The Empirical Hierarchical Bayes Approach for Pathway Integration and Gene-Environment Interactions in Genome-Wide Association Studies

Dissertation
zur Erlangung des Doktorgrades
der Mathematisch-Naturwissenschaftlichen Fakultäten
der Georg-August-Universität zu Göttingen

vorgelegt von
Melanie Sohns
aus Prüm

Göttingen, 2012

Referent: Prof. Dr. Martin Schlather
Koreferentin: Prof. Dr. Heike Bickeböller
Tag der mündlichen Prüfung:

Danksagung

Mein besonderer Dank gilt Frau Prof. Dr. Heike Bickeböller für die umfassende und engagierte wissenschaftliche Betreuung bei der Entstehung dieser Arbeit. Außerdem möchte ich mich bei Ihr bedanken für die Möglichkeit, neben der Dissertation an verschiedenen spannenden Forschungsprojekten in der Genetischen Epidemiologie mitarbeiten und so wertvolle praktische Erfahrung sammeln zu können. Darüber hinaus hat sie mir die Teilnahme an zahlreichen Workshops und Tagungen ermöglicht, sowie einen Forschungsaufenthalt an der University of Southern California, Los Angeles, USA, der entscheidend zur Entstehung dieser Arbeit beigetragen hat. In diesem Zusammenhang danke ich auch Duncan Thomas und Juan Pablo Lewinger recht herzlich für ihre Betreuung in dieser Zeit, sowie dem Deutschen Akademischen Austauschdienst (DAAD) für die finanzielle Unterstützung.
Herrn Prof. Schlather danke ich für die freundliche Übernahme des Erstgutachtens und Unterstützung im Rahmen des Promotionsstudiengangs "Angewandte Statistik und Empirische Methoden".

Ohne die Nutzung des Hochleistungsrechenclusters finanziert duch das Bundesministerium für Bildung und Forschung (BMBF) (Services@MediGrid-Projekt, Förderkennzeichen 01IG07015A) und betrieben durch den Geschäftsbereich IT der Universitätsmedizin Göttingen wären die Simulationsstudien meiner Arbeit sowie die Datenanwendungen in dieser Form nicht möglich gewesen. Daher danke ich auch diesen recht herzlich.
Zudem möchte ich mich beim Genetic Analysis Workshop (GAW) 16 sowie dem International Lung Cancer Consortium (ILCCO) und der Arbeitsgruppe Transdisciplinary Research in Cancer of the Lung (TRICL) für die Möglichkeit bedanken, die genomweiten Datensätze zur Rheumatoiden Arthritis sowie zum Lungenkrebs für meine Dissertation zu verwenden. Insbesondere gilt mein Dank dabei den Leitern der Lungenkrebsstudien Christopher Amos, Paul Brennan, Rayjean Hung und Heinz-Erich Wichman, sowie Gord Fehringer und Younghun Han für die gute Zusammenarbeit.
GAW wird finanziert durch das NIH Grant R01 GM031575 des National Institute of General Medical Sciences, TRICL durch das NIH Grant 1U19CA148127-01.
Die Arbeit wurde darüber hinaus teilweise durch das Bundesministerium für Bildung und Forschung (BMBF) im Rahmen des nationalen Genomforschungsnetz plus gefördert (Förderkennzeichen: 01GS0837).

Ein großes Dankeschön geht weiterhin an meine lieben Kollegen aus der Genetischen Epidemiologie und Medizinischen Statistik für die hervorragende Zusammenarbeit, die vielen kleinen und großen Hilfen im Alltag und die stets freundliche, angenehme Arbeitsatmosphäre. Mein Dank gilt dabei insbesondere Albert Rosenberger, der mir bei Fragen stets hilfreich war und Elena Viktorova für das sorgfältige Lesen meine Arbeit.
Schließlich möchte ich ganz herzlich meinem Schatzi, meinen Eltern und Freunden danken, die mich auf meinem Weg stets unterstützt haben, mir Kraft gegeben haben und für mich da waren.

Contents

1 Introduction

The investigation of complex diseases such as cancer, cardiovascular diseases, diabetes, rheumatoid arthritis, allergies or Alzheimer's disease is of high importance for public health and economy because of their widespread in modern western populations. Cardiovascular and cancer diseases are the main causes of death in such countries as Germany and the USA, contributing 42% and 26%, respectively, to the German overall mortality in 2008 (Robert Koch-Institut, 2011). Other common chronic diseases such as allergies or diabetes affect a large proportion of young people and lead to enormous medical and economical costs. Hence, it is important to understand complex diseases by detecting pathogenic mechanisms that cause disease development and progress. This will help to derive risk prediction models, preventive methods, medications and therapies.

In general, a disease is defined as an abnormal medical condition affecting the body of an organism, associated with specific symptoms and signs (Saunders Company, 1968). Many diseases, particularly most types of cancer, heart diseases and allergies, develop due to an internal dysfunction in the human body, e.g. in immune response or inflammatory process. These dysfunctions can arise from non-genetic factors that are in Genetic Epidemiology denoted as "environment", compassing lifestyle, external exposures and therapies, but they may also be caused partly or completely by genetic factors. For the latter we can differentiate the hereditary disposition transmitted from the parents and occurring in all body cells from changes in the genetic information occurring during lifetime affecting only some cells and their descendants. The latter is especially relevant in cancer.

The relation of genetic factors to disease development arises from the fact that the genetic information codes for proteins and regulates their synthesis. Proteins are the basic molecules of life and are responsible for all necessary tasks of the human body, e.g. metabolism, signal translation or regulation of cell growth. On the one hand, proteins preserve life. On the other hand wrong or defect proteins appearing due to an improper protein coding can be responsible for disease susceptibility and development as well. The same holds for an insufficient or excessive amount of proteins due to improper regulation of protein synthesis.

When examining the influence of genetic factors on disease development, we should differentiate between classical genetic disorders and complex diseases. Classical genetic disorders, called Mendelian diseases, are rare with a simple inheritance pattern in affected families. They are determined by a single gene only (monogenic diseases), e.g. Huntington's disease, Cystic fibrosis or red green color blindness (Bickeböller and Fischer, 2007). For such diseases, the defect or the loss of only one specific protein or the construction of one wrong gene product directly causes the disease development and provides a unique relationship between the genetic factor and the disease. In contrast, complex diseases are characterized by the absence of clear inheritance patterns and often by no obvious aggregation mechanism in families, resulting from a complicated interplay of numerous genetic and environmental factors. Most complex diseases do not show evidence for the presence of clear genetic causation, but rather a genetic sensibility to the disease given by multiple genetic factors - with an additional strong environmental component that leads to an ambiguous relation between the genetic

make-up and the disease of interest (Office of Genetics and Disease Prevention, 2000). Proteins in general do not work in isolation but together to fulfill the different biological processes of the human body. Therefore, it is assumed that for complex diseases whole biological pathways or complex molecular networks involving multiple interrelating and competing pathways are implicated in disease susceptibility and progression (Elbers *et al.*, 2009; Schadt, 2009; Thomas, 2005; Wang *et al.*, 2007). This implies that the genes involved in disease etiology will be functionally related and that the corresponding proteins cluster in several pathways, acting in concert to confer disease predisposition (Carlborg and Haley, 2004; Elbers *et al.*, 2009; Subramanian *et al.*, 2005; Wang *et al.*, 2010). The pathological mechanism of these diseases is not based on the defect or loss of a single gene product, but on multiple proteins altering the flux through a particular pathway, finally resulting in its malfunction or drop out (Subramanian *et al.*, 2005).
Beside the perhaps dozens of gene products a pathway comprises, environmental substrates can be included in biological processes (Thomas, 2005). A lack or excess of an environmental factor or the intervention of an improper environmental substrate can lead to pathway defects and furthermore to diseases. In allergies for example, the immune system reacts hypersensitive to harmless environmental substances called allergens. The environmental factor plays an important role, since an allergic disease becomes only noticeable in the presence of the allergen. Other examples of environmental factors contributing to a multitude of diseases such as diabetes, cardiovascular and cancer diseases are poor nutrition, lack of physical activity and smoking, where the latter is the main cause for lung cancer. The important role of environmental factors in complex diseases must not be neglected. The understanding of the underlying pathway, involving genetic and environmental factors is essential to counteract diseases and thus is an important research topic.
While Epidemiology concentrates on the investigation of environmental factors in diseases, the discipline that is engaged in finding internal risk factors in form of genetic predisposing factors is called Genetic Epidemiology. In Genetic Epidemiology, genetic markers are analyzed to identify variants in DNA sequence related to a disease of interest. These findings open insights into the pathological mechanism of the disease, can be used to determine disease risk models and develop new therapies. The direct examination of proteins is often inappropriate since their occurrence differs between tissues and proteins are unstable underlying synthesis and degradation at all times. In contrast, the genetic information appears to be stable and covers not only protein coding regions but also sequences responsible for proper regulation of protein biosynthesis. Therefore, the knowledge obtained by examining the genetic information directly instead of working on the protein level is advantageous.
For the identification of genetic risk factors, two different principles can be used: linkage and association. In linkage studies the cosegregation of genetic markers with the disease of interest in families is examined, resulting in a coarse candidate region on a particular chromosome. Genetic association studies investigate the joint occurrence of particular genetic variants with the disease either on a family or population level, allowing a fine mapping of the disease causing locus. The foundation for the performance of linkage studies in humans was proposed by Botstein and colleagues in 1980. They suggested that restriction enzymes could be used to obtain DNA sequence variants, characterized by a variation in the length of the produced fragments (restriction fragment-length

polymorphisms, RFLPs), that could be used to examine disease causing genetic factors (Botstein *et al.*, 1980; Maresso and Broeckel, 2008). In the following years, various other types of genetic variants in humans, e.g. microsatellites, were discovered. In the 1990s, by the development of high throughput genotyping methods determining standard marker sets of 200-800 microsatellite polymorphisms to cover the whole genome, linkage studies became available on a genome-wide level (Borecki and Province, 2008; Maresso and Broeckel, 2008; Sham and Cherny, 2010).

For Mendelian diseases characterized by a large effect of one single gene and an unambiguous relationship of this genetic factor and the disease, linkage studies were successfully performed to detect the underlying genetic risk factors. Two early examples are cystic fibrosis resulting from a defect in the gene CFTR (Cystic Fibrosis Transmembrane Conductance Regulator) located on chromosome 7 (Riordan *et al.*, 1989) and Huntington's disease caused by a gene coding for the protein huntingtin on chromosome 4 (The Huntington's Disease Collaborative Research Group, 1993).

Nevertheless, to discover the genetic risk factors of complex diseases provides a major challenge with only small success before the 21th century. Reason for that was the complexity of these diseases incorporating an unknown number of multiple genes with often moderate to low effects interacting with various environmental factors (Smith *et al.*, 2005). Although linkage studies in families are successful to identify the rare genetic variants of monogenic diseases, they seldomly have enough power to detect susceptibility genes with low or moderate effects, with the complicated interplay of numerous factors exacerbating the identification in addition. However, it is possible to find genetic markers for clear disease subformes that have their origin in only one single mutant gene (major gene) with a strong effect and are transmitted by a simple inheritance pattern with characteristic transmission comparable to Mendelian diseases (Scheuner *et al.*, 2004). These monogenic subtypes of diseases are often characterized by early age of onset in affected families and sometimes more severe clinical manifestations. The most famous gene belonging to this class is the BRCA1 gene on chromosome 17 identified for breast cancer by Hall *et al.*. It plays an important role in DNA repair and cell cycle control, and increases the breast cancer risk of mutation carriers during lifetime to nearly 65% (Antoniou *et al.*, 2003). Furthermore, it contributes to other types of cancer such as ovarian, prostate, pancreatic and colon cancer (Hall *et al.*, 1990; Online Mendelian Inheritance in Man (OMIM), 2012 #113705). Another example is Alzheimer's disease, with 3 causal subtype genes detected for early onset, a gene called APP coding for the amyloid precursor protein on chromosome 21 (Tanzi *et al.*, 1987; OMIM, 2012 #104300, #104760), the presenilin-1 gene on chromosome 14 (Clark *et al.*, 1996; OMIM, 2012 #607822, #104311) identified by linkage studies, and the presenilin-2 gene on chromosome 1 detected by a sequence comparison with presenilin-1 (Sherrington *et al.*, 1995; OMIM, 2012 #606889, #600759). However, this kind of disease subtypes is responsible only for a small fraction of the diseased individuals. Unfortunately, for identifying other non high-risk genes related to the disease, linkage methods proved to be unsuccessful, so that the genetic mechanisms of the remaining majority of the complex diseases remained unclear (Sham and Cherny, 2010).

In the late 1990s in response to this unsatisfying progress in studies of complex diseases by linkage analysis, the era of genetic population-based association studies started (Risch and Merikangas, 1996; Sham and Cherny, 2010). Although association

studies are not able to find rare variants in families for Mendelian disease in contrast to linkage studies, they provide much higher power to reveal common disease risk factors with moderate and low effects as predominantly involved in complex diseases (Sham and Cherny, 2010). Historically, association studies were only applicable in a candidate approach, restricted to a selection of a small number of candidate genes, regions or pathways. These candidates were derived from biological knowledge about the disease development or statistical hypotheses from previous, e.g. linkage, studies (Zondervan, 2010). Hence, association studies required a good choice of candidate genes to be performed successful. For example, the APOE gene on chromosome 19, coding for the Apolioprotein E that plays an important role in the lipid metabolism, was detected by Strittmatter *et al.* in 1991 in an association study of Alzheimer's disease characterized by late onset. APOE was replicated in several subsequent studies. Although the gene itself contributes to the risk of Alzheimer's disease only moderately in comparison to the effects of the genes in the monogenic subformes, it is responsible for many affected individuals because of the common occurrence of the risk increasing variant in the population (nearly 15%) (Bickeböller and Fischer, 2007; OMIM, 2012 #104310, #107741). Another gene, the *TP53* (*tumor protein p53*) was found in multiple association studies of different cancer diseases, compassing breast, cervical, endometrial, head and neck, lung and ovarian cancer (Hirschhorn *et al.*, 2002). The *TP53* is a tumor suppressor gene. It controls cell growth by inducing cell cycle arrest when DNA is damaged, activates DNA repair and initiates programmed cell death if irreparable DNA damages occur. The genetic variation contributing to disease risk enables cell division despite DNA damages, leading to uncontrolled cell growth and tumor formation.

Many susceptibility genes were revealed in candidate association studies of complex diseases, with more than several hundreds of associations found in works published between 1986 and 2000. However only for few of them successful replication was possible (Hirschhorn *et al.*, 2002). The lack of biological knowledge about many complex diseases and hence about potential pathways and genes (Zondervan, 2010) limited the ability to examine good candidates. New candidate regions could not be discovered by linkage methods either, because of the moderate to low effects of the genetic factors in the diseases of interest. Thus, the chance of missing genes that were not expected to be involved in etiology of a particular disease was very high.

Generally the success to unveil the etiology of complex disease in large parts remained limited due the lack of good candidates for association studies and due to the low power of linkage methods to find susceptibility moderate and low effect genes (Sham and Cherny, 2010; Zondervan, 2010). At the beginning of the 21st century a new, promising approach was introduced. Increasing knowledge about the human genome from the Hap Map (International HapMap Consortium, 2003, 2005) and the human genome project (International Human Genome Sequencing Consortium, 2004), as well as the technological progress in developing chips of genetic markers covering nearly the whole genome, made it possible to carry out genome-wide association studies. Genetic markers that are allocated on these genome-wide chips are single-nucleotide-polymorphisms (SNPs) - DNA sequence variations resulting from a change of a single DNA base. The new strategy of genome-wide association studies (GWAS) seemed to fulfill the needs for examining complex diseases, and it expressed a new ray of hope to reveal pathological

mechanisms of the diseases (Sham and Cherny, 2010; Zondervan, 2010). The idea of this new methodology is supported by the common disease common variant hypothesis (CDCV) (Pritchard and Cox, 2002). The CDCV states that the genetic burden of a complex disease can be conveyed by common variants, since variants influencing complex diseases harm people only later in life time, usually after reproductive years, and therefore not eliminated by natural selection (Stranger *et al.*, 2011). Common variants are defined as DNA variants that occur for at least 1% in a population (Frazer *et al.*, 2009). This hypothesis was one of the fundamentals of the Hap Map project, where the patterns of common genetic variations in different populations were characterized and provided for the chip-technology of GWAS to facilitate the genotyping of a huge number of SNPs at reasonable costs.

Although GWAS initially provided many new challenges, the first genome-wide association study of age-related macular degeneration performed in 2005 showed success and presented a promising start by identifying CFH (complement factor H) (Klein *et al.*, 2005) among 100,000 genotyped SNPs in only 96 cases and 50 controls. Nowadays, many of the initial problems have been solved. However researchers are still struggling with new issues resulting from GWAS and developing corresponding methods. At the beginning of the GWAS era, two step (Bukszár and van den Oord, 2006; Satagopan *et al.*, 2002; Skol *et al.*, 2006; Thomas *et al.*, 2004) and DNA pooling methods (Sham *et al.*, 2002) were of high interest promising to reduce genotyping costs. Due to decreasing chip expenses they lost attractiveness over the years. Availability of increased computer power and the help from computer sciences made the handling of huge amount of data possible. To guarantee high quality of the genome-wide data, different quality control criteria had to be assessed, with nowadays nearly consensus found about this issue. Methods from other disciplines were borrowed and adapted to solve such difficulties as multiple testing (Dudoit and Laan, 2008; Rice *et al.*, 2008; Westfall and Young, 1993) and meta-analyses (Trikalinos *et al.*, 2008). Several new methods were developed for new highly important challenges that specially arise in genome-wide association studies such as population stratification (Devlin and Roeder, 1999; Price *et al.*, 2006; Pritchard *et al.*, 2000) or imputation (Browning and Browning, 2009; Li *et al.*, 2009; Marchini *et al.*, 2007). General GWAS software and packages were created (Aulchenko *et al.*, 2007; Herold *et al.*, 2009; Purcell *et al.*, 2007), providing the main methods for quality control and analysis of GWA data with an efficient time and memory consumption. Specific software addressing the special issues was developed as well, e.g. EIGENSTRAT (Price *et al.*, 2006) or MACH (Li and Wang, 2010).

Numerous successful GWAS were performed, with especially the investigations of the Wellcome Trust Case Control Consortium (WTCCC) worth to mention. The WTCCC analyzed 500,000 genetic markers for 7 common diseases within 1,500 – 2,000 cases for each disease and 3,000 shared controls (Wellcome Trust Case Control Consortium, 2007). Until October 2010, 702 GWAS in humans were published, involving 421 different human traits with several hundreds of genetic markers replicated (Johnson and O'Donnell, 2009; Hindorff *et al.*, 2009, 2012; Stranger *et al.*, 2011). Nevertheless, for many complex diseases GWAS reached their limits. Although many genetic susceptibility loci have been reported so far, many of them were not replicated. Furthermore, in replicated findings, the effects are often weak and explain only a small proportion of the disease, so that the medical relevance of the results remains small (Gibson, 2010;

Ioannidis, 2007; Ioannidis *et al.*, 2007; Janssens and van Duijn, 2010; Manolio *et al.*, 2009).
During the last years, this partially unsatisfying progress gave rise to the thought, that GWAS involving only the analysis of single genes with common variants are not as sufficient as expected and mark only one step along the road. It is necessary to strike new complementary paths, e.g. compassing collaborative work, analysis of gene x gene (GxG) and gene x environment (GxE) interactions, consideration of pathways in the analysis and examination of other kinds of genetic markers not covered in current GWAS (Juran and Lazaridis, 2011; Gibson, 2010; Manolio *et al.*, 2009; Ober and Vercelli, 2011; Park *et al.*, 2008; Yang *et al.*, 2010).
By forming large consortia and working together closely, sample sizes are enlarged and an increase in power to find genetic components with only small effects is achieved (Ingelsson, 2010). The collaboration assures consistent analyses for the different participating studies, which can improve meta-analysis results further. Since several genes are found to be responsible for multiple diseases, e.g. TP53 for numerous cancer diseases, approaches that look at multiple phenotypes at once are of interest (Park *et al.*, 2011). Currently, special emphasis is placed on examining rare variants according to the common disease/rare variant hypothesis (CDRV) (Asimit and Zeggini, 2010; Basu and Pan, 2011; Dering *et al.*, 2011; Manolio *et al.*, 2009; Sun *et al.*, 2011), that opposites to the common disease/common variant hypothesis (CDCV) underlying the GWAS concept. The CDRV hypothesis postulates that common disease are rather caused by a high number of rare variants with high effects, what seems more consistent with human pathologies and population biology than the CDCV (Pritchard, 2001). Rare variants are defined by a frequency of less than 1% in a population (Frazer *et al.*, 2009). They are investigated in the ongoing 1,000 genomes project (1000 Genomes Project Consortium, 2010), where nearly 2,500 genomes are completely sequenced. Next generation sequencing will cover the whole genetic variation of a population. This will comprise not only single nucleotide changes in form of SNPs as considered in GWAS, but also structural variations. In addition, since the complexity of disease development cannot be neglected, including this complexity into the analysis gains importance. This is e.g. done by incorporating knowledge about biological pathways into the analysis (Chasman, 2008; Wang *et al.*, 2007) to relate several genes coding for proteins that work together in the same pathway, so that analysis results of single gene analyses can be improved. The examination of gene x gene (GxG) and gene x environment interactions (GxE) is another important point that gains attraction as a good complement to simple single marker analyses (Moore, 2003; Moore and Williams, 2005; Thomas, 2010a,b). Furthermore, haplotypes are considered (Liu *et al.*, 2008). Haplotypes encompass several genetic markers originating from the same parent at once.

The focus of this thesis is the integration of pathway information into the analysis of genome-wide association studies and the examination of gene x environment interactions to complement the simple single SNP results. We adapted and improved for our purpose a hierarchical Bayes model originally proposed by Lewinger *et al.* in 2007 for integrating external knowledge into genome-wide association studies.
In the last few years, the consideration of pathway information in GWAS was mainly performed by genes set analysis (GSA) methods (Chasman, 2008; De la Cruz *et al.*,

2010; Hosack *et al.*, 2003; Tintle *et al.*, 2009b; Wang *et al.*, 2007) originating from gene expression analyses. These methods assign significance to whole sets of genes or biological pathways, rather than single genetic marker, so that whole pathways contributing to pathological mechanism can be identified. In contrast, the hierarchical Bayes method (Chen and Witte, 2007; Heron *et al.*, 2011; Hung *et al.*, 2004; Lebrec *et al.*, 2009; Sohns *et al.*, 2009) concentrates on using the pathway information to relate the different genes to each other. Thereby, genetic markers in the same pathway can be prioritized by supporting each other to be detected. This helps to reveal the full spectrum of genes influencing the disease. Beside, the Bayesian approach provides the possibility to consider any other external knowledge in addition to the pathways, e.g. if the genetic marker directly results in a change of the corresponding protein or if the marker was found in another study before. For GSA methods, this is not possible. When integrating pathway information, we will not only focus on pathways expected in disease etiology, but allow a global overall search by integrating the whole available set of pathway knowledge.

Furthermore, GxE interactions play an important role in complex diseases and their consideration can improve results (Thomas, 2010a,b), especially in diseases such as lung cancer where the environmental factor smoking is known to have such a great impact on disease development. Interaction of this particular environmental factor with genetic factors, for example, could explain why some individuals who smoked during their whole life do not develop lung cancer, while some never smokers get affected by the disease.

GxE interaction can be investigated by a logistic regression model that includes a corresponding regression coefficient for the interaction term. The traditional case-control test is based on the estimation of this coefficient. Unfortunately, this classical test usually has low power to detect GxE interactions. Hence, the case-only approach, based on diseased individuals only, was suggested by Piegorsch *et al.* in 1994. It results in increased power but has one major drawback: the test is biased and leads to false positive results in the presence of an underlying G-E association on a population level independent of the disease of interest. Such population-based G-E associations can for example occur when genes influence the choice of an environmental factor, e.g. in lung cancer gene that favor smoking, but are not involved in the disease development themselves. Unfortunately, G-E associations cannot be ruled out. They are even expected to appear, especially in genome-wide context, where up to two million SNPs are tested. Therefore, during the last years, several methods were developed, trying to increase the power in finding GxE while taking G-E associations on a population level into account, e.g. two-step procedures (Albert *et al.*, 2001; Murcray *et al.*, 2009) testing first for a population-based G-E association and then in the second step for the interaction, or by empirical Bayes methods (Mukherjee *et al.*, 2008; Mukherjee and Chatterjee, 2008).

We modified and improved the hierarchical Bayes model of Lewinger *et al.* (2007) for the purpose of GxE analysis. This newly developed GxE test exploits the high power of the case-only test while considering population-based G-E associations. We worked out two strategies to combine the integration of pathway information and the analysis of GxE interactions. The first strategy integrates the available pathway information into the analysis to support markers that have only a minor interaction effect based on the

case-control test for interaction but occur in the same pathway. In the second method integrating pathway information with GxE interactions, we consider only pathways with a known or highly expected relation to the considered environmental factor. These were included into the analysis to support the correct control for population-based G-E associations, since SNPs involved in such an environment associated pathway should rather have a population-based G-E association then SNPs outside of such a pathway. For example, smoking pathways related to nicotine dependency would belong to that category.

This dissertation starts with three introductory chapters providing the necessary basic genetic and statistical concepts. In chapter 2, basic information about the human genome, population genetics and genetic diseases is given. Chapter 3 includes the statistical basics used in Genetic Epidemiology and principles of association studies. In particular, genome-wide association studies and GxE interactions are considered. Chapter 4 introduces the Bayesian approach and specifically the empirical Bayes approach as the statistical basic concept for the method we used. The hierarchical Bayes approach suggested by (Lewinger *et al.*, 2007) for genome-wide association studies, denoted as hierarchical Bayes prioritization (HBP), is discussed in the same chapter. The fifth chapter is about the integration of pathway information into genome-wide association studies. Different gene set analysis methods are presented (Chasman, 2008; De la Cruz *et al.*, 2010; Hosack *et al.*, 2003; Tintle *et al.*, 2009b; Wang *et al.*, 2007) and the comparison of the hierarchical Bayes prioritization using several strategies integrating pathway information to other gene set methods based on rheumatoid arthritis data is discussed (Lebrec *et al.*, 2009; Sohns *et al.*, 2009). Chapter 6 focuses on GxE interactions in GWAs. Different GxE interaction methods are explained (Albert *et al.*, 2001; Mukherjee *et al.*, 2008; Mukherjee and Chatterjee, 2008; Murcray *et al.*, 2009) and an improved statistical method for GxE in GWAs based on the hierarchical Bayes approach is provided. Simulation studies are presented, investigating the performance of this new method in comparison to other existing GxE approaches. In chapter 7 the hierarchical Bayes method for pathway integration, the hierarchical Bayes method for detection of GxE interactions and two strategies incorporating pathway information into the analysis of GxE interactions are applied to several lung cancer studies from the international lung cancer consortium (ILCCO) and the working group on transdisciplinary research in cancer of the lung (TRICL) (International Agency for Research on Cancer (IARC), 2012; Amos, 2007). For comparison purpose, Gene Set Enrichment Analysis (Subramanian *et al.*, 2005; Wang *et al.*, 2007), the most popular gene set analysis method, and several GxE approaches (Albert *et al.*, 2001; Mukherjee and Chatterjee, 2008; Murcray *et al.*, 2009; Piegorsch *et al.*, 1994) are applied to the same data. The last chapter gives a short summary and contains an outlook for further investigations to extend and improve this work.
Since chapter 2 and 3 are restricted to the basics necessary for the mathematically focused reader, more detailed information for several topics in genetics, genetic diseases and genome-wide association studies is given in the appendix part A for the molecular genetic comprehension and interpretation of the applications. In the appendix part B additional information and results for our data applications can be found. Finally, mathematical derivatives for the empirical hierarchical Bayes approach for GxE interaction are given in the appendix part C.

2 Fundamentals of genetics and genetic diseases

2.1 Genetic basics

"DNA makes RNA
RNA makes proteins
proteins make us"
(Ziegler and König, 2006)

2.1.1 The hereditary information

The **genome** is the entirety of the inheritable information of an organism that is necessary for its development and the specification of characteristics, biological features and traits. In organisms with cell nuclei (eukaryotes) the main part of the hereditary information is located in the nucleus and organized in separate physical units, the **chromosomes**, which build the control center of each cell. Human cells contain 23 pairs of chromosomes including 22 autosomal pairs (**autosomes**) and 1 pair of sex-chromosomes. The two copies of each pair are called **homologous** chromosomes because they have - except for the sex chromosomes - the same length and structure and are responsible for the same biological features. For the sex chromosomes, two different forms exist, the X and the Y chromosome, determining the sex of an individual with an XX pair in females and an XY pair in males.
Chromosomes consist of **deoxyribonucleic acid (DNA)** as carrier of the genetic information. A graphical presentation of the DNA is given in figure 2.1. The DNA is composed of two long linear molecules (strands) of several individual elements called **nucleotides** that form a **double helix** structure. Four different types of nucleotides occur, containing one of the bases **adenine** (A), **cytosine** (C), **guanine** (G) or **thymine** (T). Each DNA strand has two different ends, the 3' and 5' end, and the bases between the two strands form pairs by binding A to T and C to G, so that the DNA has two complementary base sequences. In total, approximately $3 \cdot 10^9$ base pairs occur in the human genome (U.S. National Library of Medicine, 2011).
The functional units of the DNA are called **genes**. They cover the genetic information by containing blueprints for **protein** construction coded by their base sequence (**genetic code**). More precisely the base sequence of a gene codes for **amino acids**, which furthermore combine to specific proteins of particular function. Each of the 20 existing amino acids is coded by 3 successive bases (**codon**) with several different codings for some of them. Additionally, there is one start codon and three stop codons that mark the beginning and the end of an amino acid sequence.

2.1.2 The synthesis of proteins

The biosynthesis of proteins using the genetic information in form of DNA gene-codes is called **gene expression**. The DNA-sequence is first transcribed to **mRNA**, **messenger ribonucleic acid (transcription)**, while the mRNA sequence is translated to a chain of amino acids that build the protein (**translation**). mRNA is only one-stranded and differs from DNA by substituting thymine with uracil (U; bounds with A)

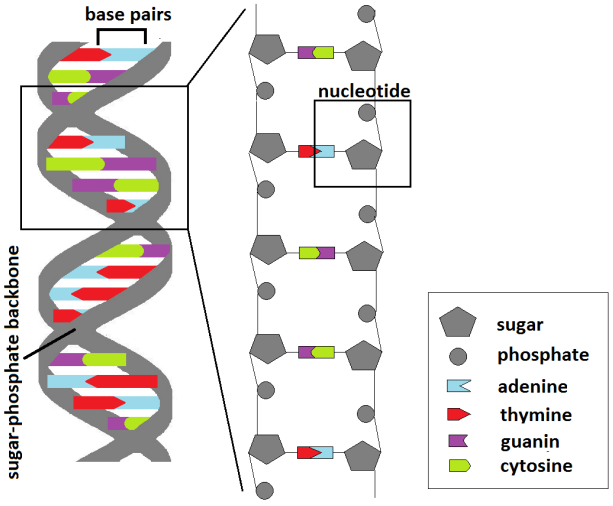


Figure 2.1: *DNA double helix and its composition*

and containing a ribose instead of a desoxyribose in each nucleotide. In addition to the DNA regions of a gene containing the information that is translated into an amino acid sequence, called **exons**, there exist intragenic regions without coding function for the protein, denoted as **introns**. At both ends of the gene we have **untranslated regions (UTR)** that can contain regulatory elements. An overview of the different components around and within a gene is given in figure 2.2.
Before translation, the introns are removed from the mRNA sequence in a process called **splicing**. By **alternative splicing** different mRNA molecules can be obtained from the same DNA sequence. Thus, one gene can code for different proteins, and the number of possible proteins clearly exceeds the number of genes. For humans 20,000 – 25,000 different genes exist (International Human Genome Sequencing Consortium, 2004), coding for more than 300,000 different proteins (Qiagen Sample and Assay Technology, 2012), that make us who we are and how we look like.
Although each single body cell contains the whole genetic information, the gene activity differs depending on the particular cell type and current need. This effectiveness of biosynthesis is guaranteed by regulatory DNA sequences located in the UTR or 3' and 5' flanking regions of a gene (gene regulation). These regulary units are furthermore controlled by the specific interplay with numerous **transcription factors**. Transcription factors are special proteins that can activate (activator) or block (repressor) the regulatory units and hence enable or inhibit the transcription. Other regulatory elements can influence the translation, e.g. by promoting or enhancing the mRNA degradation and hence determining how often the same mRNA is translated into protein.
Proteins can contain one or more amino acid chains, each comprising hundreds to several thousand amino acids. Functionality and challenges of a protein are determined by the sequence of the amino acids. Proteins are responsible for all tasks concerning sustainment and function of the human body, e.g. for the transport of substances, the regulation of ion concentrations, the catalyzation of chemical reactions or infection defense, with many proteins acting together for the different tasks. Every moment, thousands of proteins are produced.

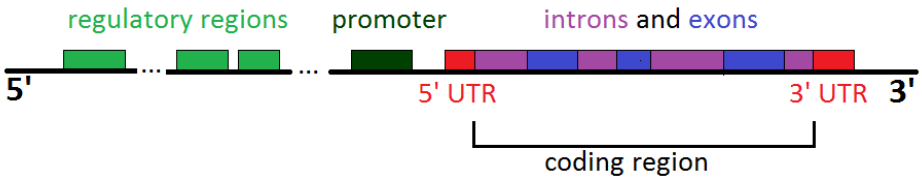


Figure 2.2: *Structure of a gene*

The protein-coding gene segments cover only around 1.2% of the DNA (International Human Genome Sequencing Consortium, 2004), non-coding introns and UTRs within the genes nearly 35%. Beside, we have about 62% intergenic regions without coding function (Brown, 2002). These can partly comprise the already mentioned regulatory units responsible for the control of the gene activity by regulating if and to which amount genes are transcribed and translated into proteins. However, for the main part of the intergenic regions, no function is known at all.

2.1.3 Genetic variability

An important process affecting the hereditary information is the cell division. Two different kinds of cell division exist, the **mitosis** which is necessary for the progeny of cells to achieve new somatic cells for growth, development and wound healing and the **meiosis**, necessary for sexual reproduction. Both types of cell division incorporate **DNA replication**: the mechanism of duplicating DNA. In this process, the double helix structure is uncoiled and each of the single strands serves as a model for a new complementary strand, so that two exact copies of the original DNA double helix are produced.
In the basic cell division, the **mitosis**, all chromosomes of a cell are duplicated and equally distributed to two daughter cells. Thereby, all body cells are identical clones of the original fertilized ovum, adapted to their special tasks by structure and function.
In the **meiosis**, a special form of the cell division, **germ cells** (gametes: egg and sperm cells) are produced. For the sexual reproduction that accomplishes the genetic information of mother and father, the presence of only a single set of chromosomes (haploid) in the responsible cells is necessary. Hence, in meiosis, only one chromosome of each pair should be transferred to one germ cell.
Since in meiosis the homologous chromosomes are randomly distributed to two germ cells, more than 8 million (2^{23}) different chromosome combinations in gametes exist. This results in a high genetic variability, with each gamete being nearly unique. In **fertilization**, a sperm and an egg cell fuse with each other and two haploid chromosome sets are merged to a diploid set, enlarging the possible combinations of genetic material as well.
Another biological phenomenon producing genetic variety during meiosis is the chromosomal **crossover** or **crossing over**. Before the actual cell division, the chromosomes of a pair arrange next to each other and partly overlap. Breaks in the DNA strands at homologous points can occur, which can be joined together the other way round, permitting an exchange of DNA segments between the two homologous chromosomes (**chromosomal recombination**). For each chromosome pair, multiple crossovers are

possible, with on average 55 crossovers per male human cell and approximately 80 in female cells (Ziegler and König, 2006). Highly related to the process of crossing over is the principle of **genetic linkage**. Genes or DNA segments are in linkage, when they tend to be inherited together over generations by staying together in meiosis and hence not being distributed independently from each other to the gametes. Genetic linkage is also known as **cosegregation**. The joint inheritance of loci on the same chromosome can be disturbed by crossover, so that only loci physically close to each other are found to be tightly linked, since the region between them where a crossing over can take place is only small.
While the chromosome combination represents a mix of complete grand paternal and grand maternal chromosomes, the crossover mixes this information in addition within the chromosomes. The combination of genetic material from all four grandparents constitutes a natural random process, leading to a large genetic variety and genetic dissimilarity in siblings. The ability of a population to develop individuals with different inheritable information is called genetic variability.
However, meiosis is not unfailing and can –beside other causes – lead to different kinds of **mutations**. Mutations are permanent changes in the genetic makeup that occur from time to time spontaneously or by external exposure (e.g. virus, radiation, chemicals). In this thesis we will concentrate on genetic variants arising from mutations that affect only one single base position, also denoted as **point mutation**. Therefore, we will skip large scale mutations affecting whole chromosomes or large chromosome segments here and address **gene mutations** that affect only one gene. Gene mutations manifest by substitution, deletion, insertion or duplication of single bases or short base sequences. Although such small-scale mutations can result from an exchange of non-homologous sequences of a chromosome pair (unbalanced crossover), the main source is defects in DNA replication. In general, most errors in DNA replication are immediately corrected by an efficient repair mechanism. Sometimes it is not possible to detect or repair these errors and mutations result. When a mutation occurs in a somatic cell, all descendants of this particular cell in this organism will be affected. Consequences will only occur when they influence the particular function of the cell tissue (e.g. transfer from a normal body cell to a proliferating cancer cell). When a DNA modification occurs in germline, it can be inherited to the offspring and all body cells of the new developing individual will contain this mutation. Therefore, germline mutations lead to genetic variation and are important for evolution and in the context of hereditary diseases.
The mutation rate, defined as the number of mutations per generation per gamete, for a gene is given by $10^{-5} - 10^{-6}$. Gene mutations occur very often and need not necessarily implicate functional consequences. The potential function of gene mutations located in intergenic regions without regulatory function is not well understood yet. While gene mutations in regulatory regions can influence gene expression and completely turn of protein synthesis, mutations in introns can alter the gene splicing. Mutations within exons may lead to wrong or defect gene products. Mutations are an important evolutionary factor responsible for the variety of species on earth. However, since mutations can influence the regulation of gene expression and hence the amount and type of protein produced, they can also affect the protein's function. This can lead to protective effects but also disadvantageous changes in the human body, the development of diseases and even death.

2.1.4 Polymorphisms and phenotypes

Mutations are DNA sequence changes away from "normal" and all sequence variations start as a mutation. When such a variant induced by a mutation causes a disease with neo-natal or childhood onset, it may reduce the fitness of the organism and therefore stays rare. However, new mutations not negatively influencing the fitness can spread out and establish in a population, what may result in a polymorphism. A **polymorphism** denotes a DNA sequence variation that is common in the population and an "acceptable, normal" alternative for the corresponding DNA sequence, that cannot be explained by a new mutation anymore. By definition, a polymorphism is a variation in DNA sequence that occurs in at least 1% of the individuals of at least one human population. Polymorphisms are responsible for many "normal" differences between people such as eye or hair color and blood type, with some also contributing to susceptibility of certain disorders. Nevertheless, 99% of our genome is the same in all humans (U.S. National Library of Medicine, 2011).
Depending on the underlying mutation, we can distinguish different kinds of polymorphisms. **Insertion or deletion polymorphisms (INDELs)** result from insertions or deletions that contain in general less than 50 nucleotides. **Copy number variations (CNV)** vary in their number of copies of a particular DNA sequence. **Single nucleotide polymorphisms (SNPs)** result from substitutions of a single base. The latter are the most common type of genetic variation among humans, accounting for 90% of the genetic variation. Most of them show only two different variants. Their mutation rate is relatively low with approximately 10^{-9} to 10^{-8}. Although SNPs are most commonly found in intergenic regions with no explainable consequences to health so far, some variants however have proven their importance in human health studies, influencing the risk of disease development or the susceptibility to environmental factors. This affects predominantly SNPs within genes or regulatory regions (U.S. National Library of Medicine, 2011).
Alternative variants of a gene or gene sequence at one locus on a chromosome are called **alleles**. A **gene locus** is the physical position of a gene in the genome. A locus is **monomorphic** when only one allele exists, a locus is **polymorphic** given at least two different alleles. A locus with exactly two different alleles is **biallelic**. The frequency of the appearance of an allele is the **allele frequency**. The frequency of the less common allele occurring at a locus in a given population is called **minor allele frequency (MAF)**. A clearly identifiable polymorphism with a known location in the genome where the different alleles can be determined is denoted as **marker** (U.S. National Library of Medicine, 2011). Markers can be used to study the relationship between a disease and its genetic causes or e.g. to predict a person's response to certain medication. The markers most commonly used in today's genome-wide association studies are SNPs.
Because human cells are diploid, we have two alleles for each genetic locus on our autosomes. The combination of such an allele pair at a particular locus is called **genotype**. Assuming that two different alleles A and a for one locus exist, we have 3 possible genotypes: AA,Aa and aa. When the two alleles at a person's locus differ we have a **heterozygous** genotype, when both homologous chromosomes have the same allele the individual is **homozygous** at that locus. Because different alleles can lead to different composition, structure and function of a protein, the genotype influences the appearance

of an organism, comprising all morphological, biochemical, physiological, psychological and behavioral properties. All these characteristics are denoted as **phenotype**. We can differentiate continuous and discrete phenotypes, e.g. hair color, weight and cholesterol level, or an affection status (affected/unaffected) by a disease such as obesity, hypertension and cancer. In this thesis, we will concentrate on disease status and hence on binary phenotypes.

2.1.5 Mendelian laws of inheritance

In the 1860s, before the physical basics of genetic factors were known, Gregor Mendel formulated three statements about the way certain characteristics in diploid organism determined by one gene are transmitted from one generation to another. His results were based on large-scale cross-breeding experiments on pea plants. The regularities for his derived inheritance patterns were later restated to describe the relationship of genotype and phenotype and are still known as Mendelian rules. These rules are the law of uniformity and the law of segregation with respect to a single gene and the law of independence dealing with the observation of two different characteristics at once. According to the first two rules, each individual transmits one of its two alleles to the offspring randomly according to a Bernoulli distribution with a probability of $p = 1/2$, with the inheritance from father and mother independent from each other. The inheritance via sex chromosomes presents a specialized rule. Mendel introduced the terms “dominant”, “recessive” and “codominant” characterizing different modes of inheritance. To visualize the relationship of phenotype and genotype, let us consider a simple example of eye color, assuming two alleles B and G coding for green and blue eye color and the possible genotypes BB, BG and GG. Given a homozygous genotype, the eye color is unambiguously blue or green. For a heterozygous genotype, a different situation can occur. When we have a **dominant-recessive inheritance**, one of the alleles (e.g. G) establishes itself compared to the other one (e.g. B), resulting in the same phenotype as the corresponding homozygous genotype (e.g. green eyes). G is the dominant allele, while B is called recessive. Nevertheless, when both alleles establish themselves, they are codominant and we have a **codominant inheritance**, e.g. resulting in blue-green eyes. An illustrating example is the ABO-blood type, with an allele for “A” and “B” resulting in blood type AB. The third rule, the law of independence postulates that two genetic factors are transmitted totally independent from each other, so that they can combine randomly and form new combinations. However, this turned out to be true only under certain conditions. Genetic factors located close to each other on a chromosome are not independently inherited.

2.2 Population genetics

Population genetics deals with the exploration of genetic structures in populations and consequences of different evolutionary factors to the genetic constitution of a population. This includes the examination of allele and genotype frequencies on a population level, including the reasons for the observed frequencies, in which population they occur and how they behave. A **population** denotes a group of reproductive individuals of the same species that live in the same area, speak the same language and have the same

culture, connected by evolution (Bickeböller and Fischer, 2007).
In the following, let M be a biallelic autosomal marker with the two alleles A and a with frequencies $p(A)$ and $p(a) = 1 - p(A)$. For an individual I of a population such a locus can be presented by two Bernoulli distributed $B(1, p(A))$ random variables X_{MIj} for the two alleles of the homologous chromosomes $j = 1$ (paternal), 2 (maternal). When allele A occurs, $X_{MIj} = 0$ and $X_{MIj} = 1$ for allele a. The genotype of the individual can be expressed as the sum of the two random variables $X_{MI} = X_{MI1} + X_{MI2}(AA = 0, Aa = 1, aa = 2)$.

2.2.1 Hardy Weinberg Equilibrium

The law of Hardy Weinberg is an important basic principle of the population genetics of diploid organisms. It describes the relationship between allele and genotype frequencies of an autosomal locus in a population. The Hardy Weinberg law indicates that the allele and genotype frequencies are in a stable equilibrium, called **Hardy Weinberg Equilibrium (HWE)**, remaining constant from generation to generation under certain assumptions. Given the biallelic marker considered above, we can derive the frequencies for the genotypes AA $(s = 0)$, Aa $(s = 1)$ and aa $(s = 2)$ from the allele frequencies $p(A)$ and $p(a)$ by

$$P(X_{MI} = s) = \binom{2}{s} p(A)^{2-s}(1 - p(A))^s \qquad \text{with} \qquad \sum_{s=0,1,2} P(X_{MI} = s) = 1.$$

Allele frequencies can be derived from given genotype frequencies as well by $p(A) = P(X_{MI} = 0) + 0.5P(X_{MI} = 1)$ and $p(a) = P(X_{MI} = 2) + 0.5P(X_{MI} = 1)$. A more detailed derivation can be found in Bickeböller and Fischer (2007).
One of the assumptions that underlie the Hardy-Weinberg-Equilibrium is that we have an infinite population where Mendel's law of segregation holds and that all pairs of different genotype carriers for reproduction are equally likely (random mating). Infinite population in the context of population genetics means that the population is really large so that the random loss of an individual does not influence the allele frequencies. Random mating excludes inbreeding or a preferential selection of a partner due to its genetic information (assortative mating). In addition, evolutionary forces which influence the allele and genotype frequencies are assumed to not occur, such as genetic drift, natural selection, immigration or emigration, population stratification and mutation. Natural selection of a particular allele results from an advantage or disadvantage for the carrier of a specific genotype or phenotype, so that not all individuals reproduce by the same probability. In contrast genetic drift is an entirely random stochastical process which changes the allele frequencies of a population strictly by chance due to random sampling. However, the effect of genetic drift is weak in large populations and therefore only relevant in very small populations. Since mutation frequencies are usually low, they also do not play such a relevant role.
Although in general the assumptions of Hardy-Weinberg-Equilibrium are not fulfilled, the law proves useful in praxis, is widely applied and many statistical methods are based on it. For testing HWE we can use a χ^2 test, comparing the expected genotype frequencies based on allele frequencies with the observed ones. Deviation from HWE may indicate the degree of evolution and can represent mixtures of different populations. In

addition, HWE deviations can be used to detect laboratory problems such as genotyping errors that express in a disequilibrium as well.

2.2.2 Linkage Disequilibrium

Linkage Disequilibrium (LD) denotes the correlation of particular alleles at nearby loci of a chromosome on a population level due to their tendency to be inherited together (Ardlie *et al.*, 2002). In the following we will concentrate on pairwise LD measures which consider only two loci at ones.
Assume that we have the alleles A/a and B/b for the two loci M_1 and M_2, with allele frequencies $p(A)$ and $p(B)$. For an individual I, let X_{M_1I} and X_{M_2I} be the genotypes at these loci. The combinations of the two alleles from the same gamete (X_{M_1I1}, X_{M_2I1}) and (X_{M_1I2}, X_{M_2I2}) are named **haplotypes**. More general, the term haplotype is not restricted to two loci and can be extended to any number of loci, up to the whole genetic information inherited from one of the parents. For the two loci M_1 and M_2, 4 possible haplotypes can be formed: AB, Ab, aB and ab, with frequencies $p(AB), p(Ab), p(aB)$ and $p(ab)$. Linkage disequilibrium expresses itself by alleles at the two loci that occur more or less often together on gametes of a population than expected from the independent combination according to their allele frequencies. Hence, the two loci M_1 and M_2 are in **linkage equilibrium** when both alleles of a haplotype are independently distributed, that means the haplotype frequencies correspond to the product of allele frequencies: $p(AB) = p(A)p(B), p(Ab) = p(A)p(b), p(aB) = p(a)p(B)$ and $p(ab) = p(a)p(b)$. A departure from independence representing a correlation of the loci is called linkage disequilibrium.
Linkage disequilibrium can be measured by the disequilibrium coefficient $D_{AB} = p(AB) - p(A)p(B)$, which equals 0 in case of linkage equilibrium and is unequal 0 when linkage disequilibrium is present. Linkage disequilibrium is a property of loci, not their alleles and considering the other haplotypes of two loci we have $D_{AB} = D_{ab}$ and $D_{aB} = D_{Ab} = -D_{AB}$ and hence only 1 degree of freedom. Because this measure highly depends on the allele frequencies, different other measures for the strength of LD were proposed, with r^2 the recommended one that is most commonly used (Ardlie *et al.*, 2002). It corresponds to the square of the correlation coefficient of the 2x2 table of haplotype frequencies, given by

$$r^2 = D^2/(p(A)p(B)p(a)p(b)) \tag{2.1}$$

The LD measure r^2 ranges from 0 to 1 and equals 1 when two markers provide identical information. On the other hand, $r^2 = 0$ denotes a perfect equilibrium. LD between two loci can arise due to a new mutation at one of the loci resulting in a new haplotype. By crossing over events between them, the loci may be inherited to different gametes what results in a decay of LD from generation to generation. The rate of crossing over between the loci determines the degree of LD reduction. Therefore, for loci physically really close to each other, the LD reduces only slightly. The region inbetween for a possible crossing over is only small, so that the loci tend to be inherited together and the existing haplotypes remain. Another source of LD is the favoring of one of the existing alleles at a locus, denoted as selection (Bickeböller and Fischer, 2007; Gillespie, 1998; Suarez and Hampe, 1994).

LD patterns are further shaped by other evolutionary forces such as random genetic drift (changing the allele and haplotype frequencies), migration, inbreeding, population admixture and stratification (different allele frequencies) (Bickeböller and Fischer, 2007; Rao and Gu, 2008). However, population admixture and stratification can cause a correlation of loci independent from their location to each other (Bickeböller and Fischer, 2007; Ziegler and König, 2006).

2.3 Genetic origin of diseases

Although our genetic information crucially contributes to our appearance, our properties and preserves our life, it also contributes to the susceptibility to diseases. Improper protein coding or regulation by the genetic information can lead to a lack or excess of the corresponding proteins, or the occurrence of wrong or defect proteins. This in turn can cause disease or at least increase the risk to develop the disease. The better understanding of participating genes and proteins in disease development can lead to advances in abatement and healing and therefore the identification of such genetic factors is of high importance. Additionally, our genetic makeup can not only partly explain the predisposition to a disease, but also individual reactions to drugs.
In this chapter we will concentrate on the genetic origin of diseases. Therefore, we will first describe simple monogenic diseases that follow quite straightforward the laws of Mendelian segregation and are characterized by a unique gene-disease relation. In section 2.3.2, factors complicating this simple pattern will follow and sections 2.3.3 and 2.4 focus on complex diseases.

2.3.1 Classical monogenic diseases

Genetic causes are easily determined for **classical monogenic** or **Mendelian diseases**. This kind of diseases follows simple Mendelian inheritance patterns (section 2.1.5) and is caused by one gene only with penetrances nearly 0 or 1. The **penetrance** relates a genotype and a phenotype to each other and is defined as the conditional probability that a person with a particular genotype develops the phenotype of interest. For discrete phenotypes, we can express the penetrance by $f_{\text{genotype}} = P(\text{phenotype}|\text{genotype})$. When a disease causing genotype always results in the development of the disease, the conditional probability equals 1 and we have **complete penetrance**. In Mendelian diseases where no further factors with an influence to the disease exist the penetrances for the remaining genotypes equal 0. We can distinguish different Mendelian modes of inheritance of the disease: dominant, recessive and codominant (section 2.1.5). Furthermore, the location of the genetic factor plays an important role. In the following, we assume that we have a biallelic locus with normal allele A and disease causing variant a. A classical monogenic disease is called **autosomal dominant** when the influencing gene is located on one of the autosomes and only one a allele at that locus suffices to cause the disease. Expressed in penetrances, we have $f_{AA} = 0$ and $f_{Aa} = f_{aa} = 1$. When the gene lies on an autosome but two disease causing a alleles are required for disease development, we have a classical **autosomal recessive** disease. In that case, the penetrances equal $f_{AA} = f_{Aa=0}$ and $f_{aa} = 1$. Chorea Huntington is an example for an autosomal dominant disease, while cystic fibrosis follows autosomal recessive inheritance patterns.

However, since heterozygous genotypes need not necessarily express the same phenotype as one of the homozygous, a **codominant** inheritance, with each genotype showing its own phenotype, can occur as well. A particular form of codominance is an additive mode of inheritance, with each susceptibility allele at a locus equally contributing to the phenotype or disease risk. An example for a codominant inheritance is a particular point mutation in the beta-hemoglobin gene (HBB) that replaces the normal hemoglobin allele HbA by a sickle cell hemoglobin allele HbS. This results in a sickle shape of red blood cells (sickle cell disease). Sickle-shaped cells can cause pain and organ damage by blocking small blood vessels and they die prematurely (http://ghr.nlm.nih.gov/gene/HBB). However, in heterozygous carriers we have genotype HbA/HbS so that both hemoglobin types are expressed and only 25%-40% of the erythrocytes are affected by the modified sickle-cell form. Therefore these persons show only few recognizable clinical symptoms. On the contrary, in homozygous individuals with genotype HbS/HbS all red blood cells are sickle-shaped, so that in general a shortage of red blood cells (anemia) occurs and serious symptoms in further organs (sickle cell anemia). Hence, the severity of the disease differs between heterozygous and homozygous individuals. Disease causing loci can be located on the sex chromosomes as well. However, we will not handle this here since our methods are restricted to the examination of autosomal markers.
Monogenic diseases are in general rare, occurring in less than 1 out of 1000 persons. This low disease frequency can be explained due to occurrence of the disease in early childhood with severe chronic progress resulting in reduced fitness or even lethal consequences. By investigating family data, genes of classical monogenic disease can be easily detected and many are already successfully examined. Although only one gene is involved in monogenic disease, one, several or even many alleles of that gene can cause disease development.

2.3.2 Departure from simple Mendelian segregation

The model of Mendelian segregation is useful to demonstrate the principle of genetic disorders. Unfortunately, even monogenic diseases are rarely subject to such straightforward models of inheritance (Bickeböller and Fischer, 2007). Several factors exist that modify this simple pattern and make the model more complicated.
One of these issues is the deviation of penetrances from the simple 0 and 1 rule. On the one hand it is possible, that not all individuals with a specific genetic predisposition necessarily develop a corresponding phenotype, but that it establishes only in a fraction of the carriers. This effect is denoted as **reduced penetrance**. Another phenomenon concerning penetrances is **phenocopies**. This is the case when the affection occurs as well in non-carriers of the genetic disposition, ascribed by other genetic and non-genetic factors with an impact to the disease development. We observe penetrances $0 < f < 1$. Furthermore, the penetrance can vary by age, with a higher probability of disease development with older age (e.g. in cancer).
In addition, **heterogeneity** can affect the inheritance of disease. This compasses **allelic heterogeneity**, denoting that different alleles of one gene can be responsible for the same disease, and **locus heterogeneity**, meaning there can be different responsible genes for disease development. **Phenotypic heterogeneity** and **pleiotropy** is given when the same disease shows diverse clinical characteristics in different individuals, or

when one gene causes different symptoms or even different phenotypic traits. Heterogeneity can occur between different families (**intra-familial heterogeneity**), but also within families (**inter-familial heterogeneity**). A useful tool to handle heterogeneity is to homogenize study samples with respect to a disease by defining subgroups that can be examined more easily. For many diseases, e.g. cancer diseases or Alzheimer's disease, concentrating on individuals with early-onset is useful for example.
Several other complicating factors exist, e.g. anticipation, genomic imprinting, gender restriction, X-inactivation in women, germ cell and somatic cell mosaics. These factors cannot be covered with the analysis methods in this thesis and are hence not covered here.

2.3.3 Complex diseases

Most common diseases such as cancer, cardiovascular diseases, allergies or psychiatric diseases are **complex diseases** that are not directly inherited according to classical Mendelian mechanisms but characterized by a complicated interplay of numerous genetic and environmental factors (Buselmaier and Tariverdian, 1999). This complexity involves different forms of heterogeneity listed above, reduced penetrance and phenocopies, and can be complicated by additional other principles not handled in this thesis. In most complex diseases we have no strict genetic causation, but rather a genetic predisposition for the disease given by multiple genetic factors and the manifestation of the disease depends on the influence of exogene factors during lifetime. This results in a misty relationship of genotype and phenotype, with no apparent inheritance pattern and even not necessarily an obvious aggregation in families. Disease etiology can be compared to a Marshalling yard: while the direction and different possibilities to change the switches are specified by the genetic factors, the environmental exposures determine which track is taken (Buselmaier and Tariverdian, 1999).
When only a low number of genetic markers is responsible for disease development, we say that the disease is oligogenic. When a high number of disease causing loci is involved the disease is polygenic. Polygenic diseases with an additional environmental contribution are denoted as multifactorial or complex. As already mentioned in the introduction, from time to time, even for complex diseases clear Mendelian subforms with one underlying mutant gene (major gene) with a strong effect can be identified. However, since these major genes of complex diseases are extremely rare and affect only a very small part of the affected people, we concentrate on oligogenes and polygenes as well as further modifying factors. A modifying factor is defined as a factor that influences the effect of another factor.
Although these genes have a much lower penetrance than the major genes, the susceptibility gene variants occur more often and affect a larger proportion of the population. Therefore, their investigation is highly important. For Alzheimer's disease for example, several oligogenes are identified besides the major genes, e.g. the Apolipoprotein-E (OMIM, 2012 #107741,#104310). Although this gene has a much lower penetrance than the major genes with a risk increased by factor 3, this gene mutation affects 15% of the population and is responsible for 30%-50% of all Alzheimer patients (Farrer and Cupples, 1998).

Detecting "non-major" genes of complex diseases may support our understanding of the underlying pathogenic mechanisms. Furthermore, the hope for the future is that it may facilitate to derive risk prediction models, new preventive strategies and more effective therapies and medications. However, it is still a long way to get there and we will concentrate here on the first step to identify the genetic risk factors. Unfortunately, such "non-major" genetic factors in common diseases etiology are difficult to reveal due to the complex mechanism involving the high number of factors with only small effects and interactions between them.
In the following section we will take a more detailed look at the complexity of the architecture of common diseases involving numerous genetic and non-genetic factors. This complexity involves the coordinated work of genes within biological pathways, genes interacting with each other (GxG) and the environment (GxE). Since the focus of this thesis is the integration of biological pathway information into a genetic analysis and the examination of GxE interactions, we will mainly focus on these two principles and touch on the topic of GxG only shortly.

2.4 The interplay of genetic and non-genetic factors

2.4.1 Biological pathways

In general, proteins do not work in isolation, but coordinate their activities to fulfill the different biological processes of the human body (Barabási and Oltvai, 2004; Li and Agarwal, 2009). They are organized in biological pathways (Li and Agarwal, 2009) that represent sequences of complex reactions at the molecular level in living cells to accomplish biological functions (Saraiya *et al.*, 2005). These biological functions can compass for example metabolism, signal transduction, immune response, as well as DNA replication and expression or cell growth and death (Kanehisa and Goto, 2000; Kanehisa *et al.*, 2012).
In metabolic pathways (e.g. glycolysis), a substrate has to coordinately pass a sequence of chemical reactions, catalyzed by enzymes and connected via their substrates and products. In a signal transduction pathway, information (e.g. nerve impulses) is transported from one cell to another. Since proteins are the main components in biological pathways, and genes and their regulatory regions are responsible for the synthesis of the proteins, the genetic information is connected by the pathways as well.
As an example of a biological pathway we can see a representation of the p53 signaling pathway from the Kyoto Encyclopedia of Genes and Genomes (KEGG) database (Kanehisa and Goto, 2000) in figure 2.3. This pathway plays a central role in the development of all kinds of cancer and clearly illustrates how different proteins work together to fulfill a particular task. For the interested reader, a more detailed description of the pathway is given in the appendix A.3.

A defect of any of the proteins involved in a pathway can be responsible for the same final pathway malfunction, that in turn may predispose disease. Since different proteins may perform the same or a similar job, the loss of only one of these is often not relevant. Depending on where in the pathway a protein is missing, different medical and clinical consequences may results. Beyond, proteins are not only connected within pathways to fulfill the different tasks, but also by different interrelating and competing pathways that

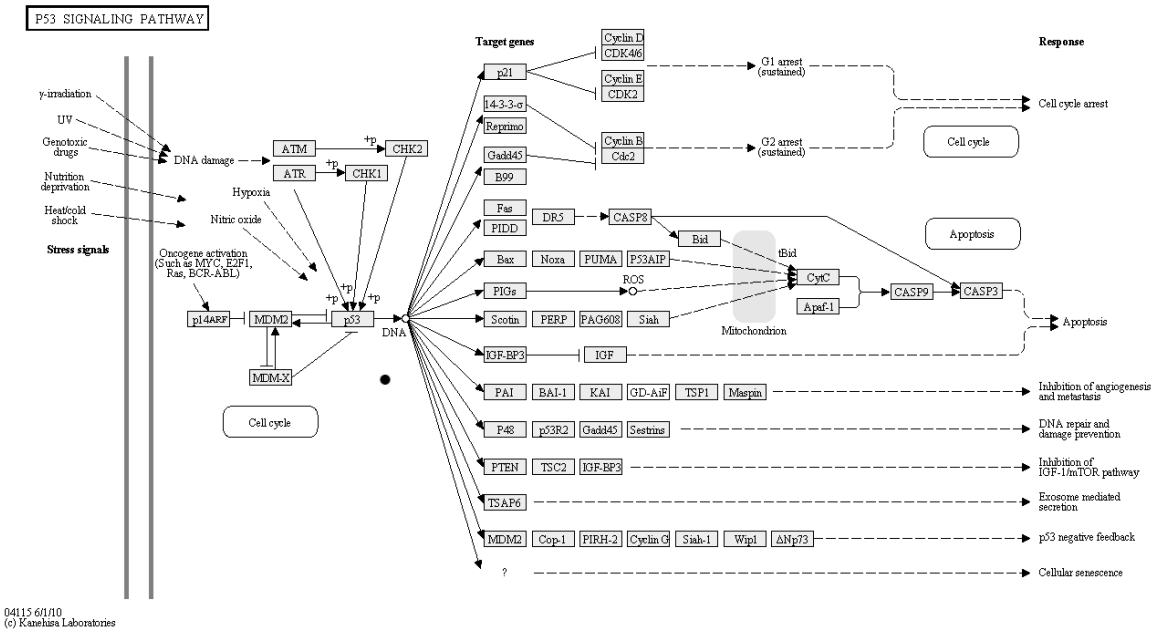


Figure 2.3: *p53 signaling pathway from the KEGG pathway database (Kanehisa and Goto, 2000)*

connect to even more complex larger networks. The p53 pathway for example is related to several other biological pathways by forwarding signals to these. Additionally, distinct alternative causal ways may lead to the development of the same disease (Brennan, 2002) and the disease complexity is enhanced by cross links between different pathways involved in different aspects of a disease. A defect of the p53 pathway alone does not cause cancer, but several other pathway defects have to accumulate in a cell and its progeny (Breuer *et al.*, 2005; Griffiths *et al.*, 2008).
To gain insights into the normal cells activities and understand the biology underlying the development of disease, it is of high importance to take a look at the relationship of the different elements to each other within a pathway or network of pathways, rather than single genes.

2.4.2 Gene x gene interactions

Beside the relationship of genes to each other defined by biological pathways, another ubiquitous component in the genetic architecture of common human diseases contributing mainly to the underlying complexity, is gene x gene interactions (GxG) (Moore, 2003; Moore and Williams, 2005). GxG interaction, also denoted as epistasis, is defined as one gene masking the effect of another gene (Cordell, 2002), so that the phenotype for a particular genotype at one locus depends on genotypes at one or more other loci (Moore and Williams, 2005).
From a biomolecular perspective, biological epistasis is defined as the result of physical interactions among biomolecules within gene regulatory networks and biochemical pathways (Moore and Williams, 2005). A protein may for example bind to another one to

modify its structure or transport it. At transcription level, transcription factors interact with DNA regulatory units, other transcription factors and further proteins enhancing or repressing their effect. Even molecules that do not directly physically interact may have epistatic masking effects if they impact the same phenotype through a hierarchy of biomolecules that affect various steps in a biochemical pathway (Chinnici, 1999). Even different genes whose products are involved in different alternative biochemical pathways may have epistatic effects (Moore and Williams, 2005).
Although the examination of individual GxG interactions is an area of research, it is not handled directly in this thesis. However, since the biological pathways relate potentially interacting genes to each other, some kinds of GxG interactions are indirectly captured by focusing on the incorporation of pathways information.

2.4.3 Environmental factors, gene x environment interactions and gene – environment associations

Biological pathways comprise not only gene products but also environmental substrates contributing to the human body functions (Thomas, 2010a,b). In the illustrated p53 pathway for example, environmental factors in form of external stress signals are responsible for the activation of the pathway.
That the development of diseases is highly driven by environmental factors as well was known long before the conduction of genetic studies (Manolio and Collins, 2007). In epidemiological studies environmental risk factors of diseases are studied with high success. The importance of the environment should not be underestimated. Any endogenous or exogenous non-genetic factor that influences the risk of disease is denoted as environmental factor (Ober and Vercelli, 2011). This involves all physical (e.g. radiation, temperature), chemical (e.g. air pollution, asbestos) and biological exposures (e.g. viruses, bacteria), as well as life events (e.g. job loss, injury), social factors (Khoury and Wacholder, 2009; Ottman, 1996; Schwartz, 2006; Vineis, 2007) and behavior patterns (e.g. habits, late age at first pregnancy) including lifestyle (e.g. diet, physical activity, stress or smoking). Therapies by drugs, hormones, chemo or radiation therapies belong to the exposures as well.
In particular, environmental substances are involved in their corresponding metabolic and signaling pathways. In a metabolic pathway, the environmental substrate, e.g. nutrients but also toxic substances, pass through a series of chemical reactions so that they are degraded and an end product is obtained. In a signal transduction pathway, a reaction to an environmental stimulus is given by a signaling cascade. Thereby all external influences are perceived, such as hearing, smelling, tasting or sensing pain.
Numerous important environmental risk factors are known so far. An influence of physical inactivity and poor nutrition compassing high fat content, few vegetables and unbalanced diet to the development of numerous diseases such as diabetes, cardiovascular diseases and cancer could be shown. In cancer it is in particular known, that environmental factors with the ability to damage the genome or disrupt cellular metabolic processes contribute majorly to the disease development. Such environmental factors are called carcinogens and encompass radiation, toxic substances as well as different infectious agents and sex hormones. Since many years, asbestos exposure and smoking are known for their high effect to lung cancer (Selikoff *et al.*, 1968).

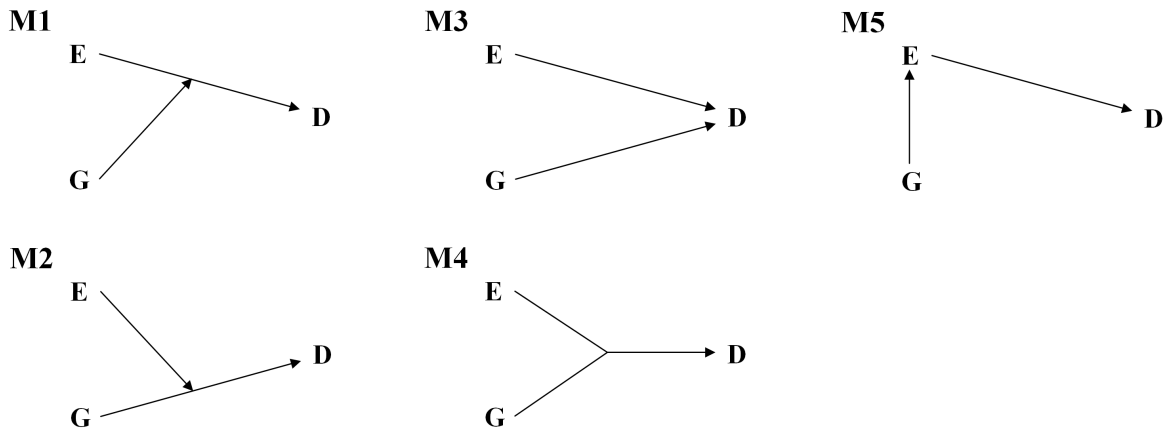


Figure 2.4: *Different models for Gene x Environment interaction (M1-M4) and Gene-Environment association (M5) according to Ottman (1990)*

The importance of environmental factors in the development of complex disease is indisputable, since the lack or excess of an environmental factor or the improper intervention of a substrate can lead to pathway defects as well as gene coding or regulation defects. In genetic epidemiological studies of complex diseases, the contribution of environmental factors (Ottman, 1990) can explain reduced penetrances as well as phenocopies. While environmental factors can have an effect of their own without any genetic predisposition, the etiology of most common disease involves not only genetic and environmental main effects, but also interactions between them (Hunter, 2005). We will concentrate in the following on GxE interaction from a biological point of view, a statistical definition is given in chapter 3.

A **GxE interaction** is given when a genetic and environmental factor work together to cause a disease (Brennan, 2002), so that the effect of the environmental factor on disease risk differs among individuals with respect to different genotypes (Brennan, 2002; Ober and Vercelli, 2011; Ottman, 1996). It is rather the sensitivity to the influence of various environmental risk factors that is inherited than the disease itself, so that differences in genetic factors cause people to respond differently to the same environmental exposure (National Institute of Environmental Health Sciences, 2011; Office of Genetics and Disease Prevention, 2000). In cancer for example, the "susceptibility" to potentially toxic compounds is heavily dependent on the efficiency with which these can be metabolized and excreted, but also on the efficiency with which small mistakes in DNA replication are repaired. This susceptibility can strongly vary between individuals of a population. The underlying susceptibility genes interact with the carcinogens. Another interpretation of GxE is that the effect of a gene varies not only with respect to the genetic background, but also by different environmental factors varying among persons (Ober and Vercelli, 2011; Ottman, 1996).

GxE interactions can be visualized by direct physical interactions. An exposure for example may react with a biomolecule initiating a signal transduction pathway (Ober and Vercelli, 2011). In metabolic pathways, an environmental substrate directly interacts with enzymes inducing its degradation. In particular, this plays a critical role in therapy,

since the genetic information can affect the response to drugs via drug metabolism and may lead to drug intolerance (Hunter, 2005). This highlights the fact that especially genes involved in these pathways are very important spots in the context of GxE interactions. However, the same biological mechanisms that apply to interactions within genetic factors apply to GxE interactions as well, and gene and environment may also interact at different steps within the same pathway or even in interrelating or competing pathways (Hunter, 2005; Rothman *et al.*, 1980).
In 1990 Ottman illustrated five different biologically plausible pathophysiological models to visualize the relationship between genetic and environmental factors in terms of their effects on the disease risk. First of all, it is possible that only one factor – either the environmental (M1) or the genetic factor (M2) - show a direct effect on their own, with the other one intensifying the effect. Furthermore, both exposure and the genotype each can have some effect on disease risk on their own, but their joint occurrence leads to an additional risk increase or reduction (M3). The fourth model describes that genetic and environmental factor both show no effect on their own, but only when they occur together (M4). In the last of these models, the genetic factor does not directly cause disease, but is associated with a disease causing environmental factor by influencing the internal dose of the exposure or the acceptance of an external dose (M5). In figure 2.4 the different types of interactions are illustrated. Examples of simple Mendelian disorders for these models can be found in Ottman (1990).
In complex diseases, these forms of GxE interactions are embedded as a single component within an even more complicated architecture. For lung cancer with smoking as the most important environmental factor, we can transfer the different models as follows. A gene involved in nicotine metabolism can exacerbate the effect of smoking – with no direct effect in non-smokers (M1). A direct lung cancer gene can result in lung cancer independent of environment, with smoking as a risk increasing agent with an effect on its own (M3) or only in combination with that particular gene (M2). A smoker without sensitivity to nicotine smoking will not have an increased lung cancer risk, as well as a non-smoker with a mutated gene responsible for nicotine sensitivity. Only both occurring at once, nicotine sensitivity and smoking, lead to a risk effect (M4). A smoking addiction gene on the contrary regulates the level of exposure, while not influencing disease directly (M5). Several examples of GxE in disease development were discovered and evaluated so far. The MC1R for example is responsible for skin color, and a fair skin color combined with UV radiation results in an increased skin cancer risk (Rees, 2004) (M1). The NAT2 gene coding for rapid acetylators increases the colorectal cancer risk only in combination with red meat intake (Chen *et al.*, 1998), while only one of these risk factors on its own shows no effect (M4).
When taking a closer look at the models of Ottman (1990), only M1-M4 represent real interactions between a genetic and environmental factor. In contrast, in M5 the genetic factor impacts only the exposure to the environmental factor but not the disease susceptibility directly. Therefore, we have no interaction, but a correlation between the genetic and environmental factor. Such a correlation totally independent of the disease status that holds in the whole population is called **population-based G-E association**. Population-based G-E associations occur when an underlying gene influences the choice of an environmental factor, e.g. a smoking addiction gene that favors smoking, or the other way around when an exposure determines genes. A population-based

G-E association can result due to causal but also non-causal mechanisms. It is well known that numerous environmental factors such as our behavior and lifestyle (e.g. social attitudes, alcohol, tobacco and other drug consumption or risk-taking behaviors) as well as specific life events and circumstances (e.g. divorce, marital quality and life support) are partially determined by our genetic make-up and heritable (Kendler and Baker, 2007). Hence, our genetic information can influence our behavior to evoke an environmental response or predispose to select or modify an environmental factor, and a population-based G-E association is given. Genes involved in addiction for example might produce causal associations, e.g. genes involved in nicotine dependency such as GPR51 and CYPR51 (Caporaso *et al.*, 2009; Thomas, 2010b; Thorgeirssson *et al.*, 2008; OMIM, 2012 #188890), or genes such as GABRA2 and ADH1C associated with alcoholism (OMIM, 2012 #103780). In addition, our childhood environment is partly influenced by the parent's genetic make-up and behavior (e.g. parental discipline or warmth, smoking of the parents, unsocial behavior), what can lead to a kind of indirect population-based G-E associations as well (Kendler and Baker, 2007). Although not that common, the environment can influence the genetic makeup. Radiation or smoking in pregnancy for example can result in genetic changes and gene defects in the child, what in turn may cause diseases. Population-based G-E associations due to non-causal mechanism can be attributed to evolutionary processes resulting in a change of an allele frequency in a particular environment. An example for this is the HbS variant of the HBB gene that was already mentioned in section 2.3.1. Although this variant leads to sickle-shaped cells in the human body and to a severe disease in homozygous individuals, it also protects against malaria (Aidoo *et al.*, 2002). Therefore, long time exposure to malaria mosquitoes in tropical and subtropical region leads to an increase of the HbS allele, and a population-based G-E association of the HBB gene and malarial environment can be observed.
Note, the difference between a population-based G-E association and a GxE interaction is a highly important aspect of this work, so that both have to be strictly distinguished. However, population-based G-E associations and GxE interactions are not mutually exclusive, but can also occur together.
The mechanism underlying GxE is not completely understood yet and there is still a long way to full knowledge about the relationship between the genetic makeup and the environment (Ober and Vercelli, 2011). Nevertheless, due to the role of GxE interactions in the development of a disease, understanding GxE is an important issue to invent more effective strategies for prevention and treatment. Carriers of particular genes for example may limit or prevent their exposure negatively interacting with the genetic predisposition (National Institutes of Health (NIH), 2012) rather than non-carriers. A regulated diet and sugar intake for example is in particular useful in genetic disposition to diabetes.

3 Genetic association studies

3.1 Association: Definition, study types and measures

In the following sections we will describe what a genetic association is, which study designs can be used to examine genetic association and how it can be measured and tested. As references we used two basic genetic epidemiological books of Bickeböller and Fischer (2007) and Ziegler and König (2006).

3.1.1 Genetic association

An association between two characteristics exists if they occur more or less often together than expected by chance. Hence, in mathematical context an association is a statistical dependency. In genetic epidemiology, genetic markers are examined with respect to an association with a phenotype. Considering a particular disease, this can be done for example by determining if a specific allele of a marker locus is over-represented in the affected individuals, so that a correlation of the genetic variant to the disease development can be assessed.
However, association does not necessarily implicate causality - hence it is important to distinguish real causal associations from non-causal and false positive results. Only the former are of biological interest in genetic studies. In genetic epidemiological studies, two causal models can be distinguished: **direct association** and **indirect association**. A **direct association** is given when the observed association reflects exactly the causal relation of the marker locus and the disease, because the examined locus contributes directly to the disease. However, more often an **indirect association** is observed which is based on a more complex dependency that involves the principle of LD as an important element. For an indirect association, the observed marker locus is not the causal variant itself, but is located close to the susceptibility locus on the same chromosome, so that LD between the marker and disease locus exists. In particular in genome-wide association studies, indirect associations due to LD play a fundamental role. Since nearby genetic variants are correlated with each other at a population level, studies covering the whole genome can be performed without examining every existing polymorphism. Due to LD, redundancies in genotyping can be avoided and the data can be minimized to a subset of SNPs (tagSNPs) representing its neighboring variants as well. Hence, even if a disease-causing variant is not genotyped directly, nearby SNPs may attract attention to the corresponding genetic region.
Nevertheless, non-causal associations can be observed as well. Such non-causal associations in the context of Epidemiology and Genetic Epidemiology are denoted as **spurious associations**. The correlation between the genetic variant and the disease is not due to the genetic factor contributing to the disease susceptibility, but usually due to a third factor not considered in the analysis, denoted as confounder. In general, a **confounder** is an unconsidered disturbing factor that is associated with the outcome variable and with the dependent variable under consideration. In the context of genetic epidemiology where our outcome is the trait of investigation and the dependent variable a genetic risk factor, an unconsidered environmental factor may act as a confounder. This can be for example an environmental factor that is favored by a particular genetic variant

(population-based G-E association, section 2.4.3) and has an association to the disease. Genes related to nicotine independence with no causal influence to lung cancer may show an association with the disease when smoking is not considered in the analysis, due to the co-occurrence with smoking. Hence, smoking presents a confounder in that context. In general, age and sex are potential main confounders, wherefore they are often integrated into the analysis. Another important issue that can lead to spurious associations is population mixture and stratification. To avoid spurious associations, it is important to control for possible external influences. This can be done beforehand by considering confounders in study design and recruitment or by integrating them into the analysis. The latter is outlined in more detail in the context of genome-wide association studies in section 3.2.
Before we will introduce the most important association measures in Genetic Epidemiology in section 3.1.3 and association tests in 3.1.4, we will describe two typical population-based study designs used to investigate genetic associations in the following: cohort studies and case-control studies.

3.1.2 Study designs

As already mentioned in the introduction, in Genetic Epidemiology we can distinguish **linkage** and **association studies**. This differentiation is based on the genetic principle used for the analysis: in linkage studies the cosegregation of genetic loci with the disease within families is examined, in association studies the joint occurrence of a particular marker allele with a disease is considered. While linkage studies are exclusively based on the examination of families, association studies provide the possibility to find an association based on family data as well as on a population level, analyzing unrelated people. In this thesis we will restrict ourselves to population-based data and outline two typical study designs from Epidemiology that are most commonly used in GWAs, as in our real data examples.
In a **cohort study**, a study population (cohort) without the disease of interest is recruited. The individuals differ with respect to the potential risk factor to estimate (genotype in genetic epidemiological studies). The cohort is **prospectively** observed over a predetermined observation period, to see who develops the disease in that timeframe. The aim is to find out if more persons with the exposition (e.g. a particular genetic variant) get the disease relatively to the number of unexposed persons developing this specific phenotype (Ziegler and König, 2006).
Even more common, the **case-control study** design can be observed, which goes the methodological reverse way by recruiting a sample of unrelated individuals with the disease and an unrelated group of individuals without the disease. The affected persons are denoted as cases, the unaffected as controls. After recruitment, the exposure of the individuals within the groups is recorded **retrospectively**, what includes the determination of the genetic status in genetic association studies. Then, the distribution of the potential risk factor can be compared between both groups.
In cohort studies, the proportion of affected individuals at the end of the time period represents the proportion in the whole population. This has the advantage that epidemiological measures can easily be derived. Unfortunately, especially for rare diseases very large cohorts observed over a long period are needed to obtain a sufficient number

Table 3.1: *Data for a cohort study with a genetic factor*

		exposure: genotype			
		G=2	G=1	G=0	
disease	affected D=1	n_{12}	n_{11}	n_{10}	$n_{1.}$
status	unaffected D=0	n_{02}	n_{01}	n_{00}	$n_{0.}$
		$n_{.2}$	$n_{.1}$	$n_{.0}$	n

of affected individuals for the analysis. In case-control studies, the number of cases and controls is fixed by the study investigator and hence does not reflect the population ratio. Caution is required with respect to epidemiological measures, but the design is adequate for rare disease, by directly selecting the cases and hence more cost effective. Furthermore, the case-control design is less time intensive than a cohort study. One does not have to wait for many individuals to e.g. acquire cancer, but can recruit persons that already have the disease.
The main disadvantage of case-control studies is that in general the retrospective recording of the exposure may lead to wrong or missing data (recall bias). However, in genetic studies this is only relevant in the context of considered environmental factors, since the genetic information in general does not change over time but stays stable (Ziegler and König, 2006). A common practice in genetic epidemiology is to use self recruited cases and an available large population-based cohort as controls, e.g. as UKBiobank (http://www.ukbiobank.ac.uk/), KORA (http://www.helmholtz-muenchen.de/kora) or POPGEN (http://www.popgen.de/) despite possible problems with such controls. When using this design, it is important that both groups are comparable with respect to other factors, since that could lead to a systematic bias. In genetic studies, regional differences in the genetic information are for example of high importance and can influence the results.

3.1.3 Measures of association: relative risks and odds ratios

An important epidemiological measure that is used in association studies to represent a disease frequency is the **risk**. The risk is defined as the probability that a randomly chosen person from a considered population at risk becomes newly affected by the disease of interest in a temporal limited period. As population at risk all individuals of a population are considered that are not yet affected by the disease and are potentially capable to develop the disease.
To reveal risk factors of a particular disease, the risks of getting the disease for people exposed to the potential risk factor and the risk for the unexposed individuals are related to each other. The corresponding measure is the relative risk (RR), which is the ratio of the risk in exposed and the risk in unexposed. A relative risk of 1 demonstrates that the exposure has no influence to the disease, while a $RR > 1$ indicates a harmful effect of the exposure and $RR < 1$ a protective one.
Transferred to the context of Genetic Epidemiology where the exposure is a genetic factor, the risks are nothing else than the disease penetrances for the different genotypes (section 2.3.1). Assuming a cohort study with n individuals and a biallelic marker with

the 3 different genotypes 0, 1 and 2 counting the potential risk alleles, the data can be illustrated in a 2x3 contingency table as shown in table 3.1. The entries n_{dg} denote the number of affected (d=1) and unaffected (d=0) individuals with the three different genotypes g=0,1,2. The total number of affected and unaffected individuals is denoted by $n_{1\cdot}$ and $n_{0\cdot}$, the total number of genotypes without considering the disease status by $n_{\cdot 0}$, $n_{\cdot 1}$ and $n_{\cdot 2}$.
We have 3 different risks $r_g = P(D = 1|G = g)$, g=0,1,2, that can be estimated by $\hat{r}_0 = n_{10}/n_{\cdot 0}$, $\hat{r}_1 = n_{11}/n_{\cdot 1}$ and $\hat{r}_2 = n_{12}/n_{\cdot 2}$. We can form two relative risks with respect to the reference genotype 0 denoted as **genotype relative risks** (GRR)

$$GRR_{\text{het}} = \gamma_1 = \frac{P(D = 1|G = 1)}{P(D = 1|G = 0)} \quad \text{and} \quad GRR_{\text{hom}} = \gamma_2 = \frac{P(D = 1|G = 2)}{P(D = 1|G = 0)}.$$

The GRRs can be estimated by $\hat{\gamma}_1 = \hat{r}_1/\hat{r}_0$ and $\hat{\gamma}_2 = \hat{r}_2/\hat{r}_0$. The relation of the two GRRs to each other gives information on the underlying mode of inheritance: while $\gamma_1 = \gamma_2 > 1$ holds for a dominant allele, we have $1 = \gamma_1 < \gamma_2$ for a recessive model. In addition, an additive effect is given when $\gamma_2 = 2\gamma_1 - 1$, a multiplicative one with $\gamma_2 = \gamma_1^2$. In the latter, γ_1 is also denoted as allelic relative risk and the relative risk is altered by this factor for each additional risk allele. While in cohort studies the number of affected and unaffected individuals per exposure group reflects the ratio in the population, this is not the case in case-control studies since the proportion is of cases and controls is arbitrarily chosen by the investigator. Therefore, it is not possible to estimate the RR directly from the data of a case-control study. However, in place of the RR, another measure, the **odds ratio** (OR) can be used. An odds or chance is the probability of an event divided by the probability of its reverse event. The OR relates two odds to each other. Considering a binary exposure E, we compare the odds of being exposed within the cases $Odds_{\text{cases}} = P(E = 1|D = 1)/P(E = 0|D = 1)$ with the corresponding odds in the controls $Odds_{\text{controls}} = P(E = 1|D = 0)/P(E = 0|D = 0)$ by the odds ratio

$$OR = \frac{P(E = 1|D = 1)P(E = 0|D = 0)}{P(E = 0|D = 1)P(E = 1|D = 0)}. \tag{3.1}$$

Based on the genetic example in the table two different odds ratios OR_{het} and OR_{hom} can be estimated by

$$\hat{OR}_{\text{het}} = \frac{n_{11}/n_{01}}{n_{10}/n_{00}} \qquad \text{and} \qquad \hat{OR}_{\text{hom}} = \frac{n_{12}/n_{02}}{n_{10}/n_{00}}.$$

When assuming a dominant or recessive model we may estimate

$$\hat{OR}_{\text{dom}} = \frac{(n_{11} + n_{12})/(n_{01} + n_{02})}{n_{10}/n_{00}} \qquad \text{or} \qquad \hat{OR}_{\text{rec}} = \frac{n_{12}/n_{02}}{(n_{10} + n_{11})/(n_{00} + n_{01})}.$$

Alternatively considering the allelic rather than the genotypic effect by counting the alleles, we obtain

$$\hat{OR}_{\text{all}} = \frac{(n_{11} + 2n_{12})/(n_{01} + 2n_{02})}{(2n_{10} + n_{11})/(2n_{00} + n_{01})}$$

In general, the OR overestimates the corresponding RR, with the OR approaching the RR with decreasing occurrence of the disease. Therefore, given a disease that is relatively rare among those with and without the risk factor of investigation (rare disease

assumption) the OR provides a good approximation of the RR and can used in case-control studies. In practice, even for a disease with a prevalence of 10%, physicians and epidemiologists still build on this assumption. The prevalence is the proportion of diseased people of the considered population at a given time point.

3.1.4 Testing for association

For the data given in table 3.1 comprising $n_{1.}$ cases and $n_{0.}$ controls, an association analysis tests the null hypothesis that the genetic variant and the disease occur independently from each other. Statistical tests to examine this null hypothesis are all methods known to analyze dichotomous outcome data, e.g. χ^2 tests or logistic regression models. In general, the χ^2 test of independence checks if two categorical or qualitative variables are independent from each other by comparing the observed frequencies for the possible combinations of variable outcomes with the expected ones assuming no association. With respect to our genetic data, we can distinguish different alternative χ^2 tests. The test can be performed allele or genotype based, with the latter providing several more alternatives, distinguishing different genetic modes of inheritance. Comparing all three genotype groups directly, we can calculate the test statistic

$$\chi^2_G = \sum_{d=0,1;g=0,1,2} \frac{(n_{dg} - e_{dg})^2}{e_{dg}},$$

with the expected counts calculated by $e_{dg} = n_{d.}n_{.g}/n$. This test statistic is asymptotically χ^2 distributed with 2 degrees of freedom (df) under the null hypothesis of independence. By assuming a dominant or recessive mode of inheritance, specific alternative hypotheses are given, that are restricted to the comparison of only two genotype groups by collapsing the heterozygotes with one of the homozygous genotypes. The test statistic assuming a dominant model is given by

$$\chi^2_{\text{dom}} = \sum_{d=0,1} \left(\frac{\left(n_{d0} - \frac{n_{.0}n_{d.}}{n}\right)^2}{\frac{n_{.0}n_{d.}}{n}} + \frac{\left((n_{d1} + n_{d2}) - \frac{(n_{.1}+n_{.2})n_{d.}}{n}\right)^2}{\frac{(n_{.1}+n_{.2})n_{d.}}{n}} \right),$$

which can be simplified to

$$\chi^2_{\text{dom}} = n\frac{(n_{10}(n_{01} + n_{02}) - n_{00}(n_{11} + n_{12}))^2}{n_{1.}n_{0.}n_{.0}(n_{.1} + n_{.2})}. \tag{3.2}$$

The corresponding statistic when assuming a recessive model is

$$\chi^2_{\text{rec}} = n\frac{((n_{10} + n_{11}) + n_{02}) - (n_{00} + n_{01})n_{12})^2}{n_{1.}n_{0.}(n_{.0} + n_{.1})n_{.2}}.$$

Under the null hypothesis of no association, both statistics are asymptotically χ^2 distributed with 1 df. As already mentioned, we can also test for association based on the alleles rather than the genotypes. We count each occurring allele resulting in twice the sample size, having $2n_{.0} + n_{.1}$ wildtype variants and $2n_{.2} + n_{.1}$ mutation variants. These are distributed to cases and controls with $2n_{10} + n_{11}, 2n_{00} + n_{01}$ and $2n_{12} + n_{12}, 2n_{02} + n_{01}$.

Plugging in these numbers in the general formula for a chi-square test and simplifying results in the test statistic

$$\chi^2_{\text{all}} = 2n\frac{((2n_{10}+n_{11})+(n_{01}+2n_{02})-(2n_{00}+n_{01})(n_{11}+2n_{12}))^2}{2n_{1.}2n_{0.}(2n_{.0}+n_{.1})(n_{.1}+2n_{.2})}$$

which is again asymptotically χ^2 distributed with 1 df. Furthermore, because of the biological plausibility that the number of risk alleles has an influence to the disease occurrence, the **Armitage-Trend-Test** is often used. This test distinguishes all 3 possible genotypes but assumes a trend in the effects with an increasing level of the risk factor. The trend statistic is given by

$$\chi^2_{\text{tr}} = \frac{\left(\sum_{g=0}^{2} w_g(n_{0g}n_{1.}-n_{1g}n_{0.})\right)^2}{\frac{n_{0.}n_{1.}}{n}\left(\sum_{g=0}^{2} w_g^2 n_{.g}(n-n_{.g}) - 2\sum_{g=0}^{2}\sum_{h=g+1}^{2} w_g w_h n_{.g} n_{.h}\right)},$$

with $w=(w_0,w_1,w_2)$ weights that can be chosen to fit different association models. The statistic is χ^2 distributed with 1 df under the null hypothesis of no association. In GWAS a linear trend with increasing number of the minor allele is often assumed, denoted as **additive effect**. Therefore, we use w=(0,1,2) and the test statistic simplifies to

$$\chi^2_{\text{tr}} = \frac{n(n(n_{11}+2n_{12})-n_{1.}(n_{.1}+2n_{.2}))^2}{n_{1.}n_{0.}(n(n_{.1}+4n_{.2})-(n_{.1}+2n_{.2})^2)} \tag{3.3}$$

In general, these weights are not only used when the trend is linear but also when the change is assumed to be monotonically (Clarke *et al.*, 2011). Weight of w=(0,0,1) would correspond to a recessive model, w=(0,1,1) to a dominant one. The advantage of the allele based approach is the doubling of the sample size. However, in general genotype based tests should be preferred, because they are robust to deviations from HWE, while the allele based test is only valid under the assumption of HWE. In addition, this is the biologically more plausible variant. Depending on the assumed biological function and mode of inheritance of a genetic variant, the corresponding test should be chosen. The trend test is suggested when no biological knowledge exists, because it often reaches the highest power ("locally optimal"). When sparse cells (expectation less than 5) occur, Fisher's exact test should be used instead of a χ^2 test.
When additional variables, e.g. age or sex, should be considered in the analysis, a logistic regression model offers a good alternative by including them as covariates. In general, a regression model describes the influence of one or more risk factors $X_1 ... X_K$ to an outcome measure Y by an equation of the form

$$f(Y) = \alpha + \beta_1 X_1 + \cdots + \beta_K X_K + \epsilon$$

with α denoted as intercept, regression coefficients $\beta_1, \ldots, \beta_K$ and $\epsilon \sim N(0, \sigma^2)$. Given a quantitative phenotype e.g. blood pressure as outcome Y and only one influencing genetic risk factor G, the model reduces to a simple linear regression of the form

$$Y = \alpha + \beta G + \epsilon, \qquad \epsilon \sim N(0, \sigma^2),$$

which is often rewritten as

$$E(Y|G) = \alpha + \beta G.$$

The equation describes which Y value is expected given a particular G with residuals $\epsilon = Y - E(Y|G)$. Given a dichotomous outcome variable, in our context affected (D=1) and unaffected (D=0) according to a disease of interest, we cannot model the outcome directly by a linear equation anymore. Therefore, a logit transformation, that is the logarithm of the odds of a disease, has to be used for D, resulting in a logistic regression model of the form

$$ln\left(\frac{P(D=1|G)}{P(D=0|G)}\right) = \alpha + \beta G + \epsilon, \qquad \epsilon \sim N(0, \sigma^2). \tag{3.4}$$

Hence, we can obtain the expected probability to become affected given genotype G

$$E(D|G) = P(D=1|G) = \frac{\exp(\alpha + \beta G)}{1 + \exp(\alpha + \beta G)}.$$

The Armitage-Trend-Test tests the same information as a logistic regression model with one regression variable for the genotype coded 0,1 and 2. Furthermore, the logistic regression coefficients are directly related to the odds ratios measuring the strength of association by

$$OR_{\text{het}} = \frac{\exp(\alpha + \beta)}{\exp(\alpha)} = \exp(\beta) \quad \text{and} \quad OR_{\text{hom}} = \frac{\exp(\alpha + 2\beta)}{\exp(\alpha)} = \exp(2\beta). \tag{3.5}$$

In this regression model $OR_{\text{hom}} = OR_{\text{het}}^2$, hence it is based on the assumption of a multiplicative allele effect. However, when no multiplicity should be assumed, two dummy variables G_{het} and G_{hom} for the heterozygous and homozygous genotype can be used alternatively

$$ln\left(\frac{P(D=1|G_{\text{het}}, G_{\text{hom}})}{P(D=0|G_{\text{het}}, G_{\text{hom}})}\right) = \alpha + \beta_{\text{het}} G_{\text{het}} + \beta_{\text{hom}} G_{\text{hom}} + \epsilon, \qquad \epsilon \sim N(0, \sigma^2).$$

with $G_{\text{hom}} = 1$ and $G_{\text{het}} = 0$ for G=1 and $G_{\text{hom}} = 1$ and $G_{\text{het}} = 0$ for G=2. Then $OR_{\text{het}} = \exp(\beta_{\text{het}})$ and $OR_{\text{hom}} = \exp(\beta_{\text{hom}})$. We will come back to the connection of regression coefficients and OR in the context of GxE interaction in section 3.1.5 and for the derivation of our approach in chapter 6. As mentioned before, the advantage of a regression model compared to a χ^2 test is that other influencing factors can be included in the analysis as well. These can be e.g. sex and age or other confounders as additional genetic or environmental factors necessary to adjust for. By including these additional factors $X_1, ..., X_K$, the model expands to a multiple logistic regression model of the form

$$ln\left(\frac{P(D=1|G, X_1, ..., X_K)}{P(D=0|G, X_1, ..., X_K)}\right) = \alpha + \beta_G G + \beta_1 X_1 + \cdots + \beta_K X_K + \epsilon, \qquad \epsilon \sim N(0, \sigma^2)$$

From this model, odds ratios adjusted for the covariates can be calculated as shown before. The regression coefficients of a logistic regression model can be estimated by the principle of maximum likelihood. In general, the maximization cannot be analytically performed since no exact solution is available, so that an iterative approach such as the Newton-Raphson algorithm for numerical optimization has to be used (Faraway, 2006). An influence of the genetic factor to the disease is given when the corresponding estimated coefficient, e.g. $\hat{\beta}_G$, is significantly different from 0. This can be tested by

dividing it by the corresponding estimated standard deviation ($\hat{\sigma}_{\beta_G}$) and using a Wald test, assuming a normal distribution under the null hypothesis of no effect. A stratified analysis of case-control data according to categorical covariates is possible as well, e.g. with the Cochrane-Mantel-Haenszel test (Agresti, 2002). Moreover, non-parametric methods exist that we will not elaborate since we are concentrating on parametric approaches.

3.1.5 Gene x environment interactions and gene-environment associations

A mathematical definition of a GxE interaction

In the following we will restrict our attention to a dominant genetic risk factor (carrier/non carrier of the susceptibility allele) and a binary environmental factor (exposed/unexposed) since it is used that way in our GxE interaction methods. The disease risks for the different risk factor combinations are given by

Table 3.2: *Disease risks for individuals with different combinations of genetic and environmental risk factor*

		environmental factor	
		exposed	unexposed
genetic	carrier	r_{11}	r_{10}
factor	non-carrier	r_{01}	r_{00}

For the definition of a gene x environment interaction in a statistical sense, the principle of conditional independence (Dawid, 1979; Jakulin and Bratko, 2004) is used. In general, two factors X and Y are called conditionally independent with respect to a third factor Z if and only if

$$P(X, Y|Z) = P(X|Z)P(Y|Z). \tag{3.6}$$

An equivalent way to express this relationship is given by

$$P(X|Y, Z) = P(X|Z). \tag{3.7}$$

In terms of gene x environment interactions, conditional independence is present when the effect of one of the risk factors (genetic or environmental) on the disease risk is the same across strata defined by the other risk factor. The absence of this independence is called interaction. Hence, a gene x environment interaction in a statistical sense is observed, when the effect of the environmental factor on disease risk differs depending on the underlying genotype, or when the genotype effect on disease risk differs in subjects depending on the environmental exposure (Ottman, 1996).
However, the existence of an interaction corresponding to this definition depends on the scale of measurement used for the disease risks. Two different scales are commonly used: additive and multiplicative. Based on a cohort, an interaction on an additive measurement scale is defined when $r_{11} - r_{10} \neq r_{01} - r_{00}$, while an interaction on a multiplicative scale is present when $r_{11}/r_{10} \neq r_{01}/r_{00}$. In terms of relative risks, an

Table 3.3: *Data for an unmatched case-control study with a binary genetic and environmental factor*

	genetic factor				
	carrier		noncarrier		
	environmental factor		**environmental factor**		
	exposed	unexposed	exposed	unexposed	Total
cases	n_{111}	n_{110}	n_{101}	n_{100}	N_{cases}
controls	n_{011}	n_{010}	n_{001}	n_{011}	N_{controls}

interaction on an additive scale is given when $RR_{11} \neq RR_{01} + RR_{10} - 1$, and $RR_{11} \neq RR_{01}RR_{10}$ represents an interaction assuming a multiplicative model. $RR_{10} = r_{10}/r_{00}$ denotes the relative risk for unexposed carriers with respect to unexposed non-carriers of the susceptibility allele as a reference group, $RR_{01} = r_{01}/r_{00}$ for exposed non-carriers and $RR_{11} = r_{11}/r_{00}$ for exposed carriers.
If we take a look at the possible kinds of biological interactions listed in section 2.4.3, we can notice that interactions of type M1, M2 and M4 express themselves in statistical interactions as defined above on both scales. A relationship of the three factors of type M5 is not reflected in a statistical interaction at all, neither on an additive nor a multiplicative scale. In this case, only the frequency of the joint occurrence of the genetic and environmental factor are influenced, but the effect of the genetic factor on the disease risk stays the same across strata of the environmental factor and the other way around. A biological interaction as described by model M3 may manifest itself in a statistical interaction, but not necessarily, and the scale of the risk measure plays an important role. For instance, when we consider a multistage process like the initiation or promotion in cancer, two factors that act at the same stage fit a risk model on additive scale and an interaction results in a departure from additivity. When both factors act at different stages, this better fits a multiplicative model, and an interaction in this case can be observed as deviation from multiplicativity (Rothman *et al.*, 1980; Siemiatycki and Thomas, 1981). Hence, both scales can be adequate depending on the underlying pathophysiological model and mechanism (Koopman, 1977; Kupper and Hogan, 1978; Ottman, 1996; Walter and Holford, 1978). A point of view that might be taken into account choosing the scale is determined by the goal of investigation, leading to the preference of a multiplicative scale when the causes of disease should be revealed (Rothman *et al.*, 1980). For our further investigations, we used the definition of GxE interaction on a multiplicative scale, which is the commonly used and adequate one in a case-control study.

Measures and testing of GxE interactions and G-E associations

In this section we will restrict to the case-control study design since the data in our applications and our simulation studies are based on that. Therefore, assume in the following that we have an unmatched case-control study for the disease D with a binary environmental exposure E and a binary genetic factor G. The data can be presented in a 2x4 table as given in table 3.3. The entries n_{dge} denote the number of cases (d=1) and controls (d=0) that are carriers (g=1) or non-carriers (g=0) of the susceptibility allele and exposed (e=1) or unexposed (e=0) to the environmental factor. The observed cell

counts for cases $n_1 = (n_{111}, n_{110}, n_{101}, n_{100})$ and controls $n_0 = (n_{011}, n_{010}, n_{001}, n_{000})$ can be viewed as realizations from two independent multinomial distributions

$$n_1 \sim MN(N_{\text{cases}}, p_1) \text{ and } n_0 \sim MN(N_{\text{controls}}, p_0),$$

where $p_1 = (p_{111}, p_{110}, p_{101}, p_{100})$ and $p_0 = (p_{011}, p_{010}, p_{001}, p_{000})$ are the cell probabilities of the underlying case-control population.
It is known that relative risks (RR) cannot be calculated for studies designed in case-control manner. However, we can calculate Odds Ratios (OR) - with unexposed non-carriers as reference group - given by

$$\begin{aligned} OR_g &= \frac{p_{000}p_{110}}{p_{010}p_{100}} \quad \text{genetic main effect,} \\ OR_e &= \frac{p_{000}p_{101}}{p_{001}p_{100}} \quad \text{environmental main effect,} \\ OR_{ge} &= \frac{p_{000}p_{111}}{p_{011}p_{100}} \quad \text{joint effect of gene and environment.} \end{aligned} \tag{3.8}$$

Under a multiplicative model with no interaction effect, we have $OR_{ge} = OR_g OR_e$, and hence the interaction effect can be measured by the interaction parameter

$$\Psi = OR_{ge}/(OR_g OR_e), \tag{3.9}$$

$$\text{with } \Psi \begin{cases} > 1 & \text{positive interaction effect - more than multiplicative} \\ = 1 & \text{no interaction effect - multiplicative} \\ < 1 & \text{negative interaction effect - less than multiplicative} \end{cases}.$$

It should be noticed, that when a GxE interaction exists, it expresses itself not only in form of the interaction itself, but also in dependent distributions of genetic factor and exposure within the cases and within the controls. In presence of a positive interaction, exposure and genetic susceptibility factor occur more often together in cases than expected and less often in controls, the contrary holds in presence of negative interaction. The dependency of a genetic and an environmental factor is called gene-environment association (G-E), and can be measured by OR as well, with

$$\begin{aligned} OR_{\text{cases}} &= \frac{p_{100}p_{111}}{p_{110}p_{101}} \quad \text{for cases} \\ OR_{\text{controls}} &= \frac{p_{000}p_{011}}{p_{010}p_{001}} \quad \text{for controls.} \end{aligned} \tag{3.10}$$

In absence of G-E association in the corresponding group, $OR_{\text{cases}} = 1$ and $OR_{\text{controls}} = 1$ respectively. A departure from one indicates an association.
The exact relationship between Ψ and the stratified G-E association measures OR_{cases} and OR_{controls} can be derived by simply rearranging the formula for Ψ, resulting in

$$\Psi = \frac{OR_{ge}}{OR_g OR_e} = \frac{\frac{p_{000}p_{111}}{p_{011}p_{100}}}{\frac{p_{000}p_{110}}{p_{010}p_{100}}\frac{p_{000}p_{101}}{p_{001}p_{100}}} = \frac{\frac{p_{100}p_{111}}{p_{110}p_{101}}}{\frac{p_{000}p_{011}}{p_{010}p_{001}}} = \frac{OR_{\text{cases}}}{OR_{\text{controls}}}. \tag{3.11}$$

According to this formula, we can distinguish two different situations where $\Psi = 1$ and hence no interaction occurs, namely when no G-E association is given at all

$$OR_{\text{cases}} = OR_{\text{controls}} = 1$$

or when the G-E association is present in the exact same magnitude in cases and controls

$$OR_{\text{cases}} = OR_{\text{controls}} \neq 1.$$

The latter is a population-based G-E association (section 2.4.3) and is given when we have a correlation between G and E that exists in the whole population to the same degree totally independent from the disease status. On the other hand, when we have a G-E association only caused by an underlying interaction effect, OR_{cases} and OR_{controls} depart from 1 in different directions. Hence, while a population-based G-E association can be observed to the same extend in cases and controls, an interaction is given when the G-E association is different in both groups.
With a decreasing prevalence of the disease, the departure from 1 of the OR_{controls} due to an interaction effect gets weaker and reduces to one under the rare disease assumption (Schmidt and Schaid (1999)). Thus, for a rare disease the interaction effect is only reflected in the association within the cases OR_{cases}.
Since odds ratios are closely connected to logistic regression, with the coefficients of the regression model corresponding to the logarithm of the respective Odds Ratios, we can test an interaction effect by such a logistic regression model

$$\text{logit}P(D = 1 \mid g, e) = log\left(\frac{P(D = 1 \mid g, e)}{P(D = 0 \mid g, e)}\right) = \alpha + \beta_e e + \beta_g g + \beta_{ge} ge. \tag{3.12}$$

The regression coefficient β_g is a measure of the genetic main effect, β_e measures the environmental main effect and β_{ge} is a measure of the interaction effect between G and E. The regression coefficients are related to the odds ratios and the interaction parameter Ψ by

$$\begin{aligned} \beta_g &= log(OR_g) \\ \beta_e &= log(OR_e) \\ \beta_{ge} &= log(\Psi). \end{aligned} \tag{3.13}$$

A regression coefficient of 0 corresponds to no effect, a coefficient >0 indicates a positive effect and a coefficient <0 a negative one. The G-E associations stratified by disease status can be measured by logistic regression as well

$$\text{logit}P(E = 1 \mid D = 1, g) = log\left(\frac{P(E = 1 \mid D = 1, g)}{P(E = 0 \mid D = 1, g)}\right) = \alpha_{\text{cases}} + \beta_{\text{cases}} g \tag{3.14}$$

$$\text{logit}P(E = 1 \mid D = 0, g) = log\left(\frac{P(E = 1 \mid D = 0, g)}{P(E = 0 \mid D = 0, g)}\right) = \alpha_{\text{controls}} + \beta_{\text{controls}} g, \tag{3.15}$$

with $\beta_{\text{cases}} = log(OR_{\text{cases}})$, $\beta_{\text{controls}} = log(OR_{\text{controls}})$ and

$$\beta_{\text{cases}} - \beta_{\text{controls}} = log(\frac{OR_{\text{cases}}}{OR_{\text{controls}}}) \overset{(3.11)}{=} log(\Psi) = \beta_{ge}. \tag{3.16}$$

The regression coefficients and hence the OR can be estimated from the data by their

maximum likelihood estimators (MLE), given by

$$\begin{aligned}
\hat{\beta}_g &= log\left(\frac{n_{000}n_{110}}{n_{010}n_{100}}\right) \\
\hat{\beta}_e &= log\left(\frac{n_{000}n_{101}}{n_{001}n_{100}}\right) \\
\hat{\beta}_{\text{cases}} &= log\left(\frac{n_{100}n_{111}}{n_{110}n_{101}}\right) \\
\hat{\beta}_{\text{controls}} &= log\left(\frac{n_{000}n_{011}}{n_{010}n_{001}}\right)
\end{aligned} \tag{3.17}$$

and

$$\hat{\beta_{ge}} \overset{(3.16)}{=} \hat{\beta}_{\text{cases}} - \hat{\beta}_{\text{controls}} = log\left(\frac{n_{001}n_{100}n_{010}n_{111}}{n_{011}n_{110}n_{101}n_{000}}\right). \tag{3.18}$$

These MLE asymptotically follow approximate normal distributions (Le, 1991; Mukherjee *et al.*, 2008)

$$\begin{aligned}
\hat{\beta}_g &\sim N(\beta_g, \sigma_g^2) \\
\hat{\beta}_e &\sim N(\beta_e, \sigma_e^2) \\
\hat{\beta}_{\text{cases}} &\sim N(\beta_{\text{cases}}, \sigma_{\text{cases}}^2) \\
\hat{\beta}_{\text{controls}} &\sim N(\beta_{\text{controls}}, \sigma_{\text{controls}}^2) \\
\hat{\beta}_{ge} &\sim N(\beta_{ge}, \sigma_{ge}^2)
\end{aligned} \tag{3.19}$$

with variance estimators given by

$$\begin{aligned}
\hat{\sigma}_g^2 &= \sum_{d=0,1}\sum_{g=0,1}\frac{1}{n_{dg0}} \\
\hat{\sigma}_e^2 &= \sum_{d=0,1}\sum_{e=0,1}\frac{1}{n_{d0e}} \\
\hat{\sigma}_{\text{cases}}^2 &= \sum_{g=0,1}\sum_{e=0,1}\frac{1}{n_{1ge}} \\
\hat{\sigma}_{\text{controls}}^2 &= \sum_{g=0,1}\sum_{e=0,1}\frac{1}{n_{0ge}} \\
\hat{\sigma}_{ge}^2 &= \sum_{d=0,1}\sum_{g=0,1}\sum_{e=0,1}\frac{1}{n_{dge}} = \hat{\sigma}_{\text{cases}}^2 + \hat{\sigma}_{\text{controls}}^2.
\end{aligned} \tag{3.20}$$

The classical case-control test for gene x environment interactions simply tests the interaction coefficient β_{ge} with null hypothesis $H_0 : \beta_{ge} = 0$. Because of its approximate normal distribution, the case-control test statistic corresponds to a standardized normal test for β_{ge} by normalizing the estimate $\hat{\beta}_{ge}$ from the data by its estimated standard deviation $\hat{\sigma}_{ge}$, resulting in

$$Z_{cc} = \frac{\hat{\beta}_{ge}}{\hat{\sigma}_{ge}} = \frac{\hat{\beta}_{\text{cases}} - \hat{\beta}_{\text{controls}}}{\sqrt{\hat{\sigma}_{\text{cases}}^2 + \hat{\sigma}_{\text{controls}}^2}}.$$

This test statistic is asymptotically $N(\beta_{ge}, 1)$ distributed, with $\beta_{ge} = 0 = \beta_{\text{cases}} - \beta_{\text{controls}}$ (standard normal distribution) under the null hypothesis of no interaction. Furthermore, we have that

$$\beta_{\text{cases}} = \beta_{\text{controls}} = 0$$

when genotype and environmental factor are independent from each other; given a population-based G-E association

$$\beta_{\text{cases}} = \beta_{\text{controls}} \neq 0$$

holds.

