Kernel-Based Pathway Approaches for Testing and Selection

Dissertation

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

vorgelegt von

Stefanie Friedrichs

aus Kassel

Göttingen, 2017

Supervisor:	Professor Heike Bickeböller Institut für Genetische Epidemiologie Universitätsmedizin Göttingen Georg-August-Universität Göttingen
Second Thesis Committee Member:	Professor Tim Beißbarth Institut für Medizinische Statistik Universitätsmedizin Göttingen Georg-August-Universität Göttingen
Third Thesis Committee Member:	Professor Thomas Kneib Professur für Statistik Wirtschaftswissenschaftliche Fakultät Georg-August-Universität Göttingen
Date of Defence:	25 September, 2017

Affidavit

I hereby declare that my doctoral thesis entitled *Kernel-Based Pathway Approaches for Testing and Selection* was written independently by myself and without the use of any other sources or aids than quoted.

Stefanie Friedrichs

Göttingen, August 2017

Acknowledgements

I would first like to thank my supervisor Professor Heike Bickeböller for her support throughout the last four years. Her door was always open to me when I needed to discuss problems, even during periods of stress. Professor Bickeböller always took the time to help me find the best possible solution for problems I encountered. Her scientific experience and guidance through the academic world was an indispensable aid. I would like to thank my second thesis committee member Professor Tim Beißbarth for listening to my reports on the state of my research and sharing his ideas and advice in stimulating discussions at many points of time on the road to completing this thesis. I would like to thank my third thesis committee member, Professor Thomas Kneib, for his valuable suggestions and discussions on the project, which contributed to the success of this work. Furthermore, I am grateful for his revision and the correction of several drafts of our kernel boosting paper.

I would like to thank my project partners for the good cooperation, in particular Benjamin Hofner for the great cooperation in the kernel boosting project. Further thanks goes to Patricia Burger for her valued assistance throughout my PhD time. I am also indebted to Andrew Entwistle for the time-intensive and precise linguistic correction of this work. Furthermore, Martin Schlather deserves a special mention for his instant help in difficult times, which was more than I could have asked for.

I am grateful to my colleagues at the Institute of Genetic Epidemiology, who, in the face of all the challenges and pitfalls, always made sure that it was most definitely a time worth remembering. My work presented in this thesis was financially supported by the Deutsche Forschungsgemeinschaft (DFG) in the context of my membership in the research training group "Scaling problems in statistics" (RTG 1644). Furthermore, I thank Professor Christopher I. Amos for the kind provision of the rheumatoid arthritis dataset.

Finally, I would like to express my gratitude to my friends and my family, who supported me throughout the time of completing this work. Those closest to me had to endure a number of repetitive talks on the difficulties I encountered en route. I sincerely thank you all for accepting this challenge patiently and still managing to find the energy to give me the courage and backing necessary to reach this point.

Abstract

With the number of single nucleotide polymorphisms (SNPs) available in genetic data currently constantly increasing, the evaluation of SNP sets has become a successful approach toward elucidating the genetic influence on various complex diseases. The joint investigation of multiple SNPs increases the probability of detecting moderate and weak association signals and bypasses the multiple testing problem inherent to testing procedures on the genome-wide scale. Furthermore, this approach assists in the biological interpretation of analysis results, which may be supported by the analysis of SNP sets representing a pathway, here denoting a set of genes fulfilling a particular biological function jointly.

The association between a pathway-representing SNP set and a phenotype may be analysed appropriately with the kernel machine approach. This evaluates the genotypes of multiple SNPs jointly by transforming them into a kernel matrix, comprising the genetic similarity measures for any pair of individuals in the study. The kernel matrix is calculated by a predefined kernel function. Multiple kernel functions have been proposed, some of which are capable of integrating further biological knowledge on a pathway and allow for varying types of effect. The network kernel function enables the direct incorporation of a pathway's network structure, while at the same time considering additive as well as interaction effects in the investigated SNP set.

A multitude of databases are available nowadays offering an increasing amount of biologically meaningful information on pathways, genes, and genetic markers. The initial work in this thesis investigates possibilities and the impact of integrating additional biological information into existing approaches in the analysis of genetic data. The impact of marker density, SNP-set aggregation with respect to linkage disequilibrium structures, and knowledge sources were considered. In this context, the software package kangar00 was developed in R, offering a wide range of functions relating to data download, pre-processing, transformation, and evaluation for single-pathway testing in the logistic kernel machine framework, implemented, and made freely available.

The identification of specific biological processes influencing disease risk is still very challenging, despite the integration of growing amounts of biological data. Single-pathway methods cannot usually discriminate causal processes influencing disease susceptibility from isolated genetic effects included in a pathway resulting from gene overlaps. Moreover, they usually lack the ability to predict any trait of interest.

The main objective of this thesis is the development of a new method in the evaluation of SNP sets, focussing on the analysis of those representing pathways. The resulting analysis approach enables the mutual investigation of multiple sets of SNPs through the adaptation of a boosting algorithm.

