Kernel Methods for Genes and Networks to Study Genome-Wide Associations of Lung Cancer and Rheumatoid Arthritis

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

vorgelegt von
Saskia Freytag
aus München

Göttingen, 2014

Supervisor:	Heike Bickeböller Institut für Genetische Epidemiologie Universitätsmedizin Georg-August-Universität Göttingen
Second Thesis Committee Member:	Martin Schlather Lehrstuhl für Mathematische Statistik, Stochastik und ihre Anwendungen Fakultät für Wirtschaftsinformatik und Wirtschaftsmathematik Universität Mannheim
Third Thesis Committee Member:	Thomas Kneib Professur für Statistik Wirtschaftswissenschaftliche Fakultät Georg-August-Universität Göttingen
Day of Disputation:	08.01.2014

Affidavit

Here I declare that my doctoral thesis entitled "Kernel Methods for Genes and Networks to Study Genome-Wide Associations of Lung Cancer and Rheumatoid Arthritis" has been written independently with no other sources and aids than quoted.

Saskia Freytag

Göttingen, Feburary 2014

Acknowledgement

On the road to completing this thesis I have been supported by numerous people. Here, I would like to highlight some who have helped me along this road and to express my warmest regards for them.

I would like to thank my supervisor Professor Heike Bickeböller for her constant support and encouragement. She always managed to provide motivation at times of need and never failed to believe in me. Her invaluable advice on science as well as her explanation of the academic world has guided me along the way. Moreover, I am indebted to her for helping me find a post-doc position in Australia. I also wish to express my deepest gratitude to my second supervisor Professor Martin Schlather. I was fortunate to work with him, as his ideas often provided the solution to my long-pondered scientific problems. His meticulous corrections of my drafts nearly cost me my sanity, but ultimately resulted in substantial improvements. He was a superb mentor. I would also like to thank Professor Thomas Kneib who in all situations had kind words of encouragement. He always listened to my problems and tried hard to find solutions. He was available whenever I requested his much sought-after time. I wish to express my thanks to all three for extensively revising and correcting several drafts of this thesis.

My research was made possible through the financial support I received from the Deutsche Forschungsgemeinschaft (DFG) within the research training group "Scaling problems in statistics" (RTG 1644). Their financial support was a crucial factor in the completion of this thesis. The data for my research were generously made available by Peter Gregerson of the National Institute of Health [AR44422] and the TRICL consortium [U19CA148127].

I am also indebted to my coauthors for joint work on my papers. Above all, I would like to mention Juliane Manitz. We stretched each other's patience several times as we drafted the manuscript for our joint paper. However, her critical approach, fierce interest and will to work long hours ultimately ensured its speedy completion. During my three years in Göttingen, Juliane and I have shared countless intellectually stimulating discussions that often pointed my work in the right direction. She was always willing to listen to my problems and helped me to get through several academic disappointments. I am also grateful to Ngoc-Thuy Ha. She started as my student helper, but along the road has become a friend, a co-author and a colleague. Even as she drafted her first manuscript, she took time to answer my at times bizarre questions on formatting with Latex. She was also responsible for transforming my R-scripts into a user-friendly R-package. Furthermore, I was lucky to collaborate with Professor Christopher I. Amos and Professor Rayjean Hung, who not only provided critical input but resources to test this R-package.

My colleagues at the Institute of Genetic Epidemiology, RTG 1644 and Chair of Statistics proved to be excellent companions during this endeavor. The colleagues at the Institute of Genetic Epidemiology certainly made Göttingen a warm and inviting place to work. Andrew Entwistle of the Institute of Genetic Epidemiology deserves a special mention, as he kindly proofread the drafts of several papers. I would like to thank all my colleagues at the Institute for their help and support. My fellow PhD students at the RTG 1644 were always open to beer and moaning. Barbara Strauss and Dörte Dede of the RTG 1644 deserve a special mention, as they made sure that I was never unduly troubled with administrative tasks. Most importantly though, my colleagues at the Chair of Statistics adopted me as one of them. This way I spent each lunch time surrounded by friendly faces and laughter. I will never forget our truly wonderful excursions.

Whenever I abandoned all hope, I took my despair to my mum. She had to bear the brunt of my tantrums and did so gracefully. During all uphill struggles, I was always able to draw inspiration from my dad and his remarkable love of science. He encouraged me to pursue a PhD and supported me unrelentingly every step along the way. In the end, their love for me made the difference.

Finally, I would like to dedicate this thesis to the many anonymous individuals who gave their genetic data. Without their contribution, this work would simply never have been possible.

Abstract

The search for genetic causes of common complex diseases has been revolutionized by the ability to genotype exceptionally large numbers of single nucleotide polymorphisms (SNPs) in hundreds of individuals at an affordable cost. Statistical analysis of the data generated in hundreds of such genome-wide association studies has been able to identify genetic risk variants with differing degrees of success. Overall, these genetic risk variants account only for a fraction of the observed genetic heritability. Reasons suggested for this shortcoming range from the identification of statistical problems with conventional analysis tools to the failure to model the complexity of the human organism properly. One proposition to uncover a portion of the 'missing heritability' is the analysis of biologically meaningful SNP sets. Methods based on SNP sets are typically powerful and aid the interpretation of results through the incorporation of biological knowledge.

A popular approach in the identification of associations between an investigated disease and SNP sets lies in kernel methods, in particular the logistic kernel machine test. Such methods formulate the estimation problem in a reproducing kernel Hilbert space of functions, which is uniquely defined by a positive semi-definite kernel. This has the benefit of facilitating the construction and estimation of a wide variety of genetic effect models. However, this immense flexibility can also prove problematic. The choice of kernel most suitable for a particular problem is seldom obvious and the choice made seriously affects the ability of the kernel method to discover genuine associations.

One of the main objectives of this thesis is the development of appropriate kernels for the analysis of SNP sets, such as genes or pathways. Here, a pathway is defined as a network of interacting genes responsible for achieving a specific cell function or regulation. In this thesis, I introduce a kernel that corrects for bias incurred through differently-sized pathways in terms of the number of SNPs or genes. This kernel also reflects the basic architecture of a pathway. This concept is expanded by constructing another kernel that integrates specific gene-gene regulations. Through simulation studies and implementation of real data on rheumatoid arthritis and lung cancer, I demonstrate both robustness as well as practical usefulness of the logistic kernel machine test with the two kernels introduced above.

Another main objective of this thesis is to compare kernel methods with other approaches in the analysis of pathways or genes. This includes comparing the performance of various multi-marker methods for ranking genes according to their strength of association with an investigated disease. In many genetic scenarios, it is possible to show that the performance of kernel methods is superior. In addition, this thesis includes a comparative study of chips currently used to genotype SNPs in the human genome. The chips are assessed with regard to their coverage of the genome, price, and efficiency.

Zusammenfassung

Die Möglichkeit Millionen von Single Nucleotide Polymorphisms (SNPs) kostengünstig in hunderten Individuen zu genotypisieren hat die Suche nach den genetischen Ursachen komplexer Krankheiten revolutioniert. Die Analyse der Daten solcher genomweiter Assoziationsstudien konnte mit sehr unterschiedlichem Erfolg Risikovarianten identifizierten. Die so identifizierten Risikovarianten konnten jedoch nur Bruchteile der geschätzten genetischen Erblichkeit erklären. Erklärungen für den bislang eingeschränkten Erfolg reichen von statistischen Probleme der konventionellen Analysemethoden zu Versäumnissen die Komplexität des menschlichen Organismus adäquat zu modellieren. Ein Vorschlag um Teile der unentdeckten Erblichkeit zu finden ist die Analyse von biologisch sinnvollen SNP-Gruppen. Methoden, die auf SNP-Gruppen basieren, sind typischerweise mächtig und helfen bei der Interpretation von Ergebnissen durch den Einbezug biologischen Wissens.

Beliebte Verfahren zur Bestimmung von Assoziationen zwischen der untersuchten Krankheit und SNP-Gruppen sind Kernmethoden, insbesondere der Kernel Logistic Machine Test. In solchen Verfahren wird das Schätzproblem in einem Reproducing-Kernel-Hilbert-Raum, welcher durch die Wahl eines positiv semidefiniten Kernes vollständig und eindeutig bestimmt ist, formuliert. Dies hat den Vorteil ,dass es die Konstruktion und Schätzung einer Vielzahl an genetischen Modellen erlaubt; allerdings kann diese immense Flexibilität auch problematisch sein. Die Wahl eines Kernes ist nur selten offensichtlich, aber wirkt sich entscheidend auf die Fähigkeit der Methode wahre genetische Assoziationen zu entdecken aus.

Diese Arbeit hat das Ziel angemessene Kerne für die Analyse von SNP-Gruppen, wie zum Beispiel Gene oder biologische Reaktionspfade, zu konstruieren. Ein Reaktionspfad ist hierbei definiert als ein Netzwerk interagierender Gene, die zusammen die spezifische Funktion oder Regulierung einer Zelle erreichen. Diese Dissertation führt einen Kern mit Korrektur für die verschiedenen Reaktionspfadsgrössen, gemessen an der Anzahl der SNPs oder Gene, ein. Dieser Kern reflektiert gleichzeitig die Grundarchitektur eines Reaktionspfades. Mit der Konstruktion eines weiteren Kernes, der spezifische Gen-Gen Interaktionen integriert, wird dieses Konzept weiterentwickelt. Die Robustheit und die praktische Relevanz dieser Entwicklung wird mit Hilfe einer Simulationsstudie sowie der Anwendung der Methode auf realen Datensätze zu Rheumatoid Arthritis und Lungenkrebs demonstriert.

Ein weiteres Ziel dieser Arbeit ist der Vergleich von Kernmethoden zu anderen gängigen Verfahren für die Analyse von Genen oder Reaktionspfaden. Dies beinhaltet den Vergleich verschiedener Verfahren für die Analyse mehrerer SNPs. Die Verfahren werden anhand ihres Vermögens die Stärke von Gen-Krankheitsassoziationen korrekt Ränge zu zuordnen verglichen. Für viele genetische Szenarien, sind Kernmethoden überlegen. Darüberhinaus enthält diese Dissertation eine komparative Studie von Chips für die Genotypisierung von SNPs im menschlichen Genom. Diese Chips wurden mit Augenmerk auf die Abdeckung des Genoms, Preis und Effizienz bewertet.