3.2 Genome-wide association studies (GWAS)

3.2.1 Genetic epidemiological study types

In **genome-wide studies** the whole genome is systematically examined by using numerous genetic markers distributed through the complete genetic information to find genes involved in disease development. The counterpart to the exploratory genome-wide approach is the hypothesis-driven **candidate gene studies**. Candidate gene studies focus on analyzing only genes or regions already known or expected to be involved in disease etiology. Candidates can come from other, e.g. experimental, studies, from knowledge in other species, or the information about functional relations of genes with the disease. For autoimmune diseases for example, the HLA system on chromosome 6 is known as the most important candidate region. While candidate gene studies can be successful when good candidates are known, a genome-wide search is the method of choice when insufficient information about the biological and biochemical processes of the disease is given and hence inadequate prior knowledge about potentially involved genes is available. Furthermore, even when good candidates are known, genome-wide studies can find additional new genes not expected before. Genome-wide studies are totally independent from pathophysiological hypotheses and therefore keep all possibilities open. As already mentioned before, two different genetic principles can be used to find genes contributing to disease development: **linkage** and **association**. Linkage studies are successful to find rare variants with high penetrances that strongly increase disease risk. On the contrary, association studies have higher power in finding common variants with a reduced penetrance and low to moderate risk effects, involved in a more complicated interplay of numerous genetic and non-genetic factors. They allow a finer mapping of potential disease causing factors while linkage analyses are only applicable to identify a coarse region. Since association studies can be performed on a population basis, the recruitment is simpler than for families. However, population approaches are more prone to confounding e.g. by population stratification, possibly leading to false positive results. Before the 21^{st} century, only linkage studies were possible genome-wide. Genome-wide linkage studies were tremendously successful for the identification of genes underlying monogenic disease, characterized by their rare occurrence, high penetrance and large relative risk (Hirschhorn, 2005; Thomas *et al.*, 2005; Thomas, 2006). Major genes involved in clear Mendelian subtypes of complex diseases showed similar properties and were detected as well, but beyond, the success in complex disease was limited (Altmüller *et al.*, 2001). For those factors of complex diseases involved in the interplay of multiple genetic and environmental factors in a complicated way (Wang *et al.*, 2005), the power was much too low due to incomplete penetrances and relatively small effects (Cardon and Bell, 2001; Hirschhorn, 2005; Risch and Merikangas, 1996; Risch, 2000; Tabor *et al.*, 2002; Thomas, 2006). Associations studies on the other hand, since only possible as a candidate approach at that time, failed due to an imperfect understanding about the fundamental biology of complex diseases and hence lack of ability to pick good candidate genes (Pearson and Manolio, 2008; Sham and Cherny, 2010). Although candidate gene association studies revealed many susceptibility genes, the replication rate was only low (Patterson and Cardon, 2005; Sham and Cherny, 2010; Todd, 2006; Zondervan, 2010). Reasons for that may be the overestimation of the ability to select adequate candidates and too low thresholds for claiming an association (Khoury and

Wacholder, 2009; Wacholder *et al.*, 2004). In a review of 2002, Hirschhorn *et al.* illustrated that in 603 candidate gene studies with a case-control design published from 1986-2000, only 6 results were independently replicated.

3.2.2 The upcoming of genome-wide association studies

Since association studies have potentially far greater power to detect genetic variants with modest effects, Risch and Merikangas suggested in 1996 that **genome-wide association studies** (GWAS) would be the answer to the problems in mapping genes of complex diseases (Risch and Merikangas, 1996). GWAS could exploit the strengths of association without having to guess the identity of causal genes (Hirschhorn, 2005). They demonstrated GWAS to be potentially feasible, with SNPs as putative genetic factors to identify. Unfortunately, at that time, GWAS were technically not feasible. Linkage extends over large distances and therefore in linkage studies the genome can be covered by only several hundred Sham and Cherny (2010) microsatellite markers. On the other hand, linkage disequilibrium which is the basis of association studies can only be observed over small distances so that an enormous large set of dense marker is necessary to cover the whole genome. However, knowledge about the human genome, common genetic variation and its LD patterns was still missing. Technologies to genotype a sufficiently comprehensive set of common variants in a large sample (Hirschhorn, 2005) for affordable costs did not exist. Then in the beginning of the 21^{st} century, mega advances in the genomic sciences set the stage for GWAS (Hirschhorn, 2005) and offer much hope for the future (Rao, 2008). By the human genome project (U.S. Department of Energy Genome Programs, 2011) and the SNP consortium (Thorisson and Stein, 2003) (appendix A.1) in the late 90s and beginning of the new millennium, some million common SNPs were discovered and publicly released. These constituted the starting point of the new SNP era (Lander *et al.*, 2001; McPherson *et al.*, 2001; Sham and Cherny, 2010). Seeing the good perspective of SNPs, the emphasis was further shifted to the investigation of SNP characteristics such as genotype frequencies and the nature of LD across the entire human genome (Sham and Cherny, 2010). This was extensively done by the International Hap Map project (International HapMap Consortium, 2003, 2005) (appendix A.1) initiated in October 2002 (Barrett, 2010; Weiss and Terwilliger, 2000). Initially one population of each European and African ancestry and two Asian populations were examined, with ongoing investigations covering additional populations (Sham and Cherny, 2010). All obtained information is published in free databases (Sachidanandam *et al.*, 2001), e.g. dbSNP database (Database of Single Nucleotide Polymorphisms, 2009). Today, the dbSNP database contains more than 40 million validated human SNPs including nearly 15 million with a MAF>1% [29 February 2012]. By using the catalog of the millions of SNPs discovered across diverse populations and considering the LD block structures obtained by Hap Map, it was possible to identify subsets of highly informative so called tag SNPs (Thomas *et al.*, 2005) to capture most of the genomic variation (Barrett, 2010; Rao, 2008; Ziegler *et al.*, 2008) without genotyping all possible SNPs. This was an important step that paved the way for the efficient practical conceptual realization of future GWAS (Barrett, 2010; Rao, 2008). As already mentioned, the concentration so far lied mainly on common SNPs, with common defined as at least 1% frequency of the minor variant in a population

(Frazer *et al.*, 2009). Nevertheless, that agrees with a substantial concept underlying the idea of GWAS in complex diseases: the common disease common variants hypothesis (CDCV) (Collins *et al.*, 1997; Reich and Lander, 2001; Zondervan, 2010). The CDCV states that the genetic origin of complex traits includes relatively common variants with modest effects on risk (RR 1.3-2.5, Thomas (2010c)), increasing the susceptibility to the disease rather than directly causing the disease (Zondervan, 2010).
However, to make GWAS possible, another major contribution was necessary: improvements in ultra-high-throughput technology. SNP genotyping chip arrays made it commercially feasible to investigate hundreds of thousands genetic SNP variants per sample simultaneously in thousands of individuals at manageable costs (Grimm *et al.*, 2011; Syvänen, 2001; Thomas *et al.*, 2005; Thomas, 2006; Thomas *et al.*, 2009; Zondervan, 2010). Genome-wide SNP platforms started with modest 10,000 SNPs, but soon several hundred thousands followed, with today's latest technological achievement of high-throughput chips comprising one million SNPs. While the number of SNPs per chip increased with time, the costs of large scale studies became even cheaper (Hirschhorn, 2005; Thomas, 2010c). The current average costs for an Affymetrix chip with 1 million SNPs are $\sim$ 400 (personal communication Affymetrix) including reagents and service. More information about the genome-wide SNP chips used is given in the appendix A.1.
Finally, nearly one decade after Risch and Merikangas' (1996) first suggestion to use GWA scans to analyze complex diseases, large-scale association studies became reality (Rao, 2008; Thomas, 2010c).

3.2.3 Data quality checks

SNP chips allow researchers to interrogate hundreds of thousands SNPs across the human genome (Weale, 2010) with the goal to identify true association signals in sea of false positive results (Christensen and Murray, 2007). A good quality of the data is an essential point to avoid false positive results and guarantee to draw accurate conclusions from the analysis (Neale and Purcell, 2008). On the one hand, quality assurance during study conduct is necessary, ensuring a good study design, good sampling protocols, good quality DNA, adequate protocols for DNA extraction and preparation (Weale, 2010). On the other hand, an additional exploratory data quality control is the first step of a GWAS analysis to evaluate the genotyping performance and is indispensable (Neale and Purcell, 2008; Thomas, 2010c). The process how to get from the chip signals to the genotype for each SNP, a process called genotype calling, is described in the appendix A.2. The assignment of a genotype to a SNP according to the corresponding chip signal is denoted as call. Genotyping errors (miscalls) as well as missing data (no-calls) can occur. Factors influencing the quality of genotyping are for example the concentration, contamination or possible degradation of the input DNA, failures or degeneration of the chip arrays, as well as differences in sample preparation (e.g. different laboratories) and plating errors (Teo, 2008; Weale, 2010). As long as wrong and missing genotypes occur at random and affect cases and controls equally, it will lead to some loss of power and bias of effect estimates, but not to an increase in the type I error (Bickeböller and Fischer, 2007; Thomas *et al.*, 2005). However, problems occur when the genotype quality differs with the phenotype because cases and controls are not genotyped in an identical manner, e.g. on separate days, separate plates, by separate laboratory assistants or even

in different laboratories (Bickeböller and Fischer, 2007; Hirschhorn, 2005). This may lead to different systematic missings and misclassifications in cases and controls and hence in bias and spurious associations (Thomas, 2010c). To avoid these problems, it is recommended to plate cases and controls together (same laboratory assistant on the same day under the same conditions) so that at least plate effects are evenly distributed. Special attention is essential when controls from a predefined reference database are used (Weale, 2010). In addition, genotyping errors and missing values can be not randomly distributed among the different genotypes of a SNP but rather over-represented in a particular genotype (Weale, 2010). By the identification of bad quality SNPs, as well as individuals that do not fulfill several quality criteria and excluding them from the analysis, inflated type I and type II errors can be avoided (Thomas, 2010c). In addition, reducing the number of SNPs leads to a decrease of the multiple testing burden and therefore to higher power for the remaining SNPs (Weale, 2010) Criteria to filter out SNPs are their proportion of missing genotypes (e.g. <95%), the MAF (e.g. <5%) and strong deviations from HWE (in GWAS: $p < 10^{-7}$). Persons should be excluded from the analysis when they show missing genotypes for many SNPs (e.g. <90%) or an excess of heterozygous or homozygous genotypes. When the reported sex is not the same as the sex determined by the X chromosome, an incorrect alignment of genetic and phenotypic data cannot be ruled out and it is recommended to remove the person prior to the analysis. Furthermore, relatedness as well as population outliers and stratification is usually investigated in the quality control step of GWAS and can lead to further exclusions of individuals. For the interested reader, more detailed information about the different quality criteria listed above is given the appendix A.2. However, since population stratification plays an important role not only in quality control but rather in the analysis of GWAS data and also for this thesis, we will consider this in the following section.

3.2.4 Analysis of genome-wide association studies

Although GWAS analyses can build on valuable lessons learned from candidate gene association and linkage studies (Pearson and Manolio, 2008), they brought also new technological, practical and statistical challenges (Thomas, 2010c). Managing the enormous amount of data was one of the first practical aspects (Neale and Purcell, 2008; Thomas, 2010c), needing large computer capacities with respect to CPU time and storage (Ziegler *et al.*, 2008). Sophisticated statistical, but also bioinformatical tools for analyzing and interpreting the data were necessary (De Bakker *et al.*, 2005; Clayton and Leung, 2007; Falush *et al.*, 2003; Marchini *et al.*, 2006, 2007; Price *et al.*, 2006; Pritchard *et al.*, 2000; Scheet and Stephens, 2006; Stephens *et al.*, 2001; Teo *et al.*, 2007; Wellcome Trust Case Control Consortium (WTCCC), 2007), skills from computer sciences essential. Due to the high number of SNPs, quality control checks need to be performed in a highly automated way, as well as the following association analyses. In this section we will outline the most important steps in the analysis of a genome-wide association study. We will start with a short description of the single step analysis methods that in general build the first step in a GWA analysis, as in the lung cancer studies of chapter 7. In addition, the pathway based methods illustrated in chapter 5 are based such results. After this, different methods to correct for the most important

confounder in genetic epidemiological studies, population stratification, are explained. In our lung cancer application in chapter 7, a corresponding adjustment for a particular study is shown in detail. Furthermore, the two most common graphical representation methods for GWAS results are presented. Finally, we will outline how significance of association signals in GWAS is assigned and how their validity is judged, since this is the main challenge in GWAS. The hierarchical method that we address with this thesis tries to improve GWAS at that point of assessing significance. The applications presented in chapter 7 are done in the scope of a consortium (Amos, 2007; International Agency for Research on Cancer (IARC), 2012), with the aim of good validation of results.

Single SNP association tests

For the association analysis, the standard first step in GWAS is to perform simple single SNP association tests (Bickeböller and Fischer, 2007). Most commonly, an additive model or a trend test is used (Pearson and Manolio, 2008). In addition, since the underlying genetic model is unknown, it prevailed to perform tests for all three standard models dominant, additive and recessive and to use their maximum. By permutation methods (Freidlin *et al.*, 2002) or a conditional test taking the correlations of the statistics into account, an adjustment can be performed (Ziegler *et al.*, 2008). Furthermore since confounders play an important role in population-based association studies, the consideration of those is an important point to avoid spurious associations. Confounders can be already considered in the study design and recruitment, e.g. by choosing homogenous groups. Alternatively, they can be integrated in the analysis by stratification or adjusting, e.g. in a regression model. One of the main confounders in population-based association studies is population stratification, also called "confounding by ethnicity" (Ziegler *et al.*, 2008). We will consider this phenomenon in more detail in the next section.

Population stratification

Population stratification is a population heterogeneity based on the presence of multiple populations or subgroups according to ethnicity or geographic origin involved in a population-based association study, where the disease prevalence differs between the subgroups and the frequencies of the genetic marker alleles and LD patterns between the markers vary (Cavalli-Sforza *et al.*, 1994; Dawson *et al.*, 2002; Hirschhorn, 2005; Jorde *et al.*, 1994; Patil *et al.*, 2001; Phillips *et al.*, 2003; Shifman *et al.*, 2003; Teo, 2008; Watkins *et al.*, 1994; Zavattari *et al.*, 2000). When population structures are undetected and not accounted in the analysis, this provides a serious issue since the variation in disease rates across the groups and the different allele frequencies have the potential to result in inflations of the test statistic (Ziegler *et al.*, 2008). This may lead to spurious associations (Palmer and Cardon, 2005). If e.g. cases tend to be over-sampled for one of these groups, all alleles more common in that group will appear to be associated with the disease. Already before the GWAS era, the problem of population stratification was widely debated (Cardon and Palmer, 2003; Freedman *et al.*, 2004; Thomas and Witte, 2002; Thomas *et al.*, 2005; Wacholder *et al.*, 2000). The simplest method to account for population stratification, the genomic control approach (Devlin and Roeder, 1999; Devlin *et al.*, 2001), corrects for the stratification without identifying the sample structure. Since population stratification leads to an

overdispersion of the statistics, the degree of inflation of the test statistics and hence the extent of population heterogeneity can be estimated (Devlin and Roeder, 1999). Therefore, χ^2 association test statistics for all SNPs are calculated and the median over all SNPs is compared with the expected theoretical median of the distribution under the null hypothesis of no association. Since the fraction of false positive results is expected to be increased, the quotient of the observed and the expected χ^2 median, denoted as inflation factor λ, is expected to be > 1 (Devlin and Roeder, 1999). All test statistics are furthermore corrected for the inflation by dividing them by the inflation factor, hence resulting in an adjusted test of association. In a study of Nelis *et al.* (2009) about the genetic structure in Europe inflation factors for the comparison of 19 samples from 16 European countries were calculated. Between Southern and Northern Germany (KORA and POPGEN cohorts) they obtained a lambda value of 1.08, both being close to the European population of Hap Map with 1.06 and 1.07 respectively. The largest genetic distance was observed for Spain and Kuusamo located in the middle of Finland (4.21), with these having inflation factors of 1.34 and 2.89 with the European Hap Map population. For the different Hap Map populations, inflation factors of 21.56 for the African and Asian population and a slightly smaller one for the African and European population, 13.27 for the European and Asian population and 1.77 between the two Asian populations were calculated. While an inflation factor of less than 1.05 is still acceptable, for higher factors a correction is recommended (Aulchenko, 2010). Another possible strategy is to initially identify the underlying population structure by determining the genetic similarity between the individual participants and then correct for the particular structure. Therefore, the SNP data should be pruned first so that only SNPs with no strong LD among them remain. A measure typically used to express the genotypic similarity between two individuals is the kinship coefficient. The kinship coefficient is defined as the probability that an allele of a particular locus that is randomly chosen from an individual is identical by descent (IBD) with an allele selected from the same locus of the other individual. Two alleles are IBD when they are copies of the same ancestral allele. The kinship coefficients for all pairs of individuals are collected in the kinship matrix and that matrix can be used as a part of the model for the correlations of the outcomes in a random effects model (Yu *et al.*, 2006). Alternatively, a principle component analysis (PCA) (Patterson *et al.*, 2006; Price *et al.*, 2006; Tian *et al.*, 2008; Tiwari *et al.*, 2008)) based on 0.5 - the kinship matrix (distance matrix) can be performed. The leading eigenvectors, the principle components (PC), can be extracted and describe informative "axes of ancestry". These axes can be represented graphically so that different populations can be identified. Furthermore, to correct for the population stratification that may exist in the data, the axes can be used as covariates in the subsequent association analysis (Patterson *et al.*, 2006; Price *et al.*, 2006; Tian *et al.*, 2008; Tiwari *et al.*, 2008; Weale, 2010). The number of PC axes to consider in the analysis can be yield by testing them for statistical significance (Weale, 2010). However, the PCA can not only detect and correct for correlations due to ancestry, but also any source of correlation in the data. Lab errors, e.g. systematic genotyping artifacts and many high-effect causal SNPs in a case-control study can be picked up as well (Weale, 2010). To clarify the source of correlation, it is possible to use external reference populations in the analysis and see how the study individuals cluster to these. This can be done PCA, but also another

approach, denoted as STRUCTURE (Pritchard *et al.*, 2000), is dealing with this very issue. In STRUCTURE, the study sample is compared with reference populations and each individual is assigned to one of these populations (Pritchard *et al.*, 2000). The population membership is determined, but also outliers, migrants and admixed individuals can be identified, not clearly belonging to one of the distinct populations. Based on the obtained knowledge, a stratified analysis can be performed. When the ancestry of all individuals is already known by the reported geographic location or ethnicity, a stratified approach can be conducted as well. The approach of genomic control has the disadvantage that a constant multiplicative factor is used to correct each SNP test statistic. This assumes that the existing population structure has an uniform influence across the whole genome (Teo, 2008). Hence, the method fails when the stratification affects certain SNPs more than the average (Teo, 2008). In comparison, the PCA correction is adjusted to each SNP individually, e.g. by the magnitude of SNP variation along each axis of ancestry, and hence corrects not only for false positive but also false negative results. PCA has been shown to be more powerful than genomic control or structured association analysis. Furthermore, it is fast to implement, intuitive and appealing (Price *et al.*, 2006), so that using PCA axes as covariates is the preferred method for handling population stratification in large genetic studies. Nevertheless, PCA and a stratified analysis have to be treated with caution when cases and controls come from different source populations. In that case the covariates could take up all the possible variance between case-control status, including true association effects. In the study of Nelis *et al.* (2009) the genetic structure within Europe showed a clear correlation with the geographic location, with the first two PCs representing the genetic diversity from northwest to southeast. In 2006, (Steffens *et al.*, 2006) investigated the genetic substructure in the German population and observed that only minor degree of population substructure (Ziegler *et al.*, 2008) exists. Nevertheless, the larger the sample size of a GWAS, the more susceptible is the study to confounding from finer levels of population differences (Teo, 2008). Hence, the greater is the potential bias from the stratification (Freedman *et al.*, 2004; Marchini *et al.*, 2004).

Visualization of GWAS results

Two popular graphical representations of GWAS results are the Manhattan plot and the Quantile-Quantile-plot (QQ-plot). The Manhattan plot is a type of scatter plot that allows the display of a high number of data points as given in genome-wide association studies. It provides a visual summary of the association test results for the examined SNPs and clearly highlights (*regions of*) significant markers. The plot displays the negative logarithm ($-log_{10}$) of the p-values for the single SNPs on the y-axis as a function of the chromosomal location on the x-axis. For visual effect, the different chromosomes are shown as blocks of different colors. Since the strongest associations have the smallest p-values, the corresponding $-log_{10}$ will be greatest, so that SNPs with significant p-values will stand out. In figure 3.1 a Manhattan plot for a meta-analysis of 4 different lung cancer GWAS is shown. We can see a clear peak by numerous neighboring SNPs on chromosome 6, as it is expected by a truly associated region. Furthermore, several genome-wide significant loci close to each other show up on chromosome 15, pinpointing to another truly positive hit. On chromosome 5, 10

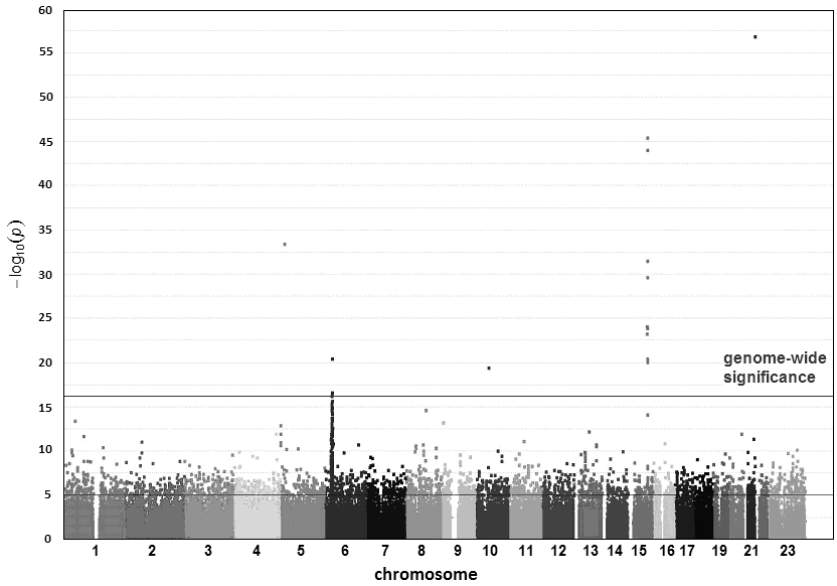


Figure 3.1: *Manhattan plot for a meta-analysis of four different lung cancer GWAS: Central Europe lung cancer GWAS of the International Agency for Research on Cancer (IARC, Prof. Brennan), Toronto lung cancer GWAS of the University of Toronto and the Samuel Lunenfeld Research Institute (SLRI, Prof. Hung), Texas genome-wide lung cancer study conducted by the M.D. Anderson Cancer Center (MDACC, Prof. Amos, Prof. Spitz) of the University of Texas and UK lung cancer GWAS at the Institute of Cancer Research (ICR, Prof. Houlston) [with kind permission of Prof. Amos]*

and 21 we can see some more SNPs that reach genome-wide significance. Since these are only single SNPs standing out, they rather indicate false positive results than true associations. The Quantile-Quantile plot (figure 3.2) is a useful tool to check the quality of the data on the one hand and assess the number and strength of the observed associations on the other hand (Pearson and Manolio, 2008). Therefore, the expected distribution of the association test statistics across all SNPs under the null hypothesis of no association (x-axis) is compared to the observed value in the data (y-axis) (Pearson and Manolio, 2008). In GWAS it is assumed that the vast majority of the genotyped SNPs is not associated with the disease. Hence, their test statistics follow the null distribution and only a minor deviation from the diagonal in the QQ-plot should be observed. Only a handful of values that deviate in the upper tail of the distribution may represent SNPs with strong evidence for a true association (Pearson and Manolio, 2008). For diseases highly associated with SNPs in a heavily genotyped region, such as Rheumatoid Arthritis associated with the HLA region on chromosome 6p21 (Pearson and Manolio, 2008), stronger deviations can be observed. However, large deviations of the observed values indicate consistent differences between the cases and controls across the whole genome. This systematic bias in the data can be due to e.g. relatedness, population stratification or genotyping artifacts (Pearson and Manolio, 2008). By filtering the data according to the different quality criteria listed in the previous section and correcting for population stratification this type of bias can be avoided. Other confounders, such as smoking in lung cancer can inflate the distribution as well when not considered in the analysis. In figure 3.2 a QQ-plot for a lung cancer GWAS that is

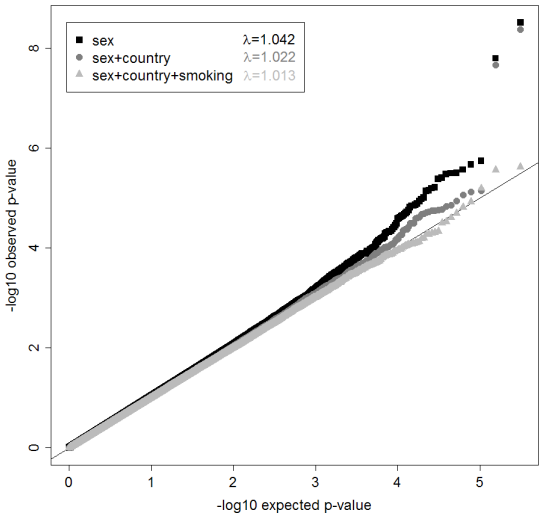


Figure 3.2: *QQ-plot for three different regression models analyzing the Central Europe lung cancer GWAS of the International Agency for Research on Cancer (IARC, Prof. Brennan). The study comprises individuals from six different central and eastern European countries. The first regression model involved only sex as a covariate, model 2 includes the country by five dummy variables. The last regression model additionally involved smoking status (never-ever) and packyears. [with kind permission of Prof. Brennan]*

composed of 6 different central and eastern European countries is shown to demonstrate the effect of population stratification and smoking as important confounding factors. Therefore, results from three different logistic regression models for the genetic markers are compared: including only sex as a covariate, including sex and country as well as including sex, country and smoking. We can clearly see that the inflation of results is reduced by including both confounders in the analysis. This is also reflected by the corresponding inflation factors of $\lambda_{GC} = 1.042$, $\lambda_{GC} = 1.022$ and $\lambda_{GC} = 1.013$.

Assigning significance of association signals and validity

The major challenge in GWAS is not the type of test to use but the number of tests. Performing hundreds of thousands of tests brings a high computational and statistical multiple testing burden. In addition, this can be complicated by having multiple phenotypes, considering further modifiers to the basic analysis (Neale and Purcell, 2008) or testing within subgroups. In a study of 500,000 SNPs conducting one single SNP test for each of them, 25,000 false positive results are expected using a nominal significance level of 5%. Therefore assessing the significance of the SNPs is an important issue in GWAS. Numerous methods to address the multiple testing problem were developed long before the GWAS era. These methods can be distinguished in methods controlling the **family wise error rate (FWER)** or the **false discovery rate (FDR)**.
The FWER is the probability of making at least one false positive result and the classical methods for controlling the FWER are characterized by simple p-value adjustment and post hoc corrections. Different popular multiple testing methods to control for the FWER are given by Bonferroni (1935, 1936), Hochberg (1988), Holm (1979b,a), Hommel (1988), Rom (1990) and Sidak (1967). In contrast, the FDR is the expectation

of the proportion of false positive results in all reported associations (Benjamini and Hochberg, 1995; Forner *et al.*, 2008). Hence, for controlling the FDR the proportion of significant associations that are actually false positives needs to be estimated (Hochberg and Benjamini, 1990; Yang *et al.*, 2005). Commonly used methods are by Benjamini and Hochberg (1995), as well as Benjamini *et al.* (2001). Controlling the FDR is less conservative than controlling the FWER (Benjamini *et al.*, 2001; Efron and Tibshirani, 2002; Sabatti *et al.*, 2003). In 2003, Storey and Tibshirani (2003) published a modified version of the FDR, called positive false discovery rate (pFDR),motivated by a Bayesian viewpoint. In this context, the **q-value** was introduced as the Bayesian posterior p-value (Storey, 2003; Wakefield, 2007, 2008). This method was further developed by Storey (2002) and Storey and Tibshirani (2003).
The methods listed so far assume independence between the different tests statistics. However, in a GWAS a large correlation between the tests of different SNPs exists due to linkage disequilibrium (Evans and Cardon, 2006; Pearson and Manolio, 2008; Ziegler *et al.*, 2008). Thus the effective number of tests in GWAS is substantially smaller. Zondervan and Cardon (2007) proposed to assume a certain constant LD across all SNPs to adjust for the effective number tests (Ziegler *et al.*, 2008). The gold standard to account for correlation structures however are resampling methods such as bootstrap or permutation based approaches. These generate appropriate experiment-wide p-values (Churchill and Doerge, 1994; Neale and Purcell, 2008; Ziegler *et al.*, 2008). Troendle (1996) and Westfall and Young (1993) proposed permutation based methods for controlling the FWER. Resampling approaches can be applied to the different FDR methods as well. Unfortunately, permutation based methods are highly computer demanding (Thomas, 2010c). A good overview of resampling methods for multiple testing is given in Ge *et al.* (2003).
In Dudoit *et al.* (2003, 2008); Dudoit and Laan (2008) and Tusher *et al.* (2001) several multiple testing techniques, encompassing FWER and FDR methods with and without correlation consideration in the context of gene expression analysis are explored. Although FDR methods in general are characterized by higher power than FWER strategies, in the context of GWAS, FDR controlling methods show no remarkable difference to FWER approaches (Dudoit *et al.*, 2003; Ziegler *et al.*, 2008), since the number of true positive results is generally expected to be very low and the number of false positive is too high. Hence, in the context of GWAS, all these methods are really conservative. Extremely small p-values are required to achieve significance (Evans and Cardon, 2006) and most GWAS studies are underpowered to achieve such stringent significance levels for true positives (Evans and Cardon, 2006; Hunter and Kraft, 2007). For causal variants in complex disease, relative risks of 2 and less are in general expected and true positive association signals are not necessarily larger than background noise or confounding effects (Teo, 2008). Detecting true associations in a GWAS is as looking for the needle in a haystack (Thomas, 2006). The investigation of the power in GWAS by Eberle *et al.* (2007), Gail *et al.* (2008), Nannya *et al.* (2007) and Thomas *et al.* (2009) showed that 1,000 cases and 1,000 controls are necessary to detect SNPs associated with a disease of OR of 1.7-2. The power to detect a SNP with an OR of 1.3, a MAF of 0.2 given 1,000 cases and 1,000 controls using a significance level of 10^{-6} is only 4% (Ziegler *et al.*, 2008). Hence, enormously large studies are necessary to detect variants of small OR using such stringent significance levels (Bickeböller and Fischer, 2007).

Although false positive results are the most visible problem in GWAS (Weale, 2010), false negative results may be even more difficult to correct (Weale, 2010). Since GWAS rather discover potential SNPs worthy for further investigations than end results (Ziegler *et al.*, 2008), false positive results are widely accepted (Hirschhorn, 2005) to avoid missing the true positive results, even though this leads to increased costs in follow-up studies (Ziegler *et al.*, 2008). Strategies common in praxis are to use a weak significance level, e.g. 10^{-4} (corresponding to 50 expected false positive associations out of 500.000 SNPs) as proposed by (Arking *et al.*, 2006). Alternatively, a particular proportion of most promising SNPs from the analysis can be chosen for further investigation, e.g. 10% of the top markers or the top 500 (Neuman and Sung, 2009; Satagopan *et al.*, 2002; Skol *et al.*, 2006). These top SNPs can be a selection of the highest ranked SNPs according to the p-values, but also according to the population attributable risk (PAR) (Hunter and Kraft, 2007; Thomas, 2010c). The latter is an estimate for the fraction of diseased people that could be avoided if the exposure (a particular genetic variant) would not exist. It measures the relative contribution of a particular genetic variant to the disease by combining information about the effect size and the frequency of the genetic variant (Hunter and Kraft, 2007). A plausible argument for a real association is when two or more SNPs in modest or strong LD in a chromosomal region show an association even for a weak threshold (Ziegler *et al.*, 2008).
A common practice in GWAS is the usage of two-stage designs (Hirschhorn, 2005; Thomas *et al.*, 2009; Thomas, 2010c; Ziegler *et al.*, 2008). In the first step, the screening process, a dense set of genotyped markers is tested in an initial part of the study sample to prioritize SNPs for the successive stage. In step two, the focus is on a small subset of most promising SNPs from step one to evaluate their effect, genotyped in the remaining sample. While some of the proposed two stage designs use the second step as a form of a statistical "built-in" replication (Thomas, 2006) of the first stage (Kraft, 2006; Saito and Kamatani, 2002; Satagopan *et al.*, 2002), others suggest to perform a joint analysis combining the information from both stages (Bukszár and van den Oord, 2006; Satagopan and Elston, 2003; Satagopan *et al.*, 2004; Skol *et al.*, 2006; Thomas *et al.*, 2004; Wang *et al.*, 2006; Yu *et al.*, 2007; Ziegler *et al.*, 2008). The joint analysis proved to be more powerful than the replication based two-stage analysis (Skol *et al.*, 2006; Yu *et al.*, 2007) since the additional information from the first stage is considered. The advantage of the multistage design is the increase in efficiency since the same power can be obtained with reduced genotyping effort or less sample (Satagopan *et al.*, 2002). Due to the necessity to separate numerous false positive results in genome-wide association studies from the few true positives, the consistency and coherence of results in this context is of high importance to establish a causal relationship (Bradford Hill criteria). Therefore, independent replication and follow-up studies are essential to confirm results and make them reliable (Neale and Sham, 2004; Palmer and Cardon, 2005). Subsequent experimental studies are necessary such as functional tests in knockout/-in animal models or expression analyses. The latter should be performed for replicated variants to understand the biological function (Thomas *et al.*, 2009) and hence the molecular and physiological basis of the disease. To avoid failures in replication because of true differences between the original and follow-up populations, heterogeneity between the different study samples should be minimized or avoided (Thomas *et al.*, 2009). For the confirmation, other GWAS, but also small scale confirmatory studies are

possible (Ziegler *et al.*, 2008). A further aspect after replication is the generalization of the finding by investigating them in other studies allowing for between study heterogeneity (Thomas *et al.*, 2009). This heterogeneity may comprise different populations, different subtypes of the same disease (e.g. subgroups according to age of onset or family history), other intermediate endpoints or even other diseases as well as different study designs (Thomas *et al.*, 2009). The generalization can be useful to confirm, refine and extend the initial finding. Fine mapping and resequencing of interesting regions are furthermore needed to find out if the detected SNP is the causing factor or another variant that is in LD (Thomas *et al.*, 2009). Not the strength of the p-value in an initial study is important, but the consistency and strength across several replication studies (Hunter and Kraft, 2007). Results from multiple studies can be combined in a meta-analysis, e.g. by Fisher's combination of p-values (Neale and Purcell, 2008).

3.2.5 Problems in GWAS

The detection of genetic susceptibility factors for complex diseases should shed light on the genetic architecture of common human diseases (Ober and Vercelli, 2011). It should lay the foundation for prevention and early diagnostic methods, as well as safer and more effective treatments and better prognostic tools (Khoury, 2010). In the last years, genetic factors for several complex diseases were identified. Through June 2011, 1,449 genome-wide associations with $p \leq 5 \cdot 10^{-8}$ were published for 237 traits in more than 900 publications (Hindorff *et al.*, 2012). However, although GWAS were successful in finding genetic risk factors, they were not that satisfying as expected. On the one hand, failures in replication as already observed in candidate gene association studies are high on the agenda (Ioannidis, 2007). On the other hand most susceptibility factors identified and replicated so far explain only a small proportion of the disease's heritability (Ober and Vercelli, 2011).

Lack of replication

The lack of replication and reproducibility is discussed in many publications (Bickeböller and Fischer, 2007; Botstein and Risch, 2003; Hirschhorn, 2005; Ioannidis, 2007; Palmer and Cardon, 2005; Pearson and Manolio, 2008; Thomas, 2010c). Beside false positive results in initial studies (Palmer and Cardon, 2005), many true genetic susceptibility factors fail to replicate as well. The sample size in replications studies is often too small and hence the power is not adequate to detect the risk variants (Palmer and Cardon, 2005; Pearson and Manolio, 2008). Genotyping errors, cryptic relatedness and population stratification may contribute to replication failures (Pearson and Manolio, 2008). Differences between populations such as different effect sizes, allele frequencies and specific LD patterns (Bickeböller and Fischer, 2007; Palmer and Cardon, 2005) as well as the allelic and genetic heterogeneity in complex diseases make the replication even more difficult (Bickeböller and Fischer, 2007; Palmer and Cardon, 2005) and may lead to different results in different studies. Heterogeneity between the different studies, e.g. due to different phenotype definitions, study designs or SNP chips used (Bickeböller and Fischer, 2007; Pearson and Manolio, 2008) as well as different statistical methods (Working Group on Replication in Association Studies *et al.*, 2007) increase the problem of replication further.

Missing and hidden heritability

Heritability is the proportion of the total phenotypic variation between the individuals of a population that is attributed to genetic factors. Most of the non-major genetic variants identified so far are characterized by weak effects and explain only a small proportion of the corresponding disease heritability (Janssens and van Duijn, 2008; Pearson and Manolio, 2008). For most complex diseases a large fraction of the heritability is still unknown. We can differentiate **missing heritability** from **hidden heritability** (Ober and Vercelli, 2011). Missing heritability denotes that traditional GWAS may miss genetic variants such as rare or structural variants (e.g. CNV) and interactions (Maher, 2008; Manolio *et al.*, 2009; Ober and Vercelli, 2011), as well as chromosome changes not determined by DNA sequence modifications (e.g. DNA methylation) that are investigated in the so called epigenetics (Allis *et al.*, 2007; Rakyan *et al.*, 2011). In contrast, hidden heritability (Gibson, 2010) denotes that the joint effect of several common risk factors might exceed the simple sum of their individual SNP effects (Park *et al.*, 2008; Yang *et al.*, 2010). For most complex diseases we expect that all of the different components mentioned above are of high importance, and that they probably all will contribute to the genetic architecture of complex disease (Ober and Vercelli, 2011). Therefore, a new post-GWAS era, considering more than only single common SNPs is necessary.

3.2.6 The post-GWAS era

As already mentioned in the introduction, different new directions can complement GWAS analyses of single SNPs, summarized as post-GWAS research. Post-GWAS research is defined by the National Cancer Institute (NCI) of the US National Institutes of Health (NIH) as "the transition from the initial GWA discovery to replication studies, epidemiologic examination of gene-gene and gene-environment interactions, and to the biological validation of the GWAS findings". Some of the methods complementing the traditional single SNP analyses still stay in the context of genome-wide association studies but consider more than only one SNP, while others go even beyond the GWAS context. Based on the traditional GWAS data, haplotype and multilocus methods may help to reveal the hidden heritability underlying complex diseases (Gibson, 2010). While haplotype methods examine segments of DNA strands comprising several nearby SNPs, multilocus methods consider several loci at once, e.g. in a regression model. It is possible to pass over from the SNP to the gene level by performing a joint analysis for all SNPs of a gene, or combine single SNP results to gene level statistics. Gene set methods go one step further by identifying even whole significant groups of genes or pathways instead of SNP markers. This may unite results from different studies and help to understand the underlying mechanisms of the disease. Furthermore, results of GWAS can be improved by integrating external information, e.g. the location or possible function of the genetic variants, information from complementary disciplines such as gene expression, as well as information about the interplay of different genes within biological pathways. By using the pathway knowledge, true positive results can be supported in the analysis. Beside, interactions are another important component in complex disease, as already discussed in section 2.4.2, indicating that interactions may account for a large proportion of the heritability. Neglecting the interplay of genetic factors with each other and environmental factors will result in incomplete risk profiles and missing heritability. Still in

the context of GWAS, but beyond the scope of today's SNP chips is the investigation of other types of variants. Rare variants and structural variants such as CNVs, short insertions and deletions, as well as translocations may be important to find some of the missing heritability. As an opposite to the CDCV hypothesis that builds a fundamental of today's GWAS, the CDRV hypothesis (common disease - rare variant) arises. The CDRV postulates that common diseases are rather caused by rare variants occurring in less than 1% of the population with much larger effects (Pritchard, 2001). CNVs for example may change the gene dose or level of gene expression (Christensen and Murray, 2007; Ober and Vercelli, 2011; Pearson and Manolio, 2008) and therefore may play an important role in disease development. More information about the importance of rare variants in complex diseases can be found in Manolio *et al.* (2009), a good overview for the analysis of rare variants can be found in (Asimit and Zeggini, 2010; Basu and Pan, 2011; Dering *et al.*, 2011; Sun *et al.*, 2011). Other disciplines such as epigenetics leave the context of GWAS but should not be disregarded (Grimm *et al.*, 2011). Epigenetics denotes the study of changes in gene expression independent from the DNA sequence that are nevertheless inherited from a cell to its children. Methylation and acetylation for example influence the activity of chromosome segments and hence affect gene expression. Gene expression is an additional important factor to complement GWAS results, as well as functional studies in animal models or experiments on a protein basis. To simplify the replication of results and the performance of meta-analysis, data sharing is encouraged to provide maximal information about association evidence (Neale and Purcell, 2008). Furthermore, the collaborative work in consortia is recommended. Thereby, consistency across the different analyses of the participating investigators can be assured and study heterogeneity can be minimized. Sample sizes can be enlarged, meta-analyses can be improved and higher power in finding the genetic disease risk factors can be achieved. Our applications in chapter 7 were performed within an international consortium for lung cancer research. The investigation of multiple related phenotypes, e.g. asthma together with lung function and related intermediate immunological phenotypes (Los *et al.*, 2001), colorectal polyps together with colorectal cancer (Croitoru *et al.*, 2004) or diabetes and related metabolic syndrome traits (Saxena *et al.*, 2007), may be useful, since related traits may show similar results and joined conclusions about underlying mechanisms.

In this thesis, we will concentrate on two of the complementary approaches based on the current genome-wide genotyping data that incorporate the complexity of diseases: the integration of biological pathway information into the analysis and the consideration of GxE interaction. The importance of pathways and gene x environment interactions in complex disease is already discussed in section 2.4, indicating that interactions may account for a large proportion of the heritability. Neglecting the interplay of genetic factors with each other and the environment will result in incomplete risk profiles and missing heritability. Due to one focus on GxE interactions, we will give a short overview of the benefits and challenges to detect GxE interaction in the genome-wide context in the following section. A good summary about these so called gene-environment-wide interaction studies (GEWIS) can be found in (Thomas, 2010a).

3.3 Gene-environment wide interaction studies (GEWIS)

An important component when confronting rather than ignoring the complex etiology of common diseases is the examination of GxE interactions (Moore, 2003). Traditionally, interactions between genetic and environmental factors were investigated in candidate gene studies. In particular, genes within biological pathways that are known to involve the environmental factor, e.g. a pathway responsible for the metabolism of the exposure, provide popular candidates to consider. In GWAS, one first approach to consider GxE was to test only those SNPs showing a genetic main effect. However, we will concentrate here on the integrated analysis of genome-wide variation and environmental factors (Khoury and Wacholder, 2009), scanning the whole available SNP data for GxE interactions (Thomas, 2010a). The analysis of GxE in the genome-wide context is denoted as gene-environment-wide interaction studies (GEWIS) (Khoury and Wacholder, 2009) and provides a complementary and important avenue of investigation (Ober and Vercelli, 2011).

3.3.1 Benefit of detecting GxE interactions in complex diseases

The investigation of GxE is worth for many reasons (Marchand, 2005; Marchand and Wilkens, 2008; Thomas, 2010a). As already mentioned in section 3.2.5, GxE interactions may account for some of the missing heritability of most complex diseases (Ober and Vercelli, 2011; Thomas, 2010a). The joint effect of a genetic and environmental factor may explain a larger proportion of the heritability (Thomas, 2010a) than the genetic main effect on its own and may even help to identify novel genetic factors without a main effect (Thomas, 2010a). Sources of heterogeneity across different studies (Greene *et al.*, 2009; Ioannidis, 2007) for the same disease may be detected, explaining failures in replication of GWAS findings (Thomas, 2010a). The revealing of GxE may substantially contribute to our understanding of the biological mechanisms underlying the development of complex diseases (Khoury and Wacholder, 2009; Thomas, 2010a) by providing insights into disease complexity and involved pathways (Thomas, 2010a). Understanding the underlying pathway may further help to determine which compounds in a complex mixture of environmental factors causes diseases (Hunter, 2005), e.g. when the identified gene of the interaction is involved in the metabolism of one of the components. In colorectal cancer for example, the interaction of the gene NAT2 with the intake of red meat cooked at high temperature indicates heterocyclic amines as the causing component (Thomas, 2010a). Additionally, interactions may support the credibility of environmental factors. Another exposure suspected to be an important cause in colon cancer are the polycyclic aromatic hydrocarbons (PAH), which are formed in red meat as well, but can also be found in cigarettes smoke and exhaust fumes. The identification of interactions with genes involved in the PAH metabolism enhance the credibility of a causal relation of PAH and cancer (Brennan, 2002). The identification of GxE may further help to find environmental factor that only affect individuals with a particular genetic predisposition (Thomas, 2010a) or to identify high-risk individuals (Brennan, 2002). Prediction models may be improved by the knowledge of GxE. New prevention strategies may be derived and the reduction of an environmental exposure to prevent disease for example may be restricted to carriers of an interacting genetic factor that

contributes to disease susceptibility. New therapeutic agents may be developed and personalized depending on the underlying genetic information (personalized medicine) to minimize adverse drug reactions and treatment failures while maximizing the response (Hunter, 2005; Thomas, 2010a).

3.3.2 Challenges of GEWIS

The conduct of gene-environment-wide interaction studies is much more difficult than the examination of purely genetic associations (Thomas, 2010a). One major additional challenge is the necessary careful collection of high-quality environmental data (Thomas, 2010a). The assessment of environmental factors is complicated due to several different reasons (Khoury and Wacholder, 2009). Many environmental factors have a multidimensional character, e.g. air pollution consisting of several different gases and particles with different biological effects (Thomas, 2010a). Environmental exposures can furthermore show different intensities, and in comparison to the genetic information, the environmental influences may vary over time (Khoury and Wacholder, 2009; Thomas, 2010a). The age at exposure as well as the duration may have an impact (Thomas, 1988, 2010a). Accurate measures of exposure may not always be possible and uncertainties in the environmental factors may occur, leading to spurious interactions (Thomas, 2010a). However, the obstacles on the environment side are not new, since for a long time epidemiological studies investigating environmental factors are familiar with these problems. Therefore, a multidisciplinary collaboration is useful to ensure good GxE studies (Hunter, 2005). Standard study designs (Thomas, 2010a) from epidemiology can be used. Another aspect already discussed in the context of GWAS is the problem of replication. In GEWIS, heterogeneity between different studies is expected to be even more severe, due to different measures of exposure, different distributions, characteristics or even definitions of the environmental factors (Thomas, 2010a). Furthermore, different confounding factors may exist. With respect to the environmental factor smoking for example, the definition of "never-smoker" is not that obvious and may in some studies only include individuals that never took a pull on a cigarette, in others all people that smoked less than a small number of cigarettes in their whole life, e.g. 100. Therefore, especially in the context of GEWIS, consortia play an even more important role, since they may ensure harmonization of study designs, exposure assessment and analysis methods across the studies already at the stage of study planning (Brennan, 2002; Thomas, 2010a). Another problem that is even bigger in GEWIS in comparison to GWAS is the lack of power to detect the influencing factors. To detect interactions, even larger sample sizes than for main effects are necessary (Ober and Vercelli, 2011). It is even more important to reproduce an interaction finding in two or more studies and to find a plausible explanation at biological level (Hunter, 2005). To solve the problem of low power, several methods were suggested so far. These are addressed in chapter 6, where in addition a new test to detect GxE interactions by a hierarchical Bayes model is derived. As a basis for this, the next chapter provides the fundamental concepts of Bayesian theory and hierarchical models needed for this approach.

4 The Bayesian approach and hierarchical modeling

The focus of this thesis is an empirical hierarchical Bayes model to consider the complexity of common diseases in genome-wide association studies. Empirical hierarchical Bayes models are characterized by a hierarchical modeling framework combined with a Bayesian flavor given by the posterior based inference. The hierarchical structure has the ability to easily include multiple sources of external information into the analysis. The empirical Bayesian treatment allows to exploit a large amount of data, as given in GWAS, more effectively by "borrowing information" between the different observations. We used the empirical hierarchical Bayes model for two aims: the integration of pathway information in GWAS of complex diseases and analysis of GxE interactions in the genome-wide context. Due to the property of hierarchical models to involve external information, they are excellently suited to consider biological pathway data in the analysis. Furthermore, in the context of GxE interaction analysis, the so called shrinkage estimators obtained by the empirical Bayesian approach may help to increase the power of detecting the interactions. Reason for that is that these shrinkage estimators are characterized by a reduced variance in comparison to frequentist estimation.
We will start this chapter with a description of the Bayesian approach to provide the reader the necessary basis for the following sections. We then move on to hierarchical Bayes models and empirical hierarchical Bayes models in section 4.2, before we will address the usage of Bayesian statistics and hierarchical modeling in the context of genome-wide association studies in section 4.3. Finally, section 4.4 presents the hierarchical Bayes model proposed by Lewinger *et al.* (2007) on which our extensions and applications are build.

4.1 The Bayesian approach

The **Bayesian approach** is an effective and practical alternative to the classical frequentist statistics for hypothesis testing and conducting statistical inferences.
In the classical statistical setting, an analysis is only based on the observed data captured by its conditional probability distribution given unknown parameters (likelihood). The data are treated as random, even though they are known and the parameters are viewed as unknown but fixed constants that are estimated by **maximum likelihood estimation (MLE)**, assuming a particular distribution.
In contrast, the Bayes approach is based not only on the observed data but also on information about the parameters to estimate, which is known before the analysis of the data, e.g. by previous studies. While here the observed data are treated as fixed, the parameters are considered as random variables with an underlying distribution function, the so called prior, specified by given a priori information. This given prior knowledge on the parameters is updated by the observed data to form the posterior distribution. If no prior information is available, this is adequately included in the analysis. In the Bayesian approach, **maximum a posteriori (MAP) estimation** can be used to achieve point estimates of the parameters instead of maximum likelihood estimation. Through the explicit use of the prior distribution, which is the essential characteristic of the Bayesian method, uncertainty on the parameters of interest can be quantified. This uncertainty is passed to the inference based on the analysis (Gelman *et al.*, 1995; Robert, 1994).

While the frequentist approach answers the question "What parameter values make the data most likely to occur?", the Bayesian analysis addresses the question "What parameter values are most likely given the data?" by using the inverse conditional probability. These very different viewpoints have always led to controversies between the frequentists and the Bayesians. The classical approach is criticized because a significance test does not answer the essential question if and with which probability a hypothesis is true based on the observed data. The Bayesian analysis is accused of being very subjective and unscientific because the prior distribution, the paradigm of Bayesian, is not only based on objective data but is influenced by a subjective perception and belief. Actually, the frequentist and the Bayesian approach can lead to different practical inferences, although based on the same observed data, because the prior in Bayesian analysis may have a strong influence on the outcome (Robert, 1994). However, Bayesians counter that even when different priors are used, the new evidence from observed data will tend to bring their posterior probabilities closer together.
Before we will go into details of the Bayesian model and inference, we will start with the fundamental equation and technical core of the Bayesian approach for parametric inference, the Bayes' theorem.

4.1.1 The Bayes' theorem

The Bayesian method is based on the well-established theorem of Reverend Thomas Bayes published in an essay of 1763 (Bayes, 1991) and relates two reverse conditional probabilities with each other (Robert, 1994). Most should be familiar with the discrete case of the Bayes' Theorem, given in its simplest form by

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}, \qquad P(B) > 0$$

representing the relationship of the conditional probability of an event A given an event B with the reverse conditional probability of event B given event A (Dehling and Haupt, 2004). The fundamental idea of the theorem is that the conditional probability of A given B depends not only on the relatedness of A and B, but also on the marginal probabilities for each of these events. By the theorem, the probability of the occurrence of event A is updated from $P(A)$ to $P(A|B)$ once B has been observed (Robert, 1994). A simple illustrative example from diagnostic testing is given by the probability $P(A|B)$ of having diabetes (A) given a positive result (B) in an urine glucose test (Werner, 1984). This probability depends not only on the accuracy of the urine test in finding the diseased people $P(B|A)$, but also on the prevalence of diabetes $P(A)$ and the probability for a positive test $P(B)$. The latter is composed of the sum of the probability of a positive test given diabetes $P(B|A)$ multiplied with the prevalence of the disease $P(A)$ and the probability of a positive test given no diabetes $P(B|A)$, multiplied with 1-prevalence $(1 - P(A))$. The prevalence for diabetes may vary e.g. depending on age or sex and hence the interpretation of a positive test result may change.
In the following, the more complicated continuous case of the Bayes' theorem for density functions is considered, because this builds the base for our purpose. Let x and y be two continuous random variables with **conditional probability density function** $f(x|y)$, the probability of x given y, and with **marginal probability density function** $g(y)$,

the probability of y regardless of x. The continuous version of the **Bayes' theorem** states that the conditional probability density function of y given x is

$$g(y|x) = \frac{f(x|y)g(y)}{\int f(x|y)g(y)dy}, \tag{4.1}$$

provided $\int f(x|y)g(y)dy \neq 0$ (Lee, 1997; Robert, 1994). A detailed derivation of this theorem can be found in Papoulis and Pillai (2002).
The numerator of that formula is the **joint probability density function** $h(x, y) = f(x|y)g(y)$ of x and y, a bivariate distribution giving the probability of the joint occurrence of x and y. Integrating the joint distribution over all possible values of y as done in the denominator yields the marginal probability density function $f(x)$ of x. This marginal probability can be interpreted as the probability of x regardless of y $f(x) = \int h(x, y)dy = \int f(x|y)g(y)dy$. Analogously, we have that $g(y) = \int h(x, y)dx$. This relationship of the marginal probabilities to the conditional ones is known as the **law of total probability**. Note, when both random variables are independent from each other, then $f(x|y) = f(x)$, $g(y|x) = g(y)$ and $h(x, y) = f(x)g(y)$. The marginal density $g(y)$ is denoted as **prior probability density function** of y, because it represents the prior belief about y neglecting any knowledge about x. The conditional probability $g(y|x)$ is denoted as **posterior probability density function**, because it reflects the probability of y considering the information given about x (Pestman, 1998). In the following, we will denote the different probability density functions shortly as prior, marginal or posterior. Furthermore, we will use the term distribution interchangeable for the probability density functions. Since conditioning on constants is not necessary, we will neglect fix values in the following notation. Traditional maximum likelihood estimates in the frequentist context will be marked by $\tilde{\cdot}$ while estimates in the Bayesian context will be marked by $\hat{\cdot}$. We will use $f(\cdot)$ for the model function, $m(\cdot)$ for a marginal function, $h(\cdot)$ for a joint distribution and $\pi(\cdot)$ for prior as well as posterior distributions.

4.1.2 The Bayesian model

In Bayesian inference, a prior probability of a hypothesis is combined with the compatibility of some observed data with this particular hypothesis, to determine the probability of the hypothesis given the observed data.
Consider a general problem where we want to specify a sampling **model** for N observations $y = (y_1, ..., y_N)$ depending on a vector of r unknown **model parameters** $\theta = (\theta_1, ..., \theta_r)$ in a known way. This dependency can be expressed in form of a probability density function $f(y|\theta)$, with $f(y|\theta) = \prod_{i=1}^{N} f(y_i|\theta)$ if the N observations $y_1, ..., y_N$ conditional on θ are independent from each other. The function $f(y|\theta)$ is a function of y that theoretically represents the probability to observe the data y under given fixed values of θ. However, in classical statistics it is regarded as a function of θ for fixed data y, also denoted by $L(\theta)$ and called **likelihood**, representing the probability to observe the given data y (Gelman *et al.*, 1995; Robert, 1994). The frequentist statistic is based on that likelihood. Estimates of the unknown parameters θ are yielded by choosing the values which maximize the likelihood (**MLE)** and hence make the data most likely to occur (Dehling and Haupt, 2004; Robert, 1994). For example, assuming normally

distributed data with known variance σ^2 and unknown expectation θ, the MLE for the expected value is given by the mean of the data $\tilde{\theta} = \frac{1}{N}\sum_{i=1}^{N} y_i$.
In the Bayesian context we are interested in the unknown quantities θ as well, but we do not want to know the parameters that make the data most likely to occur, but the parameters, that are most likely given the data. Therefore, we need to reverse the conditional probability $f(y|\theta)$ of y given θ to a conditional probability $\pi(\theta|y)$ of θ given y, what can be done by Bayes' Theorem. Hence, we need additional prior beliefs about the parameter values θ which we want to take into account, expressed in terms of a probability density function. We suppose θ is a random quantity, having a probability distribution $\pi(\theta)$, which is formalized by the available prior information. This function $\pi(\theta)$ is called **prior density function** of θ.
Regarding the prior information, two different interpretations can be opposed: the population interpretation and the knowledge interpretation. From the first perspective, the prior represents a population of possible parameter values from which the model parameters have been drawn. From the latter, more subjective viewpoint, the prior expresses the knowledge about the model parameters as if its values could be thought of as a random realization from a prior distribution (Gelman *et al.*, 1995).
Having specified the prior distribution and having observed the data, we can use Bayes' Theorem to transfer the prior belief about the model parameters before the observation into a posterior belief considering the new observed data.
Therefore, we multiply the prior $\pi(\theta)$ by the likelihood $f(y|\theta)$ (contribution of the observed data) to obtain the joint distribution $h(y,\theta) = f(y|\theta)\pi(\theta)$ of y and θ. Normalizing the joint distribution by the marginal $m(y)$ of the data, we obtain the posterior

$$\pi(\theta|y) = \frac{h(y,\theta)}{m(y)} = \frac{h(y,\theta)}{\int h(y,\theta)d\theta} = \frac{f(y|\theta)\pi(\theta)}{\int f(y|\theta)\pi(\theta)d\theta}. \tag{4.2}$$

In general, the posterior distribution has no closed form expression and in particular the computation of the normalizing constant $m(y)$ may be difficult due to the integration. Therefore, the unnormalized posterior density function simply given by model times prior $\pi(\theta|y) \propto f(y|\theta)\pi(\theta)$ is often used (Gelman *et al.*, 1995). The posterior distribution is the main tool of the Bayesian inference (Robert, 1994). It summarizes the current state of knowledge about the parameter of interest by updating or weighting the prior opinion formalized by $\pi(\theta)$ according to the new evidence given by the experimental data represented by the likelihood $f(y|\theta)$ (Gelman *et al.*, 1995; Robert, 1994).
In comparison to simple maximum likelihood (ML) parameter estimates, the Bayesian approach results in a whole distribution for the parameters from which any information can be extracted and inferences concerning can be made. To give a closer understanding of the Bayesian approach, two common simple models and corresponding examples are illustrated in the following. Since the examples will be taken up later in this chapter, we will numerate the different parts by 1a), 1b) and 2a), 2b).

Example 1a): Normal-normal model
(Gelman *et al.*, 1995)

The simplest combination of model and prior is using normal distributions for both of them. Assume that a woman wants to know her diastolic blood pressure and obtains a value of y in a blood pressure measurement. This value is normally distributed with

unknown mean θ presenting her true blood pressure and known variance σ^2 due to measurement errors

$$y|\theta \sim N(\theta, \sigma^2).$$

In a frequentist manner, the observed score would be used as an estimate for her true blood pressure. However, experts say that the diastolic blood pressure in women of that age in general is a random variable with known mean μ and variance τ^2. The woman can use this information as a prior

$$\theta \sim N(\mu, \tau^2),$$

to get a Bayesian solution for her problem. We obtain the marginal distribution $m(y)$ by multiplying the model $f(y|\theta)$ and the prior $\pi(\theta)$ and integrating over the parameter of interest θ. This results again in a normal distribution

$$y \sim N(\mu, \sigma^2 + \tau^2).$$

The posterior distribution $\pi(\theta|y)$ is obtained by dividing the joint distribution $h(y, \theta) = f(y|\theta)\pi(\theta)$ by the marginal $m(y)$ resulting in

$$\theta|y \sim N(\frac{\sigma^2}{\sigma^2 + \tau^2}\mu + \frac{\tau^2}{\sigma^2 + \tau^2}y, \frac{\sigma^2\tau^2}{\sigma^2 + \tau^2}), \tag{4.3}$$

another normal distribution.
Let us now assume that the woman repeated the measurement N times with values $y = (y_1, ..., y_N)$ that are independently and identical distributed (iid) conditioned on the true blood pressure

$$y_i|\theta \overset{iid}{\sim} N(\theta, \sigma^2) \qquad i = 1, ..., N.$$

The common distribution is given by $f(y|\theta) = \prod_{i=1}^{N} f(y_i|\theta)$. The MLE for her true blood pressure is then given by the mean of the data $\tilde{\theta} = \bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i$. The marginal distribution for the single scores is $y_i \overset{iid}{\sim} N(\mu, \sigma^2+\tau^2)$, with the marginal over all observed data given by

$$m(y) = \frac{1}{\sqrt{2\pi(\sigma^2 + \tau^2)}} e^{-\frac{\sum_{i=1}^{N}(y_i-\mu)^2}{2(\sigma^2+\tau^2)}}.$$

The corresponding posterior distribution $\pi(\theta|y)$ is given by

$$\theta|y \sim N\left(\frac{(\sigma^2/N)}{(\sigma^2/N) + \tau^2}\mu + \frac{\tau^2}{(\sigma^2/N) + \tau^2}\bar{y}, \frac{(\sigma^2/N)\,\tau^2}{(\sigma^2/N) + \tau^2}\right).$$

Example 2a): Binomial-beta model
(Gelman *et al.*, 1995)

Another Bayesian model that is often used in praxis is the combination of a binomial distribution model with a beta distributed prior. Assume that we have a study to

evaluate the risk of tumors in laboratory rats. The sample consists of N rats from the same strand that were treated under identical conditions. Of these N rats y developed a tumor. We can model the experiment as a realization of a binomial distribution $y|p \sim \text{Bin}(N, p)$. The MLE for the tumor risk is given by $\tilde{p} = \frac{y}{N}$.
From historical data we know that the tumor risk among laboratory rats from this strand under varying experimental conditions is approximately beta distributed $\text{Beta}(\alpha, \beta)$ with known mean μ and variance σ^2. The parameters α and β of the distribution are related to mean and variance by $\mu = \alpha/(\alpha+\beta)$ and $\sigma^2 = \alpha\beta/((\alpha+\beta)^2(\alpha+\beta+1))$. We can use this information as prior $p \sim \text{Beta}(\alpha, \beta)$ for our unknown probability p. For convenience, we reparameterize the distribution by the expected mean μ and $M = \alpha+\beta$ as a measure of prior precision, so that the prior density is given by

$$\pi(p) = \frac{\Gamma(M)}{\Gamma(M\mu)\Gamma(M(1-\mu))} p^{M\mu-1}(1-p)^{M(1-\mu)-1} .$$

The marginal distribution for y is a beta-binomial with density

$$m(y) = \frac{\Gamma(M)}{\Gamma(M\mu)\Gamma(M(1-\mu))} \begin{pmatrix} N \\ y \end{pmatrix} \frac{\Gamma(y+M\mu)\Gamma(N-M(1-\mu))}{\Gamma(N+M)} .$$

The posterior distribution of the tumor risk is again beta distributed given by

$$\pi(p|y) = \frac{\Gamma(M)}{\Gamma(M\mu)\Gamma(M(1-\mu))} p^{y+M\mu-1}(1-p)^{N-y+M(1-\mu)-1} .$$

The main advantage of the Bayesian approach in comparison to the classical frequentist procedure is that it provides a simple conceptual framework with high flexibility and generality that allows to deal with really complex problems (Gelman *et al.*, 1995). In addition, MLE estimates from classical statistics often have the drawback, that the estimators can be quite unstable (Robert, 1994) and may lack smoothness (Robert, 1994), whereas Bayes estimators are more stable. Furthermore, a key aspect of Bayesian methods is that it is possible to easily perform sequential analyses using Bayesian formula sequentially (Gelman *et al.*, 1995). When a posterior distribution is calculated and new data become available, the previous posterior can be used as a prior for the new data (Lee, 1997). In the blood pressure example given different measurements, we have that $f(y_1, y_2|\theta) = f(y_1|\theta)f(y_2|\theta)$ when y_1 and y_2 conditional on θ are independent from each other. We can rewrite $\pi(\theta|y_1, y_2) \propto \pi(\theta)f(y_1, y_2|\theta)$ by $\pi(\theta|y_1, y_2) \propto \pi(\theta)f(y_1|\theta)f(y_2|\theta) \propto \pi(\theta|y_1)f(y_2|\theta)$, treating the posterior given y_1 as prior for y_2.

4.1.3 The prior

The prior distribution can be determined on a subjective or theoretical basis, e.g. from previous analyses, or other external information and always includes partially subjective considerations. Since the prior can clearly influence the posterior probability, the choice of the prior is the most critical and most criticized point of Bayesian analysis. Hence, the reasonable justification of the chosen prior by the statistician is really important,

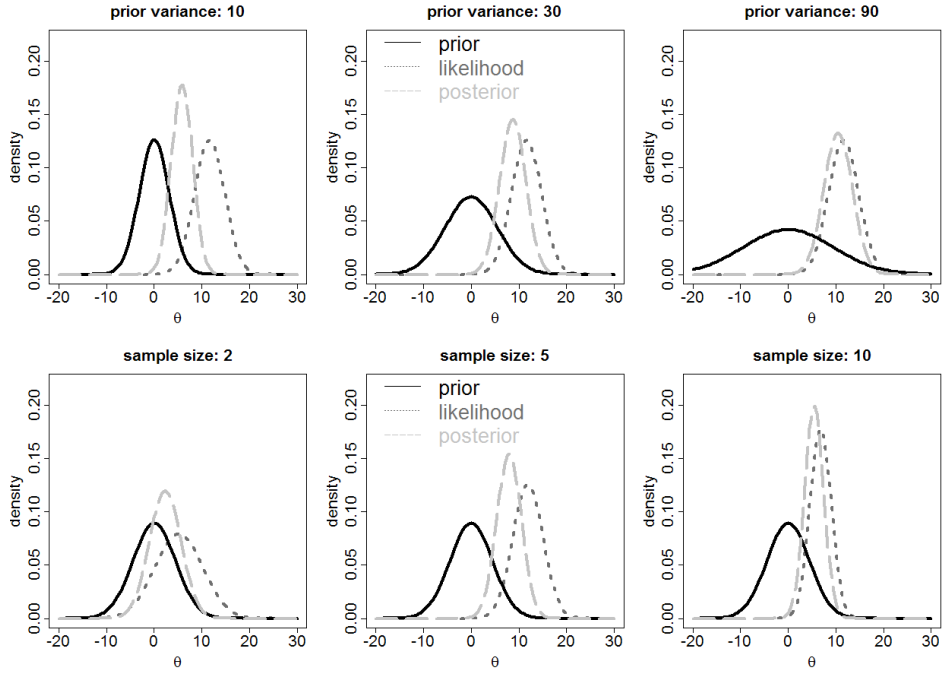


Figure 4.1: *The posterior distribution of a parameter of interest is a compromise between its prior distribution and the likelihood. With decreasing informativeness of the prior (top) – represented by its variance – and with increasing sample size of the data (bottom), the posterior gets closer to the likelihood. For the top row the data were sampled from a normal distribution N(10,50) with fixed sample size N=5 and prior probability N(0, σ^2) with varying variance σ^2=10,30,90. For the bottom row, data were sampled from N(10,50) with varying sample size N=2,5,10 and prior probability N(0, σ^2) with variance σ^2=20. Parameter of interest is the expected value of the data θ.*

based on sound or repeatable arguments, e.g. information based on physical, economical or biological mechanisms or experiments of the same type (Robert, 1994).
The influence of the prior informativeness to the posterior distribution can be clearly seen in our example 1 assuming a normal model and normal prior. Our posterior expectation

$$\frac{(\sigma^2/N)}{(\sigma^2/N)+\tau^2}\mu + \frac{\tau^2}{(\sigma^2/N)+\tau^2}\bar{y}$$

converges to the prior expectation μ with decreasing prior variance τ^2, since $(\sigma^2/N)/((\sigma^2/N)+\tau^2)$ converges to 1 and $\tau^2/((\sigma^2/N)+\tau^2)$ to 0. With increasing prior variance on the contrary, $(\sigma^2/N)/((\sigma^2/N)+\tau^2)$ approximates to 0 and $\tau^2/((\sigma^2/N)+\tau^2)$ to 1, so that the posterior expectation converges to $\bar{y}$ and the sample information becomes predominant. The influence of the prior informativeness to the posterior is illustrated in figure 4.1 (top). With an increasing variance of the prior from left to right, the informativeness decreases and the posterior approaches the likelihood.
In addition, the influence of the prior is affected by the sample size of the observed data - with decreasing impact of the prior the larger the sample size (Robert, 1994). From the posterior expectation of the normal-normal models this is obvious as well, with increasing N causing the same behavior as the increasing τ^2. We can see a graphical

presentation in figure 4.1 (bottom). The sample size increases from left to right and hence the posterior approaches the sampling distribution. We can further see that even for a relatively low sample size of only 10 observations the posterior is already very close to the likelihood.
If there is no information to integrate into the analysis, so called noninformative or flat priors can be used. These give imprecise prior information by having a large variance, e.g. given by a uniform distribution (Carlin and Louis, 2000) or by the so called Jeffreys uninformative priors Robert (1994) derived from the model distribution. In this case the data speak for themselves, with a posterior nearly the same as the likelihood and a solution close to the ML solution (Gelman *et al.*, 1995).
A special, important kind of prior distributions are the **conjugate priors**. A prior is called conjugate to a special model distribution, when the resulting posterior follows the same parametric form as the prior, meaning they belong to the same distribution family (Gelman *et al.*, 1995). The advantage of choosing a conjugate prior is that it is computationally more convenient, because the posterior distribution is given in closed form and can be analytically obtained without numerical integration (Carlin and Louis, 2000; Gelman *et al.*, 1995). For the Bernoulli, binomial and negative binomial distribution, with a probability as the unknown parameter, the beta distribution provides a conjugate prior (rat example) with suitable properties: it ranges from 0 to 1, can be symmetric or skewed, with a large or narrow peak or even U-shaped. When interested in the mean of a normal distribution, a normal distributed prior leads to a normal posterior, and we say that the normal distribution is **self-conjugated** (blood pressure example). For an unknown variance, a scaled inverse χ^2 prior or inverse gamma serve as conjugate priors. Furthermore, the gamma distribution is a conjugate prior for the Poisson distribution, the exponential distribution and itself (Lee, 1997; Robert, 1994).
When a non conjugate prior is given, the computation of the integrals in the posterior are in general not tractable analytically, even for statistical models of moderate complexity. Thus, they have to be evaluated numerically by good approximations. This can e.g. be done by sampling-based methods such as Markov chain Monte Carlo (MCMC). Although conjugate priors are mathematical convenient, more realistic priors should be preferred when available even though they are more inconvenient (Gelman *et al.*, 1995). Nevertheless, conjugate priors are in general a good starting point and e.g. mixtures of conjugate families can be useful when the simple conjugate distribution is not reasonable (Gelman *et al.*, 1995). In a noninformative setting, a compromise between a noninformative distribution which sometimes may be difficult to use or justify and a conjugate prior distribution with with analytical tractability should be found.

4.1.4 Bayesian inference

As mentioned before, Bayesian inference relies on the posterior distribution of the model parameters and appropriate inference statements can be derived from this distribution by calculating posterior quantities such as point estimates, interval estimates or probabilities (Gelman *et al.*, 1995; Robert, 1994).
While in classical frequentist statistics, point estimates for the unknown model parameters are derived by maximum likelihood estimation

$$\tilde{\theta}^{ML} = \arg\max_{\theta} L(\theta),$$

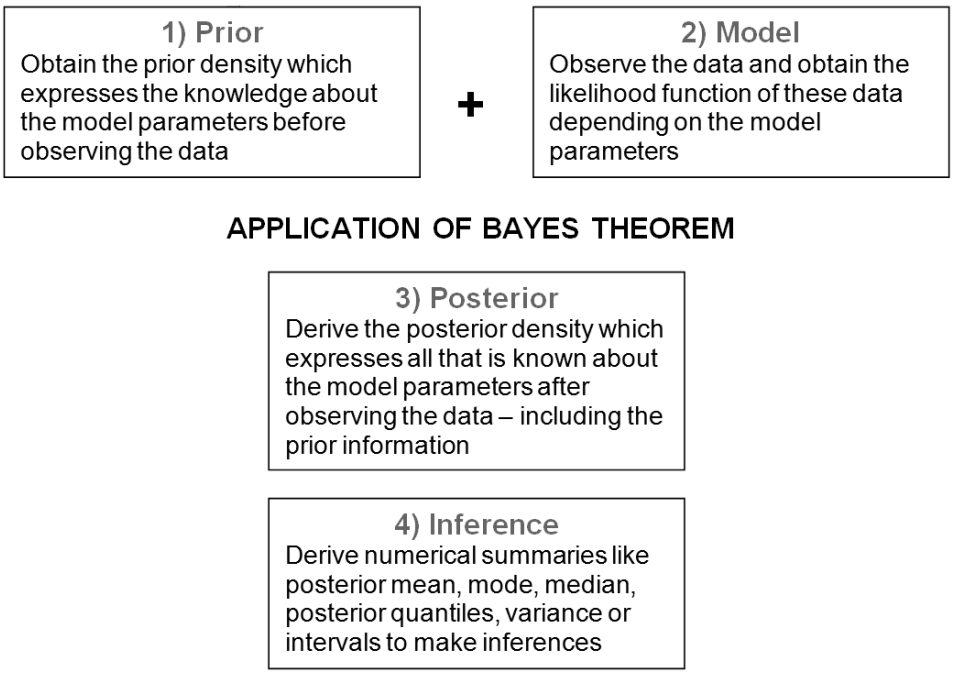


Figure 4.2: *Principle of the Bayesian approach*

this may be replaced by maximum a posterior (MAP) estimation in Bayesian statistics. MAP estimators are the values that are most likely given the data, hence the mode of the posterior distribution

$$\hat{\theta}^{MAP} = \arg\max_{\theta} \pi(\theta|y).$$

Alternatively, the posterior mean or median of the posterior distribution may be used. In order to obtain a measure of the accuracy of such a point estimate $\hat{\theta}(y)$ we might use the posterior variance with respect to the point estimate $\mathrm{E}_{\theta|y}(\theta - \hat{\theta}(y))^2$. Practically, the posterior mean $\hat{\theta}(y) = \mathrm{E}_{\theta|y}(\theta|y) = \int \theta\pi(\theta|y)d\theta$ as a point estimate for $\hat{\theta}(y)$ is preferred since it minimizes the posterior variance with respect to $\hat{\theta}(y)$, given by $\mathrm{Var}_{\theta|y}(\theta|y) = \mathrm{E}_{\theta|y}(\theta - \hat{\theta}(y))^2 = \mathrm{E}_{\theta|y}(\theta - \mathrm{E}_{\theta|y}(\theta))^2$. In the multivariate case, the same holds with respect to the posterior covariance matrix $\mathrm{Cov}_{\theta|y}(\theta|y) = \mathrm{E}_{\theta|y}((\theta - \mathrm{E}_{\theta|y}(\theta))(\theta - \mathrm{E}_{\theta|y}(\theta))')$.
In the normal-normal example, our point estimate of the woman's true blood pressure θ is given by the posterior expectation that is mode and median at once

$$\hat{\theta} = \mathrm{E}(\theta|y) = \frac{\sigma^2/N}{\sigma^2/N + \tau^2}\mu + \frac{\tau^2}{\sigma^2/N + \tau^2}\bar{y}. \tag{4.4}$$

We see that the estimate is a weighted average of the observed data itself ($\bar{y}$) and the prior mean (μ) $B\mu + (1-B)\bar{y}$, with weights proportional to the precisions $B = \frac{\sigma^2/N}{\sigma^2/N+\tau^2}$. Alternatively, the posterior mean can be expressed as the prior mean adjusted toward the observed data $\mu + (\bar{y} - \mu)(1 - B)$ or as the data 'shrunk' toward the prior mean $\bar{y} - (\bar{y} - \mu)B$. Therefore, B is often denoted as shrinkage factor, indicating how much the prior mean contributes to the estimate. The final estimate $\hat{\theta}$ is also denoted as shrinkage estimator (Heron *et al.*, 2011).

In the rat example, the posterior expectation that can be used as a point estimate of the tumor risk is given by

$$\hat{\theta} = \mathrm{E}(\theta|N, y) = \frac{y + M\mu}{N + M}.$$

This can be rewritten in the weighted average form

$$\frac{M}{N+M}\mu + \frac{N}{N+M}\frac{y}{N} = B\mu + (1-B)\tilde{\theta}$$

with $B = M/(M+N)$ and $\tilde{\theta}$ the classical maximum likelihood estimate.
The posterior mean is related to the prior mean by $\mathrm{E}(\theta) = \mathrm{E}(\mathrm{E}(\theta|y))$, the relation of the posterior variance and prior variance is given by $\mathrm{Var}(\theta) = \mathrm{E}(\mathrm{Var}(\theta|y)) + \mathrm{Var}(\mathrm{E}(\theta|y))$. Hence, the prior mean is the average over all possible posterior means over the distribution of possible data. The posterior variance is on average smaller than the prior variance by an amount that depends on the variation on posterior means over the distribution of possible data (Gelman *et al.*, 1995).As a summary of the Baysian method, the principle steps are illustrated in figure 4.2.

4.2 Empirical hierarchical Bayes methods

The main criticism of the Bayesian approach is that prior information is seldom rich enough to exactly define a prior distribution of a single known form as done in the previous section (Robert, 1994). Instead, the extent of the prior knowledge is often subject to uncertainty (Lee, 1997) and it may be necessary to involve this in the Bayes model. A possibility to consider uncertainty within the Bayesian paradigm is a particular modeling by decomposing the prior information into several distributional levels, denoted as hierarchical Bayes modeling.
In Robert (1994), a **hierarchical Bayes (HB) model** is defined as a Bayesian statistical model $(f(y|\theta), \pi(\theta))$, where the prior distribution is decomposed in conditional distributions $\pi_1(\theta|\eta_1), \pi_2(\eta_1|\eta_2), ..., \pi_l(\eta_{l-1}|\eta_l)$ and a marginal distribution $\pi_{l+1}(\eta_l)$ such that

$$\pi(\theta) = \int \pi_1(\theta|\eta_1)\pi_2(\eta_1|\eta_2)...\pi_l(\eta_{l-1}|\eta_l)\pi_{l+1}(\eta_l)d\eta_1 d\eta_2 ... d\eta_l. \tag{4.5}$$

The parameters η_j are called hyperparameters of level j $(j = 1, \ldots, l)$ to distinguish them from the model parameters θ.
From this definition, it automatically follows that the hierarchical Bayes is a special case of a Bayesian model. Hence, hierarchical Bayes models are covered in the Bayesian paradigm and underlie the conditions and properties of the Bayesian approach, with some additional advantages related to the prior decomposition (Robert, 1994).
The model specification over different levels with each new level in the hierarchy formed by a new distribution is in particular useful since statistical applications often involve multiple parameters that can be regarded as related or connected in some way by the structure of the problem. The hierarchical modeling allows the distinction between structural and subjective items of information. A special kind of structural relatedness

that we will focus on for our purpose is that the model parameters are independently and identically distributed (iid) (Lee, 1997). This relationship can be modeled in a natural way by using a prior distribution in which the single model parameters are viewed as a sample from a common population distribution, depending on unknown hyperparameters (Gelman *et al.*, 1995). Given N observations $y = (y_1, \ldots, y_N)$ depending on r unknown iid parameters $\theta = (\theta_1, \ldots, \theta_r)$, we can set up the hierarchical Bayes model with density $f(y|\theta)$ on the data level and the hierarchical prior with first stage prior $\theta_i \overset{iid}{\sim} \pi(\theta_i|\eta)$ $(i = 1, \ldots, N)$ and second stage prior $\eta \sim \pi(\eta)$. While the first stage prior represents the parameter relationship depending on hyperparameters η, the second stage prior express our beliefs about possible hyperparameter values (Gelman *et al.*, 1995; Lee, 1997). The second stage prior is also denoted as hyperprior.
In particular, in the following we will consider the case where the number of parameters r is the same as the number of observations N, with

$$y_i|\theta_i \overset{id}{\sim} f(y_i|\theta_i), \qquad i = 1, ..., N$$

since we have exactly that case in our application of the empirical hierarchical Bayes approach. For our normal-normal and binomial-beta example, we can image the following hierarchical Bayes models of that form.

Example 1b): Normal-normal model
(Berger, 1985; Robert, 1994)

With respect to our normal-normal example, we assume that N independent blood pressure measurements $y = (y_1, \ldots, y_N)$ of a woman for consecutive weeks are available. These observed values are assumed to be observations from independent distributions (id)

$$1^{st}\ level \qquad y_i|\theta_i \overset{id}{\sim} N(\theta_i, \sigma^2), \qquad i = 1, \ldots, N$$

with known variance σ^2 (measurement errors). The true blood pressure may vary from week to week but it is very reasonable that these weekly blood pressure values are from a common prior distribution $\pi_1(\theta_i|\mu, \tau)$, given by

$$2^{nd}\ level \qquad \theta_i|\mu, \tau \overset{iid}{\sim} N(\mu, \tau^2), \qquad i = 1, \ldots, N.$$

We can put a second stage prior $\pi_2(\eta) = \pi_{2,1}(\mu)\pi_{2,2}(\tau^2)$ on the hyperparameters $\eta = (\mu, \tau^2)$ assuming independence of μ and τ^2. For example, $\pi_{2,1}(\mu)$ may be specified by the overall distribution of diastolic blood pressure for a good studied population of women with $\mu \sim N(72, 120)$. When only vague knowledge about τ^2 is available, an appropriate noninformative prior may be chosen, e.g. $\pi_{2,2}(\tau^2) = 1$.

Alternatively, $y = (y_1, ..., y_N)$ may not be N blood pressure measurements for the same individual, but measurements for N different women . In that case, the same model may be used, but this time θ_i represents the true blood pressure of woman i, i=1,...,N which all come from the same common normal prior $\pi_1(\theta_i|\mu, \tau)$ representing the corresponding population.

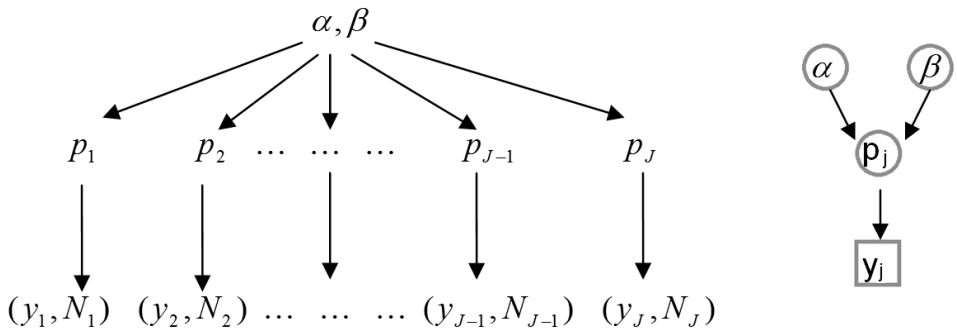


Figure 4.3: *Structure of the hierarchical Bayes model (left) and directed acyclic graph for the rat example (right)*

Example 2b): Binomial-beta model
(Gelman *et al.*, 1995; Lee, 1997)

Assume that we study the risk of tumors among laboratory rats of a special type. From J different groups of rats, we have data $(y_j, N_j), j = 1, \ldots, J$, with y_j the number of rats that developed a tumor from a total of n_j rats in group j, that follow independent binomial distributions. Due to differences between the rats and experimental conditions, the probability of a tumor is believed to vary between the different groups, but it is well reasonable to suggest these probabilities as random samples from a common beta distribution. Hence, our model is given by data level

$$1^{st}\ level \qquad y_j|p_j \overset{id}{\sim} \text{Bin}(N_j, p_j),$$

and prior distribution

$$2^{nd}\ level \qquad p_j|\alpha, \beta \overset{iid}{\sim} \text{Beta}(\alpha, \beta).$$

with $j = 1, \ldots, J$. However, the hyperparameters $\eta = (\alpha, \beta)$ are unknown and we take some appropriate second stage prior, e.g. a noninformative hyperprior of the form

$$3^{nd}\ level \qquad \pi_2(\alpha, \beta) \propto (\alpha + \beta)^{-5/2}$$

as suggested in Gelman *et al.* (1995) to represent our ignorance about the hyperparameters. In figure 4.3 (left), the structure of this hierarchical model is schematically displayed. The usual graphical tool to represent a hierarchical Bayes model is a directed acyclic graph (DAG) shown in figure 4.3 (right). In such a graph, random variables are represented as stochastic circles and known quantities are represented by squares, with the data as root of the graph.
We see that by hierarchical Bayes modeling, prior information can be separated into two parts, the structural prior knowledge and the more subjective standard form of Bayesian prior belief about the parameters of this structure (Lee, 1997; Robert, 1994). In principle, as given in the definition of a hierarchical model, the hyperprior of η can depend on unknown hyper-hyper-parameters λ as well, having a second-stage prior $\pi_2(\eta|\lambda)$ and a third-stage prior $\pi_3(\lambda)$ representing our beliefs about possible values of λ (Carlin

and Louis, 2000; Lee, 1997). This process may be repeated more often, continually adding randomness as moving down the hierarchy with parameters depending in turn on additional parameters which require their own prior. The hierarchy must stop at some point, with all remaining hyperparameters assumed to be known. On each level of the hierarchy, covariates may be involved. While the hyperprior in hierarchical Bayes models is often specified by a suitable non-informative distribution, conjugate priors are commonly used (Berger, 1985; Lee, 1997; Robert, 1994).
For simplification and since this is only necessary for our application, we will restrict to the simplest case and reduce the hierarchical structure to a 3-level model with a 2 stage prior

$$\begin{aligned} 1^{st}\ level\ &-\ model && y|\theta \sim && f(y|\theta) \\ 2^{nd}\ level\ &-\ prior && \theta|\eta \sim && \pi_1(\theta|\eta) \\ 3^{rd}\ level\ &-\ hyperprior && \eta \sim && \pi_2(\eta). \end{aligned} \tag{4.6}$$

Note that by definition all hierarchical Bayes models (formula 4.5) may be reduced to a 3-level model by eliminating the intermediate steps and additional hyperparameters $\pi_2(\eta) = \int \pi_2(\eta_1|\eta_2)...\pi_l(\eta_{l-1}|\eta_l)\pi_{l+1}(\eta_l)d\eta_1 d\eta_2 ... d\eta_l$ when $\eta_1, ..., \eta_l$ are not of interest for the inference (Robert, 1994).
The main inferential interest in hierarchical Bayes models may be the same as in the Bayesian approach, the calculation of the posterior distribution of the model parameters θ and its features Berger (1985). The hyperparameters are only a tool for its estimation Lee (1997). We can compute the posterior distribution for θ by additionally marginalizing over the hyperparameters η resulting in

$$\begin{aligned} \pi(\theta|y) &= \frac{\int f(y|\theta)\pi(\theta|\eta)g(\eta)d\eta}{\int\int f(y|u)\pi(u|\eta)g(\eta)d\eta\, du} = \frac{\int h(y,\theta|\eta)g(\eta)d\eta}{\int\int h(y,u|\eta)g(\eta)d\eta\, du} \\ &= \frac{p(y,\theta)}{m(y)} = \frac{\int h(y,\theta|\eta)g(\eta)d\eta}{\int m(y|\eta)g(\eta)d\eta} = \int \pi(\theta|y,\eta)h(\eta|y)d\eta. \end{aligned} \tag{4.7}$$

and e.g. yield point estimates for θ from this distribution. In the alternative of our blood pressure example 1b) given blood pressure measurements for different women, the estimation of the vector $\theta = (\theta_1, \ldots, \theta_N)$ is of interest, representing the women's true blood pressures. The posterior distribution of each single model parameter θ_i comes not only from the properties of those data which directly depend on it, but also from the hyperparameters η which summarize the properties of the population of the model parameters as a whole. However, in the first case of example 1 b) having different measurements for one woman, the true blood pressure of the woman is represented by the μ and we may rather focus on this hyperparameter. Hence, the interest in hierarchical Bayes analyses is not necessarily restricted to the posterior distribution of θ and corresponding quantities but may vary (Lee, 1997). Here, the posterior distribution of the hyperparameters has to be calculated instead, given by

$$\pi(\eta|y) = \frac{\int f(y|\theta)\pi(\theta|\eta)d\theta g(\eta)}{\int\int f(y|\theta)\pi(\theta|\eta)d\theta g(\eta)d\eta} = \frac{\int h(y,\theta|\eta)d\theta g(\eta)}{\int\int h(y,\theta|\eta)d\theta g(\eta)d\eta} = \frac{p(y,\eta)}{m(y)} \tag{4.8}$$

and we can obtain a point estimate of η. The rat example falls into this class of inference as well, with the tumor rate in the different single groups of the experiment not of

interest but the general tumor probability for this kind of rats. Another intention for example may lie in finding a predictive distribution rather than parameter estimation. For example, the prediction of a blood pressure measurement or the number of tumors in a new experiment, obtained by the overall distribution in the population under consideration $m(y) = \int f(y|\theta)\pi(\theta|\eta)\pi(\eta)d\eta d\theta$ may be of interest (Lee, 1997).
As in Bayesian analysis, integration steps may be explicitly carried out by a close-form expression in a simple case, while more complex models require numerical approximation methods such as MCMC (e.g. WinBUGS) (Berger, 1985).

Hierarchical Bayes models can capture dependencies within the data more realistically than non-hierarchical models (Gelman *et al.*, 1995). Very complex and flexible models can be generated and the hierarchical thinking may help to understand natural multilevel structures and enable in particular to analyze them using the information from all levels. Different sources of variability, clustered and correlated data can be modeled as well as overdispersion. Hierarchical Bayes models have enough parameters to fit the data well, nevertheless avoiding the problem of overfitting (Gelman *et al.*, 1995). Furthermore, hierarchical Bayes models are appropriate for a wide range of applications, e.g. in medicine, biology, animal breeding, economy, where the population of interest can be perceived as a subpopulation of a population (e.g. meta-analysis)(Robert, 1994).
From a practical point of view, hierarchical Bayes models play an important role in developing computational strategies (Gelman *et al.*, 1995). A computational advantage of the hierarchical Bayes method is that Bayesian calculation may be simplified by the hierarchical structure. The decomposition of the prior and the posterior may compensate the apparent complexity induced by successive levels and allow easier approximations of posterior quantities by simulations. Nevertheless, hierarchical Bayes models usually prevent an explicit derivation of the corresponding Bayes estimators as well, even when the successive levels are conjugate, and therefore they call for numerical approximation (Robert, 1994).
By hierarchical Bayes models, aspects of a population distribution of model parameters can be estimated, although these values are not directly observed (Gelman *et al.*, 1995). Hierarchical Bayes models permit the computation of individual-level parameter estimates that fit the individual outcome reasonably well but are relatively stable by borrowing information from other respondents. By this property of "borrowing strength" from the entire ensemble, even for groups with small sample size inference can be performed well. By the hierarchical modeling, the robustness of Bayes estimators is improved from a frequentist point of view since the arbitrariness of choice of hyperparameters is reduced, while still incorporating prior information (Robert, 1994). The more subjective aspect of the prior modeling is relegated to a higher level and thus the hierarchical Bayes model provides an intermediary position between a straightforward Bayesian analysis and frequentist imperatives (Robert, 1994). Nevertheless, estimators of the hierarchical Bayes approach are not more and not less admissible than normal Bayesian estimators (Robert, 1994).

In general, hierarchical modeling is not only known in the Bayesian context. Whenever we specify a model over several levels with each stage defining a stochastic model for

the previous stage, we have a hierarchical model, and the inference on such models can be frequentist as well. The most famous, classical occurrence of hierarchical models in the non-Bayesian context is the random effects model (Robert, 1994), appropriate to represent nested data, e.g. pupils nested in classrooms nested within schools. For observations y_{ij}, such a model may be given by

$$y_{ij} = \mu + \theta_i + \varepsilon_{ij}$$

with random error $\varepsilon_{ij} \sim N(0, \sigma^2)$ and random effects $\theta_i \sim N(0, \tau^2)$ modeled as drawn from a distribution. In the frequentist framework, the inference is about the fixed effect μ and the variability τ^2 of the random effects (Robert, 1994). The estimation of the random effects θ_i is not possible, since they are considered as unobserved variables and not as parameters. However, operating on the same model from a Bayesian perspective the focus of interest can shift to the estimation of the individual effects θ_i by considering the second stage as information entering the model in form of a prior distribution, specifying the uncertainty about the parameters (Robert, 1994). In a Bayesian fashion, the Bayes theorem is used to compute the posterior distribution of the θ_i and point estimates are obtained e.g. by the posterior mean. The example of the random effects model illustrates the conceptual difference of hierarchical modeling in the frequentist context to a Bayesian hierarchical modeling and that the boundary between classical and Bayesian models is sometimes fuzzy and mainly depends on the interpretation of the model (Heron *et al.*, 2011; Robert, 1994).

Further on, there exists a hybrid method of estimation for hierarchical models somewhere between a full Bayesian solution and the classical frequentist proceeding, called **empirical Bayes (EB)** (Robert, 1994). The empirical Bayes approach differs from the complete Bayesian analysis by its strategy to construct the prior distribution. As illustrated before, regarding the hierarchical model in formula 4.5, in full Bayesian treatment a hyperprior distribution is specified independently of the observed data, quantifying some uncertainty about the hyperparameters (Berger, 1985; Gelman *et al.*, 1995). The full posterior distribution for the parameters of interest is then estimated by additionally marginalizing over η (Carlin and Louis, 2000). Alternatively, the empirical Bayes approach is a procedure for statistical inference in which the unknown hyperparameters η of the hierarchical model are replaced by an estimate $\hat{\eta}$ based on the observed data. Instead of using a hyperprior distribution on these parameters, the obtained point estimates are substituted in the prior distribution. Hence, the empirical Bayes offers a good possibility to define the model without introducing a further specification of the hyperprior distribution.

In the empirical Bayes approach the estimation of the hyperparameter is done in a frequentist sense by maximizing the marginal distribution $m(y|\eta)$ of the observations viewed as a function of η (**marginal maximum likelihood estimate, MMLE**). Hence, the parameters are set to their most likely values (Berger, 1985; Carlin and Louis, 2000; Heron *et al.*, 2011).

Given data $y = (y_1, ..., y_N)$ with $y_i|\theta_i \overset{id}{\sim} f(y_i|\theta_i)$ and $\theta_i|\eta \overset{iid}{\sim} \pi(\theta_i|\eta)$, i =1,. . . ,N, the marginal likelihood of the data is obtained by marginalizing the likelihood function $\prod_{i=1}^{N} f(y_i|\theta_i)\pi(\theta_i|\eta)$ over the parameters of interest

$$m(y|\eta) = \prod_{i=1}^{N} m(y_i|\eta) = \prod_{i=1}^{N} \left(\int f(y_i|\theta_i)\pi(\theta_i|\eta)d\theta_i \right). \tag{4.9}$$

For simple distributions, an exact solution of the MMLE is possible by using standard iterative ML methods like Nelder-Mead, Quasi-Newton or conjugate gradient. Otherwise, numerical integration methods like Monte Carlo, Laplace approximation, Gibbs Sampling, Newton Raphson Iteration or Expectation-Maximum algorithm (EM) have to be used (Lee, 1997).
The empirical Bayes analysis continues as if the prior distribution is known by plugging these point estimates $\hat{\eta}$ in the prior distribution (formula 4.6) $\theta|\hat{\eta} \sim \pi(\theta|\hat{\eta})$. By using this model with the prior specification based on the data, we obtain a model of Bayesian form and we can proceed in the standard Bayesian fashion by calculating the posterior for θ_i (4.2) (Berger, 1985; Heron *et al.*, 2011)

$$\pi(\theta_i|y_i, \hat{\eta}) = \frac{f_i(y_i|\theta_i)\pi(\theta_i|\hat{\eta})}{m_i(y_i|\hat{\eta})}. \tag{4.10}$$

In comparison to the complete Bayesian analysis, the posterior for each of the parameters depends not only on y_i, those data directly related to the parameter θ_i, but also on all data depending on a whole population of parameters $\theta_j, j = 1, \ldots, N$ summarized by the hyperparameters $\hat{\eta}$. Note that the empirical Bayes approach of course can be implemented for hierarchical models with any number of levels.

In our blood pressure example 1b), the marginal likelihood is given by

$$m(y|\mu) = \frac{1}{\sqrt{2\pi(\sigma^2+\tau^2)}} e^{-\frac{\sum_{i=1}^{N}(y_i-\mu)^2}{2(\sigma^2+\tau^2)}}.$$

Maximizing this yields the MMLE $\hat{\mu} = \bar{y} = \frac{1}{N}\sum_{i=1}^{n} y_i$ for the unknown hyperparameter. Replacing the unknown μ by its MMLE, we obtain the posterior distribution $\pi(\theta_i|y_i, \hat{\mu})$

$$\theta_i|y_i \overset{iid}{\sim} N\left(\frac{\sigma^2}{\sigma^2+\tau^2}\hat{\mu} + \frac{\tau^2}{\sigma^2+\tau^2}y_i, \frac{\sigma^2\tau^2}{\sigma^2+\tau^2}\right).$$

An estimate for θ_i is given by the expected value of the posterior normal distribution

$$\hat{\theta}_i = \frac{\sigma^2}{\sigma^2+\tau^2}\hat{\mu} + \frac{\tau^2}{\sigma^2+\tau^2}y_i = B\bar{y} + (1-B)y_i \text{ with } B = \frac{\sigma^2}{\sigma^2+\tau^2}, \tag{4.11}$$

which is a weighted average of the observed data itself and the corresponding estimated prior mean $\hat{\mu}$. The inference about each single component depends not only on the corresponding data itself, but on all given data. Our single observations y_i are shrunk towards the mean of observations $\hat{\mu}$, with the impact of the mean depending on ratio of the variances (Heron *et al.*, 2011). When τ^2 is additionally unknown, we have to estimate this variance from the data as well. The total variance over all measured values $y = (y_1, ..., y_N)$ is estimated by

$$s^2 = \frac{1}{N}\sum_{i=1}^{N}(y_i - \bar{y}).$$

This is the sum of the variability between the different blood pressure measurements (τ^2) and the known measurement error (σ^2). Hence, we can estimate τ^2 by

$$\hat{\tau}^2 = \max(0, s^2 - \sigma^2) \tag{4.12}$$

and plug this estimate into formula (equation 4.11)(Carlin and Louis, 2000).
In the tumor risk example, the marginal densities are

$$m(y_i|\mu, M) = \frac{\Gamma(M)}{\Gamma(M\mu)\Gamma(M(1-\mu))} \begin{pmatrix} N_i \\ y_i \end{pmatrix} \frac{\Gamma(y_i + M\mu)\Gamma(N_i - M(1-\mu))}{\Gamma(N_i + M)},$$

for the single observations. The marginal likelihood is obtained by building the product of these densities over all groups $i = 1, \ldots, J$. The hyperparameters μ and M can be estimated by the MMLE, using a numerical method, or simpler by using the method of moments. The hyperparameter μ is estimated by the overall tumor risk neglecting group differences

$$\hat{\mu} = \frac{\sum_{i=1}^{J} y_i}{\sum_{i=1}^{J} N_i}.$$

The moment estimate for the variance is given by

$$s^2 = \frac{1}{J} \sum_{i=1}^{J} \frac{\hat{\mu}(1-\hat{\mu})}{N_i} \left[1 + \frac{N_i - 1}{\hat{M} + 1}\right],$$

and solving this for $\hat{M}$ results in

$$\hat{M} = \frac{\hat{\mu}(1-\hat{\mu}) - s^2}{s^2 - \frac{\hat{\mu}(1-\hat{\mu})}{J} \sum_{i=1}^{J} 1/N_i} \quad \text{with } s^2 = \frac{J}{J-1} \frac{\sum_{i=1}^{J} N_i(\hat{p}_i - \hat{\mu})^2}{\sum_{i=1}^{J} N_i}.$$

Our posterior distribution is given by

$$p(p_i|y_i, \hat{\mu}, \hat{M}) = \frac{\Gamma(M)}{\Gamma(M\mu)\Gamma(M(1-\mu))} p_i^{y_i + M\mu - 1} (1-p_i)^{N_i - y_u + M(1-\mu) - 1},$$

what leads to point estimates

$$\hat{p}_i = \mathrm{E}(p_i|y_i, \hat{\mu}, \hat{M}) = \frac{y_i + \hat{M}\hat{\mu}}{N_i + \hat{M}} = \frac{\hat{M}}{N_i + \hat{M}}\hat{\mu} + \frac{N_i}{N_i + \hat{M}} \frac{y_i}{N_i} = \hat{B}_i\hat{\mu} + (1 - \hat{B}_i)\tilde{p}_i,$$

which is a weighted average of the event estimate $\tilde{p}_i = y_i/N_i$ and $\hat{\mu}$ (Carlin and Louis, 2000).