Boosting originates from the field of machine learning, in which it was developed as a classification approach. Its main idea is to combine functions with poor classification performance iteratively into a strong classifying set. If the functions considered only depend on a subset of the explanatory variables available, variable selection may be performed while the model is fitted. We made use of this to perform selection on a set of pathways by employing a kernel function dependent on SNP sets representing pathways. Since all pathways of interest are investigated jointly in the boosting algorithm, correlations between them are also considered. We may therefore discriminate biological processes influential on disease susceptibility from single effect genes included in a pathway resulting from gene overlap. Our software package kangar00 includes an interface to a boosting algorithm, together with which all functionalities necessary to apply kernel boosting are available.

Thanks to its inherent properties and the freely available software implementation, kernel boosting has great potential to elucidate key biological functions involved in disease risk, while creating a directly interpretable model to predict disease status.

Contents

Introduction

Many common diseases are influenced by a complex interplay between multiple genetic and non-genetic factors. The field of genetic epidemiology aims to elucidate the genetic elements involved in disease susceptibility and development, while considering their interrelation with environmental influences. Examples of common diseases include cancer, asthma, obesity, and diabetes [1].

The genetic information of each and every individual is represented by the sequence of base pairs found in deoxyribonucleic acid (DNA). Genetic locations that display alterations in this sequence in a population are referred to as genetic markers [2]. The simplest and most frequent type of genetic marker is a variation in a single base pair and is termed a SNP. Although the occurrence of more than one form in the population is possible, a SNP typically has one of two possible manifestations, which are called alleles. The less frequent allele in a population is referred to as the minor allele and the percentage of its occurrence is defined the minor allele frequency (MAF). Since most genetic loci exist on each copy of a chromosome pair in the human genome, two alleles per SNP usually exist, forming the SNP's genotype [2]. In genome-wide SNP data, genotypes are often represented by the count of minor alleles observed at the specific locus.

1.1 Association Analysis of Genetic Data

Genetic association studies aim to understand how genetic variants and certain characteristics of an individual are related. We refer to these characteristics as traits or phenotypes and attempt to identify genetic markers associated with them. A SNP is said to be associated with a phenotype if one of its alleles occurs more frequently together with specific forms of the phenotype than might be expected by chance. We can furthermore distinguish two types of association: Direct association, in which the investigated genetic marker represents the locus influential on the disease, and indirect association, which results purely from the considered marker's correlation to a causal locus [3]. Such a correlation among different genetic loci is referred to as linkage disequilibrium (LD).

A genome-wide association study (GWAS) often involves hundreds of thousands of SNPs distributed across the whole genome. Markers under consideration are selected based on the

idea of covering the complete genetic information without genotyping every SNP in existence. This is possible, since SNPs in reasonably high LD can act as a substitute for each other, so that effects of non-genotyped SNPs can be evaluated as indirect associations [1]. Genetic association studies are often case-control studies. Consequently, diseased probands (cases) are recruited along with healthy individuals (controls). The genotypes of all study participants are determined and, along with informative environmental covariates, used to identify new genetic risk factors based on the distribution of genotypes among said cases and controls [4].

When GWAS were first developed, diseases with a high prevalence in the population were hypothesized as being caused by common genetic variants. This assumption is termed the 'common disease - common variant' hypothesis, in which common variants typically refer to those exceeding a frequency of 5%. Although several arguments in favour of this theory exist, common SNPs were only able to explain a small proportion of the phenotypic variance and cannot account for the levels of heritability discovered in family studies, a phenomenon termed 'missing heritability' [5]. Many complex diseases are probably influenced by both frequent and rare variants, each contributing slightly to the overall disease risk [4].

GWAS data were often evaluated by individually analysing each involved SNP. This procedure leads to notable statistical challenges, one of which being the problem of multiple testing [1]. This particular problem arises from an accumulation of possible type-I errors across the multitude of statistical tests conducted. To maintain the overall type-I error on the experimental level at the desired limit, the significance level for each test needs to be adjusted accordingly [6]. This usually results in a very low p-value threshold required to identify a globally significant association, which hinders the detection of moderate or weak genetic effects. The constantly increasing density of SNP data resulting from technological advances and formation of data-sharing consortia over the last decades has further aggravated this situation. One answer to this problem is the joint investigation of multiple markers, aggregated to SNP sets.

1.2 Linkage Disequilibrium

As mentioned in Section 1.1, correlated loci are said to be in LD, meaning that their alleles do not occur independently of each other. We use the term association to refer to a relation between an allele at a genetic locus and a phenotype, while we use LD to describe a correlation between two alleles at two different genetic loci [3]. A population in which allele combinations at the considered loci only occur in frequencies expected at random formation is said to be in linkage equilibrium at these loci. LD in a population is introduced and modified by various factors, for example, the appearance of a new mutation. The pattern of LD therefore provides insight into genetic processes emerging in the population [1].