Contents

1 Introduction

1.1 The Current State of Genome-Wide Association Studies

Since decoding the human genome sequence in 2003, genetic epidemiologists interested in the genetic causes of common diseases shifted their focus from mutations clustering in families to genetic variation commonly occurring in the entire population. Family-based approaches, such as linkage analysis, had mostly failed to establish causal genetic factors for diseases not characterized by the Mendelian properties of monogenic diseases. The failure of family-based approaches with regards to common diseases can mainly be attributed to the fact that such diseases evidently do not solely result from variations in one genetic region. Indeed, most diseases are believed "to have complex architectures [...], for which the phenotype is determined by the sum total of, and/or interactions between, multiple genetic and environmental factors" (Hirschhorn and Daly, 2005). In case of common diseases, the term "common disease common variant" (CDCV) is used to refer to this hypothesis and several studies have provided evidence for its plausibility (Koeleman, Al-Ali, van der Laan, and Asselbergs, 2013). Under this hypothesis, family-based approaches would have required impractically large sample sizes in order to detect genetic risk factors (Risch and Merikangas, 1996; Risch, 2000). Instead, with advances in genotyping technologies and the cataloging of millions of single nucleotide polymorphisms (SNPs) genome-wide case-control studies became affordable and increasingly popular for the detection of genetic risk variants for common complex diseases.

A genome-wide case-control study, which is one type of genome-wide association study (GWAS), surveys most of the genome for SNPs with a different frequency of genotypes between cases and controls. SNPs are the genetic marker of choice since they are the most abundant type of genetic variation, feasible for high throughput and at the same time provide comprehensive coverage of the genome. By definition, SNP refers to any bi-allelic genetic variation found in more than 1% of a population. The dbSNP database (`ncbi.nlm.nih.gov/SNP`), which collects genetic markers, included information on more than 18 million validated SNPs in July 2013. Using this catalogue and information on linkage disequilibrium (LD) structures, researchers are able to select SNPs that most effectively and efficiently capture the majority of the variation in the genome. In particular, haplotype blocks can be represented by just very few SNPs since such genetic regions are rarely broken by recombination.

Thus, SNPs located in such regions can be called with the help of simple correlation measures. Today, available genome-wide genotype arrays range from 500,000 SNPs to 4,3 million SNPs and accordingly offer different coverage of the genome (Ha, Freytag, and Bickeböller, 2014).

The knowledge gain due to GWASs has received mixed evaluations (Visscher and Montgomery, 2009). Certainly, before the beginning of the initial wave of large-scale GWAS such studies were believed to greatly advance our understanding of the genetic basis underlying common diseases (Hirschhorn and Daly, 2005; Cardon and Bell, 2001). Indeed, for traits, like inflammatory bowel disease and breast cancer, it was possible to replicate and biologically verify a tremendous amount of genetic risk factors. In 2011, there were over 2,000 genetic markers that had been shown to be robustly associated with one or more complex traits (Visscher, Brown, McCarthy, and Yang, 2012). In some cases, the discoveries resulting from the analysis of GWASs have even been translated into medical treatments. For example, psoriasis medication, which neutralizes IL-17 involved in the regulation of inflammatory circuits, was developed after an association between IL-17 and psoriasis had been determined using GWASs (Krueger, Fretzin, Suárez-Fariñas, Haslett, Phipps, Cameron, McColm, Katcherian, Cueto, White, Banerjee, and Hoffman, 2012). Nevertheless, the total amount of heritability explained by the newly discovered genetic risk factors has remained below the often unreasonably high expectations of some scientists. In general, established genetic risk factors "fail to explain the vast majority of genetic heritability for any human disease, either individually or collectively" (Manolio, Collins, Cox, Goldstein, Hindorff, Hunter, McCarthy, Ramos, Cardon, Chakravarti, Cho, Guttmacher, Kong, Kruglyak, Mardis, Rotimi, Slatkin, Valle, Whittemore, Boehnke, Clark, Eichler, Gibson, Haines, Mackay, McCarroll, and Visscher, 2009). The bulk of identified markers was of no immediate clinical utility nor was their biological relevance for the investigated disease apparent.

The reasons for the shortcomings of GWASs and "the mystery of the missing heritability" (Manolio, Collins, Cox, Goldstein, Hindorff, Hunter, McCarthy, Ramos, Cardon, Chakravarti, Cho, Guttmacher, Kong, Kruglyak, Mardis, Rotimi, Slatkin, Valle, Whittemore, Boehnke, Clark, Eichler, Gibson, Haines, Mackay, McCarroll, and Visscher, 2009) have been hotly debated in the genetic research community. Several convincing opinions concerning the source of the missing heritability and accompanying search strategies have been put forward (Eichler, Flint, Gibson, Kong, Leal, Moore, and Nadeau, 2010). Genome-wide genotype arrays do only contain some rare variants (Manolio, Collins, Cox, Goldstein, Hindorff, Hunter, McCarthy, Ramos, Cardon, Chakravarti, Cho, Guttmacher, Kong, Kruglyak, Mardis, Rotimi, Slatkin, Valle, Whittemore, Boehnke, Clark, Eichler, Gibson, Haines, Mackay, McCarroll, and Visscher, 2009) and often fail to adequately cover copy number variations (Mefford and Eichler, 2009). Both are collectively frequent variations in the genome and have so far been inadequately examined for their role in the development and progression of common diseases. However, new sequencing technology enabling the analysis of exons or even the whole genome should change this in the coming years. It has also been suggested that the failure of analyses to consider a transgenerational genetic effects, such as impriniting, is a possible source of the missing heritability (Kong et al., 2009). In this case, phased genotype data or sequence data might lead to the discovery of a substantial proportion of the missing heritability. Interaction with environmental factors has further been named as a possible explanation and their careful inclusion in analyses of GWASs has been called for (Eichler, Flint, Gibson, Kong, Leal, Moore, and Nadeau, 2010). Yet another explanation might lie in the architecture of genetic effects, in particular highly complex interconnected networks of genes responsible for the manifestation of diseases (Zuk, Hechter, Sunyaev, and Lander, 2012). Solving the mystery of the missing heritability would accordingly require the application of novel statistical methods that incorporate a systems biological perspective. Finally, missing heritability could be the consequence of low statistical power that obstructs the success of most GWASs (Gibson, 2010). Conventional individual testing for association with each SNP leads to immense problems with multiple testing. Already in case of small genome-wide arrays, the application of appropriate corrections for multiple testing leads to high false-negative rates. Here, considering the joint effect of multiple genetic markers on the risk of being affected by the disease could help to increase statistical power.

The analysis of biologically meaningful sets of SNPs has been proposed as a new methodological approach to identify further genetic risk factors and therefore explain a part of the missing heritability. By simultaneously considering biologically related SNPs this approach provides a substantial boost of power. On the one hand, this is due to the potential to jointly detect moderate to low effects of multiple SNPs (Fridley and Patch, 2011) and the subsequent reduction of the multiple testing burden. On the other hand, simultaneous analysis allows researchers to consider complex genetic architectures. Moreover, the reintroduction of biological concepts through the use of biologically meaningful sets of SNPs may aid researchers in the interpretation and utilization of their results (Wang, Li, and Hakonarson, 2010). One of the most frequently investigated SNP sets is the gene. Genes constitute a particularly attractive genetic unit, since mutations in such regions are known to directly impact the functionality of the human organism. Furthermore, genes located on the human genome have been well annotated in projects such as ENCODE (Rosenbloom, Dreszer, Pheasant, Barber, Meyer, Pohl, Raney, Wang, Hinrichs, Zweig, Fujita, Learned, Rhead, Smith, Kuhn, Karolchik, Haussler, and Kent, 2009) and their locations as well as other properties can be accessed via several online databases (Wheeler, Barrett, Benson, Bryant, Canese, Chetvernin, Church, DiCuccio, Edgar, Federhen, Geer, Kapustin, Khovayko, Landsman, Lipman, Madden, Maglott, Ostell, Miller, Pruitt, Schuler, Sequeira, Sherry, Sirotkin, Souvorov, Starchenko, Tatusov, Tatusova, Wagner, and Yaschenko, 2007). Another unit frequently of interest is the pathway. The pathway represents a collection of interacting genes that coordinate to achieve a specific cell function or cell response as part of a biological process, such as metabolic processes (Cantor, Lange, and Sinsheimer, 2010). Genes belonging to the same pathway form complex networks that regulate the particular response or function. If genetic variations disrupt a sufficient fraction of the pathway, its ability to regulate might be severely damaged, which can in turn lead to the manifestation of a disease. Similarly to genes, pathways have been collected in numerous online databases (Viswanathan, Seto, Patil, Nudelman, and Sealfon, 2008), which often provide detail about their functionality within the human organism. However, these data bases often disagree, are ambiguous about the exact structures of the networks and miss information on considerable parts of the human organism.

There exist a broad range of multiple marker methods for the analysis of sets of SNPs and several more specialized approaches targeted towards detecting pathway-disease associations. One of the earliest pathway analysis methods was gene-set enrichment analysis by Wang, Li, and Bucan (2007). In this approach, genes in the pathway are assigned their strongest empirical p-value obtained from single marker tests with all SNPs located in the gene. Using a weighted Kolmogorov-Smirnov-like running sum statistics, each pathway is assessed in order to determine whether genes with small p-values are overrepresented in the investigated pathway. There have been several extensions of this method; most notably Yu, Li, Bergen, Pfeiffer, Rosenberg, Caporaso, Kraft, and Chatterjee (2009) introduced the adaptive rank truncated product statistic in order to evaluate the association of a gene. Another pathway analysis approach, the hierarchical Bayes prioritization, relies on the Bayesian framework to prioritize markers using prior biological information such as gene and pathway membership (Lewinger, Conti, Baurley, Triche, and Thomas, 2007). Both gene-set enrichment analysis and hierarchical Bayes prioritization depend upon individual testing with each SNP. Thus, they are characterized by some of the disadvantages burdening single marker testing. Namely, the effect of a SNP is estimated independently of all others. Kernel methods do not share this dependence on single marker testing. Instead, their flexible framework, which consists of a covariance structure reflecting the genetic similarity between all pairs of individuals, enables modeling of highly complex SNP-SNP interactions. Furthermore, kernel methods applied to GWASs have proven extremely powerful (Pan, 2009; Wu, Kraft, Epstein, Taylor, Chanock, Hunter, and Lin, 2010) and their superior performance compared to other pathway-based methods, in particular gene-set enrichment analysis and hierarchical Bayes prioritization, have been empirically established (Freytag, Bickeböller, Amos, Kneib, and Schlather, 2012).