Since the empirical Bayes (EB) approach replaces the integration over the hyperparameters η by a maximization to obtain the most likely values $\hat{\eta}$, it fails to account for the uncertainty of η. The resulting posterior conditioned on $\eta = \hat{\eta}$ is just an approximation of the posterior density $\pi(\theta|y)$. This conditional posterior $\pi(\theta|y, \hat{\eta})$ is also denoted as pseudo posterior distribution (Robert, 1994).
From a modeling perspective, the empirical Bayes approach does not partake in the

Bayesian paradigm but is considered to be a problem in classical statistics since it does not strictly distinguish between data and prior information and uses the data to specify the prior (Berger, 1985; Heron *et al.*, 2011). However, by frequentists and practitioners, the empirical Bayes approach is usually classified as Bayesian (Robert, 1994). Empirical Bayes analyses naturally are very related to Bayesian methods, since after the estimation of the prior, the analysis proceeds in a typical Bayesian fashion to compute a posterior distribution and relevant posterior quantities (Berger, 1985). The empirical Bayes approach can be viewed as an approximation to the complete Bayesian treatment of a hierarchical model and is asymptotically equivalent (Gelman *et al.*, 1995; Robert, 1994). In particular in noninformative settings, EB techniques appeal (Robert, 1994).
Since the empirical Bayes approach includes both, classical estimation and a Bayesian flavor, it can simultaneously draw strength from frequentist and Bayesian methods (Robert, 1994). From the calculation perspective, the empirical Bayes approach requires the solution of a likelihood equation, while hierarchical Bayes requires additional numerical integration (Berger, 1985). This makes empirical Bayes computationally more simple (Robert, 1994), resulting in a gain of estimation efficiency and simplified treatment of complex problems. Therefore, in particular in problems where a genuine Bayes modeling is too complicated or costly it may be an acceptable approximation.
However, in general the hierarchical Bayes approach has shown to be the superior methodology. It is often preferable over the empirical procedure, since it has the clear advantage that it measures standard errors (Robert, 1994). The main drawback of EB is that it relies on classical frequentist methods such as ML to estimate the hyperparameters (Lee, 1997; Robert, 1994) and fails to incorporate any uncertainty in the prior information (Berger, 1985). Nevertheless, HB and EB often lead to comparable results especially in context of point estimation. Furthermore, Morris (1983) argues that EB analyses actually have a type of frequentist justification making such analyses more attractive to non-Bayesians (Berger, 1985). Due to the popularity of EB due to good frequentist properties of some resulting estimators (Robert, 1994), the EB method is in praxis often employed in case where we have structural relationship between parameters as described, allowing us the use of the data to estimate some features of the prior distribution (Berger, 1985).
Since all Bayesian inference is driven by the conditional posterior and the posterior distribution for each single parameter depends on all data, a point estimate $\hat{\theta}_i$ given by the corresponding conditional posterior mean for example depend on all the observed data too. It is known that the conditional posterior mean $\mathrm{E}(\theta_i|y_i, \hat{\eta})$ as an estimator for θ_i is approximately equal to the correct posterior mean $\mathrm{E}(\theta_i|y_i)$. Nevertheless, it is also well known that the corresponding conditional variance $\mathrm{Var}(\theta_i|y_i, \hat{\eta})$ underestimates the correct posterior variance $\mathrm{Var}(\theta_i|y_i)$ (Kass and Steffey, 1989). Hence, an adjustment to account for the uncertainty of η may be required to produce valid estimates for the variance.

In our given outline about the empirical Bayes, we restricted to the situation where we only had to chose a value for η to completely specify the estimated posterior distribution Carlin and Louis (2000). This approach is in particular denoted as **parametric empirical Bayes (PEB)**, since it assumes a parametric form for

the prior, given by a special type of distributions. Nevertheless, the underlying cumulative distribution function $\pi(\theta|\eta)$ of θ may have an unknown form, known as **non-parametric empirical Bayes (NPEB)**. In general, non-parametric methods have the advantage that no assumptions about the type of prior distribution have to be made. This protects against violations of assumptions, but they are less powerful than parametric approaches because they use less information in their calculations. Parametric methods on the other hand are more efficient and result in higher power, but they are only valid when the type of distribution is clear and no assumptions are violated. Since the model of Lewinger *et al.* (2007) used in this thesis is a parametrical method, with a reasonable justification of the choice of prior distribution, we will not go into more detail of the NPEB.

4.3 Bayesian methods and hierarchical models in genome-wide association studies

As already discussed in chapter 3, the main challenge of genome-wide association studies is to identify the few causal variants of a disease in a large quantity of genetic data. Usually, a genome-wide association study is used as an initial step for selecting a subset of "most promising" markers that are examined more closely in later stages of a multi-stage design or in a replication study Thomas (2006).
Traditionally, the measure most commonly used to flag SNPs as "noteworthy" in this screening SNP is the p-value by selecting the markers with the most significant associations to the disease at some cut off (Wakefield, 2008). However, due to the large number of performed tests and a low prior probability of a non-null association in a GWAS, the chance that such a significant rated result is a true positive result is only low. That is the case even in large well-designed and well-conducted studies (Wacholder *et al.*, 2004; Wakefield, 2008). The portion of false positive results in association studies is estimated to be at least 95% (Colhoun *et al.*, 2003). The chance that none of the variants further investigated has a true effect to the disease is high, implicating extreme caution when using p-value ranking (Wakefield, 2008). In addition, the truly associated SNPs are not necessarily ranked among these top SNPs due to their small risks and interrelatedness and therefore the chance to find these is only small – even with large samples.
To overcome these problems, several alternative quantities from a Bayesian perspective were suggested to decide if a finding deserves the attention of further investigation by considering the prior probability of the null hypothesis (Wakefield, 2008). In 2004, Wacholder *et al.* published the false positive report probability (FPRP). The FPRP (Hunter and Kraft, 2007; Samani *et al.*, 2007; Thomas, 2010b; Wacholder *et al.*, 2004) is the posterior probability that a statistical finding between a marker and the disease is not a true association $\mathrm{P}(H_0 \text{ is true}|H_0 \text{ is rejected})$. In addition to the observed p-value, $P(H_0 \text{ is rejected}|H_0 \text{ is true})$, the FPRP incorporates two more factors: a prior assumption about the fraction of tested variants that are truly associated with the disease, $P(H_1 \text{ is true})$, and $\mathrm{P}(H_0 \text{ is rejected } |H_1 \text{ is true})$, the statistical power of the test. The latter depends on the sample size, the frequency of the genetic marker and its observed odds ratio and in general is low (Wacholder *et al.*, 2004). By considering the prior probability, the FPRP may protect from "over interpreting" statistically significant findings that are not likely to be true (Wacholder *et al.*, 2004).

Another Bayesian alternative to assess whether an association is noteworthy is the Bayes factor. The Bayes factor is defined as the probability to observe the data given the null hypothesis divided by the probability of the data under the alternative, $P(\text{data}|H_0)/P(\text{data}|H_1)$, measuring the impact of the data to support H_0 in preference to H_1 (Wakefield, 2007). The Wellcome Trust Case Control Consortium (WTCCC) for example used the Bayes factor in its GWAS of seven different diseases (Wellcome Trust Case Control Consortium, 2007). To overcome the difficulties of specifying a prior distribution over all unknown parameters and evaluating multi-dimensional integrals necessary to calculate the Bayes factor, Wakefield proposed in 2007 an approach based on an approximated Bayes factor, denoted as Bayesian False Discovery Probability (BFDP). By using the BFDP, the number of discoveries that are followed-up but cannot be replicated in further investigation may be reduced (Wakefield, 2007).
Although these two methods, FPRP and BFDP, may improve the selection of SNPs for follow-up, they both ignore any external information that might favor a less significant association with supporting evidence over a more significant one without prior knowledge (Thomas, 2010b). Each of the considered SNPs is a priori assumed to be equally likely causal (Chen and Witte, 2007). Thereby, findings that are not likely to be true may lead to false positive results, while true associations with biological support may be still missed due to their small effects. However, giving higher priority to subsets of SNPs with greater biological plausibility would improve the SNP selection by reducing false positive findings and increasing true positive results (Botstein and Risch, 2003). By this, true causal variants may be distinguished more clearly from the noise (Chen and Witte, 2007). For example, a greater credibility of association may be given to coding SNPs or markers already reported in another study and to SNPs that are located within genes or identified linkage regions (Thomas, 2010a). Hence, by the external information the assumption that all SNPs act similarly is overcome by allowing the SNP effects to vary depending on the additional genetic information (Heron *et al.*, 2011).
Addressing this problem, Whittemore (2007) and Roeder *et al.* (2006) suggested Bayesian variants of the FDR framework (section 3.2.4) to allow for the consideration of external knowledge to up- or down-weigh each of the SNPs (Roeder *et al.*, 2006, 2007; Whittemore, 2007). The Bayesian false discovery rate (BFDR) defined by Whittemore (2007) combines the frequentist FDR approach of Benjamini *et al.* (2001) for multiple testing correction with the FPRP strategy of Wacholder *et al.* (2004). Therefore, so called **"b-values"** are used, which are based on the FPRP instead of simple p-values as decision criterion to control the FDR. While Wacholder *et al.* (2004)'s FPRP assumes that the prior probability of a true association for each tested SNP is the same, in the BFDR this prior differs between the genetic variants, depending on the given external information. Roeder *et al.* (2006) use a weighted FDR framework where information from previous linkage studies is used to modify the rejection criterion. In this approach, each p-value is divided by a weight, obtained from the linkage information known for that location. By using this prior knowledge, the FDR is spread non-uniformly across all tested markers (Thomas, 2006). Both approaches provide potentially useful frameworks for integrating external data in genome-wide association studies. Furthermore, Roeder *et al.* (2006) and Whittemore (2007) could show that the integration of uninformative prior information results in relatively small loss of power in comparison to simple p-value usage (Thomas, 2010a). When additional well-chosen prior knowledge is used, it can

lead to substantially greater statistical power and less false positive results. Nevertheless, both methods require a pre-specification of priors or weights for every marker at genome-wide level (Thomas, 2010a). This causes new problems, especially when several sources of prior information are available but only little knowledge about their relative informativeness (Lewinger *et al.*, 2007).
To avoid tedious and judgmental pre-specifications, hierarchical modeling approaches offer a valid way to incorporate multiple types of prior information into a general framework to prioritize SNPs of an initial GWAS for further investigation (Chen and Witte, 2007; Hung *et al.*, 2004; Lewinger *et al.*, 2007; Pan, 2005; Thomas, 2010b).
In a hierarchical model, the external information can be integrated via covariates containing this knowledge (prior covariates) in a prior or hyperprior distribution. Instead of ranking the SNPs according to their original p-value, they are ranked by their posterior expectation given by the hierarchical model. This re-ranking should achieve more effective results by prioritizing SNPs according to the given prior knowledge that would not have been selected before by pure p-value ranking. By using an empirical Bayes approach, the method does not rely on the subjective input of the practitioner in setting the prior parameters, but instead uses the available data to obtain the parameter estimates. It exploits the attractive feature of a GWA scan, that the large quantity of data (many hundreds of thousands of markers) makes it possible to let the data itself suggest, which prior information is correlated with the association. Therefore, the method provides a more flexible approach than the Bayesian or weighted FDR where weights for the information sources have to be prespecified.
In general, hierarchical models have the ability to easily include relevant biological information in a coherent framework. They offer better and more stable estimates of the parameters, since all data are considered for each single estimate. Thereby estimates that were unstable or extreme before become more reasonable (Heron *et al.*, 2011). Empirical hierarchical Bayes effect estimates can potentially reduce false positive results.

In the following, we will present three hierarchical models proposed in the context of GWAS. For all methods, the biological relevant external information will be modeled by the covariate matrix Z. Given N_M SNPs M_i, $(i = 1, \ldots, N_M)$ and N_C different covariates C_r, $(r = 1, \ldots, N_C)$ to represent the prior knowledge, Z has the dimension $N_M \times N_C$ and the entry $z_{M_iC_r}$ contains the information about covariate C_r of SNP M_i (Hung *et al.*, 2004). The Z matrix is the key component of the hierarchical model approaches, defined by the investigator to reflect similarities between the SNP markers (Hung *et al.*, 2004). By these similarities, strength among the SNPs can be borrowed to enrich the overall GWAS signals (Chen and Witte, 2007). The possible external information that could be incorporated might be about the functionality of the SNPs (e.g. coding, nonsynonymous), the location (e.g. intron, exon, regulatory region), conservation, previously reported linkage or association regions, information on candidate genes and pathways, gene expression, in silico predictions of potential functions (Chen and Witte, 2007; Hung *et al.*, 2004) or location relatively to known or predicted genes. This information can e.g. be used for categorization, with $z_{M_iC_r} = 1$ indicating the membership of a SNP M_i to the particular category given by covariate C_r and 0 else, given e.g. a previous association signal, a value related to the corresponding statistic from the earlier analysis. Using different categories, SNPs assigned to the same category are assumed to arise from a common distribution (Hung *et al.*, 2004).

An intuitive hierarchical model

An intuitive approach of a hierarchical model in genetic association studies was published by Hung *et al.* (2004) and picked up by Chen and Witte in 2007 to show its potential value in the context of genome-wide association studies. Furthermore, in 2011 Heron *et al.* examined the impact of the inclusion of informative and non-informative information to this model in GWAS by simulation studies (Heron *et al.*, 2011).
While Hung *et al.* (2004) and Heron *et al.* (2011) used a logistic version for the application in case-control studies, Chen and Witte (2007) illustrated a linear version for quantitative traits. In the following we will outline the logistic version since this thesis focuses on case-control studies. However, the linear version is obtained by simply replacing the first stage logistic regression model by a linear one.
The first stage of the intuitive hierarchical model is the conventional approach to estimate the main effects of each SNP M_i individually by a logistic regression model (Heron *et al.*, 2011; Hung *et al.*, 2004)

$$1^{st}\ level \qquad \ln\left(\frac{p_{M_i,I_n}}{1-p_{M_i,I_n}}\right) = \alpha_{M_i} + X_{M_i,I_n}\beta_{M_i}, \qquad i = 1,\ldots,N_M, n = 1,\ldots,N_I.$$

X_{M_i,I_n} is the genotype of individual I_n for SNP M_i. Different genetic models can be assumed for the genotype, e.g. an additive model coded by 0,1,2, presenting the copies of the minor allele. p_{M_i,I_n} is the probability of individual I_n being a case given the genotype X_{M_i,I_n} (Heron *et al.*, 2011). The intercept term α_{M_i} represents the baseline risk of disease in form of a log odds for an individual with the homozygote genotype of the major allele (Heron *et al.*, 2011). The regression coefficient β_{M_i} represents the effect of the particular genetic marker on the disease risk on a log odds scale (Hung *et al.*, 2004). More precisely, in the case of an additive model it represents the increase in odds of being a case for each additional allele. Additional covariates, e.g. age or sex, may be considered in the model as well – but we will present only the simple model without any phenotypic covariates. For the traditional frequentist strategy, the coefficients of this model are estimated by maximum likelihood estimation. A Wald statistic is formed by dividing this estimate by the corresponding standard deviation (Chen and Witte, 2007; Wald, 1943). Existing information about the SNPs is ignored and each SNP is assumed to be equally likely associated with the phenotype (Chen and Witte, 2007).
To improve the estimation of the β_{M_i} and the SNP ranking through the inclusion of additional biological information in a Bayesian manner, a second stage model (prior) is added that incorporates external marker information (Hung *et al.*, 2004) to specify the relations among the different genetic variants so that the markers can support each other. This can be given by a second-level regression of the form

$$2^{nd}\ level \qquad \beta = Z\mu + \delta, \qquad \delta \sim N(0, \tau^2 W)$$

where $\beta = (\beta_{M_1},\ldots,\beta_{M_{N_M}})$ is (Hung *et al.*, 2004) the vector of the N_M coefficients from the first stage. $\delta = (\delta_{M_1},\ldots,\delta_{M_{N_M}})$ is a vector of residual effects of the different markers, which are normally distributed with mean 0 and variance-covariance matrix $\tau^2 W$. Correlations between the SNPs can be modeled in the off-diagonal entries of the W matrix, assuming no correlation W simplifies to the identity matrix I (Heron *et al.*, 2011). Residual effects may arise due to interaction effects among the covariates of the second step or unconsidered covariates (Hung *et al.*, 2004). The external information is

incorporated in the second stage, with the effects of these prior covariates on the first stage estimates given in the vector of prior coefficients $\mu = (\mu_{C_1}, ..., \mu_{C_{N_C}})$ (Hung *et al.*, 2004).
The final estimate of the SNP effects is then given by

$$\hat{\beta}_{EB} = BZ\hat{\mu} + (I - B)\tilde{\beta}, \tag{4.13}$$

which is the shrinkage estimator of the usual estimator $\tilde{\beta} = (\tilde{\beta}_{M_1} \dots \tilde{\beta}_{M_{N_M}})$ shrunk towards the second-level mean $Z\hat{\mu}$ with shrinkage factor $B = (\tilde{V} + \tau^2 W)^{-1}\tilde{V}$ (Chen and Witte, 2007; Heron *et al.*, 2011). $\tilde{V}$ is the conventional ML estimate of the variance-covariance matrix based on the first level regression model and $\hat{\mu}$ are the empirical Bayes estimates of the prior coefficients, obtained by maximization of the marginal maximum likelihood (MML) of this model.

In her comprehensive simulation study, Heron *et al.* (2011) demonstrates that the inclusion of biologically relevant information through this hierarchical empirical Bayes model offers a more robust method of detecting associated SNPs. The resulting estimates are more stable advancing from reduced variability. The method performs better than the conventional p-values ranking. When uninformative covariates are given, the hierarchical model still performs equally to the traditional approach with respect to power and false positive rate, even given noisy information the method performs well. Hence, the approach is not adversely affected by the inclusion of unreliable prior information, thereby ensuring robustness when considering incorporating additional biological information (Heron *et al.*, 2011).

Linear regression on pathways (LRP) of Lebrec *et al.* (2009)

Lebrec *et al.* (2009) chose another strategy by not directly considering the genotype and phenotype data in the first stage of his hierarchical model, but the effect estimates of the different SNPs. First of all, the positive allelic effect on the log odds scale and corresponding variance for each SNP is estimated by a logistic regression model. In the following, these estimates are denoted by $\hat{\beta}_{M_i}$ and $\hat{\sigma}^2_{M_i}$. These effect estimates $\hat{\beta}_{M_i}$ are normally distributed with expectation μ_{M_i} representing the true underlying effect and variance $\sigma^2_{M_i}$, with the latter set to its asymptotic estimate $\hat{\sigma}^2_{M_i}$. This builds the first stage of the hierarchical model

$$1^{st}\ level \qquad \hat{\beta}_{M_i}|\mu_{M_i} \sim N(\mu_{M_i}, \hat{\sigma}^2_{M_i}) \qquad , i = 1, \dots, N_M. \tag{4.14}$$

Furthermore, these underlying SNP effects are assumed to depend on external information modeled in the second stage by

$$2^{nd}\ level \qquad \mu_{M_i}|\mu_0, \beta, Z, \tau^2 \overset{id}{\sim} N(\mu_0 + Z\gamma, \tau^2) \qquad , i = 1, \dots, N_M, \tag{4.15}$$

with μ_0 the overall average effect across all SNPs and τ^2 the between-SNP variance. $\gamma = (\gamma_{C_1}, \dots, \gamma_{C_{N_C}})^T$ denote the effects of the N_C different external information components. Since this hierarchical model has exactly the form given in example 1 in section 4.2, we will refer to that example for the further calculations of the marginal distribution, MMLE estimates of the hyperparameters and posterior distribution. The posterior estimates of μ_{M_i} are given by

$$\hat{\mu}_{M_i} = BZ\hat{\gamma} + (1 - B)\hat{\beta}_{M_i} \text{ with } B = \frac{\hat{\sigma}^2_{M_i}}{\hat{\sigma}^2_{M_i} + \hat{\tau}^2}. \tag{4.16}$$

Hierarchical Bayes prioritization (HBP) of Lewinger *et al.* (2007)

In 2007, Lewinger *et al.* published another parametric empirical Bayes approach integrating external information in a GWAS. Similar to the method of Lebrec *et al.* (2009), his hierarchical model for Bayesian SNP prioritization is based on a summary measure for each SNP rather than the direct genotype and phenotype data. However, Lewinger *et al.* (2007) chose a χ^2 distributed test statistic for each SNP as starting point instead of an effect estimate obtained by logistic regression. Since this empirical Bayes approach of Lewinger *et al.* (2007) builds the basis for this thesis, we will explain that method in more detail in the following section.

4.4 Lewinger's hierarchical Bayes prioritization for genome-wide association studies

In comparison to the two methods listed before, Lewinger *et al.* (2007)'s hierarchical model for incorporating external information consists of three rather than two levels. The prior covariates are integrated in the hyperprior distributions in the third level to influence the estimates of the hyperparameters and prioritize SNPs that would not have been selected by pure p-value ranking.
By careful considerations which model fits best in this context of integrating external information into GWAS analyses, Lewinger *et al.* (2007) found a reasonable, obvious prior distribution for modeling the parameters of interest for this purpose. Hence, the choice of a parametric approach is justifiable, leading to higher efficiency than a non-parametric approach, given a valid type of distribution and no assumption violations. Lewinger *et al.* (2007) chose an empirical approach, so that no arbitrary additional last level prior with fixed parameters has to be specified (Carlin and Louis, 2000). By simulation studies comparing the empirical approach with the full Bayesian method using MCMC, Lewinger *et al.* (2007) showed that similar results can be obtained. However, the empirical approach has the advantage that the computationally intensive MCMC method for large number of makers given in GWAS could be avoided.
Because this model was already adapted to GWAS incorporating external information and the model and prior distribution were reasonable in the given context, the model is used for this thesis, having especially the possibility to integrate pathway information into the analysis for improving the results. Furthermore, a reasonable modification for the application to detect GxE (chapter 6) was developed. In addition, both versions of the hierarchical model, the original HBP by Lewinger *et al.* (2007) and the modified GxE version, can be used to combine the detection of GxE interactions with the consideration of pathway information. All in all, this model provides a promising approach to unveil the complex etiology of diseases, considering pathway information on the one hand and GxE interactions on the other hand.

4.4.1 The hierarchical Bayes model

In this approach an association measure for each considered genetic marker is modeled in the **first stage** of a hierarchical model. As measure for association, the χ^2 statistic is chosen, depending only on one parameter, the non-centrality parameter. In the **second**

stage of the model, these non-centrality parameters are modeled by a prior composed of a large mass at zero (unassociated markers) and a continuous distribution of nonzero values (associations). The prior probability of nonzero values and their prior means themselves are modeled on the **third level** as functions of the prior covariates reflecting the prior knowledge that characterizes the markers.

For each SNPs $M_i, i = 1, \ldots, N_M$ in our GWAS, a simple single SNP association test with an asymptotic χ^2 distributed test statistic $T^2_{M_i}$ with non-centrality parameter $\lambda^2_{M_i}$ and one degree of freedom ($\chi^2_1(\lambda^2_{M_i})$) is performed. For simplification and since the direction of the effect is not of interest for our model, we will focus on the unsigned statistics $T_{M_i} = +\sqrt{T^2_{M_i}}$ which are asymptotically χ distributed with non-centrality parameters $\lambda_{M_i} = +\sqrt{\lambda^2_{M_i}}$ and 1 degree of freedom ($\chi_1(\lambda_{M_i})$). The density of a 1 df noncentral χ distribution $\chi_1(\lambda)$ is given by $f(y|\lambda) = \varphi(y-\lambda) + \varphi(y+\lambda), y \geq 0$, where φ denotes the standard normal density. Thus, it is equal in distribution to the absolute value of a normal random variable with mean λ and variance 1 (Evans *et al.*, 2000). A graphical presentation of the connection between the normal and χ distribution is given in figure 4.4.
The test statistics build the **first level** of our hierarchical model

$$1^{st}\ level \quad T_{M_i}|\lambda_{M_i} \sim \chi_1(\lambda_{M_i}) \qquad i = 1, ..., N_M. \quad (4.17)$$

The noncentrality parameters λ_{M_i}, i=1,...,N_M, are the main objects of interest and will be modeled in the **second level** of the hierarchy. For the SNPs not associated with the disease, we have that $\lambda_{M_i} = 0$ (null hypothesis), while the associated SNPs have $\lambda_{M_i} > 0$. In GWAS usually 500.000 up to 2 million SNPs are considered. Most of them will have no association, with perhaps only some to several hundred SNPs associated. Since we have a strong prior belief that most of the SNPs are not involved in the examined disease, we adopt for the $\lambda_{M_i}, i = 1, \ldots, N_M$, a mixture model of the form

$$2^{nd}\ level \quad \lambda_{M_i}|p_{M_i}, e_{M_i}, \sigma \sim p_{M_i}\sigma\chi_1(e_{M_i}) + (1-p_{M_i})\delta(0) \qquad i = 1, ..., N_M. \quad (4.18)$$

p_{M_i} is the prior probability that marker M_i is associated with the disease. Given an association ($\lambda_{M_i} > 0$), λ_{M_i} is assumed to be have a χ_1 distribution with noncentrality parameter e_{M_i} as measure for its strength of association and a scaling parameter $\sigma > 0$. $\delta(0)$ denotes a point mass concentrated at $\lambda_{M_i} = 0$ given no association. Simulation studies showed that the positive square root of the non-centrality parameters $\lambda_{M_i} = +\sqrt{\lambda^2_{M_i}}$ can be reasonably modeled by a χ distribution with 1 degree of freedom.

e_{M_i} and p_{M_i} are not declared as constant across all SNPs, but can depend on some of the prior knowledge that is given to characterize the markers. This prior information to be incorporated for marker M_i is formalized by the vectors $Z^{\mu}_{M_i} = (z^{\mu}_{M_i 0}, z^{\mu}_{M_i C_1}, \ldots, z^{\mu}_{M_i C_{N_{C_\mu}}})$ and $Z^{\beta}_{M_i} = (z^{\beta}_{M_i 0}, z^{\beta}_{M_i C_1}, \ldots, z^{\beta}_{M_i C_{N_{C_\beta}}})$, with N_{C_μ} and N_{C_β} "prior covariates" and intercept term $z^{\mu}_{M_i 0} = z^{\beta}_{M_i 0} = 1$. $Z^{\mu}_{M_i}$ and $Z^{\beta}_{M_i}$ may carry the same covariates and be identical or involve different kinds of prior information. It is even possible, that only the prior probabilities p_{M_i} or the prior noncentrality parameter e_{M_i} depend on prior information. The information is included in two regression models for the prior probabilities p_{M_i} and

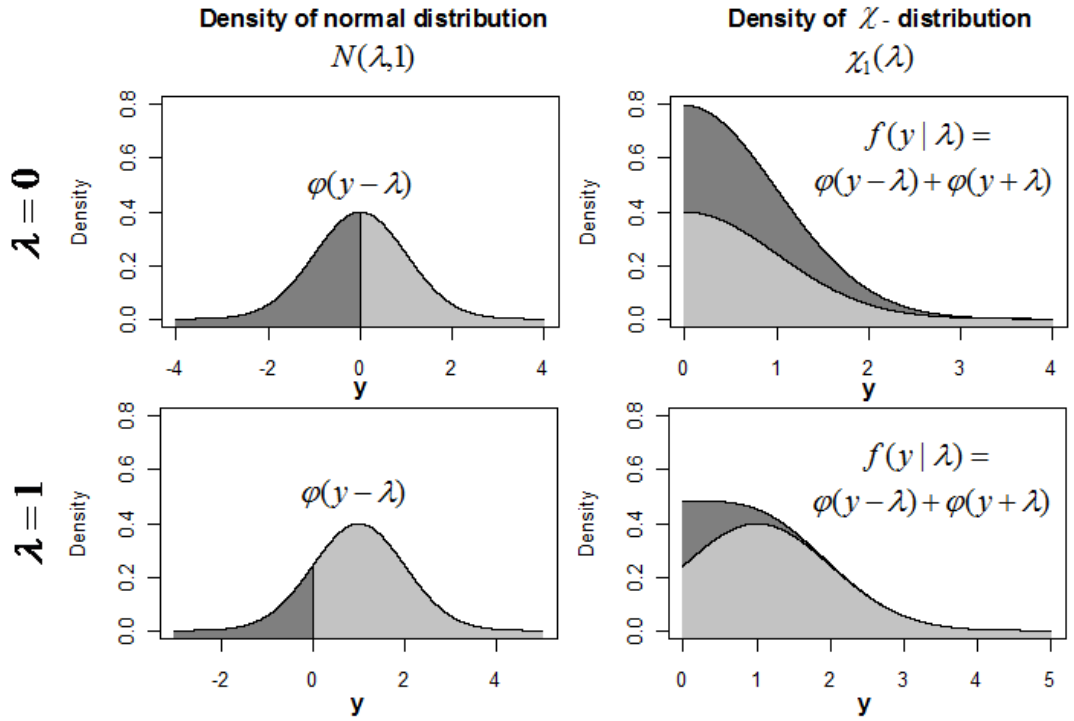


Figure 4.4: *Connection between normal and χ distribution with φ denoting the standard normal density and f the density of the χ distribution with one degree of fredom and non-centrality parameter λ.*

the prior expectations e_{M_i}, building the **third level** to complete the hierarchical model $(i = 1, \ldots, N_M)$

$$3^{rd}\ level \quad \text{logit}(p_{M_i}) = \text{logit}\ \Pr(\lambda_{M_i} > 0) = \beta^T Z^{\beta}_{M_i} \tag{4.19}$$

$$e_{M_i} = |\mu^T Z^{\mu}_{M_i}|. \tag{4.20}$$

The regression parameters $\mu^T = (\mu_0, \mu_1, \ldots, \mu_{N_{c_\mu}})$ and $\beta^T = (\beta_0, \beta_1, \ldots, \beta_{N_{C_\beta}})$ represent the influence of the different prior covariates to the prior probabilities of association and the corresponding expectations. For identifiability we have the constraint that the intercept term $\mu_0 \geq 0$.
The hierarchical model can be reduced to 2 levels, by combining the 2^{nd} and 3^{rd} stage to

$$2^{nd}\ level \quad \lambda_{M_i}|\mu, \beta, \sigma, Z^{\mu}_{M_i}, Z^{\beta}_{M_i} \sim \frac{e^{\beta^T Z^{\beta}_{M_i}}}{(1 + e^{\beta^T Z^{\beta}_{M_i}})} \sigma\chi_1(|\mu^T Z^{\mu}_{M_i}|) + \frac{1}{(1 + e^{\beta^T Z^{\beta}_{M_i}})} \delta(0).$$

Note: The most straightforward method would be to model the signed version of the test statistics $\sqrt{T^2_{M_i}}$ by a normal distribution $N(\theta_{M_i}, 1)$ and θ_{M_i} by a mixture of the point mass at zero and $N(\mu_{M_i}, \tau^2)$. Though, Lewinger *et al.* (2007)'s simulation studies showed that the distribution of θ_{M_i} given an association is not symmetric and therefore cannot be adequately modeled as normal. For a marker M_i in LD with a causal locus M_{causal}, we have that $\theta_{M_i} \approx r_{M_i M_{\text{causal}}} \theta_{M_{\text{causal}}}$ with $r_{M_i M_{\text{causal}}}$ the correlation between M_i

and M_{causal}. Since new mutations are more likely to occur in phase with a major allele at nearby loci, a major marker allele is more likely to be positively associated with a causal allele. Hence the distribution of LD measures to a causal locus is not symmetric around zero but has a higher proportion of small negative values and a smaller proportion of large positive values. This results in a non-symmetric distribution of the θ_{M_i}. By using the absolute value of both positive and negative values $\lambda_{M_i} = |\theta_{M_i}|$, this unbalance cancels out and an approximate χ distribution is obtained. Hence, to base the model on a χ distribution and not a normal distribution seems reasonable.

4.4.2 The empirical Bayes analysis

Treating the hierarchical model in an empirical Bayes manner, we have to estimate the hyperparameter σ and vectors β and μ from the data. Thus, the parameter estimation is done on the basis of the observed data, what makes the used approach to an empirical Bayes method. We can calculate the marginal distribution for each single SNP test statistic $T_{M_i}, i = 1, \ldots, N_M$, by multiplying stage one $f(T_{M_i}|\lambda_{M_i})$ and two $\pi(\lambda_{M_i}|\mu, \beta, \sigma, Z^{\mu}_{M_i}, Z^{\beta}_{M_i})$ and integrating over the unknown distribution of λ_{M_i}. Thus the likelihood for marker M_i is given by

$$\begin{aligned} L_{M_i} &= \Pr(T_{M_i}|\mu, \beta, \sigma, Z^{\mu}_{M_i}, Z^{\beta}_{M_i}) \\ &= \int f(T_{M_i}|\lambda_{M_i})\pi(\lambda_{M_i}|\mu, \beta, \sigma, Z^{\mu}_{M_i}, Z^{\beta}_{M_i})d\lambda_{M_i} \\ &= p_{M_i}\frac{f(T_{M_i}/\sqrt{1+\sigma^2}|\mu^T Z^{\mu}_{M_i}/\sqrt{1+\sigma^2})}{\sqrt{1+\sigma^2}} + (1-p_{M_i})f(T_{M_i}|0) \end{aligned} \tag{4.21}$$

From this, the marginal likelihood of the data is given by $L = \prod_{i=1}^{N_M} L_{M_i}$. By maximizing this likelihood L with respect to $\Theta = (\mu, \beta, \sigma)$ using a standard numerical maximization algorithm, we can obtain estimates $\hat{\Theta} = (\hat{\mu}, \hat{\beta}, \hat{\sigma})$ for the unknown hyperparameters. The posterior probability P_{M_i} of an association of SNP M_i with the disease can be derived by application of the Bayes formula

$$\begin{aligned} P_{M_i} &= \Pr(\lambda_{M_i} > 0|T_{M_i}, Z^{\mu}_{M_i}, Z^{\beta}_{M_i}, \Theta) \\ &= \left(1 + \frac{(1-p_{M_i})}{p_{M_i}}\frac{f(T_{M_i}|0)\sqrt{1+\sigma^2}}{f(T_{M_i}/\sqrt{1+\sigma^2}|\mu^T Z^{\mu}_{M_i}/\sqrt{1+\sigma^2})}\right)^{-1}. \end{aligned} \tag{4.22}$$

In addition, we can obtain a formula for the posterior expectation $E^{+}_{M_i}$ of the underlying noncentrality parameter λ_{M_i} given an association $(\lambda_{M_i} > 0)$

$$\begin{aligned} \mathrm{E}^{+}_{M_i} =& \mathrm{E}(\lambda_{M_i}|\lambda_{M_i} > 0, T_{M_i}, Z^{\mu}_{M_i}, Z^{\beta}_{M_i}, \Theta) \\ =& \frac{\sigma}{\sqrt{1+\sigma^2}}\frac{1}{\varphi(E_{+(M_i)}) + \varphi(E_{-(M_i)})}\left[2/\pi \exp(-(\sigma^2 T^2_{M_i} + (\mu^T Z^{\mu}_{M_i})^2)/(2\sigma^2)\right. \\ &\left. + \lambda_{+(M_i)}\varphi(E_{-(M_i)})(2\Phi(\lambda_{+(M_i)}) - 1) + \lambda_{-(M_i)}\varphi(E_{+(M_i)})(2\Phi(\lambda_{-(M_i)}) - 1)\right] \end{aligned} \tag{4.23}$$

with Φ the cumulative distribution function of a standard normal and

$$\begin{aligned}
\lambda_{+(M_i)} &= (\mu^T Z^{\mu}_{M_i} + \sigma^2 T_{M_i})/(\sigma\sqrt{1+\sigma^2}) \\
\lambda_{-(M_i)} &= (\mu^T Z^{\mu}_{M_i} - \sigma^2 T_{M_i})/(\sigma\sqrt{1+\sigma^2}) \\
E_{+(M_i)} &= (T_{M_i} + \mu^T Z^{\mu}_{M_i})/(\sqrt{1+\sigma^2}) \\
E_{-(M_i)} &= (T_{M_i} - \mu^T Z^{\mu}_{M_i})/(\sqrt{1+\sigma^2}).
\end{aligned} \tag{4.24}$$

Then we can plug in the estimates $\hat{\Theta}$ into the formulae to yield P_{M_i} and $E^+_{M_i}$ for each SNP M_i. One of these posterior quantities or their product

$$E_{M_i} = P_{M_i} E^+_{M_i} = \mathrm{E}(\lambda_{M_i} | Z^{\mu}_{M_i}, Z^{\beta}_{M_i}, Z_{M_i}, \hat{\Theta}) \tag{4.25}$$

can be used for SNP ranking and selection for further investigations.

4.4.3 Evaluation of the approach

By simulation studies comprising different parameter settings, Lewinger *et al.* (2007) evaluated the performance of the Bayes method. He compared the regular SNP ranking based on the classical χ^2 distributed statistics $T^2_{M_i}$ with the ranking according to the three posterior quantities P_{M_i},$E^+_{M_i}$ and E_{M_i}. He included the covariate information either in only one of the third level regression models, or in both of these submodels. The power for the different ranking strategies was calculated by the proportion of true positive SNPs within a given fraction of top SNPs selected for further research.
The simulation studies showed that ranking by $E^+_{M_i}$ was less powerful than the other ranking strategies for all parameter combinations. Thus $E^+_{M_i}$ should not be used to select SNPs for further investigation. In contrast, the performance of P_{M_i} and E_{M_i} depend on the power situation and the informativeness of the prior covariate. In the situation of medium to high power with respect to the GWAS χ^2 results, both ranking strategies reached nearly the same power as using the original test statistic $T^2_{M_i}$. Because of a really large effect, the raw $T^2_{M_i}$ statistic can already pick the true association signals and the prior covariates cannot gain more power although highly informative. In the case of low $T^2_{M_i}$ ranking power, P_{M_i} and E_{M_i} can improve the results when some of the prior covariates are informative and have a strong effect to the prior probability of a SNP to be associated and/or the corresponding prior strength of association. When all covariates are non informative or have only a very small contribution, the ranking by P_{M_i} and E_{M_i} shows slightly lower power than the initial ranking.
The informativeness of the covariates was shown to be the most important factor influencing the power. Furthermore, is also depends on the informativeness which Bayes ranking strategy, P_{M_i} or E_{M_i}, shows best results. However, overall P_{M_i} or E_{M_i} should be preferred, with both of them showing comparable output.
When the full Bayesian approach was used, the discrepancies in ranking compared to the empirical Bayes procedure were large for SNPs with low χ statistics and only low for SNPs with medium to large χ values. Hence, since the latter that remain consistent build the top ranking and are of interest for our application, the empirical Bayes approach offers a good alternative.

4.4.4 Conclusion

Lewinger *et al.* (2007) presented a new method for SNP selection in a first step of a genome-wide association study to control the number of false positive findings without missing too many scientifically interesting associations. Therefore, an empirical hierarchical Bayes appraoch was used, that re-ranks SNPs by prioritizing markers with external support even if their corresponding test statistics show only low to moderate association effects.
Depending on the informativeness of the covariates, this approach of marker re-ranking can reach at least comparable or even higher power than the original ranking. Particularly, in GWAS with a relatively large proportion of truly associated SNPs the method can be superior. In the context of GWAS with only a few true associations, the regression models will be driven mainly by correlations with false positive markers, what will not improve the resulting ranking. Nevertheless, this obvious problem of the approach may be overcome by the incorporation of an additional prior about the proportion of expected true associations or corresponding likelihood penalties.
The main advantage of the approach is that it is very flexible, since multiple sources of information can be included in the analysis, without prespecifying any weights for the covariates. The weights are estimated from the observed data itself. Results from earlier studies can easily be considered as corresponding covariates. Wheen prior evidence for only some specific SNPs is given, it can be included in the probability model in form of an intercept offset. Ignoring LD between the SNPs does not affect validity and efficiency of the method.

5 Integration of pathway information in the analysis of genome-wide association studies

5.1 Motivation

The importance of biological pathways in the etiology of complex diseases is obvious and should not be ignored. A more detailed description of their prominent role in disease development was outlined in section 2.4.1. Therefore, involving knowledge about biological pathways in the analysis of GWAS can be seen as an attractive approach to utilize the wealth of data more effectively to complement and improve the traditional single SNP results. The ability to find further associations not detected by using the GWA data alone is increased by the integration of the pathway data (Tintle *et al.*, 2009a). Weekly associated genetic variants that cluster in pathways can be identified using this method. This has the potential to reveal a more detailed picture of the genetic causes of diseases (Chasman, 2008; Sohns *et al.*, 2009).
In the last years, various methods to utilize biological pathway information were suggested to increase the power of GWAS. In this thesis, we are concentrating on those approaches that need a preceding traditional single SNP analysis. However, according to their main idea we can group them in two different classes: gene set identification and gene or SNP prioritization (Tintle *et al.*, 2009a).
The first class of methods directly focuses on the identification of whole sets of biologically relevant genes (gene sets) significantly associated with the disease rather than single SNPs or genes. This can be done by examining if a pathway is enriched with genes represented towards the top of a ranking list based on the information from the traditional GWAS analysis. These methods are denoted as gene set analysis methods (GSA) and originate from gene expression. A gene set is defined as a set of genes related to each other by function, structure, nomenclature or in particular biological pathways.
The conception of the second group of methods is to prioritize SNPs by re-ranking the traditional analysis results using the pathway information. SNPs in the same pathway with only small association effects can support each other to up-rank, resulting in a better ranking and SNP list. This can be done by hierarchical Bayes methods integrating external information in form of covariates (see sections 4.3 and 4.4). The external information we use is the knowledge about biological pathways. New associated SNPs may be discovered by this proceeding and thus the selection of SNPs for further investigations improved.
In our research on the topic of integrating pathway information into GWAS analyses, we concentrated on the HBP model of Lewinger *et al.* (2007) described in the previous chapter. We worked out two advanced strategies: a two-step HBP method and a combination of a gene set analysis method and the HBP. In an application, results of the original "one-step" HBP, both derived strategies and the gene set analysis method were compared to each other.
Before we will outline our new work in sections 5.4-5.6, we will address gene set methods in more detail in the following section and explain the strategies used in our international cooperations. General issues in pathway analysis will be handled in section 5.3, including our choice how to address these questions. Section 5.4 illustrates our individual work based on a Rheumatoide Arthritis GWAS. Section 5.5 includes the comparison

of our results to those of other investigators working on the same data with the same aim. We will end with a closing discussion in section 5.6.

5.2 Gene set analysis methods

Gene set analysis (GSA) methods were initially used in gene expression studies (Subramanian *et al.*, 2005). The basic principle of these approaches is to assign significance to a *priori* defined sets of genes instead of single genes. The goal is to increase the power for subtle but consistent effects within groups of genes of which several jointly account for an association with a disease (Tintle *et al.*, 2009a).
In 2007, Wang *et al.* introduced the idea of GSA methods in the field of GWAS, where they became popular in the last years (Chasman, 2008). The basic principle of GSA in the context of GWAS is to combine statistics from single SNP tests into a single statistic per gene (Tintle *et al.*, 2009a). This gene level statistics are used to evaluate gene set significance. For the assessment of significance of pathways, two different principles are commonly employed. Either, one can calculate the over-representation of each gene set among the top ranked genes (over-representation analysis methods, ORA), or alternatively compute an enrichment score for each pathway and assess the significance by a permutation method (gene set resampling methods, GSR) (Lee *et al.*, 2005).
Furthermore, we can distinguish GSA methods concerning the definition of the statistical null hypothesis in competitive and self-contained testing (Fridley *et al.*, 2010; Goeman and Buehlmann, 2007; Tian *et al.*, 2005; Wang *et al.*, 2010). Competitive methods compares the genes within a gene set to the other genes in the genome to see if the set shows the same pattern and level of association with the disease than the complement genes (Nam and Kim, 2008; Tian *et al.*, 2005; Wang *et al.*, 2011). Thereby, the relative enrichment of the set compared with the background can be evaluated (Nam and Kim, 2008). So the researcher can determine whether the genes in a set tend to be more associated with the given disease (Wang *et al.*, 2010). On the contrary, a self-contained test directly tests the gene set association with the disease by assessing if the gene set contains any genes correlated with the disease (Tian *et al.*, 2005; Wang *et al.*, 2011). These methods use only the results given for the genes in the set of interest and do not depend on the genes outside the set (Nam and Kim, 2008; Wang *et al.*, 2011). ORA methods are competitive by definition, a GSR method may be competitive or self-contained.
When causal SNPs are fully contained in one particular gene set, both hypothesis assumptions will lead to similar results (Wang *et al.*, 2011). Given causal SNPs located in multiple gene sets or genes shared by multiple gene sets, self-contained testing is more powerful (Chai *et al.*, 2009; Hong *et al.*, 2009; Wang *et al.*, 2011). However, a significant fraction of associated genes may implicate that large but irrelevant gene sets that are purely a random subset of the entire gene list contain many associated genes just by chance and hence rank high (Tian *et al.*, 2005). In addition, only a single gene can make a whole set significant (Nam and Kim, 2008). Using competitive testing, gene sets without any associated gene may be identified due to special association patterns resulting from tightly correlated but unimportant genes. For both approaches, the interpretation of results should be treated with caution (Tian *et al.*, 2005).

Table 5.1: *Overview of gene set analysis methods discussed in this thesis. ORA: over-representation analysis; GSR: gene set resampling*

	method	underlying principle	underlying hypothesis	software	publication
ORA	Fisher's test	Hypergeometric distribution	competitive	EASE	Chasman (2008), Hosack *et al.* (2003)
	EASE score	Modified Fisher's exact test	competitive	EASE	Hosack *et al.* (2003)
	Binomial	Binomial distribution	competitive		
GSR	SLAT	Sum of -log(p-values)	Self-contained		De la Cruz *et al.* (2010)
	SUMSTAT	Sum of χ^2 statistics	Self-contained		Efron and Tibshirani (2006) Tintle *et al.* (2009b)
	SUMSQ	Sum of squared χ^2 statistics	Self-contained		Dinu *et al.* (2007), Tintle *et al.* (2009b)
	GSEA	Kolmogorov-Smirnov-like running sum	competitive	GenGen	Subramanian *et al.* (2005), Wang *et al.* (2007)

In the next two sections we will present the over-representation analysis and gene set resampling methods discussed in this thesis. These also represent the most important gene set methods in GWAS so far. A short overview of the methods, their main principle and underlying null hypothesis, as well as corresponding references and available software are given in table 5.1. Although in the context of GWAS the most common approach is to summarize the SNP associations in a first step to reach gene level statistics, the presented methods will all work on a gene level as well as on a SNP level and are easy to transfer. How to obtain the gene level statistics from the SNP results is explained in section 5.3.2.

In the following we will assume that we have examined N_G genes G_j with test statistics t_{G_j} and corresponding p-values $p_{G_j}, j = 1, \ldots, N_G$. To demonstrate the idea of the different gene set methods, we will consider only one gene set S involving $N_{G(S)}$ investigated genes. In praxis, considering a higher number of gene sets, the described proceeding is repeated for each of the sets.

5.2.1 Over-representation analysis

In these methods the over-representation of a particular gene set among the most promising genes (top genes) is measured. Therefore, the genes in the list are classified in two groups (promising genes, others) according to a particular selection criterion (e.g. significance threshold) fixed by the study investigator in advance. The enrichment of the pathway within the promising genes is then evaluated by determining if the observed number of pathway genes among the top is greater than expected by chance. This can be tested by comparing the proportion of top genes within the pathway with the proportion of top genes not in the pathway. These methods are also denoted as cut-off methods. The exact ranking positions and corresponding statistics or p-values do not matter, only if the gene is above or below the threshold.

Table 5.2: *2×2 table underlying over-representation analysis methods. There are N_G different genes covered by the analysis, with $N_{G(T)}$ of these having a test statistic above a certain threshold (top genes). Considering a gene set S that comprises $N_{G(S)}$ genes, k genes of these occur within the top genes as well.*

	promising genes (above threshold)	remaining genes (below threshold)	total genes (covered by SNP chip)
genes in set S	k	$N_{G(S)} - k$	$N_{G(S)}$
genes not in set S	$N_{G_T} - k$	$N_G - N_{G(S)} - N_{G(T)} + k$	$N_G - N_{G(S)}$
	N_{G_T}	$N_G - N_{G_T}$	N_G

Assume in the following that the top list T in our example comprises N_{G_T} genes. These N_{G_T} genes and the $N_{G(S)}$ genes involved in gene set S have k genes in common. Hence, we have k top genes within S. We can present the data in a 2x2 table as given in 5.2. Based on this table, over-representation can be calculated by Fisher's exact test or the binomial approximation.

Fisher's exact test

The test of proportions is based on the cumulative hypergeometric distribution, representing sampling without replacement (promising genes, $N_{G(T)}$) from a finite population of two types of elements (within ($N_{G(S)}$) and not within ($N_G - N_{G(S)}$) the gene set) (Chasman, 2008; Hosack *et al.*, 2003).
To assign significance to the excess of gene set genes within the most promising ones, we calculate the probability to observe k or more pathway genes among the top genes when the latter are randomly drawn from the whole set of genes by

$$p_{S(\text{Fisher})}(X \geq k) = 1 - \sum_{x=0}^{k-1} \frac{\begin{pmatrix} N_{G(S)} \\ x \end{pmatrix} \begin{pmatrix} N_G - N_{G(S)} \\ N_{G(T)} - x \end{pmatrix}}{\begin{pmatrix} N_G \\ N_{G(T)} \end{pmatrix}}. \tag{5.1}$$

This corresponds to the one-tailed Fisher's exact test (FET) (Chasman, 2008; Hosack *et al.*, 2003; Tintle *et al.*, 2009b), with random variable X denoting the number of genes within the gene set of interest and the list of most promising genes.
In the GSA software EASE, a modified variant of Fisher's exact test called EASE Score is additionally implemented (Fehringer *et al.*, 2012; Hosack *et al.*, 2003). The EASE Score is obtained by removing one gene that belongs to the gene set and the top list and the modified gene set p-value $p_{S(EASEScore)}$ is calculated based on this. The EASE Score is a more conservative method that eliminates the significance of unstable gene sets by penalizing sets supported by only a small number of top genes. More robust gene sets are only slightly penalized and therefore favored. The method is inspired by the concept of jackknifing a probability that is used to evaluate the stability of a test by repeatedly removing a single observation from the data and recalculating the statistic. The resulting jackknife distribution is broad for highly variable results, while robust results show a tight distribution (Tukey, 1958).

Binomial test

When the number of examined genes is high, the hypergeometric distribution of Fisher's exact test can be approximated by the binomial distribution. The binomial distribution is the equivalent to sampling with replacement. Using this asymptotic equivalent, the p-value can be calculated more simply by

$$p_{S(\text{Binomial})}(X \geq k) = 1 - \sum_{x=0}^{k-1} \left(\begin{array}{c} N_{G(S)} \\ k \end{array} \right) \left(\frac{N_{G(T)}}{N_G} \right)^k \left(1 - \frac{N_{G(T)}}{N_G} \right)^{N_{G(S)}-k}.$$

The over-representation methods directly result in a p-value for each of the gene sets, with a correction for multiple testing necessary when several sets are tested. Therefore, the traditional FDR and FWER approaches can be used.
For the over-representation methods, the threshold to separate the promising genes from the rest plays an important role. Different thresholds lead to different results. Furthermore, this binary classification of genes ignoring the exact test results of genes and their order implies a great loss of information (Tian *et al.*, 2005).

5.2.2 Gene set resampling

Gene set resampling (GSR) methods do not required a threshold specification for "gene selection" but use the single gene results to produce a gene set score. Non-promising genes can contribute to the score as well and more information is preserved than in ORA methods (Lee *et al.*, 2005). A greater gene set score represents a greater enrichment with top resulting genes (enrichment score) and significance is assigned using a permutation method (Curtis *et al.*, 2005). GSR methods tend to be more robust than over-representation methods (Lee *et al.*, 2005). Several alternative methods were proposed to obtain the gene set score and in the following we will explain a selection of these used in our applications or by our cooperation partners relevant for this thesis. The choice of the resampling strategy for assessing significance to gene sets is one of the issues in pathway analysis and different possibilities are outlined in section 5.3.4.

Combining p-values

A popular method to combine results in meta-analyses is **Fisher's combination test**. In the context of gene set analyses, we can use this procedure to combine the p-values for all genes within a particular gene set. For p-values $p_{G_{j(S)}}, j = 1, \ldots, N_{G(S)}$ corresponding to the genes within the gene set S, the Fisher's combination test statistic is given by

$$Z_{S(\text{Fisher})} = -2 \sum_{j=1}^{N_{G(S)}} \log(p_{G_j(S)}). \tag{5.2}$$

Under the assumption of independence between the different genes, this score follows a χ^2-distribution with $2N_{G(S)}$ degrees of freedom. However, since we cannot assume that all genes are independent, permutation based methods have to be used to assign significance.

In 2010, De la Cruz *et al.* (2010) suggested a modified version of this test for the GWAS analysis on a pathway level based on single SNPs rather than gene level p-values. This method alters by using truncation to preselect p-values most likely to carry a true signal and using weights to incorporate other prior information and deal with the SNP marker correlation. Assume that $N_{M(S)}$ SNPs belong to the genes in gene set S and their ordered p-values are given by $p_{M_{(1)}(S)} \leq \cdots \leq p_{M_{(N_{M(S)})}(S)}$. The combined statistic according to De la Cruz *et al.* (2010), denoted as SLAT (set level association testing), is defined by

$$Z_{S(\text{SLAT})} = -\sum_{l=1}^{N_{M(S)}} w_l \log(p_{M_{(l)}(S)}) I_{\{p_{(M_{(l)}(S)<\alpha_l\}}}, \tag{5.3}$$

where α_l are the truncation thresholds and w_l, $l = 1, \ldots, N_{M(S)}$, the weights for the different markers. We can fix a particular threshold for all SNPs with all p-values below used for the statistic. Alternatively, we may set $\alpha_1, \ldots, \alpha_r = 1$ and $\alpha_{r+1}, \ldots, \alpha_{N_{M(S)}} = 0$, corresponding to a rank truncation selecting a fixed number of r top SNPs contributing to the statistic. In addition, the α values can be chosen inspired by the step-up method of Benjamini and Hochberg (1995) to control the FDR or the higher-criticism method that is used in model selection. More details for this can be found in De la Cruz *et al.* (2010) and Donoho and Jin (2004). The weights can incorporate prior knowledge about the markers, e.g. about their relevance or the accuracy of the p-value, and information about the dependencies among SNPs (LD). Nonsynonymous SNPs for example should have more relevance and can be up-weighted. LD structures can be considered by applying lower weights to highly correlated SNP groups to compensate their inflating effect to the overall statistic. A recommendation how to do this precisely can be found in De la Cruz *et al.* (2010).
This method of De la Cruz *et al.* (2010) was used by one of our cooperating partners to compare different gene set methods based on the same data as in our applications of chapter 7. We will shortly outline his main results there as well. For the method of De la Cruz *et al.* (2010) he used $\alpha_1, \ldots, \alpha_{N_{M(S)}} = 0.05$.

Combining test statistics

Another simple method to combine gene results to gene sets enrichment scores is to simply sum the corresponding test statistics. Efron and Tibshirani (2006) originally suggested this strategy for expression data (MAXMEAN) and Tintle *et al.* (2009b) used the same idea in the context of GWAS in 2009. For the gene set S involving genes $G_{j(S)}$, $j = 1,\ldots,N_{G(S)}$, with χ^2 test statistics $t_{G_{j(S)}}$ we can calculate the enrichment score according to the sum of the test statistics (SUMSTAT) by

$$Z_{S(\text{SUMSTAT})} = \sum_{j=1}^{N_{G(S)}} t_{G_j(S)}. \tag{5.4}$$

Based on another method originally suggested in the context of gene expression (SUMGS) (Dinu *et al.*, 2007), Tintle *et al.* (2009b) furthermore proposed to alternatively use the sum of the squared test statistics (SUMSQ)

$$Z_{S(\text{SUMSQ})} = \sum_{j=1}^{N_{G(S)}} t^2_{G_j(S)}. \tag{5.5}$$

Weighted Kolmogorov-Smirnov-like running sum

The most popular GSR method is the gene set enrichment analysis (GSEA) based on a weighted Kolmogorov-Smirnov-like running sum. This method was originally proposed for gene expression by Mootha *et al.* (2003) and Subramanian *et al.* (2005) and transferred to the context of GWAS by Wang *et al.* (2007) in 2007, who first suggested to use pathway based analyses to complement GWAS single-SNP analyses. In 2008, Holden *et al.* (2008) published the GSEA-SNP, a SNP-based version of GSEA for GWAS.
In comparison to the previous methods which used only the test statistics or corresponding p-values for the genes within the gene set, GSEA additionally considers the distribution of these genes in the entire ranked list (Tian *et al.*, 2005). Therefore, the ordered gene list is processed gene by gene starting at the top of the ranking and increasing the score of a gene set when a gene is in that set and decreasing it else (Curtis *et al.*, 2005; Tian *et al.*, 2005). The magnitude of the increment depends on the test statistic of the corresponding gene, while the decrease is of same size for all non-gene set genes corresponding to 1/(number of non-gene set genes). The enrichment score is then given by the maximum of the running-sum.
Given in total N_G examined genes $G_{(1)}, \ldots, G_{(N_G)}$ with decreasing sorted test statistics $t_{G_{(1)}} \geq \cdots \geq t_{G_{(N_G)}}$, the enrichment score (ES) for gene set S of size $N_{G(S)}$ is

$$ES(S) = \max_{1 \leq j \leq N_G} \left\{ \sum_{G_{(j*)} \in S, (j*) \leq (j)} \frac{|t_{G_{(j*)}}|^p}{\sum_{G_{(j*)} \in S} |t_{G_{(j*)}}|^p} - \sum_{G_{(j*)} \notin S, (j*) \leq (j)} \frac{1}{N_G - N_{G(S)}} \right\}, \quad (5.6)$$

with parameter p that may be varied. Given $p = 0$ it is the regular Kolmogorov-Smirnov statistic. However, using equal step sizes for all genes, gene sets clustering in the middle of the ranked list yielded high scores, although not enriched. Therefore, Subramanian *et al.* (2005) recommended in the original GSEA algorithm to use $p = 1$, so that the steps correspond to the test statistics and high scores only occur when genes from a set cluster on the top of the ranked list. The ES measures the maximum deviation of concentration of the statistic values in a particular gene set S compared to a randomly picked set. A high ES is obtained when a gene set is concentrated at the top of the ranking list.
Since not all genes of a gene set necessarily participate in disease development but only a subset, Subramanian *et al.* (2005) proposed a strategy to extract the core members of the significant gene sets. These comprise all genes of the set that occur in the ranked gene list before the point where the running sum reaches its maximum deviation from zero and hence drive the enrichment signal. This gene subset is denoted as leading edge subset (LES). Given the significant gene set S, the LES is given by

$$LES_S = \{G_{(j)}\}_{G_{(j)} \in S, j \leq j_{ES(S)}}, \quad (5.7)$$

with $j_{ES(S)}$ the maximum position of the running sum

$$j_{ES(S)} = \arg \max_{1 \leq j \leq N_G} \left\{ \sum_{G_{(j*)} \in S, j* \leq j} \frac{|t_{G_{(j*)}}|^p}{N_t} - \sum_{G_{(j*)} \notin S, j* \leq j} \frac{1}{N_G - N_{G(S)}} \right\}.$$

For a better understanding, figure 5.1 presents the calculation of the running sum as well as the graphical determination of the LES.

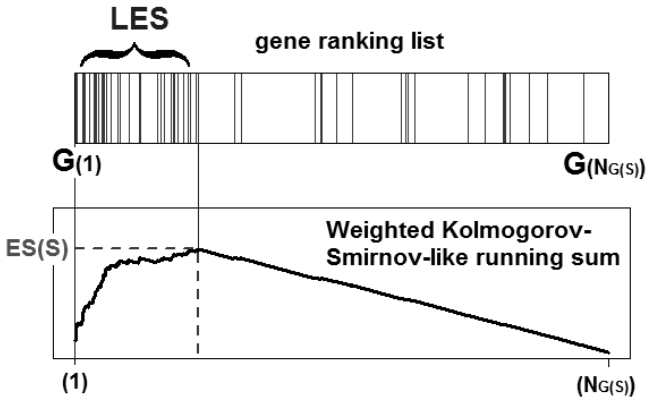


Figure 5.1: *Principle of the gene set enrichment analysis (GSEA) method*
top: *The lines represent genes within the gene set S. The ranking list starts with the gene having the highest test statistic on the left (top of the list).* ***bottom:*** *Kolmogorov-Smirnov-running sum for the gene set S. The Enrichment Score ES(S) is the maximum value of the running sum. The Leading Edge Subset (LES) genes are the genes within the set occurring on the ranking list to the left of the Enrichment Score.*
(adopted from Subramanian et al. (2005))

5.3 Practical issues in pathway based GWAS analysis

Although pathway based methods enjoy increasing popularity in the area of genome-wide association studies, their application is still in its infancy and presents several challenges. Different gene set sizes and gene lengths, as well as the strong correlation of SNPs due to the LD patterns and the presence of overlapping genes may lead to bias. Methodological issues start with assigning the single SNPs to genes and genes to gene sets as well as summarizing marker information on a gene level. Gene set analysis methods also include the construction of the test statistics and finally the assessment of statistical significance to whole gene set by considering the potential sources of bias (Wang *et al.*, 2011). Moreover, when prioritization methods are used a really important issue is how to code the covariates for the gene set information.
Except for the last point mentioned, the listed issues are not the central focus of our work, but we had to deal with them in our applications and make decisions how to solve them. Therefore, we will take a closer look at these different critical steps in pathway based GWAS analyses in the following. We will process the different issues one by one, present typical possibilities and illustrate our decision.
So far, these logistical aspects that accompany the pathway based methods demand subjective choices since more intensive research on these critical points is still necessary to come to an informed decision (Tintle *et al.*, 2009a). Some of the challenges have been explored in greater detail in other areas before and we can take advantage of the lessons learned and use the available information. However, others are still inadequately investigated. Since the coding of the pathway information given as covariates for gene or SNP prioritization methods is a more central point of our work, it will be part of our applications in section 5.4. A graphical overview about the different steps in the analysis flow is given in figure 5.2.

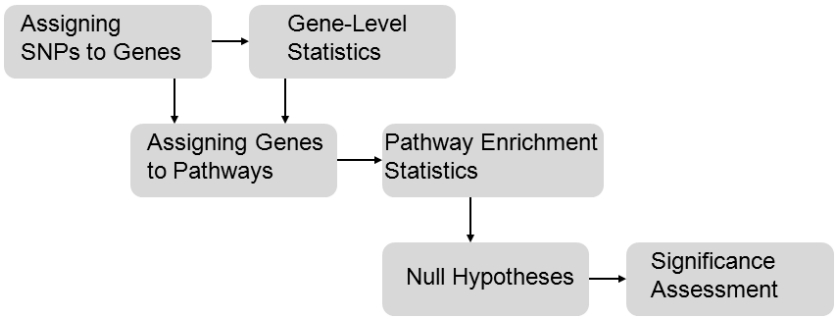


Figure 5.2: *Steps in a pathway based genome-wide association study analysis*

5.3.1 SNP to gene assignment

In comparison to gene set analysis in gene expression studies, one challenge of pathway based methods in GWAS is that we do not have one test statistic per gene but several test results for single SNPs within and close to a gene (Peng *et al.*, 2010). Hence, as a first step in a pathway based analysis, SNPs have to be assigned to genes to relate the data to the pathway information. This holds in particular for our applications.
Several data bases, such as the National Center for Biotechnology Information (NCBI) (2012) data base, the USCS genome browser (Kent *et al.*, 2002) or the Ensembl genome browser (Flicek *et al.*, 2012), provide gene annotation information. This information can include start and end position for the genes, SNP positions as well as direct SNP to gene assignments. The information can be obtained by downloading corresponding annotation files or extracting information from the data base e.g. by R (Melville, 2011) or programming languages for data management (SQL). In addition, Affymetrix (http://www.affymetrix.com) and Illumina (http://www.illumina.com) directly provide annotation files for their SNP arrays, including SNP position and the nearest gene. Since different names can occur for the same gene, caution is necessary to ensure a consistent notation.

When assigning SNPs to genes it is possible to restrict to those markers within gene coding regions. SNPs that can be mapped to the coding region of several genes are assigned to all of them. However, restricting to the coding part does not cover regulatory regions or consider LD. Therefore, the usual strategy to assign SNPs to genes is map SNPs not directly located within a coding sequence to the nearest gene within a certain distance window of +/- X kb down- and upstream. Thereby, the core gene part as well as the boundary regions containing regulatory units are covered (Tintle *et al.*, 2009a; Wang *et al.*, 2010, 2011). The reasonable window size is still unclear and the implications of different sizes are still unknown (Tintle *et al.*, 2009a). Different distances were proposed such as 500kb (Wang *et al.*, 2007, 2010), 200kb (Perry *et al.*, 2009), 100kb (Wang *et al.*, 2010), 20kb (Jia *et al.*, 2010), 10kb (Wang *et al.*, 2010) and 5kb (Chen *et al.*, 2010).
Assigning the SNPs to genes still poses problems for the analysis. So far, there is no exact definition of a gene and the corresponding gene positions vary between different databases and over time. Assigning SNPs to more than one gene induces correlation between them and may result in bias. Furthermore, SNPs may be located in a regulatory region of a gene although this is not the nearest one, resulting in a wrong assignment.

In gene set analysis methods, all SNPs that are not assigned to any gene because they are not close to any one are excluded from further analysis. This can be a severe loss of information. This especially holds when a small window size is used since it may affect several hundreds of thousands of SNPs. A wide window size on the contrary allows the assignment of numerous irrelevant SNPs without any effect to contribute to a gene set and may dilute its potential single strengths (Wang *et al.*, 2011).
To improve gene set analyses in GWAS one may use more sophisticated strategies for SNP assignment. These may for example consider information about exact LD patterns around a gene (Bush *et al.*, 2009; Hong *et al.*, 2009) or regulatory units from expression studies (Wang *et al.*, 2011). Veyrieras (Veyrieras *et al.*, 2008) estimated that most genetic variants that influence gene expression are located within 20 KB around the gene (Wang *et al.*, 2011).
For our application presented in section 5.4, we assigned SNPs to genes using the corresponding SNP annotation file available from Illumina upon request. Each SNP was assigned to the nearest gene within +/- 500kb. This relatively large distance guaranteed to miss no regulatory or LD region of the gene, even though that was accompanied by assigning many non-relevant SNPs to the gene as well. By using 500kb the number of SNPs not assigned to a gene and hence not considered at all in the gene set analysis is reduced.

5.3.2 Gene-based test statistic

Since pathway based methods in GWAS are often performed based on genes rather than SNPs, gene level summary measures of association have to be obtained from the sets of underlying SNPs. Although the HBP in our application parts was based on the SNP level, we performed the GSEA on the gene level so that this issue represents a practical aspect of our work. The best strategy for the reduction of the SNP-level information within each gene is still disputed and represents a whole area of research not pursued here.
In the following assume that we have a gene G and $N_{M(G)}$ assigned SNPs with ordered p-values $p_{M_{(1)}(G)} \leq \cdots \leq p_{M_{(N_{M(G)})}(G)}$ (ascending) and corresponding test statistics $t_{M_{(1)}(G)} \geq \cdots \geq t_{M_{(N_{M(G)})}(G)}$ (descending) from a single SNP analysis. A very simple approach to summarize the single SNP signals at the gene level is to represent each gene by its most significant SNP (maximum SNP statistic, Sidak's combination test, Peng *et al.* (2010)) $p_{G(\text{Sidak})} = p_{M_{(1)}(G)}$ and $t_{G(\text{Sidak})} = t_{M_{(1)}(G)}$. According to Simes's (1986) combination method we could also chose $p_{G(\text{Simes})} = \min_j \left\{ \frac{N_{M(G)} p_{M_{(j)}(G)}}{j} \right\}$ (Peng *et al.*, 2010). Simes procedure was proposed as a more powerful alternative to the Bonferroni correction for multiple testing. For J hypotheses $H_{(1)}, \ldots, H_{(J)}$ with p-values $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(J)}$, $H_{(j)}$, j=1,...,J, is rejected if $p_{(j)(\text{Simes})} = \frac{J p_{(j)}}{j} \leq \alpha$. This method is simple to apply and in particular advantageous over Bonferroni given highly correlated test statistics. The combination method uses the minimum of these Simes corrected p-values $p_{(j)(\text{Simes})}$, j=1,...,J, for the joint test of the whole set of hypotheses (Simes, 1986).
However, for both gene summary methods the loss of information is huge, they do not consider the correlation between the SNPs due to strong LD within a gene and are

susceptible to genotyping errors. Several weakly informative markers within one gene will not be detected (Sohns *et al.*, 2009). Nevertheless, the maximum SNP statistic is often used due to its simplicity.
Instead of picking the result of only one particular SNP to represent the gene, alternatively the results of all or a subset of most significant SNPs assigned to a gene could be combined in one statistic (Chen *et al.*, 2010; Dudbridge and Koeleman, 2003; Hoh *et al.*, 2001; Yu *et al.*, 2009; Zaykin *et al.*, 2002). This may be done by Fisher's combination method (equation 5.2) or other methods known for meta-analysis. However, the assumption of independence between the different p-values is not fulfilled. Depending on the number of SNPs involved, the resulting gene level statistics are distributed with different degrees of freedom. Thus they are not directly comparable to each other. An alternative may be a joint association test such as a multiple regression.
In general, the different gene size and hence different number of SNPs assigned per gene poses a problem and may result in a potential bias. Larger genes are more likely to have larger test statistics and more significant p-values, leading to favor large genes and discriminate small genes (Tintle *et al.*, 2009a). This again leads to privilege gene sets with many large genes rather than gene sets primarily involving small genes. Therefore, an adjustment of the gene level measures is recommended, e.g. captured by permutation methods that are discussed in section 5.3.4.
In a comparison of Ballard *et al.* (2010) among seven multi-marker association tests, including the maximum SNP statistic, principal component regression (Gauderman *et al.*, 2007; Wang and Abbott, 2008) demonstrates to be the most powerful (Wang *et al.*, 2011). Alternatively, the analysis can be performed on SNP level directly using SNP sets instead of gene sets.
In our application given in section 5.4, the maximum test statistic among all SNPs of each gene was chosen to represent the gene test statistic for the GSEA since it is the method most commonly used so far and simple to apply. Our HBP was based on the SNP level.

5.3.3 Pathway information

The most prominent component for pathway based analysis is the biological information used. The quality of gene set analysis results is highly related to the quality of the underlying gene set annotations.
Information about biological pathways can be found in different publicly available databases. Since these sources provide different knowledge about gene sets, the choice of gene set information is a challenge and has to be made arbitrarily. So far, there is no consistent information and only little advice how to correctly choose the most relevant accurate prior information (Tintle *et al.*, 2009a). Despite the high availability of gene set information, the pathway description is not yet sufficient and complete due to our incomplete knowledge about the human genes and their relationships (Sohns *et al.*, 2009; Wang *et al.*, 2011). Various human genes are not well understood or uncharacterized and cannot be mapped to pathways (Wang *et al.*, 2010). Another problem is that genes in general function in multiple ways and therefore appear several times in different pathways. This overlap of gene sets results in redundant information among the sets (Wang *et al.*, 2011), leading to problems in the pathway analysis.

For our analysis presented in this chapter, we used a file provided by Wang *et al.* (2007) in the GenGen-package combining different Web resources (Wang, 2008). This collection involves information from 2076 pathways and gene sets from Biocarta, KEGG and the Gene Ontology (January 14^{th} 2008). Some more information about the gene set databases mentioned in this thesis can be found in the appendix A.3.

5.3.4 Significance Assessment

One important issue in gene score resampling (GSR) methods is the calculation of the empirical distribution of the score under the null hypothesis. The empirical distribution is obtained by recalculating the gene set scores multiple times by a permutation procedure. Disease labels as well as SNP or gene-level statistics can be used as permutation units (Tintle *et al.*, 2009a).
When a phenotype permutation (sample randomization) test procedure is used, the case-control status is randomly shuffled keeping the total number of cases and controls fixed (Sohns *et al.*, 2009; Tintle *et al.*, 2009a). SNP statistics, gene-level statistics and gene set score are recalculated. In the SNP and gene permutation method, the SNP or gene-level statistics are shuffled across the genome and gene set scores are calculated based on these reallocated statistics (SNP or gene randomization). Hence, using gene randomization, random gene sets of the same size are generated for comparison. A comparison of advantages and disadvantages of the different permutation methods is given the appendix B.1.
For each gene set, an empirical distribution of the corresponding score is obtained by the permutations to represent the null distribution. The background distribution established by phenotype permutations represents the null hypothesis underlying self-contained GSA methods, that no SNP and hence gene is associated with the disease. In practice however, some susceptibility SNPs will occur (Wang *et al.*, 2010). SNP and gene randomization permutation represent the compatible null hypothesis of no enrichment as in competitive testing (Ballard *et al.*, 2009; Tintle *et al.*, 2009a,b). Typically, the permutation strategy is chosen ignoring the compatibility with the corresponding null hypothesis of the used gene set method (no association vs. no enrichment) (Wang *et al.*, 2011). This may lead to some bias.
Based on the empirical distribution for a gene set, a nominal p-value can be simply calculated by the fraction of permutation scores for this set that are equal or larger than the original score. Furthermore, since in GWAS numerous gene sets are tested, permutation results may be used for a FWER or FDR method to correct for multiple testing (Curtis *et al.*, 2005). This affects GSR as well as ORA methods.
In the following we will present an example of a phenotype permutation method to assign significance to GSEA scores and calculate the FDR and FWER in more detail. This approach is used in our practical applications to avoid bias due to gene size and keep the correlation structures in the data.

Sample-label permutation procedure in GSEA

This sample-label permutation procedure for GSEA in the context of genome-wide association studies was proposed by Wang in 2007. Assume that we calculated the enrichment score $ES(S_k)$ for each of N_S examined gene sets S_k, k=1,...,N_S. For each

permutation $b = 1, \ldots B$, we randomly assign the original disease labels to the samples, recalculate the SNP and gene-level statistics and recomputed the enrichment score $ES(S_k, b)$ for each gene set to obtain the null distribution of $ES(S_k)$. The nominal p-value is given by

$$p_{S_k(nom)} = \frac{\underset{b=1,\ldots,B}{\#} (ES(S_k, b) \geq ES(S_k))}{B} \tag{5.8}$$

Since gene sets of varying size are not necessarily directly comparable to each other (Wang *et al.*, 2007), the permutations can be used to adjust for these differences between gene sets. Therefore, we can calculate the mean and variance of the permutation scores per gene set by $\hat{\mu}_{S_k} = \sum_{b=1}^{B} ES(S_k, b)$ and $\hat{\sigma}^2_{S_k} = \frac{1}{B-1} \sum_{b=1}^{B} (ES(S_k, b) - \hat{\mu}_{S_k})^2$ and each enrichment score is normalized by subtracting the corresponding mean and dividing by the standard deviation

$$NES(S_k) = \frac{ES(S_k) - \hat{\mu}_{S_k}}{\hat{\sigma}_{S_k}}, \quad NES(S_k, b) = \frac{ES(S_k, b) - \hat{\mu}_{S_k}}{\hat{\sigma}_{S_k}}, b = 1, \ldots, B.$$

To control the FWER, we can calculate the p-value by comparing the true NES of the gene set of interest *NES(S_k)* with the highest NES score over all gene sets per permutation

$$p_{S_k(\text{FWER})} = \frac{\underset{b=1,\ldots,B}{\#} (\max_{l=1,\ldots,N_S} (NES(S_l, b)) \geq NES(S_k))}{B}. \tag{5.9}$$

To control the FDR, the distribution of all $NES(S_l, b)$ over all gene sets $S_l, l = 1, \ldots, N_S$ and permutations $b = 1, \ldots, B$ is used to estimate a FDR q-value for a given $NES(S_k)$. This is given by

$$q_{S_k(\text{FDR})} = \frac{\underset{\substack{b=1,\ldots,B \\ l=1,\ldots,N_S}}{\#} (NES(S_l, b) \geq NES(S_k)) \Big/ (BN_S)}{\underset{l=1,\ldots,N_S}{\#} (NES(S_l) \geq NES(S_k)) \Big/ N_S}. \tag{5.10}$$

Both FDR and FWER were calculated in our applications.

5.4 Analysis of the NARAC data for Genetic Analysis Workshop 16

The Genetic Analysis Workshops are a collaboration of researchers in the whole world to develop, evaluate and compare statistical methods for the detection of genetic effects in complex diseases. For each workshop current analytical issues in Genetic Epidemiology and Statistical Genetics are chosen and can be investigated by different groups based on the same provided data sets. In a meeting, the results from the different investigators are presented, compared and discussed (www.gaworkshop.org). The Genetic Analysis Workshop 16 (GAW 16) held in September 2008 in St. Louis, Missouri, USA focused on the analysis of genome-wide association scans (Cupples *et al.*, 2009).
As a contribution for this workshop we worked on the incorporation of pathway information into GWAS. We investigated the hierarchical Bayes prioritization method of Lewinger *et al.* (2007), since it is perfectly fitted for this purpose. Our focus was the comparison and combination of the HBP with a gene set analysis approach. The GSEA of Wang *et al.* (2007) was chosen as a representative GSA method since it is very popular in expression analysis and was proposed for GWAS shortly before our investigation. We worked out a HBP-two-step method and a combination of HBP and GSEA and applied them to provided Rheumatoid Arthritis data. The results of the four strategies were contrasted and compared with each other for coincidences and differences (Sohns *et al.*, 2009). Furthermore, the biological plausibility of the results was evaluated.
We will present our individual work and results in this section (Sohns *et al.*, 2009), starting with a short description of the provided data set we used (Amos *et al.*, 2009). The next section 5.5 describes our work within the group of all GAW 16 contributors focusing on the incorporation of gene set information, comparing our results (Tintle *et al.*, 2009a). Therefore, the individual results of the other investigators will be shortly outlined as well (Ballard *et al.*, 2009; Lebrec *et al.*, 2009; Tintle *et al.*, 2009b). We will close with an overall discussion about the methods for gene set integration in section 5.6, with the focus on the hierarchical Bayes approach.

5.4.1 Genome-wide data for Rheumatoid Arthritis

Rheumatoid arthritis (RA) is a chronic disease resulting from a complex interaction of genetic and environmental factors. It is an autoimmune disorder causing chronic inflammation primarily in joints but also other tissues and organs of the body. In RA patients, the immune system normally responsible to protect our health by attacking foreign cells mistakenly attacks the own body cells. RA is a progressive illness with painful and disabling acute episodes alternating with periods without symptoms. Long-term, RA may cause joint destruction and permanent functional disability. In Caucasians, the prevalence of RA is 0.8% and the recurrence risk ratio for siblings is estimated to nearly 6. RA can occur at any age, with the mean age of onset in the fifth decade. Women are more often affected than men.
For the diagnosis of rheumatoid arthritis, specific autoantibodies can be used, with anti-cyclic citrullinated peptide (anti-CCP) as best disease predictor and the rheumatoid factor Immunoglobulin M (IgM) representing erosive arthritis. So far, the human leukocyte antigen (HLA) region on chromosome 6p21 is known as a highly important

genetic component in the disease susceptibility. This genomic region contains numerous genes encoding for cell-surface antigens. HLA as a risk factor for RA was implicated by a large number of studies, with consistent evidence for the contribution of alleles of the HLA-DR genes. Beside high main effects of the HLA-DR genes, interactions with other HLA loci are expected (Newton *et al.*, 2004). Several other non-HLA genes that increase risk of RA are known as well, all somehow related to immune response or playing a role in inflammation processes and therefore biologically plausible. In table 5.3 a list of such genes taken from Raychaudhuri (2010) is given. Only a few environmental factors are known. Smoking increases risk by a factor of 2 (Jawaheer *et al.*, 2002) and interacts with predisposing HLA-DR alleles and high levels of anti-CCP (Klareskog *et al.*, 2006). The genome-wide RA data provided for GAW 16 were derived from a genome-wide study to identify genetic risk factors of RA (Plenge *et al.*, 2005) and involved 868 cases and 1,194 controls assayed using the Illumina 550 k platform. The cases were composed of 445 independent individuals from affected sibpairs from the North American Rheumatoid Arthritis Consortium (NARAC) and 423 independent cases not selected for family history recruited across the United States. The controls were derived from the New York Cancer Project (Mitchell *et al.*, 2004). While cases were predominantly of Northern European origin, controls were slightly enriched with individuals of Southern European or Ashkenazi Jewish ancestry. Individuals with an overall call rate <95%, first degrees relatives, duplicated and contaminated samples were already removed.
The data were included in a previous publication showing significant effects for the HLA region, non-HLA gene PTPN22 and identifying a disease causing risk locus between the genes TRAF1 and C5 (Plenge *et al.*, 2007). Beside affection status and sex, levels of anti-CCP and IgM as well as further information for HLA-DRB1 were given. A more detailed description of the dataset is given in Amos *et al.* (2009).

5.4.2 Preprocessing

Both GSEA and HBP are based on an initial ranking of single SNP association analysis. Quality checks were performed to filter out SNPs and individuals of bad quality (missingness >5%, MAF <1%, $p_{HWE} < 10^{-7}$; CR < 90%, relatedness, sex inconsistencies). For the remaining SNPs and individuals, Cochrane-Armitage's trend test was performed to obtain single SNP test statistics. In section 5.3 we presented how the different issues that appear in GWAS pathway analysis were handled. An overview of that can be found in table 5.4 (see page 108). For GSEA, 1,000 permutations were generated.
Since gene names are ambiguous, we used the Gene Name Service (GNS) (Lin *et al.*, 2007) to assure consistency of the gene names in the Illumina Annotation file and the gene sets. Furthermore, gene sets with large overlap were combined and sets with less than 11 genes excluded. Small gene sets are less informative when single SNP analyses are considered and therefore no great loss. Often, they are not well understood, involving many genes not known so far. Since very large gene sets are generally non-specific and hence little gain of information, they are often excluded as well. The reduction of gene sets examined reduces the multiple testing burden and decreases the computation time using resampling-based methods (Lee *et al.*, 2005). Some more detailed information about the preprocessing of the gene annotations and pathway information is given in appendix B.1.

We end up with a total of 876 gene sets for the analysis. Due to computational limitations with the hierarchical model at that time we had to restrict the number of considered pathways to 100. These were selected after the initial single SNP analysis, so that at least one of the top ranked genes was involved in each pathway.

5.4.3 Analysis Strategies

Initially, we directly applied GSEA (I) and HBP (II). For the HBP, the gene set information was only considered in the logistic regression model for the prior probability of association (equation 4.19), but not in the linear model for the prior strength of association (equation 4.20). It was integrated into the model by the covariate vectors $Z^{\beta}_{M_i}$, with a coding of 1 for a SNP in the gene set and a coding of 0 when not (gene set information). Since additional information about the SNP may be important as well, information about the functional position of a SNP and its coding function (SNP information) was considered. The SNP information was modeled in both, logistic ($Z^{\beta}_{M_i}$) and linear ($Z^{\mu}_{M_i}$) regression (equations 4.19 and 4.20), having an impact on the probability of association as well as the corresponding strength. SNPs within protein coding regions for example are expected to be more likely associated and to have a larger effect on the disease than SNP in intergenic regions without regulatory function. Information about the SNP location and coding status was kept from Illumina's annotation file. Illumina classified the SNP location in 7 different mutually exclusive categories: 3' UTR, 5' UTR, (other) UTR, flanking 3' UTR, flanking 5' UTR, intron and coding. The coding type is distinguished in synonymous, non-synonymous and complex. A SNP is **synonymous** when the corresponding codons code for the same amino acid, so that protein's sequence is not changed. When SNP variants lead to different amino acids, so that the SNP influences the protein's composition, it is denoted as **nonsynonymous**. A complex coding SNP results in even more complicated changes of a protein, e.g. by changing the start or end point of the amino acid sequence. The membership of the SNP to the 10 categories was coded in the covariate vectors by 0/1 as well. In figure 5.3 an extract of the covariate matrix $Z^{\beta} = (Z^{\beta}_{M_1} Z^{\beta}_{M_2}, \ldots, Z^{\beta}_{M_{N_M}})^T$ is shown. $Z^{\mu} = (Z^{\mu}_{M_1} Z^{\mu}_{M_2} \ldots Z^{\mu}_{M_{N_M}})^T$ contains only the first 11 columns of that matrix. SNPs were ranked according to their resulting posterior probabilities obtained by the hierarchical model.
Then we thought how HBP can be further improved by dwelling on both strategies advantages. Depending on the research question, there may be an interest in obtaining significance on the gene set rather than on the level of SNPs or genes. This was only obtained by GSEA, but not by HBP. However, external knowledge additionally to the gene set information could only be considered in the HBP but not in GSEA. Therefore, we developed a third strategy that combines the HBP incorporating the SNP information with the GSEA for assigning p-values to gene sets. In a first step, the HBP using the information about SNP function as prior covariates was performed. The resulting posterior probabilities from this model were then further used as the input ranking for the GSEA instead of the initial ranking list. Hence, a gene set significance was obtained while involving SNP information.
Next, HBP's covariates simply indicating if SNPs are located within a gene set (1) or not (0) may not be a particularly appropriate choice. Different other covariate values are conceivable, e.g. considering the number of SNPs within a gene and hence gene set,

	intercept	coding	3' UTR	5' UTR	UTR	3' flanking	5' flanking	intron	synonymous	nonsynonymous	complex	gene set 1	gene set 2	gene set 3	gene set 4	gene set 5	gene set 6	gene set 7	⋯	gene set N
SNP 1	1	0	0	0	0	0	0	1	0	0	0	1	0	1	1	0	0	0	…	0
SNP 2	1	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	…	1
SNP 3	1	0	0	0	0	1	0	0	0	0	0	0	1	0	1	0	0	0	…	0
SNP 4	1	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	…	0
SNP 5	1	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	1	0	…	1
SNP 6	1	1	0	0	0	0	0	0	0	1	0	0	0	0	0	1	1	1	…	0
SNP 7	1	1	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	…	0
SNP 8	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	…	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	1	⋮	⋮	⋮	⋱	⋮
SNP M	1	1	0	0	0	0	0	0	1	0	0	0	1	0	0	0	1	0	…	0

Figure 5.3: *Z matrix for the hierarchical Bayes prioritization method containing information about functional position, coding function and gene set membership for each SNP. SNP 1 for example is an intron-SNP involved in gene sets 1, 3 and 4. SNP 6 is nonsynonymous coding and involved in set 5, 6 and 7.*

or even considering LD, to improve the analysis. However, we proceeded differently and chose set-specific weights derived from the observed data (Sohns *et al.*, 2009). We first used the HBP with SNP information as external information. The resulting posterior probabilities were then used in a second step to build weights for the pathways instead of using the indicators. For each SNP the weight was calculated as a function of the first step posterior probabilities to be associated given by the other genes involved in the same gene set. We therefore picked for each gene $G_j, j = 1, \ldots, N_G$, the maximum posterior probability of all assigned SNPs $M_{i(G_j)}, i = 1, \ldots, N_{M(G_j)}$, to represent the gene (analogous to GSEA)

$$p_{\text{post}-G_j} = \max_{i=1\ldots N_{M(G_j)}} \left(p_{\text{post}-M_{i(G_j)}} \right),$$

with $p_{\text{post}-M_{i(G_j)}} = P(\lambda_{i(G_j)} > 0 | T_{M_{i(G_j)}}, Z_{M_{i(G_j)}}, \hat{\Theta})$ (formula 4.22).
For a SNP $M_{i(G_{j(S_k)})}$ that is assigned to a gene $G_{j(S_k)}$ that in turn is involved in gene set S_k, we calculated the gene set weight for this SNP

$$z^{\beta}_{M_{i\left(G_{j(S_k)}\right)},S_k} = 1 - \prod_{l=1,\ldots,N_{G_{j(S_k)}};\ l\neq j} \left(1 - p_{\text{post}-G_{l(S_k)}}\right)^{1/[(N_{G_{j(S_k)}}-1)]}. \tag{5.11}$$

This is one minus the average posterior probability to be not associated (geometric mean) of all other genes in the same gene set leaving out the particular gene that contains this SNP. Note, that p-values $p_{\text{post}-G_j} = 1$ are substituted by 0.99999, since otherwise the weight would be 1 for all sets involving at least one gene with posterior probability of 1, independent of the remaining gene information. For SNPs not assigned to a gene involved in the particular gene set, the corresponding weight is set to 0.

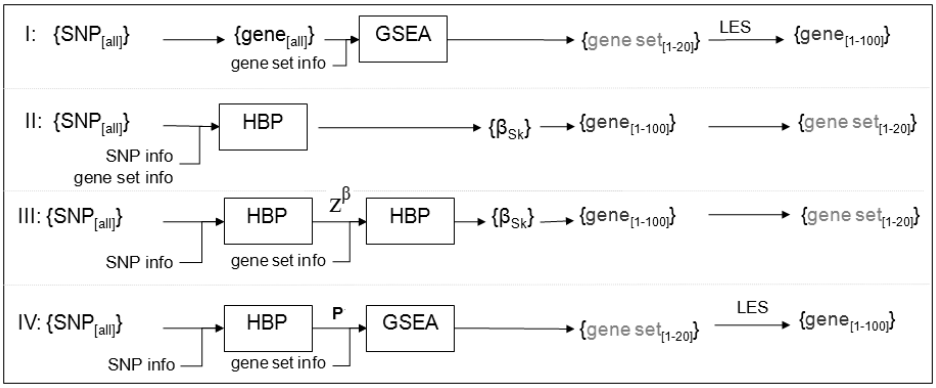


Figure 5.4: *Strategies of data analysis*
SNP info: additional information per SNP; gene set info: information about gene sets; {}: ordered list of SNPs, regression parameters, genes or gene sets; β_{S_k}: parameter of a logistic regression sub-model within HBP; P: posterior probabilities of association (page 80); z^{β} posterior gene set weights (equation 5.11); LES: leading edge subset of genes

Summarized, the four strategies we used for the data analyses were (see also figure 5.4):

I GSEA: GSEA based on single-SNP association test statistics, resulting in gene set p-values

II HBP: HBP based on single-SNP association test statistics using SNP and gene set information coded by 0/1 indicators as prior covariates, resulting in a re-ranking of the SNPs by the posterior probability

III Two-step HBP (HBP+HBP): first HBP based on single-SNP association test statistics using SNP information as prior covariates, second HBP based on single-SNP association test statistics using gene set information as prior with posterior probabilities of first step as gene set weights, resulting in re-ranking of SNPs by final posterior probabilities

IV HBP followed by GSEA (HBP+GSEA): first HBP based on single-SNP association test statistics using SNP information as prior covariates, second GSEA based on obtained posterior probabilities, resulting in gene set p-values

5.4.4 Strategies for result comparison

Results of GSEA and HBP are presented at different levels. While GSEA allocates p-values to gene sets, the output of HBP is a ranking of SNPs. Therefore presentation of results needs to be harmonized to be able to compare the strategies. First we compared the **ranking of most promising genes.** For strategy I and IV we used the leading edge subsets of the top gene sets to obtain a list of 100 genes. For strategy II and III we reduced the resulting list of ranked SNPs by considering only the highest ranked SNP per gene to obtain a ranked list of top genes, limited to the top 100. As ranking criterion, the a posteriori probability of association from the HBP model was used (equation 4.22). We decided for this posterior quantity, since the pathway information

was only included to influence the prior probability of association (equation 4.19), but not the corresponding prior association strength (equation 4.20). Secondly we compared the **ranking of identified gene sets**. For strategy I and IV we ranked the significant gene sets according to their FWER. For strategy II and III we ranked the gene sets according to their corresponding estimated regression coefficients β of the logistic sub-model $\text{logit}(p) = \beta Z$ (equation 4.19), which represent the increase or decrease of the prior probability for each SNP involved in the corresponding gene set. For comparison we considered the top 20 gene sets.
For the quantification of the overlap of the top 20 gene set or top 100 gene lists of the different methods respectively, we calculated an overlap-index. The overlap-index of top lists from different $methods \subseteq initial, I, II, III, IV$ is calculated by

$$I_{methods} = \frac{\#\text{lists}}{\#\text{lists} - 1} \left(\frac{\text{total \# elements in lists - different elements in lists}}{\text{total \# elements in lists}} \right) \quad (5.12)$$

Within the brackets, we have the ratio of the number of duplicated elements within the lists to the total number of elements. The preceding factor is a normalization factor, so that we have an index of 0 when all lists are different and no element occurs twice (worst case), while an index of 1 indicates that all lists have exactly the same elements (best case). In total, the index describes the observed number of duplicated elements in relation to the maximal possible number of duplicated elements. Given five different lists of genes (including initial results) or four lists of gene sets, we can make pairwise comparisons, as well as compare three, four or even all five of the lists at once (#lists = 2,3,4,5). The number of elements per list equals 100 for the genes and 20 for the gene sets. Comparing only two lists with each other, the index gives the proportion of elements in list two that do occur in list one as well. Note, when comparing more than 2 lists, the index may be larger than for the corresponding subsets of lists.

5.4.5 Results

By single SNP analysis of the Rheumatoid Arthritis (RA) data, we observed 334 SNPs with genome-wide significance. This large number of associated SNPs with very small p-values is a specialty of the data that results from the important role of the HLA region in RA development. The significant SNPs belong to 90 different genes including 81 genes of the HLA region. 153 of the 876 examined gene sets involve at least one of these 90 genes. A Manhattan plot of the initial single SNP results is given in figure 5.5. Taking a look at non-HLA genes known for an association with Rheumatoid Arthritis (table 5.3), we found only PTPN22 (rank 55), C5 (rank 78) and TRAF1 (rank 82) within the top 100 genes.
To reduce the number of pathways for the analysis to 100, we started selecting all pathways that involve the top gene, then added the ones that include the gene on rank two and so on, until we reached 100 selected pathways. We processed the top 75 genes to reach that final number. Only two of these gene sets were without genes from the HLA region. This leads to the preference towards HLA in our analyses.
I (GSEA): With the gene set enrichment method alone 47 gene sets reached a FDR< 0.05 (that is nearly half of the considered 100 pathways) and still 20 of these had a FWER< 0.05. These presented our top 20 gene sets used for the method comparison.

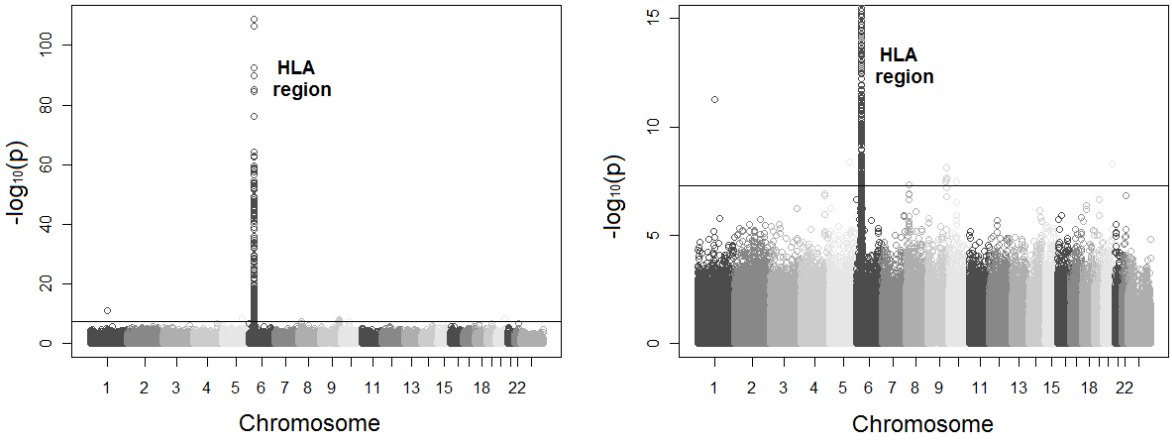


Figure 5.5: *Manhattan plots of single SNP results (initial GWAS).*
left plot: includes all results; right plot: zoom into the left plot to get a better impression of the non-HLA results

All gene sets had a nominal $p < 0.05$. To obtain 100 top genes, we started extracting the leading edge subset of the best ranked set and added the genes of the LES for the next best ranked gene set until we reached 100 genes. For this, the LES of gene set on rank 20 is not considered since the top 100 genes were already filled.
IV (HBP+GSEA): When the HBP preceded the GSEA we observed 7 gene sets with a FDR < 0.05 and 3 of these with FWER <0.05. The two best ranked pathways involved 68 different genes in their LES that were used for the top 100 gene selection. Thus only an additional 32 LES genes from the pathway on rank position 3 were considered, although its LES even involved 67 genes.
II (HBP) and III (HBP+HBP): For the two pure hierarchical Bayes strategies the SNP ranking was changed only slightly by the external gene set information. Taking a closer look at the estimated hyperparameters, we can see that for both strategies 49 of the 100 gene sets involved as covariates in the hierarchical model had positive beta-regression-coefficients in the prior probability model of association. A positive regression coefficient represents an up-ranking of all SNPs involved in the corresponding gene set. Since each of the 100 pathways incorporated in the model involves at least one highly significant gene due to the strategy used for selection of these 100 gene sets, the high number of positive regression coefficients is not surprising. Based on the re-ranked SNP lists of strategies II (HBP) and III (HBP+HBP), a new ranking list on the gene level was obtained by using the maximum posterior probability as a representative for each gene. For comparison purpose, the top 100 genes of these lists were picked.

Comparison of most promising genes
The figures of 5.6 show the comparison of the top 100 genes of all 4 strategies with the initial analysis and each other. In the left plot, the initial ranks of the top 100 genes for each of the the different methods is plotted on the y-axis. The figure on the right displays a Venn-Diagram representing the top 100 gene overlap. As a general overlap-index for all four strategies we obtain $I^{(\text{genes})}_{I,II,III,IV} = 0.51$.
The gene list we obtained by strategy II (HBP) is nearly the same as the one we get

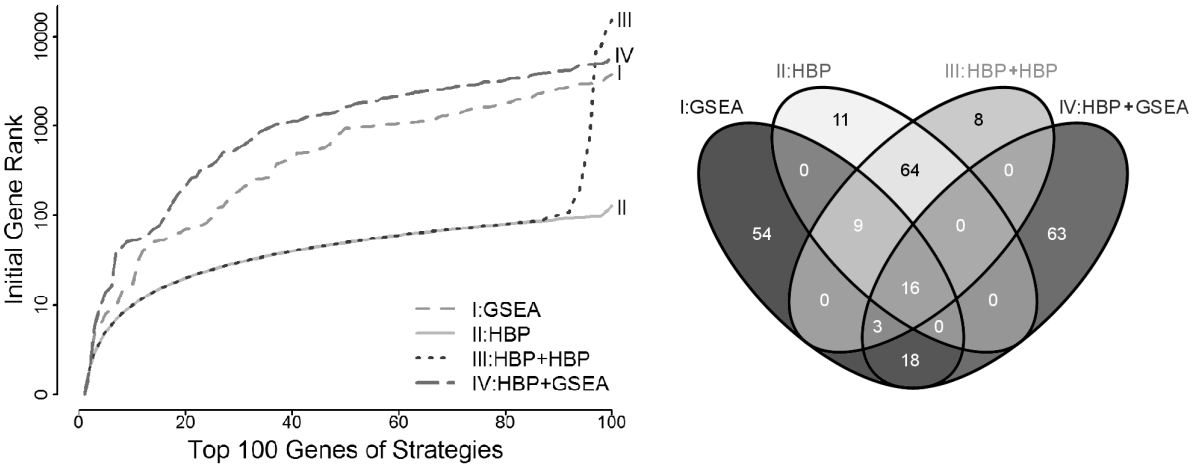


Figure 5.6: *Top 100 genes after applying GSEA, HBP, HBP+HBP and HBP+GSEA.* ***left****: Comparison of the top 100 genes with initial ranking positions. The top 100 genes per method were ordered by initial GWAS rank.* ***right****: Venn diagrams for the overlap of top 100 genes between our four strategies*

by simple single SNP test. Only 2 genes from original rank 112 and 127 are replaced by two other genes and occurred newly in the top 100. The same behavior holds for strategy III (HBP+HBP). The list of top genes is almost identical to the one from strategy II ($I^{(\text{genes})}_{II,III} = 0.89$) and stayed almost unchanged compared with the initial list ($I^{(\text{genes})}_{II,III,initial} = 0.94$, $I^{(\text{genes})}_{II,initial} = 0.98$, $I^{(\text{genes})}_{III,initial} = 0.89$). 11 new genes appeared in the top 100 lists that were initially ranked between rank 229 and 586. Taking a look at the non-HLA genes known for an association with RA (table 5.3), *PTPN22*, *CTLA4*, *CD28*, *CD40*, *PRKCQ* and *PTPRC* reached mentionable higher ranks for strategy III (HBP+HBP) than in the original ranking (table 5.3). For strategy II, no such improvement is observed. Note that for strategies II and III 72 and 82 genes respectively had a posterior probability of 100% and even 3,423 and 6,922 genes had posterior probabilities >80%. Thus the HBP only strategies yielded nearly no change on the gene level. This supports Lewinger *et al.*'s conclusion that the approach is not helpful if highly significant associations occur. Because of the high impact of the HLA genes on chromosome 6p21 and very low p-values occurring for SNPs of these genes, the HBP approach cannot change the results by much. The leading edge subsets of the GSEA strategies highlighted many new genes in comparison to the top 100, primarily HLA genes, of the initial single SNP tests. Only 25 and 16 of the initial top 100 list are also included in the top 100 genes of I (GSEA) and IV (HBP+GSEA) with an overlap index of $I^{(\text{genes})}_{I,IV,initial} = 0.31$ ($I^{(\text{genes})}_{I,initial} = 0.25$, $I^{(\text{genes})}_{IV,initial} = 0.16$). The newly occurring genes in the leading edge subsets had initially ranks of up to 10,789 and 21,361, respectively. Additionally, the LES gene lists of the two GSEA methods also differ remarkable, with an overlap index of $I^{(\text{genes})}_{I,IV} = 0.37$. They have only 37 genes in common. As shown in table 5.3, from the non-HLA genes known to be associated with RA, 11 genes belong to the LES of the significant pathways (FDR <0.05) of GSEA. For HBP+GSEA, only four of these plus one more were found in the LES.
Summarizing promising genes, the HBP only methods reveal the extraordinary impor-

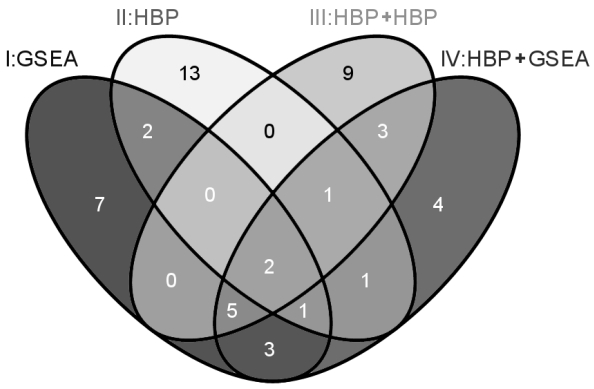


Figure 5.7: *Venn diagrams for the overlap of top 20 pathways between our strategies*

tance of the HLA region with more than 89 HLA genes retained in the top 100 lists. The ranking of genes changed only slightly compared to the initial list. In contrast, apart from only 29 and 19 HLA genes, respectively, the GSEA strategies include many additional other genes, that could be a new starting point to identify yet unknown factors influencing the disease.

Comparison of most promising gene sets

In total, we find 51 different gene sets in the four top 20 lists of the different strategies, with an overall overlap-index of $I^{(\text{gene sets})}_{I,II,III,IV} = 0.48$. Two gene sets, *GO0002460*: the *adaptive immune response* gene set (66 genes) and *hsa04612*: the *antigen processing and presentation* gene set (61 genes), appear in all four lists. Of the 61 genes in hsa04612, 22 are from the HLA region, while GO0002560 contains only 3 HLA genes. Both sets have 3 genes in common (*HLA-DMA*, *CD74* and *LTA*). Since the non-HLA genes in the latter gene set are involved in the activation or inhibition of immune reactions, this set is a reasonable candidate for RA. We can see a tendency of *hsa04612* to reach higher ranks with the GSEA methods (third and fourth rank), as with the strategies with a final HBP step (rank 11 and 17). On the contrary, *GO0002460* ranked between 11^{th} and 19^{th} rank in all four strategies.

Figure 5.7 shows the overlap of the top 20 gene sets per strategy in a Venn diagram. Strategy II yields essentially different results to the others, with list-to-list overlap indices of $I^{(\text{gene sets})}_{I,II}, I^{(\text{gene sets})}_{II,III}, I^{(\text{gene sets})}_{II,IV} \leq 0.25$. The remaining strategies comprise 38 different gene sets, have an overlap-index of $I^{(\text{gene sets})}_{I,III,IV} = 0.53$, ($I^{(\text{gene sets})}_{I,III} = 0.35$, $I^{(\text{gene sets})}_{I,IV} = 0.55$, $I^{(\text{gene sets})}_{III,IV} = 0.55$) and share 5 gene sets. The latter are composed of 11 to 63 genes and have pairwise no more than 5 genes in common. Comparing the strategies involving GSEA (I and IV) we have an overlap-index of $I^{(\text{gene sets})}_{I,IV} = 0.5$, while strategies II and III – both compassing HBP only have an overlap-index of $I^{(\text{gene sets})}_{II,III} = 0.15$. Hence, this indicates a more robust ranking of gene sets by GSEA than HBP, although strategy I and IV even share 11 sets of their top 20.