Different options exist to measure LD. Imagine two loci with possible alleles A or a and B or b, respectively, and let p_A and p_B denote the corresponding probabilities of observing allele A or B at its locus. If the two loci are not in LD and thus mathematically independent, the probability p_{AB} of observing A and B together can be derived by

$$p_{AB} = p_A \cdot p_B$$

In case of LD, this equation does not hold and we can derive the deviation from linkage equilibrium

$$D = p_{AB} - p_A \cdot p_B$$

which depends on the allele frequencies at the considered loci. This dependency renders it unsuitable as a general comparison measurement of LD, for which different standardizations have been developed [3]. The methods most commonly employed are the measures D' and r^2. D' is a relative value of disequilibrium suggested by Lewontin [7]. In this case, D is divided by the maximum possible value that it could take under the given frequencies. The measure is specified by

$$D' = \frac{D}{D_{max}} \tag{1.1}$$

with D_{max} defined as

$$D_{max} = \begin{cases} min\{p_A \cdot (1 - p_B), (1 - p_A) \cdot p_B\} & \text{if } D > 0 \\ max\{-p_A \cdot p_B, -(1 - p_A) \cdot (1 - p_B)\} & \text{if } D < 0 \end{cases}$$

The measure r^2 is the square of Pearson's product moment correlation and is given by

$$r^2 = \frac{D^2}{p_A \cdot (1 - p_A) \cdot p_B \cdot (1 - p_B)} \tag{1.2}$$

Possible LD values range from 0 to 1, with a value of 0 indicating no LD and 1 corresponding to perfect correlation. SNPs are often categorized as strongly correlated, and thus suitable to represent one another, if their LD exceeds a particular threshold. A typical requirement is an r^2 value of 0.7 or 0.8 [1].

LD is not only important in connection with the selection of GWAS SNPs, but also needs to be taken into account in various genetic analysis scenarios. Depending on the method utilized in the evaluation of SNP data, LD will either have to be corrected for or may be exploited to assist in the detection of associations.

1.3 SNP Sets and Pathway Analysis

Genes act and interact with other genes in human beings following sophisticated mechanisms to perform various biological functions. Thus, any isolated evaluation of single genes is not

sufficient to understand the complex biological systems involved in disease susceptibility [8]. One approach that takes this fact into account is the analysis of pathways. Pathways are biologically defined networks of interacting genes, jointly fulfilling a specific function [9]. As each gene can be represented by the SNPs located within its genomic boundaries, a pathway may be represented by a set of SNPs. The joint investigation of multiple SNPs forming a new unit of analysis has a number of benefits. It evidently reduces the number of tests needed to evaluate GWAS data, which is why SNP-set analysis strategies emerged as one reaction to the multiple testing problem inherent to the single testing of large numbers of markers. Since SNPs are usually aggregated to represent an element of particular biological function, SNP-set analysis more importantly assists the biological interpretation of results. The mutual evaluation of multiple markers also facilitates the detection of several moderate effects, which alone are not strong enough to be of genome-wide significance in single-SNP tests. Furthermore, the evaluation of SNP sets allows us to take interactions between markers into account [10].

The analysis of pathways has become a frequently used approach in the evaluation of GWAS data and a multitude of statistical methods to this purpose exist nowadays. These can be categorized according to their primary characteristics [11, 12, 13]. A clear and structured classification is provided by Khatri and colleagues in [14] and is briefly explained:

Over-representation analysis evaluates the effect of a pathway based on the proportion of influential genes it includes. Using a gene-level test, methods of this class firstly identify putative effect genes. In a second step, the fraction of effect genes in each pathway is evaluated, in which a proportion higher than expected at random indicates an influential pathway. Limitations of this class include the gene cutoff, involving only a part of the available information, as well as the inability to consider interactions among genes. Gene set enrichment analysis (GSEA) is a representative of this type of method [15].

Functional class scoring approaches do not rely on high-effect genes only, but aim to facilitate the detection of multiple interacting genes with moderate effects as well. These methods usually compute gene-level scores for all genes, which are then combined to pathway-level statistics and tested for their influence on disease risk. Although pathway-level statistics can be modelled to allow for interactions among genes, no direct information on the interaction structure of the network is considered. Kernel methods [16, 17], which will be described in Chapter 2, may generally be assigned to this class.

Pathway topology approaches follow a design similar to functional class scoring techniques, but differing in the fact that they incorporate topological knowledge on a pathway into the analysis in addition. They directly exploit interaction patterns of gene networks upon calculations of pathway-level statistics, thus considering interaction between genes. However, correlation between pathways remains unaccounted for. Kernels incorporating topological

information, such as the network kernel [18] presented in the next chapter, are examples in this class.