1.2 Kernel Methods for the Analysis of Genome-Wide Association Studies

Kernel methods, which include nonparametric regression, smoothing splines and support vector machines, are particularly well suited to cope with the challenges connected to the analysis of GWAS: high dimensionality, non-linearity of genetic effects and heterogeneity of data types. Kernel methods transform the data into a set of points in a high dimensional space. The kernel, the name-sake of these methods, acts thereby as a function that enables operations in this new space without the need to explicitly compute the coordinates of the data in this particular space. This is often easier than the explicit computation and thus can prove highly advantageous. This feature also ensures that high-dimensional data can usually be modeled without encountering computational problems or extreme power losses. In this new space, the data can be modeled linearly using a variety of statistical models (Hofmann, Schölkopf, and Smola, 2008). These modeled relationships need not necessarily be linear in the original space. Even though the kernel is restricted to be positive semi-definite or positive definite, there is a wide spectrum of choice for the kernel. This flexibility can also be exploited to incorporate different types of data. However, this immense flexibility can also be problematic as it is difficult to know which kernel to apply for a specific research problem.

The kernel machine test, a member of the class of kernel methods, was originally developed for gene expression data (Liu, Ghosh, and Lin, 2008), it has been extended to enable testing associations between multiple genetic variants, i.e. genotypes, and a phenotype of interest (Wu, Kraft, Epstein, Taylor, Chanock, Hunter, and Lin, 2010). To this end, the kernel converts the genetic information into pairwise similarity between individuals. At this stage, essentially a simple linear test for correlation between the pairwise similarity of phenotypes and pairwise similarity of genotypes suffices (Wu, Maity, Lee, Simmons, Harmon, Lin, Engel, Molldrem, and Armistead, 2013). This methodology has several advantages compared to ordinary tests that rely on multiple regression. Firstly, kernel machine testing has the ability to detect the joint effect of several moderately associated genetic variants even in the absence of large main effects (Wu, Kraft, Epstein, Taylor, Chanock, Hunter, and Lin, 2010). Secondly, directionality information regarding the effect of the genetic variants is not required. Finally, its versatile kernel framework enables the modeling of complex non-linear relationships as long as these can be formulated using positive definite or semi-positive definite functions (Schölkopf and Smola, 2002).

The kernel machine test can be shown to be equivalent to a score test applied to an appropriately specified generalized linear mixed model (GLMM). In brief, the influence of the genetic variants is described by random effects with a correlation matrix generated by the kernel (Schaid, 2010a). Other variables that are informative in the context of the investigated phenotype, such as age and sex, can be entered as fixed effects into the model. The score test is then used to test whether or not the common variance of the random effects is zero. A non-zero variance would point to an influence of the tested genetic variants on the phenotype. Hence, it is sometimes referred to as a variance component test. Indeed, there exist other variance component tests for multiple genetic variants that can be shown to be closely related to kernel machine testing, like the weighted sum of squared score test by Pan (Pan, 2009; Pan, Han, and Shen, 2010).

The performance of the kernel machine test crucially depends on the choice of the kernel, because it subsumes the underlying model of genotype-phenotype relationship. However, knowing the precise nature of the relationship specified in the kernel is not always possible or computationally feasible, since this requires the back-transformation of the converted data to the original space. In case of the frequently applied linear kernel the underlying genetic model is known to be linear. Using this kernel, when the underlying genetic model includes interactions between genotypes, called epistasis, leads to reduced power (Wu, Kraft, Epstein, Taylor, Chanock, Hunter, and Lin, 2010). Furthermore, Freytag, Bickeböller, Amos, Kneib, and Schlather (2012) demonstrated that this kernel does not maintain type I error when applied in combination with the Satterthwaite approximation (for an explanation of this approximation see Section 2.2).

Successful applications of the kernel machine test have included GWAS with a diverse range of traits and disorders (Freytag, Bickeböller, Amos, Kneib, and Schlather, 2012; Wu, Kraft, Epstein, Taylor, Chanock, Hunter, and Lin, 2010; Shui, Mucci, Wilson, Kraft, Penney, Stampfer, and Giovannucci, 2012; Locke, Dooley, Tinker, Cheong, Feingold, Allen, Freeman, Torfs, Cua, Epstein, Wu, Lin, Capone, Sherman, and Bean, 2010). In the case of dichotomous traits, like disease status, the kernel machine test is also called logistic kernel machine test due to its use of the logistic link function (corresponding to logistic regression). Besides dichotomous and quantitative traits, the kernel machine test has also been extended to be applicable to data types such as censored survival data (Lin, Cai, Wu, Zhou, Liu, Christiani, and Lin, 2011), multivariate (Maity, Sullivan, and Tzeng, 2012) or family data (Malzahn, Friedrichs, Rosenberger, and Bickeböller, 2013). Furthermore, Larson and Schaid (2013) demonstrated that the kernel machine test is "a powerful approach toward detecting gene-gene interactions even in the absence of marginal effects." Most importantly, the kernel machine test has emerged as one of the more frequently encountered methods for the analysis of whole genome or exon sequence data (Wu, Lee, Cai, Li, Boehnke, and Lin, 2011; Nuytemans, Bademci, Inchausti, Dressen, Kinnamon, Mehta, Wang, Züchner, Beecham, Martin, et al., 2013; Neale, Kou, Liu, Maáyan, Samocha, Sabo, Lin, Stevens, Wang, Makarov, Polak, Yoon, Maguire, Crawford, Campbell, Geller, Valladares, Schafer, Liu, Zhao, Cai, Lihm, Dannenfelser, Jabado, Peralta, Nagaswamy, Muzny, Reid, Newsham, Wu, Lewis, Han, Voight, Lim, Rossin, Kirby, Flannick, Fromer, Shakir, Fennell, Garimella, Banks, Poplin, Gabriel, DePristo, Wimbish, Boone, Levy, Betancur, Sunyaev, Boerwinkle, Buxbaum, Cook Jr, Devlin, Gibbs, Roeder, Schellenberg, Sutcliffe, and Daly, 2012). This is because it is able to handle rare variants effectively and efficiently with regards to computation time.

This PhD thesis is concerned with the performance and methodological advancement of kernel methods for the analysis of genome-wide association data using genes or pathways. In particular, the choice or developmemt of an appropriate kernel for the analysis of genome-wide association data integrating pathway information. The findings presented are not just of statistical and methodological interest, but have implications for medicine as well as genetics.

Due to its cumulative nature, the thesis is organized in the following manner: Chapter 2 provides a brief technical description of the basic principles behind kernel methods and establishes the equivalence of the logistic kernel machine model and the logistic mixed model. It further describes the construction of the test statistic as well as its asymptotic distribution, for which an appropriate approximation is introduced. The research about the performance, suggested improvements and extensions of kernel methods for GWAS analyses that have been presented in three different research articles. These articles have all been either published or submitted to international peer-reviewed journals (Freytag, Bickeböller, Amos, Kneib, and Schlather, 2012; Freytag and Bickeböller, 2013; Freytag, Manitz, Schlather, Kneib, Amos, Risch, Chang-Claude, Heinrich, and Bickeböller, 2013), is summarized in Chapter 3. Furthermore, this thesis also includes an original research article about work on coverage and efficiency in current SNP chips (Ha, Freytag, and Bickeböller, 2014). The refrences and respective urls of all original articles can be found in the Appendix A. Finally, Chapter 4 concludes this thesis by weighing the advantages and disadvantages of gene- or network-based kernel methods in the context of genome-wide association studies as well as looking at the future role of kernel methods in genetic epidemiology.

2 Kernel Machine Test and the Generalized Linear Mixed Model

2.1 Logistic Kernel Machine Model and Its Equivalence to the Logistic Mixed Model

The logistic kernel machine model assumes a nonparametric relationship between the disease status and the examined genotypes adjusted for potential environmental variables, such as age, sex or other disease relevant information. This semiparametric model for the probability of being a case can be expressed as follows:

$$\text{logit}\left(P(y_i = 1)\right) = \mathbf{x}_i^T \boldsymbol{\beta} + h(\mathbf{z}_i) \tag{2.1}$$

Here y_i indicates the disease status of the i^{th} individual ($y_i = 0$ unaffected, $y_i = 1$ affected), for $i = 1, \dots, n$. The vector $\mathbf{x}_i$ is used to denote the environmental variables, which are associated with regression coefficients summarized in vector $\boldsymbol{\beta}$. Environmental variables refers to any non-genomic variables, such as age or sex, that could be informative in the context of the disease. The vector $\mathbf{z}_i$ contains all genotypes of the genetic variants of interest. In the case of SNPs (as assumed from here on), genotypes are coded as usual by the number of minor alleles, i.e. $z_{il} \in \{0,1,2\}$ for all SNPs l. The function h is the unspecified nonparametric function that describes the influence of the examined genotypes on the probability of being diseased.

We assume that the unknown function h is in a reproducing kernel Hilbert space (RKHS), $\mathcal{H}_K$, generated by a kernel K. In order to understand the RKHS it helps to recall the more familiar idea of a vector space. A vector space is a mathematical structure formed by an additive group that consists of elements, called vectors, which can be multiplied by a real or complex number. Here, we only deal with the case of real numbers. The inner product of a vector space, $\langle \mathbf{x}, \mathbf{y} \rangle$, defines a bilinear and positive definite map of two arbitrary vectors $\mathbf{x}$ and $\mathbf{y}$ to a number. A vector space that has an inner product and is complete, i.e. a closed set, is called Hilbert space, $\mathcal{H}$. The dimension of such a Hilbert space can be finite or infinite. A typical example for an infinite Hilbert space is given when the elements of the space are quadratically integrable functions. A RKHS is a Hilbert space endowed with the inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}_K}$ when there exists a symmetric kernel K with the following properties (Schölkopf and Smola, 2002):

- K has the reproducing property, i.e $\langle f, K(x,\cdot)\rangle_{\mathcal{H}_K} = f(x)$ for all $f \in \mathcal{H}_K$
- $K(x,\cdot)$ spans $\mathcal{H}_K$,

where $x \in \mathbb{R}^d$. It can be shown that K is positive semi-definite (referred to as simply positive definite in the following), as well as unique.