Although the top 100 gene list of both strategies with HBP as last step (II and III) are almost identical, their top 20 gene sets diverge considerable. Furthermore, it has to be mentioned that the order of gene sets for the HBP only methods (II and III) is

Table 5.3: *Non-HLA genes known for an association with Rheumatoid Arthritis (Raychaudhuri, 2010; Amos et al., 2009)*

	INITIAL gene rank		I: GSEA LES		II: HBP gene rank		III: HBP+HBP gene rank	IV: HBP+GSEA LES
	include HLA	exclude HLA	include HLA	exclude HLA	include HLA	exclude HLA	include HLA	include HLA
PTPN22 Begovich *et al.* (2004)	55	1	24,38		25	1	28	
C5 Plenge *et al.* (2007)	78	4	17,19,20,25, 30,34,42,49	*	80	2	75	*,
TRAF1 Plenge *et al.* (2007)	82	6			83	3	98	
BLK Gregersen *et al.* (2009)	167	65			183	6187	273	
CTLA4 Plenge *et al.* (2005)	358	246	1		382	8452	219	2,*
CD28 Raychaudhuri *et al.* (2009)	400	286	1,6,8,15, 25,31,38	*	407	67	97	1,2,*
CD40 Raychaudhuri *et al.* (2008)	489	374	1,6,8,15, 19,30,31,38	*	510	3372	129	2,*
TNFAIP3 Plenge *et al.* (2007)	570	455	45		548	3835	679	
IL2RB Barton *et al.* (2008)	618	503			809	5439	1168	
PRKCQ Barton *et al.* (2008)	625	510	31,38		603	2342	260	*
REL Gregersen *et al.* (2009)	644	529			677	10746	711	
PRDM1 Raychaudhuri *et al.* (2009)	827	710			885	4409	1235	
AFF3 Barton *et al.* (2009)	1751	1633			1821	2497	1282	
PTPRC Raychaudhuri *et al.* (2009)	1765	1647	19,24,30, 31,38,45	*	1647	398	391	2,*
IL21 Zhernakova *et al.* (2007)	2517	2398	31,38		2694	3684	880	*
CD2 Raychaudhuri *et al.* (2009)	3579	3458	24,31		3895	3073	1011	*
IGSF2 Raychaudhuri *et al.* (2009)	3605	3484	25	*	3692	6212	3395	
TAGAP Raychaudhuri *et al.* (2009)	3674	3553			4084	954	3951	
STAT4 Remmers *et al.* (2007)	4633	4511			4748	3310	4652	
CCL21 Raychaudhuri *et al.* (2008)	4957	4835			4538	11506	5647	
CD58 Raychaudhuri *et al.* (2009)	5718	5595			4023	4575	6627	
IL2RA Thomson *et al.* (2007)	6620	6496			8357	798	4616	*
IL2 Zhernakova *et al.* (2007)	8858	8733			8849	12306	1302	1
TNFSF14 Raychaudhuri *et al.* (2008)	10669	10542			9957	11365	13669	
KIF5A Barton *et al.* (2008)	11263	11135			11793	11919	12079	
TRAF6 Raychaudhuri *et al.* (2009)	12186	12058			12234	14275	6199	
RAG1 Raychaudhuri *et al.* (2009)	12216	12088			16628	1074	17017	
FCGR2A Raychaudhuri *et al.* (2009)	12256	12128			13650	10713	13361	

For initial single SNPs analysis, HBP (II) and HBP+HBP (III) the corresponding gene ranks are given. For GSEA (I) and HBP+GSEA (IV) the ranks of significant gene sets (FDR ≤ 0.05) are given that involve these genes in their leading edge subset (LES). Genes belonging to the LES of sets significant according to the nominal p-value are marked by *. For the GESA excluding the HLA region, none of the gene sets is significant according to FDR.

only build by the estimated regression coefficients (equation of 4.19) that represent the increase or decrease of the prior probability for each SNP involved in the corresponding gene set. The standard deviation of these estimates is neglected. However, for the strategies with GSEA (I and IV), the order of the gene sets is based on gene set significance. In general, taking a closer look at the gene sets that show up in the lists, of both strategies with HBP as last step (II and III) are almost identical ($I^{(\text{genes})}_{II,III} = 0.89$), their top 20 gene sets diverge considerable. Many of them have biological plausibility. For the biological interested reader, the lists of the different top 20 gene set lists are given in appendix table B.1.
To see how the methods behave when we do not have such a huge number of very high signals, we repeated the simple HBP and GSEA analyses excluding the HLA region. Based on the recommendation of Lebrec *et al.* (2009), we excluded all genes in the region between the genes MOG and KIFC1. These comprise 128 genes and 1,336 underlying SNPs. After the exclusion, 9 genome-wide significant genes remained, including PTPN22, TRAF1 and C5 (table 5.3). For the HBP, taking a look at the estimates for the posterior expectations of the noncentrality parameters that represent the strength of association (equation 4.23), PTPN22, C5 and TRAF1 were located on the three top ranks. C5 and TRAF1 were up-ranked in comparison to the initial ranking. Their corresponding posterior estimates were given by 2.21, 1.70 and 1.67 (table 5.3). The estimates for all β parameters of the logistic regression model from level 3 (equation 4.19) were positive, representing an increase of prior probability for SNPs involved in any of the pathways. The basic prior probability from that model for a SNP in none of the pathways was estimated by 0.9. Hence, all SNPs reached a posterior probability to be associated of at least 90%. As it is not plausible that everything is associated with the disease, results have to be considered with caution. In comparison, for the analysis with the HLA region included, the basic prior probability of a SNP involved in none of the genes sets was given by 0.175 and varied for the other SNPs between 0.015 and 0.76 depending on the corresponding pathway membership and SNP information. For GSEA, none of the considered pathways reached a FDR or FWER < 0.05. Hence, GSEA was no help to select candidate genes for further investigations when HLA is excluded.

5.5 Comparison with other results from the Genetic Analysis Workshop 16

In GAW16 two more contributors (Ballard *et al.*, 2009; Tintle *et al.*, 2009b) worked on gene set analysis methods and one other author used a hierarchical model to integrate gene set information in the analysis of GWAS (Lebrec *et al.*, 2009). In the following we will outline the analyses and results of our joint work with these (Ballard *et al.*, 2009; Lebrec *et al.*, 2009; Tintle *et al.*, 2009b), which we carried out after GAW 16. In this joint work, we compared the different investigated methods to integrate gene set information in GWAS with each other. As background for this comparison, we will shortly specify the methods used by the other investigators and illustrate the main findings in their individual analyses. An overview how they handled the different practical issues is given in table 5.4.

5.5.1 Analysis

In contrast to our application of the HBP involving a global overall knowledge about gene sets, Lebrec *et al.* (2009) concentrated on the integration of gene sets already "known to be involved in rheumatoid arthritis" in his hierarchical method. We will denote this approach in the following as linear regression on pathways (LRP). Initially, Lebrec *et al.* (2009) used an empirical Bayes estimate for the between SNP-variance τ^2 of the prior distribution (equation 4.15) and calculated the corresponding posterior gene effects ("untuned" version). Since pathway information was overruled by the strong signals within the HLA region, he repeated the analysis with a reduced relative influence of the GWAS signals by setting $\tau^2_{G_j}$ for each individual gene G_j, so that the ratio $\hat{\sigma}^2_{M_i(G_j)}/\tau^2_{G_j}$ equaled 100. Thereby, the shrinkage factor B for the posterior gene effect (equation 4.16) reduced to B=1/1,0001 and the posterior gene effects were almost completely determined by the gene set knowledge ("tuned" version).
Ballard *et al.* (2009) chose to compare a competitive over-representation analysis method with a self-contained gene set resampling method. Ballard *et al.* (2009) decided to use the binomial to represent the class of competitive ORA methods and compared it with a random set scoring method analog to Fisher's combination method by combining all gene p-values within a set by summing over their negative logarithms. To assess the impact of the HLA region on the pathway results, Ballard *et al.* (2009) performed the pathway analysis including and excluding the 156 genes located in the HLA region.
Tintle *et al.* (2009b) compared four gene set analysis methods with each other: over-representation method FET, competitive gene set resampling method GSEA and self-contained gene set resampling methods SUMSTAT and SUMSQ. In the context of gene expression, GSEA and FET have been shown to be less powerful than other methods (Dinu *et al.*, 2007; Efron and Tibshirani, 2006; Tintle *et al.*, 2008) such as MAXMEAN (Efron and Tibshirani, 2006) and SAM-GS (Dinu *et al.*, 2007). The latter are the analogs to the SUMSTAT and SUMSQ (section 5.2.2). FET additionally suffered from a lack of robustness (Allison *et al.*, 2006; Tintle *et al.*, 2008, 2009b). The goal of Tintle *et al.*'s (2009b) investigation was the confirmation of these results from gene expression in the context of GWAS. Tintle *et al.* (2009b) initially did not use the NARAC data but data from the Framingham Heart Study (FHS) that included original genome-wide FHS data (Cupples *et al.*, 2009) and simulated data based on FHS (Kraja *et al.*, 2009). The Framingham Heart Study is a family-based, observational, longitudinal study for the investigation of risk factors in cardiovascular diseases ongoing since 1948.

To enhance the comparability of these methods for gene set incorporation, we performed follow-up analyses for the GSA methods and hierarchical models. We applied SUMSTAT, SUMSQ, GSEA and FET to the RA data using three different permutation methods – SNP, gene and phenotype permutation. We incorporated 825 gene sets from the GO biological processes (Harris *et al.*, 2004). The analysis was repeated with and without HLA region. To allow a closer comparison of HBP and LRP, we repeated our HBP analysis by refitting the hierarchical model using the same gene set information as Lebrec *et al.* (2009), using gene level statistics (maximum statistic per gene) and excluding the SNP information.

Table 5.4: *Characteristics of the GAW16 contributions integrating gene set information in GWAS*

	Ballard *et al.* (2009)	Lebrec *et al.* (2009)	Sohns *et al.* (2009)	Tintle *et al.* (2009b)
Data	NARAC: unrelated cases and controls	NARAC: unrelated cases and controls	NARAC: unrelated cases and controls	FHS: unrelated cases and controls real and simulated data
SNP to gene assignment	NCBI Refseq: intragenic SNPs	gene coding region +/-500kb	Illumina Annotations: gene coding region +/-500kb	Ensemble database: gene coding region +/-500kb
Measure of association and gene level statistic	p-value for multiple regression of all SNPs per gene	Maximum SNP statistic of trend test; single SNP regression coefficient	Maximum SNP statistic of trend test	Maximum SNP statistic of allelic test
Pathway source	564 pathways from KEGG, GenMAPP and Biocarta (GenGen)	27 GO biological processes involved in RA (MsigDB)	100 pathways from KEGG, GO and Biocarta (GenGen)	702 positional cytogenetic band and molecular function sets (MsigDB); self-defined gene sets for simulated data
Pathway based analysis method	Binomial test, random set method	Hierarchical modeling framework (gene level)	GSEA, HBP	GSEA FET SUMSQ SUMSTAT
Significance Assessment	Binomial test: p-value thresholds 0.01, 0.1 and 0.2 Random set test: permutation of gene statistics FDR: phenotype permutation procedure	Forward stepwise regression procedure using Akaike information criterion	phenotype permutation procedure FDR: phenotype permutation method as described in section 5.3.4	FET: q-value thresholds 5.992, 9.210, 13.816 and 18.421 other methods: permutation of gene statistics

5.5.2 Results

In the following we will summarize the main results from the individual analyses of Tintle *et al.* (2009b), Ballard *et al.* (2009) and Lebrec *et al.* (2009), before we will go into detail of our joint group work results. Some more information about the individual analysis results can be found in the corresponding publications.
Ballard *et al.* (2009) observed for both his methods consistent top-scoring gene sets dominated by HLA genes and related to immune response. The random set method identified more significant pathways than the binomial approach. When HLA was excluded, the number of significant gene sets for the random set method increased, still involving many immune-related gene sets, while the binomial test identified only a low number of sets.
In Tintle *et al.*'s (2009b) analysis of the simulated FHS data, SUMSTAT proved to be most powerful, with all methods controlling the type I error rate. The results for the original FHS data can be found in figure 5.8, with SUMSTAT identifying nearly all sets found by SUMSQ, GSEA and FET as well. The results lead to the same conclusion for GWAS as Efron and Tibshirani (2006) and Tintle *et al.* (2008) already stated for gene expression, that SUMSTAT appears to provide the most powerful and robust results.
Using the forward-stepwise regression procedure, Lebrec *et al.* (2009) selected 8 out of the 27 gene sets as relevant for RA. The gene set information is overruled by the strong HLA GWAS results and HLA genes still remained at the top of the list using the standardized posterior gene effects for ranking. The top 1% top ranked genes of the tuned and unturned LRP version shared 17 genes. Four of these genes were confirmed to be associated with RA (Harney *et al.*, 2008; Raychaudhuri *et al.*, 2008). Assuming that 100 among approximately 20,000 human genes have already been identified for RA, the probability to detect at least 4 of those 100 genes within 17 genes drawn at random is only 10^{-6}. Therefore, the list of these 17 genes seems not to be random. Since these 4 genes were not all prioritized based on the GWAS data alone, this indicates the ability of the hierarchical model to contribute new candidate genes.

In our joint group work we performed, the following results were obtained.
Regarding the comparison of the different permutation strategies in the follow-up analysis, we contrast the respective numbers of significant gene sets in table 5.5. For SUMSQ and SUMSTAT using phenotype permutation around 80% of 825 considered gene sets showed significance. It is highly implausible that all these gene sets are really involved in RA. However by highlighting everything, nothing is illuminated and hence the results of the analysis are not useful to prioritize genes for follow-up. When gene permutations were used, the number of significant sets was clearly reduced. In general, gene permutation combined with SUMSQ, SUMSTAT and GSEA found more significant sets when HLA is excluded than included. The GSEA with SNP permutation including HLA detected only a very small number of gene sets, pinpointing that this is not very sensitive. While the GSEA with phenotype permutation detects many more gene sets including HLA than excluding this region, the reverse is true for gene permutations. Comparing the results of SUMSQ and SUMSTAT with each other overlaps more than GSEA compared to one of these.
FET finds decreasing numbers of significant sets with increasing cutoff when gene permutations are used. This corresponds to the results observed by Tintle *et al.* (2008)

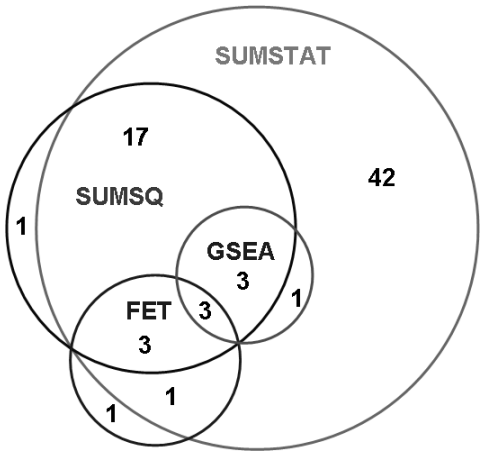


Figure 5.8: *Venn diagram of the results in Tintle et al. (2009b) comparing the significant sets of four different gene set analysis methods using the Framingham Heart study data. (adopted from Tintle et al. (2009b))*

before. However, this trend reverses with phenotype permutations. FET is fairly robust to whether or not HLA region is included.
The 27 selected pathways of Lebrec *et al.* (2009) all occurred in the GSEA of our follow-up analyses. When HLA was included and phenotype permutation was used, one of these pathways (defense response) had a FWER and FDR$<$ 0.05. Using SNP permutations, this pathway reached nominal significance only. Using gene permutations, not even this pathway was observed. Excluding HLA, gene and phenotype permutations showed no significance for this pathway, however for SNP permutation nominal significance was reached again.
Out of these 27 gene sets from Gene Ontology selected by Lebrec *et al.* (2009), only four were found in our original gene set collection, of which only one was in our selected 100 used pathways. However, this pathway (cell adhesion pathway) occurred on rank 11 of our 1-stage HBP when HLA was included, with an OR of 1.21. For the 2-stage HBP, a negative logistic regression coefficient occurred, so that the pathway had reduced prior probability of association. In our original GSEA analyses, the pathway reached no significance.
When comparing the LRP results of Lebrec *et al.* (2009) and our refitted HBP model including the same 8 gene sets selected in Lebrec *et al.*'s (2009) forward-stepwise regression procedure as associated with RA, the gene set coefficients of the LRP method showed a high correlation to the coefficients of the linear HBP submodel for the strength of association.
Comparing the top 200 genes between our HBP and the untuned LRP version of Lebrec *et al.* (2009), we have an overlap of 153 genes. Of this overlap, 102 genes are from the HLA region and 139 were also included in the initial top 200 genes. 14 non-HLA genes were newly identified by both methods. Comparing the top 200 genes of the tuned LRP version to the initial top 200 genes, only 5 genes from HLA region and one non-HLA gene were in common. These six genes were identified by HBP as well. Furthermore, HBP and tuned LRP shared 18 more genes - all from the non-HLA region. Of the 17 genes identified by the tuned and untuned LRP (page 107), all were detected by the

Table 5.5: *Number of significant gene sets identified by different GSA methods using different permutation strategies considering 825 gene sets from Gene Ontology.*

Method	Permutation strategy	Including HLA	Excluding HLA
GSEA	SNP	6	90
	Gene	38	81
	Pheno	87	26
FET	Gene	98	94
(test statistic cutoff 5.992)	Pheno	22	19
FET	Gene	8	8
(test statistic cutoff 18.421)	Pheno	203	121
SUMSQ	Gene	38	60
	Pheno	673	645
SUMSTAT	Gene	39	107
	Pheno	704	680

HBP gene level version as well.
By using the HBP on the SNP level, the consistency between the gene set parameter estimates disappears and the overlap of the top 200 genes with the initial ranking increases, while the number of newly identified genes by HBP and the two different LRP versions decreases to only 1 gene in each case (*RND3*).
The same analyses and comparisons were conducted, excluding all genes from the HLA region. The gene set coefficients between the two methods remained similar. The overlap between the two versions of Lebrec *et al.* (2009) comprises 24 genes. Surprisingly, the overlapping set results between the LRP and HBP described above reversed. The overlap of the HBP with the untuned Lebrec *et al.* (2009) version includes only 19 genes (in particular CD40), while it shared 161 genes with the tuned version. *CDKN1A* is the only gene from the initial top 200 genes (including 88 HLA-region genes) that occurs in both the 200 top genes of Lebrec *et al.* (2009) and Sohns *et al.* (2009) as well. Considering the HBP based on SNPs yields 98 genes in common with the untuned LRP version and only 2 genes in common with the tuned LRP version (*CDKN1A*, *RND3*).

5.6 Discussion

The incorporation of gene set information into the analysis of GWAS offers an attractive approach to marry the pathway-driven candidate and exploratory genome-wide association method (Thomas, 2006). GWAS aim to identify new genetic variants predisposing to diseases. However, for complex diseases this represents a challenge due to the multiplicity of genetic factors with only weak effects, locus heterogeneity, interactions and other complicating factors. Knowledge about biological processes and interrelation of genes as prior information in the analysis cannot replace the original GWAS analysis (Sohns *et al.*, 2009), but may lead to a better understanding of the

underlying biology (Ballard *et al.*, 2009). The methods increase signals of markers with weak effects that jointly occur within one gene sets, resulting in supporting, complementing and completing results (Sohns *et al.*, 2009). Locus heterogeneity within one pathway can be considered and the replication problem can be reduced (Sohns *et al.*, 2009). It may help to structure results, distinguish truly associated from false positive results and facilitate their interpretation (Chasman, 2008). Existing hypotheses may be substantiated and new hypotheses provided.
A variety of methods to increase the GWAS power by gene set information have been proposed (Tintle *et al.*, 2009a; Wang *et al.*, 2010, 2011). Two strategies are GSA methods transferred from gene expression research to GWAS by Wang *et al.* (2007) and (Chasman, 2008) or supporting GWAS data within hierarchical models to prioritize candidate genes for further studies (Lebrec *et al.*, 2009; Lewinger *et al.*, 2007; Sohns *et al.*, 2009).

Hierarchical models

Our main interest are the empirical Bayes hierarchical models as simple and flexible alternative framework to integrate external information in GWAS, particularly gene set information. These methods to prioritize SNPs and genes supported by gene set information result in a mix of known and novel genes identified as significant (Lebrec *et al.*, 2009; Sohns *et al.*, 2009; Tintle *et al.*, 2009a). Both hierarchical models presented in this chapter (Lebrec *et al.*, 2009; Sohns *et al.*, 2009) allow a natural incorporation of multiple prior gene set effects with GWAS. The weight of the GWAS data relative to the prior gene set information may be varied (Lebrec *et al.*, 2009).
For convenience Lebrec *et al.* (2009) fitted his model into a Gaussian framework. Therefore, a few approximations were necessary. Although the positive allelic effect of the maximal SNP representing a gene has a skewed distribution departing from normality (section 4.4.1), a normal distribution is assumed. The variance estimates $\hat{\sigma}^2_{M_i(G_j)}$, where M_i the SNP with highest test statistic in gene G_j, are used to present the gene variances. Although these variances are closely related, their exact relationship is not clear. The robustness of the forward stepwise fitting procedure to the normal assumption is still unknown. Initially, the between gene variance τ^2 was assumed to be equal over all genes and estimated from the data ("untuned version"). Alternatively, Lebrec *et al.* (2009) pre-specified gene specific values $\tau^2_{G_j}$ to increase the contribution of the pathway information to the final results ("tuned version"). In the context of linkage and RA, the gene set information accounts for 50% of the variation between the genes Lebrec *et al.* (2009). In GWAS however, only 1% could be explained (Lebrec *et al.*, 2009; Tintle *et al.*, 2009a).
Lewinger *et al.* (2007) adapted his hierarchical model well to the data – accepting a more complicated form of the model than Lebrec *et al.* (2009). He considered the fact that the majority of the SNPs have no effect on disease susceptibility by a mixture of a prior point mass concentrated on zero for the unassociated SNPs and a non-central χ-distribution for the remaining SNP effects. That the latter is adequately modeled was assessed in extensive simulation studies. In addition, not only the SNP effect is influenced by the external information but also the prior probability that a SNP is associated. However, Lewinger *et al.* (2007) assumed one unique variance for the SNPs in the prior distribution as well. As Lebrec *et al.* (2009), we did not consider LD

between the SNPs in the model but ignored the correlation. In HBP we considered all markers of a gene to exploit all available data, while Lebrec *et al.* (2009) chose to use the gene-level summaries instead due to computational limitations. Working on the SNP level may penalize large genes since a high number of unassociated markers within a gene may dilute true positive results (Sohns *et al.*, 2009). By the more simplistic model using the maximum association per gene, a lot of information may be discarded and large genes are favored. Lebrec *et al.* (2009) included gene size as a predictor in the marginal model and the modification left the selected model and top ranking nearly unaltered.
Although LRP and HBP translate the same idea in very different models, in the follow-up analysis presented we see large consistencies in their regression coefficients ("untuned" LRP) and hence pathway based SNP supporting. However, due to many and extremely high association signals in our data, the pathway information has relatively small influence on the results and the top 100 gene lists are almost identical to the original one. In Lewinger *et al.*'s (2007) paper proposing the HBP for integrating external information in GWAS, he stated that HBP "can be superior when the proportion of true positive associations is not too small, as in GWAS with hundreds of truly associated SNPs". However, "when the non-centrality parameters of the true associations are large enough to be picked by the raw test statistics there is little to be gained from prior covariates" Lewinger *et al.* (2007). Hence, it is not surprising that the list of top genes did not change for HBP, since this GWAS of RA has several hundred genome-wide significant SNPs assigned to nearly 100 different genes.
When we excluded the HLA region from the analysis, the results of the "untuned" LRP and HBP diverged extremely. However, this time we observed a high similarity of the HBP results to the results from the "tuned" LRP. This indicates that in HBP the pathway information anyhow influences the results more strongly than in LRP.
While we used the whole wealth of pathways available in our initial analysis, Lebrec *et al.* (2009) concentrated on a small number of expected pathways. Both strategies have their advantages and disadvantages. Lebrec *et al.* (2009) argues, that in the hierarchical Bayes context the risk of over-fitting is high when a large number of gene sets is considered and one should restrict to a set of initial candidate pathways to limit the number of sets. However, this prohibits the identification of totally new and unexpected insights to the development of a disease.

Gene set analysis methods

For our individual analysis we chose GSEA with phenotype permutations as the gene set method to compare our HBP results to. This method was proposed in the context of GWAS shortly before our investigations (Wang *et al.* (2007)). The decision for the phenotype permutation procedure was motivated by the ability to correct for several kinds of bias. In a comparison of Chasman (2008) based on GWAS, it was shown that gene sets containing a few highly significant genes are rather detected by FET, while GSEA has more power to identify sets involving a high number of weakly associated genes.
However, in the context of gene expression, GSA method comparisons had shown that GSEA and FET are both less powerful then other GSA methods (Dinu *et al.*, 2007; Efron and Tibshirani, 2006; Tintle *et al.*, 2008) and lack robustness (Allison

et al., 2006; Tintle *et al.*, 2008). In the GSA analyses for GAW16, these results were confirmed for GWAS data by Ballard *et al.* (2009) using a multilocus method to obtain a gene level statistic, as well as by Tintle *et al.* (2009b) using the maximum SNP statistic to represent a gene. Self-contained GSR methods such as SUMSTAT, SUMSQ (Tintle *et al.*, 2009b) and Ballard's random set method (Ballard *et al.*, 2009) identified more significant gene sets than FET and its binomial approximation (Ballard *et al.*, 2009; Tintle *et al.*, 2009b). This is expected, since several genes below the significance threshold are necessary for an over-representation method to identify a gene set, while sets with even one very significant gene or many genes slightly above a specified cutoff may be identified by self contained GSR as well. SUMSTAT was more powerful than SUMSQ. The analyses provide evidence that GSEA and over-representation methods for gene set analysis in GWAS are not optimal and SUMSTAT or a similar self-contained GSR method should be used instead. Based on simulated data, Tintle *et al.* (2009b) showed that the type I error was controlled by all methods.
Excluding the HLA genes from the analysis, the number of gene sets found as significant by the random set method, SUMSTAT, SUMSQ and GSEA using a gene permutation procedure increased, while the number for the over-representation tests decreased. This reflects the strategy of significance assessment. The removal of the highly significant genes reduced the number of genes below a cutoff and hence the probability that a pathway will involve such genes. This in turn leads to the identification of less gene sets using an ORA method. Although SUMSTAT, SUMSQ and random set statistics are self-contained by simply summing up the values for the single genes involved in a gene set, the methods become automatically competitive by the gene permutation procedure. The null distribution is generated based on statistics of genes not in the gene set. By excluding the numerous highly significant genes from the analysis, there is less competition for the gene sets containing only one or two significant genes compared to random gene sets of the same size. This leads to a higher number of significantly classified genes sets (Ballard *et al.*, 2009).
Phenotype permutations on the other hand are not reasonable in combination with SUMSTAT and SUMSQ, although they generate the correct self-contained null distribution of "no genes associated with the disease" corresponding to the self-contained statistics. Even one associated gene per set makes it significant resulting in the practical problem that too many things are found. In the follow-up analysis, nearly all gene sets reached significance using this combination. The problem has been documented before by Efron and Tibshirani (2006). Although the combination of self-contained method with competitive permutation procedure may lead to bias, our group work, (Wang *et al.*, 2007) as well as others (Tintle *et al.*, 2008) have shown that assessing significance with random gene sets provides reasonable results. For FET and GSEA we do not see the same problems with phenotype permutations, since they are competitive by nature, comparing the genes in a set to the gene complement. However, for GSEA in our original analysis, we observed nominal p-values of < 0.05 for all involved gene sets. Since all 100 sets were selected so that they involve at least one highly significant gene, the set significance in comparison to the complement genes is not surprising. However, for the FDR calculation, the enrichment score is compared to the other gene set scores as well, so that only pathways more enriched than the others become significant.
Although phenotype permutations have the main advantage to correct for different

types of bias, our group work has shown that this is not necessarily the best choice anyway. Depending on which GSA method is chosen, gene permutations may be the better choice. Nevertheless, the decision for the phenotype permutation for GSEA in our individual work was justified by our group results comparing the different permutation strategies. However, overall, the best choice of GSA method seems to be SUMSTAT with gene permutations.

Gene set analysis methods versus Hierarchical Bayes models

Due to the different ideologies and methodologies underlying gene set analysis methods and hierarchical models involving gene set information, none of them is a gold standard for the integration of gene set information into GWAS (Sohns *et al.*, 2009). While GSA methods use the gene ranking to find enriched gene sets e.g. by cumulating the statistics of the involved genes, hierarchical models use the prior gene set information to re-rank SNPs or genes (Ballard *et al.*, 2009; Sohns *et al.*, 2009). Hence, GSA methods lead directly to the identification of whole gene sets while hierarchical models obtain new promising genes or SNPs. However, by the LES of GSEA for example, a list of top genes can be obtained and by the regression coefficients corresponding to the different gene sets hierarchical models can lead to a list of top gene sets (Sohns *et al.*, 2009; Lebrec *et al.*, 2009).
The main advantage of the hierarchical Bayes models compared to the GSA methods is that they can integrate not only the gene set information but also additional other types of prior knowledge, e.g. SNP location, function or previous results. Hierarchical models are therefore more flexible. Furthermore, all SNPs can be considered, while GSA involves only SNPs within or close to a gene and only genes within the examined gene sets. In HBP the SNPs excluded in GSA are at least modeled as one "remaining group" and may be grouped by other additional external information (SNP information), so that they still have the chance to stand out in the re-ranking.
In GSA, it is possible to correct for different kinds of bias due to different gene length, gene set sizes or correlations by permutation methods. In hierarchical models, these issues may be directly considered in the model, e.g. by involving LD structure in the regression model or by corresponding weights. However, since the model was not extensively studied before, especially in the context of pathway analysis, we concentrated on the simplest case, considering no such correction.
The special challenge in this Rheumatoid Arthritis data set was to contrast genes within the HLA region that play a predominant role in this disease and result in a large number of highly significant associations, but also identify new non-HLA susceptibility genes (Sohns *et al.*, 2009). Comparing GSEA and HBP in that particular context, the GSEA strategies have shown to be superior to the HBP. While all four strategies of Sohns *et al.* (2009) identified the well known association of HLA in RA, only the GSEA methods were successful to enrich the top gene list with non-HLA genes as well. HBP only keeps the prominent role of the HLA complex (Sohns *et al.*, 2009). This confirms the statement of Lewinger *et al.* (2007) about the HBP, that when "true associations are large enough to be picked by the raw statistics, there is little to be gained from prior covariates, even highly informative ones". Although this specific characteristic of the HBP applied to the analyzed RA data set, GWAS normally show only small effects with often no genome-wide significant SNPs. Hence, our results are not generally applicable.

Comparing the results of the four different strategies, we found considerable differences in the most promising genes and gene sets identified. The chance that a gene appears in more than one of our top 100 gene lists or that a gene set appears in more than one of our top 20 gene set lists is only 50%. Although the HBP and HBP+HBP had nearly identical top 100 genes, their list of top 20 gene sets had only 3 sets in common. In general, the high number of top genes after the pathway analysis located in the HLA region and top pathways involved in immune response, inflammation or related other theories with respect to RA was not surprising and biological plausible.
All in all, incorporating gene set information in the analysis of GWAS has demonstrated to be a promising and useful approach (Ballard *et al.*, 2009). The techniques validate prior knowledge and produce new gene and gene set candidates not captured by the single-SNPS analyses (Ballard *et al.*, 2009; Tintle *et al.*, 2009a). The prioritization methods in form of hierarchical models have the main advantage to integrate other additional external information simultaneously to the biological pathways. Due to the different rational underlying gene set analysis methods and hierarchical Bayes approaches to integrate pathway information, the choice of methods depends on the data, study aim and observed single SNP results. In the particular application of the pathway integrating methods to the Rheumatoid Arthritis data, the hierarchical models and GSA methods were able to validate the single-SNP results. Due to the high initial effects, new candidates were only found by GSA methods. Therefore, GSA methods should be preferred in that particular context.
Since we have seen in our group comparisons that the chosen method may have a large impact on the results, further work in that area is still necessary to evaluate their relative usefulness for pathways identification and gene prioritization (Lebrec *et al.*, 2009). This comprises more comprehensive theoretical or simulation analysis to validate the robustness of the different pathway analysis techniques and their ability to fulfill the initial intention to provide increased power to find consistent but weak effect (Tintle *et al.*, 2009a). So far it is still unclear if newly identified pathways, genes or SNPs are good new candidates or only false positive results. In addition, potential modifications may help to further maximize power in GWAS (Tintle *et al.*, 2009a). The incidental issues discussed in section 5.3 are part of further research as well.

6 The empirical hierarchical Bayes approach for gene x environment interaction

6.1 Motivation

Since biological pathways comprise gene products as well as environmental substrates that contribute to the human body functions, the important role of interactions between genetic and environmental factors in the etiology of complex disease is indisputable (section 2.4). Hence, the analysis of GxE interactions gains attraction as a good complement to simple single marker analyses to improve GWAS results. Unfortunately, the detection of GxE still leaves much to be desired, as the classical case-control test outlined in section 3.1.5 has insufficient power to detect the interactions and therefore requires sample sizes of several thousands.
During the last years, several alternative GxE methods were proposed, trying to increase the power for the detection of interactions, partly coming across other problems. An important requirement for a GxE test is that interactions are clearly differentiated from G-E associations on a population level (sections 2.4.3,3.1.5). An optimal solution has not been found yet.
Based on an idea of Volk *et al.* presented in a conference poster in 2007 (IGES 2007, abstract Volk *et al.* (2007) unfortunately does not contain main idea), we developed a new promising approach for the analysis of GxE interactions. This test adopts the hierarchical model of Lewinger *et al.*'s (2007) hierarchical Bayes prioritization (section 4.4) for the purpose of GxE interaction. The approach uses advantages from different other methods and thereby reaches high power combined with the strict separation of GxE interaction effects from population based G-E associations.
Before we will present our new method in section 6.3, the following section demonstrates different alternative tests for GxE interactions established so far. We compared our new approach to these other methods in comprehensive simulation studies. The simulation set-ups and the corresponding results are given in section 6.4. We will end with a final discussion in section 6.5.

6.2 Methods for GxE interaction analysis

In the following we restrict to an unmatched case-control study with a binary environmental exposure E and a binary genetic factor G as given in table 3.3.
The coefficients from logistic regression models for a GxE interaction, a G-E association within cases and a G-E association within controls are given by β_{ge}, β_{cases} and β_{controls} (equations 3.12, 3.14 and 3.15), with MLE estimates $\hat{\beta}_{\text{ge}}$, $\hat{\beta}_{\text{cases}}$ and $\hat{\beta}_{\text{controls}}$ (equations 3.17 and 3.18). The corresponding variance estimates are given by $\hat{\sigma}^2_{\text{ge}}$, $\hat{\sigma}^2_{\text{cases}}$ and $\hat{\sigma}^2_{\text{controls}}$ (equations 3.20).
Recall that the test statistic of the classical case-control test for gene x environment interactions is given by

$$T_{\text{cc}} = \frac{\hat{\beta}_{\text{ge}}}{\hat{\sigma}_{\text{ge}}} = \frac{\hat{\beta}_{\text{cases}} - \hat{\beta}_{\text{controls}}}{\sqrt{\hat{\sigma}^2_{\text{cases}} + \hat{\sigma}^2_{\text{controls}}}}.$$

T_{cc} follows an approximate normal distribution $N(\beta_{ge}, 1)$, with $\beta_{ge} = 0$ under the null hypothesis of no interaction.

6.2.1 The case-only test

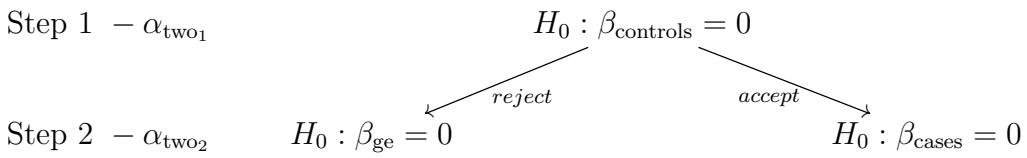
In 1994, Piegorsch *et al.* proposed the case-only approach for GxE interaction analysis to reach higher power than the traditional case-control test. This method benefits from a reduced variance by limiting to β_{cases} instead of testing $\beta_{ge} = \beta_{cases} - \beta_{controls}$. However, this test is based on two assumptions: the assumption of gene-environment independence on the population level (section 3.1.5) and the rare disease assumption. When these assumptions are fulfilled, $\beta_{controls}$, that represents the association between G and E among the disease-free subjects, reduces to 0 and can be neglected. Hence, β_{ge} can be unbiasedly estimated by the association between G and E among the cases alone (Piegorsch *et al.*, 1994).
The case-only test statistic is given by

$$T_{cases} = \frac{\hat{\beta}_{cases}}{\hat{\sigma}_{cases}}.$$

It is approximately standard normally distributed under the null hypothesis of no interaction and the assumptions stated above.
The case-only design can reach much more power than the case-control design. However, when the assumptions are not full filled, the method leads to bias and false positive results. The presence of even a small population based G-E association greatly inflates the type I error in the case-only test and makes it not recommendable. Especially in the genome-wide context, population based associations cannot be ruled out, quite the contrary, they are even expected to occur.

6.2.2 An intuitive two-step method

While low power is the weakness of the case-control method that holds type I error level, population based G-E dependencies cause an inflated type I error when the more powerful case-only method is used. Hence, an intuitive strategy is to test for G-E independence first and based on this result to decide if in a second step case-only or case-control method is used for single SNP GxE testing (Albert *et al.*, 2001).
Given a rare disease, $\beta_{controls}$ measuring the G-E association in the control population can be used as a representative for the population based G-E association. When genetic and environmental factor are independent from each other, we have $\beta_{controls} = 0$, given a population based G-E association $\beta_{controls} \neq 0$.
Hence, we can use this G-E association among controls in step 1 to test the null hypothesis $H_0 : \beta_{controls} = 0$ with a significance level α_{two_1}. The procedure in step 2 depends on the result of step 1. When the null hypothesis is rejected, a G-E dependency is likely and the case-control statistic will be used to test for an interaction to avoid false positive results. If one fails to reject the null hypothesis of G-E independence, the case-only test is used in step 2, because a dependency between environmental factor and genotype could not be shown (Albert *et al.*, 2001).



This two-step test statistic can be expressed as

$$T_{\text{two}} = T_{\text{cc}} I[\mid T_{\text{controls}} \mid > T_{\alpha_{\text{two}_1}/2}] + T_{\text{cases}} I[\mid T_{\text{controls}} \mid \leq T_{\alpha_{\text{two}_1}/2}]$$

with $T_{\alpha_{\text{two}_1}/2}$ percentile of the standard normal distribution and I[A] the indicator function if A holds and zero otherwise.
A problem of this test is the correlation of T_{controls} and T_{cc}, and hence a correlation between step 1 and step 2 for these SNPs where H_0 of step 1 has to be rejected. The pretesting is ignored when significance is assessed in the second step. This may lead to an inflated type I error rate for the overall procedure.

6.2.3 Murcray's two-step approach

An alternative two-step approach to scan for interactions with a simple concept was developed by Murcray *et al.* in 2009. This test combines the power of the case-only test with the protection from bias of the case-control test in the two-step procedure with independent test statistics.
The first step is again a screening for associations between G and E, but not based on controls only, but on the combined sample of cases and controls. In this study sample G-E association can be measured by the logistic regression model

$$\text{logit}P(E = 1 \mid g) = log\left(\frac{P(E = 1 \mid g)}{P(E = 0 \mid g)}\right) = \alpha_{\text{all}} + \beta_{\text{all}} G \quad .$$

As seen in section 3.1.5 for the other regression coefficients, the coefficient equals the logarithm of the corresponding odds ratio OR_{all}, that is given by

$$OR_{\text{all}} = \frac{(p_{000} + p_{100})(p_{011} + p_{111})}{(p_{001} + p_{101})(p_{010} + p_{110})} \quad .$$

The maximum likelihood estimator of β_{all} is determined by

$$\hat{\beta}_{\text{all}} = log\left(\frac{(n_{000} + n_{100})(n_{011} + n_{111})}{(n_{001} + n_{101})(n_{010} + n_{110})}\right) \quad .$$

It is approximately normally distributed

$$\hat{\beta}_{\text{all}} \sim N(\beta_{\text{all}}, \sigma^2_{\text{all}}) \quad ,$$

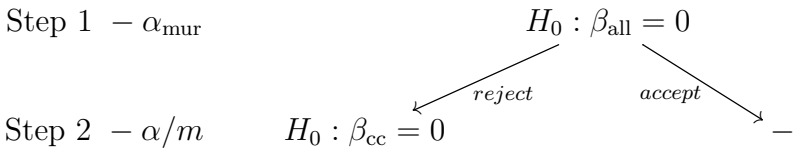
with estimated variance

$$\hat{\sigma}^2_{\text{all}} = \sum_{g=0,1} \sum_{e=0,1} \frac{1}{(n_{0ge} + n_{1ge})} \quad . \qquad .$$

Hence, the test statistic is

$$T_{\text{mur}} = \frac{\hat{\beta}_{\text{all}}}{\hat{\sigma}_{\text{all}}} \quad .$$

T_{mur} is again approximately standard normally distributed when no interaction exists. Only the subset of m SNPs exceeding a given significance threshold α_{mur} in step 1 is selected for step 2. A G-E association in the whole study sample can result from a dependency between G-E not related to the analyzed disease or an underlying GxE interaction with an effect on case-control status. To distinguish both kinds of selected SNPs, in the second step they are further analyzed by the traditional GxE interaction test with $H_0 : \beta_{\text{ge}} = 0$. Therefore, a less stringent significance level adapted to the size of the selected subset $\frac{\alpha}{m}$ is used. Murcray's two-step statistic



can be written as

$$T_{\text{mur}} = T_{\text{cc}} I[\mid T_{\text{all}} \mid > T_{\alpha_{\text{mur}}/2}].$$

The screening for G-E association in the entire study sample in step 1 eliminates the correlation between step 1 and step 2 test statistics and therefore preserves the overall type I error rate. Although step 1 is sensitive to G-E independence assumption in the population, step 2 is not and hence the overall two-step procedure provides a valid test in the presence of population-level associations between genotype and exposure. This test is more powerful than the one-step case-control test and robust in the presence of G-E dependencies.
A disadvantage of the Murcray approach is that the first step depends on the ratio of cases to controls in the sample. A higher number of controls than cases leads to a decrease of the power in step 1 and hence to a loss of power for the overall method. The choice of the step 1 significance level has a high influence on the results. Therefore, it should be chosen carefully. However, the best choice highly depends on different characteristics of the sample to analyzed and is not that clear (Mukherjee *et al.*, 2012).

6.2.4 Mukherjee's shrinkage estimator

The method for detecting GxE interactions proposed by Mukherjee and Chatterjee in 2008 is inspired by the idea of an empirical Bayes model and combines the robust case-control estimator $\hat{\beta}_{\text{ge}}$ and the powerful estimator $\hat{\beta}_{\text{cases}}$ by their weighted average

$$\hat{\beta}_{\text{muk}} = (1 - B)\hat{\beta}_{\text{cases}} + B\hat{\beta}_{\text{ge}} \tag{6.1}$$

The weight B is chosen, so that when evidence for an underlying G-E independence in the control population is given by the data $B \to 0$ and hence $\hat{\beta}_{\text{muk}} \to \hat{\beta}_{\text{cases}}$. When the

G-E independence assumption is violated, the estimator $\hat{\beta}_{\text{muk}}$ should converge to the unbiased estimator $\hat{\beta}_{\text{ge}}$ resulting from $B \to 1$.
To obtain a shrinkage factor B of that characteristic, we assume that we cannot rule out G-E dependence and provide a Bayesian framework for $\hat{\beta}_{\text{controls}}$, with the first level model

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

and the prior

$$\beta_{\text{controls}} \sim N(0, \tau^2) \quad .$$

The same form of that model was given in example 1a on page 57 and for the linear regression on pathways (LRP) method of Lebrec *et al.* (2009) (section 4.3). In the particular context here, the hyperparameter τ^2 is a quantity for the uncertainty about the G-E independence. An estimate for the asymptotic variance in the first level model is given in formula 3.20.
We can derive the marginal distribution

$$\hat{\beta}_{\text{controls}} \sim N(0, v^2), \qquad v^2 = \sigma^2_{\text{controls}} + \tau^2 \quad .$$

Based on that marginal variance, $\tau^2 = \max(0, v^2 - \sigma^2_{\text{controls}})$ (equation 4.12) and a consistent estimator of the unknown prior variance can be obtained by $\hat{\tau}^2 = \max(0, \hat{\beta}^2_{\text{controls}} - \hat{\sigma}^2_{\text{controls}})$ (Morris, 1983; Greenland, 1993). For convenience regarding the variance estimator of the Mukherjee statistic, $\hat{\tau}^2 = \hat{\beta}_{\text{controls}}$ was used instead, although it is more conservative. However, simulation studies showed that the usage of this estimate does not reduce efficiency (Mukherjee and Chatterjee, 2008).
Using the hyperparameter estimate, our weight B is set to

$$B(\hat{\tau}^2, \hat{\sigma}^2_{\text{cc}}) = \frac{\hat{\tau}^2}{\hat{\tau}^2 + \hat{\sigma}_{\text{cc}}} \quad . \tag{6.2}$$

This results in the final estimator

$$\hat{\beta}_{\text{muk}} = \frac{\hat{\sigma}^2_{\text{cc}}}{\hat{\tau}^2 + \hat{\sigma}^2_{\text{cc}}} \hat{\beta}_{\text{cases}} + \frac{\hat{\tau}^2}{\hat{\tau}^2 + \hat{\sigma}^2_{\text{cc}}} \hat{\beta}_{\text{cc}} \quad .$$

When $\hat{\beta}_{\text{controls}}$ is close to 0, indicating no G-E association, we have more weight on $\hat{\beta}_{\text{cases}}$; with increasing $\hat{\beta}_{\text{controls}}$ the estimator is shrunk towards $\hat{\beta}_{\text{ge}}$.
We can rewrite this new estimator by

$$\hat{\beta}_{\text{muk}} = \hat{\beta}_{\text{cases}} - \frac{\hat{\tau}^2}{\hat{\tau}^2 + \hat{\sigma}_{\text{cc}}} \hat{\beta}_{\text{controls}} \quad . \tag{6.3}$$

From this second perspective, $\hat{\beta}_{\text{controls}}$ is not completely subtracted from $\hat{\beta}_{\text{cases}}$ as for the traditional case-control approach, but only partly, with the G-E association part within controls shrunk to zero when G-E independence is given.
Although Mukherjee and Chatterjee's (2008) estimator is constructed from a Bayesian perspective, it is neither true Bayesian nor empirical Bayesian, but purely a simple

function of the observed data (Mukherjee and Chatterjee, 2008).
To construct a test statistic for testing the null hypothesis $H_0 : \beta_{\text{muk}} = 0$, an asymptotic variance has to be calculated. For the sake of simplicity, the variation in $\hat{\sigma}^2_{\text{ge}}$ is ignored and treated as constant, since $\hat{\sigma}^2_{\text{ge}} \to 0$ at the rate of O(1/n).
Based on formula (6.3), the first part of the estimator depends only on the cases and the second one only on the controls, so that both parts can be considered independent from each other. For the first term, the variance is simply given by $\hat{\sigma}^2_{\text{cases}}$. The second term can be viewed as a function of $\hat{\tau} = \hat{\beta}_{\text{controls}}$, and using the delta method we obtain

$$\hat{\sigma}^2_{\text{muk}} \approx \hat{\sigma}^2_{\text{cases}} + \left(\frac{\hat{\beta}^2_{\text{controls}}(\hat{\beta}^2_{\text{controls}} + 3\hat{\sigma}^2_{\text{cc}})}{(\hat{\sigma}^2_{\text{cc}} + \hat{\beta}^2_{\text{controls}})^2} \right)^2 \hat{\sigma}^2_{\text{cc}}. \tag{6.4}$$

Using this approximate estimator, we can construct the Wald test statistic

$$T_{\text{muk}} = \frac{\hat{\beta}_{\text{muk}}}{\hat{\sigma}_{\text{muk}}}. \tag{6.5}$$

Simulation studies showed that this variance approximation works fairly well, even for smaller sample sizes of 100 cases and 100 controls (Mukherjee and Chatterjee, 2008).

In Mukherjee *et al.* (2008), a slightly modified version of this test statistic was proposed, by subtracting the posterior estimator of $\hat{\beta}_{\text{controls}}$ from $\hat{\beta}_{\text{cases}}$ instead of the maximum likelihood estimator $\hat{\beta_{\text{controls}}}$. Starting from the Bayesian framework stated above, we can calculate the posterior distribution (see also 4.3):

$$\beta_{\text{controls}} \mid \hat{\beta}_{\text{controls}} \sim N\left(\frac{\tau^2}{\tau^2 + \sigma^2_{\text{controls}}}\hat{\beta}_{\text{controls}}, \frac{\tau^2\sigma^2_{\text{controls}}}{\tau^2 + \sigma^2_{\text{controls}}}\right).$$

Using the estimate of the posterior expectation, we obtain

$$\hat{\beta}_{\text{muk2}} = \hat{\beta}_{\text{cases}} - \frac{\hat{\tau}^2}{\hat{\tau}^2 + \hat{\sigma}_{\text{controls}}}\hat{\beta}_{\text{controls}}.$$

This estimator distinguishes from $\hat{\beta}_{\text{muk}}$ (equation 6.3) only by using $\hat{\sigma}^2_{\text{controls}}$ instead of $\hat{\sigma}^2_{\text{ge}}$. The corresponding variance expression $\hat{\sigma}^2_{\text{muk2}}$ is of the same form as in 6.4, substituting $\hat{\sigma}^2_{\text{ge}}$ by $\hat{\sigma}^2_{\text{controls}}$. The alternative test statistic is given by

$$T_{\text{muk2}} = \frac{\hat{\beta}_{\text{muk2}}}{\hat{\sigma}_{\text{muk2}}}.$$

In our simulation studies we chose the first alternative T_{muk}, since it is preferred over T_{muk2} in terms of mean-squared error (Mukherjee *et al.*, 2012) and was shown to reach slightly higher power (Mukherjee *et al.*, 2008).

Table 6.1: *Overview of the different GxE methods compared in our simulations*

method	test statistic
Case-control (Breslow and Day, 1994)	$T_{\text{cc}} = \frac{\hat{\beta}_{\text{cases}} - \hat{\beta}_{\text{controls}}}{\hat{\sigma}^2_{\text{cases}} + \hat{\sigma}^2_{\text{controls}}}$
Case-only (Piegorsch *et al.*, 1994)	$T_{\text{cases}} = \frac{\hat{\beta}_{\text{cases}}}{\hat{\sigma}^2_{\text{cases}}}$
Simple two-step (Albert *et al.*, 2001)	$T_{\text{two}} = T_{\text{cc}} I[\mid T_{\text{controls}} \mid > T_{\alpha_{\text{two}_1}/2}]$ $+T_{\text{cases}} I[\mid T_{\text{controls}} \mid \leq T_{\alpha_{\text{two}_1}/2}]$
Murcray's two-step (Murcray *et al.*, 2009)	$T_{\text{mur}} = T_{\text{cc}} I[\mid T_{\text{all}} \mid > T_{\alpha_{\text{mur}}/2}]$
Mukherjee's shrinkage estimator (Mukherjee and Chatterjee, 2008)	$T_{\text{muk}} = \frac{\hat{\beta}_{\text{cases}} - B(\hat{\tau}^2, \hat{\sigma}^2_{\text{cc}})\hat{\beta}_{\text{controls}}}{\hat{\sigma}^2_{\text{muk}}}$
Empirical hierarchical Bayes (alternative 1)	$T_{\text{EHB}} = \frac{\hat{\beta}_{\text{cases}} - \hat{\lambda}}{\sqrt{\hat{\sigma}^2_{\text{cases}} + \hat{\sigma}^2_{\lambda}}}$
Empirical hierarchical Bayes (alternative 2)	$T_{EHB2} = \frac{\hat{T}_{\text{cases}} - \hat{\lambda}^S}{\sqrt{\hat{\sigma}^2_{\text{cases}} + \hat{\sigma}^2_{\lambda^S}}}$

$\hat{\beta}_{\text{cases}}$, $\hat{\beta}_{conrols}$ and $\hat{\beta}_{\text{all}}$ are estimates for G-E association within cases, controls and the whole sample with corresponding variance estimates $\hat{\sigma}_{\text{cases}}$, $\hat{\sigma}_{conrols}$ and $\hat{\sigma}_{\text{all}}$. $B(\hat{\tau}^2, \hat{\sigma}^2_{\text{ge}})$ is a shrinkage factor, σ^2_{muk} denotes an approximate variance considering this factor. $\hat{\lambda}$ and $\hat{\lambda}^S$ are posterior estimates for $\hat{\beta}_{\text{controls}}$ and $\hat{T}_{\text{controls}}$, $\hat{\sigma}^2_{\lambda}$ and $\hat{\sigma}^2_{\lambda^S}$ the corresponding posterior variance estimates.

6.3 Empirical hierarchical Bayes approach for GxE interaction analysis

In 2007, Volk *et al.* presented an idea of an empirical hierarchical Bayes method to detect GxE interactions in GWAS (conference poster). Similar to Mukherjee *et al.* (2008), they proposed to use a compromise between the case-control and case-only test of GxE interaction, by subtracting a posterior estimate of the G-E association within controls from the G-E association within cases.
In contrast to Mukherjee *et al.* (2008), they chose a more complicated hierarchical modeling framework for the G-E association within controls, analog to the model provided for Lewinger *et al.*'s (2007) hierarchical Bayes prioritization (section 4.4.1). Furthermore, their posterior estimate for each marker is not only based on the data of the SNP itself, but is obtained by borrowing G-E information across all SNPs and is therefore characterized by a reduced variance. To normalize their resulting test statistic, they calculated the variance across their simulation replicates.
For this work here, we adopted that basic idea of Volk *et al.* (2007) and derived an appropriate variance for the statistic, to provide a new GxE interaction test applicable to real data where no replicates are available. Additionally, based on distributional considerations, we modified their idea to obtain better properties of the statistic.
Before we will outline our work on the new GxE test statistic in sections 6.3.1 and 6.3.2, we will shortly summarize the basic idea of Volk *et al.* (2007).

Assume that we have a genome-wide association study with N_M genetic markers M_i, $i = 1, .., N_M$. Let $\beta_{M_i}^{\text{cases}}$ and $\beta_{M_i}^{\text{controls}}$ be the corresponding regression coefficients for G-E association among cases and among controls with standard deviations $\sigma_{M_i}^{\text{cases}}$ and $\sigma_{M_i}^{\text{controls}}$. The corresponding test statistics are $T_{M_i}^{\text{cases}} = \hat{\beta}_{M_i}^{\text{cases}}/\hat{\sigma}_{M_i}^{\text{cases}}$ and $T_{M_i}^{\text{controls}} = \hat{\beta}_{M_i}^{\text{controls}}/\hat{\sigma}_{M_i}^{\text{controls}}$.
Volk *et al.* (2007) suggested to model $T_{M_i}^{\text{controls}}$, the estimated statistic for the G-E association within the controls, by the hierarchical model

$$\begin{aligned} &\text{level 1} && \left|T_{M_i}^{\text{controls}}\right| && \mid \lambda_{M_i}^S && \sim \chi_1(\lambda_{M_i}^S) && (6.6)\\ &\text{level 2} && \lambda_{M_i}^S && \mid p^S, \sigma^S, \theta^S && \sim p^S\sigma^S\chi_1(\theta^S) + (1-p)\delta(0). && (6.7) \end{aligned}$$

$\chi_1(\zeta)$ is the χ-distribution with 1 degree of freedom and non-centrality parameter ζ. p^S represents the proportion of SNPs with a population based G-E association, θ^S is the non-centrality parameter for the strength of association and σ^S a scaling parameter. The superscript S is used to distinguish this procedure modeling the test statistic $T_{M_i}^{\text{controls}}$ by the hierarchical model from our alternative recommended in the following section.
By maximization of the corresponding marginal likelihood with respect to the hyperparameters $\Theta^S = (\theta^S, \sigma^S, p^S)$, estimates for these quantities can be obtained. These can be used in the a posteriori expectations $E(\lambda_{M_i}^S | T_{M_i}^{\text{controls}}, \hat{\Theta}_T)$ to yield adequate posterior estimates $\hat{\lambda}_{M_i}^S$ for the non-centrality parameters.
The proposed test statistic $T_{M_i}^S$ subtracts $\hat{\lambda}_{M_i}^S$ from the case-only statistic $T_{M_i}^{\text{cases}}$ to remove the population based G-E association effect

$$V_{M_i}^S = \frac{(T_{M_i}^{\text{cases}} - sg_{M_i}\hat{\lambda}_{M_i}^S)}{s_{M_i}^S}. \qquad (6.8)$$

sg_{M_i} is the sign of the corresponding control-only statistic $T_{M_i}^{\text{controls}}$. The normalizing factors $s_{M_i}^S$ are empirically computed by their simulation studies as follows. For each replicate $r = 1, ..., R$, $U_r = T_{M_i}^{\text{cases}} - sg_{M_i}\hat{\lambda}_{M_i}^S$ was calculated, with $\bar{U}$ and std(U) the corresponding mean and standard deviation over the differen replicates. The normalizing factor for the test statistic was then set to $s_{M_i}^S = \rho$ std(U). In their simulations, $\rho = 1.2$ was found to work well.

6.3.1 Modification of the empirical hierarchical Bayes approach

When we took a closer look at the idea of Volk *et al.* (2007), we recognized a distributional problem with their suggested proceeding.
Assuming a rare disease, we have the following approximate distributions for the different types of SNPs:

no effect	$T_{M_i}^{\text{cases}} \sim N(0,1)$,	$T_{M_i}^{\text{controls}} \sim N(0,1)$
G-E association	$T_{M_i}^{\text{cases}} \sim N\left(\frac{\beta_{M_i}^{\text{cases}}}{\sigma_{M_i}^{\text{cases}}}, 1\right)$ $\beta_{M_i}^{\text{cases}} = \beta_{M_i}^{\text{controls}} \neq 0$	$T_{M_i}^{\text{controls}} \sim N\left(\frac{\beta_{M_i}^{\text{controls}}}{\sigma_{M_i}^{\text{controls}}}, 1\right)$,
GxE interaction	$T_{M_i}^{\text{cases}} \sim N\left(\frac{\beta_{M_i}^{\text{cases}}}{\sigma_{M_i}^{\text{cases}}}, 1\right)$ $\beta_{M_i}^{\text{cases}} \neq 0$	$T_{M_i}^{\text{controls}} \sim N(0,1)$

The two quantities $T_{M_i}^{\text{cases}}$ and $T_{M_i}^{\text{controls}}$ are used as basis for the empirical hierarchical Bayes statistic of Volk *et al.* (2007) by subtracting the a posteriori expectation of $T_{M_i}^{\text{controls}}$ of the case-only statistic $T_{M_i}^{\text{cases}}$. However, for G-E associated SNPs, $\sigma_{M_i}^{\text{cases}}$ is not necessarily equal to $\sigma_{M_i}^{\text{controls}}$. This is particularly the case, given a main effect of the environmental factor or different numbers of cases and controls. In that situation $T_{M_i}^{\text{cases}}$ and $T_{M_i}^{\text{controls}}$ have different expected values, resulting in $(T_{M_i}^{\text{cases}} - T_{M_i}^{\text{controls}})$ not distributed around 0 although the null hypothesis is true. As a consequence, the same holds for $(T_{M_i}^{\text{cases}} - sg_{M_i}\hat{\lambda}_{M_i}^S)$. Thereby, association effects may be misinterpreted as interaction effects.
To overcome this weakness, we improved the empirical hierarchical Bayes test by applying the hierarchical model to $\hat{\beta}_{M_i}^{\text{controls}}$ instead of $T_{M_i}^{\text{controls}}$ and calculating $\beta_{M_i}^{\text{cases}} - sg_{M_i}\hat{\lambda}_{M_i}$ with $\hat{\lambda}_{M_i}$ the posterior estimator of $\hat{\beta}_{M_i}^{\text{controls}}$.
The corresponding hierarchical model is given by

$$\begin{array}{lll} \text{level 1} & \left|\hat{\beta}_{M_i}^{\text{controls}}\right| \mid \lambda_{M_i} & \sim \hat{\sigma}_{M_i}^{\text{controls}}\chi_1(\lambda_{M_i}) \\ \text{level 2} & \lambda_{M_i} \mid \theta, \sigma, p & \sim p\sigma\chi_1(\theta) + (1-p)\delta(0). \end{array} \tag{6.9}$$

The corresponding density functions are given by

$$f(\left|\hat{\beta}_{M_i}^{\text{controls}}\right| \mid \lambda_{M_i}) = \left(\varphi\left(\frac{\mid \hat{\beta}_{M_i}^{\text{controls}} \mid -\lambda_{M_i}}{\hat{\sigma}_{M_i}^{\text{controls}}}\right) + \varphi\left(\frac{\mid \hat{\beta}_{M_i}^{\text{controls}} \mid +\lambda_{M_i}}{\hat{\sigma}_{M_i}^{\text{controls}}}\right)\right) / \hat{\sigma}_{M_i}^{\text{controls}}$$

$$g(\lambda_{M_i} \mid \theta, \sigma, p) = p\left(\varphi\left(\frac{\lambda_{M_i} - \theta}{\sigma}\right) + \varphi\left(\frac{\lambda_{M_i} + \theta}{\sigma}\right)\right) / \sigma + (1-p)\delta(0).$$

$\varphi(\cdot)$ represents the standard normal distribution.
To obtain estimates for the hyperparameters $\Theta = (\theta, p, \sigma)$ of that model, we need the marginal likelihood function. The marginal distribution can be written as

$$\begin{aligned} m(\mid \hat{\beta}_{M_i}^{\text{controls}} \mid\mid \theta, \sigma, p) &= \int f(\mid \hat{\beta}_{M_i}^{\text{controls}} \mid\mid \lambda_{M_i}) g(\lambda_{M_i} \mid \theta, \sigma, p) d\lambda_{M_i} \\ &= p\frac{\varphi(D_{+M_i}) + \varphi(D_{-M_i})}{\sqrt{(\hat{\sigma}_{M_i}^{\text{controls}})^2 + (\sigma)^2}} + (1-p)2\varphi\left(\frac{\mid \hat{\beta}_{M_i}^{\text{controls}} \mid}{\hat{\sigma}_{M_i}^{\text{controls}}}\right) / \hat{\sigma}_{M_i}^{\text{controls}}, \end{aligned}$$

with

$$\begin{aligned} D_{+M_i} &= \frac{\left|\hat{\beta}_{M_i}^{\text{controls}}\right| + \theta}{\sqrt{(\hat{\sigma}_{M_i}^{\text{controls}})^2 + (\sigma)^2}} \\ D_{-M_i} &= \frac{\left|\hat{\beta}_{M_i}^{\text{controls}}\right| - \theta}{\sqrt{(\hat{\sigma}_{M_i}^{\text{controls}})^2 + (\sigma)^2}} \end{aligned} \tag{6.10}$$

The likelihood of the hierarchical model is given by $L = \prod_{M_i} m(\hat{\beta}_{M_i}^{\text{controls}} \mid \theta, \sigma, p)$ and is maximized with respect to $\Theta = (\theta, \sigma, p)$ to obtain the MLE's $\hat{\theta}, \hat{\sigma}, \hat{p}$.
The posterior expected value given $\lambda_{M_i} > 0$ equals

$$\begin{aligned} E_{M_i}^{+} &= \mathrm{E}\left[\lambda_{M_i} \mid \lambda_{M_i} > 0, \hat{\beta}_{M_i}^{\text{controls}}, \Theta\right] \\ &= \frac{\sigma\hat{\sigma}_{M_i}^{\text{controls}}}{\sqrt{(\hat{\sigma}_{M_i}^{\text{controls}})^2 + (\sigma)^2}} \\ &\quad \frac{(Q + L_{+M_i}\varphi(D_{-M_i})(2\Phi(L_{+M_i}) - 1) + L_{-M_i}\varphi(D_{+M_i})(2\Phi(L_{-M_i}) - 1))}{(\varphi(D_{+M_i}) + \varphi(D_{-M_i}))}. \end{aligned}$$

The corresponding posterior probability is given by

$$\begin{aligned} P_{M_i} &= \Pr(\lambda_{M_i} > 0 \mid \hat{\beta}_{M_i}^{\text{controls}}, \Theta) \\ &= \left(1 + \frac{(1-p)}{p}\frac{2\varphi(\mid \hat{\beta}_{M_i}^{\text{controls}} \mid / \hat{\sigma}_{M_i}^{\text{controls}})/\hat{\sigma}_{M_i}^{\text{controls}}}{(\varphi(D_{+M_i}) + \varphi(D_{-M_i}))/\sqrt{(\hat{\sigma}_{M_i}^{\text{controls}})^2 + \sigma^2}}\right)^{-1} \end{aligned}$$

with

$$
\begin{aligned}
Q &= \frac{2}{\pi}\exp\left(-\frac{(\sigma)^2 \mid \hat{\beta}_{M_i}^{\text{controls}} \mid^2 + (\hat{\sigma}_{M_i}^{\text{controls}})^2(\theta)^2}{2(\sigma)^2(\hat{\sigma}_{M_i}^{\text{controls}})^2}\right)\\
L_{+M_i} &= \frac{(\hat{\sigma}_{M_i}^{\text{controls}})^2\theta + (\sigma)^2 \mid \hat{\beta}_{M_i}^{\text{controls}} \mid}{\sigma\hat{\sigma}_{M_i}^{\text{controls}}\sqrt{(\hat{\sigma}_{M_i}^{\text{controls}})^2 + (\sigma)^2}}\\
L_{-M_i} &= \frac{(\hat{\sigma}_{M_i}^{\text{controls}})^2\theta - (\sigma)^2 \mid \hat{\beta}_{M_i}^{\text{controls}} \mid}{\sigma\hat{\sigma}_{M_i}^{\text{controls}}\sqrt{(\hat{\sigma}_{M_i}^{\text{controls}})^2 + (\sigma)^2}}
\end{aligned}
$$

and D_{+M_i} and D_{-M_i} as defined above.
Hence

$$
\begin{aligned}
\mathrm{E}\left[\lambda_{M_i} \mid \hat{\beta}_{M_i}^{\text{controls}}, \Theta\right] &= \mathrm{E}\left[\lambda_{M_i} \mid \lambda_{M_i} > 0, \hat{\beta}_{M_i}^{\text{controls}}, \Theta\right] P_{M_i}\\
&\quad + E\left[\lambda_{M_i} \mid \lambda_{M_i} = 0, \hat{\beta}_{M_i}^{\text{controls}}, \Theta\right](1 - P_{M_i})\\
&= E_{M_i}^{+} + P_{M_i}.
\end{aligned}
$$

By using the MLE estimates $\hat{\Theta} = (\hat{\theta}, \hat{\sigma}, \hat{p})$, we obtain the posterior expectation of the non-centrality parameter

$$
\hat{\lambda}_{M_i} = \mathrm{E}\left[\lambda_{M_i} \mid \hat{\beta}_{M_i}^{\text{controls}}, \hat{\theta}, \hat{\sigma}, \hat{p}\right] \tag{6.11}
$$

While Volk *et al.*'s (2007) test statistic version $V_{M_i}^{S}$ subtracts the posterior estimate of $T_{M_i}^{\text{controls}}$ off the case-only statistic $T_{M_i}^{\text{cases}}$ (equation 6.8), our newly proposed alternative V_{M_i} subtracts $\hat{\lambda}_{M_i}$, the posterior estimate of $\beta_{M_i}^{\text{controls}}$, off the case parameter $\hat{\beta}_{M_i}^{\text{cases}}$ to remove the population association

$$
V_{M_i} = \frac{(\hat{\beta}_{M_i}^{\text{cases}} - sg_{M_i}\hat{\lambda}_{M_i})}{s_{M_i}^{rep}}. \tag{6.12}
$$

We have $sg_{M_i} = sgn(\hat{\beta}_{M_i}^{\text{controls}})$ and normalizing factor $s_{M_i}^{rep} = \rho\ sd(\hat{\beta}_{M_i}^{\text{cases}} - sg_{M_i}\hat{\lambda}_{M_i})$ obtained empirically as explained above.
Note that all corresponding formulae of the different distributions mentioned can be achieved for the originally proposed version of Volk *et al.* (2007) by simply replacing $\hat{\beta}_{M_i}^{\text{cases}}$ and $\hat{\beta}_{M_i}^{\text{controls}}$ by $T_{M_i}^{\text{cases}}$ and $T_{M_i}^{\text{controls}}$ and substituting $\hat{\sigma}_{M_i}^{\text{cases}}$ and $\hat{\sigma}_{M_i}^{\text{controls}}$ with 1.

6.3.2 Calculation of an appropriate variance for the statistic

To obtain a usable test statistic for the application to real data where no replicates are available, an appropriate variance for $(\hat{\beta}_{M_i}^{\text{cases}} - sg_{M_i}\hat{\lambda}_{M_i})$ and $(\hat{T}_{M_i}^{\text{cases}} - sg_{M_i}\hat{\lambda}_{M_i}^{S})$ has to be calculated. For both T-based and β-based strategy we derived that variance. In the following we will illustrate our variance derivations for the β-based statistic.
Case and control part of the differences $(\hat{\beta}_{M_i}^{\text{cases}} - sg_{M_i}\hat{\lambda}_{M_i})$ are independent from each other, and $\mathrm{Var}(\hat{\beta}_{M_i}^{\text{cases}}) = (\sigma_{M_i}^{\text{cases}})^2$, estimated by $(\hat{\sigma}_{M_i}^{\text{cases}})^2$.

Hence, we only have to determine the variance for the control part, the $\hat{\lambda}_{M_i}$, what we calculate by $\text{Var}\left[\lambda_{M_i} \mid \hat{\beta}_{M_i}^{\text{controls}}\right]$. As already mentioned in chapter 4.2, the conditional posterior expectation $\text{E}\left[\lambda_{M_i} \mid \hat{\beta}_{M_i}^{\text{controls}}, \hat{\Theta}\right]$ is approximately equal to the posterior mean $\text{E}\left[\lambda_{M_i} \mid \hat{\beta}_{M_i}^{\text{controls}}\right]$, while the conditional posterior variance $\text{Var}\left[\lambda_{M_i} \mid \hat{\beta}_{M_i}^{\text{controls}}, \hat{\Theta}\right]$ underestimates $\text{Var}\left[\lambda_{M_i} \mid \hat{\beta}_{M_i}^{\text{controls}}\right]$. Both have the relationship:

$$\text{Var}\left[\lambda_{M_i} \mid \hat{\beta}_{M_i}^{\text{controls}}\right] = \text{E}\left[Var(\lambda_{M_i} \mid \hat{\beta}_{M_i}^{\text{controls}}, \Theta)\right] + \text{Var}\left[E(\lambda_{M_i} \mid \hat{\beta}_{M_i}^{\text{controls}}, \Theta)\right].$$

To calculate the posterior variance $\text{Var}\left[\lambda_{M_i} \mid \hat{\beta}_{M_i}^{\text{controls}}\right]$ for our statistic, we used an approximation of Kass and Steffey (1989), given by

$$\text{Var}\left[\lambda_{M_i} \mid \hat{\beta}_{M_i}^{\text{controls}}\right] \approx \text{Var}\left[\lambda_{M_i} \mid \hat{\beta}_{M_i}^{\text{controls}}, \hat{\Theta}\right] + \sum_{j,h} \tilde{\tau}_{jh}\tilde{\delta}_{M_ij}\tilde{\delta}_{M_ih}.$$

$\tilde{\tau}_{jh}$ is the (j,h)-component of the inverse negative Hessian of the marginal log-likelihood evaluated at the marginal maximum likelihood estimator $\tilde{\Sigma} = (-D^2 log(L)(\hat{\Theta}))^{-1}$. $\tilde{\delta}_{M_ik}$ is given by the Jacobian of the posterior expectation with

$$\tilde{\delta}_{M_ik} = (\partial/\partial\Theta_k)\text{E}\left[\lambda_{M_i} \mid \hat{\beta}_{M_i}^{\text{controls}}, \Theta\right] \mid_{\Theta=\hat{\Theta}} \tag{6.13}$$

at $\hat{\Theta}$. According to this first order approximation, we derived the specific variance for our model using maple.
The approximation has two parts, the conditional variance and the correction term. For the first part we have

$$\text{Var}\left[\lambda_{M_i} \mid \hat{\beta}_{M_i}^{\text{controls}}, \hat{\Theta}\right] = \text{E}\left[\lambda_{M_i}^2 \mid \hat{\beta}_{M_i}^{\text{controls}}, \hat{\Theta}\right] - \text{E}\left[\lambda_{M_i} \mid \hat{\beta}_{M_i}^{\text{controls}}, \hat{\Theta}\right]^2.$$

$\text{E}\left[\lambda_{M_i} \mid \hat{\beta}_{M_i}^{\text{controls}}, \hat{\Theta}\right]$ is known already, so we only have to derive

$$\text{E}\left[\lambda_{M_i}^2 \mid \hat{\beta}_{M_i}^{\text{controls}}, \hat{\Theta}\right] = \text{E}\left[\lambda_{M_i}^2 \mid \lambda_{M_i} > 0, \hat{\beta}_{M_i}^{\text{controls}}, \hat{\Theta}\right] P\left[\lambda_{M_i} > 0 \mid \hat{\beta}_{M_i}^{\text{controls}}, \hat{\Theta}\right] \tag{6.14}$$

$$+\text{E}\left[\lambda_{M_i}^2 \mid \lambda_{M_i} = 0, \hat{\beta}_{M_i}^{\text{controls}}, \hat{\Theta}\right] P\left[\lambda_{M_i} = 0 \mid \hat{\beta}_{M_i}^{\text{controls}}, \hat{\Theta}\right]. \tag{6.15}$$

$P\left[\lambda_{M_i} > 0 \mid \hat{\beta}_{M_i}^{\text{controls}}, \hat{\Theta}\right]$ is given in equation 4.22,
$\text{E}\left[\lambda_{M_i}^2 \mid \lambda_{M_i} = 0, \hat{\beta}_{M_i}^{\text{controls}}, \hat{\Theta}\right] = 0.$
For the unknown part we get

$$\text{E}\left[\lambda_{M_i}^2 \mid \lambda_{M_i} > 0, \hat{\beta}_{M_i}^{\text{controls}}, \hat{\Theta}\right] = \frac{(\sigma)^2(\sigma_{M_i}^{\text{controls}})^2}{(\sigma)^2 + (\sigma_{M_i}^{\text{controls}})^2} \cdot \frac{((L_{+M_i}+1)\varphi(D_{+M_i}) + (L_{-M_i}+1)\varphi(D_{-M_i}))}{\varphi(D_{+M_i}) + \varphi(D_{-M_i})}.$$

The Jacobian and Hessian used in the correction term are given by

$$\tilde{\nabla}_{M_i} = \begin{pmatrix} \tilde{\delta}_{M_i\theta} \\ \tilde{\delta}_{M_i\sigma} \\ \tilde{\delta}_{M_ip} \end{pmatrix} \textbf{ and } \tilde{\Sigma} = \begin{pmatrix} \tilde{\tau}_{\theta\theta} & \tilde{\tau}_{\theta\sigma} & \tilde{\tau}_{\theta p} \\ \tilde{\tau}_{\sigma\theta} & \tilde{\tau}_{\sigma\sigma} & \tilde{\tau}_{\sigma p} \\ \tilde{\tau}_{p\theta} & \tilde{\tau}_{p\sigma} & \tilde{\tau}_{pp} \end{pmatrix},$$

with the detailed formulas for the individual parts listed in the appendix part C.
All together, we have

$$\text{Var}\left[\lambda_{M_i} \mid \hat{\beta}_{M_i}^{\text{controls}}\right] = \text{E}\left[(\lambda_{M_i})^2 \mid \lambda_{M_i} > 0, \hat{\beta}_{M_i}^{\text{controls}}; \hat{\Theta}\right] P_{M_i} - (E_{M_i}^{+b} P_{M_i})^2 + (\tilde{\nabla}_{M_i})^S \tilde{\Sigma} \tilde{\nabla}_{M_i}. \tag{6.16}$$

Our final test statistic is given by

$$T_{M_i}^{EHB} = \frac{\hat{\beta}_{M_i}^{\text{cases}} - sg_{M_i}\hat{\lambda}_{M_i}}{\sqrt{(\hat{\sigma}_{M_i}^{\text{cases}})^2 + \text{Var}\left[\lambda_{M_i} \mid \hat{\beta}_{M_i}^{\text{controls}}\right]}} \tag{6.17}$$

For the corresponding variance for the T-based version of the test statistic, we have to replace $\hat{\beta}_{M_i}^{\text{cases}}$ and $\hat{\beta}_{M_i}^{\text{controls}}$ in all formulas by $T_{M_i}^{\text{cases}}$ and $T_{M_i}^{\text{controls}}$ and substitute $\hat{\sigma}_{M_i}^{\text{cases}}$ and $\hat{\sigma}_{M_i}^{\text{controls}}$ by 1. The final statistic is given by

$$T_{M_i}^{EHB2} = \frac{T_{M_i}^{\text{cases}} - sg_{M_i}\hat{\lambda}_{M_i}^{S}}{\sqrt{1 + \text{Var}\left[\lambda_{M_i}^{S} \mid \hat{T}_{M_i}^{\text{controls}}\right]}}. \tag{6.18}$$

In the following, we will abbreviate our new empirical hierarchical Bayes approach by EHB, while we denote the version based on T-statistics extended by our new estimated variance with $EHB2$.

6.4 Simulation studies

For the investigation of the performance of our method, we generated data sets under different parameter settings and compared the results of our approach with the results of the case-control (CC), case-only (CASES), the two different two-step methods described (TWO, MUR) and the approach of Mukherjee (MUK). An overview of the different test statistics is given in table 6.1.
We will compare the ability of the methods to identify an interacting marker within the top rankings according to the corresponding method. The first rank only, the top 10, top 25, top 50 and top 100 are considered. We denote the percentage of replicates where the interacting SNP is within the considered top positions by **rank power**.
We concentrated on the rank power and not on the type I error and power in a conventional sense, since GWAS are typically considered as a screening process, selecting a subset of top SNPs for further investigation (see section 3.2.4). The rank power addresses this issue, evaluating the quality of a method not to miss a true positive effect for follow-up, even given a small effect not reaching significance. The discovery step procedure should really guarantee to find true effects. False positive findings are weed out in the following independent replication.

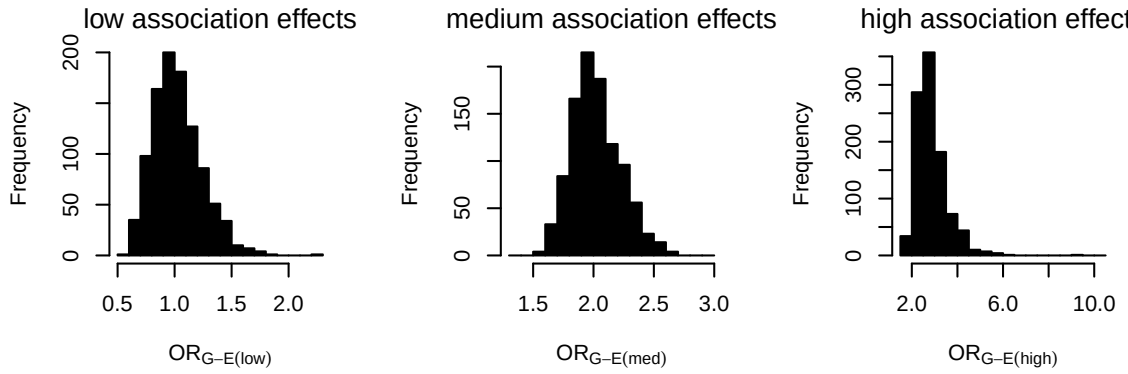


Figure 6.1: *Distribution of the G-E association effects for the 1,000 simulated markers with a G-E association. We distinguished between low, medium and high association scenarios, with values sampled from* $log(OR_{\text{G-E(low)}}) \sim N(0, log(1.5)/2)$, $log(OR_{\text{G-E(med)}}) \sim N(0.7, 0.1)$ *and* $log(log(OR_{\text{G-E(high)}})) \sim N(0, log(1.5)/2)$.

6.4.1 Simulation set-up

In our simulated data we restricted to one environmental factor and one interacting SNP. For both, we varied the frequency for their corresponding risk variant between 10, 30 and 50%. For each combination of exposure and marker frequency, five different interacting odds ratios OR_{GxE} between 1.2 and 3 were chosen. Based on that information, cohorts of the general population were generated by using the logistic regression model

$$log\left(\frac{p}{1-p}\right) = \alpha + \beta_{\text{GxE}}\text{GxE}, \qquad \beta_{\text{GxE}} = log(OR_{\text{GxE}}).$$

α was chosen such that the overall disease prevalence equals 1, 5 or 10%. Initially, main effects were not considered.
From the generated cohort, cases and controls were randomly selected, with case:control ratio 1,500:1,500, 1,000:2,000 and 2,000:1,000 (short: 1:1,1:2,2:1). For these cases and controls, up to 1,000 SNPs with a population based G-E association were simulated ($N_{\text{G-E}} \in \{0,\ 1, 5, 10, 25, 50, 75, 100, 200, 300, 400, 500, 600, 700, 800, 900, 1,000\}$). The strength of these population based G-E associations was varied between low, medium and high associations by the following distributions

$$\begin{aligned}
OR_{\text{G-E(low)}} &= \exp(\beta_{\text{G-E(low)}}) \text{ with} & \beta_{\text{G-E(low)}} &\sim N\left(0, \frac{log(1.5)}{2}\right) \\
OR_{\text{G-E(med)}} &= \exp(\beta_{\text{G-E(med)}}) \text{ with} & \beta_{\text{G-E(med)}} &\sim N(0.7, 0.1) \\
OR_{\text{G-E(high)}} &= \exp(\beta_{\text{G-E(high)}}) \text{ with} & log(\beta_{\text{G-E(high)}}) &\sim N\left(0, \frac{log(1.5)}{2}\right)
\end{aligned} \tag{6.19}$$

In figure 6.1 the generated $OR_{\text{G-E}}$ for all three situations are graphically represented. The first distribution (low association) was adopted from simulation studies in Mukherjee *et al.* (2008), comparing their methods to CC, CASES and TWO. As we can see in the corresponding histogram, there are positive and negative G-E associations. However, most of the SNPs will only have a really low effect close to 1, nearly representing

G-E independence. To reach odds ratios varying around 2, we chose the second alternative (medium association). In that case, each of the G-E association markers has a recognizable effect of at least $OR_{\text{G-E}} = 1.5$. For environmental factors such as smoking, associated strongly to numerous genetic factors, even more extreme G-E associations may be given. To represent such a high association case, we used the exponential values from the low association, reaching from an OR of 1 up to 10.
To reach a total number of 10,000 SNPs per generated data set, additional SNPs with no effect to the disease and no underlying G-E independence were generated. Except for the one GxE SNP, all other minor allele frequencies were randomly chosen from a Beta distribution B(1,3) truncated to the interval $[0.01 - 0.5]$. For each parameter setting, 1,000 replicates were generated.

In view of our data application to lung cancer GWAS where a main effect of the environmental factor smoking is given, we furthermore performed simulation scenarios including an environmental main effect $OR_e = 2$, 5 or 10. Due to efficiency reasons, we restricted to a subset of the scenarios given above. We varied the exposure frequency between 30 and 50% and chose a frequency of 10% and 30% for the interacting marker. The interaction effect OR_{GxE} was given by 1.5, 2 and 3. We simulated two different disease prevalences, 1% and 10%. For each of the prevalences, we picked 3 scenarios with respect to the numbers of cases and controls, fitting to our different analyzes in chapter 7. For 0.01 we chose 300 cases and 500 controls, 500 cases and 500 controls as well as 2,000 cases and 2,500 controls. Given a disease prevalence of 0.1, the samples consist of 250 cases and 250 controls, 500 cases and 250 controls or 1,500 cases and 1,500 controls. In table 6.2, the relation of these scenarios to our data application is illustrated. We restricted to $N_{\text{G-E}} \in \{0,\ 50,\ 100,\ 200,\ 500,\ 1,000\}$ with G-E association effect sizes given by $OR_{\text{G-E(low)}}$, $OR_{\text{G-E(med)}}$ or $OR_{\text{G-E(high)}}$. For comparison purpose, the same situations were simulated given $OR_e = 1$.

6.4.2 Simulation results

Behavior of the empirical hierarchical Bayes (EHB) approach

Before we will compare the performance of our new empirical hierarchical Bayes approach to other GxE interaction methods outlined in this chapter, we will first take a look at the behavior of the empirical hierarchical Bayes approach with respect to different parameters.
Interaction effect: As expected of a GxE interaction testing procedure, the ranking power of the EHB increases with increasing effect size of the interaction (OR_{GxE}). We can see this behavior in figures 6.2-6.4 for several settings of the other simulation parameters. Furthermore, we see higher rank power with increasing frequency of the environmental (p_e) or genetic factor (p_g) up to 50%. The underlying rational of this is that a higher balance between the different risk groups is given with the frequency approaching 50%. Note, we did not simulate environmental or genetic factors with a frequency exceeding 50%. In that case, a decrease of rank power would be observed.
Prevalence: However, comparing the different prevalences assumed for the underlying disease, we observe an uncommon characteristic with the rank power decreasing

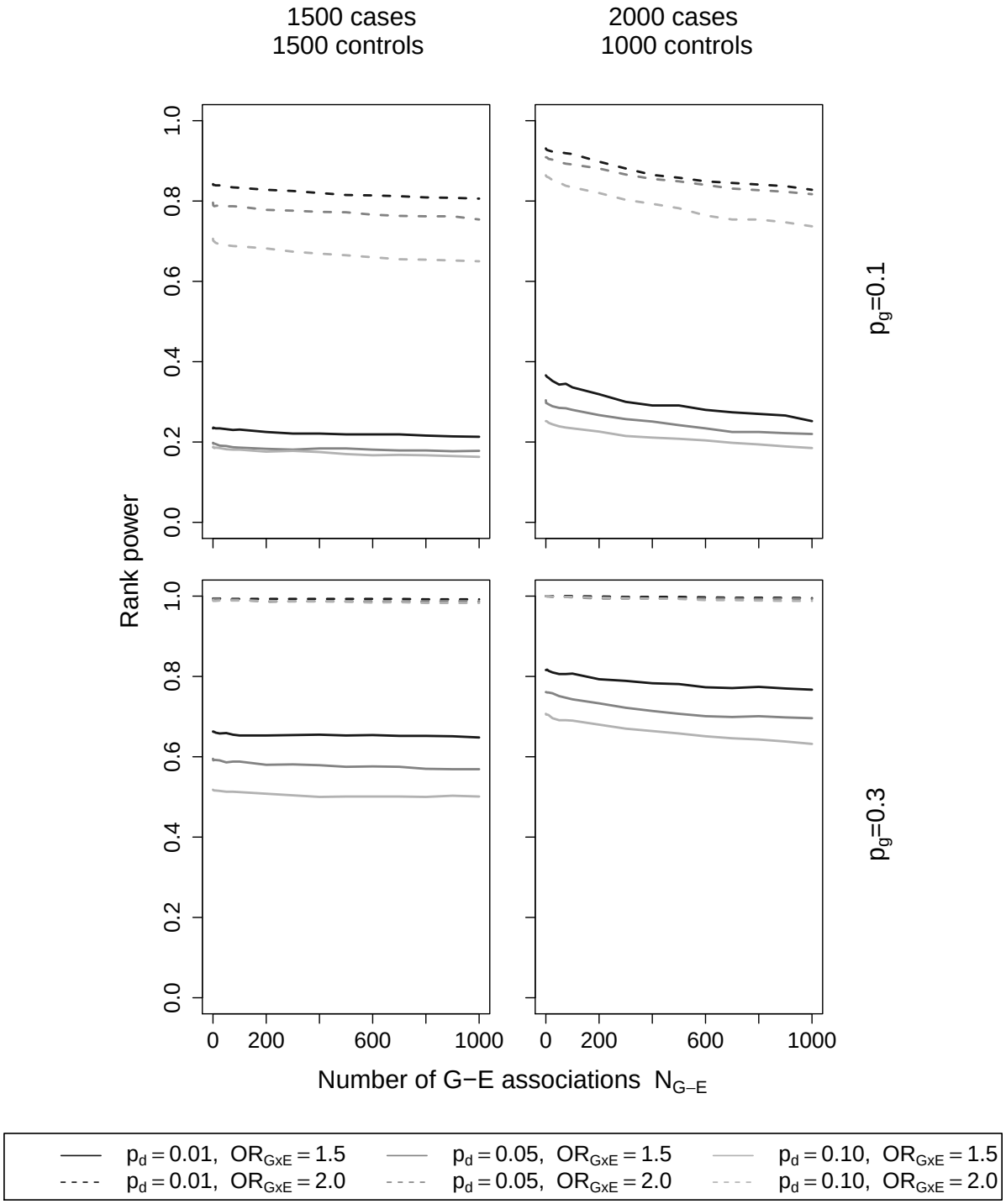


Figure 6.2: *Rank power to detect a GxE interaction in the top 25 ranking SNPs using the EHB method given* ***different prevalence*** p_d *of the underlying disease. A varying number of population based G-E associations* $N_{\mathrm{G-E}}$ *with a high effect* ($OR_{\mathrm{G-E(high)}}$) *is assumed. The environmental factor has a frequency of* p_e*=0.3.*

Table 6.2: *Connection of our simulated scenarios given an environmental main effect to our data applications in chapter 7. GLC, CE-IARC, MDACC and SLRI denote the four different lung cancer GWAS considered. Two analyses are performed for each study, using never and ever smokers as binary classification of the environmental factor or moderate and heavy smokers.*

Simulated prevalence: 0.01				
	lung cancer in the general population testing never vs. ever smokers			
	GLC	CE-IARC	MDACC	SLRI
Given number cases:controls	467:468	1,901:2,503	-	331:499
Simulated number cases:controls	500:500	2,000:2,500	-	300:500
Simulated prevalence: 0.1				
	lung cancer within ever smokers testing moderate vs. heavy smokers			
	GLC	CE-IARC	MDACC	SLRI
Given number cases:controls	411:253	1,752:1,617	1,150:1,134	183:228
Simulated number cases:controls	500:250	1,500:1,500	1,500:1,500	250:250

with increasing disease occurrence. This behavior is given for nearly all different combinations of other parameters and we can see some examples in figure 6.2. The decrease ranges up to 10% comparing a prevalence of 1% and 5% and we reach up to 15% more rank power for the prevalence of 1% in comparison to 10%. This behavior is adapted from case-only test, where the increase with decreasing prevalence is even stronger. This characteristic can be explained by a stronger enrichment of individuals with underlying genetic and environmental susceptibility factor in cases and hence a better balance of the different risk groups. For case-control, rank power increases with increasing prevalence.

Case-control ratio: The behavior of the method with respect to different given ratios of cases and controls contained in the underlying sample highly depends on the number of the given population based G-E associations and their effect size. In figure 6.3 we compare the three different combinations of cases and controls to each other for low, medium and high effects of the G-E association, given 5%, 10% and 30% frequency for disease, environmental factor and interacting marker.

We clearly see, that given a low number of G-E associations (left plot of figure 6.3), the test involving more cases than controls outperforms the two other situations. Having twice as much controls as cases is the most unfavorable case. This behavior persists independent of the number of G-E association effect.

However, given stronger G-E association effects, as in the middle and right plot of figure 6.3, the rank power for the scenario with 2:1 ratio decreases clearly with an increasing number of the population based associations. This effect is even stronger for the medium association strength situation than for the high one. Since the case control ratios of 1:2 and 1:1 decrease only slightly, 2:1 cannot keep the advantage in that case so that 1:1 is the best proportion of cases and controls for a higher number of G-E

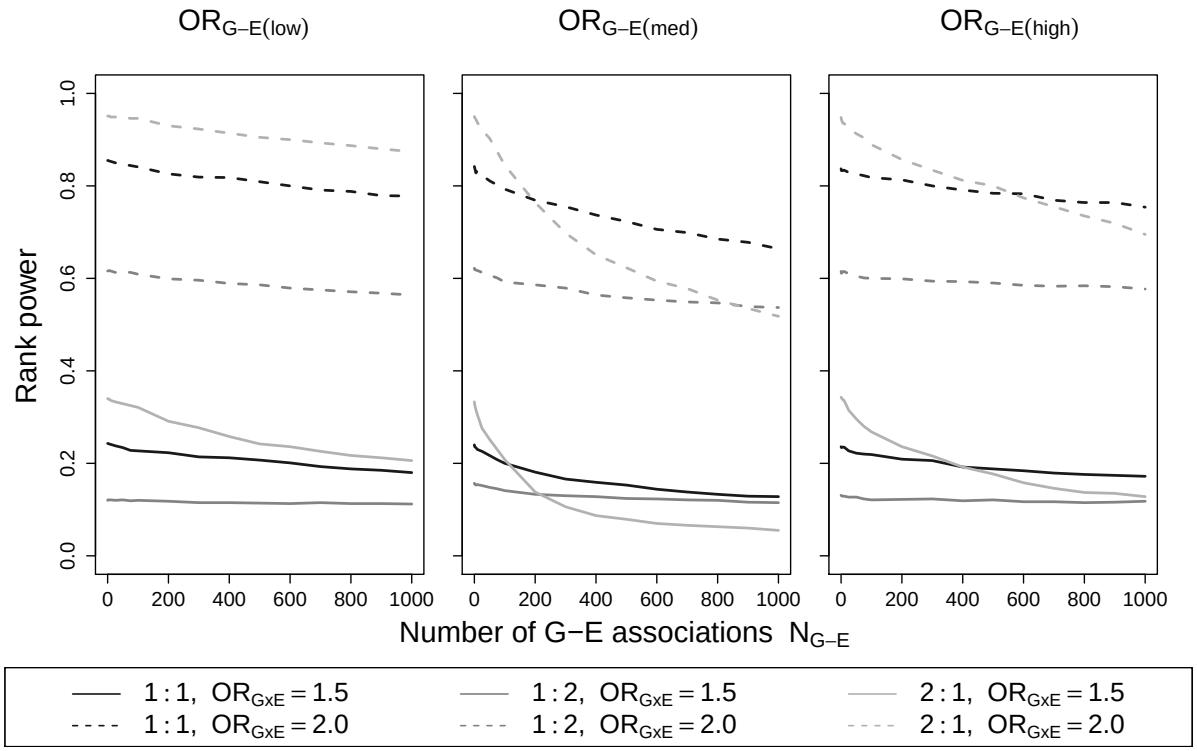


Figure 6.3: *Rank power to detect a GxE interaction in the top 25 ranking SNPs using the EHB method given* ***different ratios of the underlying cases and controls****. A varying number of population based G-E associations* $N_{\mathrm{G-E}}$ *with different effect sizes* $OR_{\mathrm{G-E(low)}}$,$OR_{\mathrm{G-E(med)}}$ *and* $OR_{\mathrm{G-E(high)}}$ *is assumed. Frequency of disease, environmental factor and interacting marker are given by 5%, 10% and 30%. The case control ratios 1:1, 1:2 and 2:1 represent 1,500 cases and 1,500 controls, 1,000 cases and 2,000 controls as well as 2,000 cases and 1,000 controls.*

associations. For the medium association strength, this crosspoint is reached earlier as for the high association case. Furthermore, for a smaller OR of the interaction effect, where the rank power is generally lower and the difference between 2:1, 1:1 and 1:1 smaller, we also see an earlier advantage of the 1:1 ratio.
The situation in the plots is representative for all combinations of prevalence, environmental factor and genetic factor combinations considered, with the concrete crosspoint varying (data not shown).
G-E association effects: In figure 6.4 we see in each of the plots the comparison of rank power between low, medium and high G-E association situation for fixed values of the other parameters. On the left side, where we have the situation of 1,500 cases and 1,500 controls, we see that given an environmental factor with a frequency of 10%, the low association case reaches most rank power. In the high association situation the rank power is even larger than for medium association. Given a more common environmental factor, the method even reaches highest rank power given high association effects, the lowest rank power is reached given only low G-E effects. The same trend is visible in the 2,000 cases : 1,000 controls situation shown on the right part of figure 6.4. The underlying reason for that may be that given a more frequent environmental factor,

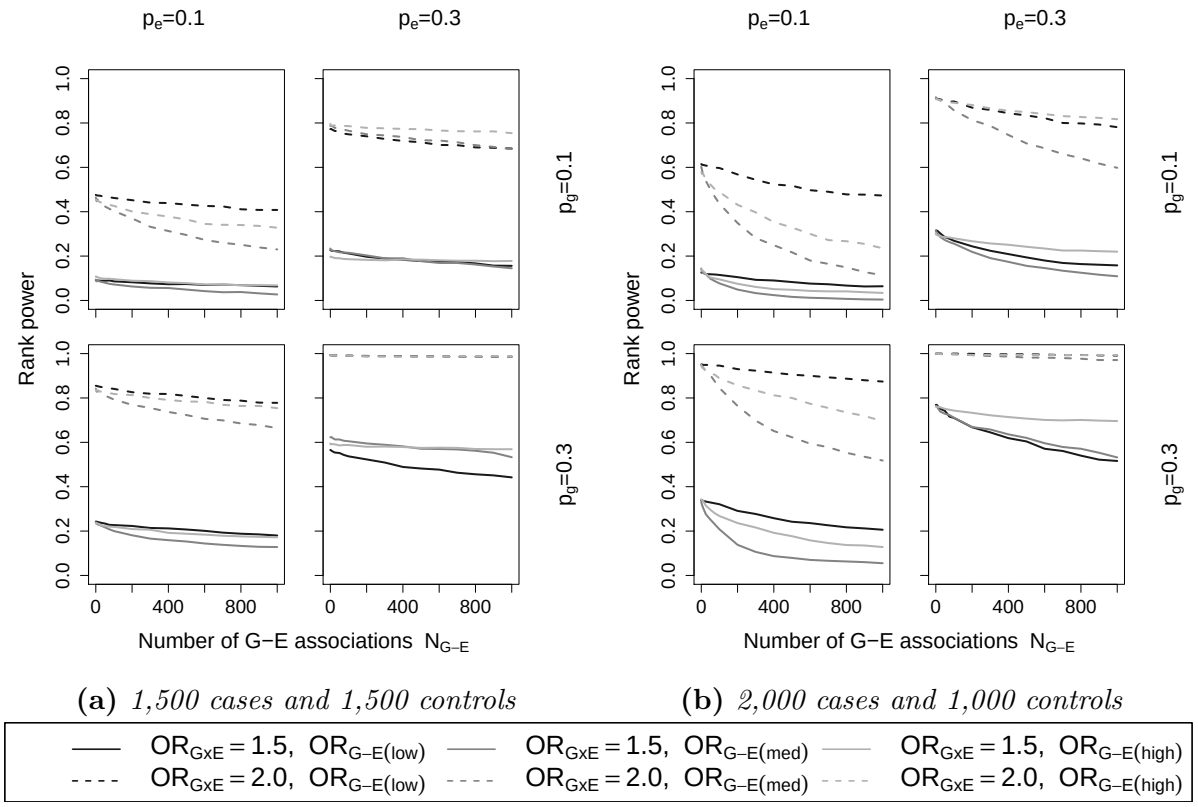


Figure 6.4: *Rank power to detect a GxE interaction in the top 25 ranking SNPs using the EHB method given* ***different strengths of the association effect*** *($OR_{\text{G−E(low)}}$,$OR_{\text{G−E(med)}}$,$OR_{\text{G−E(high)}}$). The underlying sample contains 1,500 cases and 1,500 controls for the four plots on the left and cases and 1,000 controls for the four plots on the right. A disease prevalence of 5% is assumed.*

high association effects are better detectable by our method and therefore a better correction for the G-E association can be done. Comparing 1:1 with 2:1, we see clearly, that the difference between the rank power of the association-strength situations deviates much stronger from each other for 2:1 than for 1:1.

Environmental main effect: In figure 6.5 we can see how the rank power of the EHB is influenced by an environmental main effect. We see that given an environmental factor of frequency p_e=0.3, the rank power for an environmental main effect of OR_e=2 is larger than given no main effect. For higher strength of OR_e=5 and 10, a decreased rank power is observed. Given an exposure frequency of p_e=0.5, the rank power increase observed for OR_e=2 diminishes. These plots are representative for the other considered simulation scenarios as well. The trend of decreasing power with increasing environmental main effect is also observed for the other GxE rank methods.

Comparison of the EHB to other GxE interaction methods

In table 6.3 we can see a part of the results comparing the top 25 ranking power for the different GxE interaction methods with our EHB when *no population based G-E associations occur*. With MUR, the interacting SNP ranks in the top 25 for every situation and hence this test shows highest ranking power considering the top 25.

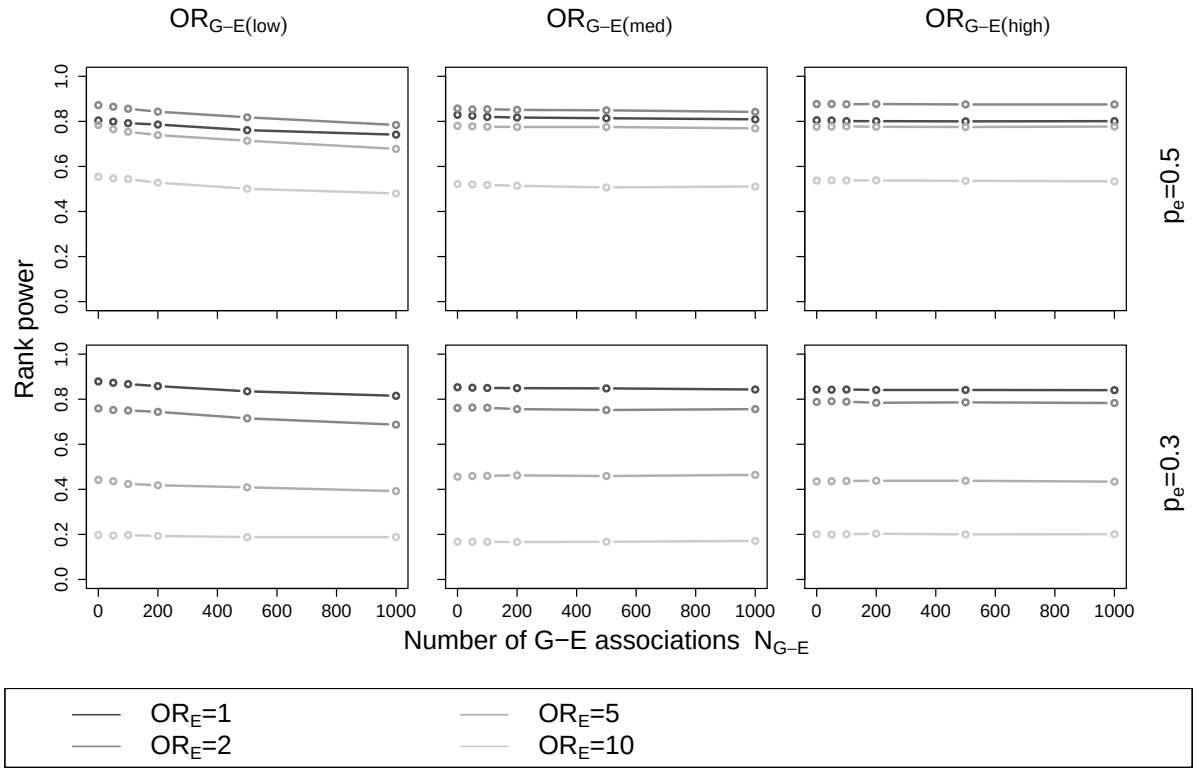


Figure 6.5: *Rank power to detect a GxE interaction in the top 25 ranking SNPs using the EHB method given* ***different environmental main effects****. The underlying sample contains 2,000 cases and 2500 controls, a disease prevalence of 5% is assumed. The GxE interaction has an effect of* $OR_{\mathrm{GxE}} = 1.5$ *with marker frequency* $p_g = 0.3$.