Two types of hypothesis may be considered when evaluating the influence of a pathway by use of a statistical test. The self-contained null hypothesis evaluates the pathway's effect based on the genes of which it is composed, isolated from other pathways. The competitive approach tests the pathway's importance in comparison with all other pathways under investigation [19]. However, both approaches still only test a single pathway at a time without accounting for any correlation between pathways. Such correlations between pathways may occur as a result of genes being included in several different pathways. A particular implication of this is that individual genes involved in a network influential on disease risk may also be found in other, non-causal pathways. Single-pathway analysis methods are incapable of discriminating causal biological processes from these overlapping effect genes. The joint investigation of multiple pathways is therefore a rather promising strategy to consider correlations among gene networks, to assist in the identification, and foster the understanding of biological systems affecting disease risk.

1.4 Objective and Outline of this Thesis

This work aims to enhance the existing toolbox in the analysis of SNP sets representing pathways through the improvement of existing and the development of new investigation approaches. The integration of additional biological information into tests of a single genetic unit is explored. However, the focus here mainly lies on the development of a new, joint-analysis approach for multiple pathways, aiming to overcome some of the limitations inherent to single-pathway testing procedures. The kernel-boosting approach developed here facilitates the mutual evaluation of the information represented by several pathways in a variable selection framework, resulting in a prediction model for the investigated outcome.

This thesis is organized as follows: Chapter 1 introduces the genetic topics required to follow further sections. Chapter 2 presents the kernel association approach for single-pathway testing, along with its theoretical background. Chapter 3 provides a concise overview of the idea behind boosting algorithms and their functionality, whereas Chapter 4 describes the data considered in application examples. Chapter 5 comprises summaries of the peer-reviewed journal publications constituting the main body of this thesis. As the work focusses on method development rather than the application of analysis methods, the summaries centre on the methodological aspects of the publications.

Kernel Methods in Pathway Analysis

Kernel methods are a machine learning approach especially well suited to the evaluation of pathways. They cope well with the high-dimensional data arising in GWAS analysis, while remaining computationally efficient. Kernel methods are flexible in terms of incorporated information without requiring any direct modelling of interaction structures [20]. The joint association of a SNP set, as representative of a pathway, with a phenotype may be evaluated conveniently in a kernel score test.

2.1 Kernel Machine Approach

On the introduction of pathway analysis as a novel approach, the information on interaction patterns was very limited. Complex and partially or completely unknown network structures rendered non-parametric analysis approaches advisable. A particularly suitable approach in this context is the kernel machine method. It employs a regression framework, in which a trait of interest is explained by parametrically modelled environmental covariates and the effect of a SNP set incorporated perhaps parametrically or non-parametrically [16]. For a study on $i = 1, \ldots, n$ participants, we consider the following regression model

$$y_i = \boldsymbol{x}_i^t \boldsymbol{\beta} + h(\boldsymbol{z}_i) + e_i \tag{2.1}$$

in which y_i denotes the trait measurement of individual i, $\boldsymbol{x}_i^t$ is the transposed $n_q \times 1$ vector of n_q environmental covariates (including intercept) with corresponding coefficient vector $\boldsymbol{\beta}$, and $\boldsymbol{z}_i$ denotes the $n_s \times 1$ dimensional genotype vector of individual i for the n_s SNPs in the SNP set under investigation. Furthermore, $h(\cdot)$ denotes an unspecified function of the genotypes, and e_i the error term of the regression model (assumed to follow $e_i \sim \mathcal{N}(0, \sigma^2)$). The investigated trait may also be binary, for example indicating case-control status, in which case we consider the logistic approach

$$logit(P(y_i = 1)) = \boldsymbol{x}_i^t \boldsymbol{\beta} + h(\boldsymbol{z}_i). \tag{2.2}$$

Here, the trait's expected value $E(y_i) = P(y_i = 1)$ is incorporated by use of the logit link function $logit(x) = log\left(\frac{x}{1-x}\right)$ [21, 22]. This logistic version is often referred to as logistic kernel machine test (LKMT).

In both regression models, the genetic influence is incorporated by an unspecified function $h(\cdot)$. It is assumed to lie in a function space $\mathcal{H}_K$, generated by a chosen kernel function $K(\cdot)$. This kernel function can be selected with a large degree of flexibility. It is however required to be positive semidefinite. Owing to the mathematical characteristics of the space, any function $h(\cdot) \in \mathcal{H}_K$ may be represented as $h(\boldsymbol{z}_i) = \sum_{j=1}^{J} \alpha_j K(\boldsymbol{z}_i, \boldsymbol{z}_j)$, a linear combination of $j = 1, \ldots, J$ suitable parameters α_j, and the kernel function evaluated at sample points $\boldsymbol{z}_j$. As the form of $h(\cdot)$ is restricted to elements of the function space spanned by $K(\cdot)$, the kernel function determines which kind of effect, such as linear effects or interactions, will be considered for the investigated SNP set.