Alternatively, we consider a symmetric and positive definite kernel $K(x,y)$, $x,y \in \mathbb{R}^d$. Under certain regulatory assumptions, K spans a RKHS, $\mathcal{H}_K$, and has an eigen-expansion

$$K(x,y) = \sum_{i=1}^{\infty} \lambda_i \phi_i(x)\phi_i(y),$$

with eigenvalues $\lambda_i > 0$ and normalized eigenfunctions ϕ_i. Furthermore, $\sum_{i=1}^{\infty} \lambda_i^2 < \infty$ holds. If the space is finite, this decomposition is nothing but the singular value decomposition of a positive definite matrix. Because the ϕ_i's build an orthogonal basis of the RKHS, every function in this space has an expansion in terms of theses eigenfunctions

$$f(x) = \sum_{i=1}^{\infty} f_i \phi_i(x),$$

with the constraint that $\|f\|^2_{\mathcal{H}_K} := \sum_{i=1}^{\infty} \frac{c_i}{\lambda_i} < \infty$, where $\|f\|_{\mathcal{H}_K}$ is the norm induced by K and $f_i, c_i \in \mathbb{R}$. Taking the inner product of two linear combinations f and g, $f, g \in \mathcal{H}_K$, results in

$$\langle f,g\rangle_{\mathcal{H}_K} = \sum_{i=1}^{\infty} \frac{1}{\lambda_i} f_i g_i.$$

The kernel by defining an RKHS represents the core of any kernel method. In the context of applications of kernel methods to GWAS, the following kernel choices are frequently encountered: Up to now for GWAS, the linear kernel $K(\mathbf{z}_i,\mathbf{z}_j) = \mathbf{z}_i^T \mathbf{z}_j$, was probably the most frequently applied kernel. Using this kernel in the logistic kernel machine test, is equivalent to using a logistic regression with a linearly defined random effect for all SNPs. In order to verify this, the reader is referred to the relationship with the logistic mixed model discussed later (compare equation (2.6)). Because of its linear property, this kernel fails in case of interaction between SNPs. Another kernel successfully applied in the context of GWAS is the identity-by-state (IBS) kernel that is constructed by evaluating the proportion of the number of alleles shared between two individuals i and j. The IBS kernel is defined as $K(\mathbf{z}_i,\mathbf{z}_j) = \sum_{l=1}^{p} \frac{2\mathbb{I}(z_{il}=z_{jl})+\mathbb{I}(|z_{il}-z_{jl}|=1)}{2p}$, where $\mathbb{I}$ denotes an indicator function taking the values 0 or 1 and p is the number of SNPs under consideration. This kernel has been shown to be more robust in case of non-linearity of genotype effects than the linear kernel (Wu, Kraft, Epstein, Taylor, Chanock, Hunter, and Lin, 2010), as it is related to piece-wise linear regression. Like the IBS measure, other genetic similarity measures have been used to construct kernels (Schaid, 2010a; Schaid, 2010b). In fact, Schaid (2010a) interpreted the kernel evaluated at $K(\mathbf{z}_i,\mathbf{z}_j)$ as converting information on genotypes of individual i and j to a quantitative value reflecting their similarity.

The flexibility provided by the RKHS can be exploited to fit nonparametric models, like the semiparametric model (2.1). The penalization term used to fit such models typically has the form $\langle f,f\rangle_{\mathcal{H}_K}$. Thus, its behavior is determined by the choice of the kernel K that defines the RKHS. Since $h \in \mathcal{H}_K$, Liu, Ghosh, and Lin (2008) obtained this penalized log-likelihood function

$$J(\boldsymbol{\beta},h) = \sum_{i=1}^{n} \Big(y_i(\mathbf{x}_i^T\boldsymbol{\beta} + h(\mathbf{z}_i)) - \log\{1 + \exp(\mathbf{x}_i^T\boldsymbol{\beta} + h(\mathbf{z}_i))\}\Big) - \frac{1}{2}\lambda\langle h, h\rangle_{\mathcal{H}_K} \quad (2.2)$$

where the parameter λ controls the magnitude of the penalization. Like other penalization approaches, which operate by minimizing a combination of a loss function and a penalty, the solution is restricted in order to provide more robust and accurate estimators. Here, the solution is constrained by the additional biological information in form of the observed genotypes. Penalization is particularly advantageous in situations where variables are highly correlated or when the number of variables exceeds the sample size as in our context. In both cases the non-penalized models would be unidentifiable, but solutions for alternative penalized models can be found.

The representer theorem of Kimeldorf and Wahba (1971) proves that a general solution for $h(\mathbf{z}_i)$, where $\mathbf{z}_i$ for $i = 1,\ldots,n$, in the penalized maximization equation (2.2) has the form

$$h(\mathbf{z}_i) = \sum_{j=1}^{n} \alpha_j K(\mathbf{z}_i,\mathbf{z}_j) = \boldsymbol{\alpha}\mathbf{k}_i^T, \quad (2.3)$$

where $\alpha_j \in \mathbb{R}$ for $j = 1,\ldots,n$ are unknown and $\mathbf{z}_j$ represent the genotype observations of the j^{th} individual. The vector $\mathbf{k}_i$ is a collection of elements $K(\mathbf{z}_i,\mathbf{z}_1),\ldots,K(\mathbf{z}_i,\mathbf{z}_n)$ and the vector $\boldsymbol{\alpha}$ denotes the n unknown parameters $\alpha_1,\ldots,\alpha_n$. Intuitively, the nonparametric function can be interpreted as the linear closure of the similarity of $\mathbf{z}_i$ with the genotype realizations of all other individuals.

Substituting expression (2.3) into (2.2) one obtains,

$$J(\boldsymbol{\beta},\boldsymbol{\alpha}) = \sum_{i=1}^{n} \Big(y_i(\mathbf{x}_i^T\boldsymbol{\beta} + \mathbf{k}_i^T\boldsymbol{\alpha}) - \log\{1 + \exp(\mathbf{x}_i^T\boldsymbol{\beta} + \mathbf{k}_i^T\boldsymbol{\alpha})\}\Big) - \frac{1}{2}\lambda\boldsymbol{\alpha}^T\mathbf{K}\boldsymbol{\alpha}, \quad (2.4)$$

where $\mathbf{K}$ is an $n \times n$ matrix whose element (i,j) is $K(\mathbf{z}_i,\mathbf{z}_j)$. Setting $\sigma_K^2 = \frac{1}{\lambda}$ and $\mathbf{b} = \mathbf{K}\boldsymbol{\alpha}$ the log-likelihood (2.4) becomes

$$J(\boldsymbol{\beta},\mathbf{b}) = \sum_{i=1}^{n} \Big(y_i(\mathbf{x}_i^T\boldsymbol{\beta} + \mathbf{b}_i) - \log\{1 + \exp(\mathbf{x}_i^T\boldsymbol{\beta} + \mathbf{b}_i)\}\Big) - \frac{1}{2\sigma_K^2}\mathbf{b}^T\mathbf{K}^{-1}\mathbf{b}. \quad (2.5)$$

It can be verified that this is the exact penalized-quasi-likelihood obtained from the logistic mixed model of the class of generalized linear mixed models (Liu, Ghosh, and Lin, 2008):

$$\begin{aligned} \operatorname{logit}\left(P(y_i = 1)\right) &= \mathbf{x}_i^T \boldsymbol{\beta} + \mathbf{b}_i \\ \mathbf{b} &\sim \mathcal{N}(\mathbf{0}, \sigma_K^2 \mathbf{K}), \end{aligned} \tag{2.6}$$

where $\mathbf{b}$ is a random effect, which is normally distributed with mean $\mathbf{0}$ and covariance $\sigma_K^2\mathbf{K}$.

The logistic mixed model (2.6) has an intuitive Bayesian interpretation due to its hierarchical expression. The distribution associated with the random effect $\mathbf{b}$ can be viewed as a prior. In particular, the kernel represents a prior correlation structure for the dependencies among the genetic variants. Furthermore, this viewpoint also emphasizes the dimension reduction: The kernel matrix has dimension $n \times n$ but encompasses all genetic variants.

2.2 Score Statistics and Their Asymptotic Distribution

In order to test whether a genetic effect influences the risk of developing the investigated disease, one can consider the null hypothesis $\mathrm{H}_0 : h(\cdot) = 0$ against the alternative $\mathrm{H}_1 : h(\cdot) \neq 0$. Via the connection to the logistic mixed model, we know that this is equivalent to testing the variance component σ_K^2 as $\mathrm{H}_0 : \sigma_K^2 = 0$ against the alternative $\mathrm{H}_1 : \sigma_K^2 > 0$. Note that the null hypothesis places σ_K^2 on the boundary of the parameter space. To test whether all variance components are indeed zero, Lin (1997) proposed to use a score test. Furthermore, Lin developed a unifying theory for testing in the GLMM framework. The score test statistic, unlike the likelihood ratio, has the advantage that estimates under the alternative hypothesis are not required. One can simply estimate the regression coefficients $\boldsymbol{\beta}$ related to the environmental variables under the null hypothesis. This makes the test extremely fast to calculate — an invaluable property for GWAS data analysis where hundreds or thousands of such tests need to be conducted. However, it should be noted that software limitations can increase computation time for large samples.

In the case of the logistic mixed model with the canonical link, as considered here (compare model (2.6)), the score test statistic is

$$Q = \frac{1}{2}(\mathbf{y} - \hat{\boldsymbol{\mu}}^{(0)})^T \mathbf{K} (\mathbf{y} - \hat{\boldsymbol{\mu}}^{(0)}), \tag{2.7}$$

where $\hat{\boldsymbol{\mu}}^{(0)}$ is a vector with elements $\hat{\mu}_i^{(0)} = \operatorname{logit}^{-1}(\mathbf{x}_i^T \hat{\boldsymbol{\beta}})$, the maximum likelihood estimate under the null hypothesis. For a detailed derivation of the score statistic (2.7) for the logistic mixed model refer to le Cessie and van Houwelingen (1995). Using the linear kernel $\mathbf{K} = \mathbf{Z}^T\mathbf{Z}$, where $\mathbf{Z}$ is the collection of the genotype vectors for all individuals, the score statistic can be shown to be equivalent to the sum of squares marginal scores statistics for multiple markers proposed by Pan (Pan, 2009;

Pan, Han, and Shen, 2010). The sum of squares marginal scores statistics as well as a weighted version have been shown to be very powerful and efficient for testing association of multiple markers with a disease (Basu and Pan, 2011).

The score statistic Q is a quadratic form because $(\mathbf{y} - \hat{\boldsymbol{\mu}}^{(0)})$ is a vector of random variables and $\mathbf{K}$ is, by definition, symmetric. Since, $\mathrm{E}(\mathbf{y} - \hat{\boldsymbol{\mu}}^{(0)}) = \mathbf{0}$ and $\mathrm{Cov}(\mathbf{y} - \hat{\boldsymbol{\mu}}^{(0)}) = \mathbf{V}$ with rank of $\mathbf{V}$ smaller or equal to n

$$Q = \sum_{i=1}^{l} \zeta_i U_i^2,$$

where ζ_i are the eigenvalues of $\mathbf{KV}$ and U_i for $i = 1, \ldots, l$ are random variables. By the central limit theorem all U_i are approximately normally distributed. It is well known that the distribution of Q is therefore a weighted sum of l independent χ_1^2 distributions.