Between EHB and EHB2 we see no difference in ranking power. In all situations, both reach nearly the same ranking power as the case-only approach. MUK and TWO always show a little less ranking power than EHB, with the rank power difference depending on the case-control ratio. Given 1,000 cases and 2,000 controls, the EHB reaches up to 5 and 10% more rank power than MUK and TWO, given 2,000 cases and 1,000 controls, the rank power advantage of the EHB increases further.
Of higher interest however, is the performance of the methods given *G-E associations on a population level*. We observed that in nearly all situations EHB reaches similar or even higher rank power than all other approaches including MUR. In particular when a higher number of G-E associated markers or strong G-E association effects occur, EHB is the superior method.
In table 6.4, 6.5 and 6.6 we see the comparison of the top 25 ranking powers of the different methods representative for a prevalence of 1%, genetic and environmental factor frequencies each of 10 and 30% and an effect size of the interaction of 1.5 and 2.
Table 6.4 shows the results for 1,500 cases and 1,500 controls. Since we observed that the EHB2 has almost identical rank power as the EHB given a case control ratio of 1:1, we neglected EHB2 in this table.
We see that for an interaction effect of size 2 and $p_e = p_g = 0.3$ the case-control method reaches an adequate ranking power of around 85% to detect the interaction

Table 6.3: *Power to detect a GxE interacting SNP in the top 25 ranking when* ***no population based G-E associations*** *occur. The underlying disease is assumed to have a prevalence of 1%.*

cases:controls	p_e	p_g	OR_{GxE}	EHB	CC	CASES	TWO	MUK	MUR
1,500:1,500	0.1	0.1	1.5	0.100	0.019	0.099	0.078	0.067	1
			2	0.458	0.119	0.459	0.402	0.302	1
		0.3	1.5	0.267	0.081	0.268	0.235	0.222	1
			2	0.867	0.405	0.867	0.813	0.729	1
	0.3	0.1	1.5	0.215	0.071	0.215	0.184	0.167	1
			2	0.839	0.412	0.837	0.794	0.721	1
		0.3	1.5	0.665	0.264	0.668	0.605	0.554	1
			2	0.997	0.855	0.997	0.977	0.961	1
1,000:2,000	0.1	0.1	1.5	0.070	0.031	0.071	0.060	0.061	1
			2	0.315	0.138	0.314	0.282	0.250	1
		0.3	1.5	0.159	0.094	0.158	0.150	0.150	1
			2	0.711	0.397	0.709	0.654	0.633	1
	0.3	0.1	1.5	0.117	0.072	0.116	0.107	0.109	1
			2	0.607	0.348	0.608	0.549	0.543	1
		0.3	1.5	0.403	0.225	0.401	0.353	0.353	1
			2	0.971	0.828	0.972	0.941	0.931	1
2,000:1,000	0.1	0.1	1.5	0.143	0.010	0.142	0.111	0.057	1
			2	0.656	0.055	0.657	0.591	0.356	1
		0.3	1.5	0.401	0.061	0.402	0.347	0.230	1
			2	0.975	0.325	0.975	0.940	0.729	1
	0.3	0.1	1.5	0.382	0.062	0.380	0.314	0.227	1
			2	0.946	0.336	0.945	0.894	0.706	1
		0.3	1.5	0.811	0.237	0.811	0.751	0.590	1
			2	1	0.802	1	0.981	0.933	1

cases:controls = number of cases and controls, p_e = frequency of environmental factor, p_g = frequency of genetic factor, OR_{GxE} interaction effect, EHB = empirical hierarchical Bayes, CC = case-control, CASES = case-only,TWO = intuitive two-step, MUK = Mukherjee, MUR = Murcray

SNP within its top 25. For all other situations shown, the rank power of the CC is substantially lower, so that the interacting marker is often missed for follow-up. In all situations, EHB performs better than case-control, resulting in a high rank power to find an interaction given the parameter combinations ($p_g = 0.1,\ p_e = 0.3,\ OR_{GxE} = 2$), ($p_g = 0.3,\ p_e = 0.1,\ OR_{GxE} = 2$) and ($p_g = p_e = 0.3\ OR_{GxE} = 1.5$) and nearly 100% rank power to detect the interaction marker with ($p_g = p_e = 0.3,\ OR_{GxE} = 2$).
Comparing EHB to the case-only method, we see that given a low number of low effect associations the rank power is very similar. However, with increasing size and strength of association, the case-only method fails tremendously and is clearly inferior to EHB that can nearly keep the rank power level.
TWO and MUK both show rank power slightly less than EHB. For TWO we can see differences up to 5%, for MUK we observe in some situations up to 10% less rank power than for EHB. This rank power increase of 5 to 10% given by the EHB may often be crucial and responsible if the interacting marker is further investigated or not.
The performance of MUR highly depends on the number of population based G-E associations and their strength. When the number or strength of G-E associations increases, the rank power of MUR decreases extremely so that EHB is much better in

Table 6.4: *Rank power to detect a GxE interacting SNP in the top 25 ranking in the* ***presence of population based G-E associations****. The underlying disease is assumed to have a prevalence of 1%. The given sample consists of* ***1,500 cases and 1,500 controls****.*

$\mathbf{p}_e$	$\mathbf{OR}_{\mathrm{G-E}}$	$\mathbf{N}_{\mathrm{G-E}}$	$\mathbf{p}_g$	$\mathbf{OR}_{\mathrm{GxE}}$	**EHB**	**CC**	**CASES**	**TWO**	**MUK**	**MUR**
0.1	low	100	0.1	1.5	0.099	0.019	0.097	0.078	0.067	0.998
				2	0.443	0.119	0.446	0.401	0.302	0.992
			0.3	1.5	0.258	0.081	0.256	0.231	0.217	0.996
				2	0.856	0.405	0.856	0.809	0.727	0.997
		500	0.1	1.5	0.080	0.019	0.067	0.069	0.062	0.022
				2	0.414	0.119	0.389	0.384	0.291	0.128
			0.3	1.5	0.234	0.079	0.203	0.213	0.206	0.050
				2	0.830	0.408	0.812	0.792	0.719	0.403
	med	100	0.1	1.5	0.074	0.017	0.003	0.069	0.069	0.010
				2	0.433	0.115	0.099	0.415	0.336	0.121
			0.3	1.5	0.204	0.079	0.016	0.190	0.204	0.031
				2	0.841	0.438	0.331	0.825	0.745	0.429
		500	0.1	1.5	0.045	0.017	0	0.045	0.058	0.004
				2	0.332	0.114	0.008	0.336	0.318	0.089
			0.3	1.5	0.158	0.080	0.003	0.132	0.193	0.022
				2	0.768	0.437	0.055	0.746	0.727	0.365
	high	100	0.1	1.5	0.072	0.021	0	0.061	0.058	0.006
				2	0.471	0.114	0.001	0.447	0.355	0.109
			0.3	1.5	0.258	0.080	0	0.219	0.202	0.033
				2	0.898	0.445	0.012	0.861	0.767	0.442
		500	0.1	1.5	0.059	0.019	0	0.054	0.057	0.003
				2	0.414	0.116	0	0.428	0.345	0.079
			0.3	1.5	0.218	0.079	0	0.194	0.201	0.021
				2	0.867	0.446	0	0.836	0.763	0.376
0.3	low	100	0.1	1.5	0.199	0.070	0.178	0.177	0.162	0.418
				2	0.823	0.411	0.814	0.792	0.720	0.538
			0.3	1.5	0.645	0.264	0.622	0.599	0.550	0.463
				2	0.996	0.858	0.996	0.976	0.961	0.889
		500	0.1	1.5	0.169	0.072	0.068	0.150	0.150	0.031
				2	0.794	0.411	0.607	0.765	0.710	0.316
			0.3	1.5	0.597	0.264	0.363	0.562	0.530	0.187
				2	0.991	0.858	0.976	0.975	0.960	0.843
	med	100	0.1	1.5	0.226	0.059	0	0.206	0.190	0.033
				2	0.812	0.365	0.009	0.771	0.681	0.343
			0.3	1.5	0.646	0.266	0.001	0.599	0.556	0.156
				2	0.998	0.856	0.301	0.978	0.965	0.865
		500	0.1	1.5	0.200	0.060	0	0.198	0.183	0.021
				2	0.762	0.364	0	0.759	0.677	0.291
			0.3	1.5	0.610	0.263	0	0.586	0.551	0.113
				2	0.998	0.855	0.025	0.978	0.964	0.857
	high	100	0.1	1.5	0.231	0.074	0	0.202	0.179	0.042
				2	0.833	0.393	0	0.783	0.705	0.324
			0.3	1.5	0.653	0.277	0	0.611	0.539	0.170
				2	0.993	0.864	0.001	0.977	0.961	0.852
		500	0.1	1.5	0.219	0.074	0	0.201	0.177	0.023
				2	0.815	0.393	0	0.781	0.700	0.274
			0.3	1.5	0.653	0.279	0	0.609	0.535	0.127
				2	0.993	0.866	0	0.976	0.961	0.838

p_e = frequency of environmental factor, $OR_{\mathrm{G-E}}$ = strength of G-E association effect, $N_{\mathrm{G-E}}$ = number of G-E association effects, p_g = frequency of genetic factor, OR_{GxE} interaction effect, EHB = empirical hierarchical Bayes, CC = case-control, CASES = case-only,TWO = intuitive two-step, MUK = Mukherjee, MUR = Murcray

these situations. However, given 100 low associations MUR reaches nearly 100% rank power. This is the case even for a very small effect of $OR_{\text{GxE}} = 1.5$ and $p_g = p_e = 0.1$. Hence, in this particular situation, MUR is clearly superior to all other methods. When we take a look at the corresponding results given a higher prevalence, we observe this high rank power of MUR for $p_e = 0.1$ as well. Given a more frequent environmental factor of $p_e = 0.3$, this effect diminishes and EHB reaches better rank power in most situations (results not shown).
Note, the corresponding conventional power for case-control test of GxE interaction for the situations shown in table 6.4 is less than 36%. The power of MUR is at most approximately 65%. The case-only test has only for $OR_{\text{GxE}} = 2$ with $p_g = p_e = 0.3$ a power of 91-93%. For all other parameter combinations, genome-wide significance is rarely reached. In reality, low power is commonly encountered. Therefore, there is much higher practical value to detect interacting SNPs for follow-up using the top 25 markers.
When we take a look at table 6.5 presenting the corresponding results based on 1,000 cases and 2,000 controls and 6.6 for 2,000 cases and 1,000 controls , we observe the same behavioral trends of the CC, MUR and CASES method with respect to EHB. Given a high number or high effect of G-E association, CASES even performs worse than case-control in most situations. For TWO and MUK we see again that they generally show a little bit less rank power than EHB. However, this time in some situations we see a really small superiority.
As expected, having an unbalanced number of cases and controls, EHB shows its advantage compared to EHB2. With increasing number and strength of associations, EHB reaches higher rank power than EHB2. While for 1:2 ratio, it makes only some percent given $p_e = 0.1$, the difference increases up to 16% for $p_e = 0.3$ (table 6.6). For 2:1 ratio, between 10 and 25 % increase in rank power of the EHB compared to EHB2 is observed several times, with a maximum of nearly 50% rank power difference.
To make sure that all other simulation settings with $p_e = 0.5, p_g = 0.5, OR_{\text{GxE}} = 1.2, 2.5, 3$ and prevalence of 5% and 10% behave to the same rules than seen for the scenarios picked out, we plotted for each method the rank power of the EHB ranking against the difference between the ranking power of EHB and other method (ranking power EHB - ranking power other method) on the y-axis. Hence, positive values on the y-axis represent a rank power improve by the EHB method. The corresponding plots can be seen in figures 6.6-6.11. The different points represent all different simulated scenarios according to disease prevalence, frequency of environmental and genetic factor and OR of the GxE interaction. The results are presented separately for different case-control ratios and different numbers and strength of association are distinguishable by color and point symbol. Furthermore, we did not only consider the ranking power of the top 25, but for the top 1, 10 and 25 in the different rows.
In figure 6.6 comparing the rank power of EHB and EHB2, we see that for 1,500 cases and 1,500 controls and the medium association situation, sometimes EHB detects the interacting SNP more often on its top rank, sometimes the $EHB2$. For low associations, nearly no difference is observed, for the high association situation, EHB is superior. Considering more of the top SNPs, these tendencies diminish - given a high similarity of both within a range of +/- 5%. For the unbalanced case control samples, we see no difference between the methods with respect to the low association

Table 6.5: *Power to detect a GxE interacting SNP in the top 25 ranking in the presence of population based G-E associations. The underlying disease is assumed to have a prevalence of 1%. The given sample consists of* ***1,000 cases and 2,000 controls****.*

p_e	OR_{G-E}	N_{G-E}	p_g	OR_{GxE}	EHB	EHB2	CC	CASES	TWO	MUK	MUR
0.1	low	100	0.1	1.5	0.068	0.066	0.031	0.068	0.059	0.059	0.996
				2	0.315	0.311	0.138	0.312	0.277	0.249	0.995
			0.3	1.5	0.158	0.16	0.094	0.157	0.148	0.150	0.997
				2	0.704	0.702	0.396	0.702	0.649	0.633	0.997
		500	0.1	1.5	0.063	0.066	0.031	0.058	0.056	0.059	0.010
				2	0.301	0.301	0.136	0.286	0.273	0.248	0.050
			0.3	1.5	0.146	0.149	0.093	0.135	0.141	0.144	0.013
				2	0.685	0.688	0.397	0.645	0.636	0.623	0.135
	med	100	0.1	1.5	0.069	0.065	0.033	0.020	0.057	0.062	0.002
				2	0.311	0.311	0.135	0.132	0.293	0.258	0.042
			0.3	1.5	0.139	0.138	0.071	0.038	0.134	0.141	0.011
				2	0.641	0.642	0.393	0.310	0.611	0.594	0.104
		500	0.1	1.5	0.054	0.053	0.034	0	0.053	0.059	0.001
				2	0.265	0.264	0.135	0.016	0.274	0.250	0.029
			0.3	1.5	0.136	0.122	0.072	0.004	0.120	0.137	0.011
				2	0.606	0.584	0.394	0.065	0.591	0.582	0.082
	high	100	0.1	1.5	0.066	0.062	0.033	0.002	0.064	0.052	0.005
				2	0.300	0.299	0.126	0.011	0.287	0.256	0.037
			0.3	1.5	0.156	0.148	0.079	0	0.133	0.141	0.008
				2	0.685	0.676	0.412	0.043	0.642	0.608	0.130
		500	0.1	1.5	0.064	0.050	0.033	0	0.063	0.051	0.004
				2	0.287	0.252	0.129	0	0.284	0.248	0.024
			0.3	1.5	0.149	0.115	0.078	0	0.131	0.139	0.003
				2	0.679	0.622	0.400	0	0.639	0.602	0.105
0.3	low	100	0.1	1.5	0.111	0.112	0.073	0.105	0.105	0.107	0.369
				2	0.595	0.598	0.350	0.579	0.546	0.539	0.385
			0.3	1.5	0.391	0.396	0.223	0.378	0.348	0.352	0.384
				2	0.968	0.968	0.827	0.962	0.939	0.930	0.582
		500	0.1	1.5	0.105	0.101	0.072	0.058	0.099	0.106	0.010
				2	0.571	0.570	0.345	0.446	0.533	0.535	0.077
			0.3	1.5	0.367	0.371	0.221	0.261	0.334	0.346	0.047
				2	0.958	0.959	0.828	0.929	0.933	0.930	0.352
	med	100	0.1	1.5	0.129	0.123	0.054	0	0.114	0.119	0.007
				2	0.597	0.584	0.347	0.019	0.557	0.546	0.090
			0.3	1.5	0.449	0.432	0.234	0.005	0.400	0.400	0.043
				2	0.968	0.963	0.823	0.295	0.943	0.949	0.336
		500	0.1	1.5	0.118	0.082	0.052	0	0.111	0.116	0.002
				2	0.582	0.533	0.345	0	0.553	0.541	0.074
			0.3	1.5	0.444	0.361	0.234	0	0.398	0.393	0.032
				2	0.965	0.947	0.823	0.036	0.942	0.949	0.331
	high	100	0.1	1.5	0.124	0.099	0.072	0	0.107	0.113	0.004
				2	0.571	0.536	0.349	0	0.539	0.534	0.093
			0.3	1.5	0.430	0.391	0.214	0	0.380	0.387	0.042
				2	0.967	0.958	0.798	0	0.936	0.930	0.360
		500	0.1	1.5	0.122	0.054	0.073	0	0.105	0.110	0.003
				2	0.565	0.399	0.349	0	0.539	0.532	0.087
			0.3	1.5	0.428	0.268	0.213	0	0.380	0.384	0.032
				2	0.965	0.925	0.797	0	0.934	0.926	0.350

p_e = frequency of environmental factor, OR_{G-E} = strength of G-E association effect,
N_{G-E} = number of G-E association effects, p_g = frequency of genetic factor,
OR_{GxE} interaction effect, EHB = empirical hierarchical Bayes based on regression coefficient,
EHB2 = empirical hierarchical Bayes based on test statistic, CC = case-control, CASES = case-only,
TWO = intuitive two-step, MUK = Mukherjee, MUR = Murcray

Table 6.6: *Power to detect a GxE interacting SNP in the top 25 ranking in the presence of population based G-E associations. The underlying disease is assumed to have a prevalence of 1%. The given sample consists of* ***2,000 cases and 1,000 controls****.*

p_e	OR_{G-E}	N_{G-E}	p_g	OR_{GxE}	EHB_Z	EHB	CC	CASES	TWO	MUK	MUR
0.1	low	100	0.1	1.5	0.133	0.133	0.010	0.133	0.108	0.057	0.994
				2	0.641	0.641	0.055	0.641	0.583	0.353	0.99
			0.3	1.5	0.382	0.385	0.061	0.383	0.34	0.227	0.994
				2	0.970	0.971	0.325	0.971	0.939	0.721	0.997
		500	0.1	1.5	0.096	0.090	0.010	0.084	0.081	0.054	0.043
				2	0.58	0.569	0.056	0.553	0.546	0.339	0.336
			0.3	1.5	0.299	0.286	0.061	0.272	0.289	0.214	0.144
				2	0.947	0.942	0.325	0.936	0.926	0.701	0.761
	med	100	0.1	1.5	0.083	0.067	0.005	0.002	0.084	0.064	0.03
				2	0.478	0.452	0.066	0.076	0.504	0.328	0.26
			0.3	1.5	0.222	0.181	0.059	0.007	0.246	0.219	0.096
				2	0.906	0.882	0.340	0.347	0.913	0.731	0.729
		500	0.1	1.5	0.024	0.021	0.005	0	0.009	0.057	0.012
				2	0.234	0.207	0.067	0.008	0.161	0.304	0.16
			0.3	1.5	0.101	0.047	0.059	0	0.030	0.187	0.047
				2	0.728	0.638	0.337	0.052	0.544	0.699	0.592
	high	100	0.1	1.5	0.098	0.077	0.01	0	0.103	0.061	0.019
				2	0.555	0.497	0.061	0.001	0.556	0.353	0.28
			0.3	1.5	0.328	0.224	0.061	0	0.314	0.221	0.099
				2	0.928	0.864	0.313	0.008	0.909	0.707	0.684
		500	0.1	1.5	0.044	0.018	0.008	0	0.055	0.059	0.009
				2	0.390	0.241	0.060	0	0.422	0.342	0.165
			0.3	1.5	0.209	0.054	0.060	0	0.184	0.214	0.047
				2	0.826	0.566	0.313	0	0.8	0.693	0.556
0.3	low	100	0.1	1.5	0.338	0.333	0.062	0.321	0.293	0.219	0.424
				2	0.926	0.917	0.336	0.913	0.883	0.704	0.793
			0.3	1.5	0.760	0.763	0.236	0.753	0.734	0.583	0.631
				2	1	1	0.801	1	0.981	0.931	0.996
		500	0.1	1.5	0.250	0.211	0.063	0.079	0.216	0.197	0.087
				2	0.876	0.844	0.333	0.658	0.835	0.679	0.641
			0.3	1.5	0.667	0.626	0.236	0.387	0.642	0.559	0.416
				2	0.998	0.999	0.799	0.998	0.979	0.926	0.993
	med	100	0.1	1.5	0.298	0.254	0.046	0	0.286	0.213	0.089
				2	0.876	0.856	0.332	0.006	0.883	0.692	0.624
			0.3	1.5	0.757	0.686	0.217	0	0.721	0.539	0.401
				2	0.994	0.998	0.790	0.281	0.978	0.932	0.988
		500	0.1	1.5	0.182	0.087	0.046	0	0.211	0.203	0.045
				2	0.759	0.659	0.330	0	0.823	0.683	0.494
			0.3	1.5	0.667	0.425	0.214	0	0.644	0.524	0.276
				2	0.988	0.985	0.788	0.016	0.978	0.927	0.967
	high	100	0.1	1.5	0.336	0.238	0.062	0	0.294	0.21	0.068
				2	0.917	0.871	0.304	0	0.879	0.683	0.626
			0.3	1.5	0.807	0.689	0.211	0	0.749	0.558	0.383
				2	1	1	0.808	0.001	0.976	0.941	0.984
		500	0.1	1.5	0.291	0.042	0.060	0	0.268	0.205	0.031
				2	0.858	0.592	0.302	0	0.871	0.675	0.480
			0.3	1.5	0.781	0.293	0.207	0	0.744	0.549	0.269
				2	0.998	0.986	0.807	0	0.976	0.941	0.960

p_e = frequency of environmental factor, OR_{G-E} = strength of G-E association effect,
N_{G-E} = number of G-E association effects, p_g = frequency of genetic factor,
OR_{GxE} interaction effect, EHB = empirical hierarchical Bayes based on regression coefficient,
EHB2 = empirical hierarchical Bayes based on test statistic, CC = case-control, CASES = case-only,
TWO = intuitive two-step, MUK = Mukherjee, MUR = Murcray

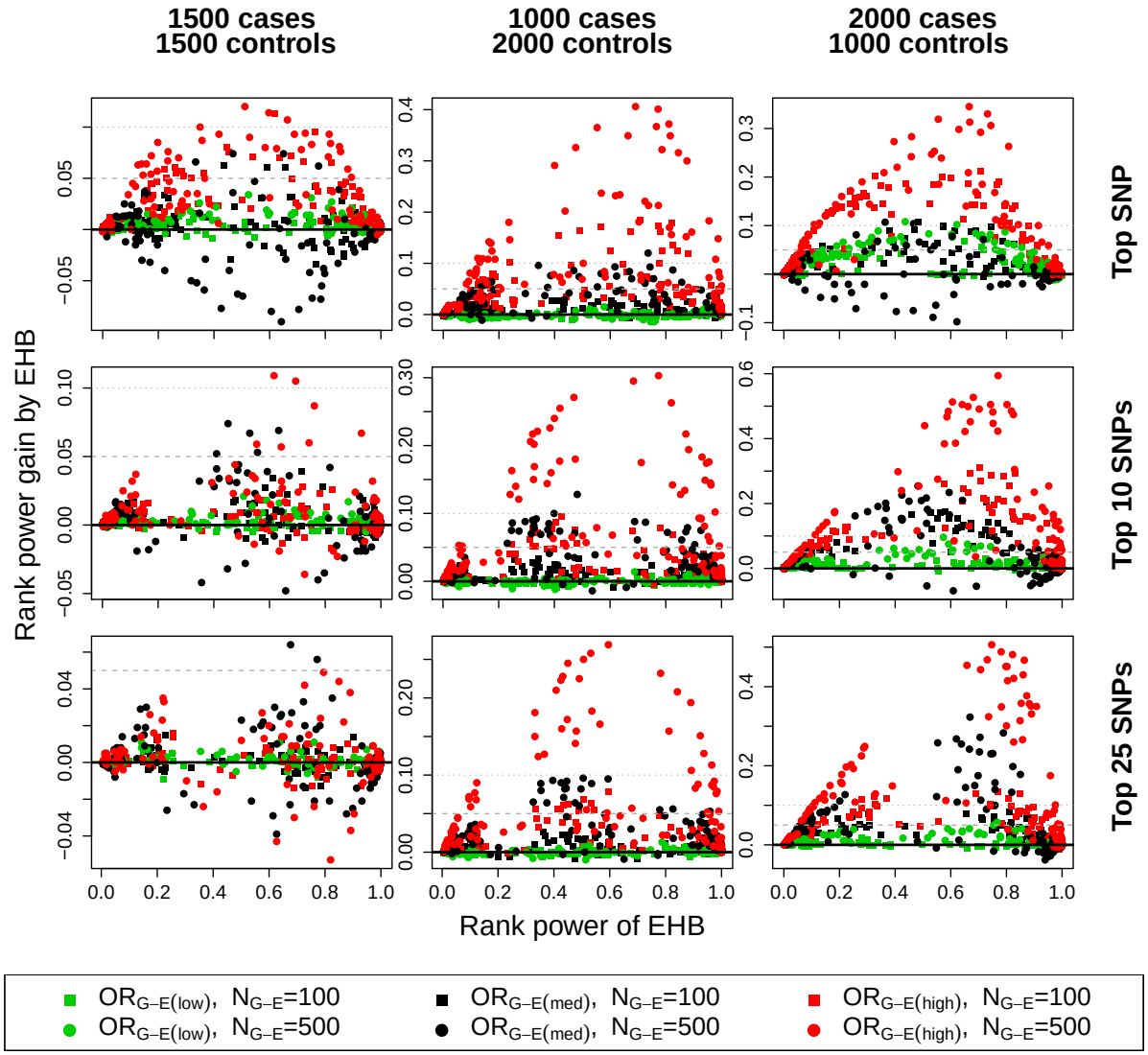


Figure 6.6: *Comparison of the rank power to detect a GxE interacting marker in the top ranking positions between* ***EHB and EHB2****. On the x-axis, the EHB ranking power is plotted against the difference between the ranking power of EHB and EHB2 (EHB - EHB2) on the y-axis. Hence, positive values on the y-axis represent a rank power improve of the EHB method. The different points represent all different simulated scenarios according to disease prevalence, frequency of environmental and genetic factor and OR of the GxE interaction. The different ratios of cases and controls are represented in the different columns, with first column 1,500 cases and 1,500 controls, second column 1,000 cases and 2,000 controls, third column 2,000 cases and 1,000 controls. In the upper row, the ranking power considering only the first rank is given, in the middle row with respect to the top 10, in the lower row with respect to the top 25.*

strength, for the high one however EHB is highly favored with up to 60% more rank power. Hence, the EHB has an enormous improvement compared to the EHB2. For the medium association, we see a slight advantage of up to 10 % in most cases, only for top 1 of 2,000:1,000 and 500 G-E associations $EHB2$ seems to be slightly better.
Taking a look at the CC plot 6.7, we see that EHB has more rank power for nearly all situations, even when only looking at the top ranking SNP. The largest differences can be seen for 2,000:1,000, followed by 1,500:1,500. In particular for the 2,000:1,000 samples, we see groupings of the points representing the same number and strength of association.
The superiority of the EHB compared to CASES observed before, can be confirmed by figure 6.8. In that figure comparing the rank power of EHB and CASES, a clear triangular structure can be observed. The diagonal is build by the high association scenarios, representing that independently of the choice of further parameters, CASES reaches nearly no rank power in these situations. The horizontal red line represents the similarity of CASES to EHB given a low number and strength of G-E associations. The vertical line on the right represents situations where EHB reaches nearly 100% rank power and CASES any number between 0 and 100%. This effects particular scenarios of any association situation.
For the top 1 rank power of TWO in figure 6.9, we see no clear preference for EHB or TWO. Both methods show situations reaching higher power than the other one. However, in practice usually more than only a hand full of top SNPs from a GWAS scan are selected. When we increase the number of selected top markers, an overall advantage of EHB emerges.
Taking a look at the top 1 rank power between EHB and MUK in figure 6.10, EHB performs better for the low association situation, while MUK has up to 30% more rank power for medium association and part of high association. With increasing number of top ranks considered, the superiority of MUK decreases clearly. For the top 25, EHB is remarkably better.
Finally, for MUR in figure 6.11 we see that given 2,000 cases and 1,000 controls, the interacting SNP is up to 50% more often on the top position for MUR than for EHB in the medium association situation and some scenarios of high association. With increasing number of top SNPs considered this effect disappears and reverse to a benefit of the EHB. For the top 25 SNPs, the situations with 100 low associations profits clearly in using the MUR. For a higher number of stronger association effects however, EHB reaches up to 50% more rank power. Having 1,000 cases and 2,000 controls, MUR has less rank power than EHB for nearly all situations when looking at no more than the top 10 ranks. When the top 25 are considered we see again the 100 low association situations showing clearly higher rank power of nearly 100% with MUR. In the top 25% of the 1,500 cases and 1,500 controls situations, we see the same pattern. Here, we had a tendency of some stronger association situations to higher MUR rank power within the top 1 as well. However, the effect was much lower than observed for 2,000 cases and 1,000 controls.
Taking a look at the corresponding plot for 200 and 1,000 population based G-E associations (results not shown), we see the same overall behavior between the EHB and the different methods as in our shown plots. The same is true looking at the top 50 and top 100 markers (data not shown).

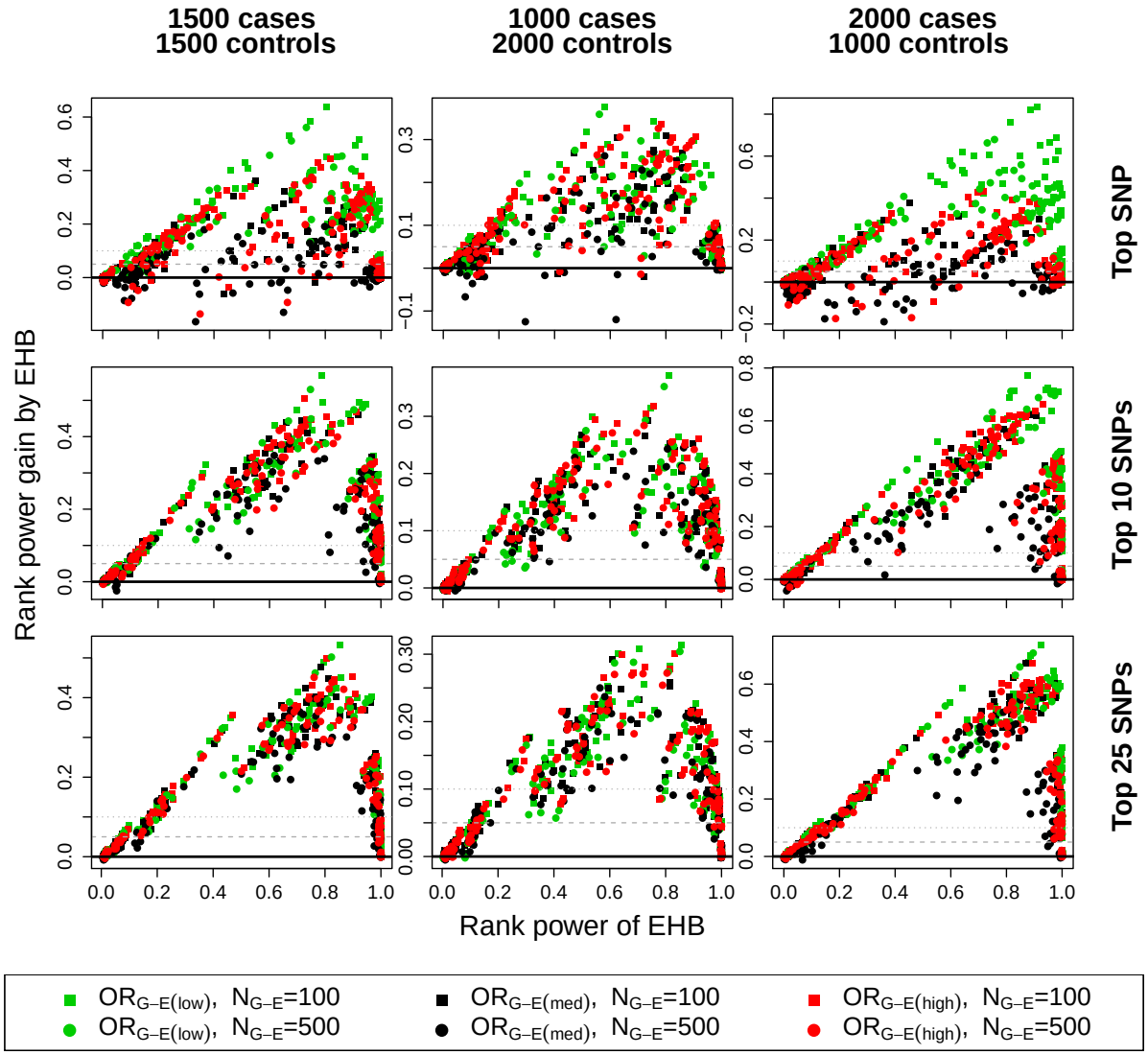


Figure 6.7: *Comparison of the power to detect a GxE interacting marker in the top ranking positions between* ***EHB and case-control test****.On the x-axis, the EHB ranking power is plotted against the difference between the ranking power of EHB and CC (EHB - CC) on the y-axis. Hence, positive values on the y-axis represent a power improve by the EHB method. The different points represent all different simulated scenarios according to disease prevalence, frequency of environmental and genetic factor and OR of the GxE interaction. The different ratios of cases and controls are represented in the different columns, with first column 1,500 cases and 1,500 controls, second column 1,000 cases and 2,000 controls; third column 2,000 cases and 1,000 controls. In the upper row, the ranking power considering only the first rank is given, in the middle row with respect to the top 10, in the lower row with respect to the top 25.*

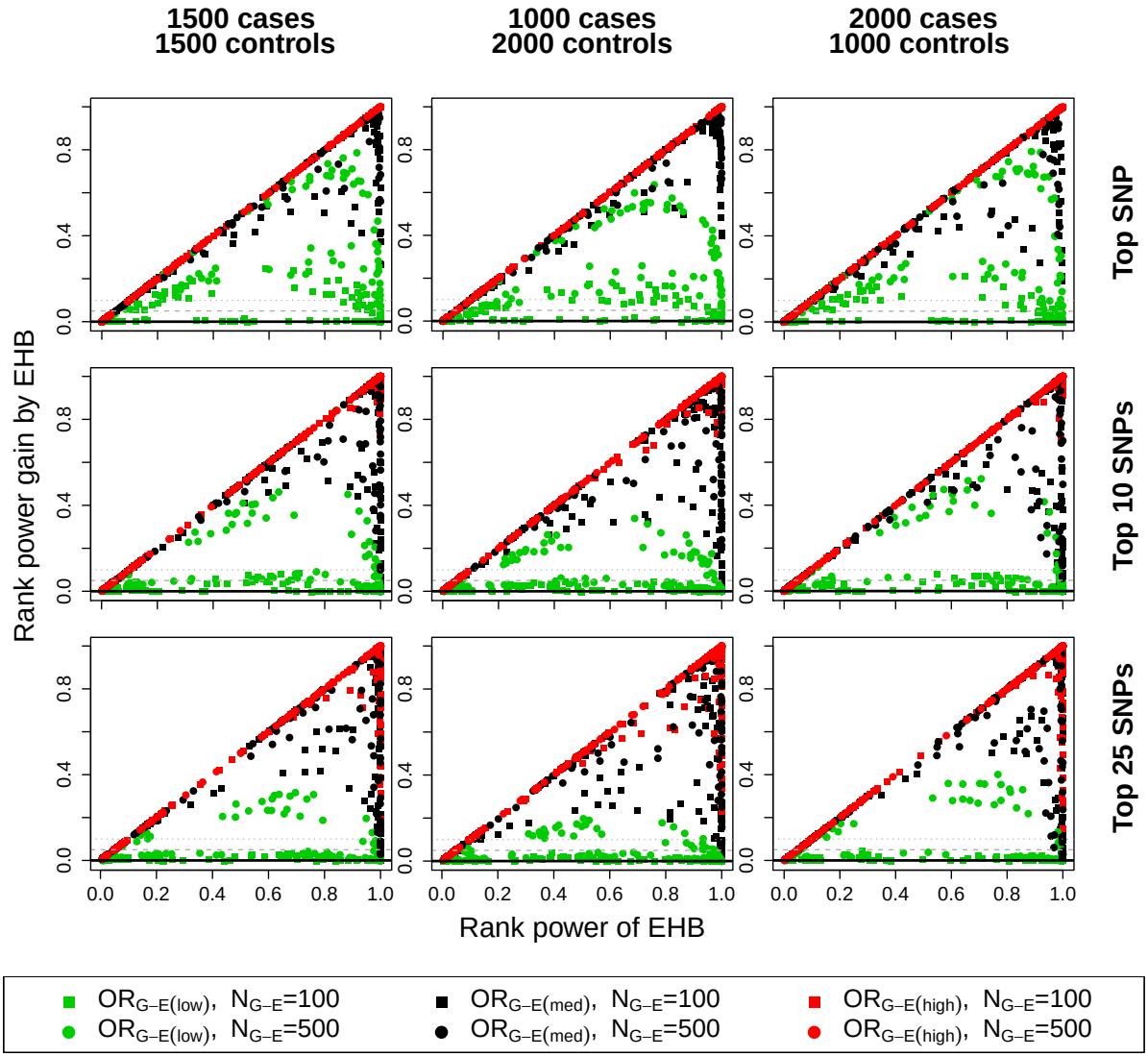


Figure 6.8: *Comparison of the power to detect a GxE interacting marker in the top ranking positions between* ***EHB and case-only test****. On the x-axis, the EHB ranking power is plotted against the difference between the ranking power of EHB and case-only (EHB - CASES) on the y-axis. Hence, positive values on the y-axis represent a power improve by the EHB method. The different points represent all different simulated scenarios according to disease prevalence, frequency of environmental and genetic factor and OR of the GxE interaction. The different ratios of cases and controls are represented in the different columns, with first column 1,500 cases and 1,500 controls, second column 1,000 cases and 2,000 controls; third column 2,000 cases and 1,000 controls. In the upper row, the ranking power considering only the first rank is given, in the middle row with respect to the top 10, in the lower row with respect to the top 25.*

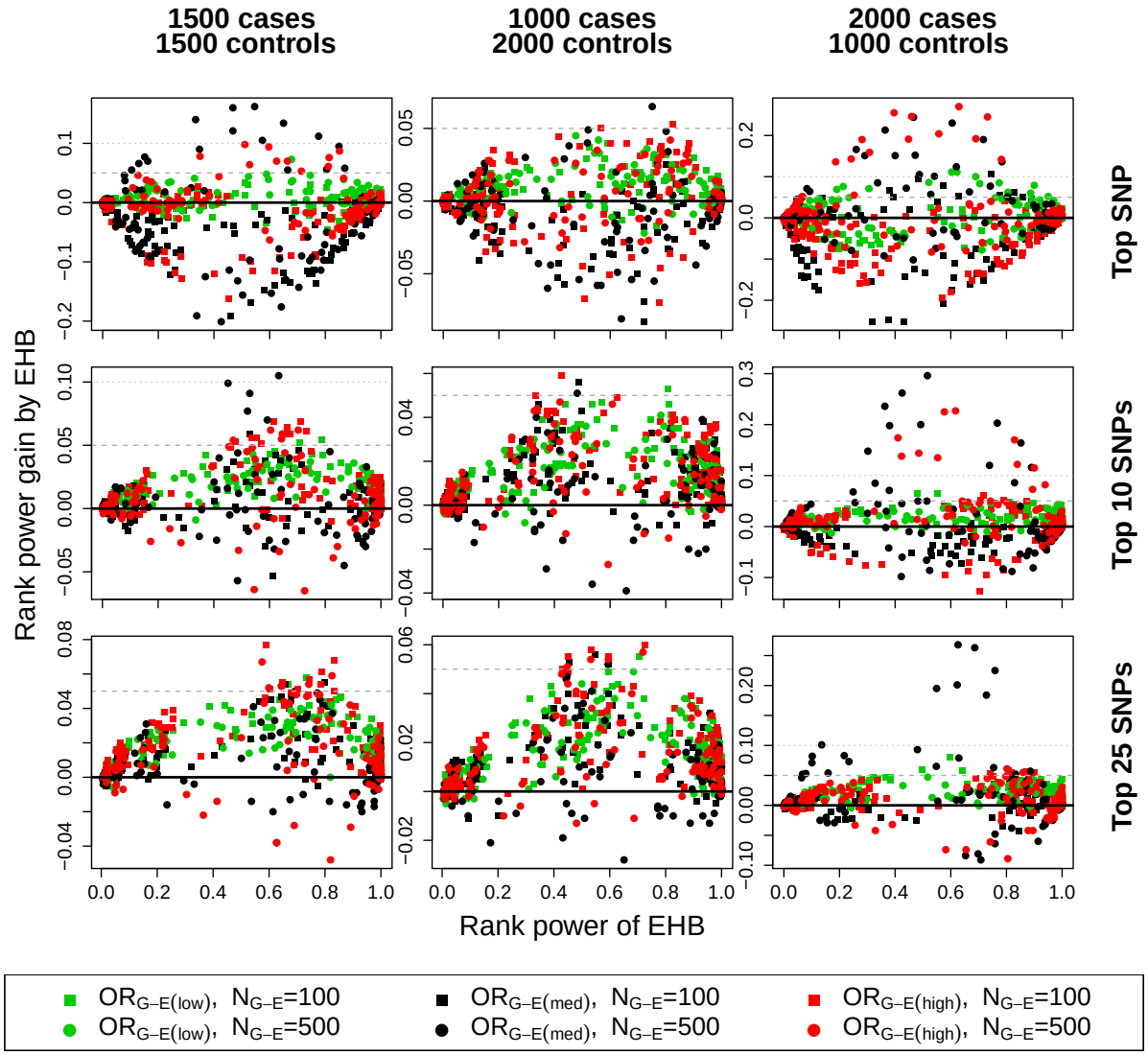


Figure 6.9: *Comparison of the power to detect a GxE interacting marker in the top ranking positions between* ***EHB and simple two-step method****. On the x-axis, the EHB ranking power is plotted against the difference between the ranking power of EHB and TWO (EHB - TWO) on the y-axis. Hence, positive values on the y-axis represent a power improve by the EHB method. The different points represent all different simulated scenarios according to disease prevalence, frequency of environmental and genetic factor and OR of the GxE interaction. The different ratios of cases and controls are represented in the different columns, with first column 1,500 cases and 1,500 controls, second column 1,000 cases and 2,000 controls; third column 2,000 cases and 1,000 controls. In the upper row, the ranking power considering only the first rank is given, in the middle row with respect to the top 10, in the lower row with respect to the top 25.*

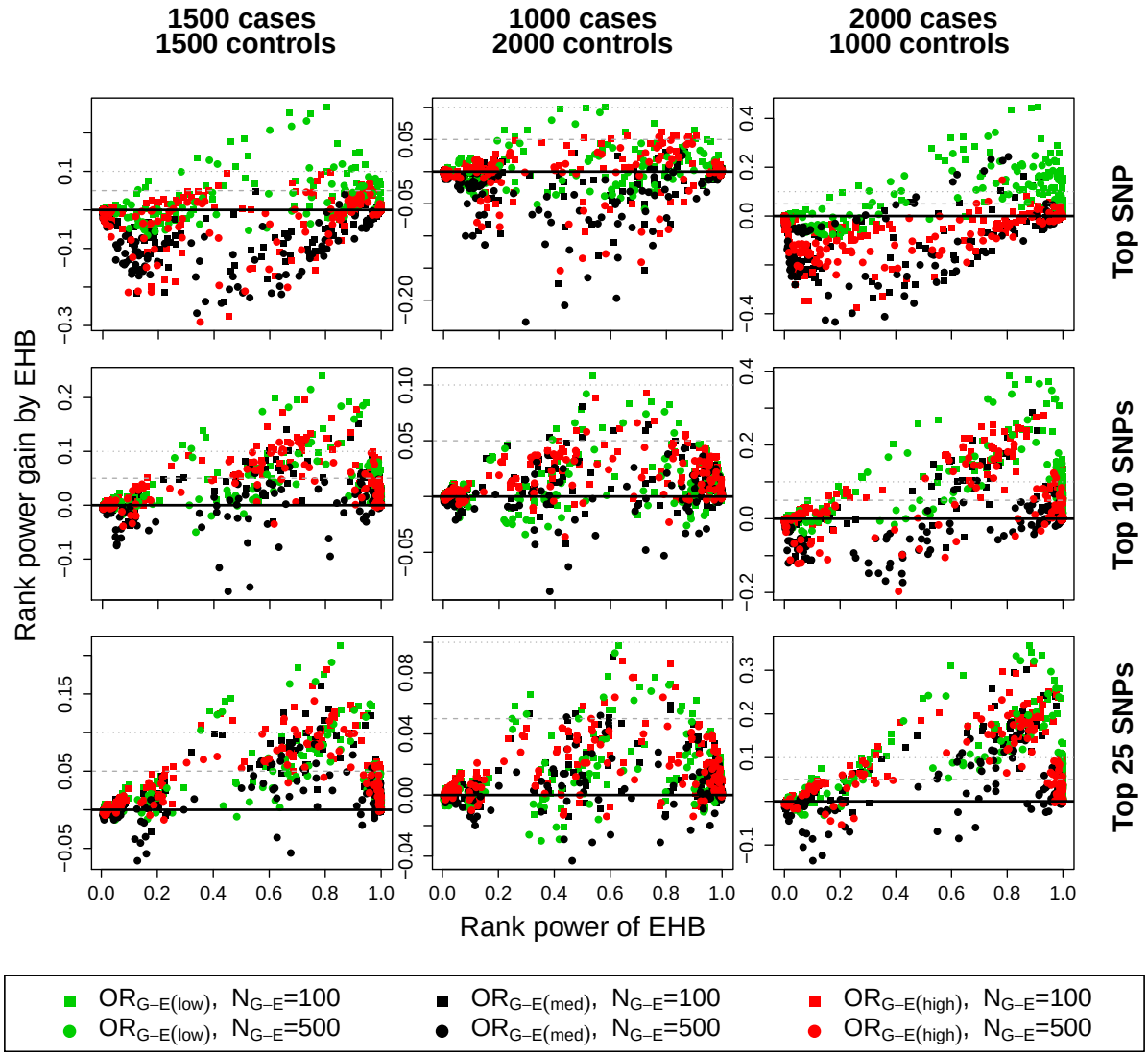


Figure 6.10: *Comparison of the power to detect a GxE interacting marker in the top ranking positions between* ***EHB and Mukherjee's method****. On the x-axis, the EHB ranking power is plotted against the difference between the ranking power of EHB and MUK (EHB - MUK) on the y-axis. Hence, positive values on the y-axis represent a power improve by the EHB method. The different points represent all different simulated scenarios according to disease prevalence, frequency of environmental and genetic factor and OR of the GxE interaction. The different ratios of cases and controls are represented in the different columns, with first column 1,500 cases and 1,500 controls, second column 1,000 cases and 2,000 controls; third column 2,000 cases and 1,000 controls. In the upper row, the ranking power considering only the first rank is given, in the middle row with respect to the top 10, in the lower row with respect to the top 25.*

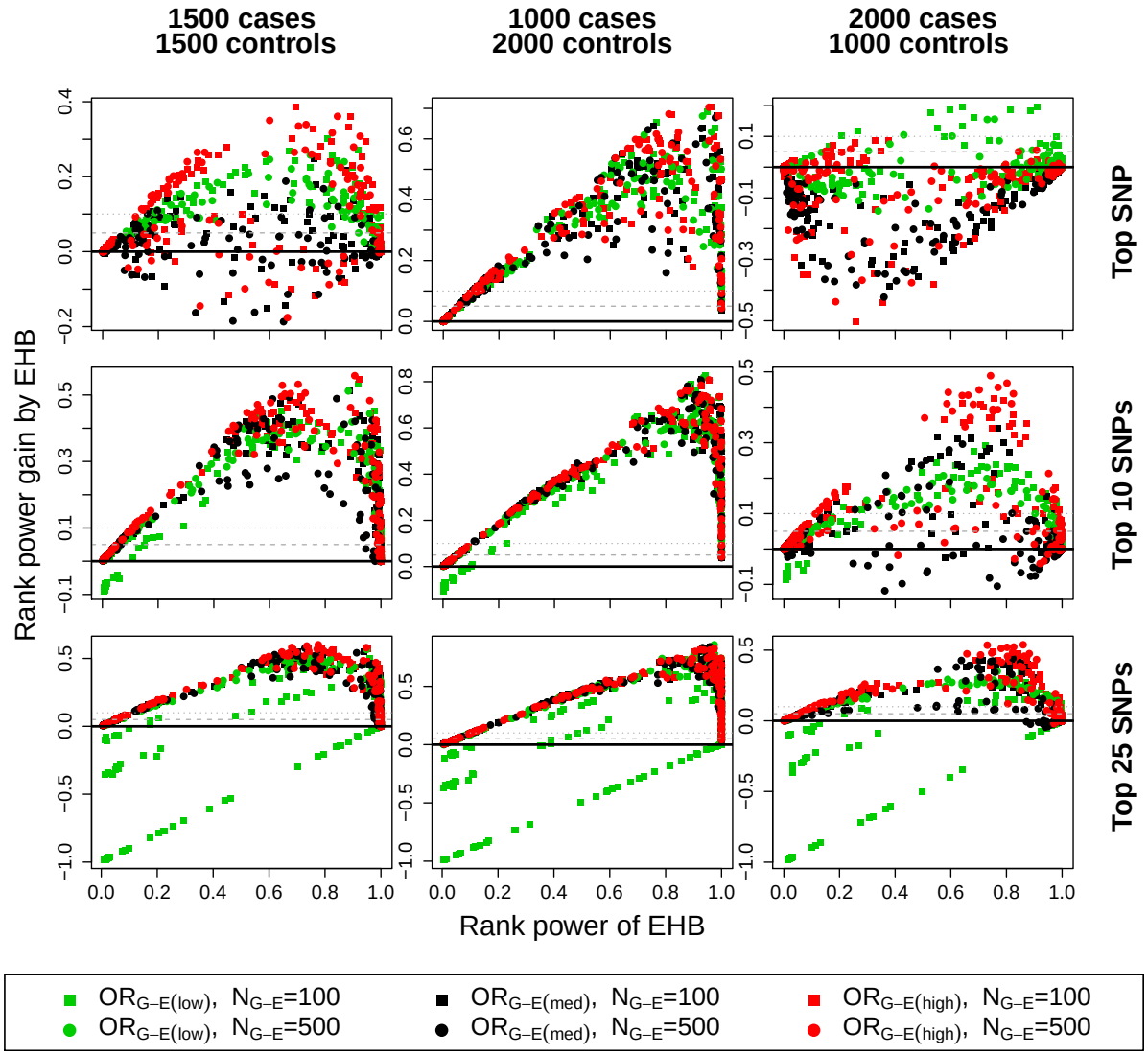


Figure 6.11: *Comparison of the power to detect a GxE interacting marker in the top ranking positions between* ***EHB and Murcray's method****. On the x-axis, the EHB ranking power is plotted against the difference between the ranking power of EHB and CC (EHB - MUR) on the y-axis. Hence, positive values on the y-axis represent a power improve by the EHB method. The different points represent all different simulated scenarios according to disease prevalence, frequency of environmental and genetic factor and OR of the GxE interaction. The different ratios of cases and controls are represented in the different columns, with first column 1,500 cases and 1,500 controls, second column 1,000 cases and 2,000 controls; third column 2,000 cases and 1,000 controls. In the upper row, the ranking power considering only the first rank is given, in the middle row with respect to the top 10, in the lower row with respect to the top 25.*

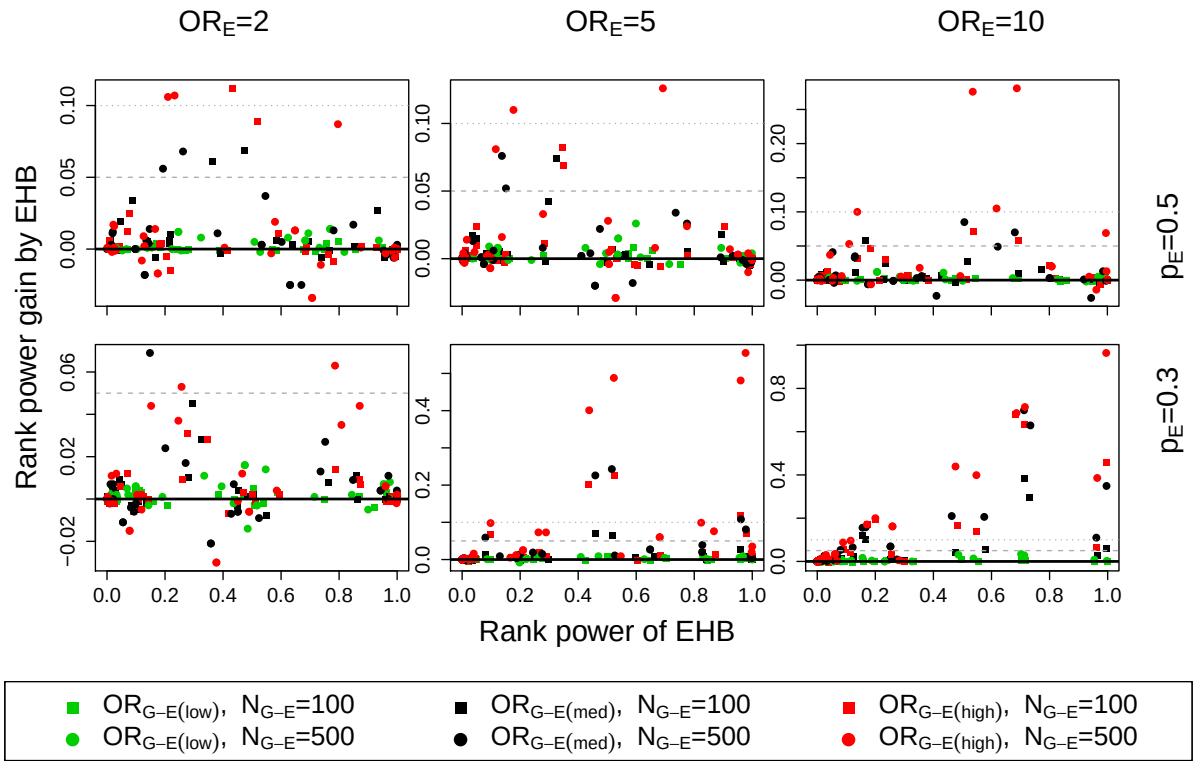


Figure 6.12: *Comparison of the power to detect a GxE interacting marker in the top 25 ranking positions between* ***EHB and EHB2*** *given an* ***environmental main effect****. On the x-axis, the EHB ranking power is plotted against the difference between the ranking power of EHB and EHB2 (EHB - EHB2) on the y-axis. Hence, positive values on the y-axis represent a power improve by the EHB method. The different points represent all different simulated scenarios according to disease prevalence, frequency of the genetic factor, OR of the GxE interaction and different case control ratios. The different environmental main effects are represented in the different columns, with first column* $OR_e = 2$*, second column* $OR_e = 5$*, third column* $OR_e = 10$*. In the upper row, the ranking power considering situation with* p_e*=0.3 is given, in lower row* $p_e = 0.5$*.*

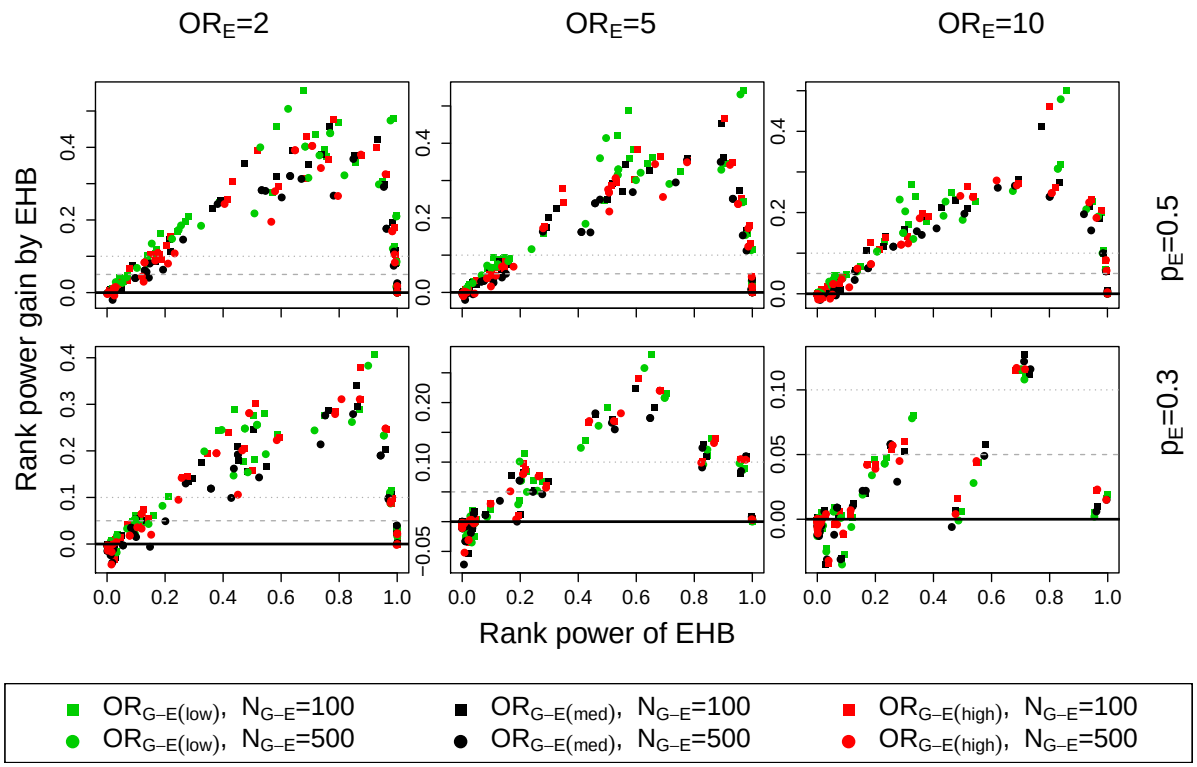


Figure 6.13: *Comparison of the power to detect a GxE interacting marker in the top 25 ranking positions between* ***EHB and case-control method*** *given an* ***environmental main effect****. On the x-axis, the EHB ranking power is plotted against the difference between the ranking power of EHB and CC (EHB - CC) on the y-axis. Hence, positive values on the y-axis represent a power improve by the EHB method. The different points represent all different simulated scenarios according to disease prevalence, frequency of the genetic factor, OR of the GxE interaction and different case control ratios. The different environmental main effects are represented in the different columns, with first column* $OR_e = 2$*, second column* $OR_e = 5$*, third column* $OR_e = 10$*. In the upper row, the ranking power considering situation with* p_e*=0.3 is given, in lower row* $p_e = 0.5$*.*

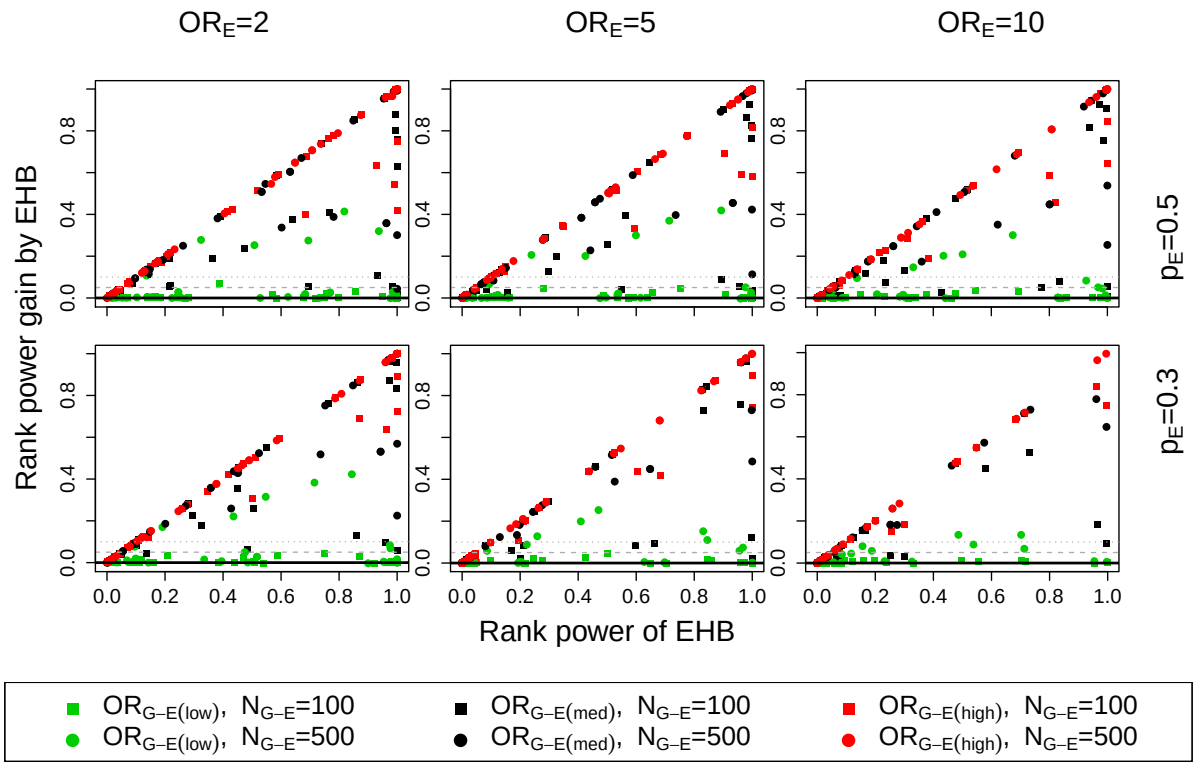


Figure 6.14: *Comparison of the power to detect a GxE interacting marker in the top 25 ranking positions between* ***EHB and case-only method*** *given an* ***environmental main effect****. On the x-axis, the EHB ranking power is plotted against the difference between the ranking power of EHB and case-only (EHB - CASES) on the y-axis. Hence, positive values on the y-axis represent a power improve by the EHB method. The different points represent all different simulated scenarios according to disease prevalence, frequency of the genetic factor, OR of the GxE interaction and different case control ratios. The different environmental main effects are represented in the different columns, with first column* $OR_e = 2$*, second column* $OR_e = 5$*, third column* $OR_e = 10$*. In the upper row, the ranking power considering situation with* p_e*=0.3 is given, in lower row* $p_e = 0.5$*.*

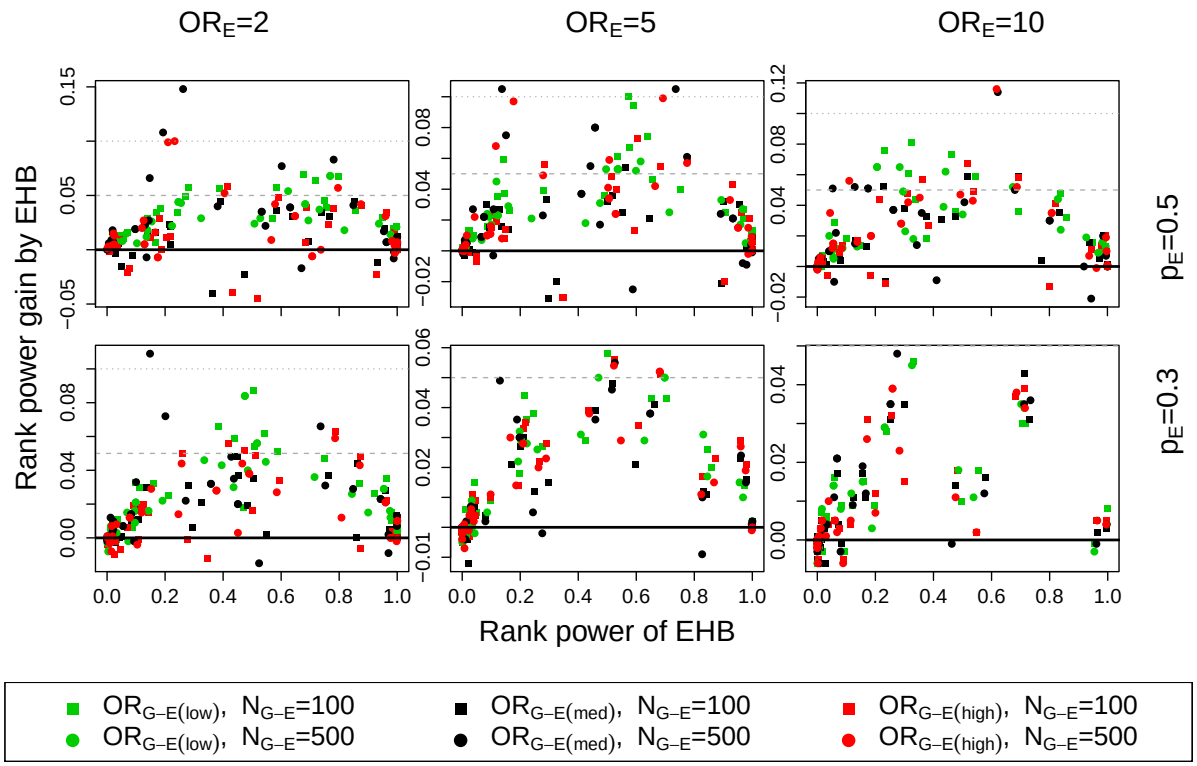


Figure 6.15: *Comparison of the power to detect a GxE interacting marker in the top 25 ranking positions between* ***EHB and simple two-step method*** *given an* ***environmental main effect****. On the x-axis, the EHB ranking power is plotted against the difference between the ranking power of EHB and TWO (EHB - TWO) on the y-axis. Hence, positive values on the y-axis represent a power improve by the EHB method. The different points represent all different simulated scenarios according to disease prevalence, frequency of the genetic factor, OR of the GxE interaction and different case control ratios. The different environmental main effects are represented in the different columns, with first column* $OR_e = 2$*, second column* $OR_e = 5$*, third column* $OR_e = 10$*. In the upper row, the ranking power considering situation with* p_e*=0.3 is given, in lower row* $p_e = 0.5$*.*

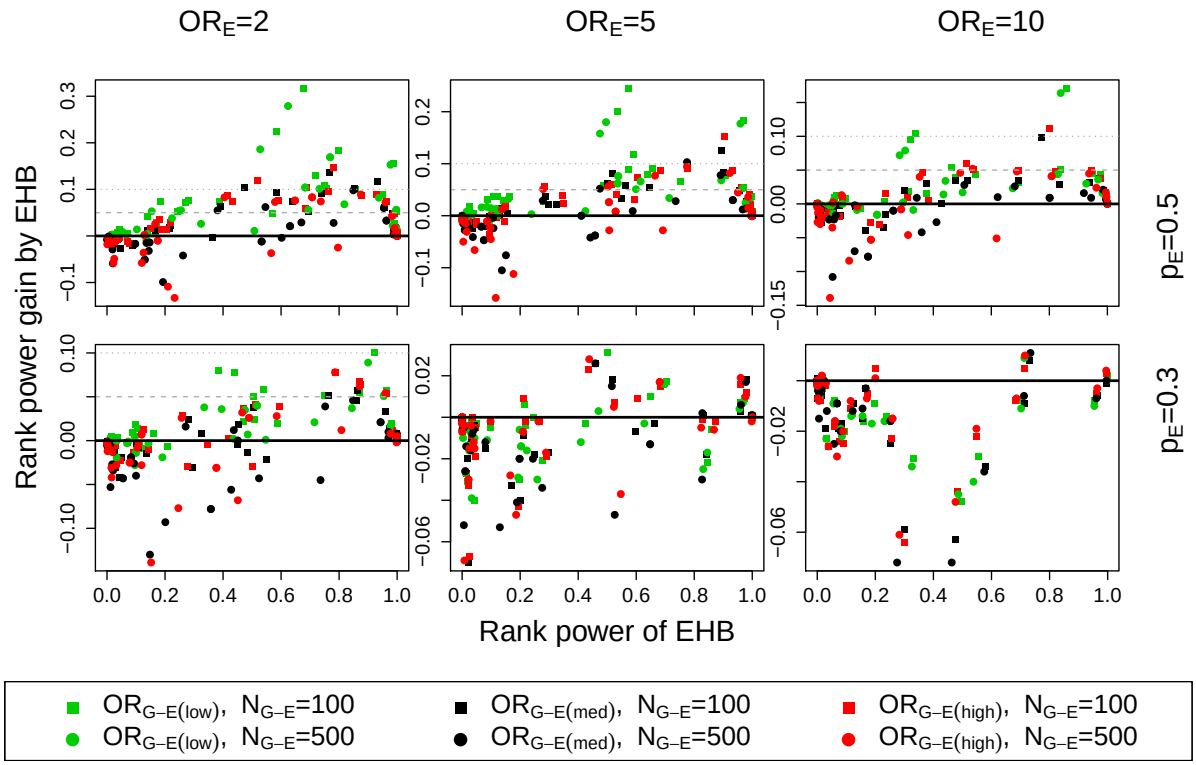


Figure 6.16: *Comparison of the power to detect a GxE interacting marker in the top 25 ranking positions between* ***EHB and Mukherjee's method*** *given an* ***environmental main effect****. On the x-axis, the EHB ranking power is plotted against the difference between the ranking power of EHB and MUK (EHB - MUK) on the y-axis. Hence, positive values on the y-axis represent a power improve by the EHB method. The different points represent all different simulated scenarios according to disease prevalence, frequency of the genetic factor, OR of the GxE interaction and different case control ratios. The different environmental main effects are represented in the different columns, with first column* $OR_e = 2$*, second column* $OR_e = 5$*, third column* $OR_e = 10$*. In the upper row, the ranking power considering situation with* p_e*=0.3 is given, in lower row* $p_e = 0.5$*.*

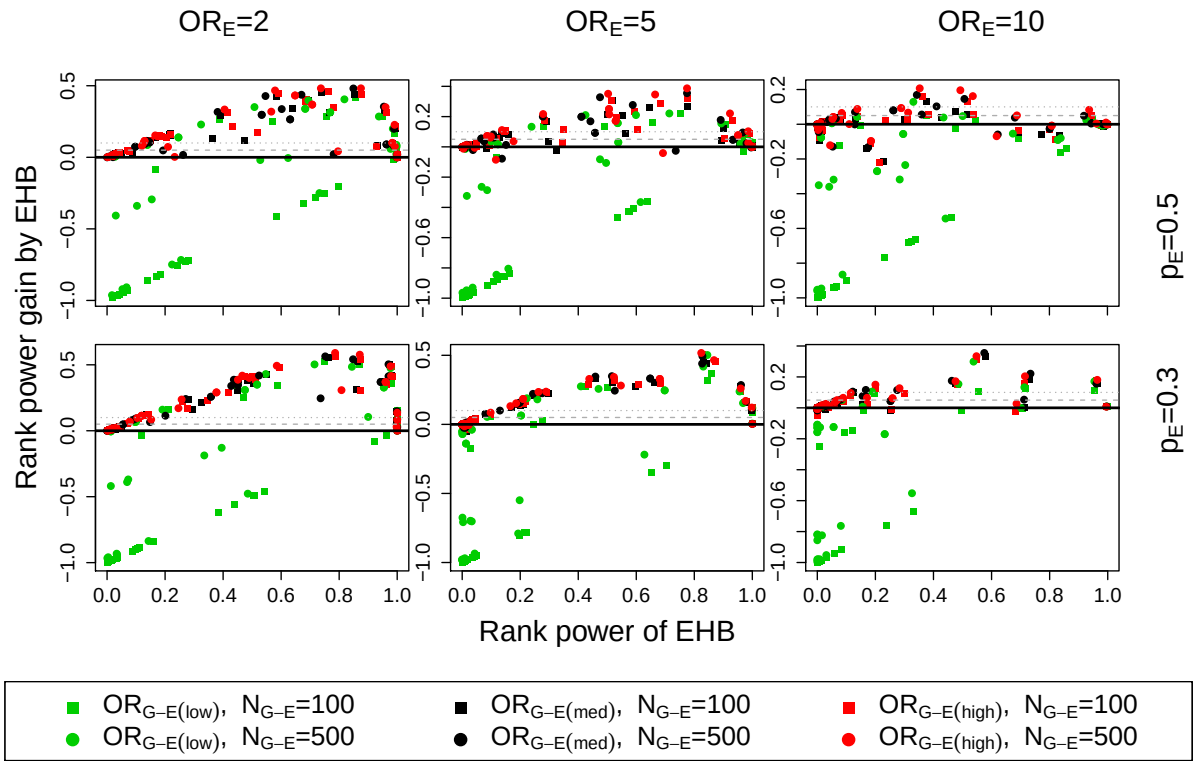


Figure 6.17: *Comparison of the power to detect a GxE interacting marker in the top 25 ranking positions between* ***EHB and Murcray's method*** *given an* ***environmental main effect****. On the x-axis, the EHB ranking power is plotted against the difference between the ranking power of EHB and CC (EHB - MUR) on the y-axis. Hence, positive values on the y-axis represent a power improve by the EHB method. The different points represent all different simulated scenarios according to disease prevalence, frequency of the genetic factor, OR of the GxE interaction and different case control ratios. The different environmental main effects are represented in the different columns, with first column* $OR_e = 2$*, second column* $OR_e = 5$*, third column* $OR_e = 10$*. In the upper row, the ranking power considering situation with* p_e*=0.3 is given, in lower row* $p_e = 0.5$*.*

With respect to our data applications in chapter 7, the performance of the EHB given an **environmental main effect** is of high interest. In figures 6.12 - 6.17 the corresponding method comparisons are investigated. For EHB vs. EHB2 (figure 6.12) the advantage of the EHB seen before is confirmed and even expands clearly with increasing environmental main effect for the medium and high association situations.
For CC illustrated in figure 6.13, EHB shows superiority in most situations. However, given a frequent environmental factor of p_e=0.5, the rank power advantage of the EHB is higher the lower the environmental main effect is. Furthermore, several situations show up where the case-control test has slightly increased ranking power than the EHB. Since this affects only low rank power situations, the effect is negligible.
In figure 6.14, showing the comparison of CASES with EHB, the same trends as seen before having no environmental main effect are observed. The same is true for MUR given in figure 6.16. The advantage of EHB compared to TWO is confirmed as well (figure 6.15). While for $p_e = 0.3$ EHB reaches more ranking power than MUK in most situations, MUK has a slight advantage given a more frequent environmental factor of $p_e = 0.5$.

6.5 Discussion

We proposed a new empirical hierarchical Bayes test for the detection of GxE interactions in GWAS. Therefore, Lewinger *et al.*'s (2007) hierarchical Bayes prioritization model (section 4.4) is adopted for the purpose of GxE interaction. By borrowing G-E information across all SNPs of a GWAS, posterior estimates for the G-E association within controls are obtained, characterized by a reduced variance. These posterior estimates are subtracted of the G-E association within cases, so that the empirical hierarchical Bayes method is a compromise between the case-control and case-only test of GxE interaction. The test reaches high power to detect markers with a GxE interaction effect, while correcting for G-E associations on the population level.
The idea for this test was proposed by Volk *et al.* in a poster in 2007. However, their test statistic was build on a variance estimate obtained by simulation replicates that are not available in real applications. To obtain a usable test statistic for the application in a real genetic epidemiological study, we calculated an appropriate variance by using a variance approximation of Kass and Steffey (1989). Furthermore, based on distributional considerations, we modified the idea of Volk et al. to obtain better properties of the statistic.
In comprehensive simulation studies comparing our new empirical hierarchical Bayes test to different established GxE interaction methods, the EHB showed overall the best performance.
As a measure for the methods' performance, we used their power to detect an interacting marker at the top ranking positions, denoted as rank power. We chose this quantity since we do not need to be concerned about type I error in a first GWAS step and since the selection of top ranking SNPs for further investigations is a common practice. This choice to consider the ranking power and disregard the conventional power and type I error, was recently supported by an invited commentary of Thomas *et al.* (2012). Thomas *et al.* (2012) argued that since a large amount of GWAS SNPs is neither related to the disease nor the exposure and since an independent replication should be done anyway

before publication, it seems reasonable to use the powerful case-only test as a screening tool in an initial discovery sample. The fact that interactions as well as associations are detected with that test is weed out by performing the case-control test not relying on the assumption of G-E independence in the replication data (Thomas *et al.*, 2012).
When we do not care about the false positive results investigated in a replication sample, the ability of a GxE interacting method to find interacting markers in the discovery step within the top ranks is even of higher importance.
In our simulation studies we have shown that our EHB is a useful alternative strategy for the selection of top SNPs in an initial GWAS. In all situations, EHB reaches higher or at least the same ranking power as the case-only approach and is therefore the better choice.
Thomas *et al.* (2012) added to his argumentation that in special circumstances where a disease has a strong behavioral component and hence high expected number of G-E associations, caution is warranted. This is the case in particular for our data application, with smoking as an environmental factor of lung cancer. However, in particular in this situation given a high number of strong G-E association markers the EHB was superior to all other methods. Investigating the behavior of our EHB method, we observed that with increasing number of G-E associations, the ranking power remains nearly the same. Given an exposure of 30% frequency we observe that the ranking power was even higher given stronger G-E association effects. We assume that this is achieved since given a more frequent environmental factor, the chance of the EHB model to identify the G-E associations correctly is higher, and hence the correction for these can be better conducted.
Another characteristic of the EHB is higher rank power for a case control ratio of 2,000:1,000 then for 1,500:1,500, at least when there is no high number of strong G-E associations. Since the cases have the more impact to the EHB test statistic than controls, this effect is not surprising. Vice versa, reducing the number of cases to 1,000 and increasing the number of controls, leads to the reverse effect. However, when a high number of strong associations is given, the ranking power for 1:1 and 1:2 remains similar, while the ranking power of 2:1 rather decreases.
Comparing the EHB and EHB2, EHB reaches higher power given an unbalanced number of cases and controls as expected while both perform similar given 1,500 cases and 1,500 controls. In all situations, the EHB outperforms the traditional case-control test of interaction.
In simulation studies of Mukherjee *et al.* (2008), MUK was compared with CC, CASES and TWO. MUK showed higher power than CC and was less powerful than CASES and TWO. As expected TWO and CASES clearly exceeded the type I error. Given larger population based G-E association effects, MUK also does not keep the type I error.
In our simulation studies considering the rank power, TWO and MUK performed similar and were better than CC and CASES in most of the situations. EHB reached generally slightly higher rank power than TWO and MUK, and a clear advantage in several scenarios.
In the original work of Murcray *et al.* (2009), they compared their approach to the traditional case-control method in simulation studies and found that it was more powerful across a wide variety of parameter setting while keeping type I error in the presence of population based G-E association effects. Therefore, they recommended it as preferable

to case-control and case-only test of GxE interaction. In a recent paper of Mukherjee *et al.* (2012), comparing CC, CA, MUK, MUR and MUR among others, MUR showed higher power when no or positive G-E associations effects were simulated, while given negative G-E associations, case-control method performs best. Overall, the results indicate that there is no most powerful procedure across all possible model parameters (Thomas *et al.*, 2012).
In our simulations, when only a low number of weak G-E associations is given, Murcray *et al.*'s (2009) methods reaches nearly 100% ranking power even for small interaction effects. However, the ranking power decreases clearly with increasing number or strength of G-E associations. In these situations, the EHB is obviously superior.
Based on all results, for the selection of top SNPs from GxE analyzes in a GWAS study, we recommend to use the EHB as a ranking tool, reaching the overall best ranking power of the considered methods. When only a low number of weak G-E association effects is expected, our advise is to use Murcray *et al.*'s (2009) method as a complementary approach, since it has high ranking power to detect small interaction effects in that situation. However, when a high number of G-E associations is expected, as given for behavioral environmental factors such as smoking, EHB should be the first and only choice.

7 TRICL lung cancer GWAS integrating pathways and GxE interaction

7.1 Motivation

Lung cancer is the most common cancer worldwide affecting nearly 1.35 million new individuals each year. Due to the limited efficacy of treatment strategies and the resulting poor 5-year survival rate of only 10%, it is the leading cause of worldwide cancer death (Parkin *et al.*, 2005). In Germany, lung cancer is the third most common cancer in both men and women after prostate, breast and bowel cancer. In 2008, nearly 34,000 men and 15,000 women newly developed lung cancer. The 5-year survival was given by 15 % in men and 19% in women (Robert Koch-Institut (RKI) und die Gesellschaft der epidemiologischen Krebsregister in Deutschland e.V. (GEKID), 2012).
Numerous environmental factors increasing the risk of lung cancer are known, e.g. radon, radiation, asbestos, air pollution or previous diseases affecting the lung such as tuberculosis (Aoki, 1993) or chronic respiratory diseases (Gao *et al.*, 1987; Wu *et al.*, 1995; Mayne *et al.*, 1999). However, the predominant factor increasing the lung cancer risk is smoking, with 90% of German lung cancer cases in men and 60% in women attributed to active tobacco smoking (RKI and GEKID, 2012). Tobacco smoke contains nearly 7,000 substances involving multiple carcinogens damaging DNA. To control and prevent lung cancer, exposure to lung carcinogens should be avoided, with cessation of tobacco consumption as the primary prevention method (Hung *et al.*, 2008b,a). However, even the risk among those who quit smoking remains elevated (Dresler *et al.*, 2006).
Although smoking plays the major role in lung cancer, only around 10% of the heavy smokers develop the disease (Sauter *et al.*, 2008). This indicates that an interindividual variability due to a genetic susceptibility to carcinogens leads to particularly higher risk in some individuals (Amos, 2007; Sun *et al.*, 2007; Matakidou *et al.*, 2005). The aggregation of lung cancer within families was evidenced in several studies (Tokuhata and Lilienfeld, 1963; Amos *et al.*, 1992; Yang *et al.*, 1997; Etzel *et al.*, 2003). Beside environmental factors, heritability is an important component in lung cancer etiology (Sun *et al.*, 2007; Matakidou *et al.*, 2005). Detecting genes involved in the disease development may help to identify groups at high risk and find new chemoprevention targets (Hung *et al.*, 2008b). Rare Mendelian cancer syndromes such as Bloom's and Werner's syndromes (Takemiya *et al.*, 1987; Yamanaka *et al.*, 1997) have shown to be related to lung cancer. Bailey-Wilson *et al.* (2004) reported a region on chromosome 6q23-25 linked to strongly familial lung cancer. Genes involved in carcinogen activation and detoxification, DNA repair of damage caused by tobacco smoke and regulation of the inflammatory response are furthermore biological plausible candidates (Sauter *et al.*, 2008). TP53, RB1 (Amos *et al.*, 1992; Etzel *et al.*, 2003) and MMP1 (Yang *et al.*, 1997) for example are identified so far. In three GWAS of lung cancer published in 2008 and subsequent pooled GWAS, susceptibility variants on chromosome 15q25 (Hung *et al.*, 2008b; Amos *et al.*, 2008; Thorgeirsson *et al.*, 2008), 5p15 (McKay *et al.*, 2008; Wang *et al.*, 2008; Rafnar *et al.*, 2009) and 6p21(Wang *et al.*, 2008) were identified with $OR \approx 1.2 - 1.3$. Landi *et al.* refined and confirmed these results in 2009 in a GWAS and meta-analysis examining different lung cancer histologies, with 5p15 showing an effect only in adenocarcinoma. The region on 15q25 involves genes coding for three nicotine acetylcholine receptor sub-

units (Hung *et al.*, 2008b). So far, the relative impact of these variants to the propensity to smoke or a direct carcinogenic effect is not clear (Landi *et al.*, 2009).

With the aim to share comparable data from ongoing lung cancer studies, to increase statistical power especially for subgroup analyses and to replicate novel findings, an international group of lung cancer researchers established in 2004 the **International Lung Cancer Consortium (ILCCO)** under the leadership of the International Agency for Research on Cancer (IARC) (Hung *et al.*, 2008a; Truong *et al.*, 2010). So far, 56 primary population- or hospital-based case-control studies from North America, Europe and Asia/Oceania participate in ILCCO. Working groups for different research areas such as genetic susceptibility, young onset or never smokers are build. The consortium provides the opportunity to share results, plan pooled analyses and discuss replication studies. It is a major step to improve our understanding of the causes and mechanisms of lung cancer and the beginning of a longstanding cooperation (Hung *et al.*, 2008a).
Furthermore, in 2010 a multidisciplinary project of worldwide lung cancer researchers funded by a grant of the National Cancer Institute (NCI) was formed, called **transdisciplinary research in cancer of the lung (TRICL)** (Amos, 2007). The NCI is the US cancer research center that conducts and supports research in the area of cancer as part of the National Institutes of Health (NIH). TRICL is one of five transdisciplinary research projects within the Post-Genome-Wide Association Initiative (U19) funded by the NCI with the goal to proceed from the initial GWAS findings to replication studies, examine gene-gene and gene-environment interactions, biologically validate GWAS findings and translate findings into clinical and preventive applications. TRICL is lead by Chris Amos at the MD Anderson Cancer Center, Houston, Texas, and its scientific goals can be grouped in 3 independent research areas. Area 1, related to the work of this thesis, is the discovery arm to identify new variants, replicate and pool GWAS findings and conduct fine mapping. Specific subsets such as early onset cases, specific histological sets, gender-defined groups or never smokers are studied, as well as gene-gene and gene-environment interactions. In addition, pathway based analyses are performed in a genome-wide manner. A collaboration of 8 lung cancer GWAS established within ILCCO contributes to this TRICL area. Area 2 concentrates on the biological understanding of precise mechanisms such as the tobacco carcinogenic process by evaluating specific genes (e.g. nicotine acetylcholine receptor subunits). Area 3 works on the in depth epidemiological modeling, characterizing genetic and environmental risk factors and constructing risk assessment models in cohorts. Overall, new insights into the etiology of lung cancer should be obtained, as well as public health benefits by identification of individual groups at high risk for lung cancer for whom screening and early detection would highly reduce the burden of lung cancer.

As a part of ILCCO and TRICL area 1, we analyzed four of the participating GWAS studies with the hierarchical model proposed for pathway analysis and GxE interaction analysis. We furthermore combined both tasks and applied other methods for comparison. After these comparisons, the consortium intends to analyze all GWAS in 2012.
In the following a short description of the four analyzed genome-wide association studies is given. The quality control procedure is exemplarily outlined for the German Lung Cancer Study in section 7.3. In section 7.4 we give an overview of the different analyses performed, that are then presented in more detail in the following three sections 7.5-7.7. We will end with a closing discussion.

7.2 Study populations

The **German Lung Cancer Study** (**GLC**) is a genome-wide, population-based case-control study in Caucasians. The investigated sample compasses 514 cases diagnosed before the age of 51 and 488 controls matched by sex and age genotyped on Illumina HumanHap 550K SNP chips (Landi *et al.*, 2009). It is composed of subsets from three independent German studies: 201 cases from the Heidelberg lung cancer study, 313 cases from the LUCY study and 488 controls form the KORA study (Sauter *et al.*, 2008; Holle *et al.*, 2005; Wichmann *et al.*, 2005).
The LUCY study, **Lu**ng **C**ancer in the **Y**oung, is a multicenter study with 31 participating hospitals all over Germany, conducted by the Helmholtz Zentrum Munich (HMGU, Prof. Wichmann) and the University Medicine Göttingen (Prof. Bickeböller). Only patients with a new diagnosis of histologically or cytologically confirmed primary lung cancer were included in the study. The data collection finished in 2011, with 847 recruited lung cancer patients and 5,524 family members. The data comprise detailed information about family history, tobacco and smoking exposure, education, occupational exposure and blood samples (Sauter *et al.*, 2008).
The Heidelberg lung cancer study is an ongoing hospital-based case-control study conducted by the German Cancer Research Center (DKFZ, PD Risch). Since 1997, more than 2,000 lung cancer cases were recruited in collaboration with the Thoraxklinik Heidelberg involving nearly 300 cases with onset of disease before the age of 51. 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 *et al.*, 2008).
The KORA study (Cooperative Health Research in the Augsburg Region) is a population-based study in the area of Augsburg, Southern Germany, conducted by the Helmholtz Zentrum Munich (Prof. Wichmann). 18,000 participants were recruited between 1984 and 2001 in four stages with the aim to examine environmental and genetic risk factors of human diseases. The data comprise multiple phenotypes, medical and laboratory data, as well as blood samples (Sauter *et al.*, 2008). Since a major population stratification between Southwest Germany and two other cohorts from Northern Germany could not be detected in a genomic control approach, KORA is accepted as a representative sample of German Caucasians (Steffens *et al.*, 2006).

The **Central Europe lung cancer GWAS** of the **IARC** (**CE-IARC**) (Prof. Brennan) is based on a multicenter hospital-based case-control study conducted with Cancer Institutions from 6 central and eastern European countries between 1998 and 2002. In total, 2,633 newly diagnosed lung cancer cases and 2,884 controls were recruited. Controls were frequency matched to cases by sex, age, geographical area and period of recruitment (Scélo *et al.*, 2004). 1,989 of the lung cancer cases and a group of 2,625 comparable hospital controls were genotyped on Illumina HumanHap 300K platforms. Data on lifestyle risk factors, occupational history, medical and family history is available (Hung *et al.*, 2008b).

The **Texas genome-wide lung cancer study** ascertained 1,150 histologically confirmed non-small cell lung cancer cases and 1,134 controls from an ongoing hospital-based case-control study in Caucasians conducted by the M.D. Anderson Cancer Center (**MDACC**) (Prof. Amos, Prof. Spitz) of the University of Texas, Houston. Lung cancer cases were newly diagnosed at MDACC since 1991, controls were from routine care at

Table 7.1: *Overview of the different TRICL lung cancer GWAS analyzed in this thesis*

	GLC	**CE-IARC**	**MDACC**	**SLRI**
Principal Investigator	Wichmann, Risch, Bickeböller	Brennan	Amos	Hung
Location	Germany	Czech Republic, Hungary, Poland, Romania, Russia, Slovakia	Houston, Texas, USA	greater Toronto area, Canada
Design	Population-based	Hospital-based	Hospital-based	Hospital-based
Matching Factors	Age, sex, residence	Age, sex, residence	Age, sex, ethnicity, smoking	Age, sex, ethnicity
Control recruitment	KORA	Non-tobacco related diseases	Cancer-free patients; only ever smokers	Hospital family medicine clinic
Number of cases/controls	514/488	1,989/2,625	1,181/1,184	332/505
Genome-wide SNP chip	HumanHap 550K	HumanHap 300K	HumanHap 300K	HumanHap 300K

the Kelsey-Seybold clinics in the Houston Metropolitan area. Only former and current smokers are involved in the sample with controls frequency matched to the cases according to their smoking behavior including age, ethnicity and sex and years of cessation for former smokers. For the genome-wide genotyping, Illumina HumanHap 300K SNP chips were used (Amos *et al.*, 2008; Wang *et al.*, 2008; Hung *et al.*, 2008a).

The **Toronto lung cancer GWAS** involves 332 cases and 505 controls of European ancestry from a case-control study conducted by the University of Toronto and the Samuel Lunenfeld Research Institute (**SLRI**) (Prof. Hung) in the greater Toronto area between 1997 and 2002. The case-control study is hospital-based and involves 445 lung cancer cases recruited at the hospitals in the network of University of Toronto and SLRI and 962 controls randomly selected from individuals visiting family medicine clinics. Controls are frequency matched by age, sex and ethnicity. The data comprise lifestyle risk factors, occupational, medical and family history as well as blood samples of more than 85% of the participants. Genome-wide genotyping was done on Illumina HumanHap 300K SNP chips (Hung *et al.*, 2008b).

In the following we will abbreviate the different genome-wide studies by **GLC** (German GWAS), **CE-IARC** (Central Europe), **MDACC** (Texas GWAS) and **SLRI** (Toronto GWAS). A short overview of the studies is given in table 7.1.

7.3 Preprocessing of the data

7.3.1 Quality Control

We conducted a systematic quality control as a first step of our analysis at each study center separately, using comparable quality criteria (section 3.2.3). We will outline the procedure for the German Lung Cancer Study.
Of the 514 cases and 488 controls selected for GLC study, genotyping failed for two of the cases. 12 additional individuals (8 cases, 4 controls) were excluded since genotypes for more than 10% of the 561,466 SNPs were missing. 966 of the remaining 988 persons had a call rate of more than 95% and the overall genotyping rate was 99.4%. We started checking the sex of the individuals based on the X-chromosomal information. For one case, the determined sex did not agree with the reported one, so that we excluded this person. Two cases showed a low rate of heterozygous genotypes in comparison to the other individuals, two controls showed an excess of heterozygotes. These were excluded as well. As a next step cryptic relatedness between the different participants was investigated by determining pairwise similarities. We identified 3 case pairs as duplicates or monozygotic twins and 17 pairs of second degree relatives. For each of the identical pairs the smoking status agreed and we removed one individual at random. The second degree pairs included 13 different persons, most of them involved in a complex network of relatedness as shown in figure 7.1. Of this group, we removed as few individuals as possible (2 cases and 4 controls) so that no second degree relatives remained in the sample.
As a last step of quality control for the individuals, we performed a principal component analysis (PCA) on a subset of nearly 100,000 SNPs to assess population structure and identify ethnic outliers (section 3.2.4). Therefore we used the software EIGENSOFT (Price *et al.*, 2006). Since we restricted our analyses to Caucasians, six individuals (4 cases, 2 controls) with a Non-Caucasian self-reported ethnicity (Arabs, Asians) were removed in advance. For a previous analysis, STRUCTURE was applied to the data set to assign the different probes to European, African or Asian ancestry with HapMap Phase II data as reference sample (International HapMap Consortium, 2005). Two controls with 40% African background were identified. These were the controls strongly deviating from the sample heterozygosity distribution that was mentioned above and they were already excluded. Furthermore, one of the self-rated Arabs was clearly identified. We also discovered that some of the cases and controls with an Eastern European and Russian background include a low percentage of Asian background (up to 16%). A graphical presentation can be seen in figure 7.2. The SNP subset used for EIGENSOFT was obtained by selecting markers from the whole set, so that no high LD between the chosen markers remained. Additionally, non-autosomal SNPs were removed as well as monomorphic SNPs. The principle component axes were tested for statistical significance by Tracy-Widom statistic (Tracy and Widom, 1992). Our PCA provided 20 eigenvectors with a p-value≤ 0.05, of whom 17 even had p-values $\leq 10^{-7}$. In figure 7.3 we can see plots of the first 8 principle components with the single individuals colored according to their self-reported origin. All four plots show a main core cluster involving most of the individuals with some outliers in the different directions.

We repeated the PCA using an iterative procedure integrated in EIGENSTRAT to automatically remove outliers. In the first iteration, 18 persons were removed and 15 more

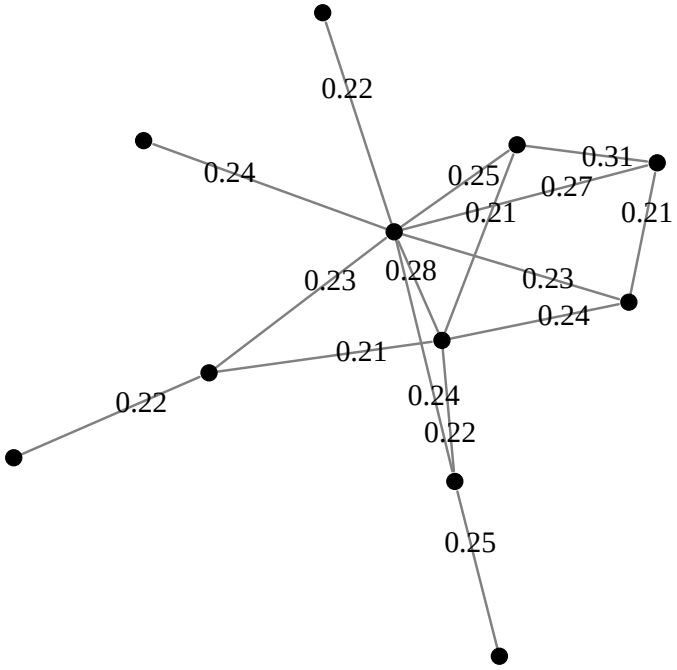


Figure 7.1: *Overview of the 2^{nd} degree relatives in the GLC. The nodes of this plot represent the different individuals denoted by their study ID. The edges represent the relatedness between the individuals with the given number the corresponding similarity measure. For genetically identical individuals, this measure equals 1, while values close to 0.5 denote first degree relatives and values close to 0.25 are given for second degree relatives.*

in a second iteration. Thereby, the number of significant eigenvectors was reduced to 4 with a p-value<0.05, involving 3 highly significant ones. In 4 more iteration steps, 3, 6, 1 and 1 more individuals were removed before the outlier removal terminated. 3 PC axes still were significant. When checking the reported ethnicities for the identified outliers, we found that several were of East European or Russian origin, as also observable from figure 7.3. We decided to remove the outliers from the first two iterations and to use the first 4 principle components (PCs) in following analyses when possible. These four PCs are displayed in figure 7.4. Individuals are colored depending on the originating study (LUCY, Heidelberg, KORA). We see no major genetic differences between these groups. In total, the final sample involved 935 individuals for the analyses (467 cases and 468 controls).
Subsequently SNPs were filtered according to their proportion of missings, minor allele frequency or deviation from HWE within the controls. We removed 7,889 SNPs with more than 5% missing genotypes, 23,778 SNPs with a MAF $< 1\%$ and 405 SNPs with a HWE p-value within controls $< 10^{-7}$. Furthermore, 2.728 heterozygous haploid genotypes (SNPs on X or Y chromosome in men) were detected and set to missing. Finally, after frequency and genotype pruning 529,730 SNPs remained.
All quality procedures – except of the identification of outliers and population structure – were performed with the GWAS software PLINK. More detailed information on the

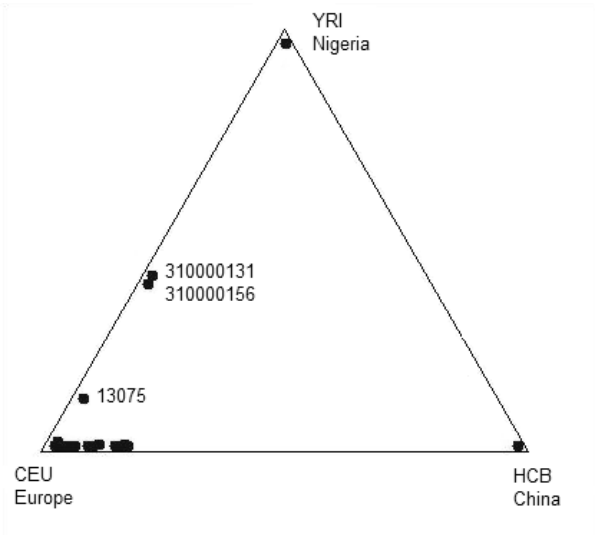


Figure 7.2: *Assignment of the GLC individuals to Caucasian, Asian and African genetic background represented by Hap Map phase II reference populations CEU, HCB and YRI using the population structure software STRUCTURE (Pritchard et al., 2000).*

Table 7.2: *Quality criteria used for our TRICL GWAS analyses*

SNP specific quality checks	
Call rate	$\geq 95\%$
Minor allele frequency	$\geq 1\%$
Hardy Weinberg Equilibrium in controls	$p_{\text{HWE}-\text{controls}} \geq 10^{-7}$
Individual specific quality checks	
Call rate	$\geq 90\%$
Sex mismatch	female F < 0.2 and male F > 0.8
Heterozygosity	[mean F +/- 6 standard deviation F]
Cryptic relatedness	proportion alleles IBD < 0.20
Population outliers	Caucasian ancestry,
	| PLINKs nearest neighbor Z score | <4

motivation for the different filter criteria and the corresponding usage of PLINK can be found in the appendix A.2. An overview of the thresholds used for the quality filtering process is given in table 7.2.
Since we did not strictly fix how outliers should be identified, the methods varied for the different studies. For the Central Europe study, STRUCTURE was used, defining population outliers as individuals with an ancestry probability rate of being Caucasian $< 80\%$. MD Anderson used the outlier detection diagnostic in PLINK (absolute value of the nearest neighbor Z score > 4).
For the SLRI, MDACC and CE-IARC study 331, 1,150 and 1,901 lung cancer cases, 499, 1,134 and 2,503 controls and 314,072, 312,452 and 310,045 SNP remained for the analysis after excluding subjects and markers based on the different quality criteria.

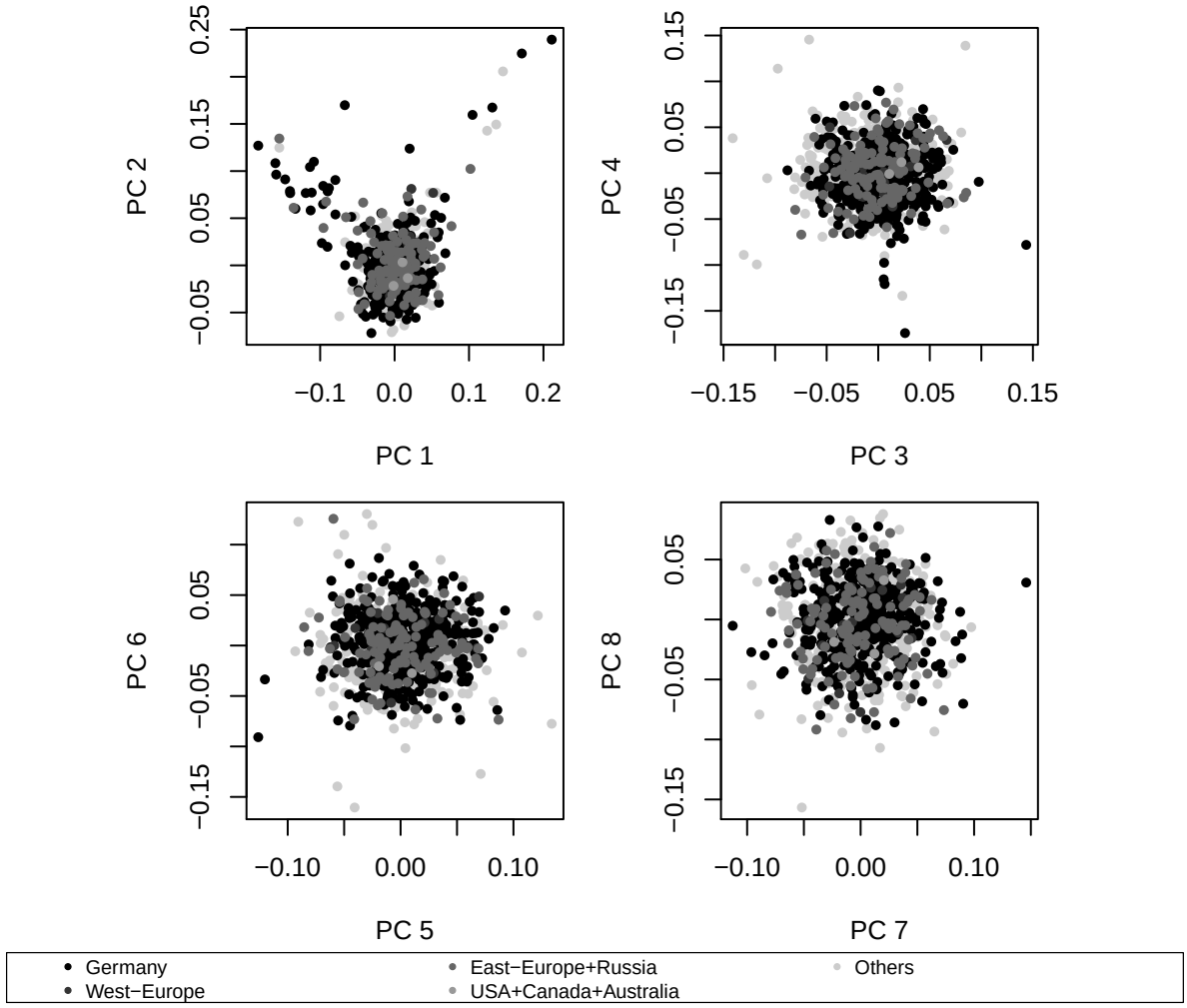


Figure 7.3: *Principle component analysis of GLC. Plots of the first 8 principle components with outliers included. The different colors represent the different reported ethnicities.*

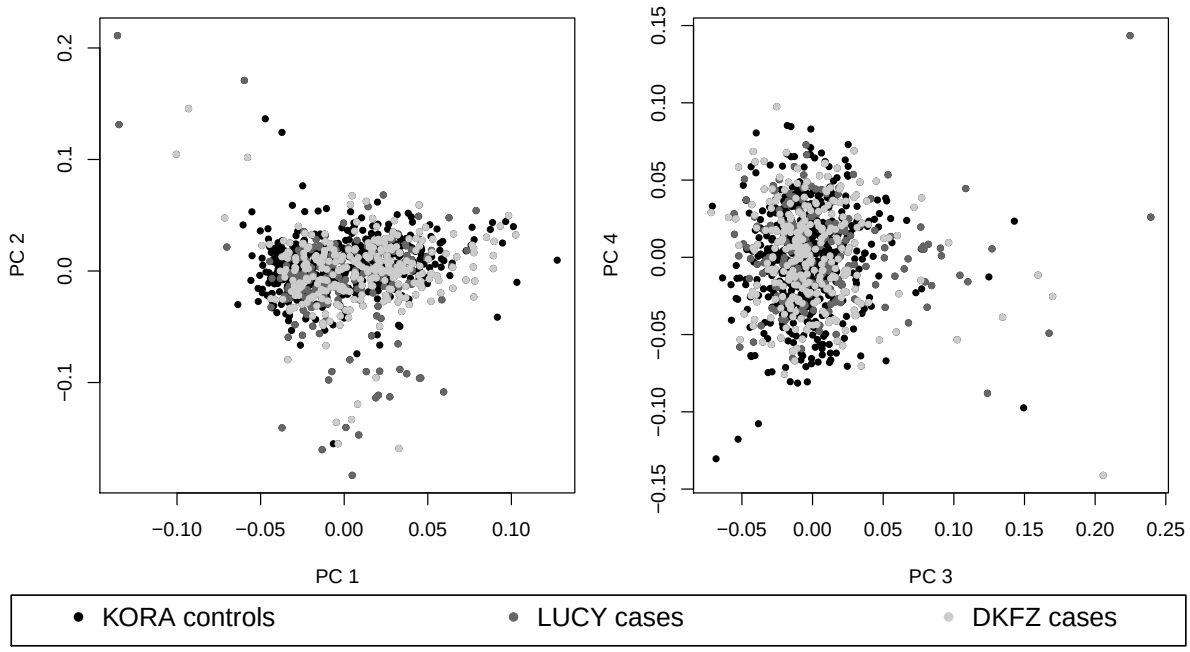


Figure 7.4: *Principle component analysis of GLC. Plots of the first 4 principle components with outliers excluded. The different colors represent the three different underlying studies.*

7.3.2 Age, sex, smoking and ethnicity

Beside the information about case-control status, additional phenotypic data considered as covariates are sex, age, smoking status and quantity as well as ethnicity. An overview about the corresponding distributions of these characteristics within the different studies is given in table 7.3.
Sex and age as important confounders for clinical questions are typically considered. Sex was coded by an indicator with 0 for men and 1 for women. With respect to age, we defined five-year intervals for age at diagnosis for cases or at interview for controls. Age intervals are usually used to give adequate odds ratios for the different groups. We chose intervals < 50, $50-54$, $55-59$, $60-64$, $65-69$, $70-74$ and > 75 years. Within GLC, involving young people only with age of diagnosis before 51, we furthermore splitted the youngest age class and used < 45, $45-49$ and $50-54$. In SLRI, age was missing for 1 case and 1 control, in GLC 9 persons were without a record of age.
Since smoking is the most important risk factor in lung cancer, with a strong dose-response relationship (Ruano-Ravina *et al.*, 2003), smoking status as well as quantity of smoking should be considered in the analysis. The smoking status was coded by never, former, current or any smoker. Never smokers were defined as no more than 100 cigarettes in life when possible or by the original definition of the single studies else. Former smokers were individuals that stopped smoking for at least 2 years. Any smokers involved all individuals that had smoked before with the current smoking state missing. Current, former and any smokers were combined to ever smokers. MDACC involved only ever smokers. The number of never smoking cases in GLC, SLRI and CE-IARC was low. Smoking status was missing only for 5 controls of SLRI and 2 controls of CE-IARC. A usual measure for the smoking quantity is pack-years. Pack-years are

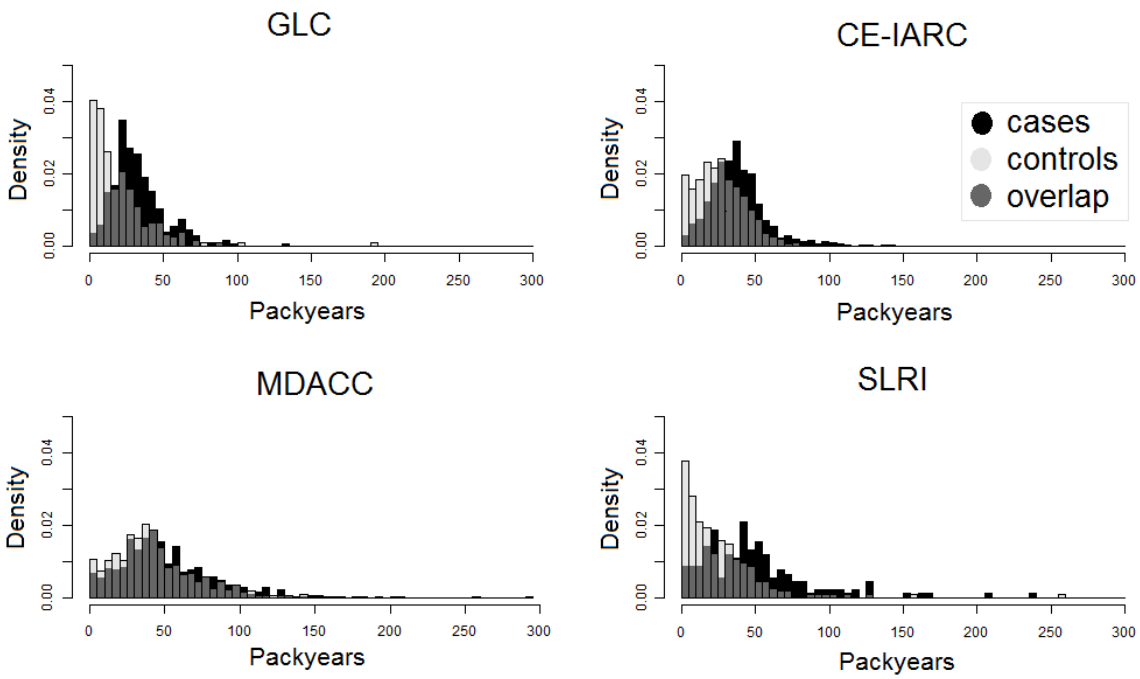


Figure 7.5: *Distribution of ever smoker pack-years for the different analyzed lung cancer GWAS*

defined by the number of packs smoked per day multiplied by the years as a smoker (Amos *et al.*, 2008). One pack was assumed to include 20 cigarettes. Hence, pack-years describe the total amount of smoked packs. In figure 7.5 we can see the distribution of pack-years for ever smokers of the different studies separately for cases and controls. Cases tend to higher numbers of pack-years than controls. Using these numbers we furthermore defined another binary variable for the smoking quantity, categorizing ever smokers in moderate or heavy smokers. Moderate smokers were defined as less or equal to 20 pack-years, while heavy smokers were defined with more than 20 pack-years. While in heavy smokers the genetic predisposition for lung cancer is nearly negligible due to the strong effect of smoking, in moderate smokers, protective or risk increasing genes may have a meaningful contribution to the development of lung cancer. We observe that the proportion of moderate smokers within the cases is lower in comparison to the controls, what confirms our graphical impression of the increased smoking behavior in cases. Note, for the SLRI data, only for 183 of the 240 ever-smoker cases information about the smoking amount was given, while 228 of 279 controls had available pack-years information. Most of the subjects with missing pack-years (about 100) are from the any smoker category, with the current status unspecified since no smoking stop dates are available. For a few more subjects from the other smoking categories either smoking start or stop date was missing and hence pack-years could not be obtained. In the GLC study pack-years were missing for 21 ever smoking cases and only 1 ever smoking control. For CE-IARC, 5 ever smoking cases and 10 ever smoking controls had no specification of smoking quantity.
The last covariate concerns the principal components for the genetic background repre-

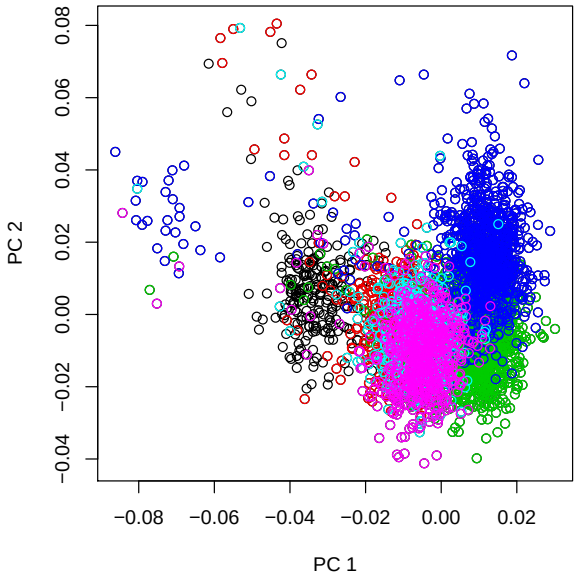


Figure 7.6: *First two principle components of the IARC Central Europe study, CE-IARC, comprising individuals from 6 different central and eastern European countries (colors). The PCs reflect the genetic differences due to the individuals origin.*

senting the ethnic origin of the individuals. For the different studies, we used 4 (GLC), 6 (CE-IARC), 2 (MDACC) and 3 (SLRI) PCs to adjust for population structure in analyses where this was possible. For the CE-IARC, the first two PCs clearly represented the 6 different European countries involved (figure 7.6).