In order to evaluate a pathway's influence on the trait, it is helpful to see that the models given in (2.1) and (2.2) may also be interpreted as mixed models

$$y_i = \boldsymbol{x}_i^t \boldsymbol{\beta} + h_i + e_i \qquad \text{or} \qquad logit(P(y_i = 1)) = \boldsymbol{x}_i^t \boldsymbol{\beta} + h_i \tag{2.3}$$

with fixed covariate effects $\boldsymbol{x}_i^t \boldsymbol{\beta}$, a genetic random effect h_i, and normally distributed error terms e_i in the linear mixed model. The random effect is assumed to follow $h_i \sim \mathcal{N}(0, \tau \boldsymbol{K})$, with an unknown variance component τ, and $\boldsymbol{K}$ the $n \times n$ kernel matrix derived by the application of $K(\cdot)$ to the genotypes of each pair of study participants.

This connection was established by demonstrating that the estimators of $\boldsymbol{\beta}$, $\boldsymbol{\alpha}$ and $h(\cdot)$ derived by maximizing the penalized likelihood in the semiparametric framework are equal to the estimators obtained for the corresponding mixed models. The technical details may be found in [16, 21].

2.2 Variance Component Test

Considering the regression models mentioned above, we wish to examine the overall association of a SNP set with the investigated trait. Thus we are interested in the null hypothesis $H_0 : h(\boldsymbol{z}) = 0$. From the mixed model representation (2.3), it can be seen that this is equivalent to testing for a significant variance component of the random effect

$$H_0 : \tau = 0 \qquad \text{vs} \qquad H_1 : \tau > 0 \tag{2.4}$$

This null hypothesis may be investigated efficiently in a score test which only requires estimation of the null model. This is advantageous, since $h(\cdot)$ disappears under the null hypothesis and thus does not have to be estimated. The test statistic is given by

$$Q = \frac{(\boldsymbol{y} - \hat{\boldsymbol{y}})^t \boldsymbol{K} (\boldsymbol{y} - \hat{\boldsymbol{y}})}{2} \tag{2.5}$$

where $\boldsymbol{y}$ denotes the vector of all n trait observations, $\hat{\boldsymbol{y}}$ the estimated values obtained by fitting the null model of the corresponding regression model, and $\boldsymbol{K}$ the kernel matrix derived

on the genotypes of the SNP set to be evaluated [17]. Not only is this test convenient to compute, it is also not affected by the directionality of the SNP effects [10]. This characteristic makes it well suited for evaluating multiple SNPs that might contain signals of opposite association direction.

Test statistic Q follows a scaled χ^2-distribution $\kappa\chi^2_\nu$, which may be approximated efficiently by the Satterthwaite method [23, 24] or Davies' algorithm [25]. The first approach estimates the unknown parameters by equating the mean and variance of Q with those of the unknown distribution, while the latter uses a numerical inversion of the characteristic function to compute the quadratic forms distribution [26]. Both approaches were implemented in this work and yielded very similar results.

2.3 Kernel Functions and Pathway Information

The previous sections indicate the importance of kernel function $K(\cdot)$ in the kernel machine approach. It implicitly determines which kind of effect will be considered for the investigated SNP set. A linear kernel function restricts the included signals to additive effects of each SNP, while a multiplicative kernel function may allow for interactions among SNPs. The selected kernel function $K(\cdot)$ is applied on the genotype vectors of each two individuals, generating an $n \times n$ matrix $\boldsymbol{K}$ with entries

$$\boldsymbol{K}_{ij} = K(z_i, z_j) \tag{2.6}$$

for $i, j = 1, \ldots, n$. The resulting kernel matrix may be interpreted intuitively as a genetic similarity matrix [17]. This implies that entry $\boldsymbol{K}_{ij}$ may be seen as a numeric value representing the genetic similarity between study participants i and j.

Linear Kernel: A frequent choice of kernel function is that of the linear kernel. If $\boldsymbol{Z}$ denotes the $n \times n_s$ matrix formed by n genotype vectors of length n_s for the considered SNP set, the kernel matrix is calculated by

$$\boldsymbol{K} = \boldsymbol{Z}\boldsymbol{Z}^t \tag{2.7}$$

The linear kernel function evaluates the joint effect of all markers forming a particular SNP set and thus can evaluate a pathway's effect on the investigated outcome. However, it does not include any interactions. All SNP effects are modelled in an additive fashion, implicitly assuming a multiple (logistic) regression model [10]. Since SNPs are involved in complex interactions within human beings, only considering their membership in a pathway alone will not be sufficient to understand biological processes fully [8].

Network Kernel: A large extent of the available knowledge on a considered SNP set or pathway may be incorporated by use of the network kernel function [18]. It also investigates a SNP set representative of a pathway. However, in contrast to the linear kernel, it includes

additional information during calculation of the kernel matrix. The network kernel assigns SNPs to individual genes within the pathway and adjusts this mapping for the total number of markers included per gene. Known interactions between genes are directly incorporated and can be categorised as activating or inhibiting type. Furthermore, the network kernel function allows for pair-wise interactions among the analysed SNPs. The corresponding kernel matrix is calculated by

$$\boldsymbol{K} = \boldsymbol{Z}\boldsymbol{A}\boldsymbol{N}\boldsymbol{A}^t\boldsymbol{Z}^t \tag{2.8}$$

$\boldsymbol{Z}$ again denotes the $n \times n_s$ genotype matrix as in (2.7). For n_g genes in the considered pathway, $\boldsymbol{A}$ maps the n_s SNPs representing the pathway to the genes and therefore is of dimension $n_s \times n_g$. Interactions between the genes are incorporated via the $n_g \times n_g$ network matrix $\boldsymbol{N}$.