For GLMM mixing probabilities are difficult to obtain, instead a Satterthwaite approximation can be used (Schaid, 2010a). This procedure compares Q to a scaled chi-square distribution with scale parameter κ and effective degrees of freedom ν. The values of κ and ν are calculated by moment matching, i.e. the mean and variance of Q are equated to the mean and variance of the scaled χ_ν^2 distribution. The mean and the variance of a quadratic form have the following expressions (Ravishanker and Dey, 2002)

$$\mathrm{E}(Q) = \frac{1}{2}\mathrm{trace}(\mathbf{PK}) \tag{2.8}$$

$$\begin{aligned}\mathrm{Var}(Q) &= \frac{1}{2}\mathrm{trace}(\mathbf{PKPK}) \\ &\quad - \frac{1}{4}\sum_{i=1}^{n} A_{ii}^2 \hat{\mu}_i^{(0)}(1 - \hat{\mu}_i^{(0)})(1 + 6(\hat{\mu}_i^{(0)})^2 - 6\hat{\mu}_i^{(0)}),\end{aligned} \tag{2.9}$$

where $\mathbf{P} = \mathbf{W} - \mathbf{WX}(\mathbf{X}^T\mathbf{WX})^{-1}\mathbf{X}^T\mathbf{W}$. The diagonal matrix $\mathbf{W}$ is defined as $\mathrm{diag}(\hat{\boldsymbol{\mu}}^0(1 - \hat{\boldsymbol{\mu}}^{(0)}))$ and $\mathbf{A}$ is given by $\mathbf{WPKPW}$. The mean and variance of a χ_ν^2 distribution scaled with κ are known to be $\kappa\nu$ and $2\kappa^2\nu$, respectively. Combining this as well as Equation (2.8) and Equation (2.9) one can easily find expressions for the scale parameter κ and the degrees of freedom ν:

$$\kappa = \frac{\mathrm{Var}(Q)}{2\mathrm{E}(Q)} \tag{2.10}$$

$$\nu = \frac{2\mathrm{E}(Q)^2}{\mathrm{Var}(Q)} \tag{2.11}$$

More exact methods for the approximation of the distribution that take into account the unique setting at the boundary of the parameter space have been suggested. One example is an approximation method by Davies (1980), which is based on the inversion of the characteristic functions.

3 Summaries

3.1 Comparison of Three Summary Statistics for Ranking Genes in Genome-Wide Association Studies

Conventional analysis of GWAS data, like single marker analysis, often yields few SNPs that reach genome-wide significance. Therefore, researchers supplement their results by ranking genes according to their association with the development and progression of the investigated disease. This is typically measured by the combined association of all SNPs in the gene with the disease. The advantages of such an approach over single marker analysis are the disposal of the conventional significance threshold, the potential to identify genes with many moderately associated SNPs as well as the availability of biological context to aid interpretation. Even though ranking genes is common practice and methods such as gene-set enrichment analysis rely on it, there have been no comparative studies that investigate the performance of different multiple-marker methods in the context of ranking. Thus, with the help of simulations we set out to investigate the following questions:

- Which multi-marker method most accurately ranks genes according to their relative strength of association? Here we investigated three different multi-marker methods; a powerful sum test called the weighted sum of marginal score statistics, a collapsing strategy named RareCover and the well-known prediction method elastic net regularization. The methods were selected to be representative of the most common classes of multiple-marker approaches.
- Do LD structures between causal SNPs, the direction of the genetic effect and interactions between causal SNPs influence the ability to accurately rank the genes?
- Are any of the methods able to consistently detect association with low frequency variants? Such ability would suggest an applicability of the method to modern sequencing data.

We simulated 9 different scenarios, all investigating different genetic aspects. In the basic scenarios we simulated one causal gene which contained causal SNPs either in strong LD or not in LD at all. We also simulated more advanced scenarios where the causal gene contained an interaction between two causal SNPs, causal SNPs with opposing effects and causal SNPs with low minor allele frequencies. Additionally, we simulated scenarios with two causal genes to investigate whether the different multiple-marker methods correctly rank the causal genes according to their strength of association with the disease. All simulations were conducted with the program HAPGEN2 (Su, Marchini, and Donnelly, 2011) and the CEU reference sample of the International HapMap Project (The International HapMap Consortium, 2007), which mimics real genetic data. In order to guarantee the computational feasibility of the simulations, we limited our investigations to genes located in a region on the first chromosome. Instead of using known genes, we split all observed genes in this region into "sub-genes" of equal size in terms of number of SNPs. This strategy allowed us to avoid the use of costly permutations when calculating the association of a gene.

Overall, the weighted sum of squared score test was demonstrated to be the most robust multi-marker methods of all investigated approaches. In particular, it was the only method to rank the causal gene consistently at the top when interactions between the causal SNPs were present. However, this approach struggled when causal genes included low frequency SNPs with small effects. Under such circumstances, the RareCover algorithm performed significantly better. Despite identifying such associated genes reliably, the RareCover algorithm sometimes failed to rank the causal genes in the correct order. The elastic net regularization proved computationally extremely demanding and its performance was unimpressive in all but the simplest scenarios. Furthermore, we showed that LD had a strong positive effect on the performance of all methods. Note that none of the approaches was able to accurately rank genes when the causal gene contained SNPs with opposing effects.

The results from the simulation study as well as details on the different multiple-marker methods can be found in Freytag and Bickeböller (2013).

3.2 A Novel Kernel for Correcting Size Bias in the Logistic Kernel Machine Test with an Application to Rheumatoid Arthritis

In their review paper on the analysis of biological pathways in GWAS, Wang, Li, and Hakonarson (2010) named the lack of adjustment for pathway size, the failure to incorporate LD structure and the inadequate accommodation of pathway architecture among other challenges facing pathway-based analysis methods. Many pathway-based methods are known to produce deflated p-values when the size of a pathway in terms of number of genes as well as number of SNPs is not properly accounted for. Pathways with more SNPs or genes are more likely to generate small p-values solely by random chance. Similarly, the failure to incorporate LD structures can lead to false-positives, as SNPs are wrongly assumed to be independent. Finally, inadequate accommodation of pathway architecture might have a severe impact on the sensitivity and power of an analysis approach. Since the logistic kernel machine test using the linear or identity-by-state kernel in combination with the Saitterthwaite approximation as applied to pathways does not explicitly deal with these challenges, the following questions arise:

- How does the size of a pathway as measured by the number of SNPs or genes influence the type I error of the logistic kernel machine test with a linear kernel?
- Can one construct a kernel that is invariant to differential pathway sizes? Can this kernel simultaneously account for basic architecture of a pathway, i.e. membership of SNPs in genes, LD patterns between SNPs and interaction between genes?
- How well does it perform compared to other approaches, such as the hierarchical Bayes prioritization (Lewinger, Conti, Baurley, Triche, and Thomas, 2007) and gene-set enrichment analysis (Wang, Li, and Bucan, 2007)?

At first, we investigated the questions using a combination of simulation study, permutation study and application to a real data set on rheumatoid arthritis, which was collected by the North American Rheumatoid Arthritis Consortium (Amos, Chen, Seldin, Remmers, Taylor, Criswell, Lee, Plenge, Kastner, and Gregersen, 2009). Indeed all three approaches confirmed that p-values are deflated for pathways including more than 1,000 SNPs. In particular, for small type I error rates the method was not able to maintain the correct type I error level. Thus, we constructed a novel kernel that included an adjustment for the number of genes contained in the pathway as well as for the LD-corrected number of SNPs in each gene located in the pathway. Moreover, due to its multiplicative nature on the gene level our kernel integrates the basic architecture of a pathway. It models interaction between all genes and groups SNPs according to their membership in genes. We were able to verify the size-invariance property of this kernel by the aforementioned permutation study. This was supported by the results from the application to the GWAS on rheumatoid arthritis.

For the rheumatoid arthritis data, the logistic kernel machine test with our kernel identified several biologically reasonable pathways as associated using a Bonferroni-corrected significance threshold. Of these pathways the majority had previously been mentioned to be significantly associated in at least one scientific publication concerning GWAS and rheumatoid arthritis. Interestingly, we were able to establish two unknown associations with pathways for APT-binding cassette transporters and extracellular matrix receptor interaction, which are both promising and biologically plausible. In contrast, the logistic kernel machine test employing the linear kernel detected an improbably large number of significant pathways. Comparisons to the two other pathway-based approaches, hierarchical Bayes prioritization and gene set enrichment analysis, for which we used results from the analysis of the real rheumatoid arthritis GWAS obtained by Sohns, Rosenberger, and Bickeböller (2009), demonstrated that the logistic kernel machine test with our novel kernel was more robust.

The results from the simulation study, permutation study and the real GWAS data, as well as details on the construction of the novel size-invariant kernel and similar but inferior kernels have been published in Freytag, Bickeböller, Amos, Kneib, and Schlather (2012). Furthermore, the R-package "PathLKMT" implementing the logistic kernel machine test with several kernels for the analysis of GWAS data is available upon request.

3.3 A Network-Based Kernel Machine Test for the Identification of Risk Pathways in Genome-Wide Association Studies

A major shortcoming of the size-invariant kernel is its inability to capture the complex nature of biological pathways. Even though the size-invariant kernel allows the integration of basic pathway structure, it fails to exploit information on precisely which genes interact and whether an interaction is inhibiting or activating. However, several studies have demonstrated that knowledge about regulatory relationships among genes can indeed be helpful for discovering associations with diseases (Chen, Cho, and Zhao, 2011; Lee, Li, Li, Rebman, Achour, Regan, Gamazon, Chen, Yang, Cox, and Lussier, 2013; Lim, Hao, Shaw, Patel, Szabó, Rual, Fisk, Li, Smolyar, Hill, Barabási, Vidal, and Zoghbi, 2006; Lin, Gan, Zhang, Jones, Sjoblom, Wood, Parsons, Papadopoulos, Kinzler, Vogelstein, Parmigiani, and Velculescu, 2007). In particular, SNP-trait associations are often enriched in genes neighboring disease-associated genes and genes occupying structurally relevant positions in the network. Hence, our investigations were guided by the following questions:

- Can one achieve the integration of network topology via an appropriately selected kernel that also adequately corrects for different gene sizes?
- Does the logistic kernel machine test with such a kernel maintain type I error? In which genetic situations is there a power gain from using such a network-based kernel compared to the linear kernel?
- How does the performance of the logistic kernel machine test change?