7.3.3 Gene and biological pathway information

For our pathway based analyses, we decided to use biological pathways from one systematic database and agreed on KEGG with TRICL. In February 2012, the KEGG database comprised 245 pathways with 5,981 genes that we extracted for our purpose using the R-package KEGGSOAP (Zhang and Gentleman, 2011).
For the SNP to gene assignment, we used the gene information available on the NCBI homepage (ftp://ftp_ncbi.gov/gene/DATA/GENE_INFO /Mammalia/Homo_sapiens_gene_info.gz, May 16^{th} 2011). Information for 45,650 human genes was available including 21.545 protein-coding genes, 12,163 pseudo-genes and different kinds of non-protein-coding RNA genes. We defined a gene region as the given start and end position extended by 20,000 base pairs in each direction. A SNP located within this region was assigned to the gene. An assignment of one SNP to several genes was possible. When no direct gene was available, a SNP was assigned to its nearest gene within +/- 480,000 base pairs. Of the 538.860 SNPs on chromosomes 1-22, X and Y that occurred in our studies, 9,025 SNPs could not be assigned to a gene since no gene within +/-480,000 bp was available. In total 35,270 of the genes involved at least 1-1,767 SNPs directly or within the 480K neighborhood. The median number of SNPs assigned per gene was 10.
For one of our strategies integrating pathway information for the analysis of GxE

Table 7.3: *Characteristics of the investigated lung cancer GWAS. (QC = quality control);moderate smoker:* ≤ 20 *pack-years; heavy smoker:* > 20 *pack-years*

Study	GLC		IARC-CE		MDACC		SLRI	
# SNPs after QC	529730		312452		314072		310045	
	cases	**controls**	**cases**	**controls**	**cases**	**controls**	**cases**	**controls**
# individuals after QC	467	468	1,901	2,503	1,150	1,134	331	499
Sex								
Male	286	237	1,493	1,821	655	644	159	190
Female	181	231	408	682	495	490	172	309
Age								
<45	169	112	246	415	176	120	41	233
45-49	239	283						
50-54	50	73	272	378	109	107	32	62
55-59	-	-	329	394	158	210	28	46
60-64	-	-	386	435	186	278	46	32
65-69	-	-	353	430	202	236	62	35
70-74	-	-	286	368	184	134	69	41
≥75	-	-	29	83	135	49	52	49
missing	9	0	-	-	-	-	1	1
Smoking status								
Never	35	214	144	884	-	-	91	215
Former	45	121	373	656	601	655	95	143
Current	377	133	1380	954	549	479	90	90
Any	10	-	4	7	-	-	55	46
missing	-	-	-	2	-	-	-	5
Smoking quantity (ever smokers only)								
Moderate	83	152	248	619	160	230	38	122
Heavy	328	101	1,504	988	990	904	145	106
missing	21	1	5	10	-	-	58	51

interactions, smoking related pathways were selected from the collection of biological KEGG pathways by a specialist in this area (Xifeng Wu, MDACC) based on literature. The gene and biological pathway information was applied identical to all four data sets.

7.4 Aims of analysis and presentation of results

In the four presented lung cancer GWAS our interest was the application of the empirical hierarchical Bayes model for

1. analysis of main effects integrating pathway information (section 7.5),
2. analysis of GxE interaction effects (section 7.6),
3. analysis of GxE interaction effects integrating pathway information (section 7.7).

For comparison purpose, we applied two gene set analysis methods and carried out a variety of tests for the analysis of GxE interactions. The pathway analyses as well as the GxE analyses are building upon simple single SNP test results. Therefore, different logistic regression models were used as an initial step, followed by the analyses with

the empirical hierarchical Bayes method and other approaches for comparison. In the following three sections, the different strategies for the analyses are outlined in more detail and the corresponding results are presented.

For genome-wide significance in the initial regression results, we used a Bonferroni correction with a global alpha-level of 0.05.
When comparing different methods with each other, we used SNP lists if the approaches work on the SNP level, as the different tests for GxE interactions do. However, for study comparison we used gene rankings since not each SNP is available for every study and furthermore it is sufficient if an effect of the same genetic region is detected within different studies, not necessarily due to exactly the same SNP. Each gene was represented by the top SNP assigned to that gene, as done for GSEA and SUMSTAT. We concentrated on the top ranked 100 SNPs or genes. For the HBP, the posterior quantity E_{M_i}, $i = 1, \ldots, N_M$, was used as re-ranking criterion (section 4.4.2). This decision was based on the simulation results of Lewinger *et al.* (2007). He showed that the posterior expectation of the strength of an association effect E_{M_i} and the posterior probability of association P_{M_i} perform better than the conditional expectation $E^+_{M_i}$. We preferred E_{M_i} since it considers both information, the posterior expectation for an association and the corresponding probability (section 4.4.3). For GSEA, the LES genes are considered as "significant" genes. We talk about a "replicated gene", if it occurs in the top 100 or LES for at least two different studies.
For the hierarchical Bayes prioritization, we considered pathways occurring at least for two different studies in the top 10 when ranked according to their corresponding β or μ coefficient as done in section 5.4. As a reminder, the β coefficients (equation 4.19) represent the increase or decrease of the prior probability of association for each SNP involved in the corresponding pathway. The μ coefficients (equation 4.20) represent the increase or decrease of the prior strength of association for each SNP involved in the corresponding pathway. We talk about a "replicated pathway" with respect to HBP, if it occurs in the top 10 according to β or μ for at least two different studies.
For SUMSTAT and GSEA nearly none of the pathways reached FDR significance ($FDR \leq 0.05$). Therefore, we considered pathways with a nominal p-value ≤ 0.05 and denoted a pathway as "replicated", when it was nominally significant in at least two different studies.
For graphical representation of our results we used among others a list comparison plot as recently proposed by Antosh *et al.* (2011) to assess the similarity of two ranked gene lists. Starting with the highest ranks of two such ranking lists of N_G different genes each, a small number of top genes (t_G) for each of the lists is selected. These top gene subsets are then compared to each other. The fraction of the selected genes (t_G/N_G) per list is plotted against the fraction of both lists' common genes within these top ones (c_G/t_G). In the following, the latter will be denoted as **proportional overlap**. Then, we march down the ranks by adding further genes to the selected top ones step by step and plot again the fraction of genes in common between the two lists of selected genes as function of fraction of genes selected in each list. Furthermore, we investigated if the proportional overlap for a particular number of top ranks between the two lists can be explained by chance alone or if it is higher than expected, pinpointing to true gene effects. In the first case the proportion of genes in the overlap should be roughly equal to the fraction of genes selected. For this purpose, we simply used the hypergeometric

distribution for the first top gene comparison to assess significance of the proportional overlap. For any further step, not the proportional overlap itself is evaluated, but the increase in the proportional overlap considering the situation of the last step. We used that kind of plot not only for gene lists, but also for other kinds of ranking list, e.g. based on SNPs or pathways.
Many results are exemplarily displayed for GLC, since this is the German GWAS. When figures and tables for a second study are shown, we chose the Central Europe data as the largest study with presumably highest power.

7.5 Analysis of main effects integrating pathway information

For the analysis of main effects integrating pathway information, two different logistic regression models were carried out for each SNP assuming a log-additive effect (SNP = 0,1,2). In the first model, we adjusted for sex, age and principal components. In the second model, smoking status coded by 0 for never smokers and by 1 for ever smokers as well as the number of pack-years representing the dose-response relationship were additionally involved. The models are given by

Model 1 (M1):

$$\text{logit(case/control)} \quad = \alpha + \beta_1 SNP + \beta_2 gender + \sum_{i=3}^{9} \beta_i age_{i-2} + \sum_{j=10}^{15} \beta_j PC_{j-9}$$

Model 2 (M2):

$$\begin{aligned}\text{logit(case/control)} \quad &= \alpha + \beta_1 SNP + \beta_2 gender + \sum_{i=3}^{9} \beta_i age_{i-2} + \sum_{j=10}^{15} \beta_j PC_{j-9} \\ &+ \beta_{16} smokingstatus + \beta_{17} packyears\end{aligned}$$

$age_1, \ldots, age_7$ are dummy variables representing the different age classes, $PC_1, \ldots, PC_6$ the principle components with optionally $PC_j = 0$ for $j = 3, 4, 5, 6$ depending on the particular study. In the following results part, we will use the abbreviations **M1** and **M2** to represent our two different **pathway models**.
Based on the logistic regression results, the hierarchical Bayes prioritization including pathway information (Lewinger *et al.*, 2007) and two other gene set analysis methods were applied. We chose the GSEA as the most popular gene set method and the SUM-STAT approach that has shown to be more powerful (Tintle *et al.*, 2009a; Fehringer *et al.*, 2012). For both, 1000 permutations were performed using PLINK (Purcell *et al.*, 2007). FDR was calculated as described in section 5.3.4. Pathways with less than 5 genes occurring in the data were excluded from the analysis, so that 234 pathways were analyzed. For the hierarchical Bayes prioritization, the Z matrix was built by 235 columns representing an intercept and the 234 pathways. The matrix entry for each SNP-pathway combination was set to 1 for SNPs involved in the pathway and to 0 when the gene(s) corresponding to the SNP did not occur within the pathway.

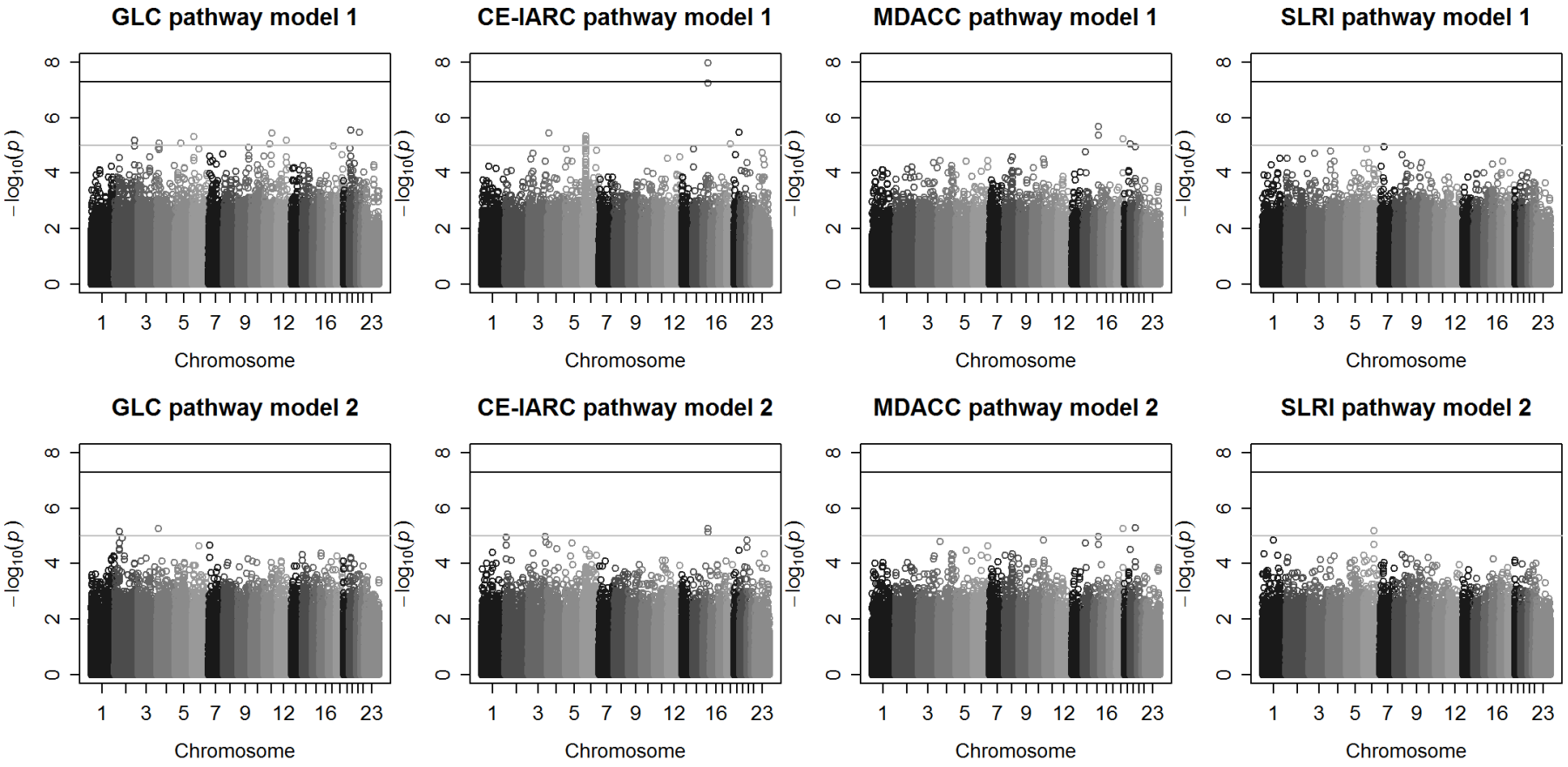


Figure 7.7: *Manhattan plots for SNP main effects of both pathway models for all four studies. The upper line represents genome-wide significance. The lower line indicates a p-value of* 10^{-5}.

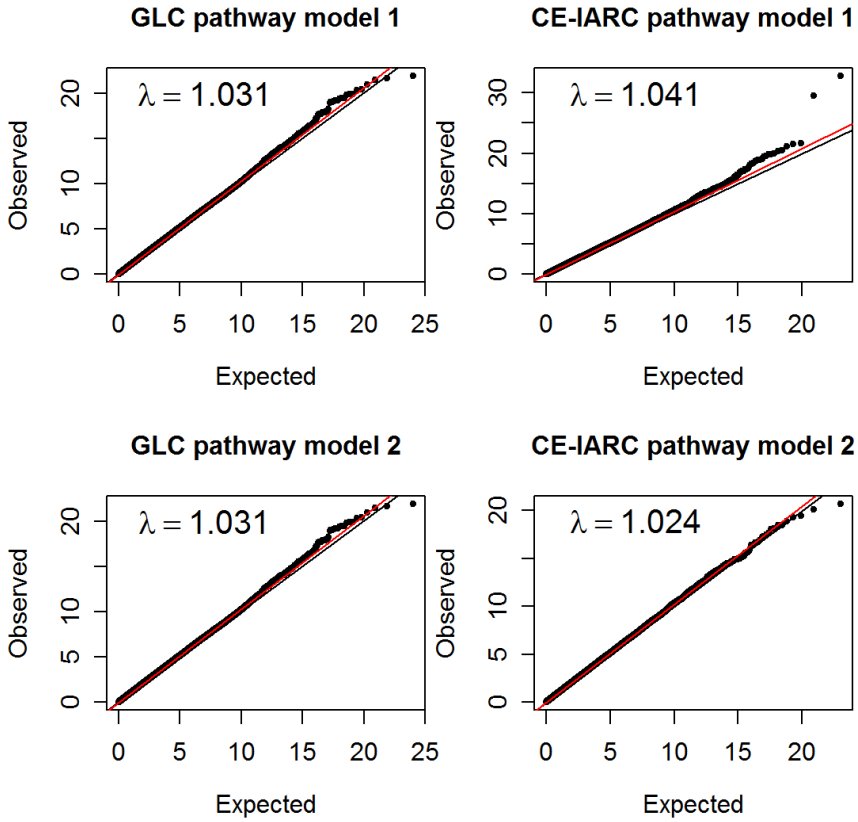


Figure 7.8: *QQ plots of SNP main effects from both pathway regression models for GLC and CE-IARC*

7.5.1 Initial main effect results

In figure 7.7 the Manhattan plots for both pathway models of all studies are given to convey an impression of the single SNP results. None of the studies showed a critical inflation factor λ_{GC}, with the largest value of $\lambda_{GC} = 1.041$ occurring for CE-IARC M1. The QQ-plots with the λ_{GC} inflation factors for GLC and CE-IARC are shown in figure 7.8. The other QQ-plots are given in the appendix figure B.1.
For GLC, we have some single SNPs spread across the genome with a $p \leq 10^{-5}$, but no striking region. For MDACC and CE-IARC we clearly identify the two SNPs on chromosome 15q25 that were published for lung cancer some years ago (Amos *et al.*, 2008; Hung *et al.*, 2008b; Thorgeirsson *et al.*, 2008). This is not surprising, since these two data sets are underlying two of the corresponding publications (MDACC: Amos *et al.* (2008) and CE-IARC: Hung *et al.* (2008b)). For the CE-IARC M1, *rs8034191* reached genome-wide significance, the neighboring SNP *rs1051730* barely missed the genome-wide level with a p-value of $5.7 \cdot 10^{-8}$. Although the signal is much weaker given the pathway model considering smoking as covariate, both SNPs are the top ones in that analysis as well. These SNPs furthermore stand out for both pathway models of MDACC. They are the two top SNPs for M1 and within the top 10 SNPs for M2. For CE-IARC, a peak of 10 moderately associated SNPs in the 6p22 region is furthermore found. This was already mentioned in Hung *et al.* (2008b), but still has to be verified.

Beyond that, only single SNPs reached a p-value $\leq 10^{-5}$ and no additional interesting region was observed. For SLRI M1, none of the SNPs even reached a $p \leq 10^{-5}$.

7.5.2 Pathway hyperparameter estimates of HBP

Of the 234 investigated pathways, between 150 and 205 had positive β-regression coefficients for the different studies and pathway models and between 167 and 206 had positive μ-regression coefficients, related to the prior probability of association and the corresponding prior strength of association effect for SNPs within a pathway. For each analysis, at least 80% of the pathways had the same sign of β- and μ-coefficient. For a particular pathway, this means that for all its SNPs the prior probability and prior strength of association are either both increased or both decreased. The same holds for the μ coefficients between the different pathway models and studies. For β, we had the same sign across studies for at least 70%. The prior probability of association for SNPs belonging to none of the pathways (β_0) was in the range of 10^{-9} for both pathway models in all studies. For M1 of GLC and M2 of CE-IARC and SLRI, the corresponding "baseline" μ was 0, for the others it ranged from 0.69 to 2.21.

7.5.3 Comparison of top pathways between studies

Hierarchical Bayes Prioritization

When comparing the pathway rankings according to β- or μ-coefficients, the correlation between the different studies is high. In particular, the analyses split into two groups, within which we see higher similarities on the top of the ranking lists than expected by chance. The first group involves model 1 of GLC and model 2 of CE-IARC and SLRI (group A), the second group compasses model 1 of CE-IARC, MDACC, SLRI and model 2 of MDACC and GLC (group B). In all cases the correlation according to β is higher than with respect to μ. In figure 7.9 we can see list comparison plots between the models of GLC and CE-IARC representative for the correlation within both groups and between them. The corresponding numbers of overlapping top 10 pathways between the different studies and pathway models can be seen in table 7.4.

The possible reasons for this particular grouping may be the role of smoking as a confounding factor in the given context and the differences in populations underlying the four studies. MDACC includes no never smokers, so that M1 and M2 differ only in the adjustment for the amount of smoking. This results in a very similar pathway ranking for both models, so that they fall into one group. For GLC, SLRI and CE-IARC the pathway ranking for both models differs more severe, since the differentiation between never and ever smokers is additionally relevant. For model M1, smoking status as a confounder is not considered at all, resulting in top pathways that may be rather related to smoking than to lung cancer directly. M2 however accounts for the smoking status, leading to different, lung cancer relevant pathways. Therefore, both models split to the two groups. The list comparison plot comparing the pathway ranking of M1 and M2 for GLC and MDACC can be seen in figure 7.10. The fact that GLC involves only young individuals may result to the contrary distribution of the models to the groups, since the importance of smoking may be different at younger ages.

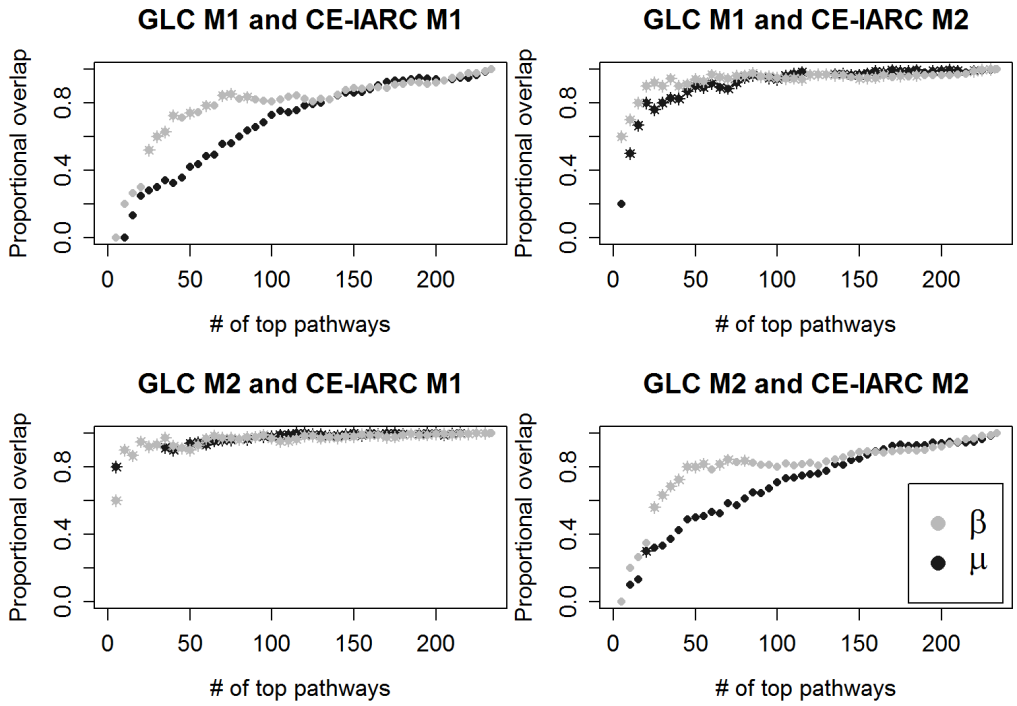


Figure 7.9: *List comparison plots of pathway rankings according to β and μ of HBP between different studies. The y-axis shows the proportion of common pathways for a particular number of top pathways given on the x-axis. The stars indicate a significant overlap.*

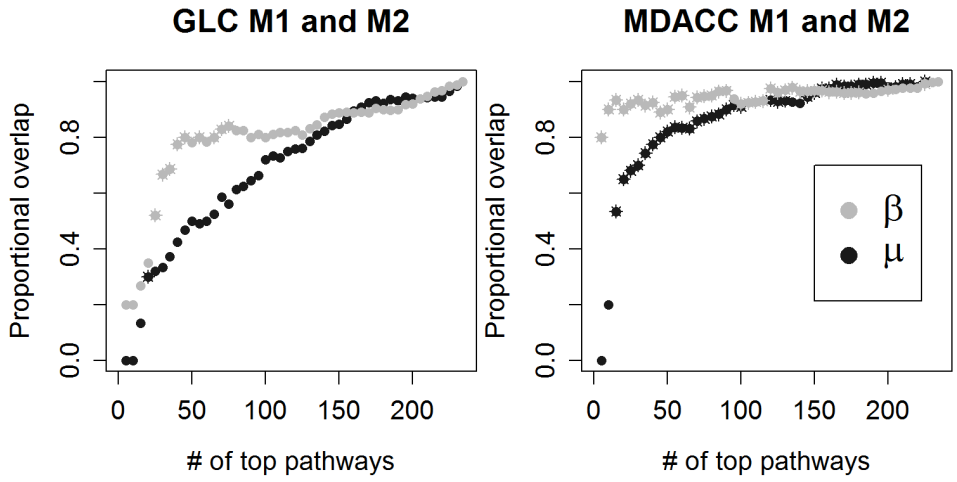


Figure 7.10: *List comparison plots of pathway rankings according to β and μ of HBP between the different pathway models in GLC and MDACC. The y-axis shows the proportion of common pathways for a particular number of top pathways given on the x-axis. The stars indicate a significant overlap.*

Table 7.4: *Numbers of common top 10 pathways between the different studies and pathway models using β regression coefficients as ranking criterion on the upper triangle or μ regression coefficients on the lower triangle.*

			β							
			GLC		CE-IARC		MDACC		SLRI	
			M1	M2	M1	M2	M1	M2	M1	M2
μ	GLC	M1		2	2	9	2	2	2	9
		M2	0		7	2	10	9	9	2
	CE-IARC	M1	0	5		2	7	6	7	2
		M2	9	1	0		2	2	2	9
	MDACC	M1	0	9	4	1		9	9	2
		M2	0	2	6	0	2		9	2
	SLRI	M1	0	9	4	1	9	1		2
		M2	6	1	0	7	1	0	1	

Table 7.5: *Numbers of common top 10 pathways between the different studies and pathway models using β regression coefficients as ranking criterion or μ regression coefficients. Both pathway models (M1 and M2) were combined for this comparison.*

	one study	two studies	three studies	four studies
β	6	2	9	7
μ	10	7	11	5

Appendix tables B.2 and B.3 give lists of the pathways that belong to the top 10 for at least two different studies. These are 16 pathways using β as pathway ranking criterion and 15 using μ. All of these pathways were not even for 2 but at least 3 of the studies in the top 10. Even 7 (β) and 4 (μ) of the pathways were in the top 10 for all 4 studies. In table 7.5 the numbers of pathways occurring in only one study, two, three or four different studies regardless of the corresponding pathway model (M1 and M2) are given. Comparing the top pathways for β and μ, 2 pathways occurred in both lists.

Gene set enrichment analysis

The gene set enrichment analysis identified overall only one pathway as significant according to FDR$\leq$ 0.05 in CE-IARC. The corresponding enrichment score was driven by 35 of the 123 genes totally involved. However, this pathway not even reached a nominal p-value ≤ 0.05 for any of the other studies.

Several pathways reached nominal significance for each of the studies ($p_{\text{nominal}} \leq 0.05$). The number of significant pathways for each of the 8 analyses as well as the overlap between model 1 and 2 per study is shown in table 7.6. The overlap between the two different pathway models was significant for the two larger studies MDACC and CE-IARC.

Table 7.6: *Number of nominal significant pathways for both smoking models of the different lung cancer studies with GSEA and SUMSTAT. The gene set analysis is based on single SNP main effects.*

	GSEA				SUMSTAT			
	GLC	**CE-IARC**	**MDACC**	**SLRI**	**GLC**	**CE-IARC**	**MDACC**	**SLRI**
model 1	8	14	5	10	26	51	15	16
model 2	8	15	10	14	5	49	14	19
model 1 ∩ 2	1	8	5	2	4	39	12	9
model 1 ∪ 2	15	21	10	22	27	61	17	26

In figure 7.11 we see the overlap of nominal significant pathways between the four different studies for GSEA. In total, 8 pathways were identified in two different studies and one pathway in three of the studies (CE-IARC M1+M2, SLRI M1, GLC M2). In the appendix table B.4 a list of these 9 pathways and their corresponding nominal p-values can be found. The overlap is significant for none of the study pairs considering the sum of pathways for model 1 and model 2. However, considering the two different models and separately looking at the common pathway per model, we have a significant overlap for CE-IARC M1 and MDACC M1 with 2 common pathways, CE-IARC M1 and SLRI M2 with 3 shared pathways, SLRI M1 and GLC M2 with an overlap of 2.

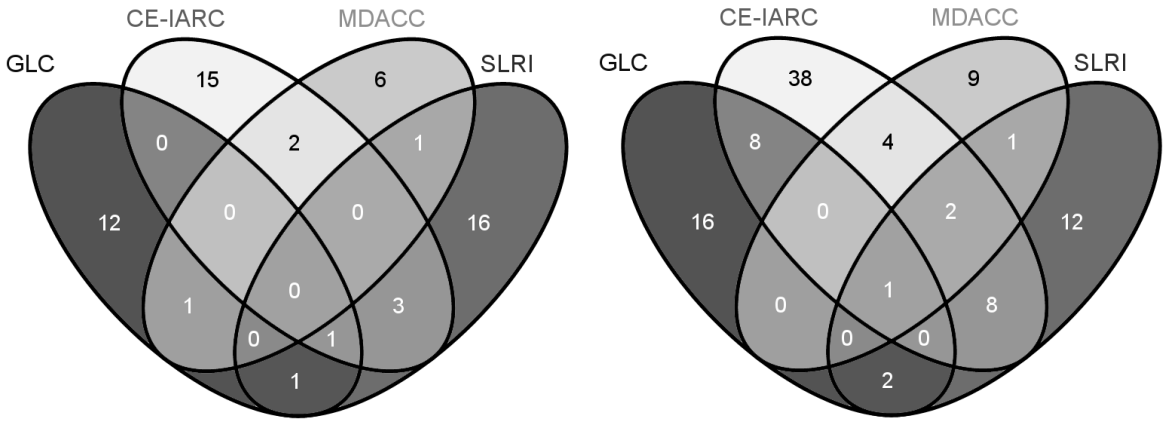
SUMSTAT method

As for GSEA, the SUMSTAT method identified only for CE-IARC M1 significant pathways according to FDR (≤ 0.05). These were the pathway that was found with GSEA as well and two additional ones. Only one of the latter had a nominal significant result in one of the other studies (SLRI M2).

Several pathways reached nominal significance for each of the studies ($p_{\text{nominal}} \leq 0.05$). Their number is shown in table 7.6. We clearly see that the numbers are much higher than for GSEA with exception of GLC M2. In particular, CE-IARC showed a really high number of significant pathways, making more than 20% of all considered pathways. The overlap between the two models was significant for all four studies ($p \leq 0.05$). Comparing the identified pathways between the different studies, only SLRI and CE-IARC M2 had a significant overlap with 8 common pathways. This leads to a significant overlap of the sum of pathways over model 1 and model 2 for SLRI and CE-IARC as well. The consistency between the different studies is illustrated in figure 7.11b. In total, 26 pathways were identified in at least two different studies. Two of these were detected in three of the studies and one pathway had a p-value ≤ 0.05 for all four. The latter was found with both models for the larger studies CE-IARC and MDACC, and with M1 only for SLRI and GLC. Most of the 26 pathways were found with at least one of the models for the CE-IARC data. Appendix table B.5 gives the 26 pathways and their corresponding nominal p-values.

7.5.4 Comparison of top pathways between methods

Of the pathways in both HBP ranking lists (table B.2 for β, B.3 for μ) we find three of them in the GSEA list (table B.4) as well. Two of them were found by SUMSTAT for



(a) *GSEA* **(b)** *SUMSTAT*

Figure 7.11: *Overlap of nominal significant pathways for the different lung cancer studies. The analysis is based on single SNP main effects. We build the sum over both pathways models.*

two different studies as well (table B.5).
Comparing the numbers of pathways with nominal p-value ≤ 0.05 between GSEA and SUMSTAT, many of the pathways found with GSEA were detected by SUMSTAT as well. However, the GSEA pathways are not a subset of the SUMSTAT ones, but GSEA also found some pathways not identified by SUMSTAT. In table 7.7 we can see the corresponding numbers.
Comparing the lists of replicated results involving 9 pathways for GSEA and 26 pathways for SUMSTAT (table B.4 and table B.5), we find more than half of the GSEA pathways for SUMSTAT as well.

7.5.5 Resulting pathways

The two pathways occurring on both lists (β and μ) of replicated HBP pathways (appendix tables B.3 and B.2) were *hsa000430* and *hsa003020*. *hsa000430* (*taurine and hypotaurine metabolism*) was within the β top 10 for all 8 analyses and in the μ top 10 for GLC model 2 and MDACC and SLRI model 1, with a maximum rank of 23 for the other analyses. The other pathway, *hsa003020* (*RNA polymerase*), was within the β top 10 for GLC model 1 and CE-IARC and SLRI model 2 with a maximum rank of

Table 7.7: *Number of nominal significant pathways for GSEA and SUMSTAT and the corresponding overlap between both methods. The gene set analysis is based on single SNP main effects.*

	GLC		CE-IARC		MDACC		SLRI	
	M1	M2	M1	M2	M1	M2	M1	M2
GSEA	8	8	14	15	5	10	10	14
SUMSTAT	26	5	51	49	15	14	19	19
GSEA ∩ SUMSTAT	6	0	12	13	2	4	7	8

32 and within the μ top 10 for GLC model 2, MDACC model 1 and SLRI model 1 and ranked up to 59.
The pathways identified by GSEA and SUMSTAT as significant according to FDR was the *systemic lupus erythematosus* pathway (*hsa05322*) in CE-IARC M1 (GSEA $FDR = 0.011$, SUMSTAT $FDR = 0.017$). Additionally, using SUMSTAT, *type 1 diabetes mellitus* (*hsa04940*, $FDR = 0.046$) and *Wnt signaling pathway* (*hsa04310*, $FDR = 0.048$) reached FDR significance for CE-IARC M1 as well.
With GSEA, the (*alpha-Linolenic acid metabolism* pathway; *hsa00592*) reached nominal significance in three of the studies (CE-IARC (M1+M2), SLRI (M1) and GLC (M2) (appendix table B.4). Using SUMSTAT, 2 such pathways were detected, *cholinergic synapse* (*hsa04725*) and *HTLV-I infection* (*hsa05166*). *Neuroactive ligand-receptor interaction* (*hsa04080*), had a p-value ≤ 0.05 for all four studies. The latter was found with both models for the larger studies CE-IARC and MDACC, and with M1 only for SLRI and GLC (appendix table B.3). Of the pathways in both HBP lists in tables B.3 and B.2, we find three of them in the GSEA list as well. These are *hsa003450* (*non-homologous end joining*), *hsa000270* (*cysteine and methionine metabolism*) and *hsa000592* (*alpha-linolenic acid metabolism*). *hsa003450* was in the β top 10 for all analyses of group B (see page 7.5.3), with maximum rank of 33 for the others, and was detected by GSEA for GLC M2, as well as MDACC M1 and M2. *hsa000270* was in the β top 10 for group A with maximum rank of 54 and identified in MDACC and SLRI M2 by GSEA. The third pathway, *hsa000592* was in μ top 10 for the group B - except for MDACC M2 (rank 11) - and ranked on 117, 118 and 120 for group A. Nominal significance using GSEA was given for GLC M2, SLRI M1 and CE-IARC M1 as well as model 2. The first two of these pathways were on the SUMSTAT list in table B.5 as well.
Pathways additionally replicated by GSEA and SUMSTAT were *neuroactive ligand-receptor interaction* (*hsa04080*), *colorectal cancer* (*hsa05210*) and *cell adhesion molecules* (*hsa04514*).

7.5.6 Comparison of top genes between studies

Hierarchical Bayes Prioritization
As mentioned in section 7.4, we used E_M for SNP re-ranking after the analysis with HBP for pathway integration. In figure 7.12 we can see a comparison of the rankings according to the three different posterior quantities P_{M_i}, ,$E^+_{M_i}$ and E_{M_i}, $i = 1, \ldots, N_M$. For all studies and both pathway models, P_{M_i} and E_{M_i} were highly correlated to each other, while E^+_M had a much lower similarity with both. This supports Lewinger *et al.* (2007) observations in his simulation studies that P_{M_i} and E_{M_i} show similar power while $E^+_{M_i}$ performs worse. Although the results presented in the following are based on E_{M_i}, the results are representative for the ranking according to P_{M_i} as well due to the high correlation.
When comparing the initial regression gene ranking to the gene ranking after integration of pathway information, we have a gain of consistency between the studies. In figure 7.13, the pairwise comparison of rankings between studies for HBP and initial regression are graphically contrasted in list comparison-plot. Considering M1, the concordance between CE-IARC, GLC and MDACC is clearly increased. SLRI

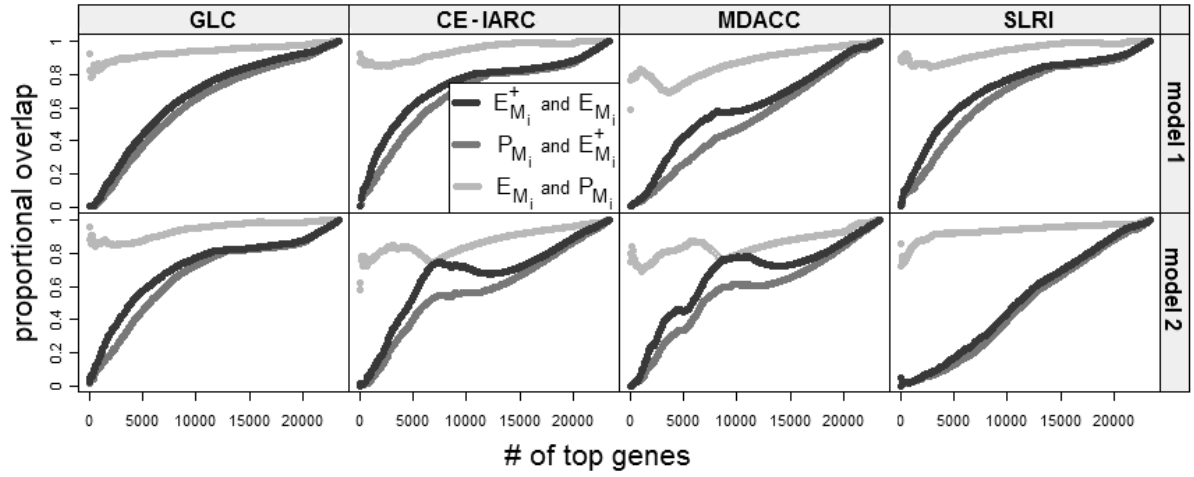
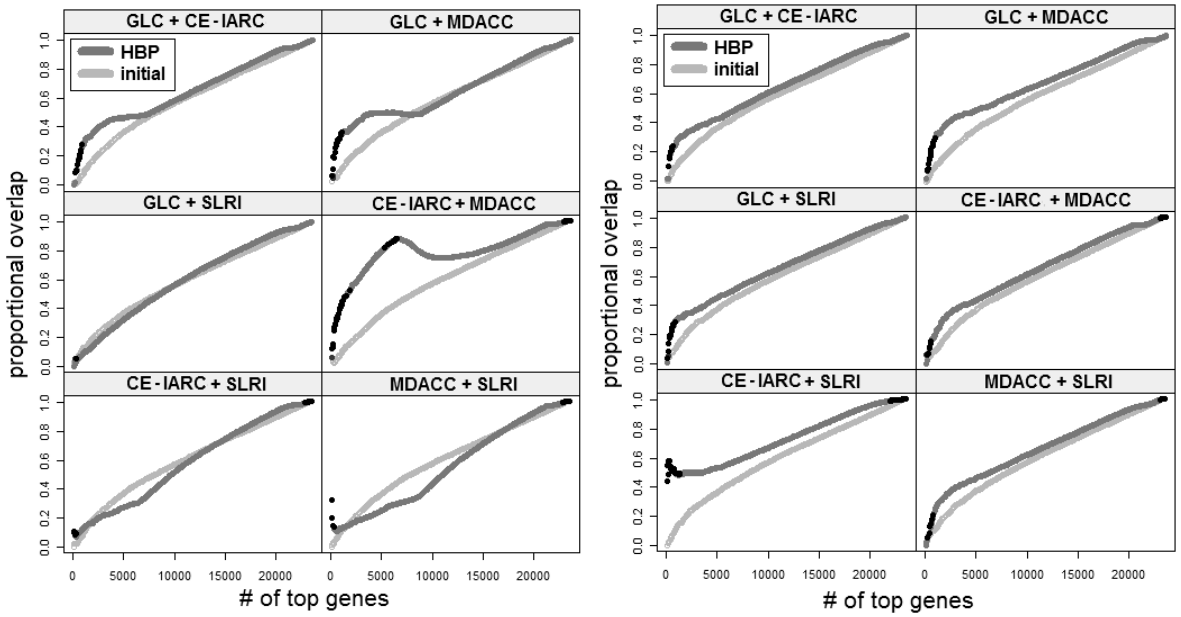


Figure 7.12: *List comparison plots of gene rankings according to different HBP posterior quantities $E^+_{M_i}$,P_{M_i} and E_{M_i}. The y-axis shows the proportion of common genes for a particular number of top genes given on the x-axis.*



(a) *Pathway model 1* **(b)** *Pathway model 2*

Figure 7.13: *List comparison plots of gene rankings between different studies for HBP (E_M) and initial regression results. The y-axis shows the proportion of common genes for a particular number of top genes given on the x-axis. The darker points indicate a significant overlap.*

Table 7.8: *Number of overlapping top 100 genes between studies and models contrasted for initial regression results (upper triangle) and HBP (lower triangle)*

			Initial regression analysis							
			GLC		CE-IARC		MDACC		SLRI	
			M1	M2	M1	M2	M1	M2	M1	M2
Hierarchical Bayes Prioritization	GLC	M1	-	32	2	1	5	4	3	3
		M2	12	-	1	0	1	1	2	2
	CE-IARC	M1	2	8	-	56	3	3	2	1
		M2	52	2	5	-	3	3	1	0
	MDACC	M1	11	4	13	8	-	81	1	1
		M2	3	8	32	2	34	-	2	1
	SLRI	M1	0	4	11	0	20	10	-	34
		M2	60	3	1	55	13	2	0	-

is the only study behaving differently. For SLRI and GLC, similar consistencies are observed with initial regression and HBP, comparing SLRI ranking with CE-IARC and MDACC an improvement on the top ranks is observed, while the similarity gets worse with increasing ranks. For M2, all pairs show a gain in consistency using HBP, with CE-IARC and SLRI in particular worth to mention, with at least 40% overlap over the whole gene ranking lists.

In table 7.8, the overlap of the top 100 genes between the different studies of initial ranking and ranking based on the posterior quantity are shown. Again, we see a clear increase in consistency by pathway integration. For the initial results, 3 overlapping genes were given for CE-IARC and MDACC, for all other combinations we had only 1 or even no gene in common. After the integration of pathway information and re-ranking of the SNPs, the number of common genes between the different studies increased for most combinations to up to 60 genes. In particular, for all pairwise comparisons of GLC M1, CE-IARC M2 and SLRI M2, as well as MDACC M1 with CE-IARC M2 and SLRI M1 a high number of common genes is observed.

Between the two models per study a higher number of common genes was observed with initial regression. A possible reason, as already mentioned before, may be that the top results for M1, for all studies except of MDACC, are not necessarily related to lung cancer directly but possibly to the unconsidered confounding factor smoking. While this already affects the top genes for the initial regression, the effect may be even increased by the additional pathway information used for the HBP. This even more severe emerge of the smoking related genes for M1 leads to less common genes between M1 and M2 per study.

Considering the top 100 genes for each of the analyses, we observe 521 different genes. Of these, 112 occur in 2 different studies, 52 occur in 3 studies, and 4 are in top 100 for at least one model of all four studies. For the initial regression, 577 different genes were observed considering the top 100 genes for each analysis. Only 20 of these occur in 2 different studies. None occurred in 3 or four studies.

Gene set enrichment analysis

Table 7.9 shows the total number of LES genes extracted from the pathways with a nominal $p \leq 0.05$ in the GSEA for each of the analyses. While GLC, CE-IARC and SLRI had around 300-400 LES genes for each of the analyses, for the MDACC study

Table 7.9: *Number of LES genes of GSEA for the different studies*

	GLC	CE-IARC	MDACC	SLRI
model 1	304	301	47	321
model 2	284	423	94	343
model 1 ∩ 2	108	204	48	66
model 1 ∪ 2	481	521	94	599

Table 7.10: *Number of overlapping LES genes of GSEA between the different studies and corresponding p-values for the overlap in brackets*

	GLC CE-IARC	GLC MDACC	GLC SLRI	CE-IARC MDACC	CE-IARC SLRI	MDACC SLRI
model 1	12 (0.8599)	3 (0.4507)	99 (**$2.2\cdot 10^{-54}$**)	17 (**$6.7\cdot 10^{-11}$**)	44 (**$8.1\cdot 10^{-10}$**)	3 (0.4870)
model 2	38 (**0.0001**)	10 (**0.0151**)	24 (**0.0396**)	17 (**0.0003**)	70 (**$1.1\cdot 10^{-16}$**)	14 (**0.0010**)
model 1 ∪ 2	44 (**$8.1\cdot 10^{-10}$**)	66 (**$7.2\cdot 10^{-21}$**)	47 (**$1.3\cdot 10^{-06}$**)	75 (**$1.7\cdot 10^{-29}$**)	204 (**$3.6\cdot 10^{-179}$**)	70 (**$1.1\cdot 10^{-16}$**)

only 47 (M1) and 94 (M2) LES genes occurred. A reason for the identification of less LES genes may be the lack of never smokers in the MDACC data.
In the same table we furthermore see the number of intersecting LES genes between M1 and M2 per study. The intersect is highly significant for all studies. This is not surprising since the same data are underlying the slightly different analyses. In particularly noticeable is MDACC, with the LES genes for M1 a subset of these for M2. This result fits well to our very similar HBP pathway ranking for both models of MDACC and our hypothesis, that results are much more similar as for the other studies due to the missing never smokers.
In table 7.10, the overlap of LES genes between the different studies is given. Measured by the total number of genes occurring within the considered 234 pathways, the overlap of LES genes between GLC and SLRI, SLRI and CE-IARC as well as CE-IARC and MDACC M1 is significant. For model 2, between all pairs of studies we have a significant overlap. The same holds for the combined sets of model 1 and model 2. Comparing the LES genes of M1 of one study and M2 of another study, for most combinations (9 out of 12) significance is given as well. In total, 1354 different LES genes occurred. Of these, 299 were LES genes for at least 2 different studies, 40 were found by 3 different studies. Two genes were identified in the LES of all four studies.

SUMSTAT method
To evaluate the results of SUMSTAT on the gene level, we built for each analysis a new ranking list containing only these genes occurring in the corresponding significant pathways. Furthermore, the top 100 genes for each of the new lists was considered and compared between the different studies. In table 7.11 the number of common top genes between the different analyses is shown. We clearly see that the number of common top genes is larger than based on the initial regression results - shown in the upper part of table 7.8. However, the consistency given by HBP exceeds the one by SUMSTAT. In

Table 7.11: *Number of overlapping top 100 genes between studies and models contrasted for initial regression results (upper triangle) and SUMSTAT (lower triangle)*

			Initial regression analysis							
			GLC		CE-IARC		MDACC		SLRI	
			M1	M2	M1	M2	M1	M2	M1	M2
	GLC	M1	-	32	2	1	5	4	3	3
		M2	21	-	1	0	1	1	2	2
SUMSTAT	CE-IARC	M1	6	2	-	56	3	3	2	1
		M2	5	3	55	-	3	3	1	0
	MDACC	M1	10	6	6	8	-	81	1	1
		M2	9	9	8	8	79	-	2	1
	SLRI	M1	11	3	5	5	4	9	-	34
		M2	4	4	3	3	2	1	43	-

total, 537 different genes occurred in the 8 different top 100 gene lists. 43 were found for two and only 11 for three different studies. No gene is identified in the top 100 for all four studies.

7.5.7 Comparison of top genes between methods

The overlap between the replicated genes of the different methods is shown in figure 7.14. The initial list of replicated genes has only one gene in common with GSEA and SUMSTAT, as well as two more with SUMSTAT. While the initial regression has nearly no overlap with any of the other methods, GSEA has multiple genes in common with HBP and SUMSTAT. Two occurred in the replicated gene lists of GSEA, SUMSTAT and HBP. Between SUMSTAT and HBP, no additional common genes occurred.
Of the 21 genes shared by GSEA and HBP, 9 genes occurred for 3 studies of one method and 2 studies of the other method. One gene was for both methods within the top 100 genes for all four studies. The remaining 11 genes were found in both methods for two different studies. Of the 33 genes shared by GSEA and SUMSTAT, 8 genes occurred for 3 studies of GSEA and 2 studies of SUMSTAT or vice versa. 5 genes occurred for 3 studies in GSEA and SUMSTAT. One of the genes found with GSEA for all 4 studies occurred on the replicated SUMSTAT gene list, identified for 2 of the studies. All other genes were found with GSEA and SUMSTAT for 2 different studies.

7.5.8 Resulting genes

For HBP, the genes *FADS1*,*FADS2*, *HES1* and *MIR1908* occurred in the top 100 in all four studies. FADS2 was in the LES of GSEA for all four studies as well, in addition to GOT2.
Comparing the replicated genes between the different methods, the initial list of replicated genes has the *CHRNA3* gene in common with GSEA and SUMSTAT, as well as *TBL1XR1* and *IL7* with SUMSTAT only. Note, *CHRNA3* was initially detected for CE-IARC and MDACC with both models. With SUMSTAT it was found for the same two studies. The GSEA method however had this gene in the LES of the other two studies, SLRI and GLC M2.

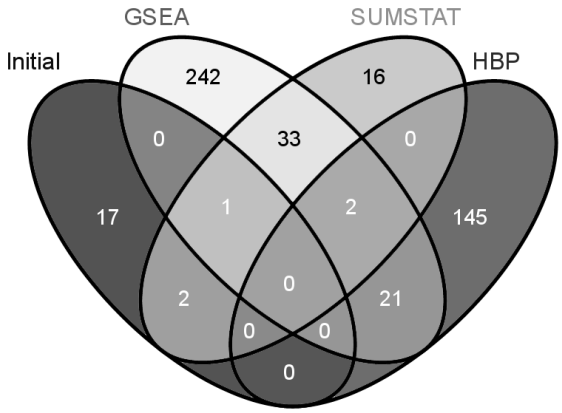


Figure 7.14: *Comparison of replicated genes between different methods. A gene is replicated, when it occurs in the top 100 genes for at least two different studies for initial, SUMSTAT and HBP or at least for two different studies in the LES genes for GSEA. Initial = initial regression results of the pathway models for SNP main effect.*

The genes *ADCY3* and *ADCY5* occurred in the replicated gene lists of GSEA, SUMSTAT and HBP. The 8 genes that occurred for 3 studies of GSEA and 2 studies of SUMSTAT or the other way around were *HLA-C*, *PARD3*, *PGM2*, *PLCB1* and *CTNNA2*, *GALR1*, *GRIK1*, *GABRG3*. *CTH*, *EGFR*, *HLA-B*, *PGM1*, *PIK3R1* occurred for both methods for 3 studies. *GOT2* found with GSEA for all 4 studies occurred on the replicated SUMSTAT gene list, identified for 2 of the studies. For HBP and GSEA 9 genes, *FEN1*, *PRKAA2*, *ITPR1*, *ADCY2*, *ADCY3*, *ADCY8*, *ADCY9*, *ATP1B1* and *PRKACG*, were for one of the methods within the top 100 genes of 3 different studies and 2 different studies for the other method. *FADS2* was for both methods within the top 100 genes for all four studies.

7.6 Analysis of GxE interaction effects

Due to the high importance of smoking in lung cancer development not only as an environmental main effect but also as a component interacting with genetic factors, we investigated GxE interactions with smoking as the environmental factor. For smoking classification we chose two different approaches, comparing never with ever smokers (coded by 0/1) (NE) and moderate with heavy smokers (coded by 0/1) (MH). In the following results part we will use the abbreviations **NE** and **MH** to distinguish the two different **smoking models**.
Genetic effects were classified into a binary variable as well by choosing a dominant inheritance model that pools the heterozygous genotype with the minor homozygous one. The binary classification of the genetic factor is an assumption for our GxE approach and we chose the dominant model since the power is much higher than based on a recessive model, and it is also more plausible in cancer.
For each SNP, we tested for a G-E association separately in cases and controls. The corresponding logistic regression models are given by

$$\text{logit}(\text{exposed/not exposed}) = \alpha_{\text{controls}} + \beta_{\text{controls}} G \tag{7.1}$$

and

$$\text{logit}(\text{exposed/not exposed}) = \alpha_{\text{cases}} + \beta_{\text{cases}} G. \tag{7.2}$$

The latter corresponds to the case-only test of GxE interaction. The resulting parameter estimates and their corresponding variance estimates were used as input for our empirical hierarchical Bayes method for GxE interactions as given in section 6.3.
For comparison purpose, the simple case-control test for interaction

$$\text{logit}(\textit{case/control}) = \alpha_{\text{ge}} + \beta_{\text{g}} G + \beta_{\text{e}} E + \beta_{\text{ge}} GE, \tag{7.3}$$

as well as the empirical Bayes approach of Mukherjee and the two-step approaches of Albert and Murcray (sections 6.2.2-6.2.4) were applied. For the latter, we additionally carried out the model testing for an overall G-E association ignoring the disease status

$$\text{logit}(\text{exposed/not exposed}) = \alpha_{\text{all}} + \beta_{\text{all}} G. \tag{7.4}$$

The different GxE methods are abbreviated as in chapter 6, with EHB for our new empirical hierarchical Bayes approach, CC (case-control), CASES (case-only), TWO (simple two-step), MUK (Mukherjee) and MUR (Murcray).

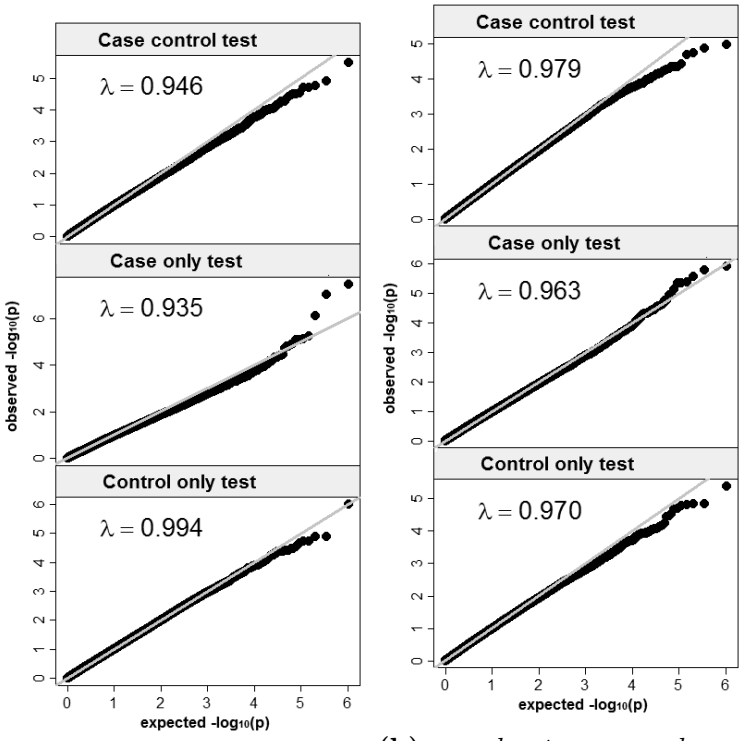
7.6.1 Initial GxE effect results

Using the traditional case-control test for GxE interactions, none of the SNP markers reached genome-wide significance in any of the studies. For CE-IARC and SLRI, some signals build by several SNPs with $p < 10^{-5}$ were detected. Beyond that, only some single SNPs reached a p-value level of 10^{-5}. The corresponding Manhattan plots for both smoking models of GLC and SLRI are shown in the upper row of figure 7.15. The corresponding plots for CE-IARC and MDACC can be found in the appendix figure B.2. Although the case-only test has been shown to be much more powerful than the case-control test for interaction, only one genome-wide significant SNP showed up, testing never vs. ever smokers in GLC. One more neighboring SNP just missed the genome-wide significance level. We can see this signal in the corresponding Manhattan plot on the right in the middle row of figure 7.15a. Furthermore, we see another prominent region on chromosome 13 for this analysis, formed by 3 different SNPs. One of these SNPs had a $p < 10^{-5}$ for case-control test as well. In GLC smoking model MH, 3 SNPs on chromosome 9 were at the top positions. For SLRI NE four SNPs on chromosome 18 that were in the top for case-control test, showed up in case-only as well. For moderate vs. heavy smokers, two SNPs on chromosome 16 are worth to mention. For CE-IARC, two regions with three SNPs each for smoking model MH were noticeable on chromosome 14 and 3 on chromosome 16.
The plots in the lower row of figure 7.15 show the results testing for a G-E association within controls only that is often used as representative for the population-based G-E association. A strong G-E association signal is not observed in any of the studies and smoking models. None of the SNPs at a top position of the case-only test of the different studies pinpoints to a G-E association within controls.

(a) *GLC*

(b) *SLRI*

Figure 7.15: *Manhattan plots for GxE interaction effects of GLC and SLRI never vs. ever smokers and moderate vs. heavy smokers. Upper row: Case-control test; Middle row: Case-only test; Lower row: Control only test. The upper line represents genome-wide significance. The lower line indicates a p-value of* 10^{-5}.



(a) *never vs. ever smokers* **(b)** *moderate vs. heavy smokers*

Figure 7.16: *QQ plots for GxE interaction effects of GLC never vs. ever smokers and moderate vs. heavy smokers. Upper row: Case-control test; Middle row: Case-only test; Lower row: Control-only test*

The QQ-plots for the several analyses do not show any inflation of case-only test results although expected. All λ values for genomic control are even less than 1. For illustration, the QQ-plots for GLC are shown in figure 7.16.

Due to the failure of even case-only test for GxE interaction to find genome-wide significant results in these studies, the selection of top SNPs for follow-up is an appropriate proceeding. To obtain a better SNP ranking, our new empirical hierarchical Bayes method (EHB) seems useful. In the following we will compare the top 100 SNPs with respect to several different GxE interaction tests. Afterwards, we will take a look at the consistency of the top findings between the studies.

7.6.2 Comparison of different GxE methods by their top SNPs

Comparing the top 100 GxE interacting SNPs of the different GxE methods within study and smoking model, we see for all four lung cancer studies quite similar trends. In tables 7.12 and 7.13 the data for the GLC and CE-IARC are shown exemplarily. For all studies, the overlap of top 100 SNPs between the different methods is in general larger for NE as for MH. While analyzing never vs. ever smokers the overlap of the top 100 SNPs between the different methods in Central Europe is always larger than for the GLC, we see a reverse trend in moderate vs. heavy smokers. SLRI generally tends to less common SNPs for both analyses, for MDACC no clear trend can be observed.

Table 7.12: *Comparison of the top 100 SNPs between the different G × E interaction methods for GLC. CC: case-control, CASES: case-only, TWO: intuitive two-step, MUK: Mukherjee's, MUR: Murcrays, EHB: empirical hierarchical Bayes, HBP-GxE: hierarchical Bayes prioritization based on GxE interaction effects (see section 7.7), EHB-PW: empirical hierarchical Bayes integrating pathway information (see section 7.7)*

		moderate vs. heavy							
		HBP-GxE	**CC**	**CASES**	**TWO**	**MUK**	**MUR**	**EHB**	**EHB-PW**
never vs. ever	**HBP-GxE**		59	9	4	15	0	9	2
	CC	24		9	43	15	0	9	1
	CASES	13	30		64	60	16	100	5
	TWO	22	59	66		56	11	64	3
	MUK	18	31	77	63		4	60	6
	MUR	0	0	2	1	0		16	5
	EHB	13	30	100	66	77	2		5
	EHB-PW	12	28	96	65	78	2	96	

Table 7.13: *Comparison of the top 100 SNPs between the different G × E interaction methods for CE-IARC. CC: case-control, CASES: case-only, TWO: intuitive two-step, MUK: Mukherjee's, MUR: Murcrays, EHB: empirical hierarchical Bayes, HBP-GxE: hierarchical Bayes prioritization based on GxE interaction effects (see section 7.7), EHB-PW: empirical hierarchical Bayes integrating pathway information (see section 7.7)*

		moderate vs. heavy							
		HBP-GxE	**CC**	**CASES**	**TWO**	**MUK**	**MUR**	**EHB**	**EHB-PW**
never vs. ever	**HBP-GxE**		57	7	32	13	0	7	8
	CC	6		8	45	12	0	8	8
	CASES	2	41		60	65	15	97	91
	TWO	3	63	72		11	48	50	58
	MUK	2	46	88	71		1	67	64
	MUR	0	0	3	2	1		13	13
	EHB	2	41	99	72	89	3		94
	EHB-PW	2	43	94	72	88	3	95	

Focusing on our new empirical hierarchical Bayes method, its results are nearly the same as for the case-only test in all analyses. For both models of the smaller studies SLRI and GLC, the top 100 SNPs are even identical, for CE-IARC and MDACC, 1-3 discordant SNPs occurred. We also observe a really high correlation of both tests with the simple two-step method and the approach of Mukherjee. For MUK the concordance in NE is stronger than in MH. The simple two-step method also tends to the same effect, but not as strongly as MUK. Around 75-90 of our empirical hierarchical Bayes top 100 SNPs for never vs. ever smokers are in the top 100 of MUK, while we have 60-70 for moderate vs. heavy smokers. For TWO, we observe 48-72 common SNPs with EHB for NE and 41-64 common ones for MH.

In particular, when comparing the empirical hierarchical Bayes approach to the traditional case-control test of interaction, we see a strong difference between never vs. ever and moderate vs. heavy. While the similarity is even lower than 10% for never vs. ever smokers, we have 20-40 common SNPs in the moderate vs. heavy model, constituting

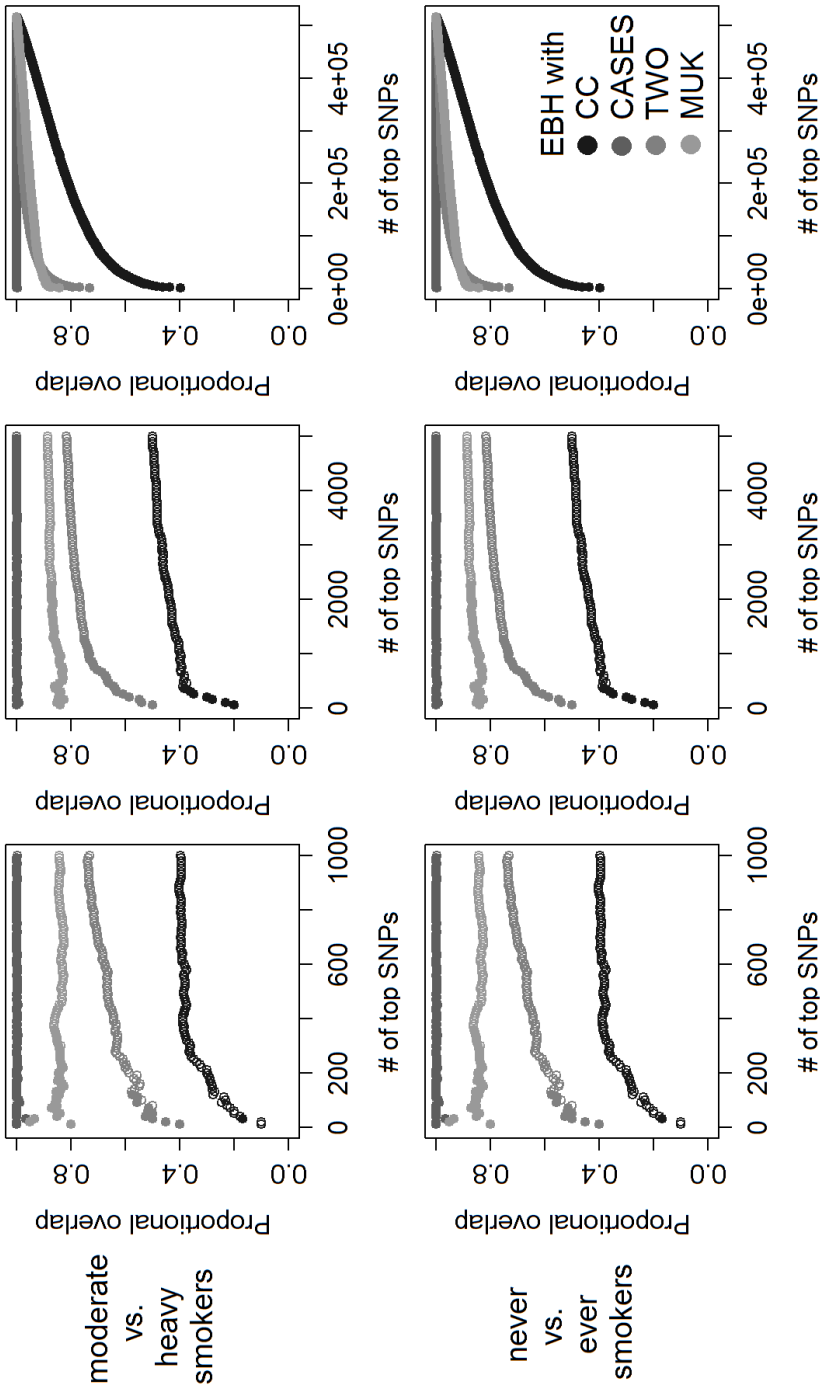


Figure 7.17: *List comparison plots of the EHB SNP ranking with rankings of different other GxE interaction methods for GLC. The y-axis shows the proportion of common SNPs for a particular number of top SNPs given on the x-axis.*

a higher consistency. Notably, in SLRI we see a strongly decreased number of common SNPs for both models compared to the other three data sets. A possible reason for that may be the low number of cases and controls. The similarity between Mukherjee and Chatterjee's (2008) method and the case-control test is only slightly enhanced in comparison to the empirical hierarchical Bayes method. On the contrary, the simple two-step method shows a much larger overlap of top 100 SNPs with case-control of 40-60 SNPs for moderate vs. heavy smokers and around 60 SNPs for never vs. ever.
Although Murcray's method has shown to be more powerful than other GxE interaction methods (Murcray *et al.*, 2009; Mukherjee *et al.*, 2012) for none of the analyses a significant result occurs. This is not surprising, since even the very powerful but biased case-only test showed nearly no such results. We have no common top SNPs of that method with case-control for any of the analyses. For never vs. ever, the overlap is limited to only a few SNPs with other methods as well. However, for moderate vs. heavy at least up to 17 common SNPs with empirical hierarchical Bayes are observed. Hence, while comparing the other methods with each other, never vs. ever shows the larger overlap, the effect is reversed for Murcray's method.
Taking a closer look at the top 100 SNPs of case-only and empirical hierarchical Bayes method, the ranking order stayed nearly constant. Only some single neighbor entries switched their ranking positions (results not shown). We went further and took a look not only at the top 100 SNPs of the different methods, but considered the overall ranking as well. In figure 7.17 we see the list comparison plot of case-only, case-control, Mukherjee's and the simple two-step method with our new EHB for the top 1000, top 5000 and all SNPs exemplarily for GLC. For never vs. ever smokers we observed a higher consistency of EHB and CC for the different studies than for moderate vs. heavy smokers within the top 5,000. Note, this holds for case-control and case-only as well, since EHB and CASES are highly correlated. While we start with 10-30% common SNPs with never vs. ever and go up to 30 to 40% within first 1,000 and even 40-50% for the top 5,000, moderate-heavy starts with 0-20%, stabilized at around 20% for the top 1,000 in GLC, MDACC and CE-IARC and increases only slightly up to 30% within top 5,000. For SLRI, the consistency reached only 5% within top 1,000 only around 15% for top 5,000. For all analyses, a strong increase of consistency is only seen in the plots considering all SNPs. The case-only and EHB overlap is from the beginning at around 100% and keeps that level with slight deviations only for all of the studies and both models. The consistency of MUK as well as TWO and the EHB lies somewhere between case-only and case-control. In all analyses, Mukherjee starts with a higher consistency to EHB than the two-step method. However, since the overlap with EHB ranking increases stronger for TWO than for MUK, we see a reverse of that effect in each case. In general, this reversing is earlier seen in moderate-heavy than for never-ever. For CE-IARC and SLRI, we see that switch for moderate-heavy already after the top 1,000 and 4,000 SNPs, for the other studies it is somewhere around 10,000. For Mukherjee's method we observe for GLC, that for the top 50 SNPs we have a slightly higher consistency to EHB (around 10% more), that then decreases a little bit, before it increases slowly again. This is in particular outstanding in GLC analyzing moderate vs. heavy smokers. Mukherjee starts here at around 70% and goes then back a little bit, before it stabilizes at 60-65%. In comparison to all other analyses, we see this effect for the two-step method in this analysis as well and even more strongly. The 9 of the

Table 7.14: *Comparison of top 100 genes between smoking models and studies for case-control test (CC) (upper triangle) and HBP-GxE (hierarchical Bayes prioritization based on GxE interaction effects, see section 7.7) (lower triangle).*

			CC							
			GLC		CE-IARC		MDACC	SLRI		
			MH	NE	MH	NE	MH	MH	NE	
HBP-GxE	GLC	MH	-	1	0	2	2	3	0	
		NE	1	-	2	2	1	0	1	
	CE-IARC	MH	3	0	-	4	4	3	3	
		NE	1	3	8	-	2	2	3	
	MDACC	MH	2	0	2	6	-	1	1	
	SLRI	MH	2	7	4	0	0	-	3	
		NE	2	6	5	10	1	2	-	

10 first ranked SNPs are identical and even 80% of the top 30 are the same. However, considering a larger amount of top SNPs, the consistency decreases to around 50%, where it nearly stays for the top 1,000.

7.6.3 Comparison of top genes between studies

When comparing the top 100 genes per GxE method between the different lung cancer studies, we see a low number of common genes not exceeding 5 in all cases. Furthermore, the consistency between the results of the studies is similar for all different methods. Hence, we do not see any method harmonizing the different study results. For the traditional case-control test of GxE interaction and our new empirical hierarchical modeling approach, the results are shown in the upper triangles of tables 7.14 and 7.15.
The number of genes occurring for at least two studies within the top 100 genes varies per method between 14 and 26. In total we have 73 such genes across the different methods. 40 of these are replicated by one method (CC, TWO, MUK or MUR), 18 by two different methods (mainly CASES and EHB or CC and TWO), 7 genes by CASES, EHB and MUK or TWO, 1 gene by CC, MUK and TWO. We observe 6 genes supported by four different methods. One gene occurred in the top gene lists for two different studies for all methods with exception of Murcray. Five genes were identified with one method only (MUR,2xCC,2xMUK), but for three different studies.

7.6.4 Resulting SNPs and genes

Taking a look at the case-control GxE test results with a $p - value \leq 10^{-5}$, we have a noticeable signal of 3 SNPs (*rs4563628*, *rs7708669*, *rs4392618*) on chromosome 5 for CE-IARC MH. Two of these SNPs are within 500kb +/- of the gene *TAG* (*tumor antigen gene*, miscellaneous RNA), that interacts with *TP53*, the third SNP is close to *CTNND2*, involved in cell adhesion. Another signal of two SNPs (*rs145910*, *rs4939359*) for the same analysis is identified in gene *OR4C15* of chromosome 11. For never vs. ever smokers, two SNPs (*rs404074* and *rs403746*) on chromosome 21 between micRNA gene *LOC100506471* and protein coding gene *PSMG1* had p-values$\leq 10^{-5}$. In the SLRI

Table 7.15: *Comparison of top 100 genes between the different smoking models and studies for EHB (empirical hierarchical Bayes) (upper triangle) and EHB-PW (empirical hierarchical Bayes integrating pathway information, section 7.7) (lower triangle).*

			EHB							
			GLC		CE-IARC		MDACC	SLRI		
			MH	NE	MH	NE	MH	MH	NE	
EHB-PW	GLC	MH	-	2	1	0	2	0	0	
		NE	1	-	0	1	1	3	2	
	CE-IARC	MH	0	0	-	4	2	0	2	
		NE	0	1	3	-	0	0	1	
	MDACC	MH	6	1	1	0	-	0	4	
	SLRI	MH	0	0	2	2	2	-	5	
		NE	12	0	0	1	8	3	-	

data, 3 SNPs in gene *AGBL1* on chromosome 15 were identified for moderate vs. heavy smoker (*rs11631489*, *rs1452454*, *rs4608306*), and two signals on chromosome 18 showed up for SLRI never vs. ever smokers with *rs12956176*, *rs1403762*, *rs1880113*, *rs9646509* and *rs4486983* of gene *KLHL14* and *rs1005419* and *rs573399* in *TXNL1*.
The SNP that showed up as genome-wide significant testing never vs. ever smokers in GLC was *rs13244987*, with neighboring SNP *rs13438768*. Both SNPs belong to the miscellaneous RNA gene *LOC645249* located on chromosome 7. The 3 SNPs that formed the prominent signal on chromosome 13 are *rs7982922*, *rs10492573*, *rs10492572* of the gene *ENOX1*. In GLC moderate vs. heavy smokers testing, 3 SNPs in gene *TRPM3* (*rs656875*, *rs1421156*, *rs672801*) on chromosome 9 were at the top positions. For SLRI never vs. heavy smokers, four of the SNPs in gene *KLHL14* on chromosome 18 that were in the top for case-control test, showed up in case-only as well (*rs12956176*, *rs4486983*, *rs1880113*, *rs9646509*). For moderate vs. heavy smokers, two SNPs (*rs1876761*, *rs9927953*) close to *WWOX* on chromosome 16 are worth to mention. For CE-IARC, two regions with three SNPs each for moderate vs. heavy smokers were noticeable, *rs2302591*, *rs175891*, *rs175888* in gene *TTLL5* of chromosome 14, that functions as a co-regulator in gene induction and repression and *rs2112783*, *rs3803716*, *rs200528* in gene *TNRC6A* on chromosome 16, responsible for gene silencing.
On our list of genes replicated by any of the different GxE methods (within top 100 for at least two studies), the genes *CSMD1*, *EML6*, *TRPM3*; *F3*, *ID100653216*, *MIR548G* and *MIR548X2* were identified by CASES, EHB and MUK or TWO, *DNAH5* by CC, MUK and TWO. The genes *ATP8A1*, *DAB1*, *ERBB4*, *KCNIP4*, *CDH2* and *LRRC16A* were supported by four different methods. *SPRY2* was in the top gene lists for two different studies for all methods with exception of MUR. We have five genes that were identified with one method only, but for three different studies. These are *CTNND2* (MUR), *FRMD4A* and *ID83879* (CC), *KCNIP4* and *MAGI2* (MUK).

7.7 Analysis of GxE interaction effects integrating pathway information

Since the integration of pathway information can improve the results of GWAS single SNP analyses, it may also help to find GxE interactions. Therefore, we developed two strategies to detect GxE interaction effects integrating pathway information into the hierarchical Bayes framework.

The first, **"global" strategy** uses the HBP for pathways as described in section 4.4, but replaces the SNP main effect statistics T_{M_i} as an input for the model by the according traditional GxE effect statistics $T^{cc}_{M_i}, i = 1, \ldots, N_M$. Hence, the mixture distribution in level 2 is composed of $p_{M_i}\sigma\chi_1(e_{M_i})$ that represents a GxE interaction, and $(1-p_{M_i})\delta(0)$ representative for no interaction effect. In the third level, general pathway information is incorporated as prior information to strengthen GxE interactions that show only a minor effect on their own but occur in the same pathway.

Level 1 $\quad | \, T^{cc}_{M_i} \, | \, | \;\; \lambda_{M_i} \sim \quad \chi_1(\lambda_{M_i})$

Level 2 $\quad \lambda_{M_i} \mid p_{M_i}, e_{M_i}, \sigma \;\sim\; \boxed{p_{M_i}}\,\sigma\chi_1(\,\boxed{e_{M_i}}\,) \;+\; (1-p_{M_i})\delta(0)$

GxE interaction — no GxE interaction

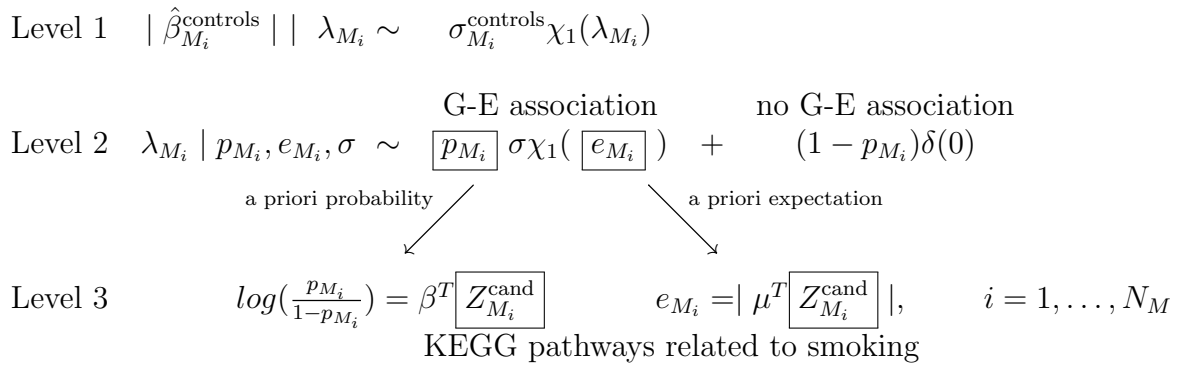
a priori probability — a priori expectation

Level 3 $\quad log(\frac{p_{M_i}}{1-p_{M_i}}) = \beta^T \boxed{Z_{M_i}} \qquad e_{M_i} = | \, \mu^T \boxed{Z_{M_i}} \, | \qquad i = 1, \ldots, N_M$

all KEGG pathways

For comparison purpose, the GSEA based on the interaction statistics instead of main effect results was performed as well.

The second, **"candidate G-E association" strategy** to combine the integration of pathway information with GxE interaction effects incorporates biological pathway information into our new GxE interaction version of the empirical hierarchical model EHB as described in section 6.3. The aim of that proceeding is to support the plausible control for population-based G-E associations. Therefore, pathways with a known or highly expected relation to the considered environmental factor smoking (e.g. nicotine dependency) are incorporated into the model to strengthen the G-E associations within such candidate pathways. SNPs involved in such smoking related pathway should rather have a population-based G-E association, then SNPs in no such pathway. In our hierarchical Bayes model for the detection of GxE interactions (section 6.3, formula 6.9) we change the proportion p of SNPs with a population-based G-E association to individual SNP properties of having a population-based G-E association p_{M_i}. Furthermore, the strength of association, given by the noncentrality parameter of the prior, is varied to individual values e_{M_i}. A third level including pathway information as in the original HBP is added to the model, so that the prior probability of a G-E association and the corresponding strength depends on its candidate pathway membership:



In the following we will abbreviate the first, global strategy by HBP-GxE since it applies the HBP to GxE interaction test statistics. Strategy two, the candidate G-E association strategy is shortly denoted as EHB-PW since it is our new EHB approach extended by pathway information.

7.7.1 Pathway hyperparameter estimates

Global GxE pathway integration - HBP-GxE
Of the 234 investigated pathways, between 93 and 193 had positive β-regression coefficients for the different pathway analyses and between 164 and 208 had positive μ-regression coefficients. For each analysis 77-91% of the pathways had the same sign of β and μ, except for MDACC with only 65%. For consistency between studies, we observed 88 pathways with positive sign of β for all 7 analyses, and 22 with only negative β regression coefficients. For μ we had 150 (=64%) pathways with only positive and 15 with only negative signs. The prior probability of association for SNPs belonging to none of the pathways was in the range of 10^{-8} to 10^{-11}, again with an exception for MDACC with 10^{-25}. For never vs. ever of GLC, the μ_0, the prior strength of association for SNPs involved in none of the pathways, was 0 and 0.0215 for moderate vs. heavy, for the others studies it ranged from 0.64 to 2.32.

Candidate G-E association pathway integration - EHB-PW
While for the GxE pathway integration strategy 1, HBP-GxE, the pathway information should strengthen GxE effects based on the traditional test of interaction, the second strategy uses pathways with a known or highly expected relation to the environmental factor smoking (candidate G-E association pathways) to support the plausible control for population-based G-E associations. Hence, a positive regression coefficient indicates in that case an increases of the prior probability of population-based G-E association effect and an increase of this association effect strength.
Of the 40 candidate G-E association pathways considered, for each of the smoking models and studies at least 30 lead to an increase in the G-E association effect of its involved SNPs. For never-ever in CE-IARC, all pathways had a positive μ coefficient. Between 28 and 31 pathways contributed positively to the prior probability of a population-based G-E association. The exact numbers for the different analyses are shown in table 7.16. The prior probability of population-based G-E association for a SNP included in none of the pathways is really high with nearly 50% for all of the different analyses. The basic non-centrality parameter for a SNP in none of the

Table 7.16: *Characteristics of the hyperparameter estimates of GxE pathway integration strategy 2. $i = 1, ..., N_M$ with N_M number of SNPs. $k = 1, ..., N_S$ with N_S number of considered pathways.*

	GLC		CE-IARC		SLRI		MDACC
	MH	NE	MH	NE	MH	NE	MH
$\# \mu > 0$	32	31	37	40	32	31	35
μ_0	10^{-6}	0.00066	10^{-6}	0.01261	10^{-6}	10^{-6}	10^{-6}
$min(\mu_k)$	-0.11282	-0.16372	-0.97989	-1.0478	-0.22673	-0.12249	-0.66864
$max(\mu_k)$	0.22062	0.33700	1.11496	1.57405	0.21946	0.93395	0.99900
$min(\mu Z_{M_i})$	-0.23434	-0.23885	-1.01674	-1.03519	-0.22673	-0.19777	-0.66864
$max(\mu Z_{M_i})$	0.37351	0.50189	10.05976	15.12712	0.44207	1.16255	3.03554
$\# \beta > 0$	30	30	28	28	31	28	28
β_0	0.50205	0.50468	0.50000	0.500000	0.51066	0.50000	0.50000
$min(\beta_k)$	0.07006	0.05672	$1.29 \cdot 10^{-8}$	$8.7 \cdot 10^{-25}$	0.10934	0.03521	0.00821
$max(\beta_k)$	0.67267	0.67711	0.50000	0.50000	0.70343	0.58191	0.50098
$min(\beta Z_{M_i})$	0.00104	$7.53 \cdot 10^{-5}$	$8.56 \cdot 10^{-41}$	$3.41 \cdot 10^{-128}$	0.00388	$5.05 \cdot 10^{06}$	$1.58 \cdot 10^{-08}$
$max(\beta Z_{M_i})$	0.72949	0.682	0.50000	0.50000	0.84124	0.583	0.501

pathways given a G-E association however is extremely low - with 10^{-6} for never-ever of GLC and CE-IARC, MDACC and both models for SLRI. This value was given as a lower bound for μ in the nonlinear optimization method estimating the hyperparameters from the marginal likelihood. For moderate-heavy of GLC and CE-IARC, the basic effect is higher with 0.00066 and 0.0126. Although the prior probability of G-E association is that high the really low μ values cause no remarkable difference to the case of no G-E association. Taking a look at the contribution of the single pathways to the prior probability of association and size of the corresponding effect, we observed values between the ranges given in table 7.16. The minimal prior probability of G-E association for a SNP that is involved in exactly one of the pathways ranged between 10% for moderate-heavy of SLRI and both models of GLC, and is nearly 0 for both CE-IARC analyses. The corresponding maximal prior probability stays close to the basic prior probability with 50% for CE-IARC and MDACC and 60-70% for SLRI and GLC. The noncentrality parameter of model level 2 maximally increased to 1.57 for CE-IARC NE and around 1 for CE-IARC MH and MDACC MH. For SLRI and GLC, values between 0.2 and 0.35 are maximal reached. Accounting that the SNPs may occur not only in one of the pathways but several, we observed for the different SNPs lower as well as higher prior probabilities than these values. For μ, values up to 10 for CE-IARC MH and 15 for CE-IARC NE can occur. For the other analyses, only a slight increase is possible.

7.7.2 Comparison of top pathways between studies

Global GxE pathway integration - HBP-GxE

Comparing the overall rankings between the different smoking models and studies pairwise, the correlation with respect to β is larger than according to μ for nearly all situations. In particular, both models of CE-IARC and moderate-heavy of SLRI and

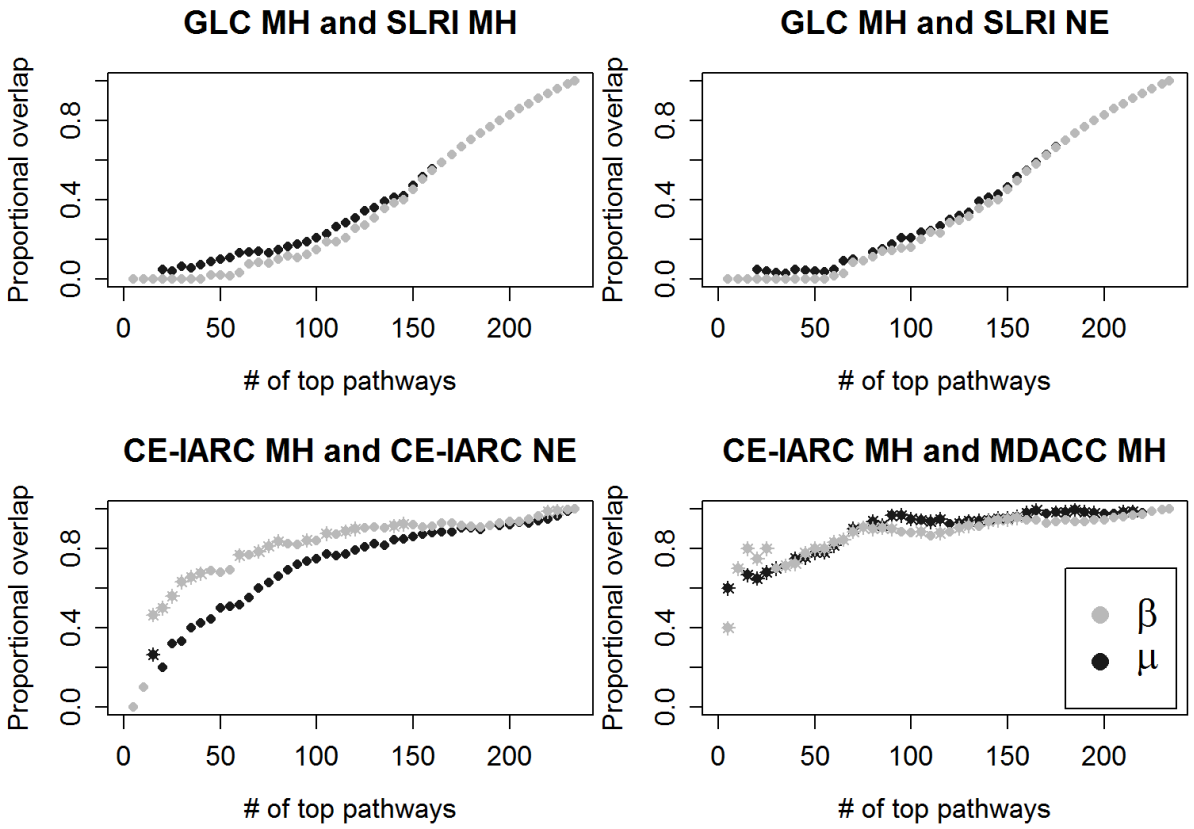


Figure 7.18: *List comparison plots of the HBP-GxE pathway ranking between different studies. The y-axis shows the proportion of common pathways for a particular number of top pathways given on the x-axis. The stars indicate a significant overlap.*

GLC show a high consistency for β as well as μ. The same is observed for never vs. ever smokers in GLC and SLRI. For β, we also see a high correlation of never-ever in GLC and SLRI with their moderate-heavy analysis and both smoking models of CE-IARC. The effect is larger for GLC than for SLRI. With respect to μ however, we have no mentionable significant correlation. The ranking of the MDACC β coefficients has much in common with GLC never vs. ever. This does not hold for the corresponding μ rankings. For MDACC and the remaining analyses we see a moderate correlation with respect to β, and no consistency for μ. Figure 7.18 shows the list comparison plots for some combinations of analyses exemplarily.
In Appendix tables B.6 and B.7 lists of the pathways are given, that belong to the top 10 pathways for at least two different studies. These are 9 pathways using β as pathway ranking criterion and 14 using μ. For the μ ranking, none of the top 10 pathways of MDACC occurred in any of the other studies' top 10. Nevertheless, 8 of the 14 pathways replicated with respect to μ ranking occurred for all other 3 studies. In general, we see for these that most either occur in the top 10 of moderate-heavy of GLC, CE-IARC and SLRI, or never-ever of GLC and SLRI. Hence, we have not much mix between pathways for never-ever and moderate-heavy. The only exception is seen for the CE-IARC study, which may be due to the much larger size of that study in

Table 7.17: *Number of nominal significant pathways for different analyses with GSEA based on GxE interaction effects.*

	GLC	CE-IARC	MDACC	SLRI
NE	7	10	-	17
MH	17	10	17	7
NE ∩ MH	1	1	-	0
NE ∪ MH	23	19	17	24

comparison to GLC and SLRI. In particular, the number of never-smokers is generally low. One pathway was in the top 10 even for both models of GLC, CE-IARC and SLRI. That pathway was found on the β replicated pathway list as well, were it ranked in the top ten for all analyses despite of GLC (rank 11). In general, all replicated β pathways were in the top 10 of never-ever GLC and moderate-heavy CE-IARC. Four more pathways were in the top 10 of all four studies, with three of them on the μ-list as well. Furthermore, one more pathways is found on the μ and β list. This pathway and all additional pathways on the β list were significant not in 2 but 3 different studies. In addition, all pathways on that list are at least in the top 20 for all studies, despite of twice a ranking of 29. In contrast, in the μ list we observe for the given pathways also ranks in a higher double-digit up to triple digit level. This is in particular the case for the pathways that are on the list due to moderate-heavy of SLRI and GLC and affects mainly never-ever of SLRI, GLC and MDACC. The pathways supported by never-ever of SLRI and GLC reach relatively high ranks for moderate-heavy of GLC and both models of CE-IARC most of the time.

Candidate G-E association pathway integration - EHB-PW
Taking a look at the consistency of pathway ranking between the different studies, we see that both models of CE-IARC show nearly the same results with respect to μ and a significant overlap for the μ top 5 pathways. The overall β ranking furthermore looks similar for SLRI MH and GLC MH, SLRI MH and GLC NE and GLC MH and GLC NE. A significant overlap of β top 5 pathways is furthermore observed for SLRI NE with these 3 analyses. The MH analysis for MDACC has a significant number of β top 5 pathways with both SLRI and GLC MH. For GLC NE and MDACC MH we do not observe a significant overlap of top 5 with respect to β, but for μ.

Gene set enrichment analysis
The gene set enrichment analysis identified overall two pathway as significant according to FDR (≤ 0.05) in SLRI never vs. ever smokers. The corresponding enrichment scores were driven by 21 and 20 out of the 37 and 49 involved genes. However, both pathways did not even reach nominal significance for any of the other studies. Several pathways reached nominal significance for each of the studies ($p_{\text{nominal}} \leq 0.05$).
The number of significant pathways for never vs. ever smokers and moderate vs. heavy smokers of each study, as well as the overlap between both smoking models per study is given in table 7.17. In figure 7.19 we see the overlap between the studies. The overlap is significant for none of the study pairs. In total, 11 pathways were identified in two different studies. None was significant in three or four of the studies. Between MDACC

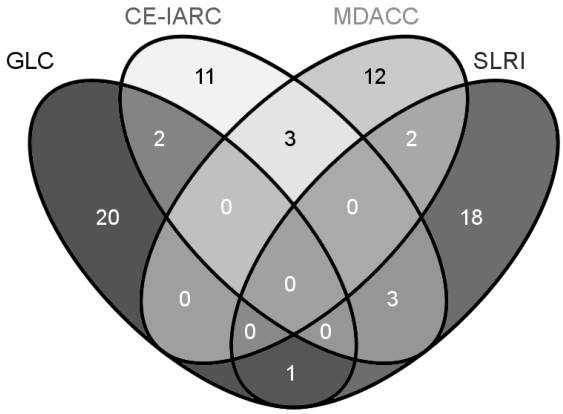


Figure 7.19: *Overlap of nominal significant pathways for different lung cancer studies with GSEA based on GxE interaction effects*

and GLC we found no common pathways. In the appendix table B.8 a list of these 11 pathways and their corresponding nominal p-values can be found. We see that for most of the pathways that are on the replicated pathway list due to GLC, CE-IARC or SLRI, the significance was both times for NE or both times for MH. This was observed for HBP as well. However, the pathways of MDACC that were found in another study as well not necessarily occurred there in moderate vs. heavy smokers, but also using the never-ever smoking as environmental factor.
Of the pathways on the replicated GSEA list (table B.8), we find none of them in any of the two HBP-GxE lists (tables B.6 and B.7).

7.7.3 Pathway results

For the global GxE pathway integration strategy HBP-GxE, the *taurine and hypotaurine metabolism* was in the μ and β top 10 for both smoking models of all three studies, except for GLC MH with a β rank of 11. *One carbon pool by folate*, *steorid biosynthesis*, *folate biosynthesis* and *sulfur relay system* were in the β top 10 of all four studies, with the three latter pathways on the μ replicated pathway list as well. Furthermore, *riboflavin metabolism* is found on the μ and β list.
The gene set enrichment analysis identified *Tryptophan metabolism* pathway (*hsa00380*, FDR = 0.04) and the *Taste transduction* pathway (*hsa04742*; FDR = 0.037) in the SLRI NE as significant according to FDR (≤ 0.05).

7.7.4 Comparison of top SNPs/genes between methods

Global GxE pathway integration - HBP-GxE
For the GxE pathway integration strategy 1, E_M was used for SNP re-ranking as done for main effects. We observed the same behavior as before - with E_M and P_M highly correlated to each other and E_M^+ having much lower similarity to both. The results for this are not shown.
While for the GxE interaction methods compared in the previous section, the consistency between methods was generally larger for never vs. ever analysis than for

moderate vs. heavy, we see the reverse effect for strategy 1 integrating pathway information with GxE interaction effects. In tables 7.12 and 7.13 the comparison of the top 100 genes of HBP-GxE to the other GxE interaction methods from the previous section are exemplarily shown for GLC and CE-IARC. The highest overlap is given with case-control test of GxE interaction, what is not surprising since exactly this statistic is used as input for the HBP-GxE model. Considering the moderate vs. heavy analyses, the largest overlap is seen for SLRI, with 82 common SNPs. GLC and CE-IARC show a moderate consistency with nearly 60 shared SNPs. With respect to the further GxE interaction test, this pathway integration strategy has a moderate overlap with the simple two-step-method as well and no common SNPs with Murcray's method. For the other methods, only up to 15 SNPs overlap occur. MDACC behaves different, with even having only 2 common genes with case-control test. For the never vs. ever smoker analyses, we see the same trend as in moderate vs. heavy, with CC and TWO having the highest overlap with the pathway integration method. However, for CE-IARC these are no more than a handful of SNPs, at most 12 for SLRI and even 24 for GLC.

GxE candidate G-E association pathway integration - EHB-PW

For the second strategy integrating pathway information with GxE effects, we see that for some of the analyses the top 100 SNPs are nearly identical to the ones with the hierarchical empirical Bayes approach not integrating pathway information (tables 7.12 and 7.13). This is the case for never vs. ever smokers of CE-IARC and GLC, as well as moderate-heavy of CE-IARC. Hence, we have no additional benefit due to the pathway information for these analyses. For GLC and MDACC with environmental factor moderate vs. heavy smoking as well as SLRI never vs. ever however, the situation looks totally different. Here, EHB-PW leads to totally different top SNPs in comparison to all other methods, with only a hand full of genes in common (table 7.12 exemplarily for GLC moderate-heavy). For the remaining analysis of SLRI with moderate vs. heavy smokers, the results lie somewhere in between. We have 60 common genes with case-only method and empirical hierarchical Bayes without pathway information and 40 shared SNPs with Mukherjee and the two-step method. The overlap to the further methods can be neglected (results not shown).

Gene set enrichment analysis

On the diagonal of table 7.18 we can see the total number of LES genes extracted from the pathways with a nominal $p \leq 0.05$ in the GSEA based on GxE interaction effects for each of the analyses. Several hundred LES genes are available in each case. The intersection of LES genes between the analyses of never-ever and moderate-heavy smokers is shown in the same table and is highly significant in every case. Comparing the LES genes with the results of the other different GxE interaction methods, we find only a handful of these genes on the top rankings. The results are given in 7.19. Exceptions are both GxE pathway integration strategies for GLC NE and MDACC MH. For both models of SLRI, we see at least a tendency to that effect as well, for GLC MH and CE-IARC NE only for the global pathway integration strategy HBP-GxE.

Table 7.18: *Overlap of LES genes for GxE GSEA between the different analyses*

		GLC		CE-IARC		MDACC	SLRI	
		NE	MH	NE	MH	MH	NE	MH
GLC	NE	194	25	26	12	28	28	13
	MH		245	19	39	26	53	26
CE-IARC	NE			159	35	37	49	17
	MH				462	85	46	95
MDACC	MH					368	64	41
SLRI	NE						444	66
	MH							332

Table 7.19: *Overlap of LES genes of GSEA based on GxE interaction statistics with top 100 genes of other GxE methods*

	GLC		CE-IARC		MDACC	SLRI	
	MH	**NE**	**MH**	**NE**	**MH**	**MH**	**NE**
CC	5	3	1	5	9	2	3
MUR	0	1	1	1	4	0	2
CASES	2	4	3	2	3	2	3
EHB	2	4	3	2	3	2	3
EHB-PW	2	12	3	1	21	5	7
HBP-GxE	7	23	9	0	13	6	5

7.7.5 Comparison of top genes between studies

Global GxE pathway integration - HBP-GxE

Comparing the results of the different studies to each other, we observe that the consistency is slightly increased in using the global GxE pathway integration strategy in comparison to the case-control GxE test only. The results are contrasted in table 7.14. However, the effect is less than observed for main effects before. The consistency of results is higher in never vs. ever smokers than in moderate vs. heavy smokers.
Considering the top 100 genes for each of the smoking models and studies, we observed in total 644 different genes, with 38 observed for two different studies and 3 for 3 studies. Of the genes occurring in the top 100 for two different studies, 14 were found in the combination GLC and SLRI, 13 in the combination SLRI and CE-IARC.

Candidate G-E association pathway integration - EHB-PW

For the candidate G-E association pathway integration strategy sometimes nearly none of the top 100 SNPs changed compared to the same approach not integrating pathway information (EHB), sometimes it leads to totally different results than all other GxE methods. The corresponding numbers are shown in table 7.15. Between SLRI NE, GLC MH and MDACC MH we observe that some more common genes show up. However, for the remaining analysis pairs, we see no improvement with still only 1-2 common genes. For EHB-PW, we obtained a total list of 660 different top 100 genes among the smoking models and studies, with 29 genes identified in two different studies - mainly GLC-MH and SLRI-NE, and SLRI and MDACC. Only two genes were found for three different studies.

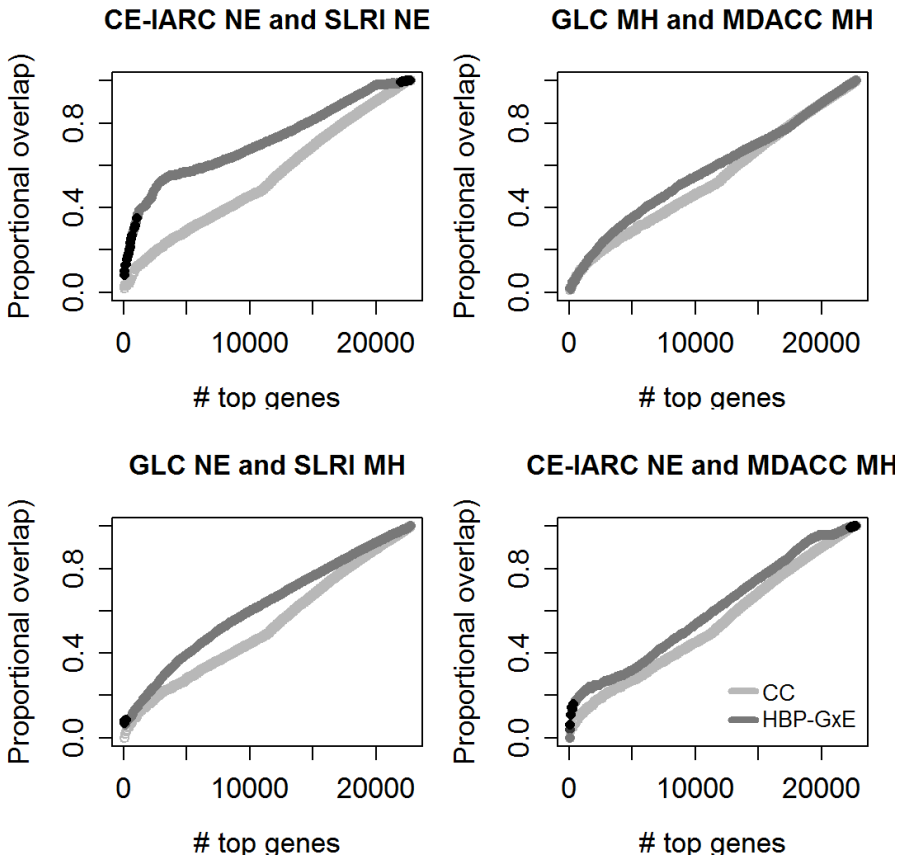


Figure 7.20: *List comparison plots of gene rankings between the different studies for* ***GxE case-control*** *test and* ***HBP-GxE****. The y-axis shows the proportion of common genes for a particular number of top genes given on the x-axis. The darker points indicate a significant overlap.*

When comparing the overall ranking consistency between different studies, we see no mentionable difference between CASES, TWO, MUK and our new EHB. However, comparing the HBP method based on GxE interaction effect (HBP-GxE) with the case-control test, a clear gain of consistency by this additional usage of pathway information is reached. In figure 7.20, the comparison of both methods for some chosen study/model combination is shown, representing the overall behavior. In particular remarkable is the comparison of the ranking lists for the CE-IARC study and SLRI never-ever analyses, with a strong increase of consistency between the two studies especially on the top rankings, followed by SLRI NE and CE-IARC NE with MDACC MH, as well as GLC NE with both models of CE-IARC and SLRI MH.
Taking a look at EHB-PW, we see minor increases of common top genes for some comparisons of studies - as for moderate-heavy of GLC and MDACC (figure 7.20 upper right graphic), for other situations nothing changes (left graphics) or even worsens (lower right graphic).

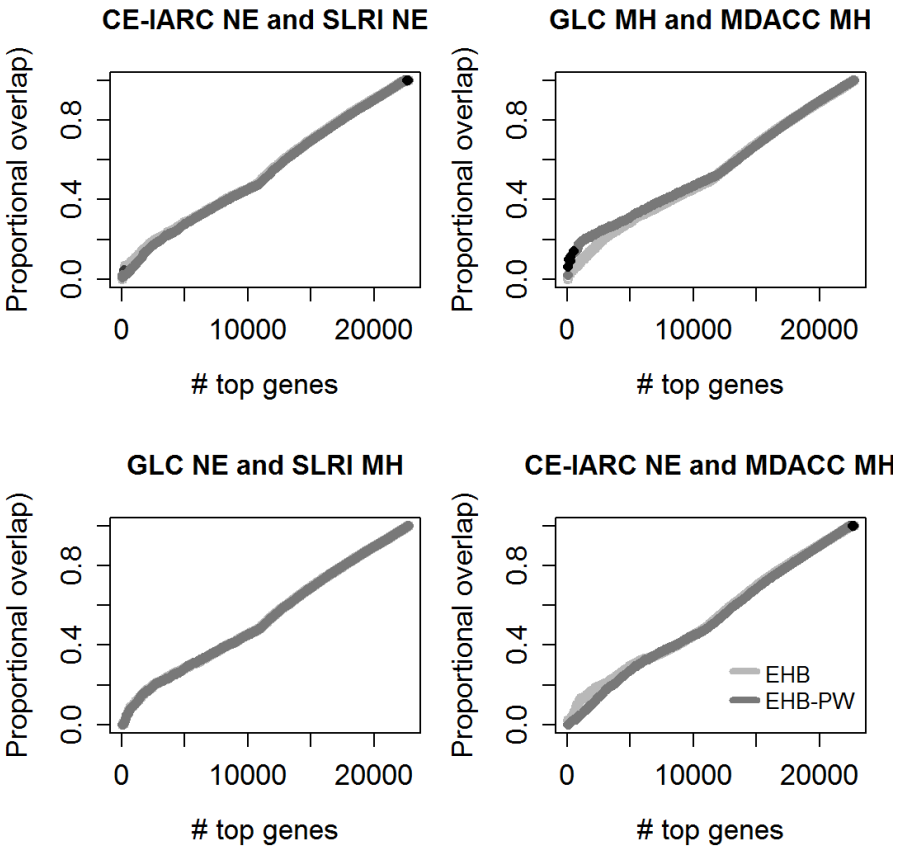


Figure 7.21: *List comparison plot of gene rankings between the different studies for* ***GxE EHB*** *test and* ***EHB-PW****. The y-axis shows the proportion of common genes for a particular number of top genes given on the x-axis. The darker points indicate a significant overlap.*

Gene set enrichment analysis

In table 7.18, the overlap of LES genes between the different studies for GSEA based on interaction effects is given. Measured by the total number of genes occurring within the considered 234 pathways, the overlap of LES genes is significant for all situations except for never ever of GLC with moderate heavy of CE-IARC and SLRI.
In total, 1607 different LES genes occurred. Of these, 333 were LES genes for 2 different studies and 49 were found by 3 different studies. We have twelve genes identified in the LES of all four studies.

Taking a look at the replicated genes of HBP-GxE and EHB-PW across different methods, we find 3 of the HBP-GxE genes for other methods replicated as well. Furthermore, 7 genes were in the replicated LES genes. 5 of the genes replicated with EHB-PW were in the replicated gene set of CASES and EBH as well. Of these, 4 were additionally in the MUK replicated set, one in TWO and two in GSEA. One more gene occurred in the replicated set of MUR and GSEA. Furthermore, we have 10 more replicated LES genes in the EHB-PW. One gene was replicated with MUK and GSEA. For genes replicated with different methods, not necessarily the same studies were

underlying. For example, a gene was on the replicated gene list of MUR due to GLC and SLRI, for HBP-GxE however GLC and CE-IARC uncovered the gene. A situation like this is observed in nearly half of the cases. A gene was even found in GLC and SLRI with HBP-GxE, but MDACC and CE-IARC with GSEA. In addition, while one strategy may find a genes due to never vs. ever testing, it may be the moderate-heavy testing for another method. Furthermore, both phenomena occured together as well.