In order to visualise how information on the pathway is incorporated more effectively, let us look at one particular entry of the resulting matrix. For individuals i and j, it is equal to

$$\boldsymbol{K}_{ij} = \sum_{u=1}^{n_g}\sum_{v=1}^{n_g} n_{uv} \cdot \sum_{r=1}^{n_s} g_{ir}a_{rv} \cdot \sum_{s=1}^{n_s} g_{js}a_{su} \tag{2.9}$$

The $\sum_{r=1}^{n_s} g_{..}a_{..}$ part of the formula sums the genotypes per gene of a specific individual. Here, $g_{..}$ denotes the minor allele count, where $a_{..}$ is an adjusted indicator variable mapping SNPs to a gene while taking into account the gene's size. It is equal to the reciprocal square root of the number of SNPs in the gene if the SNP maps to the gene and 0 otherwise. For each two genes, the corresponding gene-level sums of the two regarded individuals are multiplied, with an additional factor accounting for the interaction between the genes. Here, $n_{..}$ may equal 1 for an activating interaction, -1 for an inhibiting one, and 0 for no interaction. To ensure the involvement of all SNPs, every gene is modelled as self-interacting by setting $n_{uu} = 1$, for all $u = 1, \ldots, p$.

This function is of particular use, as it has been shown to be superior in terms of performance in the analysis of interconnected effects, which are assumed to occur in pathways influential on disease susceptibility [18].

Pathway Databases: Biological pathways are designed to map molecular reactions occurring inside the cells of an organism. They are involved in numerous aspects, such as metabolism, information processing, disease development or cellular processes, and are usually responsible for a specific product or cell function [9, 27]. More and more information on pathways beyond mere gene membership is available nowadays and may be retrieved from numerous online databases. A comprehensive overview of online resources with relation to pathways can be found on the `pathguide.org` website [28]. The site currently lists over 640 resources, with more than 350 including knowledge on human pathways.

This abundance of sources makes it difficult to decide from where information should be retrieved. The pathway databases available to date differ in a number of ways, such as available species, interaction types, focus, or employed pathway definition [8]. Thus, databases currently demonstrate surprisingly little overlap in information, which may also be attributed partially to the fact that they are still a work in progress [29]. As it is highly likely that no database holds the full information on a pathway (yet), the integration of the knowledge available is desired. However, this poses quite a challenge, given the use of various data formats, naming conventions, and lack of clarity as to whether differing information can be regarded as either complementary or contradictory [29]. A well-considered choice is essential, as the database selection may well influence the results of analysis. No gold standard pathway database exists; however, several quality criteria can assist in selecting a suitable resource. Knowledge should be updated periodically, in order to keep pace with new findings. Manually curated experimental data is considered to be of the highest quality, with computationally inferred and electronically annotated data being viewed as lower in quality. Furthermore, the coverage of a database should be taken into consideration, that is, determining how many known genes are involved in one of the given interactions [8]. Finally, the database's focus should match the research question to provide the best fitting information possible for the individual analysis.

One of the first pathway databases to be established was the Kyoto Encyclopedia of Genes and Genomes (KEGG) database [27, 29]. It was first released in 1995 with the aim of facilitating the biological interpretation of genetic information via pathway mapping. Reference pathway maps were, and still are, manually drawn. KEGG focusses on pathways, but has expanded in various directions over the last 20 years. It now consists of 16 databases, providing knowledge on various types of -omics data, mostly based on information derived from published research articles [27]. KEGG is updated on a weekly to daily [8, 27] basis and changes may be followed by reading the release notes on the website. The pathway information considered in this work was derived solely from this database.

Boosting

Boosting emerged out of the field of machine learning, in which it was designed as a classification approach. The main idea of boosting is to combine weak classifying functions with poor performance into one new classifier with strong prediction ability. The algorithm assigns more weight to the measurements difficult to classify [30]. The development of boosting is attributed to the work of Schapire [31] and Freund and Schapire [32, 33]. They introduced the first boosting algorithm, the now famous *AdaBoost*, which laid the foundations for all the subsequent boosting algorithms.

The concept of boosting has been successfully adapted to the field of statistical modelling, in which it provides a flexible framework for model fitting and variable selection. In this context, boosting is especially well suited to biomedical applications such as the analysis of GWAS data. This may be explained by its inherent properties: Boosting algorithms cope well with high-dimensional data that can include more explanatory variables than observations, various types of variable may be incorporated into one prediction model jointly, and model fitting can automatically include variable selection, thus reducing the set of available predictors to those most relevant that are included in the model [34].