We were indeed able to construct a network-based kernel that integrates the mathematical graph representation in form of the adjacency matrix of the network under consideration. For this, we were required to find a method that alters any given non-positive definite adjacency matrix to a positive definite adjacency matrix that reflects the exact same network topology. Furthermore, the kernel includes appropriate correction terms standardizing each gene with respect to its size as measured by the number of SNPs. Using a null simulation we were able to show that the logistic kernel machine test with our novel kernel maintains type I error levels. A simple power simulation study, which was based on non-interacting causal SNPs in different genes, demonstrated that there is a power increase when using the network-based kernel compared to the linear kernel in certain scenarios. In these scenarios the causal genes were neighbors or located close to each other with regards to network topology. Applications to two real GWAS, of which one was the previously discussed GWAS on rheumatoid arthritis and the other a GWAS on lung cancer (German Lung Cancer Study) (Sauter, Rosenberger, Beckmann, Kropp, Mittelstrass, Timofeeva, Wolke, Steinwachs, Scheiner, Meese, Sybrecht, Kronenberg, Dienemann, The LUCY-Consortium, Chang-Claude, Illig, Wichmann, Bickeböller, and Risch, 2008), also indicated clear benefits from using the network-based kernel compared to other popular pathway based approaches.

The analysis of the rheumatoid arthritis GWAS with the logistic kernel machine test and the network-based kernel revealed several associated pathways (Bonferroni-corrected significance level). To our knowledge, an associated pathway for drug metabolism, was identified for the first time as susceptibility pathway for rheumatoid arthritis. For the lung cancer GWAS, we could not establish any pathways to be significantly associated at Bonferroni-corrected significance level. However, this was also the case when using the logistic kernel machine test with the linear kernel. Interestingly, the pathway with the smallest p-values was related to the Warburg effect known to be a characteristic of cancer-causing mutations. Additionally, using the real data we also investigated whether network topology alone and not genotypes were driving association results. Fortunately, we did not observe any abnormal correlation between p-values and common descriptive statistics for networks.

For an explanation of the construction of the network-based kernel and the findings from applications to the GWAS on lung cancer and rheumatoid arthritis refer to Freytag, Manitz, Schlather, Kneib, Amos, Risch, Chang-Claude, Heinrich, and Bickeböller (2013).

3.4 Coverage and Efficiency in Current SNP Chips

Genetic epidemiologists typically rely on commercial whole-genome SNP chips to genotype individuals participating in a GWAS. Nowadays, multiple different SNP chips produced by different companies are on offer. They differ greatly in the number of SNPs covered and of course the price. Moreover, strategies of selecting SNPs to be part of a SNP chip vary from chip to chip and are markedly different between companies. This can result in one chip covering more of the genome than another chip of equal size, because LD allows investigators to infer the genotype of SNPs not on the chip with the help of an appropriate reference set. Furthermore, as different populations exhibit different LD patterns, the performance of a chip need not be equal across different ethnicities. Since the success of a GWAS in part depends on cost-effective and thorough coverage of the whole genome, we tried to answer the following questions in our analysis of coverage and efficiency of the more recent SNP chips of Affymetrix and Illumina:

- Which chip covers the greatest number of variation in the European, Asian or African human genome?
- Which chip is the most efficient with regards to exploiting LD structures, to covering SNPs which are not in LD with other SNPs and to cost-benefit analysis?
- In particular, we were interested in the performance of Affymetrix' population-optimized chips, which cover SNPs selected based on the LD patterns observed

in the population of interest. So, are population-optimized chips truly able to provide better coverage of the genome for their targeted ethnicity than chips of comparable size or chips of equal price?

In order to answer these questions, we calculated gene coverage as well as local coverage using the definition of coverage rate by Li, Li, and Guan (2008). Since this coverage rate does not account for chip size, we developed efficiency measures standardized with regards to chip size. Efficiency was measured on the one hand by how well LD is exploited and on the other hand by how many SNPs not in LD with any other SNPs are covered. Comparison of different chips that also take different chip prices into account were provided through our newly introduced cost-benefit ratio. For all calculation, we used samples of appropriate populations found in the 1000 Genome Project Version 3 (The 1000 Genomes Project Consortium, 2012) as a reference set.

None of the investigated Illumina or Affymetrix SNP chips reached the advertised coverage rate stated on the website of the respective company. In general, the local as well as gene coverage of SNPs with minor allele frequency greater than or equal to 5% achieved by chips including more than two million SNPs was excellent. However, coverage of SNPs with minor allele frequency down to 1% was considerably lower even for such extremely big chips. In particular, the performance of the chips for the African population was poor, as this population exhibits lower levels of LD. Interestingly, while the population-targeted chips counted towards the chips with the lowest local and gene coverage in their particular size range, they were able to outperform all other chips in terms of efficiency and cost-benefit (Asian and African population).

The definitions of the newly introduced measures and the results on coverage, efficiency and cost-benefit of the different chips are published in Ha, Freytag, and Bickeböller (2014).

4 Discussion

Kernel methods, like the logistic kernel machine test, possess many characteristics that recommend them for the analysis of association between a disease and biologically meaningful sets of multiple SNPs. Mathematically, such methods are founded in the rigorous framework of the reproducing kernel Hilbert space. In this framework, a model can be fitted with the help of a loss function and a penalty function that is defined by the choice of a kernel. Such an approach proves extremely flexible as well as powerful. The kernel, which aims to capture the genetic similarity between any two individuals in the study, can be chosen from the enormous range of symmetric and positive definite functions. Thus, kernel methods enjoy expressive power to reflect assumptions about genetic relationships. Furthermore, prior biological knowledge, such as gene-gene interaction networks, can be incorporated in the kernel without the need to change the statistical procedure of the model fitting. Also, such incorporation of prior biological knowledge offers biological context for the interpretation of association results and might therefore lead to treatment options. Finally, kernel methods are often faster than comparable multi-marker approaches, because they involve a considerable reduction in the dimensionality of the problem. Instead of the number of markers in the model, the speed of the analysis is determined by the number of individuals in the study.

Kernel methods are not without disadvantages. Even though, calculations required to fit kernel models are generally fast for GWAS of moderate size, this changes dramatically when GWAS include 10,000 individuals or more. Studies of this size are frequently encountered in meta-analysis, where several studies are pooled in order to gain more statistical power. However, at least for the logistic kernel machine test a solution to this problem might be the use of modern graphics processing units, which can be modified to handle large matrix calculations (Fatahalian, Sugerman, and Hanrahan, 2004). Despite offering tremendous flexibility, the generality of the kernel constitutes another disadvantage. This generality can make it difficult to know how to construct a good kernel for a specific problem (Schaid, 2010a). Furthermore, it can sometimes be challenging to translate a model to a symmetric and positive definite function reflecting genetic similarity. For example, including the direction of gene-gene interactions in a network-based kernel for the detection of pathway-disease association violates the requirement of symmetry. A further disadvantage of the logistic kernel machine test is that we can only obtain information about association between the disease and gene or pathway. Even if we discover an association, we do not have information which genes or which SNPs in the pathway or gene are precisely involved in the pathogenesis of a particular disease. Additionally, this limitation to a single level prevents a broader systems biological perspective. Genes do not act independently in the human organism. The same is true for pathways. Therefore, it would be of great interest to be able to model the associations of genes or pathways imbedded in the broader context of the human organism. This should also enable proper consideration of SNPs or genes which are shared among many pathways or genes, respectively.

More challenging than the statistical problems associated with kernel methods are problems of biological nature associated with all pathway-based methods. Modern genotyping chips do not exclusively include SNPs in genes or regulatory sequences. Even though there exist several strategies for assignment of SNPs outside of such regions, assignment rules are likely to be considered arbitrary (Cantor, Lange, and Sinsheimer, 2010). In general, SNPs located more than 500kbp from the nearest known gene often need to be excluded from the analysis. Hence, important associations outside coding regions might be missed. Moreover, the lack of formal assignment may limit the reproducibility of results obtained by pathway-based methods. However, in the future with the help of sequencing studies it can be hoped that the functionality of more and more SNPs will have been predicted (Petersen, Alvarez, DeClaire, and Tintle, 2013). The other major challenges facing pathway-based approaches, and even more so network-based approaches, are inaccuracy and incompleteness of the available regulatory network models. However, in the coming years, it is expected that models will become better, dramatically increasing their ability to yield important insights into the development and progression of diseases (Califano, Butte, Friend, Ideker, and Schadt, 2012). Furthermore, a great abundance of ever evolving pathway databases addressed to different research audiences also hinders the reproduction of results (Cantor, Lange, and Sinsheimer, 2010). Not only are pathway databases inconsistent, but the choice of a database severely biases analyses (Elbers, van Eijk, Franke, Mulder, van der Schouw, Wijmenga, and Onland-Moret, 2009). A solution to this particular problem might lie "[...in the] combined use of manually curated pathways and electronically compiled pathways to ensure comprehensive coverage as well as high-quality information for well-studied pathways" (Wang, Li, and Hakonarson, 2010).

Because of their immense flexibility, kernel methods are one promising approach in order to overcome the challenges posed by modern GWAS data. They are ideally suited to reflect the complex understanding of the human organism that researchers are in the process of elucidating. The ability of kernel methods to adjust to novel biological concepts is demonstrated by our research into modifications of the kernel aiming at a systems biological perspective. In particular, we show that it is possible to incorporate a model of complex interactions — a network — into analyses, which points to the possibility of combining other types of microbiological data with the help of kernel methods. Moreover, the flexibility provided by the kernel can be exploited in order to solve problems of statistical nature. For example, we conducted simulation studies showing that for some kernels a bias owing to the different sizes of the pathways or genes can be incurred. An appropriately chosen kernel, like our size-invariant kernel, can help to prevent such biases.

There already exist several empirical as well as simulation studies concerned with the performance of kernel methods compared to other approaches in the context of GWAS data. As part of our work, we compared the performance of the logistic kernel machine test and the closely related sum of squared score test to other approaches frequently applied to GWAS. Even though for the majority of simulation scenarios and for all real data examples the performance of kernel methods was superior, further studies are needed for validation. This is especially the case for kernel methods used for pathway-based analysis, as our understanding of pathways quickly evolves. In the near future, as more knowledge about the regulatory processes in gene-gene networks becomes available, kernel methods applied to pathway-based analysis of GWAS data should begin to catch the dynamic and cell-specific nature of pathways. Ideally, approaches should incorporate dynamic networks specifically adapted towards the investigated disease rather than integrating static networks. While this will certainly require cooperation between many different disciplines, statisticians will need to reevaluate the appropriateness of the currently used methods. Because of their flexibility, it is likely that modification of kernel methods in order to deal with dynamic biological process regulation is possible.