7.7.6 Resulting genes

For the global GxE pathway integration strategy HBP-GxE, the three genes observed for three different studies within the top 100 are *ELOVL6* (SLRI NE, MDACC MH, CE-IARC NE), *RFC3* (NE of GLC, SLRI, CE-IARC) and *WRAP53* (SLRI NE, GLC and CE-IARC MH). With the candidate G-E association pathway integration strategy EHB-PW, *CADM1* and *ERBB4* were found for three different studies (GLC MH, MDACC MH and SLRI NE).
Taking a look at the replicated genes of strategy 1 and strategy 2 across different methods, we find 3 of HBP-GxE genes for other methods replicated as well, namely *ACCN1* (MUR), *MAML2* (CC and TWO) and *PDE10A* (CC). Furthermore, 7 genes were in the replicated LES genes (*ATP1B2*, *BAAT*, *CLTCL1*, *GRIN2A*, *MAGOHB*, *SLC8A1*, *THBS2*). 5 of the genes replicated with EHB-PW were in the replicated gene set of CASES and EHB as well (*DAB1*, *ID100653216*, *MIR548G*, *NRG3*, *ERBB4*). Of these, 4 were additionally in the MUK replicated set, one in TWO and two in GSEA. One more gene, *PLCB1*, occurred in the replicated set of MUR and GSEA as well. Furthermore, we have 10 more replicated LES genes in the strategy 2 set (*ADCY8*, *CCL1*, *CTNNA3*, *CTNNB1*, *CXCL13*, *ITGA11*, *ITGA9*, *SRC*, *TLR4*, *VAV3*). One gene was replicated with MUK and GSEA (*TGFB2*).
For genes replicated with different methods, not necessarily the same studies were underlying. *ACCN1* for example was on the replicated gene list of MUR due to GLC and SLRI, for HBP-GxE however GLC and CE-IARC were responsible. A situation like this is observed in nearly half of the cases. *MAGOHB* was even found in GLC and SLRI with strategy 1, but MDACC and CE-IARC with GSEA. In addition, while one strategy may find a genes due to never vs. ever testing, it may be the moderate-heavy testing for another method, e.g. for *ATP1B2* (MH of GLC and CE-IARC for strategy 1, NE of these studies for GSEA). Furthermore, both phenomena occurred together, e.g. for the gene *GRIN2A*, identified by HBP-GxE in CE-IARC NE and GLC MH and by GSEA in MDACC MH. For GSEA, the twelve genes identified in the LES of all four studies are *ADCY2*, *ADCY8*, *ADCY9*, *CREBBP*, *GRIN2B*, *ITPR1*, *PIK3R1*, *PIK3R5*, *PLCB1*, *PRKCA*, *PRKCB* and *PTK*.

7.8 Discussion

The integration of pathway information and GxE interaction in four lung cancer GWAS lead to new prospective insights into the development of the disease. Candidate genes and pathways that may be involved in lung cancer etiology were identified and should be further investigated.

In particular, when integrating pathway information with main effects in the hierarchical Bayes prioritization approach, several pathways were found as main contributors to the prior probability of association and to the corresponding association effect strength across three or four different studies. The consistency of top genes between the different studies was clearly increased compared to the initial single SNP regression results.
In the analysis of GxE interactions, EHB lead to similar results than the powerful case-only test. The case-only test is biased in the presence of G-E association on a population level. In our simulation studies, we found similar results for EHB and case-only when no or only a low number of weak G-E association effects were given. Although population-based G-E associations were expected in the context of smoking, we did not observe strong G-E association effects in our data.
As observed for main effects, the integration of pathway information with GxE interaction effects by the HBP increased the consistency between the top genes of the different studies. Again, for each of the different studies, similar pathways were found on the top 10 using the β and μ regression coefficients as ranking criterion. When integrating the candidate G-E association pathways to the EHB to support the correct control for the population-based G-E associations, we observed no change in the top 100 genes for some of the analyses compared to EHB without pathway information. However, for some other analyses totally different results appeared, with more common top genes between the studies.

Analysis of main effects integrating pathway information

In table 7.20 the pathways consistently identified for all four studies with our HBP or at least 3 studies with GSEA or SUMSTAT are given with some further biological information. Nearly all can be somehow related to lung cancer risk and partly have been even reported to be associated before. This biological plausibility of the results supports the strength of our methods integrating pathway information.
In a recent publication of Fehringer *et al.* (2012), four different pathway approaches were compared using the same lung cancer GWAS as well. For this comparison, GSEA, SUMSTAT, SLAT and the modified Fisher test were chosen, since they are all widely used and representative for others. Fehringer *et al.* (2012) built two data sets by combining Central Europe study (CE-IARC) with the Toronto study (SLRI) (CETO) and German (GLC) with MDACC study (GRMD) to reach adequate sample size and higher statistical power. As pathway information, gene ontology level 4 pathways (Ashburner *et al.*, 2000) with 15-200 genes were used. SNPs were assigned to genes within a region of +/- 20kb. A logistic regression was performed, assuming an additive SNP effect, adjusted for sex, age and country of origin. As criterion for a replicated pathway, a FDR$\leq$ 0.05 in both data sets was used.
Table 7.21 gives the main results of Fehringer *et al.* (2012). With the GSEA, none of the pathways reached a FDR$\leq$ 0.05. This fits to the results we have seen in our analyses. In comparison to us, they found several pathways with SUMSTAT, what is explainable by the merge of the studies, larger sample size and hence higher power. Note, we considered different pathway information, based on manually curated KEGG pathways (Kanehisa and Goto, 2000), while Fehringer *et al.* (2012) used the predominantly bioinformatically generated gene sets in GO. This may lead to discrepancies in results

Table 7.20: *Interesting pathways replicated based on SNP main effects*

Linolenic acid metabolism, *alpha-Linolenic acid metabolism*	essential fatty acid α-Linolenic acid = omega-3 acid and γ-Linolenic acid = omega-6 acid; Yehuda *et al.* (2005): association to lower risk of cardiovascular disease; reduces anxiety, stress levels and cortisol levels; preliminary research: omega-3 fatty acid supplements may decrease inflammation and improve lung function in some people with asthma
cysteine and methionine metabolism	methionine = essential a-amino acid that cannot be synthesized by organisms; cysteine = semi-essential a-amino acid that can be biosynthesized based on methionine; Johansson *et al.* (2010): inverse association of methionine with lung cancer; Pöschl and Seitz (2004); Salaspuro *et al.* (2006); Sprince *et al.* (1975): cysteine eliminates mutagen and carcinogen acetaldehyde, that e.g. occurs due to smoking and drinking; Salaspuro (2003): acetaldehyde in particular associated with alcohol related gastrointestinal tract carcinogenesis
Taurine and hypotaurine metabolism	taurine = derivative of cysteine, an amino acid; typotaurine = intermediate in biosynthesis of taurine; act as endogenous neurotransmitter; Maher *et al.* (2005): taurine attenuates peripheral apoptosis and cell death of T cells; taurine helps slow the loss of lymphocytes associated with skin and renal cancer cells; potential use of taurine in cancer immunotherapy
Riboflavin metabolism	known as vitamin B2; Bassett *et al.* (2012): vitamine B and methionine reduces lung cancer risk, higher riboflavin intake associated with lower risk of lung cancer among current smokers
Folate biosynthesis	known as vitamin B9; responsibility: DNA synthesis, repair and methylation; Johansson *et al.* (2010): lung cancer prevention; combination of Vitamin B6, methionine and folate reduces the chances of lung cancer
Ascorbate and aldorate metabolism	Known as vitamin C; acts as an antioxidant by protecting the body against oxidative stress
non-homologous end joining	Kanehisa and Goto (2000): repair of double-strand breaks in DNA to maintain genomic stability in response to irradiation; Tseng *et al.* (2009): asssociation with nonsmall cell lung cancer
RNA polymerase	RNA polymerase: constructs RNA from DNA in transcription; Huang *et al.* (2010): regulation of DNA polymerase POLD4 influences genomic instability in lung cancer
cholinergic synapse	choline = primary component of the neurotransmitter acetylcholine, binds among others to nicotinic acetylcholine receptor
Neuroactive ligand-receptor interaction	involves several genes of nicotinic acetylcholine receptor family proteins, e.g. CHRNA3 and CHRNA5
HTLV-I infection	HTLV-I: pathogenic retrovirus; Kanehisa and Goto (2000): associated with adult T-cell leukemia/lymphoma (ATL), implicated non-neoplastic chronic inflammatory diseases; Nomori *et al.* (2011): association to bronchioloalveolar carcinoma
Wnt signaling pathway	signal transduction pathway; initiates cell reaction to external signals; Mazieres *et al.* (2005); Tennis *et al.* (2007); Uematsu *et al.* (2003): critical pathway in lung carcinogenesis as already demonstrated in many cancer, in particular colorectal cancer
Steroid biosynthesis	Chaudhuri *et al.* (1982): receptors for all classes of steroid hormones identified in cytosols of adenocarcinoma of the lung

Table 7.21: *Number of significant pathways in Fehringer et al. (2012) for different GSA methods. CETO = CE-IARC+SLRI, GRMD = GLC+MDACC*

	modified Fisher	**GSEA**	**SUMSTAT**	**SLAT**
CETO	7	0	8	2
GRMD	5	0	1	0
CETO ∪ GRMD	2	0	1	0

as well. The pathway identified by SUMSTAT in CETO and GRMD - the *acetylcholine receptor activity pathway* - for example is not involved in our analyses. However, this pathway is a biologically highly plausible candidate, since it involves the *CHRNA3* region that is known to be associated with lung cancer (Amos *et al.*, 2008; Hung *et al.*, 2008b) and nicotine addiction (Thorgeirsson *et al.*, 2008).
The two replicated pathways for the modified Fisher test were the *nerve impulse* pathway and *Ras-GEF*. By investigating the influence of the number of SNPs per gene to the pathway results, Fehringer *et al.* (2012) explored that the modified Fisher method more likely detects pathways with a greater median number of SNPs per gene. This is not surprising since the test statistic of the top SNPs per gene is used representatively and thereby genes with a greater number of SNPs tend to higher association statistics (gene size bias). Since a normalization routine and phenotype permutations was used for GSEA and SUMSTAT, these methods were protected against the bias. SLAT uses all SNPs in a pathway for the analysis and a phenotype shuffling routine.
For the hierarchical Bayes prioritization approach we found a very high consistency in top pathways between the four different studies. Based upon the investigations of gene size in Fehringer *et al.* (2012), we analyzed the influence of the number of SNPs per pathway to its ranking according to β or μ coefficient. In figure 7.22, the β and μ coefficients for the GLC study are plotted as a function of the number of SNPs of the corresponding pathway. In the appendix figure B.3 the corresponding plot for CE-IARC can be found. We see a clear tendency to higher regression coefficient given a lower number of SNPs. Analyzing this behavior with a linear regression model with β/μ as the outcome and number of SNPs as the dependent variable confirmed that impression. We found a highly significant association for both models of all studies: with increasing number of SNPs, the β and μ regression coefficients decrease. This is not surprising, since an increasing number of SNPs indicates a higher number of SNPs without any effect even given a pathway with some associated SNPs. Hence, for future use of the HBP, it is highly recommended to consider the size of pathway in the model, e.g. by appropriate weights in the pathway covariate matrix. We did not consider this in our analyses presented here.

Fehringer *et al.* (2012) noted that their identified pathway was only driven by signals within the CHRNA3 region on chromosome 15q25 (Amos *et al.*, 2008; Hung *et al.*, 2008b). When this region was removed from the analysis, the pathway lost its significance. However, when interested in the pathways resulting from a pathway based analysis, the goal is not to find a pathway based on a single gene or signal only. Furthermore, one may not be interest in the biological pathways itself but proceed on SNP or gene level in further investigations with results supported by the pathway

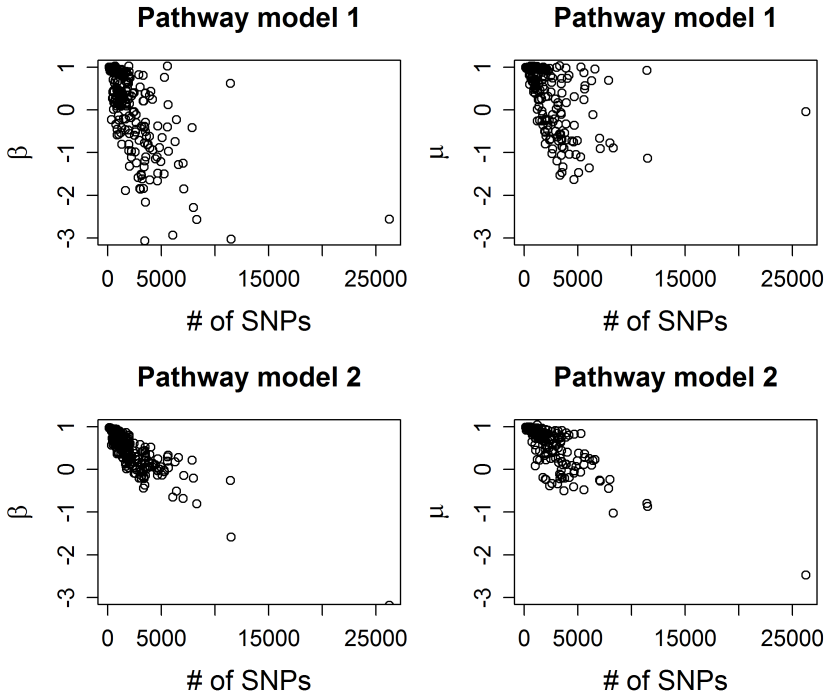


Figure 7.22: *Correlation of HBP β and μ regression coefficients and number of SNPs per pathways in GLC study*

information. For the GSEA method, one may use LES genes. For SUMSTAT we selected the top genes at a ranking list containing only the genes from the nominal significant pathways. However, a main drawback of that proceeding is that only SNPs involved in any of the pathway have a chance for follow-up. SNPs too far from any gene and genes involved in none of the considered pathways are neglected and may be missed although having a high effect. For our given pathway set, only 1/4 of the genes are involved. Hence, the hierarchical Bayes prioritization may have a big advantage benefitting from the pathway information but keeping all SNPs as potential units for further investigation. In our data application, we have seen that in comparison to the initial regression ranking the consistency of top genes between the studies was increased and hence genes found in at least two different studies are promising candidates. In table 7.22 a list of these candidates obtained by HBP is given. Many of these are biologically plausible as well. To name just a few: HES1 for example controls cell proliferation (Murata *et al.*, 2005) and is therefore a plausible candidate in lung cancer. FEN1 was found to be over-expressed in lung tumors (Nikolova *et al.*, 2009) and functional polymorphisms were found to be associated with DNA damage and lung cancer risk (Yang *et al.*, 2009). The oncogene RAB13 may be another potential candidate. RPA1 and RPA2 play an essential role in the replication, recombination and repair of DNA. RPA2 was found to be implicated in non-small cell carcinoma (Murphy and Borowiec, 2010). RFC4 was identified in a meta-analysis as over-expressed in lung adenocarcinoma (Erdogan *et al.*, 2009), and LIG1 participating in DNA repair was recently shown to be modestly associated with lung cancer in smokers (Sakoda *et al.*, 2012).

Table 7.22: *Genes occurring in the top 100 for at least 3 different studies based on HBP analysis of main SNP effects*

ACIN1	DPCR1	GTF2H4	LIG1	NUP210L	RPA1	SMPDL3B
ACP2	EIF4A2	HES1	MADD	PIGZ	RPA2	SMYD4
BTK	FADS1	HNRNPH2	MIOS	PLA2G4C	RPA3	SNORA4
C1orf38	FADS2	ID100529097	MIR135A1	PRKAA2	RPA4	SNORA63
CDH24	FEN1	ID152217	MIR1908	PSMB11	RPL36A	SNORA81
CREB3L4	GCK	ID729852	MIR611	PSMB5	RPS27	SNORD2
DDB2	GLA	ID746	NCBP2	RAB13	SENP5	VARS2
DIAPH2	GLYCTK	JTB	NR1H3	RFC4	SFTA2	WDR82

Table 7.23: *EHB hyperparameter estimates for the different smoking models and studies.* $p = \exp(\beta)/(1+\exp(\beta))$

	GLC		CE-IARC		MDACC	SLRI	
	NE	MH	NE	MH	MH	NE	MH
μ	$3 \cdot 10^{-6}$	10^{-6}	0.017	0.066	0.079	10^{-6}	10^{-6}
p	0.004	0.005	0.152	0.039	0.003	0.004	0.006

Analysis of GxE interaction effects with and without integrating pathway information

In our application of the different GxE interaction methods to the lung cancer data, we found a high similarity between case-only test and our new empirical hierarchical Bayes approach. Since the case-only approach is prone to false positive results due to population based G-E associations, and the empirical hierarchical Bayes approach is designed to be protected against that bias, these results are not obvious.
However, when taking a look at the controls only test of G-E association, no striking signals were found pinpointing to a population-based G-E association, although such effects are expected in this particular context with smoking as an environmental factor. Taking a look at the estimates of hyperparameters of the empirical hierarchical Bayes method, it is not surprising that the results are very similar to the ones of case-only test. The estimates are shown in table 7.23. For the analyses of GLC and SLRI, the μ estimate representing the strength of a population-based G-E association are less than 10^{-5}. Combined with prior probabilities of association of around 0.005, the posterior estimates λ_{M_i} for the single markers M_i that are subtracted of the $\beta_{M_i}^{\text{cases}}$ are close to 0. As a consequence the top 100 SNPs are the same as for case-only. For the larger studies MDACC and CE-IARC, the μ estimates reach values with a little more impact. For never-ever of CE-IARC, $\mu = 0.017$, for moderate-heavy of CE-IARC and MDACC $\mu = 0.066$ and 0.079. One possible reason is that the size of the study is responsible for the larger μ values, and the larger estimates for MH in comparison to NE may be due to a stronger effect of population-based G-E associations in that case. The β estimates and hence estimates of the prior probability of a population-based G-E association are larger for CE-IARC and MDACC as well. After all, for CE-IARC and MDACC we see only 1-3 SNPs not identical between CASES and EHB top 100.
Integrating the pathway information with the GxE interaction effects, we see for the global GxE pathway integration strategy HBP-GxE again a slight improve in consis-

Table 7.24: *Number of top 100 genes involved in smoking related pathways*

		HBP-GxE	CC	CASES	TWO	MUK	MUR	EHB	EHB-PW
GLC	**M1**	24	14	11	15	11	11	11	9
	M2	25	10	6	5	9	10	6	62
CE-IARC	**M1**	20	5	4	6	4	9	4	4
	M2	11	14	8	9	8	10	9	10
MDACC	**M2**	7	15	15	12	14	16	15	69
SLRI	**M1**	20	12	8	9	11	7	8	65
	M2	12	7	12	5	11	9	12	31

tency of the top genes between the different analyses analogously to the main effects. Furthermore, interesting top pathways occurred, e.g. the *taurine and hypotaurine metabolism* and *folate biosynthesis*. Both pathways are listed in table 7.20 with some additional information about their connection to lung cancer. The former was within the top 10 for nearly all studies with respect to β and μ coefficient. Taurine is known to reverse damage done by smoking (Fennessy *et al.*, 2003). The latter occurred within the β top 10 for nearly all analyses and twice in the μ top 10, for GLC and SLRI never vs. ever smokers. The pathway is related to smoking since folate status is decreased by chronic cigarette smoking (Piyathilake *et al.*, 1994).

For the EHB-PW pathway strategy integrating smoking related candidate pathways, for some of the analyses we observed no difference in top 100 SNPs compared to EHB without pathway information. One explanation could be that none of the genes involved in these pathways occur at the top. The numbers of genes from the smoking related pathways involved in the top 100 for the different GxE methods are shown in table 7.24. The case-only and EHB method have even less of these genes in their top 100 than the case-control method in most of the analyses. This supports that the case-only method does not seem to have false positive results due to G-E associations on a population level in our application.

Furthermore, regarding our candidate G-E pathway strategy EHB-PW for the situations where the pathway information changed the top 100 rankings we see that the numbers of genes involved in smoking related pathways increase highly. Although this was not expected, thinking about the nature of the model more sophistically, it is not so surprising. The SNPs involved in the integrated pathways have different prior probabilities and prior effect strengths of population-based G-E association compared to the main bulk of general SNPs. However, this change can go both directions and therefore does not necessarily mean an increase of the population-based G-E association and hence a decrease of final test statistic. While some of the smoking related candidate pathways may turn out to be true G-E effect clusters and hence will support the correct control for G-E association of the corresponding SNPs , this may even completely switch for pathways that involve SNPs with less evidence for a population-based G-E association than the main part of SNPs involved in none of the incorporated candidate pathways, so that the genes of these pathways will even occur at our top positions.

For the situations with no difference between the EHB and EHB-PW top results, we also compared the overall rankings of both methods. An example for GLC never-ever

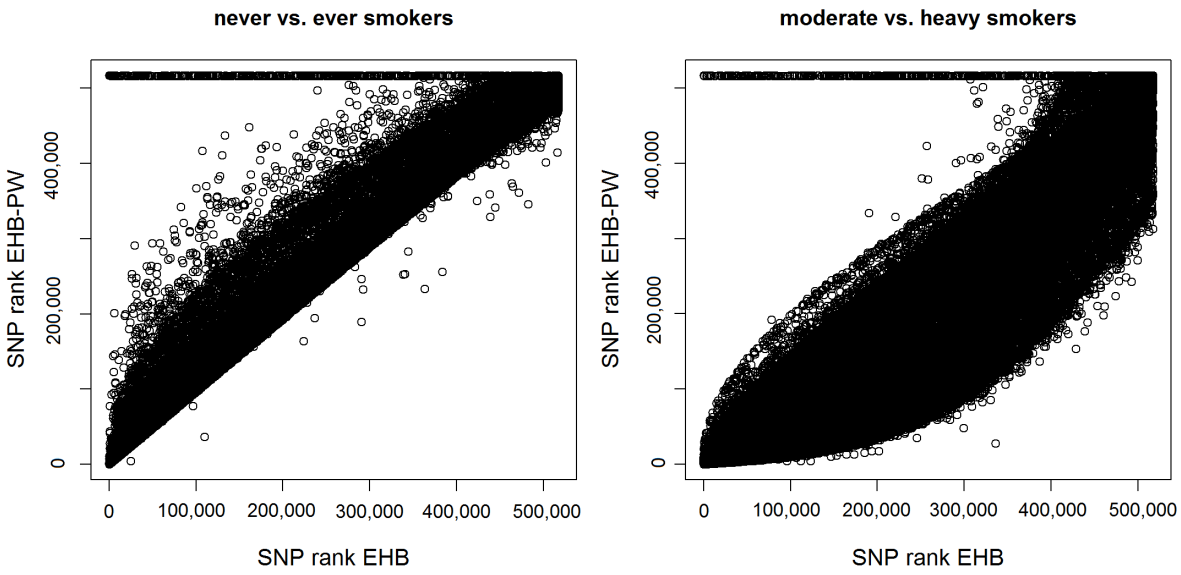


Figure 7.23: *Comparison of SNP ranks between EHB and EHB-PW for both smoking models of GLC.*

can be seen in figure 7.23 on the left. Although the top positions did not change, a lot of re-ranking can be observed all over the whole list. In particular noticeable is the line on the top of the plot. This indicates, that a set of SNPs distributed across the whole list with EHB was down-ranked to the last ranks with EBH-PW. For comparison, we also checked the over all ranking list for the situations with a large difference of EHB and EHB-PW, e.g. GLC MH in figure 7.23 on the right. Here, the re-ranking all over the list is more extremely, and again this "top line " is observed.

Overall, one of our main conclusions from the analyses of the different lung cancer studies is that the pathway information can indeed improve GWAS findings. This is in particular supported by the biological plausibility of the results. We see that different methods lead to different results and as already stated by Fehringer *et al.* (2012) so far none of the pathway analysis methods can be clearly established as superior. Therefore they should be considered for confirming results but also complementing these. However, the high harmonization of the results between the different studies is a remarkable characteristic of the HBP approach and made HBP a very useful tool in this lung cancer application. Furthermore, depending on the goal of the pathway integration method, a gene set method or the HBP method may be preferred. The HBP may be in particular preferred when concentrating on the follow-up of a subset of SNPs or genes of a GWAS as is currently done as the next step in TRICL, while GSEA is perhaps a little bit more intuitive when the main interest is in pathways itself. Nevertheless, also on the pathway level HBP has shown promising and plausible results. Although the results of the GxE interaction analysis using our new EHB are nearly the same as the potentially biased case-only results, the good performance of that method is supported since no obvious population-based G-E associations can be seen. Furthermore, these results have biological plausibility as well. This indicates that

no false positive results due to G-E associations are given in the top of CASES and EHB and that EHB can reach the same true positive results as the powerful CASES approach. Integrating pathway information with GxE effects may improve the results as seen for main effects and therefore should be investigated as complement to the simple single SNP GxE interaction analysis.

8 Summary and Outlook

Complex diseases such as cancer result from a complicated interplay of multiple genetic and environmental factors. To unveil their genetic component, simple single SNP analysis as done in genome-wide association studies is not sufficient. Complementary approaches considering the complexity of disease, such as the incorporation of biological pathway information or detection of gene-environment interaction, are necessary.

In this thesis we focused on an empirical hierarchical Bayes model, the hierarchical Bayes prioritization (HBP), originally proposed by Lewinger *et al.* (2007) for the integration of external information into genome-wide association studies. We used the approach to incorporate biological pathway information and provided a new test for GxE interaction by adapting the method for that purpose. In an application, we furthermore integrated pathway information with GxE interaction effects by two different strategies.
The integration of pathway information by the HBP in a GWAS for Rheumatoid Arthritis that was characterized by an extremely high number of strong association signals supported the initial results. In an application to four lung cancer GWAS it led to higher consistency of resulting top genes between the different studies compared to the initial ranking. In both cases, the strength of the HBP approach was supported by the biological plausibility of the results.
In comprehensive simulation studies, our new empirical hierarchical Bayes approach (EHB) for GxE interactions outperformed other GxE methods, having high power to identify GxE interactions in the presence of population-based G-E associations. In our real data application to the lung cancer GWAS, no strong G-E associations with smoking as environmental factor were observed. Therefore, the EHB led to similar results as the powerful case-only test. By additional integration of pathway information with GxE effects, the consistency of results between the studies increased. Due to the biological plausibility of the results, good candidates for further investigation were identified.

In the Rheumatoid Arthritis GWAS we integrated pathway information by the empirical hierarchical Bayes model (Lewinger *et al.*, 2007) and compared it the gene set analysis (GSA) method GSEA. The hierarchical Bayes prioritization (HBP) approach uses the pathway information to relate the different genetic markers to each other, so that SNPs in the same pathway can support each other to up-rank. GSA focuses on the identification of whole sets of genes rather than single markers and became popular for GWAS in the last years. For the comparison of the methods, genes were ranked using the posterior probability of association for HBP and the leading edge subset genes for GSEA. Pathways were ranked by the β regression coefficients of HBP that represent the increase or decrease of the prior probability of association for each SNP involved in the corresponding pathway and by nominal p-value for GSEA.
Due to the high importance of the HLA-gene region in Rheumatoid Arthritis, a high number of very strong genetic main effects was detected in that study by simple single SNP analyses. Therefore, with HLA-SNPs included in the HBP analysis, other SNPs with small to moderate effects had little chance to reach the top rank positions. Thus the top ranking SNPs changed only slightly compared to the initial ranking based on prior covariates. However, although we have no gain of information by the HBP in that case, it supports the findings from the initial regression analysis by keeping the prominent role of the HLA genes. To obtain some additional findings not detected by the single

SNP analysis, the HLA region had to be neglected in the HBP analysis. In contrast to the hierarchical Bayes prioritization, the gene set analysis method highlighted many new non-HLA genes that may be a good starting point for further research. Therefore, in that particular context, GSA methods should be preferred.

Based on the same hierarchical model, Volk *et al.* (2007) proposed an idea for a test of gene-environment interactions. By borrowing G-E information across all SNPs of a GWAS, posterior estimates for the G-E association within controls are obtained, characterized by a reduced variance. These posterior estimates are subtracted of the G-E association within cases, so that the empirical hierarchical Bayes method is a compromise between the case-control and case-only test of GxE interaction. The test reaches high power to detect markers with a GxE interaction effect, while correcting for G-E associations on the population level.
Incorporating this idea we build a test statistic usable in real applications by calculating an appropriate variance using a variance approximation of Kass and Steffey (1989). Furthermore, based on distributional considerations, we modified the idea of Volk *et al.* (2007) to obtain better properties of the statistic.
In comprehensive simulation studies, our new empirical hierarchical Bayes method was compared to several other GxE interaction methods, including the traditional case control test and the powerful but biased case-only test. In all situations, compared to other GxE methods our new test has similar or higher ranking power (ranking power = finding the GxE SNP within the top ranking SNPs) to detect an interacting SNP when strong population based G-E associations or a high number of these associations are present. In the particular case where no or only a low number of weak G-E association effects are given, a two-step method developed by Murcray performs better - in particular for low interaction effects. Since population based G-E associations are highly expected in genome-wide association studies, we recommend to use the empirical hierarchical Bayes method so that true GxE interaction effects are not missed. In particular in a disease such as lung cancer with the very strong environmental factor smoking that is known to be associated to the genetic makeup (nicotine addiction), caution is warranted. When such a strong association is not known, Murcrays method may be used as a complement to the empirical hierarchical Bayes method.

We applied both, the HBP for pathway integration and the empirical hierarchical Bayes method for GxE interactions to four lung cancer GWAS. Furthermore, we combined both purposes by using traditional GxE interaction statistics as an input for the HBP, and integrating candidate pathway information into the EHB.
Using the HBP integrating pathway information based on single marker main effects, we observed an increase in consistency of the top genes between the different studies. This and the biological plausibility of the results encourage the benefit of using HBP for SNP prioritization. For the EHB for the identification of GxE interactions, we observe that for all four studies the top genes are nearly the same as for the case-only test. This is surprising since in our simulation studies the EHB and case-only showed similar performance only given a low number of low association effects. In that particular context however, large G-E association effects are expected. Nevertheless, looking at the G-E association effects within controls, no evidence for population based G-E associations was given.
Using the HBP based on the GxE interaction effects, a slight increase of concordance

between the studies top genes is observed compared to case-control test and the other GxE methods - although to a lower extend than observed for main effects. For the EHB integrating biological pathways, the additional information changed the SNP ranking, but not necessarily at the top positions. Also here an increased consistency between the different study results is given.

Although HBP and EHB have shown to be promising tools for complementing GWAS single SNP results, there are still open questions and further possible evaluations and developments. In simulation studies Lewinger *et al.* (2007) showed that his hierarchical Bayes prioritization method integrating external information performs well in comparison to an initial single SNPs analysis. However, to evaluate the performance of the HBP in particular in the context of pathway based analyses, simulation studies should be investigated to validate the ability of the HBP to fulfill the initial intention to find consistent but weak effects. Furthermore, the relative usefulness for pathway identification and gene prioritization compared to GSA methods should be evaluated in simulations as well (Lebrec *et al.*, 2009). Compared to gene set analysis methods, the HBP has some general advantages by nature. On the one hand, it is possible to integrate some additional external information in the HBP as well, such as information about SNP function or information from a previous analysis. On the other hand, all SNPs are considered in the pathway based analysis and have a chance to reach the top ranking positions. In gene set analysis, only these genes involved in the analyzed pathways are considered. This means a loss of a high number of SNPs and hence information and prohibits the identification of new and unexpected genes with so far unknown function.
Since the hierarchical Bayes method prefers to up-rank SNPs in pathways with a small number of SNPs, as seen for the lung cancer data, the consideration of the number of SNPs or genes per pathway within the model would be another aspect of further research. Additionally, linkage disequilibrium between the different SNPs may be integrated. The model of the hierarchical Bayes prioritization approach is designed for an input of χ-distributed test statistics with one degree of freedom. A generalization of the model for two or more degrees of freedom could be worked out, e.g. for the application of the HBP to a multilocus or haplotype analysis.

In our simulation studies to evaluate the performance of our new empirical hierarchical Bayes approach for GxE interactions, we did not carry out the situation with a genetic main effect of our interacting SNPs. We neglected these kinds of interacting markers, since these SNPs should be detected and further investigated base on the analysis of single SNP main effects. Given an important environmental factor, a subgroup analysis with respect to this exposure will likely be performed. Our main focus in this thesis lies on interacting SNPs that are missed by the analysis of single SNP main effects. Nevertheless, the influence of an additional genetic main effect may be another interesting aspect to investigate.
Furthermore, a marker may be associated to an environmental factor and have a GxE interaction effect on disease development at the same time. In our lung cancer application, the nicotine acetylcholine receptor CHRNA3 for example is involved in tobacco dependence and has been identified as a lung cancer risk factor. This situation was also not covered by our simulation studies and may be investigated.
Additionally, we did not simulate linkage disequilibrium between the different SNPs. The behavior of the method given dependencies between the different observed SNPs

is a another aspect that may be investigated in prospective simulations. Furthermore, linkage disequilibrium may be integrated into the hierarchical model of the EHB.
In our empirical hierarchical Bayes model for GxE interactions, we did not consider covariate information. However, there may be the need to adjust for confounding factors, e.g. a correction for population stratification may be necessary. Therefore, one further challenge is the extension of the model to that case of having additional covariates included in the analysis. Population stratification is an issue for GxE if the population membership is associated with the disease, the genetic marker and the environmental factor (Engelman *et al.*, 2009). The impact of population stratification in the detection of GxE interactions considering different GxE tests is currently investigated in our group (Abstract to be published in Annals of Human Genetics, EMGM 2012).
Beside we restricted with our EHB to a binary classification of the genetic and environmental factor. A generalization of the method to an additive genetic effect and more complex considerations of the environmental factor is of practical interest and an issue for further research.

As already mentioned in section 7, the analysis of further lung cancer GWAS from the participating consortium is intended in the future. It will be very interesting to see how their results approve the current findings. An aspect that will play a role in that context and constitutes an interesting research area for the future is how the HBP and EHB results across the different GWAS can be combined by a meta-analytical approach. For gene set analysis methods, a member of our group is currently working on a meta-GSA approach to combine GSA findings. The challenge is to combine the different pathway results while considering the direction of the underlying single SNP effects across studies (Abstract to be published in Annals of Human Genetics, EMGM 2012).

Appendices

A Fundamentals of genome-wide association studies and data resources

A.1 SNP databases and arrays

The human genome project
In 1990 the human genome project (HGP) coordinated by the National Institutes of Health (NIH) started with the goal to sequence the whole human genome and to identify all approximately 25,000 genes in the human DNA. In 1998, as part of their last years plan, they furthermore established the goal to identify genetic variants and build a publicly available map of at least 100,000 common SNP markers by systematically cataloging them. The HGP was completed in 2003 with great success (Human Genome Project, 2003; Sham and Cherny, 2010; U.S. Department of Energy Genome Programs, 2011).

The SNP consortium
In 1999 a collaboration of large pharmaceutical companies and the UK Wellcome Trust Case Control Consortium (WTCCC) formed the SNP consortium (TSC), with the aim to discover 300,.000 SNPs of the human genome and provide them as a public resource. Finally, more than 1,8 million SNPs were discovered in total, with 1,4 million SNPs made publicly available at the end of 2001 (Thorisson and Stein, 2003; U.S. Department of Energy Genome Programs, 2011).

The Hap Map project
Seeing the good perspective of SNPs in GWAS, the emphasis was further shifted to study them in more detail. That was extensively done by the International Hap Map project, initiated in October 2002, comprising researchers from 20 groups in 6 countries (International HapMap Consortium, 2003, 2005). Aim of that project was to determine genotype frequencies of millions of common DNA sequence variants (MAF >5%) and investigate the nature of linkage disequilibrium (LD) across the entire human genome in different populations of European, African and Asian ancestry (Barrett, 2010). In phase I of the project, completed in 2005, 1,2 million common SNPs were validated in 270 individuals from Utah with European ancestry (CEU), Japan (JPT), China (CHB) and Nigeria (YRI). Block–like patterns of LD that were already observed before were verified to occur in the entire genome. Furthermore, crossing over hotspots were identified. In the second phase of HapMap that ended in 2007 (International HapMap Consortium *et al.*, 2007), the project was extended to more than 2 million additional SNPs in the same samples, to obtain more precise LD information, more insights into history of the human populations and better tag SNP selection. In the ongoing third phase, additional populations are genotyped (Barrett, 2010; Sham and Cherny, 2010). All obtained information is published in free databases, e.g. dbSNP database of NCBI (National Center for Biotechnology Information), with today nearly 12,5 million common SNPs (MAF >1%) in the dbSNP database (NCBI).

Genome-wide SNP chips
Two big companies lead the market of genome-wide SNP platforms, Affymetrix (http://www.affymetrix.com/) and Illumina (http://www.illumina.com/). While in the beginning Affymetrix chose the markers on their chips physically randomly distributed on the whole genome, Illumina based their SNP selection on the observed LD structures (Bickeböller and Fischer, 2007). The 500K Affymetrix chip, comprising 500,000 SNPs covers 65% of all known common SNPs in CEU with at least an r^2 for LD of 0.8 with one SNP, while the Illumina Human Hap 300, containing 300,000 SNPs, has an coverage of 75% (Bickeböller and Fischer, 2007). Affymetrix's 1 million SNP chip covers nearly 85% of the genetic variation in Europeans and Asians, and 62% in Africans, while the Illumina chips of the same size covers nearly 93% and 68% (Li *et al.*, 2008). Although the great majority of SNPs are shared between different populations, significant differences in allele frequencies and local LD structures due to different evolutionary development exist (Li and Wang, 2010). Because of the older history of the African population, they show more genetic diversity and tend to have less LD than European and Asians. Recently, Affymetrix introduced a new generation of chips, the new Axiom Genome-Wide array plates. These chips are population-optimized with different available plates for the European, East Asian, Chinese and African population and offer best genetic coverage of rare and common variants.

A.2 Genotype calling and data Quality Checks

Genotype calling
The main underlying principle of the chip technology is to measure hybridization intensities for the two different occurring alleles of each SNP. Therefore, short specific pieces of DNA for the different alleles of any SNP and the adjacent nucleotides (known from the human genome project) are arrayed on the small chip. These DNA sequences are called probes. Millions of copies of sample single strand DNA are put on the SNP chip and can interact with the complementary probe strands on the chip (hybridization) fitting to the present SNP alleles, but not with probes corresponding to non-occuring alleles. Afterwards, fluorescent molecules are washed over the array and stick only to hybridized DNA spots. These molecules glow when a laser shines on them, so that fluorescence signals can be read out that represent hybridization intensities for each allele per SNP and show where the sample DNA has stucked to the probes. These measures are normalized and genotypes are assigned to each individual according to the signal intensities (genotype calling) (Ziegler *et al.*, 2008).
For homozygous genotypes, one of the allele intensities is high and the other one is low, while heterozygous subjects show similar intensities for both alleles (Ziegler *et al.*, 2008). The three different genotypes can be visualized as three clouds in a scatter plot of the allele signal intensities of one SNP for all persons, with each cluster representing one genotype. Due to the high number of SNPs in GWAS, this genotype calling step has to be performed by an automated procedure (Ziegler *et al.*, 2008). For this, different genotype calling algorithms were developed (M. Inouye, 2010), e.g. Birdseed (Korn *et al.*, 2008), BRLMM (Affymetrix, 2006) or Chiamo (Wellcome Trust Case Control Consortium (WTCCC), 2007). Unfortunately, since the signal intensities between the subjects can vary due to factors like DNA concentration or degradation, different preparation

of the samples, e.g. in different laboratories, plating errors and hybridization failures of the chips, e.g. by degeneration of arrays over time, overlaps of the three genotype clouds may arise, leading to failure of the genotype calling algorithm at the cloud edges, resulting in missing or even misclassified genotypes (Teo, 2008; Weale, 2010; Ziegler *et al.*, 2008). Affymetrix quotes the call rate for their Affymetrix 6.0 chip with 99,8%, with a correct genotyping rate of 99,97% (Affymetrix, 2009).

Missingness

Since non-random missingness with respect to phenotype or genotype can lead to false positive results, checking SNPs as well as individuals for the number of missings is of high importance. The subject-wise missing frequency is also denoted as call rate. Low call rates imply poor DNA quality or hybridization problems caused by faulty arrays (Ziegler *et al.*, 2008). As a threshold for filtering out individuals with low call rates, Weale (2010) recommended 97% to 98 % as appropriate. In small studies 90% is often used. The missing rate for a SNP is a good marker for genotyping accuracy and SNP performance and hence identification of problematic SNPs (Weale, 2010). A high missing rate indicates that the genotyping of the SNP failed for a high number of probands, i.e. that the calling algorithm was not able to assign genotypes to signal intensities (Neale and Purcell, 2008). This can be closely related to the overlap of clouds illustrated above. Furthermore, since particular differences in missingness between cases and controls can lead to false positive results, comparing the missing rates between cases and controls is a popular strategy. When different study groups are considered, the missingness should furthermore be separately investigated in each of these groups (Ziegler *et al.*, 2008). A common threshold for filtering out bad quality SNPs due to missingness is between 2% and 5% (Weale, 2010).

Minor allele frequency

The data quality for SNPs tends to decrease with a decreasing minor allele frequency. Less information for genotypes with a low MAF is available for the calling algorithm, resulting in less certain results and hence only poor performance for SNPs with rare alleles (Weale, 2010). Informative missingness can affect low MAF SNPs more strongly and increase the chance for false positives. Furthermore, low MAF SNPs are generally not informative in GWAS, since the power in these studies is too low to detect associations of such low frequency SNPs (Teo, 2008; Weale, 2010). Therefore, it is reasonable to exclude these SNPs, which also reduces the multiple testing burden. Depending on the sample size MAF thresholds of 1%-5% are generally used (Ziegler *et al.*, 2008), with Weale (2010) suggesting 10/n with n=number of samples as a reasonable threshold.

Hardy Weinberg Equilibrium

Since genotype calling algorithms can not only fail to assign a genotype and produce missing values but have also the potential to make incorrect calls (Teo, 2008) leading to genotype misclassifications, identifying such kinds of conspicuous SNPs is important as well. Again, overlapping clouds is the origin of this problem. Such kinds of genotyping errors express in deviations from HWE (Teo, 2008). HWE was described in detail in chapter 2.2.1, relating the allele and genotype frequencies of a SNP to each other. Nevertheless, not only genotyping errors can cause departures from HWE, but deviations can also occur due to population stratification, selection and non-random mating. In addition, strong signals of true association can express by HWE deviation as well due to the fact that disease disposing alleles are favored in cases and hence corresponding

genotypes are enriched (Weale, 2010; Neale and Purcell, 2008; Ziegler *et al.*, 2008). Therefore, HWE is often only checked within controls (Neale and Purcell, 2008; Ziegler *et al.*, 2008). Furthermore, since extreme departures are more likely caused by failures, more stringent deviation thresholds (Neale and Purcell, 2008) are used to exclude SNPs according to HWE deviations, often the significance level as for finding associated signals is used.

Sex mismatches

A quality control on the subject level is to check if the genetic sex determined by the X chromosome data matches with the sex given in the clinical data. This can help to make sure that genetic and phenotypic data are correctly aligned (Weale, 2010). A mismatch can occur due to labeling errors, where a wrong DNA probe is assigned to a wrong clinical record, an error in ascribing the sex in the clinical record, due to sample contamination, X chromosome mosaics in females or due to rare medical conditions (e.g. Kinefelter's XXY). The latter affects only less than 0.1% of the population and hence occurs only in 1 out of 1,000 persons (Weale, 2010). The genetic sex can be determined by the extent of heterozygous and homozygous genotypes for the X-chromosomal SNPs, since no heterozygotes should occur in men. An estimate for the homozygosity is given by Wright's inbreeding coefficient F. The inbreeding coefficient F is the probability that two alleles given at a randomly chosen locus are identical by descent (IBD), what means that they are copies of the same allele from a common ancestor. An estimate can be obtained based on the observed number of homozygous genotypes versus the expected one under Hardy-Weinberg Equilibrium. A positive F indicates the excess of homozygotes, while a negative F indicates the excess of heterozygous genotypes. For females this value should be close to zero, while males show values near one, representing no heterozygous genotypes. Intermediate values are often observed in women, what can be explained by very large copy number variation deletions on the X-chromosome, but can also indicate DNA contamination (Weale, 2010). Individuals with a mismatch that cannot be clarified should be removed from the analysis (Weale, 2010). Recommended values to assign the sex are <0.2 for females and >0.8 for males.

Heterozygosity

When checking sex a measure for the homo- and heterozygosity based on the X-chromosome is calculated. Another useful indicator for poorly genotyped samples is the extent of heterozygosity based on the autosomal SNPs. Negative F values represent an excess of heterozygous genotypes due to contamination of the DNA probe (Weale, 2010; Teo, 2008; Ziegler *et al.*, 2008). An excess of homozygotes expressed by a positive F value can result from membership to different populations and inbreeding (Weale, 2010). A typical approach to exclude persons according to the heterozygosity measure is to calculate the corresponding mean and standard deviation across all subjects in the study and exclude individuals outside the range of mean +/- 3 standard deviations (Ziegler *et al.*, 2008).

Cryptic relatedness

An assumption of association statistics used in population-based studies is the independence of observational units. The relatedness of the participants in a population-based study may result in biased statistics (Weale, 2010) and false positive or false negative results. Therefore, another important issue is to check if individuals are more closely related than population average and hence are close family members (Weale, 2010). Fur-

Table A.1: *Recommended Quality Filters for population-based GWAS data*

Individual specific quality checks	
Call rate	$\geq 90\%$
Sex mismatch	female F<0.2 and male F>0.8
Heterozygosity	within mean F +/- 3 standard deviation F
Cryptic relatedness	proportion allele IBD <0.1875
Population stratification	$\lambda_{GenomicControl}$<1.05
Population outliers	| PLINKS nearest neighbor Z score | <4
SNP specific quality checks	
Call rate	$\geq 95\%$
Minor allele frequency	$\geq 1\%$
Hardy Weinberg Equilibrium	$p_{HWE} \geq 10{-}7$

thermore, in large-scale studies, sample duplications can occur by accident (Teo, 2008). To uncover cryptic relatedness, a LD pruned data set with no strong LD among the remaining SNPs has to be prepared, and the allele sharing by each pair of subjects is calculated by the proportion of alleles identical to descent (IBD) for these SNPs (Weale, 2010). Two alleles are identical by descent when they originate from the same ancestral allele. By comparing the observed proportion of alleles that are the same between both individuals (identity by state, IBS) and the corresponding expectation for two unrelated subjects given the allele frequencies, IBD can be estimated. A value of 1 denotes that all alleles are IBD and represents monozygotic twins or a replicate. For first degree relatives, we have $p_{IBD} = 0.5$, for second degree relatives $p_{IBD} = 0.25$ and so on. A commonly used threshold is p_{IBD}<0.1875.

Population outliers

The proportion of concordant alleles can furthermore be used to detect population outliers. Population outliers are characterizes by a different ethnicity than the remaining sample and can lead to biased test statistics. While the previous part concentrates to identify pairs of individuals with higher allele sharing than two unrelated persons from the same population, outliers express by an outstanding low number of alleles IBS in comparison to the rest of the sample. Population outliers have the potential to results in inflation of the test statistic and false results and should therefore be removed from the sample before the analysis of the data. A method to detect population outliers is outlined in the main text (principle component analysis, section 3.2.4). Although this method can be used to adjust for outliers as well, it is recommended to rather remove than correct for them.

Population stratification is another important point handled in the process of quality control. Due to its high importance in the association analysis as well, it is outlined in the main text of this thesis (section 3.2.4).

A.3 Pathway databases

One of the databases most commonly used and subject of our pathway based analyses in chapter 7 is the **Kyoto Encyclopedia of Genes and Genomes** (KEGG) (Kanehisa and Goto, 2000). KEGG is a free access database with structured information about biomolecules and genes, particularly involving a collection of 249 manually drawn maps of biological pathways (release 61.1, February 1 2012). These represent our knowledge on the molecular interaction and reaction networks in metabolism, genetic information processing, environmental information processing, cellular processes, human diseases and drug development. Of the more than 20,000 known human protein coding genes, around 6,200 currently occur in at least one of these pathways.
Another popular pathway collection is Biocarta (BioCarta LLC, 2011). Biocarta provides molecular relationships of genes within pathways and their interactions by dynamic graphical models. More than 120,000 genes from different species are cataloged and summarized. Information about the proteome collected by the research community is integrated.
In addition, Gene Ontology (GO) is often used (Ashburner *et al.*, 2000). GO classifies genes into a hierarchy of categories (GO terms), placing gene products with familiar function together. The GO consists of three biological domains: cellular component, molecular function and biological process. Hence, the gene products are characterizes by where they act, which function they have and in which process they are involved. In total, 34,940 terms are defined (21,401 biological process, 2,896 cellular component, 9,063 molecular function), involving 37,957 human gene products (proteins, different kinds of RNA) (GO version 1.248, September 13th 2011). To illustrate the hierarchical concept, we will take a look on "death", one kind of the biological processes. Death can be distinguished in cell and tissue death. Programmed cell death is a subitem of the former, which in turn involves apoptosis. Apoptosis can be further differentiated according to the kind of cells affected, e.g. leukocyte apoptosis. This furthermore refines into positive or negative regulation of B cell apoptosis. Due to the hierarchical structure, genes in a category are part of all parent classifications (Curtis *et al.*, 2005). To reduce the problem of high correlations between the gene sets, it is recommended to restrict to one level of the hierarchy when using GO (Wang *et al.*, 2010).
While KEGG and Biocarta are based on manually curated pathways, the gene sets in GO are predominantly bioinformatically generated (Wang *et al.*, 2010). The former yields high-quality information for well-studied pathways, the latter ensures comprehensive coverage of pathways.
The molecular signatures database MSigDB (Subramanian *et al.*, 2005) is a comprehensive collection of gene sets involving positional gene sets (chromosome, chromosome region), curated pathways from other online databases, among others KEGG and Biocarta, publications and expert knowledge, gene motif gene sets, computational gene sets defined by expression neighbourhoods and GO sets. In February 2012, the total number of gene sets was close to 7,000.
The Gene Map Annotator and Pathway Profiler (GenMAPP) (Salomonis *et al.*, 2007) is a computer application for visualization of gene groups and pathway diagrams from genomic data including software for pathway analysis. New custom pathway maps can be constructed with the tool and existing pathways and gene sets are provided for download. This archive contains 100 hand-curated human pathways created at GenMAPP.org

or submitted by GenMAPP users based on textbooks or review articles, as well as a pathways from public databases such as GO or KEGG.
A good overview of 325 different resources related to biological pathway or molecular interaction information, including links to the several databases, is given by pathguide (Bader *et al.*, 2006). Pathguide lists manually curated pathway databases as well as computationally predicted ones, open source providers as well as commercial ones and general as well as more specialized pathway databases.

The p53 signaling pathway

The p53 signaling pathway is important in cell growth and health and is responsible for genomic stability (Melino *et al.*, 2002; Soussi, 2010; Vogelstein *et al.*, 2000; Vousden and Lu, 2002). The pathway is induced by stress signals, with p53 as its key protein. The name of the corresponding gene is TP53. In normal, unstressed cells, the protein MDM2 binds to the p53 and inactivates it. p53 is a cell stress sensor molecule and responds to numerous different stress factors such as DNA damage induced by e.g. radiation or chemical agents (Soussi, 2010) or deregulations of genes responsible for cell metabolism, growth and division (oncogenes). The stress signals are transmitted to p53 by various upstream mediator proteins that induce splitting of the p53-MDM2 complex and activate the p53 as a transcription factor for numerous genes (Karp, 2002). These genes are e.g. involved in cell cycle arrest during cell division (keeping damaged cells from progressing through the cell cycle) (BioCarta LLC, 2011), DNA repair (repairing the genetic damage before the DNA replication is initiated) and apoptosis (the programmed cell death, initiated when the damage proves to be irreparable) (BioCarta LLC, 2011; Karp, 2002).
Due to the important role in cell growth, malfunctions of the p53 can lead to enormous negative consequences. A defect or missing p53 protein can inhibit the ability to bind DNA regulatory regions in an effective way and activate specific gene's expression. The cell division of damaged cells will not be stopped, DNA damage cannot be repaired and the cell is not destroyed by apoptosis (Karp, 2002), what leads to an unrestrained replication of the damaged DNA. This again involves the production of numerous abnormal cells with the potential to divide uncontrollable, form tumors and become malignant (Karp, 2002).
However, not only defects of the TP53 gene directly result in the replication of damaged DNA, but also other proteins involved in the same pathway can be responsible for the same final malfunction. Changes in MDM2 concentration or binding affinity for example can influence the p53 pathway (Michael and Oren, 2002; Soussi, 2010), as well as defects in the stress signal mediators and even failures in up- or downstream pathways.
The non-functioning of the p53 pathway destabilizes the cells genetic information (Buselmaier and Tariverdian, 1999) by allowing its division and survival despite a DNA damage (Alberts, 2004). This leads to an increased general mutation rate of that cell and its progenitors, so that mutations that promote cell proliferation or block apoptosis can arise more easily as well (Buselmaier and Tariverdian, 1999; Griffiths *et al.*, 2008).
Although the p53 pathway plays a very important role in cancer development, a defect of this pathway alone does not cause cancer. The mechanism underlying disease development is even more complicated than only a drop out of one particular pathway.

Cancer arises from a series of special mutations that accumulate in a cell and its progeny (Griffiths *et al.*, 2008) and lead to different specific properties. Beside the genomic instability and increased mutation rate achieved by p53 pathway defects, the characteristics of a malign tumor are uncontrolled cell division and cell growth (proliferation), blocked apoptosis, sustained angiogenesis, ability of invasion and metastazation (Alberts, 2004). Angiogenesis denotes the development of new blood vessels. These are necessary to provide the tumor cells with oxygen and nutrients. Invasion is the ability of a cell and its progeny to digest their way through the underlying tissue and to displace the "normal" neighbors. Metastazation is the ability of cells to get in and out of the blood or lymph circulation and spread to distant sites (Alberts, 2004).
The example of the p53 signaling pathway clearly illustrates how different proteins work together in biological pathways to fulfill a particular task. Although in some cases a damaged TP53 gene is directly responsible for the blocking of the DNA repair and apoptosis, the malfunction needs not necessarily to be a consequence of the defected tumor suppressor gene itself (Karp, 2002). Even if a "crucial" gene is not mutated, the function of this gene can be affected as a result of an alteration in other genes whose products are part of the same pathway or an upstream or downstream pathway. Proteins at different positions within a pathway can drop out due to mutations in their coding DNA sequences and regulatory regions or due to inadequate regulation of expression by transcription factors, all finally leading to the defect of the pathway.
Several examples of GxG interactions are found in the p53 signaling pathway. MDM2 binds to the p53 to form a protein complex and inactivate it, the mediator proteins interact with the MDM2-p53 complex to solve the connection, and finally p53 interacts with the DNA regulatory sequences to induce gene expression.

B Data applications

B.1 Genetic Analysis Workshop 16

Preprocessing

The Gene Name Service (GNS) (Lin *et al.*, 2007) we used to assure consistency of gene names in our application, retrieves and organizes data from different gene databases (HUGO, NCBI and GeneCard) to provide gene aliases for gene identification. In our SNP-to-gene assignment and gene set file, 99,896 different gene names occurred in total. These were ascribed to 17,513 different genes in the human genome. 4 genes had to be removed since they were deleted from the NCBI gene database or since they annotated SNPs that were located on different chromosomes according to the NCBI SNP database Database of Single Nucleotide Polymorphisms, 2009.
We further took a closer look at the gene sets to combine those with a large overlap and excluded those with less than 11 genes to reduce the multiple testing burden. Five gene sets were completely composed of other genes than the genes covered by the SNPs and 381 sets included only one of the covered genes. These gene sets were removed. We checked the overlap of any pair of the remaining 1,690 gene sets and found 772 gene sets composed by less than 11 genes being completely part of a larger gene set. This reflects the hierarchical structure of the gene ontology (GO) (Ashburner *et al.*, 2000). We excluded these small sets. To reduce the multiple testing problem in addition, we merged pairs of gene sets where at least 90% of the genes from the smaller set were shared with the larger set and these shared genes made up at least 66% of the larger gene set. By this we replaced 83 sets with 41 merged ones and end up with a total of 876 gene sets for the analysis.

Comparison of different permutation strategies

The comparison of different permutation procedures was extensively discussed in the context of gene expression (Efron and Tibshirani, 2006; Goeman and Buehlmann, 2007). However, due to the hierarchical SNP to gene and gene to set assignment, in GWAS context this challenge is even more complicated.
The permutation of phenotypes has the very important advantage that it adjusts for the different gene sizes while the complex correlation structure of the data is preserved (Wang *et al.*, 2007)). Genes vary in size and LD block structure and especially when using the maximum SNP statistic to represent a gene, the gene size is a potential source of bias. Smaller p-values tend to occur more likely in larger genes just by chance and therefore large genes are more likely to show a significant effect. As a consequence, gene sets containing large genes may be inflated (Wang *et al.*, 2007). In GWAS data, SNP correlations occur due to LD and genes may overlap or interact with each other. To keep the correlation structures between SNPs and genes when permuting potentially may increase power, while ignoring the correlations may lead to biased results (Wang *et al.*, 2007)). Unfortunately, phenotype permutation is extremely time-consuming and memory intensive. Furthermore, the hierarchical structure compassing SNP to gene assignment and gene to gene set assignment is not straightforward modeled (Wang *et al.*, 2011). The alternatives of permuting the SNP or gene statistics (Wang *et al.*, 2011) are less computationally intensive than the disease label permutation and do not

Table B.1: ***Results for our analysis of the NARAC data:*** *Top 20 gene sets after applying GSEA, one/two-step HBP, or HBP+GSEA*
[a] in two strategies; [b] in three strategies; [c] in all four strategies.

rank	Strategy I GSEA	Strategy II HBP	Strategy III HBP+HBP	Strategy IV HBP+GSEA
1	*hsa*04514[a]	*hsa*04330[a]	*GO*0032393[a]	*hsa*04940[a]
2	*hsa*04640	*GO*0032395[a]	*GO*0002504[b]	*hsa*04514[a]
3	*hsa*04612[c]	*GO*0006956[a]	*GO*0048002[b]	*GO*0008236
4	*hsa*04940[a]	*GO*0016820	*GO*0051327	*hsa*04612[c]
5	*inflamPathway*	*GO*0051028	*asbcellPathway*[b]	*GO*0032395[a]
6	*th1th2Pathway*[a]	*GO*0004004	*GO*0042287	*GO*0002504[b]
7	*CSKPathway*	*GO*0030554	*GO*0032395	*GO*0048002[b]
8	*ctla4Pathway*	*GO*0000279	*GO*0001569	*GO*0051249[a]
9	*blymphoPathway*	*GO*0051276	*GO*0051249	*hsa*04512[a]
10	*hsa*04650[a]	*GO*0019199	*GO*0002526[b]	*GO*0032393[a]
11	*tcraPathway*	*GO*0002460[c]	*GO*0006957	*th1th2Pathway*[a]
12	*GO*0048002[b]	*GO*0007160	*GO*0002460[c]	*hsa*04650[a]
13	*GO*0046982	*hsa*04940[a]	*ctla4Pathway*[b]	*hsa*04610
14	*GO*0009405	*hsa*04612[c]	*hsa*00310	*GO*0002526[b]
15	*asbcellPathway*[b]	*hsa*04010	*hsa*04512	*GO*0002460[c]
16	*hsa*04330[a]	*GO*0043069	*GO*00051169	*ctla4Path*[b]
17	*GO*0006956[a]	*GO*0002521	*hsa*04612[c]	*GO*0002443[a]
18	*GO*0002504[b]	*GO*0002443[a]	*hsa*04320	*GO*0004175
19	*GO*0002460[c]	*GO*0006281	*GO*0006643	*asbcellPathway*[b]
20	*GO*0002526[b]	*GO*0009952	*GO*0016301	*GO*0003779

need the raw genotype data. However, these proceedings violate the key assumption of permutation methods that the permuted units have to be independent from each other. SNPs are correlated due to LD - even between different genes and genes may overlap or interact. Although different individuals are distantly related as well (Wang *et al.*, 2010), this correlation is ineffectively compared to the strong SNP and gene dependencies. The permutation of both, SNP and gene-level statistics, disrupts LD patterns, does not consider GxG interactions and results in a biased null distribution. While permuting SNP statistics at least considers gene size and preserves gene correlations due to an overlap, permuting gene statistics do not keep this neither. Although ORA methods not necessarily require a permutation routine, for the purpose of considering correlation structures and other potential sources of bias, permutation methods are useful as well (Wang *et al.*, 2010).

B.2 Supplementary results of the lung cancer GWAS

This appendix gives the following additional results:

- QQ-lots of the initial SNP main effect results from MDACC and SLRI
- Table of replicated pathways of HBP based on SNP main effects using β-regression coefficients as pathway ranking criterion
- Table of replicated pathways of HBP based on SNP main effects using μ-regression coefficients as pathway ranking criterion
- Table of replicated pathways of GSEA based on SNP main effects
- Table of replicated pathways of SUMSTAT based on SNP main effects
- Manhattan plots of the initial SNP interaction effects for CE-IARC and MDACC
- Table of replicated pathways of HBP-GxE based on SNP interaction effects using β-regression coefficients as pathway ranking criterion
- Table of replicated pathways of HBP-GxE based on SNP interaction effects using μ-regression coefficients as pathway ranking criterion
- Table of replicated pathways of GSEA based on SNP interaction effects
- Correlation of HBP regression coefficients and number of SNPs per pathways for the CE-IARC study

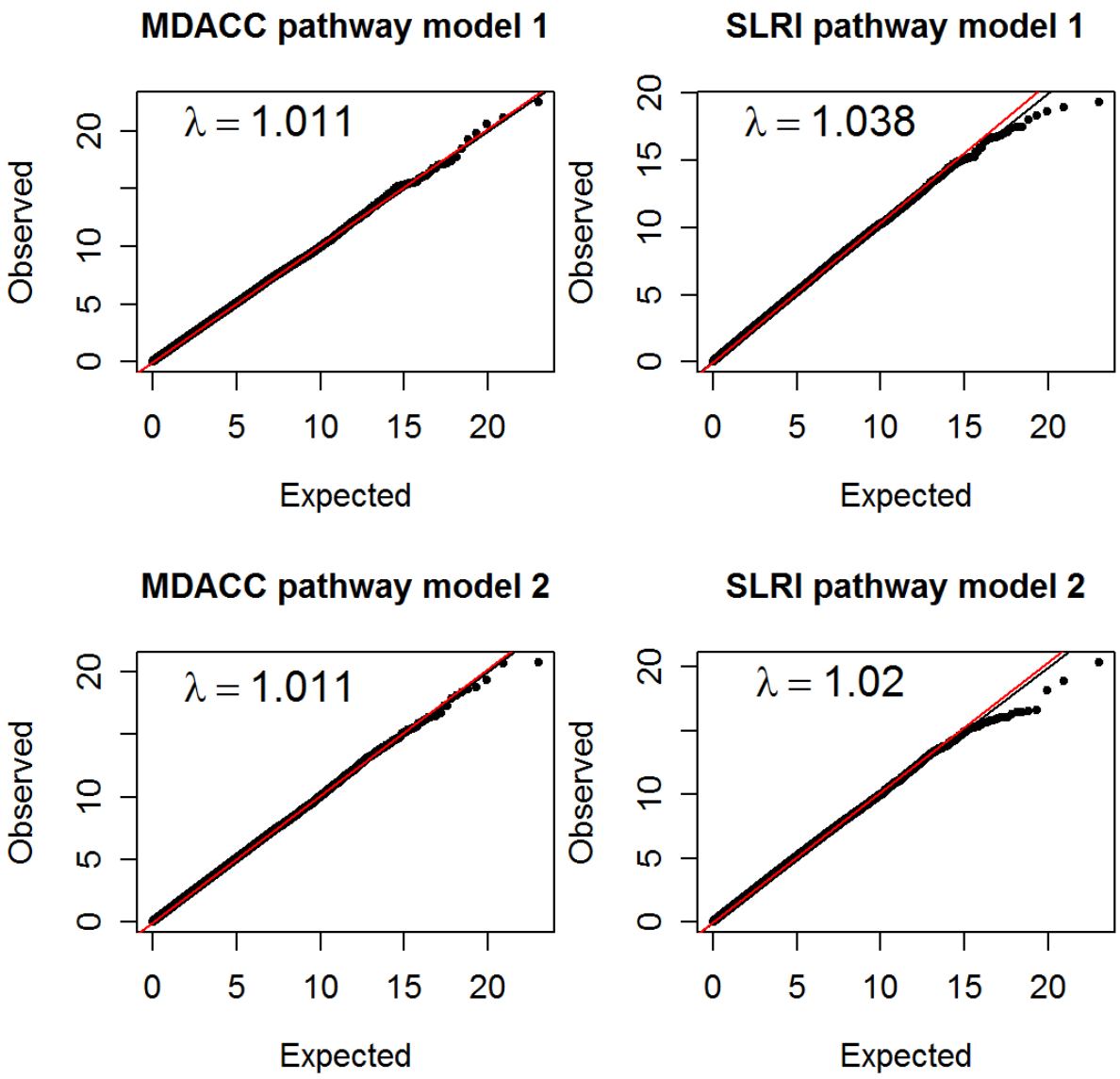


Figure B.1: *QQ plots of initial SNP main effect results from both pathway regression models for MDACC and SLRI*

Table B.2: ***Hierarchical Bayes prioritization****: Pathways within the top 10 for at least two different studies using* ***β*-regression coefficients** *as pathway ranking criterion. The β coefficients (equation 4.20) represent the increase or decrease of the prior probability of association for each SNP involved in the corresponding pathway. The numbers denote the corresponding ranks. Top 10 ranks are printed in bold. Pathways within the top 10 of all four studies are printed in bold.*

	Rank of *β* regression coefficient							
	GLC		**CE-IARC**		**MDACC**		**SLRI**	
	model 1	**model 2**	**model 1**	**model 2**	**model 1**	**model 2**	**model 1**	**model 2**
Alanine, aspartate and glutamate metabolism	**2**	48	54	**1**	45	52	71	**7**
Cysteine and methionine metabolism	**8**	40	45	**4**	38	43	54	**10**
Folate biosynthesis	27	**5**	**5**	25	**4**	**7**	**7**	19
Glycosylphosphatidylinositol(GPI)-anchor biosynthesis	15	**7**	12	17	**8**	**8**	**9**	16
Hedgehog signaling pathway	**3**	118	108	**2**	90	120	122	**1**
Nitrogen metabolism	**10**	22	24	**8**	23	26	29	**5**
Non-homologous end-joining	31	**4**	**6**	33	**7**	**2**	**2**	32
Other glycan degradation	11	**8**	**2**	12	**9**	11	**8**	23
Pantothenate and CoA biosynthesis	**7**	30	33	**7**	26	34	34	**3**
Pentose phosphate pathway	**4**	33	37	**3**	32	35	40	**2**
Riboflavin metabolism	9	**6**	**9**	**9**	**5**	**5**	**4**	**8**
RNA polymerase	**6**	28	32	**5**	25	31	32	**4**
Steroid biosynthesis	23	**3**	**4**	26	**2**	**3**	**3**	12
Sulfur relay system	25	**1**	**1**	24	**1**	**1**	**1**	14
Taurine and hypotaurine metabolism	5	**2**	**7**	**6**	**3**	**4**	**6**	**6**
Terpenoid backbone biosynthesis	48	**9**	11	45	**6**	**6**	**5**	40

Table B.3: ***Hierarchical Bayes prioritization****: Pathways within the top 10 for at least two different studies using* ***μ-regression coefficients*** *as pathway ranking criterion. The μ coefficients (equation 4.20) represent the increase or decrease of the prior strength of association for each SNP involved in the corresponding pathway. A pathway is considered as replicated, when it ranks among the top 10 for at least two different studies, independent of the particular pathway model. The numbers denote the corresponding ranks. Top 10 ranks are printed in bold. Pathways within the top 10 of all four studies are printed in bold.*

	Rank of μ regression coefficient							
	GLC		**CE-IARC**		**MDACC**		**SLRI**	
	model 1	**model 2**	**model 1**	**model 2**	**model 1**	**model 2**	**model 1**	**model 2**
alpha-Linolenic acid metabolism	117	**4**	**8**	118	**4**	11	**4**	120
Ascorbate and aldarate metabolism	13	**3**	**9**	16	**2**	13	**2**	84
beta-Alanine metabolism	96	**6**	10	90	**3**	12	**3**	98
Collecting duct acid secretion	14	**7**	13	**10**	**7**	15	**6**	**4**
Drug metabolism - cytochrome P450	**8**	38	40	**4**	30	39	26	**6**
ECM-receptor interaction	**1**	149	135	**1**	133	146	131	**1**
Galactose metabolism	47	**10**	14	42	**10**	18	**8**	19
Linoleic acid metabolism	115	**2**	**7**	115	**1**	**9**	**1**	116
Phenylalanine metabolism	17	**9**	15	13	**9**	17	**9**	14
Porphyrin and chlorophyll metabolism	**2**	54	60	**3**	54	67	58	**9**
Retinol metabolism	**4**	67	77	**5**	67	80	73	**7**
RNA polymerase	57	**5**	12	58	**5**	14	**5**	59
Starch and sucrose metabolism	**5**	83	69	**6**	73	85	69	**8**
Steroid hormone biosynthesis	**3**	95	86	**2**	82	95	77	**5**
Taurine and hypotaurine metabolism	23	**8**	11	22	**6**	16	**7**	23

Table B.4: ***GSEA****: Pathways with nominal p-value ≤ 0.05 for at least two different studies. The numbers denote the corresponding p-values. Significant pathways are printed in bold. Pathways significant for at least three different studies are printed in bold.*

	Nominal p-value							
	GLC		CE-IARC		MDACC		SLRI	
	model 1	model 2	model 1	model 2	model 1	model 2	model 1	model 2
Adherens junction	0.742	0.072	**0.032**	0.635	0.617	0.917	**0.015**	**0.05**
Allograft rejection	0.568	0.854	**0.047**	0.098	**0.044**	**0.04**	0.318	0.363
alpha-Linolenic acid metabolism	0.706	**0.041**	**0.011**	**0.018**	0.191	0.201	**0.044**	0.328
Cell adhesion molecules	0.846	0.984	**0.011**	0.238	0.163	0.109	0.064	**0.019**
Colorectal cancer	0.933	0.924	**0.006**	**0.028**	0.44	0.321	0.759	**0.048**
Cysteine and methionine metabolism	0.991	0.824	0.284	0.36	0.125	**0.05**	0.117	**0.026**
Neuroactive ligand-receptor interaction	0.038	**0.014**	0.216	0.164	0.391	0.406	**0.046**	0.117
Non-homologous end-joining	0.846	**0.005**	0.381	0.178	**0.006**	**0.018**	0.615	0.645
Tyrosine metabolism	0.87	0.829	**0.022**	0.325	**0.047**	**0.047**	0.892	0.438

Table B.5: ***SUMSTAT****: Pathways with nominal p-value ≤ 0.05 for at least two different studies. The numbers denote the corresponding p-values. Significant pathways are printed in bold. Pathways significant for at least three different studies are printed in bold.*

	Nominal p-value							
	GLC		CE-IARC		MDACC		SLRI	
	model 1	model 2	model 1	model 2	model 1	model 2	model 1	model 2
Alzheimers disease	0.009	0.345	0.006	0.022	0.677	0.612	0.69	0.878
Amino sugar and nucleotide sugar metabolism	0.011	0.015	0.021	0.008	0.161	0.161	0.475	0.749
Carbohydrate digestion and absorption	0.139	0.213	0.057	0.026	0.693	0.699	0.082	0.043
Cell adhesion molecules (CAMs)	0.482	0.707	0.002	0.048	0.136	0.108	0.049	0.046
Cholinergic synapse	0.437	0.934	0.01	0.018	0.003	0.003	0.038	0.215
Colorectal cancer	0.404	0.896	0.001	0.002	0.074	0.109	0.282	0.015
Cysteine and methionine metabolism	0.991	0.876	0.206	0.202	0.08	0.048	0.032	0.003
Fructose and mannose metabolism	0.244	0.04	0.092	0.014	0.736	0.656	0.645	0.768
Glyoxylate and dicarboxylate metabolism	0.596	0.934	0.015	0.041	0.157	0.21	0.311	0.023
Graft-versus-host disease	0.322	0.22	0.002	0.007	0.062	0.042	0.245	0.546
Herpes simplex infection	0.02	0.117	0.004	0.004	0.239	0.196	0.152	0.175
HTLV-I infection	0.247	0.739	0.001	0	0.047	0.073	0.101	0.033
Huntingtons disease	0	0.326	0.013	0.059	0.84	0.825	0.287	0.397
Influenza A	0.037	0.51	0.021	0.021	0.183	0.2	0.231	0.761
Long-term depression	0.737	0.661	0.047	0.022	0.53	0.51	0.125	0.035
Lysosome	0.039	0.282	0.025	0.03	0.059	0.089	0.568	0.378
mRNA surveillance pathway	0.011	0.196	0.086	0.431	0.948	0.935	0.028	0.009
Neuroactive ligand-receptor interaction	0.016	0.106	0.006	0.008	0.037	0.043	0.018	0.107
Non-homologous end-joining	0.516	0.205	0.253	0.021	0.049	0.062	0.274	0.343
Notch signaling pathway	0.549	0.759	0.028	0.093	0.03	0.038	0.354	0.125
Pancreatic secretion	0.429	0.305	0.006	0.009	0.052	0.052	0.006	0.012
Pathways in cancer	0.019	0.306	0.015	0.016	0.121	0.141	0.998	0.839
Pertussis	0.995	0.99	0.06	0.05	0.008	0.008	0.202	0.444
Primary bile acid biosynthesis	0.258	0.262	0.1	0.034	0.223	0.137	0.032	0.052
Pyrimidine metabolism	0.037	0.929	0.903	0.983	0.56	0.454	0.016	0.039
Wnt signaling pathway	0.114	0.845	0.001	0.003	0.627	0.725	0.305	0.048

(a) *CE-IARC*

(b) *MDACC*

Figure B.2: *Manhattan plots for GxE interaction effects of CE-IARC never vs. ever smokers, moderate vs. heavy smokers and MDACC moderate vs. heavy smokers. Upper row: Case-control test; Middle row: Case-only test; Lower row: Control only test*

Table B.6: ***Hierarchical Bayes prioritization based on GxE interaction effects (HBP-GxE)****: Pathways within the top 10 for at least two different studies using* ***β-regression coefficients*** *as pathway ranking criterion. The β coefficients (equation 4.20) represent the increase or decrease of the prior probability of association for each SNP involved in the corresponding pathway. The numbers denote the corresponding ranks. Top 10 ranks are printed in bold. Pathways within the top 10 of all four studies are printed in bold.*

	Rank of β regression coefficient						
	GLC		**CE-IARC**		**MDACC**	**SLRI**	
	MH	**NE**	**MH**	**NE**	**MH**	**MH**	**NE**
Folate biosynthesis	12	**4**	**4**	**5**	**5**	**5**	16
Glycosylphosphatidylinositol(GPI)-anchor biosynthesis	14	**1**	**7**	**8**	14	**7**	**6**
One carbon pool by folate	19	**9**	**9**	**15**	**7**	11	**5**
Primary bile acid biosynthesis	18	**7**	**10**	11	12	**9**	**7**
Riboflavin metabolism	15	**8**	**6**	**4**	16	**6**	**8**
Steroid biosynthesis	13	**5**	**3**	**3**	**8**	**3**	**4**
Sulfur metabolism	29	**10**	**8**	18	29	12	**9**
Sulfur relay system	**9**	**3**	**2**	**1**	**6**	**1**	13
Taurine and hypotaurine metabolism	11	**2**	**5**	**7**	**3**	**4**	**3**

Table B.7: ***Hierarchical Bayes prioritization based on GxE interaction effects (HBP-GxE)****: Pathways within the top 10 for at least two different studies using* ***μ-regression coefficients*** *as pathway ranking criterion. The μ coefficients (equation 4.20) represent the increase or decrease of the prior strength of association for each SNP involved in the corresponding pathway. A pathway is considered as replicated, when it ranks among the top 10 for at least two different studies, independent of the particular pathway model. The numbers denote the corresponding ranks. Top 10 ranks are printed in bold. Pathways within the top 10 of all four studies are printed in bold.*

	Rank of μ regression coefficient						
	GLC		**CE-IARC**		**MDACC**	**SLRI**	
	MH	**NE**	**MH**	**NE**	**MH**	**MH**	**NE**
alpha-Linolenic acid metabolism	**7**	96	10	**2**	131	**8**	112
Ascorbate and aldarate metabolism	**4**	69	**3**	**3**	42	**5**	12
Carbohydrate digestion and absorption	**1**	176	**1**	87	172	**2**	170
Collecting duct acid secretion	**5**	56	**5**	**5**	13	**6**	13
Folate biosynthesis	27	**6**	20	31	55	23	**6**
Inositol phosphate metabolism	**3**	155	14	91	161	**1**	161
Linoleic acid metabolism	**10**	98	**7**	**1**	128	**7**	110
Non-homologous end-joining	25	**2**	23	20	78	30	**5**
Riboflavin metabolism	14	**8**	15	14	61	15	**10**
RNA polymerase	**8**	31	**9**	**9**	44	**9**	41
Sulfur relay system	12	**1**	11	12	53	12	**3**
Taurine and hypotaurine metabolism	**6**	**5**	**4**	**4**	46	**10**	**8**
Terpenoid backbone biosynthesis	15	**4**	**2**	11	45	14	**4**

Table B.8: ***GSEA-GxE****: Pathways with nominal p-value ≤ 0.05 for at least two different studies. The numbers denote the corresponding p-values. Significant pathways are printed in bold. Pathways significant for at least three different studies are printed in bold.*

	Nominal p-value						
	GLC		CE-IARC		MDACC	SLRI	
	MH	NE	MH	NE	MH	MH	NE
Alanine, aspartate and glutamate metabolism	0.993	**0.029**	0.648	0.992	0.395	0.466	**0.03**
Butanoate metabolism	0.222	0.186	0.296	**0.006**	0.483	0.86	**0.011**
Cytokine-cytokine receptor interaction	0.749	0.861	**0.03**	0.763	0.84	**0.028**	0.693
Endocrine and other factor-regulated calcium reabsorption	0.499	**0.031**	0.1	**0.038**	0.684	0.463	0.205
ErbB signaling pathway	0.852	0.916	0.185	0.95	**0.005**	0.623	**0.001**
Gap junction	0.211	0.918	0.451	0.316	**0.029**	0.533	**0.017**
Gastric acid secretion	0.35	0.778	0.207	**0.012**	**0.01**	0.306	0.113
Leukocyte transendothelial migration	0.688	0.998	**0.036**	0.697	**0.021**	0.184	0.721
Nicotinate and nicotinamide metabolism	**0.019**	0.721	**0.047**	0.755	0.099	0.689	0.382
RNA transport	0.092	0.533	**0.006**	0.247	**0.021**	0.619	0.07
TGF-beta signaling pathway	0.631	0.534	0.446	**0.042**	0.249	0.603	**0.018**
Valine, leucine and isoleucine degradation	**0.05**	0.228	0.888	0.353	0.671	0.207	**0.008**

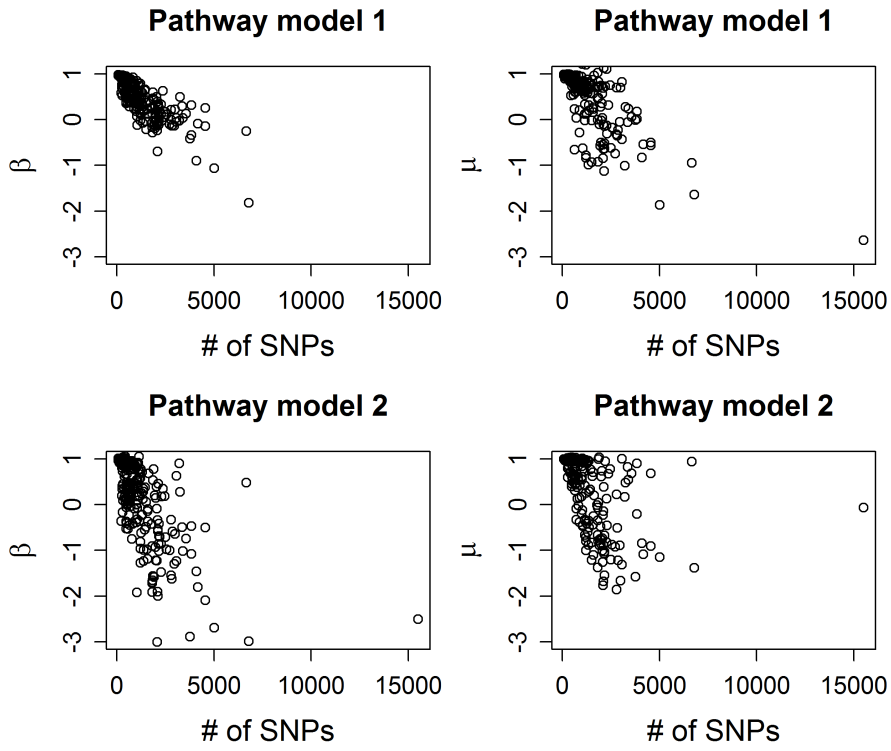


Figure B.3: *Correlation of HBP β and μ regression coefficients and number of SNPs per pathway for the CE-IARC study*

C Correction term for the posterior variance

In the following we have $\varphi(\cdot)$ probability density function and $\Phi(\cdot)$ probability distribution of the standard normal distribution and $\operatorname{erf}(x) = 2\Phi(x\sqrt{2}) - 1$.
Furthermore let

$$\begin{aligned} D_+ &= \frac{(x+\theta)}{\sqrt{\sigma_x^2+\sigma^2}} \\ D_- &= \frac{(x-\theta)}{\sqrt{\sigma_x^2+\sigma^2}}. \end{aligned}$$

C.1 Jacobian of the posterior expectation

Posterior expectation (equation 4.25):

$$E = \mathrm{E}\left[\lambda \mid x, \theta, b, \sigma\right]$$

Jacobian:

$$\tilde{\nabla} = \begin{pmatrix} \tilde{\delta}\theta \\ \tilde{\delta}\sigma \\ \tilde{\delta}b \end{pmatrix}$$

We can split the posterior expectation:

$$E = E_+ P_+ = \left(\frac{f_1}{f_2}\frac{E_{num}}{E_{den}}\right) P_+ = \left(\frac{f_1}{f_2}\frac{E_{num1} + E_{num2} + E_{num3}}{E_{den}}\right) P_+$$

with

$$\begin{aligned} f_1 &= \sigma_x \sigma \\ f_2 &= \frac{1}{\sqrt{\sigma^2+\sigma_x^2}} \\ E_{num1} &= \frac{2}{\pi} \exp - \frac{\sigma^2 x^2 + \sigma_x^2 \theta^2}{2\sigma_x^2 \sigma^2} \\ E_{num2} &= L_+ \varphi(D_-) \operatorname{erf}(\frac{L_+}{\sqrt{2}}) \\ E_{num3} &= L_- \varphi(D_+) \operatorname{erf}(\frac{L_-}{\sqrt{2}}) \\ E_{den} &= \varphi(D_+) + \varphi(D_-) \\ P_+ &= \left(1 + \frac{1}{\exp(b)} 2\varphi\left(\frac{x}{\sigma_x}\right) / \sigma_x \frac{\sqrt{\sigma_x^2+\sigma^2}}{\varphi(D_+)+\varphi(D_-)}\right)^{-1} \end{aligned}$$

and

$$\begin{aligned} L_+ &= \frac{\sigma_x^2\theta + \sigma^2 x}{\sqrt{\sigma_x^2}\sigma\sqrt{\sigma_x^2+\sigma^2}} \\ L_- &= \frac{\sigma_x^2\theta - \sigma^2 x}{\sqrt{\sigma_x^2}\sigma\sqrt{\sigma_x^2+\sigma^2}}. \end{aligned}$$

We will further use

$$P_{inv} = 1/P_+.$$

Derivative with respect to θ

$$\frac{\partial E}{\partial \theta} = E\left(\frac{\partial E_{num}}{\partial \theta}/E_{num} - \frac{\partial E_{den}}{\partial \theta}/E_{den} - \frac{\partial P_{inv}}{\partial \theta}/P_{inv}\right)$$

with

$$\begin{aligned} \frac{\partial E_{num}}{\partial \theta} &= \frac{\partial E_{num1}}{\partial \theta} + \frac{\partial E_{num2}}{\partial \theta} + \frac{\partial E_{num3}}{\partial \theta} \\ \frac{\partial E_{num1}}{\partial \theta} &= -\frac{\theta}{\sigma^2} E_{num1} \\ \frac{\partial E_{num2}}{\partial \theta} &= L_+\varphi(D_-)\frac{\partial \operatorname{erf}(L_+)}{\partial \theta} + L_+\frac{\partial \varphi D_-}{\partial \theta}\operatorname{erf}(\frac{L_+}{\sqrt{2}}) + \frac{\partial L_+}{\partial \theta}\varphi(D_-)\operatorname{erf}(\frac{L_+}{\sqrt{2}}) \\ \frac{\partial E_{num3}}{\partial \theta} &= L_-\varphi(D_+)\frac{\partial \operatorname{erf}(L_-)}{\partial \theta} + L_-\frac{\partial \varphi D_+}{\partial \theta}\operatorname{erf}(\frac{L_-}{\sqrt{2}}) + \frac{\partial L_-}{\partial \theta}\varphi(D_+)\operatorname{erf}(\frac{L_-}{\sqrt{2}}) \\ \frac{\partial E_{den}}{\partial \theta} &= \frac{\partial \varphi(D_+)}{\partial \theta} + \frac{\partial \varphi(D_-)}{\partial \theta} \\ \frac{\partial P_{inv}}{\partial \theta} &= -(P_{inv}-1)\frac{\partial E_{den}}{\partial \theta}/E_{den} \end{aligned}$$

and

$$\begin{aligned} \frac{\partial L_+}{\partial \theta} &= \frac{\sqrt{\sigma_x^2}}{\sigma\sqrt{\sigma_x^2+\sigma^2}} \\ \frac{\partial L_-}{\partial \theta} &= \frac{\sqrt{\sigma_x^2}}{\sigma\sqrt{\sigma_x^2+\sigma^2}} \\ \frac{\partial D_+}{\partial \theta} &= \frac{1}{\sqrt{\sigma_x^2+\sigma^2}} \\ \frac{\partial D_-}{\partial \theta} &= -\frac{1}{\sqrt{\sigma_x^2+\sigma^2}} \\ \frac{\partial \varphi(D_+)}{\partial \theta} &= -\varphi(D_+)D_+\frac{\partial D_+}{\partial \theta} \\ \frac{\partial \varphi(D_-)}{\partial \theta} &= -\varphi(D_-)D_-\frac{\partial D_-}{\partial \theta} \\ \frac{\partial \operatorname{erf}(L_+)}{\partial \theta} &= 2\varphi L_+\frac{\partial L_+}{\partial \theta} \\ \frac{\partial \operatorname{erf}(L_-)}{\partial \theta} &= 2\varphi L_-\frac{\partial L_-}{\partial \theta} \end{aligned}$$

Derivative with respect to σ

$$\frac{\partial E}{\partial \sigma} = E\left(-\frac{\partial P_{inv}}{\partial \sigma}/P_{inv} + \frac{\partial E_{num}}{\partial \sigma}/E_{num} - \frac{\partial E_{den}}{\partial \sigma}/E_{den} + \frac{\partial f_1}{\partial \sigma}/f_1 + \frac{\partial f_2}{\partial \sigma}/f_2\right)$$

with

$$\begin{aligned}
\frac{\partial E_{num}}{\partial \sigma} &= \frac{\partial E_{num1}}{\partial \sigma} + \frac{\partial E_{num2}}{\partial \sigma} + \frac{\partial E_{num3}}{\partial \sigma} \\
\frac{\partial E_{num1}}{\partial \sigma} &= E_{num1}\left(\frac{-x^2}{\sigma_x^2\sigma} + \frac{\sigma^2x^2 + \sigma_x^2\theta^2}{\sigma_x^2\sigma^3}\right) \\
\frac{\partial E_{num2}}{\partial \sigma} &= L_+\varphi(D_-)\frac{\partial \operatorname{erf}(L_+)}{\partial \sigma} + L_+\frac{\partial D_-}{\partial \sigma}\operatorname{erf}(\frac{L_+}{\sqrt{(2)}}) + \frac{\partial L_+}{\partial \sigma}\varphi(D_-)\operatorname{erf}(\frac{L_+}{\sqrt{(2)}}) \\
\frac{\partial E_{num3}}{\partial \sigma} &= L_-\varphi(D_+)\frac{\partial \operatorname{erf}(L_-)}{\partial \sigma} + L_-\frac{\partial D_+}{\partial \sigma}\operatorname{erf}(\frac{L_-}{\sqrt{(2)}}) + \frac{\partial L_-}{\partial \sigma}\varphi(D_+)\operatorname{erf}(\frac{L_-}{\sqrt{(2)}}) \\
\frac{\partial E_{den}}{\partial \sigma} &= \frac{\partial D_+}{\partial \sigma} + \frac{\partial D_-}{\partial \sigma} \\
\frac{\partial P_{inv}}{\partial \sigma} &= (Pinv - 1)\frac{\sigma}{\sigma_x^2 + \sigma^2} - (Pinv - 1)\frac{\partial E_{den}}{\partial \sigma}/E_{den}
\end{aligned}$$

and

$$\begin{aligned}
\frac{\partial f_1}{\partial \sigma} &= \sqrt{\sigma_x^2} \\
\frac{\partial f_2}{\partial \sigma} &= -\frac{\sigma}{(\sigma^2 + \sigma_x^2)^{3/2}} \\
\frac{\partial L_+}{\partial \sigma} &= \sigma_x\frac{\sigma^2x - \sigma_x^2\theta - 2\sigma^2\theta}{(\sigma_x^2 + \sigma^2)^{3/2}\sigma^2} \\
\frac{\partial L_-}{\partial \sigma} &= -\sigma_x\frac{\sigma^2x + \sigma_x^2\theta + 2\sigma^2\theta}{(\sigma_x^2 + \sigma^2)^{3/2}\sigma^2} \\
\frac{\partial D_+}{\partial \sigma} &= D_+\frac{\partial f_2}{\partial \sigma}/f_2 \\
\frac{\partial D_-}{\partial \sigma} &= D_-\frac{\partial f_2}{\partial \sigma}/f_2 \\
\frac{\partial \varphi D_+}{\partial \sigma} &= -\varphi(D_+)D_+\frac{\partial D_+}{\partial \sigma} \\
\frac{\partial \varphi D_-}{\partial \sigma} &= -\varphi(D_-)D_-\frac{\partial D_-}{\partial \sigma} \\
\frac{\partial \operatorname{erf}(L_+)}{\partial \sigma} &= 2\varphi(L_+)\frac{\partial L_+}{\partial \sigma} \\
\frac{\partial \operatorname{erf}(L_-)}{\partial \sigma} &= 2\varphi(L_-)\frac{\partial L_-}{\partial \sigma}
\end{aligned}$$

Derivative with respect to $\mathbf{b} = \operatorname{logit}(p/1-p)$

$$\frac{\partial E}{\partial b} = -E\frac{\partial P_{inv}}{\partial b}/P_{inv}$$

with

$$\frac{\partial P_{inv}}{\partial b} = \frac{-1}{\exp(b)} 2\varphi\left(\frac{x}{\sqrt{\sigma_x^2}}\right) / \sqrt{\sigma_x^2}\sqrt{\sigma_x^2 + \sigma^2}\frac{f_2}{E_{den}}$$

C.2 Hessian of the marginal log likelihood

Marginal likelihood (equation 4.21):

$$L = m(x \mid \theta, \sigma, b)$$

Hessian:

$$\tilde{\Sigma} = \begin{pmatrix} \tilde{\tau}_{\theta\theta} & \tilde{\tau}_{\theta\sigma} & \tilde{\tau}_{\theta b} \\ \tilde{\tau}_{\sigma\theta} & \tilde{\tau}_{\sigma\sigma} & \tilde{\tau}_{\sigma b} \\ \tilde{\tau}_{b\theta} & \tilde{\tau}_{b\sigma} & \tilde{\tau}_{bb} \end{pmatrix},$$

We can split the likelihood

$$L = p(U + V) + (1 - p)2W = p\frac{(U_n + V_n)}{\sqrt{\sigma_x^2 + \sigma^2}} + (1 - p)2W$$

with

$$\begin{aligned} p &= \frac{\exp(b)}{1 + \exp(b)} \\ U_n &= \varphi(D_+) \\ V_n &= \varphi(D_-) \\ W &= \varphi\left(\frac{x}{\sigma_x}\right)/\sigma_x \end{aligned}$$

Mixed second order partial derivatives of log(L):

$$\begin{aligned} \frac{\partial^2 logL}{\partial\theta\partial\theta} &= \left[L\frac{\partial^2 L}{\partial\theta\partial\theta} - \left(\frac{\partial L}{\partial\theta}\right)^2\right]/L^2 \\ \frac{\partial^2 logL}{\partial\theta\partial b} &= \left[L\frac{\partial^2 L}{\partial\theta\partial b} - \left(\frac{\partial L}{\partial\theta}\frac{\partial L}{\partial b}\right)\right]/L^2 \\ \frac{\partial^2 logL}{\partial\theta\partial\sigma} &= \left[L\frac{\partial^2 L}{\partial\theta\partial\sigma} - \left(\frac{\partial L}{\partial\theta}\frac{\partial L}{\partial\sigma}\right)\right]/L^2 \\ \frac{\partial^2 logL}{\partial b\partial\theta} &= \left[L\frac{\partial^2 L}{\partial b\partial\theta} - \left(\frac{\partial L}{\partial b}\frac{\partial L}{\partial\theta}\right)\right]/L^2 \\ \frac{\partial^2 logL}{\partial b\partial b} &= \left[L\frac{\partial^2 L}{\partial b\partial b} - \left(\frac{\partial L}{\partial b}\right)^2\right]/L^2 \\ \frac{\partial^2 logL}{\partial b\partial\sigma} &= \left[L\frac{\partial^2 L}{\partial b\partial\sigma} - \left(\frac{\partial L}{\partial b}\frac{\partial L}{\partial\sigma}\right)\right]/L^2 \\ \frac{\partial^2 logL}{\partial\sigma\partial\theta} &= \left[L\frac{\partial^2 L}{\partial\sigma\partial\theta} - \left(\frac{\partial L}{\partial\sigma}\frac{\partial L}{\partial\theta}\right)\right]/L^2 \\ \frac{\partial^2 logL}{\partial\sigma\partial b} &= \left[L\frac{\partial^2 L}{\partial\sigma\partial b} - \left(\frac{\partial L}{\partial\sigma}\frac{\partial L}{\partial b}\right)\right]/L^2 \\ \frac{\partial^2 logL}{\partial\sigma\partial\sigma} &= \left[L\frac{\partial^2 L}{\partial\sigma\partial\sigma} - \left(\frac{\partial L}{\partial\sigma}\right)^2\right]/L^2 \end{aligned}$$

First partial derivative with respect to θ:

$$\frac{\partial L}{\partial \theta} = p\left(\frac{\partial U_n}{\partial \theta} + \frac{\partial V_n}{\partial \theta}\right) / \sqrt{\sigma^2 + \sigma_x^2}$$

with

$$\begin{aligned}
\frac{\partial D_+}{\partial \theta} &= \frac{1}{\sqrt{\sigma_x^2 + \sigma^2}} \\
\frac{\partial D_-}{\partial \theta} &= -\frac{1}{\sqrt{\sigma_x^2 + \sigma^2}} \\
\frac{\partial U_n}{\partial \theta} &= -\varphi(D_+) D_+ \frac{\partial D_+}{\partial \theta} \\
\frac{\partial V_n}{\partial \theta} &= -\varphi(D_-) D_- \frac{\partial D_-}{\partial \theta}
\end{aligned}$$

Second order partial derivative with respect to θ

$$\frac{\partial^2 L}{\partial \theta \partial \theta} = p\left(\frac{\partial^2 U_n}{\partial \theta \partial \theta} + \frac{\partial^2 V_n}{\partial \theta \partial \theta}\right) / \sqrt{\sigma_x^2 + \sigma^2}$$

with

$$\begin{aligned}
\frac{\partial^2 D_+}{\partial \theta \partial \theta} &= 0 \\
\frac{\partial^2 D_-}{\partial \theta \partial \theta} &= 0 \\
\frac{\partial^2 U_n}{\partial \theta \partial \theta} &= -\left(\frac{\partial U_n}{\partial \theta} D_+ \frac{\partial D_+}{\partial \theta} + \varphi(D_+) \frac{\partial D_+}{\partial \theta} \frac{\partial D_+}{\partial \theta} + \varphi(D_+) D_+ \frac{\partial^2 D_+}{\partial \theta \partial \theta}\right) \\
\frac{\partial^2 V_n}{\partial \theta \partial \theta} &= -\left(\frac{\partial V_n}{\partial \theta} D_- \frac{\partial D_-}{\partial \theta} + \varphi(D_-) \frac{\partial D_-}{\partial \theta} \frac{\partial D_-}{\partial \theta} + \varphi(D_-) D_- \frac{\partial^2 D_-}{\partial \theta \partial \theta}\right)
\end{aligned}$$

Mixed second order partial derivative with respect to θ and σ:

$$\frac{\partial L}{\partial \theta \partial \sigma} = \frac{\partial L}{\partial \theta}\left(\frac{\partial U_n}{\partial \theta \partial \sigma} + \frac{\partial V_n}{\partial \theta \partial \sigma}\right) / \left(\frac{\partial U_n}{\partial \theta} + \frac{\partial V_n}{\partial \theta}\right) + \frac{\partial L}{\partial \theta}\left(-\frac{\sigma}{\sigma_x^2 + \sigma^2}\right)$$

with

$$\begin{aligned}
\frac{\partial D_+}{\partial \theta \partial \sigma} &= \frac{\partial D_+}{\partial \theta} \frac{-\sigma}{\sigma_x^2 + \sigma^2} \\
\frac{\partial D_-}{\partial \theta \partial \sigma} &= \frac{\partial D_-}{\partial \theta} \frac{-\sigma}{\sigma_x^2 + \sigma^2} \\
\frac{\partial U_n}{\partial \theta \partial \sigma} &= -\left(\frac{\partial U_n}{\partial \sigma} D_+ \frac{\partial D_+}{\partial \theta} + \varphi(D_+) \frac{\partial D_+}{\partial \sigma} \frac{\partial D_+}{\partial \theta} + \varphi(D_+) D_+ \frac{\partial D_+}{\partial \theta \partial \sigma}\right) \\
\frac{\partial V_n}{\partial \sigma} &= -\left(\frac{\partial V_n}{\partial \sigma} D_- \frac{\partial D_-}{\partial \theta} + \varphi(D_-) \frac{\partial D_-}{\partial \sigma} \frac{\partial D_-}{\partial \theta} + \varphi(D_-) D_- \frac{\partial D_-}{\partial \theta \partial \sigma}\right)
\end{aligned}$$

Mixed second order partial derivative with respect to θ and b:

$$\frac{\partial^2 L}{\partial \theta \partial b} = \frac{\partial L}{\partial \theta} \frac{\partial p}{\partial b} / p$$

First partial derivative with respect to σ:

$$\frac{\partial L}{\partial \sigma} = p \left[\left(\frac{\partial V_n}{\partial \sigma} + \frac{\partial V_n}{\partial \sigma} \right) - (U + V) \frac{\partial sd}{\partial \sigma} \right] / \sqrt{\sigma_x^2 + \sigma^2}$$

with

$$\begin{aligned}
\frac{\partial sd}{\partial \sigma} &= \frac{\sigma}{\sqrt{\sigma_x^2 + \sigma^2}} \\
\frac{\partial D_+}{\partial \sigma} &= -D_+ \frac{\partial sd}{\partial \sigma} / \sqrt{\sigma_x^2 + \sigma^2} \\
\frac{\partial D_-}{\partial \sigma} &= -D_- \frac{\partial sd}{\partial \sigma} / \sqrt{\sigma_x^2 + \sigma^2} \\
\frac{\partial U_n}{\partial \sigma} &= -\varphi(D_+) D_+ \frac{\partial D_+}{\partial \sigma} \\
\frac{\partial V_n}{\partial \sigma} &= -\varphi(D_-) D_- \frac{\partial D_-}{\partial \sigma}
\end{aligned}$$

Mixed second order partial derivative with respect to σ and θ:

$$\frac{\partial^2 L}{\partial \sigma \partial \theta} = \frac{\exp(b)}{1 + \exp(b)} \left[\left(\frac{\partial^2 U_n}{\partial \sigma \partial \theta} + \frac{\partial^2 V_n}{\partial \sigma \partial \theta} \right) - \left(\frac{\partial U_n}{\partial \theta} + \frac{\partial V_n}{\partial \theta} \right) \frac{\partial sd}{\partial \sigma} / \sqrt{\sigma_x^2 + \sigma^2} \right] / \sqrt{\sigma_x^2 + \sigma^2}$$

with

$$\begin{aligned}
\frac{\partial^2 D_+}{\partial \sigma \partial \theta} &= \frac{\partial D_+}{\partial \sigma} \frac{\partial D+}{\partial \theta} / D_+ \\
\frac{\partial^2 D_-}{\partial \sigma \partial \theta} &= \frac{\partial D_-}{\partial \sigma} \frac{\partial D-}{\partial \theta} / D_- \\
\frac{\partial^2 U_n}{\partial \sigma \partial \theta} &= -\left(\frac{\partial U_n}{\partial \theta} \frac{\partial D_+}{\partial \sigma} D_+ + \varphi(D_+) \frac{\partial D_+}{\partial \theta} \frac{\partial D_+}{\partial \sigma} + \varphi(D+) D + \frac{\partial^2 D_+}{\partial \sigma \partial \theta} \right) \\
\frac{\partial^2 V_n}{\partial \sigma \partial \theta} &= -\left(\frac{\partial V_n}{\partial \theta} \frac{\partial D_-}{\partial \sigma} D_- + \varphi(D_-) \frac{\partial D_-}{\partial \theta} \frac{\partial D_-}{\partial \sigma} + \varphi(D-) D - \frac{\partial^2 D_-}{\partial \sigma \partial \theta} \right)
\end{aligned}$$

Mixed second order partial derivative with respect to σ:

$$\begin{aligned}
\frac{\partial^2 L}{\partial \sigma \partial \sigma} = \frac{\exp(b)}{1 + \exp(b)} \Bigg[& \left(\frac{\partial U_n}{\partial \sigma} + \frac{\partial V_n}{\partial \sigma} \right) / \sqrt{\sigma_x^2 + \sigma^2} \\
& - \left(\frac{\partial U_n}{\partial \sigma} + \frac{\partial V_n}{\partial \sigma} \right) / \sqrt{\sigma_x^2 + \sigma^2} \frac{\partial sd}{\partial \sigma} / \sqrt{\sigma_x^2 + \sigma^2} \\
& - (U_n + V_n) \left(\frac{-3\sigma^2}{(\sigma_x^2 + \sigma^2)^{5/2}} + \frac{1}{(\sigma_x^2 + \sigma^2)^{3/2}} \right) \\
& - \left(\frac{\partial U_n}{\partial \sigma} + \frac{\partial V_n}{\partial \sigma} \right) \frac{\partial sd}{\partial \sigma} / (\sigma_x^2 + \sigma^2) \Bigg]
\end{aligned}$$

with

$$
\begin{aligned}
\frac{\partial^2 sd}{\partial\sigma\partial\sigma} &= \frac{\sigma_x^2}{(\sqrt{\sigma^2+\sigma_x^2})^3} \\
\frac{\partial^2 D_+}{\partial\sigma\partial\sigma} &= \frac{(x+\theta)(2\sigma^2-\sigma_x^2)}{(\sigma_x^2+\sigma^2)^{5/2}} \\
\frac{\partial^2 D_-}{\partial\sigma\partial\sigma} &= \frac{(x-\theta)(2\sigma^2-\sigma_x^2)}{(\sigma_x^2+\sigma^2)^{5/2}} \\
\frac{\partial^2 U_n}{\partial\sigma\partial\sigma} &= \frac{\partial U_n}{\partial\sigma} D_+^2 \frac{\partial sd}{\partial\sigma} \frac{1}{\sqrt{\sigma_x^2+\sigma^2}} \\
&\quad + \varphi(D_+) 2 \frac{\partial D_+}{\partial\sigma} D_+ \frac{\partial sd}{\partial\sigma} \frac{1}{\sqrt{\sigma_x^2+\sigma^2}} \\
&\quad + \varphi(D_+) D_+ D_+ \left(\frac{-2\sigma^2}{(\sqrt{\sigma^2+\sigma_x^2})^4} + \frac{1}{(\sqrt{\sigma_x^2+\sigma^2})^2} \right) \\
\frac{\partial^2 V_n}{\partial\sigma\partial\sigma} &= \frac{\partial V_n}{\partial\sigma} D_-^2 \frac{\partial sd}{\partial\sigma} \frac{1}{\sqrt{\sigma_x^2+\sigma^2}} \\
&\quad + \varphi(D_-) 2 \frac{\partial D_-}{\partial\sigma} D_- \frac{\partial sd}{\partial\sigma} \frac{1}{/\sqrt{\sigma_x^2+\sigma^2}} \\
&\quad + \varphi(D_-) D_- D_- \left(\frac{-2\sigma^2}{(\sqrt{\sigma^2+\sigma_x^2})^4} + \frac{1}{(\sqrt{\sigma_x^2+\sigma^2})^2} \right)
\end{aligned}
$$

Mixed second order partial derivative with respect to σ and b:

$$
\frac{\partial^2 L}{\partial\sigma\partial b} = \frac{\partial L}{\partial\sigma} \frac{1}{1+\exp(b)}
$$

First partial derivative with respect to b:

$$
\begin{aligned}
\frac{\partial L}{\partial b} &= \frac{\partial p}{\partial b}(U+V) - \frac{\partial p}{\partial b} 2W \\
\frac{\partial p}{\partial b} &= \frac{\exp(b)}{1+\exp(b)} \frac{1}{1+\exp(b)}
\end{aligned}
$$

Mixed second order partial derivative with respect to b and θ:

$$
\frac{\partial^2 L}{\partial b\partial\theta} = \frac{\partial p}{\partial b} \left(\frac{\partial U_n}{\partial\theta} + \frac{\partial V_n}{\partial\theta} \right) (\sqrt{\sigma_x^2+\sigma^2})
$$

Mixed second order partial derivative with respect to b and σ:

$$
\frac{\partial^2 L}{\partial b\partial\sigma} = \frac{\partial^2 p}{\partial b} \left(\frac{\partial U_n}{\partial\theta} + \frac{\partial V_n}{\partial\theta} \right) / \sqrt{\sigma_x^2+\sigma^2} - \frac{U+V}{\sqrt{\sigma_x^2+\sigma^2}} \frac{\partial sd}{\partial\sigma}
$$

Second order partial derivative with respect to b:

$$\frac{\partial^2 L}{\partial b \partial b} = \frac{\partial^2 p}{\partial b \partial b}(U + V) - \frac{\partial^2 p}{\partial b \partial b} 2W$$
$$\frac{\partial^2 p}{\partial b \partial b} = -\frac{\partial p}{\partial b} \frac{\exp(b) - 1}{1 + \exp(b)}$$

References

1000 Genomes Project Consortium (2010). A map of human genome variation from population-scale sequencing. *Nature*, 467(7319):1061–1073.