3.1 Introduction to Boosting

As mentioned above, boosting aims to combine weak classifiers in order to 'boost' their performance. We assume having data from a study of $i = 1, \ldots, n$ participants with observations of a binary trait $y_i \in \{0, 1\}$ and a $q \times 1$ dimensional vector $\tilde{\boldsymbol{x}}_i$ of measurements of q predictors. The latter may be of differing kinds, such as continuous or categorical variables.

A weak classifier is a function $f(\tilde{\boldsymbol{x}}_i)$ that predicts y_i with an error rate only slightly better than random guessing. The error rate can be derived as the number of falsely classified observations divided by the total number of classifications performed [30]. In the boosting framework, the weak classifiers $f_j(\cdot)$ are typically referred to as *base-learners*.

From a statistical point of view, the boosting algorithm models the influence of the prediction variables on the investigated trait by fitting a structured additive predictor

$$\eta(\tilde{\boldsymbol{x}}_i) = \beta_0 + \sum_{j=1}^{J} f_j(\tilde{\boldsymbol{x}}_i) \tag{3.1}$$

where β_0 is the intercept and $f_j(\cdot)$ are the $j = 1, \ldots, J$ base-learners considered. One base-learner $f_j(\cdot)$ often does not depend on all predictors in $\tilde{\boldsymbol{x}}_i$, but only on a part $\tilde{\boldsymbol{x}}_{i_j} = (\tilde{x}_{i_{j,1}}, \ldots, \tilde{x}_{i_{j,n_j}})$ ($n_j \leq n_q$). This implies that the J base-learners can incorporate differing effects for the same (subset of) variables. A dependency of several base-learners on the same variables is possible and may be interpreted as modelling alternatives for the particular prediction variable [35].

The quality of a predictor's prognosis of the trait may be measured by an appropriate loss function $\rho(\cdot)$, which indicates the discrepancy between $\eta(\tilde{\boldsymbol{x}}_i)$ and y_i. Different options for $\rho(\cdot)$ exist, among which the squared error loss or a likelihood-based loss function are common choices [30]. The optimal predictor $\eta^*(\cdot)$ would be the function minimizing the expected value of the loss function for general $(y, \tilde{\boldsymbol{x}})$-values [36]. In practice, an approximation $\hat{\eta}(\cdot)$ for $\eta^*(\cdot)$ is determined by minimizing the empirical risk [37], that is, the loss function summed over the (training) data

$$\hat{\eta}(\tilde{\boldsymbol{x}}) = \underset{\eta(\tilde{\boldsymbol{x}})}{\operatorname{argmin}} \sum_{i=1}^{n} \rho(y_i, \eta(\tilde{\boldsymbol{x}}_i)) \tag{3.2}$$

A solution for (3.2) can be derived efficiently using a gradient descent algorithm, which considers the steepest descent of the loss function to determine iteratively an estimate $\hat{\eta}(\cdot)$. This procedure may be combined with a stagewise inclusion of single base-learners into the model, which is of particular interest in statistical modelling [34].

3.2 Component-Wise Functional Gradient Descent Boosting

If we interpret $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_n)^t = (\eta(\tilde{x}_1), \ldots, \eta(\tilde{x}_n))^t$ as n-dimensional parameter vector obtained by applying the additive predictor $\eta(\cdot)$ on the data points $\tilde{\boldsymbol{x}}_i$ [30, 35], problem (3.2) can be seen as searching for the minimizing vector of parameters

$$\hat{\boldsymbol{\eta}} = \underset{\boldsymbol{\eta}}{\operatorname{argmin}} \sum_{i=1}^{n} \rho(y_i, \eta_i) \tag{3.3}$$

In each iteration m, the negative gradient of the loss function evaluated at the current parameter vector $\hat{\eta}^{[m-1]}(\tilde{\boldsymbol{x}}_i)$ (the estimate obtained in the previous iteration) is derived. This results in an $n \times 1$ dimensional gradient vector $\boldsymbol{u}^{[m]} = (u_i^{[m]}, \ldots, u_n^{[m]})$ with entries

$$u_i^{[m]} = -\frac{\delta \rho(y_i, \eta)}{\delta \eta}\bigg|_{\eta = \hat{\eta}^{[m-1]}(\tilde{\boldsymbol{x}}_i)} \tag{3.4}$$

for $i = 1, \ldots, n$. The estimate of the additive predictor is initialized by a starting value $\hat{\boldsymbol{\eta}} = \hat{\boldsymbol{\eta}}^{[0]}$, such as $\hat{\eta}_i{}^{[0]} = 0$ for $i = 1, \ldots, n$ [35], and updated in each iteration according to the steepest descent of the loss function. In gradient descent boosting, this update is given by

$$\hat{\boldsymbol{\eta}}^{[m]} = \hat{\boldsymbol{\eta}}^{[m-1]} + \nu^{[m]} \boldsymbol{u}^{[m]} \tag{3.5}$$

where $\nu^{[m]}$ denotes the step length for the update in iteration m. A suitable choice will be discussed below.