Bibliography

Amos, C. I., Chen, W., 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 Proceedings* **3**(Suppl 7): S2.

Basu, S. and Pan, W. (2011). 'Comparison of statistical tests for disease association with rare variants'. *Genetic Epidemiology* **35**(7): 606–619.

Califano, A., Butte, A. J., Friend, S., Ideker, T., and Schadt, E. (2012). 'Leveraging models of cell regulation and GWAS data in integrative network-based association studies'. *Nature Genetics* **44**(8): 841–847.

Cantor, R. M., Lange, K., and Sinsheimer, J. S. (2010). 'Prioritizing GWAS Results: A Review of Statistical Methods and Recommendations for Their Application'. *The American Journal of Human Genetics* **86**(1): 6–22.

Cardon, L. R. and Bell, J. I. (2001). 'Association study designs for complex diseases'. *Nature Reviews Genetics* **2**(2): 91–99.

Cessie, S. le and Houwelingen, H. C. van (1995). 'Testing the Fit of a Regression Model Via Score Tests in Random Effects Models'. *Biometrics* **51**(2): 600.

Chen, M., Cho, J., and Zhao, H. (2011). 'Incorporating Biological Pathways via a Markov Random Field Model in Genome-Wide Association Studies'. *PLoS Genetics* **7**(4): e1001353.

Davies, R. B. (1980). 'Algorithm AS 155: The Distribution of a Linear Combination of Chi-Squared Random Variables'. English. *Journal of the Royal Statistical Society. Series C (Applied Statistics)* **29**(3): 323–333.

Eichler, E. E., Flint, J., Gibson, G., Kong, A., Leal, S. M., Moore, J. H., and Nadeau, J. H. (2010). 'Missing heritability and strategies for finding the underlying causes of complex disease'. *Nature Reviews Genetics* **11**(6): 446–450.

Elbers, C. C., Eijk, K. R. van, Franke, L., Mulder, F., Schouw, Y. T. van der, Wijmenga, C., and Onland-Moret, N. C. (2009). 'Using genome-wide pathway analysis to unravel the etiology of complex diseases'. *Genetic Epidemiology* **33**(5): 419–431.

Fatahalian, K., Sugerman, J., and Hanrahan, P. (2004). 'Understanding the efficiency of GPU algorithms for matrix-matrix multiplication'. *Proceedings of the ACM SIGGRAPH/EUROGRAPHICS Conference on Graphics Hardware*. ACM, 133–137.

Freytag, S. and Bickeböller, H. (2013). ‘Comparison of three summary statistics for ranking genes in genome-wide association studies’. *Statistics in Medicine.*

Freytag, S., Bickeböller, H., Amos, C. I., Kneib, T., and Schlather, M. (2012). ‘A Novel Kernel for Correcting Size Bias in the Logistic Kernel Machine Test with an Application to Rheumatoid Arthritis’. *Human Heredity* **74**(2): 97–108.

Freytag, S., Manitz, J., Schlather, M., Kneib, T., Amos, I. C., Risch, A., Chang-Claude, J., Heinrich, J., and Bickeböller, H. (2013). ‘A network-based kernel machine test for the identification of risk pathways in genome-wide association studies’. *Human Heredity* **76**(2): 64–75.

Fridley, B. L. and Patch, C. (2011). ‘Gene set analysis of SNP data: benefits, challenges, and future directions’. *European Journal of Human Genetics* **19**(8): 837–843.

Gibson, G. (2010). ‘Hints of hidden heritability in GWAS’. *Nature Genetics* **42**(7): 558–560.

Ha, N.-T., Freytag, S., and Bickeböller, H. (2014). ‘Coverage and efficiency in current SNP chips’. *European Journal of Human Genetics (Major Revision).*

Hirschhorn, J. N. and Daly, M. J. (2005). ‘Genome-wide association studies for common diseases and complex traits’. *Nature Reviews Genetics* **6**(2): 95–108.

Hofmann, T., Schölkopf, B., and Smola, A. J. (2008). ‘Kernel methods in machine learning’. *The Annals of Statistics* **36**(3): 1171–1220.

Kimeldorf, G. and Wahba, G. (1971). ‘Some results on Tchebycheffian spline functions’. *Journal of Mathematical Analysis and Applications* **33**(1): 82–95.

Koeleman, B. P., Al-Ali, A, Laan, S. van der, and Asselbergs, F. W. (2013). ‘A concise history of genome-wide association studies’. *Saudi Journal of Medicine and Medical Sciences* **1**(1): 4.

Kong, A. et al. (2009). ‘Parental origin of sequence variants associated with complex diseases’. *Nature* **462**(7275): 868–874.

Krueger, J. G., Fretzin, S., Suárez-Fariñas, M., Haslett, P. A., Phipps, K. M., Cameron, G. S., McColm, J., Katcherian, A., Cueto, I., White, T., Banerjee, S., and Hoffman, R. W. (2012). ‘IL-17A is essential for cell activation and inflammatory gene circuits in subjects with psoriasis’. *Journal of Allergy and Clinical Immunology* **130**(1): 145–154.

Larson, N. B. and Schaid, D. J. (2013). ‘A Kernel Regression Approach to Gene-Gene Interaction Detection for Case-Control Studies’. *Genetic Epidemiology.*

Lee, Y., Li, H., Li, J., Rebman, E., Achour, I., Regan, K. E., Gamazon, E. R., Chen, J. L., Yang, X. H., Cox, N. J., and Lussier, Y. A. (2013). ‘Network models of genome-wide association studies uncover the topological centrality of protein interactions in complex diseases’. *Journal of the American Medical Informatics Association* **20**(4): 619–629.

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’. *Genetic Epidemiology* **31**(8): 871–882.

Li, M., Li, C., and Guan, W. (2008). ‘Evaluation of coverage variation of SNP chips for genome-wide association studies’. *European Journal of Human Genetics* **16**(5): 635–643.

Lim, J., Hao, T., Shaw, C., Patel, A. J., Szabó, G., Rual, J.-F., Fisk, C. J., Li, N., Smolyar, A., Hill, D. E., Barabási, A.-L., Vidal, M., and Zoghbi, H. Y. (2006). 'A Protein-Protein Interaction Network for Human Inherited Ataxias and Disorders of Purkinje Cell Degeneration'. *Cell* **125**(4): 801–814.

Lin, J., Gan, C. M., Zhang, X., Jones, S., Sjoblom, T., Wood, L. D., Parsons, D. W., Papadopoulos, N., Kinzler, K. W., Vogelstein, B., Parmigiani, G., and Velculescu, V. E. (2007). 'A multidimensional analysis of genes mutated in breast and colorectal cancers'. *Genome Research* **17**(9): 1304–1318.

Lin, X (1997). 'Variance component testing in generalised linear models with random effects'. *Biometrika* **84**(2): 309–326.

Lin, X., Cai, T., Wu, M. C., Zhou, Q., Liu, G., Christiani, D. C., and Lin, X. (2011). 'Kernel machine SNP-set analysis for censored survival outcomes in genome-wide association studies'. *Genetic Epidemiology* **35**(7): 620–631.

Liu, D., Ghosh, D., and Lin, X. (2008). 'Estimation and testing for the effect of a genetic pathway on a disease outcome using logistic kernel machine regression via logistic mixed models'. *BMC Bioinformatics* **9**(1): 292.

Locke, A. E., Dooley, K. J., Tinker, S. W., Cheong, S. Y., Feingold, E., Allen, E. G., Freeman, S. B., Torfs, C. P., Cua, C. L., Epstein, M. P., Wu, M. C., Lin, X., Capone, G., Sherman, S. L., and Bean, L. J. (2010). 'Variation in folate pathway genes contributes to risk of congenital heart defects among individuals with Down syndrome'. *Genetic Epidemiology* **34**(6): 613–623.

Maity, A., Sullivan, P. F., and Tzeng, J.-Y. (2012). 'Multivariate Phenotype Association Analysis by Marker-Set Kernel Machine Regression'. *Genetic Epidemiology* **36**(7): 686–695.

Malzahn, D., Friedrichs, S., Rosenberger, A., and Bickeböller, H. (2013). 'Kernel score statistic for dependent data'. *BMC Proceedings (In Press).*

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.

Mefford, H. C. and Eichler, E. E. (2009). 'Duplication hotspots, rare genomic disorders, and common disease'. *Current Opinion in Genetics & Development* **19**(3): 196–204.

Neale, B. M., Kou, Y., Liu, L., Maáyan, A., Samocha, K. E., Sabo, A., Lin, C.-F., Stevens, C., Wang, L.-S., Makarov, V., Polak, P., Yoon, S., Maguire, J., Crawford, E. L., Campbell, N. G., Geller, E. T., Valladares, O., Schafer, C., Liu, H., Zhao, T., Cai, G., Lihm, J., Dannenfelser, R., Jabado, O., Peralta, Z., Nagaswamy, U., Muzny, D., Reid, J. G., Newsham, I., Wu, Y., Lewis, L., Han, Y., Voight, B. F., Lim, E., Rossin, E., Kirby, A., Flannick, J., Fromer, M., Shakir, K., Fennell, T., Garimella, K., Banks, E., Poplin, R., Gabriel, S., DePristo, M., Wimbish, J. R., Boone, B. E., Levy, S. E., Betancur, C., Sunyaev, S., Boerwinkle, E., Buxbaum, J. D., Cook Jr, E. H., Devlin, B., Gibbs, R. A., Roeder, K., Schellenberg, G. D., Sutcliffe, J. S., and Daly, M. J. (2012). 'Patterns and rates of exonic de novo mutations in autism spectrum disorders'. *Nature* **485**(7397): 242–245.

Nuytemans, K., Bademci, G., Inchausti, V., Dressen, A., Kinnamon, D. D., Mehta, A., Wang, L., Züchner, S., Beecham, G. W., Martin, E. R., et al. (2013). 'Whole exome sequencing of rare variants in EIF4G1 and VPS35 in Parkinson disease'. *Neurology* **80**(11): 982–989.

Pan, W. (2009). 'Asymptotic tests of association with multiple SNPs in linkage disequilibrium'. *Genetic Epidemiology* **33**(6): 497–507.