Affymetrix (2006). BRLMM: an improved genotype calling method for the GeneChip Human Mapping 500K Array Set. Technical report.

Affymetrix (2009). Genome-wide human SNP array 6.0 data sheet. Available at http://media.affymetrix.com/support/technical/datasheets/genomewide_snp6_datasheet.pdf [Accessed 5 March 2012].

Agresti, A. (2002). *Categorical data analysis*. Wiley series in probability and statistics. Wiley-Interscience.

Aidoo, M., Terlouw, D. J., Kolczak, M. S., McElroy, P. D., ter Kuile, F. O., Kariuki, S., Nahlen, B. L., Lal, A. A., and Udhayakumar, V. (2002). Protective effects of the sickle cell gene against malaria morbidity and mortality. *Lancet*, 359(9314):1311–1312.

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

Alberts, B. (2004). *Essential cell biology*. Number Bd. 1 in Essential Cell Biology. Garland Science Pub.

Allis, C., Jenuwein, T., and Reinberg, D. (2007). *Epigenetics*. Cold Spring Harbor Laboratory Press.

Allison, D. B., Cui, X., Page, G. P., and Sabripour, M. (2006). Microarray data analysis: from disarray to consolidation and consensus. *Nat Rev Genet*, 7(1):55–65.

Altmüller, J., Palmer, L. J., Fischer, G., Scherb, H., and Wjst, M. (2001). Genomewide scans of complex human diseases: true linkage is hard to find. *Am J Hum Genet*, 69(5):936–950.

Amos, C. I. (2007). Transdisciplinary research in cancer of the lung. Website http://u19tricl.org.

Amos, C. I., Caporaso, N. E., and Weston, A. (1992). Host factors in lung cancer risk: a review of interdisciplinary studies. *Cancer Epidem Biomar*, 1(6):505–513.

Amos, C. I., Chen, W. V., Seldin, M. F., Remmers, E. F., Taylor, K. E., Criswell, L. A., Lee, A. T., Plenge, R. M., Kastner, D. L., and Gregersen, P. K. (2009). Data for genetic analysis workshop 16 problem 1, association analysis of rheumatoid arthritis data. *BMC Proc*, 3 Suppl 7:S2.

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

Antoniou, A., Pharoah, P. D. P., Narod, S., Risch, H. A., Eyfjord, J. E., Hopper, J. L., Loman, N., Olsson, H., Johannsson, O., Borg, A., Pasini, B., Radice, P., Manoukian, S., Eccles, D. M., Tang, N., Olah, E., Anton-Culver, H., Warner, E., Lubinski, J., Gronwald, J., Gorski, B., Tulinius, H., Thorlacius, S., Eerola, H., Nevanlinna, H., Syrjäkoski, K., Kallioniemi, O.-P., Thompson, D., Evans, C., Peto, J., Lalloo, F., Evans, D. G., and Easton, D. F. (2003). Average risks of breast and ovarian cancer associated with BRCA1 or BRCA2 mutations detected in case series unselected for family history: a combined analysis of 22 studies. *Am J Hum Genet*, 72(5):1117–1130.

Antosh, M., Fox, D., Helfand, S. L., Cooper, L. N., and Neretti, N. (2011). New comparative genomics approach reveals a conserved health span signature across species. *Aging (Albany NY)*, 3(6):576–583.

Aoki, K. (1993). Excess incidence of lung cancer among pulmonary tuberculosis patients. *Jpn J Clin Oncol*, 23(4):205–220.

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

Arking, D. E., Pfeufer, A., Post, W., Kao, W. H. L., Newton-Cheh, C., Ikeda, M., West, K., Kashuk, C., Akyol, M., Perz, S., Jalilzadeh, S., Illig, T., Gieger, C., Guo, C.-Y., Larson, M. G., Wichmann, H. E., Marbán, E., O'Donnell, C. J., Hirschhorn, J. N., Kääb, S., Spooner, P. M., Meitinger, T., and Chakravarti, A. (2006). A common genetic variant in the NOS1 regulator NOS1AP modulates cardiac repolarization. *Nat Genet*, 38(6):644–651.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., and et al. (2000). Gene ontology: tool for the unification of biology. the gene ontology consortium. *Nat Genet*, 25(1):25–29.

Asimit, J. and Zeggini, E. (2010). Rare variant association analysis methods for complex traits. *Annu Rev Genet*, 44:293–308.

Aulchenko, Y. S. (2010). Effects of population structure in genome-wide association studies. In Zeggini, E. and Morris, A., editors, *Analysis of Complex Disease Association Studies: A Practical Guide*, volume 60 of *Academic Press*, pages 123–156. Elsevier Science.

Aulchenko, Y. S., Ripke, S., Isaacs, A., and Van Duijn, C. M. (2007). GenABEL: an R library for genome-wide association analysis. *Bioinformatics*, 23(10):1294–1296.

Bader, G. D., Cary, M. P., and Sander, C. (2006). Pathguide: a pathway resource list. *Nucleic Acids Res*, 34(Database issue):D504–D506.

Bailey-Wilson, J. E., Amos, C. I., Pinney, S. M., Petersen, G. M., de Andrade, M., Wiest, J. S., Fain, P., Schwartz, A. G., You, M., Franklin, W., Klein, C., Gazdar, A., Rothschild, H., Mandal, D., Coons, T., Slusser, J., Lee, J., Gaba, C., Kupert, E., Perez, A., Zhou, X., Zeng, D., Liu, Q., Zhang, Q., Seminara, D., Minna, J., and Anderson, M. W. (2004). A major lung cancer susceptibility locus maps to chromosome 6q23-25. *Am J Hum Genet*, 75(3):460–474.

Ballard, D. H., Aporntewan, C., Lee, J. Y., Lee, J. S., Wu, Z., and Zhao, H. (2009). A pathway analysis applied to genetic analysis workshop 16 genome-wide rheumatoid arthritis data. *BMC Proc*, 3 Suppl 7:S91.

Ballard, D. H., Cho, J., and Zhao, H. (2010). Comparisons of multi-marker association methods to detect association between a candidate region and disease. *Genet Epidemiol*, 34(3):201–212.

Barabási, A.-L. and Oltvai, Z. N. (2004). Network biology: understanding the cell's functional organization. *Nat Rev Genet*, 5(2):101–113.

Barrett, J. C. (2010). Population genetics and linkage disequilibrium. In Zeggini, E. and Morris, A., editors, *Analysis of Complex Disease Association Studies: A Practical Guide*, volume 60 of *Academic Press*, pages 15–23. Elsevier Science.

Barton, A., Eyre, S., Ke, X., Hinks, A., Bowes, J., Flynn, E., Martin, P., Consortium, Y. E. A. R., Consortium, B. I. R. A. C., Wilson, A. G., Morgan, A. W., Emery, P., Steer, S., Hocking, L. J., Reid, D. M., Harrison, P., Wordsworth, P., Thomson, W., and Worthington, J. (2009). Identification of AF4/FMR2 family, member 3 (AFF3) as a novel rheumatoid arthritis susceptibility locus and confirmation of two further pan-autoimmune susceptibility genes. *Hum Mol Genet*, 18(13):2518–2522.

Barton, A., Thomson, W., Ke, X., Eyre, S., Hinks, A., Bowes, J., Plant, D., Gibbons, L. J., Consortium, W. T. C. C., Consortium, Y. E. A. R., Consortium, B. I. R. A. C., Wilson, A. G., Bax, D. E., Morgan, A. W., Emery, P., Steer, S., Hocking, L., Reid, D. M., Wordsworth, P., Harrison, P., and Worthington, J. (2008). Rheumatoid arthritis susceptibility loci at chromosomes 10p15, 12q13 and 22q13. *Nat Genet*, 40(10):1156–1159.

Bassett, J. K., Hodge, A. M., English, D. R., Baglietto, L., Hopper, J. L., Giles, G. G., and Severi, G. (2012). Dietary intake of B vitamins and methionine and risk of lung cancer. *Eur J Clin Nutr*, 66(2):182–187.

Basu, S. and Pan, W. (2011). Comparison of statistical tests for disease association with rare variants. *Genet Epidemiol*, 35(7):606–619.

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

Begovich, A. B., Carlton, V. E. H., Honigberg, L. A., Schrodi, S. J., Chokkalingam, A. P., Alexander, H. C., Ardlie, K. G., Huang, Q., Smith, A. M., Spoerke, J. M., and et al. (2004). A missense single-nucleotide polymorphism in a gene encoding a protein tyrosine phosphatase (PTPN22) is associated with rheumatoid arthritis. *Am J Hum Genet*, 75(2):330–7.

Benjamini, Y., Drai, D., Elmer, G., Kafkafi, N., and Golani, I. (2001). Controlling the false discovery rate in behavior genetics research. *Behav Brain Res*, 125(1-2):279–284.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J R Stat Soc Series B Stat Methodol*, 57(1):289–300.

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

Bickeböller, H. and Fischer, C. (2007). *Einführung in die Genetische Epidemiologie*. Statistik und ihre Anwendungen. Springer.

BioCarta LLC (2011). BioCarta - Charting pathways of life. Website http://www.biocarta.com/genes/index.asp [Accessed 1 August 2011].

Bonferroni, C. (1935). *Il calcolo delle assicurazioni su gruppi di teste*. Tipografia del Senato.

Bonferroni, C. E. (1936). Teoria statistica delle classi e calcolo delle probabilità. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze*, 8:3–62.

Borecki, I. B. and Province, M. A. (2008). Linkage and association: Basic concepts. In Rao, D. C. and Gu, C. C., editors, *Genetic Dissection of Complex Traits*, volume 60 of *Advances in Genetics*, pages 51 – 74. Academic Press.

Botstein, D. and Risch, N. (2003). Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. *Nat Genet*, 33 Suppl:228–237.

Botstein, D., White, R. L., Skolnick, M., and Davis, R. W. (1980). Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *Am J Hum Genet*, 32(3):314–331.

Brennan, P. (2002). Gene-environment interaction and aetiology of cancer: what does it mean and how can we measure it? *Carcinogenesis*, 23(3):381–387.

Breslow, N. and Day, N. (1994). *Statistical Methods in Cancer Research: The Design and Analysis of Cohort Studies*. IARC Scientific Publications. International Agency for Research on Cancer.

Breuer, R. H. J., Postmus, P. E., and Smit, E. F. (2005). Molecular pathology of non-small-cell lung cancer. *Respiration*, 72(3):313–330.

Brown, T. (2002). *Genomes*. Wiley-Liss.

Browning, B. L. and Browning, S. R. (2009). A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *Am J Hum Genet*, 84(2):210–223.

Bukszár, J. and van den Oord, E. J. C. G. (2006). Optimization of two-stage genetic designs where data are combined using an accurate and efficient approximation for pearson's statistic. *Biometrics*, 62(4):1132–1137.

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

Bush, W. S., Chen, G., Torstenson, E. S., and Ritchie, M. D. (2009). LD-spline: mapping SNPs on genotyping platforms to genomic regions using patterns of linkage disequilibrium. *BioData Min*, 2(1):7.

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

Cardon, L. R. and Bell, J. I. (2001). Association study designs for complex diseases. *Nat Rev Genet*, 2(2):91–99.

Cardon, L. R. and Palmer, L. J. (2003). Population stratification and spurious allelic association. *Lancet*, 361(9357):598–604.

Carlborg, O. and Haley, C. S. (2004). Epistasis: too often neglected in complex trait studies? *Nat Rev Genet*, 5(8):618–625.

Carlin, B. P. and Louis, T. A. (2000). *Bayes and Empirical Bayes Methods for Data Analysis*. Chapman & Hall, London.

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

Chai, H.-S., Sicotte, H., Bailey, K. R., Turner, S. T., Asmann, Y. W., and Kocher, J.-P. A. (2009). Glossi: a method to assess the association of genetic loci-sets with complex diseases. *BMC Bioinformatics*, 10:102.

Chasman, D. I. (2008). On the utility of gene set methods in genomewide association studies of quantitative traits. *Genet Epidemiol*, 32(7):658–668.

Chaudhuri, P. K., Thomas, P. A., Walker, M. J., Briele, H. A., Gupta, T. K. D., and Beattie, C. W. (1982). Steroid receptors in human lung cancer cytosols. *Cancer Lett*, 16(3):327–332.

Chen, G. K. and Witte, J. S. (2007). Enriching the analysis of genomewide association studies with hierarchical modeling. *Am J Hum Genet*, 81(2):397–404.

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

Chen, X., Wang, L., Hu, B., Guo, M., Barnard, J., and Zhu, X. (2010). Pathway-based analysis for genome-wide association studies using supervised principal components. *Genet Epidemiol*, 34(7):716–724.

Chinnici, J. P. (1999). Using the biochemical pathway model to teach the concepts of gene interaction & epistasis. *Am Biol Teach*, 61(3):pp. 207–213.

Christensen, K. and Murray, J. C. (2007). What genome-wide association studies can do for medicine. *N Engl J Med*, 356(11):1094–1097.

Churchill, G. A. and Doerge, R. W. (1994). Empirical threshold values for quantitative trait mapping. *Genetics*, 138(3):963–971.

Clark, R. F., Hutton, M., Talbot, C., Wragg, M., Lendon, C., Busfield, F., Han, S. W., Perez-Tur, J., Adams, M., Fuldner, R., Roberts, G., Karran, E., Hardy, J., and Goate, A. (1996). The role of presenilin 1 in the genetics of Alzheimer's disease. *Cold Spring Harb Sym*, 61:551–558.

Clarke, G. M., Anderson, C. A., Pettersson, F. H., Cardon, L. R., Morris, A. P., and Zondervan, K. T. (2011). Basic statistical analysis in genetic case-control studies. *Nat Protoc*, 6(2):121–133.

Clayton, D. and Leung, H. (2007). An R package for analysis of whole-genome association studies. *Hum Hered*, 64(1):45–51.

Colhoun, H. M., McKeigue, P. M., and Smith, G. D. (2003). Problems of reporting genetic associations with complex outcomes. *Lancet*, 361(9360):865–872.

Collins, F. S., Guyer, M. S., and Charkravarti, A. (1997). Variations on a theme: cataloging human DNA sequence variation. *Science*, 278(5343):1580–1581.

Cordell, H. J. (2002). Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. *Hum Mol Genet*, 11(20):2463–2468.

Croitoru, M. E., Cleary, S. P., Nicola, N. D., Manno, M., Selander, T., Aronson, M., Redston, M., Cotterchio, M., Knight, J., Gryfe, R., and Gallinger, S. (2004). Association between biallelic and monoallelic germline MYH gene mutations and colorectal cancer risk. *J Natl Cancer Inst*, 96(21):1631–1634.

Cupples, L. A., Heard-Costa, N., Lee, M., Atwood, L. D., and Investigators, F. H. S. (2009). Genetics analysis workshop 16 problem 2: the framingham heart study data. *BMC Proc*, 3 Suppl 7:S3.

Curtis, R. K., Oresic, M., and Vidal-Puig, A. (2005). Pathways to the analysis of microarray data. *Trends Biotechnol*, 23(8):429–435.

Database of Single Nucleotide Polymorphisms (dbSNP) (2009). Bethesda (MD): National Center for Biotechnology Information, National Library of Medicine. (dbSNP Build ID: 136). Available at `http://www.ncbi.nlm.nih.gov/SNP/` [Accessed 29 February 2012].

Dawid, A. P. (1979). Conditional Independence in Statistical Theory. *J Roy Stat Soc B Met*, 41(1):1–31.

Dawson, E., Abecasis, G. R., Bumpstead, S., Chen, Y., Hunt, S., Beare, D. M., Pabial, J., Dibling, T., Tinsley, E., Kirby, S., Carter, D., Papaspyridonos, M., Livingstone, S., Ganske, R., Lõhmussaar, E., Zernant, J., Tõnisson, N., Remm, M., Mägi, R., Puurand, T., Vilo, J., Kurg, A., Rice, K., Deloukas, P., Mott, R., Metspalu, A., Bentley, D. R., Cardon, L. R., and Dunham, I. (2002). A first-generation linkage disequilibrium map of human chromosome 22. *Nature*, 418(6897):544–548.

De Bakker, P. I. W., Yelensky, R., Pe'er, I., Gabriel, S. B., Daly, M. J., and Altshuler, D. (2005). Efficiency and power in genetic association studies. *Nat Genet*, 37(11):1217–1223.

De la Cruz, O., Wen, X., Ke, B., Song, M., and Nicolae, D. L. (2010). Gene, region and pathway level analyses in whole-genome studies. *Genet Epidemiol*, 34(3):222–231.

Dehling, H. and Haupt, B. (2004). *Einführung in die Wahrscheinlichkeitstheorie und Statistik*. Statistik und ihre Anwendungen. Springer.

Dering, C., Ziegler, A., Konig, I., and Hemmelmann, C. (2011). Comparison of collapsing methods for the statistical analysis of rare variants. *BMC Proc*, 5(Suppl 9):S115.

Devlin, B. and Roeder, K. (1999). Genomic control for association studies. *Biometrics*, 55(4):997–1004.

Devlin, B., Roeder, K., and Wasserman, L. (2001). Genomic control, a new approach to genetic-based association studies. *Theor Popul Biol*, 60(3):155–166.

Dinu, I., Potter, J. D., Mueller, T., Liu, Q., Adewale, A. J., Jhangri, G. S., Einecke, G., Famulski, K. S., Halloran, P., and Yasui, Y. (2007). Improving gene set analysis of microarray data by SAM-GS. *BMC Bioinformatics*, 8:242.

Donoho, D. and Jin, J. (2004). Higher criticism for detecting sparse heterogeneous mixtures. *Ann Stat*, 32(3):962–994.

Dresler, C. M., León, M. E., Straif, K., Baan, R., and Secretan, B. (2006). Reversal of risk upon quitting smoking. *Lancet*, 368(9533):348–349.

Dudbridge, F. and Koeleman, B. P. C. (2003). Rank truncated product of p-values, with application to genomewide association scans. *Genet Epidemiol*, 25(4):360–366.

Dudoit, S., Gilbert, H. N., and van der Laan, M. J. (2008). Resampling-based empirical Bayes multiple testing procedures for controlling generalized tail probability and expected value error rates: focus on the false discovery rate and simulation study. *Biom J*, 50(5):716–744.

Dudoit, S. and Laan, M. (2008). *Multiple testing procedures with applications to genomics*. Springer series in statistics. Springer.

Dudoit, S., Shaffer, J. P., and Boldrick, J. C. (2003). Multiple hypothesis testing in microarray experiments. *Stat Sci*, 18(1):71–103.

Eberle, M. A., Ng, P. C., Kuhn, K., Zhou, L., Peiffer, D. A., Galver, L., Viaud-Martinez, K. A., Lawley, C. T., Gunderson, K. L., Shen, R., and Murray, S. S. (2007). Power to detect risk alleles using genome-wide tag SNP panels. *PLoS Genet*, 3(10):1827–1837.

Efron, B. and Tibshirani, R. (2002). Empirical Bayes methods and false discovery rates for microarrays. *Genet Epidemiol*, 23(1):70–86.

Efron, B. and Tibshirani, R. (2006). On testing the significance of sets of genes. Technical report, Annals of Applied Statistics.

Elbers, C. C., van Eijk, K. R., Franke, L., Mulder, F., van der Schouw, Y. T., Wijmenga, C., and Onland-Moret, N. C. (2009). Using genome-wide pathway analysis to unravel the etiology of complex diseases. *Genet Epidemiol*, 33(5):419–431.

Engelman, C. D., Baurley, J. W., Chiu, Y.-F., Joubert, B. R., Lewinger, J. P., Maenner, M. J., Murcray, C. E., Shi, G., and Gauderman, W. J. (2009). Detecting gene-environment interactions in genome-wide association data. *Genet Epidemiol*, 33 Suppl 1:S68–S73.

Erdogan, E., Klee, E. W., Thompson, E. A., and Fields, A. P. (2009). Meta-analysis of oncogenic protein kinase Ciota signaling in lung adenocarcinoma. *Clin Cancer Res*, 15(5):1527–1533.

Etzel, C. J., Amos, C. I., and Spitz, M. R. (2003). Risk for smoking-related cancer among relatives of lung cancer patients. *Cancer Res*, 63(23):8531–8535.

Evans, D. M. and Cardon, L. R. (2006). Genome-wide association: a promising start to a long race. *Trends Genet*, 22(7):350–354.

Evans, M., Hastings, N., and Peacock, B. (2000). *Statistical Distributions*. Wiley-Interscience.

Falush, D., Stephens, M., and Pritchard, J. K. (2003). Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics*, 164(4):1567–1587.

Faraway, J. (2006). *Extending the linear model with R: generalized linear, mixed effects and nonparametric regression models*. Texts in statistical science. Chapman & Hall/CRC.

Farrer, L. A. and Cupples, L. A. (1998). Determiniing the genetic componenet of a disease. In Haines, J. L. and Pericak-Vance, M. A., editors, *Approaches to Gene Mapping in complex Human Diseases*, pages 223–233. Wiley: New York.

Fehringer, G., Liu, G., Briollais, L., Brennan, P., Amos, C. I., Spitz, M. R., Bickeböller, H., Wichmann, H. E., Risch, A., and Hung, R. J. (2012). Comparison of pathway analysis approaches using lung cancer GWAS data sets. *PLoS One*, 7(2):e31816.

Fennessy, F. M., Moneley, D. S., Wang, J. H., Kelly, C. J., and Bouchier-Hayes, D. J. (2003). Taurine and vitamin C modify monocyte and endothelial dysfunction in young smokers. *Circulation*, 107(3):410–415.

Flicek, P., Amode, M. R., Barrell, D., Beal, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fairley, S., Fitzgerald, S., Gil, L., Gordon, L., Hendrix, M., Hourlier, T., Johnson, N., Kähäri, A. K., Keefe, D., Keenan, S., Kinsella, R., Komorowska, M., Koscielny, G., Kulesha, E., Larsson, P., Longden, I., McLaren, W., Muffato, M., Overduin, B., Pignatelli, M., Pritchard, B., Riat, H. S., Ritchie, G. R. S., Ruffier, M., Schuster, M., Sobral, D., Tang, Y. A., Taylor, K., Trevanion, S., Vandrovcova, J., White, S., Wilson, M., Wilder, S. P., Aken, B. L., Birney, E., Cunningham, F., Dunham, I., Durbin, R., Fernández-Suarez, X. M., Harrow, J., Herrero, J., Hubbard,

T. J. P., Parker, A., Proctor, G., Spudich, G., Vogel, J., Yates, A., Zadissa, A., and Searle, S. M. J. (2012). Ensembl 2012. *Nucleic Acids Res*, 40(Database issue):D84–D90.

Forner, K., Lamarine, M., Guedj, M., Dauvillier, J., and Wojcik, J. (2008). Universal false discovery rate estimation methodology for genome-wide association studies. *Hum Hered*, 65(4):183–194.

Frazer, K. A., Murray, S. S., Schork, N. J., and Topol, E. J. (2009). Human genetic variation and its contribution to complex traits. *Nat Rev Genet*, 10(4):241–251.

Freedman, M. L., Reich, D., Penney, K. L., McDonald, G. J., Mignault, A. A., Patterson, N., Gabriel, S. B., Topol, E. J., Smoller, J. W., Pato, C. N., Pato, M. T., Petryshen, T. L., Kolonel, L. N., Lander, E. S., Sklar, P., Henderson, B., Hirschhorn, J. N., and Altshuler, D. (2004). Assessing the impact of population stratification on genetic association studies. *Nat Genet*, 36(4):388–393.

Freidlin, B., Zheng, G., Li, Z., and Gastwirth, J. L. (2002). Trend tests for case-control studies of genetic markers: power, sample size and robustness. *Hum Hered*, 53(3):146–152.

Fridley, B. L., Jenkins, G. D., and Biernacka, J. M. (2010). Self-contained gene-set analysis of expression data: an evaluation of existing and novel methods. *PLoS One*, 5(9).

Gail, M. H., Pfeiffer, R. M., Wheeler, W., and Pee, D. (2008). Probability of detecting disease-associated single nucleotide polymorphisms in case-control genome-wide association studies. *Biostatistics*, 9(2):201–215.

Gao, Y. T., Blot, W. J., Zheng, W., Ershow, A. G., Hsu, C. W., Levin, L. I., Zhang, R., and Fraumeni, J. F. (1987). Lung cancer among chinese women. *Int J Cancer*, 40(5):604–609.

Gauderman, W. J., Murcray, C., Gilliland, F., and Conti, D. V. (2007). Testing association between disease and multiple SNPs in a candidate gene. *Genet Epidemiol*, 31(5):383–395.

Ge, Y., Dudoit, S., and Speed, T. P. (2003). Resampling-based multiple testing for microarray data analysis. *Test*, 12(1):1–77.

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

Gibson, G. (2010). Hints of hidden heritability in GWAS. *Nat Genet*, 42(7):558–560.

Gillespie, J. (1998). *Population genetics: a concise guide*. A Johns Hopkins paperback. The Johns Hopkins University Press.

Goeman, J. J. and Buehlmann, P. (2007). Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics*, 23(8):980–987.

Greene, C. S., Penrod, N. M., Williams, S. M., and Moore, J. H. (2009). Failure to replicate a genetic association may provide important clues about genetic architecture. *PLoS One*, 4(6):e5639.

Greenland, S. (1993). Methods for epidemiologic analyses of multiple exposures: a review and comparative study of maximum-likelihood, preliminary-testing, and empirical-Bayes regression. *Stat Med*, 12(8):717–736.

Gregersen, P. K., Amos, C. I., Lee, A. T., Lu, Y., Remmers, E. F., Kastner, D. L., Seldin, M. F., Criswell, L. A., Plenge, R. M., Holers, V. M., and et al. (2009). REL, encoding a member of the NF-kappaB family of transcription factors, is a newly defined risk locus for rheumatoid arthritis. *Nat Genet*, 41(7):820–823.

Griffiths, A., Wessler, S., Lewontin, R., and Carroll, S. (2008). *Introduction to Genetic Analysis*. W. H. Freeman.

Grimm, D., Blum, H. E., and Thimme, R. (2011). Genomweite Assoziationsstudien. *Deutsche medizinische Wochenschrift 1946*, 713(3):407–409.

Hall, J. M., Lee, M. K., Newman, B., Morrow, J. E., Anderson, L. A., Huey, B., and King, M. C. (1990). Linkage of early-onset familial breast cancer to chromosome 17q21. *Science*, 250(4988):1684–1689.

Harney, S. M. J., Vilariño-Güell, C., Adamopoulos, I. E., Sims, A.-M., Lawrence, R. W., Cardon, L. R., Newton, J. L., Meisel, C., Pointon, J. J., Darke, C., Athanasou, N., Wordsworth, B. P., and Brown, M. A. (2008). Fine mapping of the MHC Class III region demonstrates association of AIF1 and rheumatoid arthritis. *Rheumatology (Oxford)*, 47(12):1761–1767.

Harris, M. A., Clark, J., Ireland, A., Lomax, J., Ashburner, M., Foulger, R., Eilbeck, K., Lewis, S., Marshall, B., Mungall, C., Richter, J., Rubin, G. M., Blake, J. A., Bult, C., Dolan, M., Drabkin, H., Eppig, J. T., Hill, D. P., Ni, L., Ringwald, M., Balakrishnan, R., Cherry, J. M., Christie, K. R., Costanzo, M. C., Dwight, S. S., Engel, S., Fisk, D. G., Hirschman, J. E., Hong, E. L., Nash, R. S., Sethuraman, A., Theesfeld, C. L., Botstein, D., Dolinski, K., Feierbach, B., Berardini, T., Mundodi, S., Rhee, S. Y., Apweiler, R., Barrell, D., Camon, E., Dimmer, E., Lee, V., Chisholm, R., Gaudet, P., Kibbe, W., Kishore, R., Schwarz, E. M., Sternberg, P., Gwinn, M., Hannick, L., Wortman, J., Berriman, M., Wood, V., de la Cruz, N., Tonellato, P., Jaiswal, P., Seigfried, T., White, R., and Consortium, G. O. (2004). The gene ontology (GO) database and informatics resource. *Nucleic Acids Res*, 32(Database issue):D258–D261.

Herold, C., Steffens, M., Brockschmidt, F. F., Baur, M. P., and Becker, T. (2009). INTERSNP: genome-wide interaction analysis guided by a priori information. *Bioinformatics*, 25(24):3275–3281.

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

Hindorff, L. A., MacArthur, J., Wise, A., Junkins, H. A., Hall, P., Klemm, A. K., and Manolio, T. (2012). A catalog of published genome-wide association studies. Available at www.genome.gov/gwastudies/ [Accessed 26 April 2012].

Hindorff, L. A., Sethupathy, P., Junkins, H. A., Ramos, E. M., Mehta, J. P., Collins, F. S., and Manolio, T. A. (2009). Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci U S A*, 106(23):9362–9367.

Hirschhorn, J. N. (2005). Genetic approaches to studying common diseases and complex traits. *Pediatr Res*, 57(5 Pt 2):74R–77R.

Hirschhorn, J. N., Lohmueller, K., Byrne, E., and Hirschhorn, K. (2002). A comprehensive review of genetic association studies. *Genet Med*, 4(2):45–61.

Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. *Biometrika*, pages 800–803.

Hochberg, Y. and Benjamini, Y. (1990). More powerful procedures for multiple significance testing. *Stat Med*, 9(7):811–818.

Hoh, J., Wille, A., and Ott, J. (2001). Trimming, weighting, and grouping SNPs in human case-control association studies. *Genome Res*, 11(12):2115–2119.

Holden, M., Deng, S., Wojnowski, L., and Kulle, B. (2008). GSEA-SNP: applying gene set enrichment analysis to SNP data from genome-wide association studies. *Bioinformatics*, 24(23):2784–2785.

Holle, R., Happich, M., Löwel, H., and Wichmann, H. E. (2005). KORA- a research platform for population based health research. *Gesundheitswesen Bundesverband der Arzte des öffentlichen Gesundheitsdienstes Germany*, 67 Suppl 1:S19–S25.

Holm, S. (1979a). A simple sequentially rejective Bonferroni test procedure. *Scand J Statist*, 6:65–70.

Holm, S. (1979b). A simple sequential rejective multiple test procedure. *Scand J Statist*, 6:65–70.

Hommel, G. (1988). A stagewise rejective multiple test procedure based on a modified Bonferroni test. *Biometrika*, 75(2):383–386.

Hong, M.-G., Pawitan, Y., Magnusson, P. K. E., and Prince, J. A. (2009). Strategies and issues in the detection of pathway enrichment in genome-wide association studies. *Hum Genet*, 126(2):289–301.

Hosack, D. A., Dennis, G., Sherman, B. T., Lane, H. C., and Lempicki, R. A. (2003). Identifying biological themes within lists of genes with ease. *Genome Biol*, 4(10):R70.

Huang, Q. M., Tomida, S., Masuda, Y., Arima, C., Cao, K., Kasahara, T.-A., Osada, H., Yatabe, Y., Akashi, T., Kamiya, K., Takahashi, T., and Suzuki, M. (2010). Regulation of DNA polymerase POLD4 influences genomic instability in lung cancer. *Cancer Res*, 70(21):8407–8416.

Human Genome Project (2003). Human Genome Project (HGP) Physical Map. *Genomics*, pages 1–10.

Hung, R. J., Brennan, P., Malaveille, C., Porru, S., Donato, F., Boffetta, P., and Witte, J. S. (2004). Using hierarchical modeling in genetic association studies with multiple markers: application to a case-control study of bladder cancer. *Cancer Epidem Biomar*, 13(6):1013–1021.

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

Hung, R. J., McKay, J. D., Gaborieau, V., Boffetta, P., Hashibe, M., Zaridze, D., Mukeria, A., Szeszenia-Dabrowska, N., Lissowska, J., Rudnai, P., and et al. (2008b). 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.

Hunter, D. J. and Kraft, P. (2007). Drinking from the fire hose–statistical issues in genomewide association studies. *N Engl J Med*, 357(5):436–439.

Ingelsson, E. (2010). Large-scale genome-wide association studies consortia: blessing, burden, or necessity? *Circ Cardiovasc Genet*, 3(5):396–398.

International Agency for Research on Cancer (IARC) (2012). International Lung Cancer Consortium (ILCCO). Website `ilcco.iarc.fr`.

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, Frazer, K. A., Ballinger, D. G., Cox, D. R., Hinds, D. A., Stuve, L. L., Gibbs, R. A., Belmont, J. W., Boudreau, A., Hardenbol, P., Leal, S. M., Pasternak, S., Wheeler, D. A., Willis, T. D., Yu, F., Yang, H., Zeng, C., Gao, Y., Hu, H., Hu, W., Li, C., Lin, W., Liu, S., Pan, H., Tang, X., Wang, J., Wang, W., Yu, J., Zhang, B., Zhang, Q., Zhao, H., Zhao, H., Zhou, J., Gabriel, S. B., Barry, R., Blumenstiel, B., Camargo, A., Defelice, M., Faggart, M., Goyette, M., Gupta, S., Moore, J., Nguyen, H., Onofrio, R. C., Parkin, M., Roy, J., Stahl, E.,

Winchester, E., Ziaugra, L., Altshuler, D., Shen, Y., Yao, Z., Huang, W., Chu, X., He, Y., Jin, L., Liu, Y., Shen, Y., Sun, W., Wang, H., Wang, Y., Wang, Y., Xiong, X., Xu, L., Waye, M. M. Y., Tsui, S. K. W., Xue, H., Wong, J. T.-F., Galver, L. M., Fan, J.-B., Gunderson, K., Murray, S. S., Oliphant, A. R., Chee, M. S., Montpetit, A., Chagnon, F., Ferretti, V., Leboeuf, M., Olivier, J.-F., Phillips, M. S., Roumy, S., Sallee, C., Verner, A., Hudson, T. J., Kwok, P.-Y., Cai, D., Koboldt, D. C., Miller, R. D., Pawlikowska, L., Taillon-Miller, P., Xiao, M., Tsui, L.-C., Mak, W., Song, Y. Q., Tam, P. K. H., Nakamura, Y., Kawaguchi, T., Kitamoto, T., Morizono, T., Nagashima, A., Ohnishi, Y., Sekine, A., Tanaka, T., Tsunoda, T., Deloukas, P., Bird, C. P., Delgado, M., Dermitzakis, E. T., Gwilliam, R., Hunt, S., Morrison, J., Powell, D., Stranger, B. E., Whittaker, P., Bentley, D. R., Daly, M. J., de Bakker, P. I. W., Barrett, J., Chretien, Y. R., Maller, J., McCarroll, S., Patterson, N., Pe'er, I., Price, A., Purcell, S., Richter, D. J., Sabeti, P., Saxena, R., Schaffner, S. F., Sham, P. C., Varilly, P., Altshuler, D., Stein, L. D., Krishnan, L., Smith, A. V., Tello-Ruiz, M. K., Thorisson, G. A., Chakravarti, A., Chen, P. E., Cutler, D. J., Kashuk, C. S., Lin, S., Abecasis, G. R., Guan, W., Li, Y., Munro, H. M., Qin, Z. S., Thomas, D. J., McVean, G., Auton, A., Bottolo, L., Cardin, N., Eyheramendy, S., Freeman, C., Marchini, J., Myers, S., Spencer, C., Stephens, M., Donnelly, P., Cardon, L. R., Clarke, G., Evans, D. M., Morris, A. P., Weir, B. S., Tsunoda, T., Mullikin, J. C., Sherry, S. T., Feolo, M., Skol, A., Zhang, H., Zeng, C., Zhao, H., Matsuda, I., Fukushima, Y., Macer, D. R., Suda, E., Rotimi, C. N., Adebamowo, C. A., Ajayi, I., Aniagwu, T., Marshall, P. A., Nkwodimmah, C., Royal, C. D. M., Leppert, M. F., Dixon, M., Peiffer, A., Qiu, R., Kent, A., Kato, K., Niikawa, N., Adewole, I. F., Knoppers, B. M., Foster, M. W., Clayton, E. W., Watkin, J., Gibbs, R. A., Belmont, J. W., Muzny, D., Nazareth, L., Sodergren, E., Weinstock, G. M., Wheeler, D. A., Yakub, I., Gabriel, S. B., Onofrio, R. C., Richter, D. J., Ziaugra, L., Birren, B. W., Daly, M. J., Altshuler, D., Wilson, R. K., Fulton, L. L., Rogers, J., Burton, J., Carter, N. P., Clee, C. M., Griffiths, M., Jones, M. C., McLay, K., Plumb, R. W., Ross, M. T., Sims, S. K., Willey, D. L., Chen, Z., Han, H., Kang, L., Godbout, M., Wallenburg, J. C., L'Archeveque, P., Bellemare, G., Saeki, K., Wang, H., An, D., Fu, H., Li, Q., Wang, Z., Wang, R., Holden, A. L., Brooks, L. D., McEwen, J. E., Guyer, M. S., Wang, V. O., Peterson, J. L., Shi, M., Spiegel, J., Sung, L. M., Zacharia, L. F., Collins, F. S., Kennedy, K., Jamieson, R., and Stewart, J. (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.

Ioannidis, J. P. A. (2007). Non-replication and inconsistency in the genome-wide association setting. *Hum Hered*, 64(4):203–213.

Ioannidis, J. P. A., Patsopoulos, N. A., and Evangelou, E. (2007). Heterogeneity in meta-analyses of genome-wide association investigations. *PLoS One*, 2(9):e841.

Jakulin, A. and Bratko, I. (2004). Quantifying and visualizing attribute interactions: An approach based on entropy. *http://arxiv.org/abs/cs.AI/0308002 v3*, 308002:3.

Janssens, A. C. and van Duijn, C. (2010). An epidemiological perspective on the future of direct-to-consumer personal genome testing. *Investig Genet*, 1(1):10.

Janssens, A. C. J. W. and van Duijn, C. M. (2008). Genome-based prediction of common diseases: advances and prospects. *Hum Mol Genet*, 17(R2):R166–R173.

Jawaheer, D., Li, W., Graham, R. R., Chen, W., Damle, A., Xiao, X., Monteiro, J., Khalili, H., Lee, A., Lundsten, R., Begovich, A., Bugawan, T., Erlich, H., Elder, J. T., Criswell, L. A., Seldin, M. F., Amos, C. I., Behrens, T. W., and Gregersen, P. K. (2002). Dissecting the genetic complexity of the association between human leukocyte antigens and rheumatoid arthritis. *Am J Hum Genet*, 71(3):585–594.

Jia, P., Wang, L., Meltzer, H. Y., and Zhao, Z. (2010). Common variants conferring risk of schizophrenia: a pathway analysis of GWAS data. *Schizophr Res*, 122(1-3):38–42.

Johansson, M., Relton, C., Ueland, P. M., Vollset, S. E., Midttun, Ø., Nygård, O., Slimani, N., Boffetta, P., Jenab, M., Clavel-Chapelon, F., Boutron-Ruault, M.-C., Fagherazzi, G., Kaaks, R., Rohrmann, S., Boeing, H., Weikert, C., Bueno-de Mesquita, H. B., Ros, M. M., van Gils, C. H., Peeters, P. H. M., Agudo, A., Barricarte, A., Navarro, C., Rodríguez, L., Sanchez, M.-J., Larranaga, N., Khaw, K.-T., Wareham, N., Allen, N. E., Crowe, F., Gallo, V., Norat, T., Krogh, V., Masala, G., Panico, S., Sacerdote, C., Tumino, R., Trichopoulou, A., Lagiou, P., Trichopoulos, D., Rasmuson, T., Hallmans, G., Riboli, E., Vineis, P., and Brennan, P. (2010). Serum B vitamin levels and risk of lung cancer. *JAMA*, 303(23):2377–2385.

Johnson, A. D. and O'Donnell, C. J. (2009). An open access database of genome-wide association results. *BMC Med Genet*, 10:6.

Jorde, L. B., Watkins, W. S., Carlson, M., Groden, J., Albertsen, H., Thliveris, A., and Leppert, M. (1994). Linkage disequilibrium predicts physical distance in the adenomatous polyposis coli region. *Am J Hum Genet*, 54(5):884–898.

Juran, B. D. and Lazaridis, K. N. (2011). Genomics in the post-GWAS era. *Semin Liver Dis*, 31(2):215–222.

Kanehisa, M. and Goto, S. (2000). KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res*, 28(1):27–30.

Kanehisa, M., Goto, S., Sato, Y., Furumichi, M., and Tanabe, M. (2012). KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res*, 40(Database issue):D109–D114.

Karp, G. (2002). *Cell and molecular biology: concepts and experiments*. J. Wiley.

Kass, R. E. and Steffey, D. (1989). Approximate Bayesian inference in conditionally independent hierarchical models (parametric empirical Bayes models). *JAMA*, 84(407):717–726.

Kendler, K. S. and Baker, J. H. (2007). Genetic influences on measures of the environment: a systematic review. *Psychol Med*, 37(5):615–626.

Kent, W., Sugnet, C., Furey, T., Roskin, K., Pringle, T., Zahler, A., and Haussler, D. (2002). The human genome browser at UCSC. *Genome Res*, 12:996–1006.

Khoury, M. (2010). *Human genome epidemiology: building the evidence for using genetic information to improve health and prevent disease*. Oxford University Press.

Khoury, M. J. and Wacholder, S. (2009). Invited commentary: from genome-wide association studies to gene-environment-wide interaction studies–challenges and opportunities. *Am J Epidemiol*, 169(2):227–30; discussion 234–5.

Klareskog, L., Stolt, P., Lundberg, K., Källberg, H., Bengtsson, C., Grunewald, J., Rönnelid, J., Harris, H. E., Ulfgren, A.-K., Rantapää-Dahlqvist, S., Eklund, A., Padyukov, L., and Alfredsson, L. (2006). A new model for an etiology of rheumatoid arthritis: smoking may trigger HLA-DR (shared epitope)-restricted immune reactions to autoantigens modified by citrullination. *Arthritis Rheum*, 54(1):38–46.

Klein, R. J., Zeiss, C., Chew, E. Y., Tsai, J.-Y., Sackler, R. S., Haynes, C., Henning, A. K., SanGiovanni, J. P., Mane, S. M., Mayne, S. T., Bracken, M. B., Ferris, F. L., Ott, J., Barnstable, C., and Hoh, J. (2005). Complement factor H polymorphism in age-related macular degeneration. *Science*, 308(5720):385–389.

Koopman, J. S. (1977). Causal models and sources of interaction. *Am J Epidemiol*, 106(6):439–444.

Korn, J. M., Kuruvilla, F. G., McCarroll, S. A., Wysoker, A., Nemesh, J., Cawley, S., Hubbell, E., Veitch, J., Collins, P. J., Darvishi, K., Lee, C., Nizzari, M. M., Gabriel, S. B., Purcell, S., Daly, M. J., and Altshuler, D. (2008). Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. *Nat Genet*, 40(10):1253–1260.

Kraft, P. (2006). Efficient two-stage genome-wide association designs based on false positive report probabilities. *Pac Symp Biocomput*, pages 523–534.

Kraja, A. T., Culverhouse, R., Daw, E. W., Wu, J., Brunt, A. V., Province, M. A., and Borecki, I. B. (2009). The genetic analysis workshop 16 problem 3: simulation of heritable longitudinal cardiovascular phenotypes based on actual genome-wide single-nucleotide polymorphisms in the framingham heart study. *BMC Proc*, 3 Suppl 7:S4.

Kupper, L. L. and Hogan, M. D. (1978). Interaction in epidemiologic studies. *Am J Epidemiol*, 108(6):447–453.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J. P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, N., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J. C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R. H., Wilson, R. K., Hillier, L. W., McPherson, J. D.,

Marra, M. A., Mardis, E. R., Fulton, L. A., Chinwalla, A. T., Pepin, K. H., Gish, W. R., Chissoe, S. L., Wendl, M. C., Delehaunty, K. D., Miner, T. L., Delehaunty, A., Kramer, J. B., Cook, L. L., Fulton, R. S., Johnson, D. L., Minx, P. J., Clifton, S. W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J. F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., Gibbs, R. A., Muzny, D. M., Scherer, S. E., Bouck, J. B., Sodergren, E. J., Worley, K. C., Rives, C. M., Gorrell, J. H., Metzker, M. L., Naylor, S. L., Kucherlapati, R. S., Nelson, D. L., Weinstock, G. M., Sakaki, Y., Fujiyama, A., Hattori, M., Yada, T., Toyoda, A., Itoh, T., Kawagoe, C., Watanabe, H., Totoki, Y., Taylor, T., Weissenbach, J., Heilig, R., Saurin, W., Artiguenave, F., Brottier, P., Bruls, T., Pelletier, E., Robert, C., Wincker, P., Smith, D. R., Doucette-Stamm, L., Rubenfield, M., Weinstock, K., Lee, H. M., Dubois, J., Rosenthal, A., Platzer, M., Nyakatura, G., Taudien, S., Rump, A., Yang, H., Yu, J., Wang, J., Huang, G., Gu, J., Hood, L., Rowen, L., Madan, A., Qin, S., Davis, R. W., Federspiel, N. A., Abola, A. P., Proctor, M. J., Myers, R. M., Schmutz, J., Dickson, M., Grimwood, J., Cox, D. R., Olson, M. V., Kaul, R., Raymond, C., Shimizu, N., Kawasaki, K., Minoshima, S., Evans, G. A., Athanasiou, M., Schultz, R., Roe, B. A., Chen, F., Pan, H., Ramser, J., Lehrach, H., Reinhardt, R., McCombie, W. R., de la Bastide, M., Dedhia, N., Blöcker, H., Hornischer, K., Nordsiek, G., Agarwala, R., Aravind, L., Bailey, J. A., Bateman, A., Batzoglou, S., Birney, E., Bork, P., Brown, D. G., Burge, C. B., Cerutti, L., Chen, H. C., Church, D., Clamp, M., Copley, R. R., Doerks, T., Eddy, S. R., Eichler, E. E., Furey, T. S., Galagan, J., Gilbert, J. G., Harmon, C., Hayashizaki, Y., Haussler, D., Hermjakob, H., Hokamp, K., Jang, W., Johnson, L. S., Jones, T. A., Kasif, S., Kaspryzk, A., Kennedy, S., Kent, W. J., Kitts, P., Koonin, E. V., Korf, I., Kulp, D., Lancet, D., Lowe, T. M., McLysaght, A., Mikkelsen, T., Moran, J. V., Mulder, N., Pollara, V. J., Ponting, C. P., Schuler, G., Schultz, J., Slater, G., Smit, A. F., Stupka, E., Szustakowski, J., Thierry-Mieg, D., Thierry-Mieg, J., Wagner, L., Wallis, J., Wheeler, R., Williams, A., Wolf, Y. I., Wolfe, K. H., Yang, S. P., Yeh, R. F., Collins, F., Guyer, M. S., Peterson, J., Felsenfeld, A., Wetterstrand, K. A., Patrinos, A., Morgan, M. J., de Jong, P., Catanese, J. J., Osoegawa, K., Shizuya, H., Choi, S., Chen, Y. J., Szustakowki, J., and Consortium, I. H. G. S. (2001). Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921.

Landi, M. T., Chatterjee, N., Yu, K., Goldin, L. R., Goldstein, A. M., Rotunno, M., Mirabello, L., Jacobs, K., Wheeler, W., Yeager, M., Bergen, A. W., Li, Q., Consonni, D., Pesatori, A. C., Wacholder, S., Thun, M., Diver, R., Oken, M., Virtamo, J., Albanes, D., Wang, Z., Burdette, L., Doheny, K. F., Pugh, E. W., Laurie, C., Brennan, P., Hung, R., Gaborieau, V., McKay, J. D., Lathrop, M., McLaughlin, J., Wang, Y., Tsao, M.-S., Spitz, M. R., Wang, Y., Krokan, H., Vatten, L., Skorpen, F., Arnesen, E., Benhamou, S., Bouchard, C., Metspalu, A., Metsapalu, A., Vooder, T., Nelis, M., Välk, K., Field, J. K., Chen, C., Goodman, G., Sulem, P., Thorleifsson, G., Rafnar, T., Eisen, T., Sauter, W., Rosenberger, A., Bickeböller, H., Risch, A., Chang-Claude, J., Wichmann, H. E., Stefansson, K., Houlston, R., Amos, C. I., Fraumeni, J. F., Savage, S. A., Bertazzi, P. A., Tucker, M. A., Chanock, S., and Caporaso, N. E. (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.

Le, C. (1991). *Fundamentals of Biostatistical Inference*. Statistics, textbooks and monographs. Taylor & Francis.

Lebrec, J. J., Huizinga, T. W., Toes, R. E., Houwing-Duistermaat, J. J., and van Houwelingen, H. C. (2009). Integration of gene ontology pathways with North American Rheumatoid Arthritis Consortium genome-wide association data via linear modeling. *BMC Proc*, 3 Suppl 7:S94.

Lee, H. K., Braynen, W., Keshav, K., and Pavlidis, P. (2005). ErmineJ: tool for functional analysis of gene expression data sets. *BMC Bioinformatics*, 6:269.

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

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

Li, J. and Wang, W.-B. (2010). Tag SNP Selection. In Zeggini, E. and Morris, A., editors, *Analysis of Complex Disease Association Studies: A Practical Guide*, volume 60 of *Academic Press*, pages 49 – 67. Elsevier Science.

Li, M., Li, C., and Guan, W. (2008). Evaluation of coverage variation of SNP chips for genome-wide association studies. *Eur J Hum Genet*, 16(5):635–643.

Li, Y. and Agarwal, P. (2009). A pathway-based view of human diseases and disease relationships. *PLoS One*, 4(2):e4346.

Li, Y., Willer, C., Sanna, S., and Abecasis, G. (2009). Genotype imputation. *Annu Rev Genomics Hum Genet*, 10:387–406.

Lin, K.-T., Liu, C.-H., Chiou, J.-J., Tseng, W.-H., Lin, K.-L., and Hsu, C.-N. (2007). Gene name service: No-nonsense alias resolution service for homo sapiens genes. In *Proceedings of the 2007 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology - Workshops*, WI-IATW '07, pages 185–188, Washington, DC, USA. IEEE Computer Society.

Liu, N., Zhang, K., and Zhao, H. (2008). Haplotype-association analysis. In Rao, D. C. and Gu, C. C., editors, *Genetic Dissection of Complex Traits*, volume 60 of *Advances in Genetics*, pages 335 – 405. Academic Press.

Los, H., Postmus, P. E., and Boomsma, D. I. (2001). Asthma genetics and intermediate phenotypes: a review from twin studies. *Twin Res*, 4(2):81–93.

M. Inouye, Y. Y. T. (2010). Genotype calling. In Zeggini, E. and Morris, A., editors, *Analysis of Complex Disease Association Studies: A Practical Guide*, volume 60 of *Academic Press*, pages 69–86. Elsevier Science.

Maher, B. (2008). Personal genomes: The case of the missing heritability. *Nature*, 456(7218):18–21.

Maher, S. G., Condron, C. E. M., Bouchier-Hayes, D. J., and Toomey, D. M. (2005). Taurine attenuates CD3/interleukin-2-induced T cell apoptosis in an in vitro model of activation-induced cell death (AICD). *Clin Exp Immunol*, 139(2):279–286.

Manolio, T. A. and Collins, F. S. (2007). Genes, environment, health, and disease: facing up to complexity. *Hum Hered*, 63(2):63–66.

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

Marchand, L. L. (2005). The predominance of the environment over genes in cancer causation: implications for genetic epidemiology. *Cancer Epidem Biomar*, 14(5):1037–1039.

Marchand, L. L. and Wilkens, L. R. (2008). Design considerations for genomic association studies: importance of gene-environment interactions. *Cancer Epidem Biomar*, 17(2):263–267.

Marchini, J., Cardon, L. R., Phillips, M. S., and Donnelly, P. (2004). The effects of human population structure on large genetic association studies. *Nat Genet*, 36(5):512–517.

Marchini, J., Cutler, D., Patterson, N., Stephens, M., Eskin, E., Halperin, E., Lin, S., Qin, Z. S., Munro, H. M., Abecasis, G. R., Donnelly, P., and Consortium, I. H. (2006). A comparison of phasing algorithms for trios and unrelated individuals. *Am J Hum Genet*, 78(3):437–450.

Marchini, J., Howie, B., Myers, S., McVean, G., and Donnelly, P. (2007). A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat Genet*, 39(7):906–913.

Maresso, K. and Broeckel, U. (2008). Genotyping Platforms for Mass-Throughput Genotyping with SNPs, Including Human Genome-Wide Scans. In Rao, D. C. and Gu, C. C., editors, *Genetic Dissection of Complex Traits*, volume 60 of *Advances in Genetics*, pages 107 – 139. Academic Press.

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

Mayne, S. T., Buenconsejo, J., and Janerich, D. T. (1999). Previous lung disease and risk of lung cancer among men and women nonsmokers. *Am J Epidemiol*, 149(1):13–20.

Mazieres, J., He, B., You, L., Xu, Z., and Jablons, D. M. (2005). Wnt signaling in lung cancer. *Cancer Lett*, 222(1):1–10.

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

McPherson, J. D., Marra, M., Hillier, L., Waterston, R. H., Chinwalla, A., Wallis, J., Sekhon, M., Wylie, K., Mardis, E. R., Wilson, R. K., and et al. (2001). A physical map of the human genome. *Nature*, 409(6822):934–41.

Melino, G., Laurenzi, V. D., and Vousden, K. H. (2002). p73: Friend or foe in tumorigenesis. *Nat Rev Cancer*, 2(8):605–615.

Melville, S. (2011). *NCBI2R: NCBI2R-An R package to navigate and annotate genes and SNPs*. R package version 1.3.3.

Michael, D. and Oren, M. (2002). The p53 and MDM2 families in cancer. *Curr Opin Genet Dev*, 12(1):53–59.

Mitchell, M. K., Gregersen, P. K., Johnson, S., Parsons, R., Vlahov, D., and Project, N. Y. C. (2004). The new york cancer project: rationale, organization, design, and baseline characteristics. *J Urban Health*, 81(2):301–310.

Moore, J. H. (2003). The ubiquitous nature of epistasis in determining susceptibility to common human diseases. *Hum Hered*, 56(1-3):73–82.

Moore, J. H. and Williams, S. M. (2005). Traversing the conceptual divide between biological and statistical epistasis: systems biology and a more modern synthesis. *Bioessays*, 27(6):637–646.

Mootha, V. K., Lindgren, C. M., Eriksson, K.-F., Subramanian, A., Sihag, S., Lehar, J., Puigserver, P., Carlsson, E., Ridderstraale, M., Laurila, E., Houstis, N., Daly, M. J., Patterson, N., Mesirov, J. P., Golub, T. R., Tamayo, P., Spiegelman, B., Lander, E. S., Hirschhorn, J. N., Altshuler, D., and Groop, L. C. (2003). PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat Genet*, 34(3):267–273.

Morris, C. N. (1983). Parametric empirical Bayes inference: Theory and applications: Rejoinder. *J Am Stat Assoc*, 78(381):63–65.

Mukherjee, B., Ahn, J., Gruber, S. B., and Chatterjee, N. (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., Ahn, J., Gruber, S. B., Rennert, G., Moreno, V., and Chatterjee, N. (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 Chatterjee, N. (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.

Murata, K., Hattori, M., Hirai, N., Shinozuka, Y., Hirata, H., Kageyama, R., Sakai, T., and Minato, N. (2005). Hes1 directly controls cell proliferation through the transcriptional repression of p27Kip1. *Mol Cell Biol*, 25(10):4262–4271.

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

Murphy, A. and Borowiec, J. (April 2010). RPA2 (replication protein A2, 32kDa). Atlas Genet Cytogenet Oncol Haematol. /url-http://crittweb.ensma.fr/Atlas/Genes/RPA2ID42146ch1p35.html.

Nam, D. and Kim, S.-Y. (2008). Gene-set approach for expression pattern analysis. *Brief Bioinform*, 9(3):189–197.

Nannya, Y., Taura, K., Kurokawa, M., Chiba, S., and Ogawa, S. (2007). Evaluation of genome-wide power of genetic association studies based on empirical data from the hapmap project. *Hum Mol Genet*, 16(20):2494–2505.

National Cancer Institute (NCI) (2012). Comprehensive cancer information. Website http://cancer.gov.

National Center for Biotechnology Information (NCBI) (2012). U.S. National Library of Medicine, 8600 Rockville Pike, Bethesda MD, 20894 USA. Website http://www.ncbi.nlm.nih.gov/.

National Institute of Environmental Health Sciences (NIEHS) (2011). Gene-environment interaction. Available at http://www.niehs.nih.gov/health/topics/science/gene-env/index.cfm[Accessed 30 August 2011].

National Institutes of Health (NIH) (2012). NIH ...Turning Discovery into Health. U.S. Department of Health & Human Services. Website http://www.nih.gov/.

Neale, B. M. and Purcell, S. (2008). The positives, protocols, and perils of genome-wide association. *Am J Med Genet B Neuropsychiatr Genet*, 147B(7):1288–1294.

Neale, B. M. and Sham, P. C. (2004). The future of association studies: gene-based analysis and replication. *Am J Hum Genet*, 75(3):353–362.

Nelis, M., Esko, T., Mägi, R., Zimprich, F., Zimprich, A., Toncheva, D., Karachanak, S., Piskácková, T., Balascák, I., Peltonen, L., Jakkula, E., Rehnström, K., Lathrop, M., Heath, S., Galan, P., Schreiber, S., Meitinger, T., Pfeufer, A., Wichmann, H.-E., Melegh, B., Polgár, N., Toniolo, D., Gasparini, P., D'Adamo, P., Klovins, J., Nikitina-Zake, L., Kucinskas, V., Kasnauskiene, J., Lubinski, J., Debniak, T., Limborska, S., Khrunin, A., Estivill, X., Rabionet, R., Marsal, S., Julià, A., Antonarakis, S. E., Deutsch, S., Borel, C., Attar, H., Gagnebin, M., Macek, M., Krawczak, M., Remm, M., and Metspalu, A. (2009). Genetic structure of Europeans: a view from the North-East. *PLoS One*, 4(5):e5472.

Neuman, R. J. and Sung, Y. J. (2009). Multistage analysis strategies for genome-wide association studies: summary of group 3 contributions to Genetic Analysis Workshop 16. *Genet Epidemiol*, 33 Suppl 1:S19–S23.

Newton, J. L., Harney, S. M. J., Wordsworth, B. P., and Brown, M. A. (2004). A review of the MHC genetics of rheumatoid arthritis. *Genes Immun*, 5(3):151–157.

Nikolova, T., Christmann, M., and Kaina, B. (2009). FEN1 is overexpressed in testis, lung and brain tumors. *Anticancer Res*, 29(7):2453–2459.

Nomori, H., Mori, T., Iyama, K., Okamoto, T., and Kamakura, M. (2011). Risk of bronchioloalveolar carcinoma in patients with human T-cell lymphotropic virus type 1 (HTLV-I): case-control study results. *Ann Thorac Cardiovasc Surg*, 17(1):19–23.

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

Office of Genetics and Disease Prevention (2000). Gene-environment interaction fact sheet. Available at http://www.ashg.org/pdf/CDCGene-EnvironmentInteractionFactSheet.pdf [Accessed 3 March 2012].

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

Ottman, R. (1990). An epidemiologic approach to gene-environment interaction. *Genet Epidemiol*, 7(3):177–185.

Ottman, R. (1996). Gene-environment interaction: definitions and study designs. *Prev Med*, 25(6):764–770.

Palmer, L. J. and Cardon, L. R. (2005). Shaking the tree: mapping complex disease genes with linkage disequilibrium. *Lancet*, 366(9492):1223–1234.

Pan, W. (2005). Incorporating biological information as a prior in an empirical Bayes approach to analyzing microarray data. *Stat Appl Genet Mol Biol*, 4:Article12.

Papoulis, A. and Pillai, S. (2002). *Probability, random variables, and stochastic processes*. McGraw-Hill electrical and electronic engineering series. McGraw-Hill.

Park, C. C., Ahn, S., Bloom, J. S., Lin, A., Wang, R. T., Wu, T., Sekar, A., Khan, A. H., Farr, C. J., Lusis, A. J., Leahy, R. M., Lange, K., and Smith, D. J. (2008). Fine mapping of regulatory loci for mammalian gene expression using radiation hybrids. *Nat Genet*, 40(4):421–429.

Park, S. H., Lee, J. Y., and Kim, S. (2011). A methodology for multivariate phenotype-based genome-wide association studies to mine pleiotropic genes. *BMC Syst Biol*, 5(Suppl 2):S13.

Parkin, D. M., Bray, F., Ferlay, J., and Pisani, P. (2005). Global cancer statistics, 2002. *CA -Cancer J Clin*, 55(2):74–108.

Patil, N., Berno, A. J., Hinds, D. A., Barrett, W. A., Doshi, J. M., Hacker, C. R., Kautzer, C. R., Lee, D. H., Marjoribanks, C., McDonough, D. P., Nguyen, B. T., Norris, M. C., Sheehan, J. B., Shen, N., Stern, D., Stokowski, R. P., Thomas, D. J., Trulson, M. O., Vyas, K. R., Frazer, K. A., Fodor, S. P., and Cox, D. R. (2001). Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science*, 294(5547):1719–1723.

Patterson, M. and Cardon, L. (2005). Replication publication. *PLoS Biol*, 3(9):e327.

Patterson, N., Price, A. L., and Reich, D. (2006). Population structure and eigenanalysis. *PLoS Genet*, 2(12):e190.

Pearson, T. A. and Manolio, T. A. (2008). How to interpret a genome-wide association study. *JAMA*, 299(11):1335–1344.

Peng, G., Luo, L., Siu, H., Zhu, Y., Hu, P., Hong, S., Zhao, J., Zhou, X., Reveille, J. D., Jin, L., Amos, C. I., and Xiong, M. (2010). Gene and pathway-based second-wave analysis of genome-wide association studies. *Eur J Hum Genet*, 18(1):111–117.

Perry, J. R. B., McCarthy, M. I., Hattersley, A. T., Zeggini, E., Consortium, W. T. C. C., Weedon, M. N., and Frayling, T. M. (2009). Interrogating type 2 diabetes genome-wide association data using a biological pathway-based approach. *Diabetes*, 58(6):1463–1467.

Pestman, W. (1998). *Mathematical statistics: an introduction*. De Gruyter textbook. Walter de Gruyter.

Phillips, M. S., Lawrence, R., Sachidanandam, R., Morris, A. P., Balding, D. J., Donaldson, M. A., Studebaker, J. F., Ankener, W. M., Alfisi, S. V., Kuo, F.-S., Camisa, A. L., Pazorov, V., Scott, K. E., Carey, B. J., Faith, J., Katari, G., Bhatti, H. A., Cyr, J. M., Derohannessian, V., Elosua, C., Forman, A. M., Grecco, N. M., Hock, C. R., Kuebler, J. M., Lathrop, J. A., Mockler, M. A., Nachtman, E. P., Restine, S. L., Varde, S. A., Hozza, M. J., Gelfand, C. A., Broxholme, J., Abecasis, G. R., Boyce-Jacino, M. T., and Cardon, L. R. (2003). Chromosome-wide distribution of haplotype blocks and the role of recombination hot spots. *Nat Genet*, 33(3):382–387.

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

Piyathilake, C. J., Macaluso, M., Hine, R. J., Richards, E. W., and Krumdieck, C. L. (1994). Local and systemic effects of cigarette smoking on folate and vitamin B-12. *Am J Clin Nutr*, 60(4):559–566.

Plenge, R. M., Padyukov, L., Remmers, E. F., Purcell, S., Lee, A. T., Karlson, E. W., Wolfe, F., Kastner, D. L., Alfredsson, L., Altshuler, D., Gregersen, P. K., Klareskog, L., and Rioux, J. D. (2005). Replication of putative candidate-gene associations with rheumatoid arthritis in >4,000 samples from North America and Sweden: association of susceptibility with PTPN22, CTLA4, and PADI4. *Am J Hum Genet*, 77(6):1044–1060.

Plenge, R. M., Seielstad, M., Padyukov, L., Lee, A. T., Remmers, E. F., Ding, B., Liew, A., Khalili, H., Chandrasekaran, A., Davies, L. R. L., Li, W., Tan, A. K. S., Bonnard, C., Ong, R. T. H., Thalamuthu, A., Pettersson, S., Liu, C., Tian, C., Chen, W. V., Carulli, J. P., Beckman, E. M., Altshuler, D., Alfredsson, L., Criswell, L. A., Amos, C. I., Seldin, M. F., Kastner, D. L., Klareskog, L., and Gregersen, P. K. (2007). TRAF1-C5 as a risk locus for rheumatoid arthritis–a genomewide study. *N Engl J Med*, 357(12):1199–1209.

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

Pritchard, J. K. (2001). Are rare variants responsible for susceptibility to complex diseases? *Am J Hum Genet*, 69(1):124–137.

Pritchard, J. K. and Cox, N. J. (2002). The allelic architecture of human disease genes: common disease-common variant...or not? *Hum Mol Genet*, 11(20):2417–2423.

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

Pöschl, G. and Seitz, H. K. (2004). Alcohol and cancer. *Alcohol Alcohol*, 39(3):155–165.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., Maller, J., Sklar, P., Bakker, P. I. W. d., Daly, M. J., and et al. (2007). PLINK: A tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet*, 81(3):559–575.

Qiagen Sample and Assay Technology (2012). The Bench Guide: What is a protein? Available at http://www.qiagen.com/literature/benchguide/pdf/1017778_benchguide_chap_4.pdf [Accessed March 13th 2012].

Rafnar, T., Sulem, P., Stacey, S. N., Geller, F., Gudmundsson, J., Sigurdsson, A., Jakobsdottir, M., Helgadottir, H., Thorlacius, S., Aben, K. K. H., Blöndal, T., Thorgeirsson, T. E., Thorleifsson, G., Kristjansson, K., Thorisdottir, K., Ragnarsson, R., Sigurgeirsson, B., Skuladottir, H., Gudbjartsson, T., Isaksson, H. J., Einarsson, G. V., Benediktsdottir, K. R., Agnarsson, B. A., Olafsson, K., Salvarsdottir, A., Bjarnason, H., Asgeirsdottir, M., Kristinsson, K. T., Matthiasdottir, S., Sveinsdottir, S. G., Polidoro, S., Höiom, V., Botella-Estrada, R., Hemminki, K., Rudnai, P., Bishop, D. T.,

Campagna, M., Kellen, E., Zeegers, M. P., de Verdier, P., Ferrer, A., Isla, D., Vidal, M. J., Andres, R., Saez, B., Juberias, P., Banzo, J., Navarrete, S., Tres, A., Kan, D., Lindblom, A., Gurzau, E., Koppova, K., de Vegt, F., Schalken, J. A., van der Heijden, H. F. M., Smit, H. J., Termeer, R. A., Oosterwijk, E., van Hooij, O., Nagore, E., Porru, S., Steineck, G., Hansson, J., Buntinx, F., Catalona, W. J., Matullo, G., Vineis, P., Kiltie, A. E., Mayordomo, J. I., Kumar, R., Kiemeney, L. A., Frigge, M. L., Jonsson, T., Saemundsson, H., Barkardottir, R. B., Jonsson, E., Jonsson, S., Olafsson, J. H., Gulcher, J. R., Masson, G., Gudbjartsson, D. F., Kong, A., Thorsteinsdottir, U., and Stefansson, K. (2009). Sequence variants at the TERT-CLPTM1L locus associate with many cancer types. *Nat Genet*, 41(2):221–227.

Rakyan, V. K., Down, T. A., Balding, D. J., and Beck, S. (2011). Epigenome-wide association studies for common human diseases. *Nat Rev Genet*, 12(8):529–541.

Rao, D. (2008). An overview of the genetic dissection of complex traits. In Rao, D. C. and Gu, C. C., editors, *Genetic Dissection of Complex Traits*, volume 60 of *Advances in Genetics*, pages 3 – 34. Academic Press.

Rao, D. and Gu, C. (2008). *Genetic dissection of complex traits*. Advances in genetics. Academic Press.

Raychaudhuri, S. (2010). Recent advances in the genetics of rheumatoid arthritis. *Curr Opin Rheumatol*, 22(2):109–118.

Raychaudhuri, S., Remmers, E. F., Lee, A. T., Hackett, R., Guiducci, C., Burtt, N. P., Gianniny, L., Korman, B. D., Padyukov, L., Kurreeman, F. A. S., Chang, M., Catanese, J. J., Ding, B., Wong, S., van der Helm-van Mil, A. H. M., Neale, B. M., Coblyn, J., Cui, J., Tak, P. P., Wolbink, G. J., Crusius, J. B. A., van der Horst-Bruinsma, I. E., Criswell, L. A., Amos, C. I., Seldin, M. F., Kastner, D. L., Ardlie, K. G., Alfredsson, L., Costenbader, K. H., Altshuler, D., Huizinga, T. W. J., Shadick, N. A., Weinblatt, M. E., de Vries, N., Worthington, J., Seielstad, M., Toes, R. E. M., Karlson, E. W., Begovich, A. B., Klareskog, L., Gregersen, P. K., Daly, M. J., and Plenge, R. M. (2008). Common variants at CD40 and other loci confer risk of rheumatoid arthritis. *Nat Genet*, 40(10):1216–1223.

Raychaudhuri, S., Thomson, B. P., Remmers, E. F., Eyre, S., Hinks, A., Guiducci, C., Catanese, J. J., Xie, G., Stahl, E. A., Chen, R., Alfredsson, L., Amos, C. I., Ardlie, K. G., Consortium, B. I. R. A. C., Barton, A., Bowes, J., Burtt, N. P., Chang, M., Coblyn, J., Costenbader, K. H., Criswell, L. A., Crusius, J. B. A., Cui, J., Jager, P. L. D., Ding, B., Emery, P., Flynn, E., Harrison, P., Hocking, L. J., Huizinga, T. W. J., Kastner, D. L., Ke, X., Kurreeman, F. A. S., Lee, A. T., Liu, X., Li, Y., Martin, P., Morgan, A. W., Padyukov, L., Reid, D. M., Seielstad, M., Seldin, M. F., Shadick, N. A., Steer, S., Tak, P. P., Thomson, W., van der Helm-van Mil, A. H. M., van der Horst-Bruinsma, I. E., Weinblatt, M. E., Wilson, A. G., Wolbink, G. J., Wordsworth, P., Consortium, Y. E. A. R., Altshuler, D., Karlson, E. W., Toes, R. E. M., de Vries, N., Begovich, A. B., Siminovitch, K. A., Worthington, J., Klareskog, L., Gregersen, P. K., Daly, M. J., and Plenge, R. M. (2009). Genetic variants at CD28, PRDM1 and CD2/CD58 are associated with rheumatoid arthritis risk. *Nat Genet*, 41(12):1313–1318.

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

Reich, D. E. and Lander, E. S. (2001). On the allelic spectrum of human disease. *Trends Genet*, 17(9):502–510.

Remmers, E. F., Plenge, R. M., Lee, A. T., Graham, R. R., Hom, G., Behrens, T. W., de Bakker, P. I. W., Le, J. M., Lee, H.-S., Batliwalla, F., Li, W., Masters, S. L., Booty, M. G., Carulli, J. P., Padyukov, L., Alfredsson, L., Klareskog, L., Chen, W. V., Amos, C. I., Criswell, L. A., Seldin, M. F., Kastner, D. L., and Gregersen, P. K. (2007). STAT4 and the risk of rheumatoid arthritis and systemic lupus erythematosus. *N Engl J Med*, 357(10):977–986.

Rice, T. K., Schork, N. J., and Rao, D. (2008). Methods for handling multiple testing. In Rao, D. C. and Gu, C. C., editors, *Genetic Dissection of Complex Traits*, volume 60 of *Advances in Genetics*, pages 293 – 308. Academic Press.

Riordan, J. R., Rommens, J. M., Kerem, B., Alon, N., Rozmahel, R., Grzelczak, Z., Zielenski, J., Lok, S., Plavsic, N., and Chou, J. L. (1989). Identification of the cystic fibrosis gene: cloning and characterization of complementary DNA. *Science*, 245(4922):1066–1073.

Risch, N. and Merikangas, K. (1996). The future of genetic studies of complex human diseases. *Science*, 273(5281):1516–1517.

Risch, N. J. (2000). Searching for genetic determinants in the new millennium. *Nature*, 405(6788):847–856.

Robert, C. P. (1994). *The Bayesian Choice: a decision-theoretic motivation*. Springer, New York.

Robert Koch-Institut (RKI) (2011). Sterblichkeit, Todesursachen und regionale Unterschiede. Gesundheit in Deutschland. Gesundheitsberichterstattung des Bundes. Heft 52. Robert Koch-Institut, Berlin.

Robert Koch-Institut (RKI) und die Gesellschaft der epidemiologischen Krebsregister in Deutschland e.V. (GEKID) (2012). Krebs in Deutschland 2007/2008. 8. Ausgabe. Gesundheitsberichterstattung des Bundes. Berlin. Available at http://www.krebsdaten.de/Krebs/DE/Content/Publikationen/Krebs_in_Deutschland/krebs_in_deutschland_node.html [Accessed 15 March 2012].

Roeder, K., Bacanu, S.-A., Wasserman, L., and Devlin, B. (2006). Using linkage genome scans to improve power of association in genome scans. *Am J Hum Genet*, 78(2):243–252.

Roeder, K., Devlin, B., and Wasserman, L. (2007). Improving power in genome-wide association studies: weights tip the scale. *Genet Epidemiol*, 31(7):741–747.

Rom, D. M. (1990). A sequentially rejective test procedure based on a modified Bonferroni inequality. *Biometrika*, 77(3):663.

Rothman, K. J., Greenland, S., and Walker, A. M. (1980). Concepts of interaction. *Am J Epidemiol*, 112(4):467–470.

Ruano-Ravina, A., Figueiras, A., Montes-Martínez, A., and Barros-Dios, J. M. (2003). Dose-response relationship between tobacco and lung cancer: new findings. *Eur J Cancer Prev*, 12(4):257–263.

Sabatti, C., Service, S., and Freimer, N. (2003). False discovery rate in linkage and association genome screens for complex disorders. *Genetics*, 164(2):829–833.

Sachidanandam, R., Weissman, D., Schmidt, S. C., Kakol, J. M., Stein, L. D., Marth, G., Sherry, S., Mullikin, J. C., Mortimore, B. J., Willey, D. L., Hunt, S. E., Cole, C. G., Coggill, P. C., Rice, C. M., Ning, Z., Rogers, J., Bentley, D. R., Kwok, P. Y., Mardis, E. R., Yeh, R. T., Schultz, B., Cook, L., Davenport, R., Dante, M., Fulton, L., Hillier, L., Waterston, R. H., McPherson, J. D., Gilman, B., Schaffner, S., Etten, W. J. V., Reich, D., Higgins, J., Daly, M. J., Blumenstiel, B., Baldwin, J., Stange-Thomann, N., Zody, M. C., Linton, L., Lander, E. S., Altshuler, D., and Group, I. S. M. W. (2001). A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature*, 409(6822):928–933.

Saito, A. and Kamatani, N. (2002). Strategies for genome-wide association studies: optimization of study designs by the stepwise focusing method. *J Hum Genet*, 47(7):360–365.

Sakoda, L. C., Loomis, M. M., Doherty, J. A., Julianto, L., Barnett, M. J., Neuhouser, M. L., Thornquist, M. D., Weiss, N. S., Goodman, G. E., and Chen, C. (2012). Germ line variation in nucleotide excision repair genes and lung cancer risk in smokers. *Int J Mol Epidemiol Genet*, 3(1):1–17.

Salaspuro, M. P. (2003). Alcohol consumption and cancer of the gastrointestinal tract. *Best Pract Res Clin Gastroenterol*, 17(4):679–694.

Salaspuro, V. J., Hietala, J. M., Marvola, M. L., and Salaspuro, M. P. (2006). Eliminating carcinogenic acetaldehyde by cysteine from saliva during smoking. *Cancer Epidem Biomar*, 15(1):146–149.

Salomonis, N., Hanspers, K., Zambon, A. C., Vranizan, K., Lawlor, S. C., Dahlquist, K. D., Doniger, S. W., Stuart, J., Conklin, B. R., and Pico, A. R. (2007). Genmapp 2: new features and resources for pathway analysis. *BMC Bioinformatics*, 8:217.

Samani, N. J., Erdmann, J., Hall, A. S., Hengstenberg, C., Mangino, M., Mayer, B., Dixon, R. J., Meitinger, T., Braund, P., Wichmann, H.-E., Barrett, J. H., Koenig, I. R., Stevens, S. E., Szymczak, S., Tregouet, D.-A., Iles, M. M., Pahlke, F., Pollard, H., Lieb, W., Cambien, F., Fischer, M., Ouwehand, W., Blankenberg, S., Balmforth, A. J., Baessler, A., Ball, S. G., Strom, T. M., Brnne, I., Gieger, C., Deloukas, P., Tobin, M. D., Ziegler, A., Thompson, J. R., and Schunkert, H. (2007). Genomewide association analysis of coronary artery disease. *N Engl J Med*, 357(5):443–453.

Saraiya, P., North, C., and Duca, K. (2005). Visualizing biological pathways: requirements analysis, systems evaluation and research agenda. *Inf Vis*, 4(3):191–205.

Satagopan, J. M. and Elston, R. C. (2003). Optimal two-stage genotyping in population-based association studies. *Genet Epidemiol*, 25(2):149–157.

Satagopan, J. M., Venkatraman, E. S., and Begg, C. B. (2004). Two-stage designs for gene-disease association studies with sample size constraints. *Biometrics*, 60(3):589–597.

Satagopan, J. M., Verbel, D. A., Venkatraman, E. S., Offit, K. E., and Begg, C. B. (2002). Two-stage designs for gene-disease association studies. *Biometrics*, 58(1):163–170.

Saunders Company, W. (1968). *Dorland's medical dictionary*. American Printing House for the Blind.

Sauter, W., Rosenberger, A., Beckmann, L., Kropp, S., Mittelstrass, K., Timofeeva, M., Wölke, G., Steinwachs, A., Scheiner, D., Meese, E., and et al. (2008). Matrix metalloproteinase 1 (MMP1) is associated with early-onset lung cancer. *Cancer Epidem Biomar*, 17(5):1127–1135.

Saxena, R., Voight, B. F., Lyssenko, V., Burtt, N. P., de Bakker, P. I. W., Chen, H., Roix, J. J., Kathiresan, S., Hirschhorn, J. N., Daly, M. J., Hughes, T. E., Groop, L., Altshuler, D., Almgren, P., Florez, J. C., Meyer, J., Ardlie, K., Boström, K. B., Isomaa, B., Lettre, G., Lindblad, U., Lyon, H. N., Melander, O., Newton-Cheh, C., Nilsson, P., Orho-Melander, M., Råstam, L., Speliotes, E. K., Taskinen, M.-R., Tuomi, T., Guiducci, C., Berglund, A., Carlson, J., Gianniny, L., Hackett, R., Hall, L., Holmkvist, J., Laurila, E., Sjögren, M., Sterner, M., Surti, A., Svensson, M., Svensson, M., Tewhey, R., Blumenstiel, B., Parkin, M., Defelice, M., Barry, R., Brodeur, W., Camarata, J., Chia, N., Fava, M., Gibbons, J., Handsaker, B., Healy, C., Nguyen, K., Gates, C., Sougnez, C., Gage, D., Nizzari, M., Gabriel, S. B., Chirn, G.-W., Ma, Q., Parikh, H., Richardson, D., Ricke, D., and Purcell, S. (2007). Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. *Science*, 316(5829):1331–1336.

Schadt, E. E. (2009). Molecular networks as sensors and drivers of common human diseases. *Nature*, 461(7261):218–223.

Scheet, P. and Stephens, M. (2006). A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *Am J Hum Genet*, 78(4):629–644.

Scheuner, M. T., Yoon, P. W., and Khoury, M. J. (2004). Contribution of mendelian disorders to common chronic disease: opportunities for recognition, intervention, and prevention. *Am J Med Genet C Semin Med Genet*, 125C(1):50–65.

Schmidt, S. and Schaid, D. J. (1999). Potential misinterpretation of the case-only study to assess gene-environment interaction. *Am J Epidemiol*, 150(8):878–885.

Schwartz, D. A. (2006). The importance of gene-environment interactions and exposure assessment in understanding human diseases. *J Expo Sci Environ Epidemiol*, 16(6):474–476.

Scélo, G., Constantinescu, V., Csiki, I., Zaridze, D., Szeszenia-Dabrowska, N., Rudnai, P., Lissowska, J., Fabiánová, E., Cassidy, A., Slamova, A., and et al. (2004). Occupational exposure to vinyl chloride, acrylonitrile and styrene and lung cancer risk (Europe). *Cancer causes control*, 15(5):445–452.

Selikoff, I. J., Hammond, E. C., and Churg, J. (1968). Asbestos exposure, smoking, and neoplasia. *JAMA*, 204(2):106–112.

Sham, P., Bader, J. S., Craig, I., O'Donovan, M., and Owen, M. (2002). DNA pooling: a tool for large-scale association studies. *Nat Rev Genet*, 3(11):862–871.

Sham, P. and Cherny, S. (2010). Genetic architecture of complex diseases. In Zeggini, E. and Morris, A., editors, *Analysis of Complex Disease Association Studies: A Practical Guide*, volume 60 of *Academic Press*, pages 1–13. Elsevier Science.

Sherrington, R., Rogaev, E. I., Liang, Y., Rogaeva, E. A., Levesque, G., Ikeda, M., Chi, H., Lin, C., Li, G., Holman, K., Tsuda, T., Mar, L., Foncin, J. F., Bruni, A. C., Montesi, M. P., Sorbi, S., Rainero, I., Pinessi, L., Nee, L., Chumakov, I., Pollen, D., Brookes, A., Sanseau, P., Polinsky, R. J., Wasco, W., Silva, H. A. D., Haines, J. L., Perkicak-Vance, M. A., Tanzi, R. E., Roses, A. D., Fraser, P. E., Rommens, J. M., and George-Hyslop, P. H. S. (1995). Cloning of a gene bearing missense mutations in early-onset familial Alzheimer's disease. *Nature*, 375(6534):754–760.

Shifman, S., Kuypers, J., Kokoris, M., Yakir, B., and Darvasi, A. (2003). Linkage disequilibrium patterns of the human genome across populations. *Hum Mol Genet*, 12(7):771–776.

Sidak, Z. (1967). Rectangular confidence regions for the means of multivariate normal distributions. *J Am Stat Assoc*, 62(318):626–633.

Siemiatycki, J. and Thomas, D. C. (1981). Biological models and statistical interactions: an example from multistage carcinogenesis. *Int J Epidemiol*, 10(4):383–387.

Simes, R. J. (1986). An improved Bonferroni procedure for multiple tests of significance. *Biometrika*, 73(3):751–754.

Skol, A. D., Scott, L. J., Abecasis, G. R., and Boehnke, M. (2006). Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies. *Nat Genet*, 38(2):209–213.

Smith, G. D., Ebrahim, S., Lewis, S., Hansell, A. L., Palmer, L. J., and Burton, P. R. (2005). Genetic epidemiology and public health: hope, hype, and future prospects. *Lancet*, 366(9495):1484–1498.

Sohns, M., Rosenberger, A., and Bickeböller, H. (2009). Integration of a priori gene set information into genome-wide association studies. *BMC Proc*, 3 Suppl 7:S95.

Soussi, T. (2010). The TP53 Web Site. Available at `http://p53.free.fr/` [Accessed 3 August 2011].

Sprince, H., Parker, C. M., Smith, G. G., and Gonzales, L. J. (1975). Protective action of ascorbic acid and sulfur compounds against acetaldehyde toxicity: implications in alcoholism and smoking. *Agents Actions*, 5(2):164–173.

Steffens, M., Lamina, C., Illig, T., Bettecken, T., Vogler, R., Entz, P., Suk, E.-K., Toliat, M. R., Klopp, N., Caliebe, A., König, I. R., Köhler, K., Ludemann, J., Lacava, A. D., Fimmers, R., Lichtner, P., Ziegler, A., Wolf, A., Krawczak, M., Nürnberg, P., Hampe, J., Schreiber, S., Meitinger, T., Wichmann, H.-E., Roeder, K., Wienker, T. F., and Baur, M. P. (2006). SNP-based analysis of genetic substructure in the German population. *Hum Hered*, 62(1):20–29.

Stephens, M., Smith, N. J., and Donnelly, P. (2001). A new statistical method for haplotype reconstruction from population data. *Am J Hum Genet*, 68(4):978–989.

Storey, J. D. (2002). A direct approach to false discovery rates. *J R Stat Soc Series B Stat Methodol*, 64(3):479–498.

Storey, J. D. (2003). The positive false discovery rate: a Bayesian interpretation and the q -value. *Ann Stat*, 31(6):2013–2035.

Storey, J. D. and Tibshirani, R. (2003). Statistical significance for genomewide studies. *Proc Natl Acad Sci U S A*, 100(16):9440–9445.

Stranger, B. E., Stahl, E. A., and Raj, T. (2011). Progress and promise of genome-wide association studies for human complex trait genetics. *Genetics*, 187(2):367–383.

Strittmatter, W. J., Saunders, A. M., Schmechel, D., Pericak-Vance, M., Enghild, J., Salvesen, G. S., and Roses, A. D. (1993). Apolipoprotein E: high-avidity binding to beta-amyloid and increased frequency of type 4 allele in late-onset familial Alzheimer disease. *Proc Natl Acad Sci U S A*, 90(5):1977–1981.

Suarez, B. K. and Hampe, C. L. (1994). Linkage and association. *Am J Hum Genet*, 54(3):554–9; author reply 560–3.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., and Mesirov, J. P. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A*, 102(43):15545–15550.

Sun, S., Schiller, J. H., and Gazdar, A. F. (2007). Lung cancer in never smokers - a different disease. *Nat Rev Cancer*, 7(10):778–790.

Sun, Y. V., Sung, Y. J., Tintle, N., and Ziegler, A. (2011). Identification of genetic association of multiple rare variants using collapsing methods. *Genet Epidemiol*, 35 Suppl 1:S101–S106.

Syvänen, A. C. (2001). Accessing genetic variation: genotyping single nucleotide polymorphisms. *Nat Rev Genet*, 2(12):930–942.

Tabor, H. K., Risch, N. J., and Myers, R. M. (2002). Candidate-gene approaches for studying complex genetic traits: practical considerations. *Nat Rev Genet*, 3(5):391–397.

Takemiya, M., Shiraishi, S., Teramoto, T., and Miki, Y. (1987). Bloom's syndrome with porokeratosis of mibelli and multiple cancers of the skin, lung and colon. *Clin Genet*, 31(1):35–44.

Tanzi, R. E., Bird, E. D., Latt, S. A., and Neve, R. L. (1987). The amyloid beta protein gene is not duplicated in brains from patients with Alzheimer's disease. *Science*, 238(4827):666–669.

Tennis, M., Scoyk, M. V., and Winn, R. A. (2007). Role of the Wnt signaling pathway and lung cancer. *J Thorac Oncol*, 2(10):889–892.

Teo, Y. Y. (2008). Common statistical issues in genome-wide association studies: a review on power, data quality control, genotype calling and population structure. *Curr Opin Lipidol*, 19(2):133–143.

Teo, Y. Y., Inouye, M., Small, K. S., Gwilliam, R., Deloukas, P., Kwiatkowski, D. P., and Clark, T. G. (2007). A genotype calling algorithm for the Illumina BeadArray platform. *Bioinformatics*, 23(20):2741–2746.

The Huntington's Disease Collaborative Research Group (1993). A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. *Cell*, 72(6):971–983.

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. *Annu Rev Public Health*, 31:21–36.

Thomas, D., Xie, R., and Gebregziabher, M. (2004). Two-stage sampling designs for gene association studies. *Genet Epidemiol*, 27(4):401–414.

Thomas, D. C. (1988). Models for exposure-time-response relationships with applications to cancer epidemiology. *Annu Rev Public Health*, 9:451–482.

Thomas, D. C. (2005). The need for a systematic approach to complex pathways in molecular epidemiology. *Cancer Epidem Biomar*, 14(3):557–559.

Thomas, D. C. (2006). Are we ready for genome-wide association studies? *Cancer Epidem Biomar*, 15(4):595–598.

Thomas, D. C. (2010c). Design and analysis issues in genome-wide association studies. In *Human Genome Epidemiology, 2nd Edition*. Khoury MJ, Oxford University Press.

Thomas, D. C., Casey, G., Conti, D. V., Haile, R. W., Lewinger, J. P. P., and Stram, D. O. (2009). Methodological issues in multistage genome-wide association studies. *Stat Sci*, 24(4):414–429.

Thomas, D. C., Haile, R. W., and Duggan, D. (2005). Recent developments in genomewide association scans: a workshop summary and review. *Am J Hum Genet*, 77(3):337–345.

Thomas, D. C., Lewinger, J. P., Murcray, C. E., and Gauderman, W. J. (2012). Invited commentary: Ge-whiz! ratcheting gene-environment studies up to the whole genome and the whole exposome. *Am J Epidemiol*, 175(3):203–7; discussion 208–9.

Thomas, D. C. and Witte, J. S. (2002). Point: population stratification: a problem for case-control studies of candidate-gene associations? *Cancer Epidem Biomar*, 11(6):505–512.

Thomson, W., Barton, A., Ke, X., Eyre, S., Hinks, A., Bowes, J., Donn, R., Symmons, D., Hider, S., Bruce, I. N., Consortium, W. T. C. C., Wilson, A. G., Marinou, I., Morgan, A., Emery, P., Consortium, Y. E. A. R., Carter, A., Steer, S., Hocking, L., Reid, D. M., Wordsworth, P., Harrison, P., Strachan, D., and Worthington, J. (2007). Rheumatoid arthritis association at 6q23. *Nat Genet*, 39(12):1431–1433.

Thorgeirsson, T. E., Geller, F., Sulem, P., Rafnar, T., Wiste, A., Magnusson, K. P., Manolescu, A., Thorleifsson, G., Stefansson, H., Ingason, A., Stacey, S. N., Bergthorsson, J. T., Thorlacius, S., Gudmundsson, J., Jonsson, T., Jakobsdottir, M., Saemundsdottir, J., Olafsdottir, O., Gudmundsson, L. J., Bjornsdottir, G., Kristjansson, K., Skuladottir, H., Isaksson, H. J., Gudbjartsson, T., Jones, G. T., Mueller, T., Gottsäter, A., Flex, A., Aben, K. K. H., de Vegt, F., Mulders, P. F. A., Isla, D., Vidal, M. J., Asin, L., Saez, B., Murillo, L., Blondal, T., Kolbeinsson, H., Stefansson, J. G., Hansdottir, I., Runarsdottir, V., Pola, R., Lindblad, B., van Rij, A. M., Dieplinger, B., Haltmayer, M., Mayordomo, J. I., Kiemeney, L. A., Matthiasson, S. E., Oskarsson, H., Tyrfingsson, T., Gudbjartsson, D. F., Gulcher, J. R., Jonsson, S., Thorsteinsdottir, U., Kong, A., and Stefansson, K. (2008). A variant associated with nicotine dependence, lung cancer and peripheral arterial disease. *Nature*, 452(7187):638–642.

Thorisson, G. A. and Stein, L. D. (2003). The SNP Consortium website: past, present and future. *Nucleic Acids Res*, 31(1):124–127.

Tian, C., Gregersen, P. K., and Seldin, M. F. (2008). Accounting for ancestry: population substructure and genome-wide association studies. *Hum Mol Genet*, 17(R2):R143–R150.

Tian, L., Greenberg, S. A., Kong, S. W., Altschuler, J., Kohane, I. S., and Park, P. J. (2005). Discovering statistically significant pathways in expression profiling studies. *Proc Natl Acad Sci U S A*, 102(38):13544–13549.

Tintle, N., Lantieri, F., Lebrec, J., Sohns, M., Ballard, D., and Bickeböller, H. (2009a). Inclusion of a priori information in genome-wide association analysis. *Genet Epidemiol*, 33 Suppl 1:S74–S80.

Tintle, N. L., Best, A. A., DeJongh, M., Bruggen, D. V., Heffron, F., Porwollik, S., and Taylor, R. C. (2008). Gene set analyses for interpreting microarray experiments on prokaryotic organisms. *BMC Bioinformatics*, 9:469.

Tintle, N. L., Borchers, B., Brown, M., and Bekmetjev, A. (2009b). Comparing gene set analysis methods on single-nucleotide polymorphism data from genetic analysis workshop 16. *BMC Proc*, 3 Suppl 7:S96.

Tiwari, H. K., Barnholtz-Sloan, J., Wineinger, N., Padilla, M. A., Vaughan, L. K., and Allison, D. B. (2008). Review and evaluation of methods correcting for population stratification with a focus on underlying statistical principles. *Hum Hered*, 66(2):67–86.

Todd, J. A. (2006). Statistical false positive or true disease pathway? *Nat Genet*, 38(7):731–733.

Tokuhata, G. K. and Lilienfeld, A. M. (1963). Familial aggregation of lung cancer in humans. *J Natl Cancer Inst*, 30:289–312.

Tracy, C. A. and Widom, H. (1992). Level-spacing distributions and the airy kernel. *Commun Math Phys*, 159(1):35.

Trikalinos, T. A., Salanti, G., Zintzaras, E., and Ioannidis, J. P. (2008). Meta-analysis methods. In Rao, D. C. and Gu, C. C., editors, *Genetic Dissection of Complex Traits*, volume 60 of *Advances in Genetics*, pages 311 – 334. Academic Press.

Troendle, J. F. (1996). A permutational step-up method of testing multiple outcomes. *Biometrics*, 52(3):846–859.

Truong, T., Hung, R. J., Amos, C. I., Wu, X., Bickeböller, H., Rosenberger, A., Sauter, W., Illig, T., Wichmann, H.-E., Risch, A., Dienemann, H., Kaaks, R., Yang, P., Jiang, R., Wiencke, J. K., Wrensch, M., Hansen, H., Kelsey, K. T., Matsuo, K., Tajima, K., Schwartz, A. G., Wenzlaff, A., Seow, A., Ying, C., Staratschek-Jox, A., Nürnberg, P., Stoelben, E., Wolf, J., Lazarus, P., Muscat, J. E., Gallagher, C. J., Zienolddiny, S., Haugen, A., van der Heijden, H. F. M., Kiemeney, L. A., Isla, D., Mayordomo, J. I., Rafnar, T., Stefansson, K., Zhang, Z.-F., Chang, S.-C., Kim, J. H., Hong, Y.-C., Duell, E. J., Andrew, A. S., Lejbkowicz, F., Rennert, G., Müller, H., Brenner, H., Marchand, L. L., Benhamou, S., Bouchardy, C., Teare, M. D., Xue, X., McLaughlin, J., Liu, G., McKay, J. D., Brennan, P., and Spitz, M. R. (2010). Replication of lung cancer susceptibility loci at chromosomes 15q25, 5p15, and 6p21: a pooled analysis from the international lung cancer consortium. *J Natl Cancer Inst*, 102(13):959–971.

Tseng, R.-C., Hsieh, F.-J., Shih, C.-M., Hsu, H.-S., Chen, C.-Y., and Wang, Y.-C. (2009). Lung cancer susceptibility and prognosis associated with polymorphisms in the nonhomologous end-joining pathway genes: a multiple genotype-phenotype study. *Cancer*, 115(13):2939–2948.

Tukey, J. W. (1958). Bias and confidence in not-quite large samples. *Ann Math Stat*, 29(2):614.

Tusher, V. G., Tibshirani, R., and Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci U S A*, 98(9):5116–5121.

Uematsu, K., He, B., You, L., Xu, Z., McCormick, F., and Jablons, D. M. (2003). Activation of the Wnt pathway in non small cell lung cancer: evidence of dishevelled overexpression. *Oncogene*, 22(46):7218–7221.

U.S. Department of Energy Genome Programs (2011). Human genome project. Available at `http://genomics.energy.gov` [Accessed 12 October 2011].

U.S. National Library of Medicine (2011). Handbook: Help me understand genetics. Available at http://ghr.nlm.nih.gov/handbook [Accessed 14 May 2011].

Veyrieras, J.-B., Kudaravalli, S., Kim, S. Y., Dermitzakis, E. T., Gilad, Y., Stephens, M., and Pritchard, J. K. (2008). High-resolution mapping of expression-QTLs yields insight into human gene regulation. *PLoS Genet*, 4(10):e1000214.

Vineis, P. (2007). Methodological approaches to gene-environment interactions in occupational epidemiology [electronic article]. *Occup Environ Med*, 64(e3).

Vogelstein, B., Lane, D., and Levine, A. J. (2000). Surfing the p53 network. *Nature*, 408(6810):307–310.

Volk, H., Lewinger, J., and Thomas, D. (2007). Two-stage strategies for detecting gene-environment interactions in a genome-wide association study (IGES abstract 152). *Genet Epidemiol*, 31(6):649.

Vousden, K. H. and Lu, X. (2002). Live or let die: the cell's response to p53. *Nat Rev Cancer*, 2(8):594–604.

Wacholder, S., Chanock, S., Garcia-Closas, M., Ghormli, L. E., and Rothman, N. (2004). Assessing the probability that a positive report is false: an approach for molecular epidemiology studies. *J Natl Cancer Inst*, 96(6):434–442.

Wacholder, S., Rothman, N., and Caporaso, N. (2000). Population stratification in epidemiologic studies of common genetic variants and cancer: quantification of bias. *J Natl Cancer Inst*, 92(14):1151–1158.

Wakefield, J. (2007). A Bayesian measure of the probability of false discovery in genetic epidemiology studies. *Am J Hum Genet*, 81(2):208–227.

Wakefield, J. (2008). Reporting and interpretation in genome-wide association studies. *Int J Epidemiol*, 37(3):641–653.

Wald, A. (1943). *Tests of statistical hypotheses concerning several parameters when the number of observations is large*. American Mathematical Society.

Walter, S. D. and Holford, T. R. (1978). Additive, multiplicative, and other models for disease risks. *Am J Epidemiol*, 108(5):341–346.

Wang, H., Thomas, D. C., Pe'er, I., and Stram, D. O. (2006). Optimal two-stage genotyping designs for genome-wide association scans. *Genet Epidemiol*, 30(4):356–368.

Wang, K. (2008). GenGen Website. Available at http://www.openbioinformatics.org/gengen/ [Accessed June 2008].

Wang, K. and Abbott, D. (2008). A principal components regression approach to multilocus genetic association studies. *Genet Epidemiol*, 32(2):108–118.

Wang, K., Li, M., and Bucan, M. (2007). Pathway-based approaches for analysis of genomewide association studies. *Am J Hum Genet*, 81(6).

Wang, K., Li, M., and Hakonarson, H. (2010). Analysing biological pathways in genome-wide association studies. *Nat Rev Genet*, 11(12):843–854.

Wang, L., Jia, P., Wolfinger, R. D., Chen, X., Grayson, B. L., Aune, T. M., and Zhao, Z. (2011). An efficient hierarchical generalized linear mixed model for pathway analysis of genome-wide association studies. *Bioinformatics*, 27(5):686–692.

Wang, W. Y. S., Barratt, B. J., Clayton, D. G., and Todd, J. A. (2005). Genome-wide association studies: theoretical and practical concerns. *Nat Rev Genet*, 6(2):109–118.

Wang, Y., Broderick, P., Webb, E., Wu, X., Vijayakrishnan, J., Matakidou, A., Qureshi, M., Dong, Q., Gu, X., Chen, W. V., Spitz, M. R., Eisen, T., Amos, C. I., and Houlston, R. S. (2008). Common 5p15.33 and 6p21.33 variants influence lung cancer risk. *Nat Genet*, 40(12):1407–1409.

Watkins, W. S., Zenger, R., O'Brien, E., Nyman, D., Eriksson, A. W., Renlund, M., and Jorde, L. B. (1994). Linkage disequilibrium patterns vary with chromosomal location: a case study from the von Willebrand factor region. *Am J Hum Genet*, 55(2):348–355.

Weale, M. E. (2010). Quality control for genome-wide association studies. *Methods Mol Biol*, 628(Genetic Variation):341–372.

Weiss, K. M. and Terwilliger, J. D. (2000). How many diseases does it take to map a gene with SNPs? *Nat Genet*, 26(2):151–157.

Wellcome Trust Case Control Consortium (WTCCC) (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447(7145):661–678.

Werner, J. (1984). *Biomathematik und Medizinische Statistik.: Eine praktische Anleitung für Studierende, Doktoranden, Ärzte und Biologen.* Urban & Schwarzenberg.

Westfall, P. and Young, S. (1993). *Resampling-based multiple testing: examples and methods for p-value adjustment.* Wiley series in probability and mathematical statistics: Applied probability and statistics. Wiley.

Whittemore, A. (2007). A Bayesian false discovery rate for multiple testing. *J Appl Stat*, 34(1):1–9.

Wichmann, H. E., Gieger, C., and Illig, T. (2005). KORA-gen - Resource for population genetics, controls and a broad spectrum of disease phenotypes. *Das Gesundheitswesen*, 67 Sonderh:26–30.

Working Group on Replication in Association Studies, N. C. I.-N. H. G. R. I., Chanock, S. J., Manolio, T., Boehnke, M., Boerwinkle, E., Hunter, D. J., Thomas, G., Hirschhorn, J. N., Abecasis, G., Altshuler, D., Bailey-Wilson, J. E., Brooks, L. D., Cardon, L. R., Daly, M., Donnelly, P., Fraumeni, J. F., Freimer, N. B., Gerhard, D. S., Gunter, C., Guttmacher, A. E., Guyer, M. S., Harris, E. L., Hoh, J., Hoover, R., Kong, C. A., Merikangas, K. R., Morton, C. C., Palmer, L. J., Phimister, E. G.,

Rice, J. P., Roberts, J., Rotimi, C., Tucker, M. A., Vogan, K. J., Wacholder, S., Wijsman, E. M., Winn, D. M., and Collins, F. S. (2007). Replicating genotype-phenotype associations. *Nature*, 447(7145):655–660.

Wu, A. H., Fontham, E. T., Reynolds, P., Greenberg, R. S., Buffler, P., Liff, J., Boyd, P., Henderson, B. E., and Correa, P. (1995). Previous lung disease and risk of lung cancer among lifetime nonsmoking women in the United States. *Am J Epidemiol*, 141(11):1023–1032.

Yamanaka, A., Hirai, T., Ohtake, Y., and Kitagawa, M. (1997). Lung cancer associated with Werner's syndrome: a case report and review of the literature. *Jpn J Clin Oncol*, 27(6):415–418.

Yang, J., Benyamin, B., McEvoy, B. P., Gordon, S., Henders, A. K., Nyholt, D. R., Madden, P. A., Heath, A. C., Martin, N. G., Montgomery, G. W., Goddard, M. E., and Visscher, P. M. (2010). Common SNPs explain a large proportion of the heritability for human height. *Nat Genet*, 42(7):565–569.

Yang, M., Guo, H., Wu, C., He, Y., Yu, D., Zhou, L., Wang, F., Xu, J., Tan, W., Wang, G., Shen, B., Yuan, J., Wu, T., and Lin, D. (2009). Functional FEN1 polymorphisms are associated with DNA damage levels and lung cancer risk. *Hum Mutat*, 30(9):1320–1328.

Yang, P., Schwartz, A. G., McAllister, A. E., Aston, C. E., and Swanson, G. M. (1997). Genetic analysis of families with nonsmoking lung cancer probands. *Genet Epidemiol*, 14(2):181–197.

Yang, Q., Cui, J., Chazaro, I., Cupples, L. A., and Demissie, S. (2005). Power and type I error rate of false discovery rate approaches in genome-wide association studies. *BMC Genet*, 6(Suppl 1):S134.

Yehuda, S., Rabinovitz, S., and Mostofsky, D. I. (2005). Mixture of essential fatty acids lowers test anxiety. *Nutr Neurosci*, 8(4):265–267.

Yu, J., Pressoir, G., Briggs, W. H., Bi, I. V., Yamasaki, M., Doebley, J. F., McMullen, M. D., Gaut, B. S., Nielsen, D. M., Holland, J. B., Kresovich, S., and Buckler, E. S. (2006). A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat Genet*, 38(2):203–208.

Yu, K., Chatterjee, N., Wheeler, W., Li, Q., Wang, S., Rothman, N., and Wacholder, S. (2007). Flexible design for following up positive findings. *Am J Hum Genet*, 81(3):540–551.

Yu, K., Li, Q., Bergen, A. W., Pfeiffer, R. M., Rosenberg, P. S., Caporaso, N., Kraft, P., and Chatterjee, N. (2009). Pathway analysis by adaptive combination of p-values. *Genet Epidemiol*, 33(8):700–709.

Zavattari, P., Deidda, E., Whalen, M., Lampis, R., Mulargia, A., Loddo, M., Eaves, I., Mastio, G., Todd, J. A., and Cucca, F. (2000). Major factors influencing linkage disequilibrium by analysis of different chromosome regions in distinct populations:

demography, chromosome recombination frequency and selection. *Hum Mol Genet*, 9(20):2947–2957.

Zaykin, D. V., Zhivotovsky, L. A., Westfall, P. H., and Weir, B. S. (2002). Truncated product method for combining p-values. *Genet Epidemiol*, 22(2):170–185.

Zhang, J. and Gentleman, R. (2011). *KEGGSOAP: Client-side SOAP access KEGG*. R package version 1.26.1.

Zhernakova, A., Alizadeh, B. Z., Bevova, M., van Leeuwen, M. A., Coenen, M. J. H., Franke, B., Franke, L., Posthumus, M. D., van Heel, D. A., van der Steege, G., Radstake, T. R. D. J., Barrera, P., Roep, B. O., Koeleman, B. P. C., and Wijmenga, C. (2007). Novel association in chromosome 4q27 region with rheumatoid arthritis and confirmation of type 1 diabetes point to a general risk locus for autoimmune diseases. *Am J Hum Genet*, 81(6):1284–1288.

Ziegler, A., König, I. R., and Thompson, J. R. (2008). Biostatistical aspects of genome-wide association studies. *Biom J*, 50(1):8–28.

Ziegler, A. and König, I. (2006). *A statistical approach to genetic epidemiology: concepts and applications*. Wiley-VCH.

Zondervan, K. T. (2010). Genetic association study design. In Zeggini, E. and Morris, A., editors, *Analysis of Complex Disease Association Studies: A Practical Guide*, volume 60 of *Academic Press*, pages 25 – 48. Elsevier Science.

Zondervan, K. T. and Cardon, L. R. (2007). Designing candidate gene and genome-wide case-control association studies. *Nat Protoc*, 2(10):2492–2501.