Pan, W., Han, F., and Shen, X. (2010). 'Test Selection with Application to Detecting Disease Association with Multiple SNPs'. *Human Heredity* **69**(2): 120–130.

Petersen, A., Alvarez, C., DeClaire, S., and Tintle, N. L. (2013). 'Assessing Methods for Assigning SNPs to Genes in Gene-Based Tests of Association Using Common Variants'. *PloS One* **8**(5): e62161.

Ravishanker, N. and Dey, D. (2002). *A first course in linear model theory.* Texts in statistical science series. Boca Raton: Chapman & Hall/CRC.

Risch, N. J. (2000). 'Searching for genetic determinants in the new millennium'. eng. *Nature* **405**(6788): 847–856.

Risch, N. and Merikangas, K. (1996). 'The future of genetic studies of complex human diseases'. *Science-AAAS-Weekly Paper Edition* **273**(5281): 1516–1517.

Rosenbloom, K. R., Dreszer, T. R., Pheasant, M., Barber, G. P., Meyer, L. R., Pohl, A., Raney, B. J., Wang, T., Hinrichs, A. S., Zweig, A. S., Fujita, P. A., Learned, K., Rhead, B., Smith, K. E., Kuhn, R. M., Karolchik, D., Haussler, D., and Kent, W. J. (2009). 'ENCODE whole-genome data in the UCSC Genome Browser'. *Nucleic Acids Research* **38**(Database): D620–D625.

Sauter, W., Rosenberger, A., Beckmann, L., Kropp, S., Mittelstrass, K., Timofeeva, M., Wolke, G., Steinwachs, A., Scheiner, D., Meese, E., Sybrecht, G., Kronenberg, F., Dienemann, H., The LUCY-Consortium, Chang-Claude, J., Illig, T., Wichmann, H.-E., Bickeböller, H., and Risch, A. (2008). 'Matrix Metalloproteinase 1 (MMP1) Is Associated with Early-Onset Lung Cancer'. *Cancer Epidemiology Biomarkers & Prevention* **17**(5): 1127–1135.

Schaid, D. J. (2010a). 'Genomic Similarity and Kernel Methods I: Advancements by Building on Mathematical and Statistical Foundations'. *Human Heredity* **70**(2): 109–131.

– (2010b). 'Genomic Similarity and Kernel Methods II: Methods for Genomic Information'. *Human Heredity* **70**(2): 132–140.

Schölkopf, B. and Smola, A. J. (2002). *Learning with kernels: support vector machines, regularization, optimization, and beyond.* Adaptive computation and machine learning. Cambridge, Mass: MIT Press.

Shui, I. M., Mucci, L. A., Wilson, K. M., Kraft, P., Penney, K. L., Stampfer, M. J., and Giovannucci, E. (2012). 'Common Genetic Variation of the Calcium-Sensing Receptor and Lethal Prostate Cancer Risk'. *Cancer Epidemiology Biomarkers & Prevention* **22**(1): 118–126.

Sohns, M., Rosenberger, A., and Bickeböller, H. (2009). 'Integration of a priori gene set information into genome-wide association studies'. *BMC Proceedings* **3**(Suppl 7): S95.

Su, Z., Marchini, J., and Donnelly, P. (2011). 'HAPGEN2: simulation of multiple disease SNPs'. *Bioinformatics* **27**(16): 2304–2305.

The 1000 Genomes Project Consortium (2012). 'An integrated map of genetic variation from 1,092 human genomes'. *Nature* **491**(7422): 56–65.

The International HapMap Consortium (2007). 'A second generation human haplotype map of over 3.1 million SNPs'. *Nature* **449**(7164): 851–861.

Visscher, P. M., Brown, M. A., McCarthy, M. I., and Yang, J. (2012). 'Five Years of GWAS Discovery'. *The American Journal of Human Genetics* **90**(1): 7–24.

Visscher, P. M. and Montgomery, G. W. (2009). 'Genome-wide association studies and human disease'. *JAMA: The Journal of the American Medical Association* **302**(18): 2028–2029.

Viswanathan, G. A., Seto, J., Patil, S., Nudelman, G., and Sealfon, S. C. (2008). 'Getting started in biological pathway construction and analysis'. *PLoS Computational Biology* **4**(2): e16.

Wang, K., Li, M., and Bucan, M. (2007). 'Pathway-based approaches for analysis of genomewide association studies'. *The American Journal of Human Genetics* **81**(6): 1278–1283.

Wang, K., Li, M., and Hakonarson, H. (2010). 'Analysing biological pathways in genome-wide association studies'. *Nature Reviews Genetics* **11**(12): 843–854.

Wheeler, D. L., Barrett, T., Benson, D. A., Bryant, S. H., Canese, K., Chetvernin, V., Church, D. M., DiCuccio, M., Edgar, R., Federhen, S., Geer, L. Y., Kapustin, Y., Khovayko, O., Landsman, D., Lipman, D. J., Madden, T. L., Maglott, D. R., Ostell, J., Miller, V., Pruitt, K. D., Schuler, G. D., Sequeira, E., Sherry, S. T., Sirotkin, K., Souvorov, A., Starchenko, G., Tatusov, R. L., Tatusova, T. A., Wagner, L., and Yaschenko, E. (2007). 'Resources of the National Center for Biotechnology Information'. *Nucleic Acids Research* **35**(Database): D5–D12.

Wu, M. C., Kraft, P., Epstein, M. P., Taylor, D. M., Chanock, S. J., Hunter, D. J., and Lin, X. (2010). 'Powerful SNP-Set Analysis for Case-Control Genome-wide Association Studies'. *The American Journal of Human Genetics* **86**(6): 929–942.

Wu, M. C., Lee, S., Cai, T., Li, Y., Boehnke, M., and Lin, X. (2011). 'Rare-variant association testing for sequencing data with the sequence kernel association test'. *The American Journal of Human Genetics* **89**(1): 82–93.

Wu, M. C., Maity, A., Lee, S., Simmons, E. M., Harmon, Q. E., Lin, X., Engel, S. M., Molldrem, J. J., and Armistead, P. M. (2013). 'Kernel Machine SNP-Set Testing Under Multiple Candidate Kernels'. *Genetic Epidemiology* **37**(3): 267–275.

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'. *Genetic Epidemiology* **33**(8): 700–709.

Zuk, O., Hechter, E., Sunyaev, S. R., and Lander, E. S. (2012). 'The mystery of missing heritability: Genetic interactions create phantom heritability'. *Proceedings of the National Academy of Sciences* **109**(4): 1193–1198.

A References, Web-Links and Digital Object Identifiers of Original Articles

Freytag S., Bickeböller H.
Comparison of Three Summary Statistics for Ranking Genes in Genome-Wide Association Studies.
Statistics in Medicine (2013).
URL: `http://onlinelibrary.wiley.com/doi/10.1002/sim.6063/full`
DOI: 10.1002/sim.6063

Freytag S., Bickeböller H., Amos I. C., Kneib T., Schlather M.
A Novel Kernel for Correcting Size Bias in the Logistic Kernel Machine Test with an Application to Rheumatoid Arthritis.
Human Heredity (2012); **74**(2), 97-108.
URL: `http://www.karger.com/Article/Abstract/347188`
DOI: 10.1159/000347188

Freytag S., Manitz J., Schlather M., Kneib T., Amos I. C.,Risch A., Chang-Claude J., Heinrich J., Bickeböller H.
A Network-Based Kernel Machine Test for the Identification of Risk Pathways in Genome-Wide Association Studies.
Human Heredity (2014); **76**(2), 64-75.
URL: `http://www.karger.com/Article/Abstract/357567`
DOI: 10.1159/000357567

Ha N.-T., Freytag S., Bickeböller H.
Coverage and Efficiency in Current SNP Chips.
European Journal of Human Genetics (2014).
URL: `http://www.nature.com/ejhg/journal/vaop/ncurrent/full/ejhg2013304a.html`
DOI: 10.1038/ejhg.2013.304

B Curriculum Vitae

Saskia Freytag	born June 5, 1988, in Munich, Germany
Nationality:	German

Education

since 01/2014	**Post-doc position** at the Bioinformatics Division, The Walter and Eliza Hall Institute, Melbourne, Australia; supervised by Dr. M. Bahlo
10/2010-01/2014	**PhD position** at the Institute of Genetic Epidemiology, Medical School, Georg-August-University Göttingen; supervised by Prof. H. Bickeböller and Prof. M. Schlather
10/2010-01/2014	Member of the Deutsche Forschungsgemeinschaft (DFG) Research Training Group "Scaling Problems in Statistics" (GRK 1644)
10/2006-08/2010	**MSci Statistical Science (International Programme)** University College London, United Kingdom
09/2008-05/2009	Year Abroad Programme within MSci Statistical Science Purdue University, West Lafayette, Indiana, USA
08/2006	International Baccalaureate
08/2004-08/2006	Sixth Form Education, Hockerill Anglo European College, Bishop's Stortford, United Kingdom
07/1994-07/2004	School Education in Germany

Awards

2010	**Royal Statistical Society Award for Master Thesis**

Publications

2014 **Freytag S.*, Manitz J.*, Schlather M., Kneib T., Amos I. C.,Risch A., Chang-Claude J., Heinrich J., Bickeböller H.**
A Network-based Kernel Machine Test for Identification of Risk Pathways in Genome-Wide Association Studies.
Human Heredity; **76**(2), 64-75.
* equal contribution

2014 **Ha N.-T.*, Freytag S.*, Bickeböller H.**
Coverage and Efficiency in Current SNP Chips.
European Journal of Human Genetics.
* equal contribution

2013 **Freytag S., Bickeböller H.**
Comparison of Three Summary Statistics for Ranking Genes in Genome-Wide Association Studies.
Statistics in Medicine.

2012 **Freytag S., Bickeböller H., Amos I. C., Kneib T., Schlather M.**
A Novel Kernel for Correcting Size Bias in the Logistic Kernel Machine Test with an Application to Rheumatoid Arthritis.
Human Heredity; **74**(2), 97-108.

2011 **Freytag S., Sohns M., Bickeböller H.**
Coverage of Exon-Targeted Sequencing.
In Hemmelmann C. *et al.*, Biometrische Aspekte der Genomanalyse IV: Next Generation Sequencing Data Analysis, Shaker Verlag; 91-98.

2010 **Freytag S.**
Conservative Confidence Intervals with Nonresponse.
Unpublished Master thesis, Department of Statistical Science, University College London, United Kingdom;
supervised by Dr. C. Henning.