Functional Gradient Descent: The updating step (3.5) may be adjusted such that it makes use of the steepest descent direction, but at the same time connects the update to the desired class of functions of the covariates given by the base-learners [36]. This link is established by fitting a base-learner to the negative gradient of the loss function, e.g. via least squares estimation [30]. The result is a constrained estimate

$$\hat{\boldsymbol{u}}^{[m]} = (\hat{u}_1^{[m]}, \ldots, \hat{u}_n^{[m]}) = (\hat{f}^{[m]}(\tilde{\boldsymbol{x}}_1), \ldots, \hat{f}^{[m]}(\tilde{\boldsymbol{x}}_n)) = \hat{\boldsymbol{f}} \tag{3.6}$$

of the steepest descent direction, in which $\hat{f}(\cdot)$ denotes the fitted base-learner. By making use of $\hat{\boldsymbol{f}}$ instead of the negative gradient $\boldsymbol{u}^{[m]}$ directly, the update in iteration m is changed to

$$\hat{\boldsymbol{\eta}}^{[m]} = \hat{\boldsymbol{\eta}}^{[m-1]} + \nu^{[m]} \hat{\boldsymbol{f}} \tag{3.7}$$

which is known as functional gradient boosting.

Component-Wise Boosting: A single function $f(\cdot)$ was considered above to estimate the negative gradient of the loss function. However, the inclusion of multiple base-learners in functional gradient boosting is possible and often desired, as it allows for a component-wise approach facilitating variable selection. Bühlmann and Yu introduced the concept of component-wise functional gradient boosting [37]. It differs from the above outlined approach as it fits each base-learner $f_j(\cdot)$ seperately to the negative gradient. This results in estimates $\hat{\boldsymbol{f}}_j$, $j = 1, \ldots, J$. The best fitting base-learner $\hat{\boldsymbol{f}}_{j^*}$ is determined via

$$j^* = \underset{j}{\operatorname{argmin}} \sum_{i=1}^{n} (u_i^{[m]} - \hat{f}_j(\tilde{\boldsymbol{x}}_i))^2 \tag{3.8}$$

as the one minimizing the residual sum of squares. In each iteration, the identified $\hat{\boldsymbol{f}}_{j^*}$ is added to the current estimate of the additive predictor according to

$$\hat{\boldsymbol{\eta}}^{[m]} = \hat{\boldsymbol{\eta}}^{[m-1]} + \nu^{[m]} \hat{\boldsymbol{f}}_{j^*} \tag{3.9}$$

in a stagewise fashion, leaving previously added function estimates unchanged [30]. In each iteration, a single base-learner, multiplied by the step length, is incorporated into the model. However, repeated selection of the identical base-learner is possible and will lead to an increased weight of the corresponding function in the estimate of $\hat{\eta}(\cdot)$. Thus, the final additive predictor is a weighted sum over all base-learners selected in at least one iteration.

Since different base-learners typically depend on differing subsets of the considered variables, not selecting a particular base-learner indicates the exclusion of the respective variables from the model. Hence, the sufficiently (but not too) early stopping of the procedure automatically leads to variable selection. The algorithm then returns a prediction model for the trait of interest and simultaneously identifies the most influential variables during model estimation [34].

Choice of Parameters: The maximum number of iterations, m_{stop}, is an important tuning parameter of the algorithm. Additional iterations usually decrease the training risk. However, this may lead to overfitting [30]. This phenomenon occurs if the training data are fitted to such an extent that the determined predictor performs poorly for new observations. A well advised choice of m_{stop} is crucial to prevent overfitting [34]. An optimum number for m_{stop} may be determined in a single dataset by use of cross-validation techniques. Herein, the data are repeatedly divided into training and test samples and subsequently used in parts to fit (training data) and validate (test data) the model. The optimum m_{stop} is the parameter leading to the lowest empirical risk on the data [34].

The number of iterations is influenced by the step length $\nu^{[m]}$ employed in the updating step of the algorithm. For $0 < \nu < 1$, the step length is a shrinkage factor scaling the contribution of each incorporated base-learner [30]. One way to derive an appropriate value for $\nu^{[m]}$ in a gradient descent approach is to define it as the minimizer

$$\nu^{[m]} = \underset{\nu}{\operatorname{argmin}}\, \rho(y, \hat{\eta}^{[m-1]} + \nu u^{[m]}) \tag{3.10}$$

in each iteration step. The step length $\nu^{[m]}$ can be understood as learning rate of the procedure. It has been found empirically that smaller values ($\nu \leq 0.1$) are favourable, as they improve the algorithm's performance considerably as compared to no shrinkage ($\nu = 1$) [30, 36]. Decreasing the step length, however, leads to a higher number of performed iterations and thus increases the computational burden for the algorithm. In practice, (3.10) does not have to be derived in every iteration. Instead, a small constant may be chosen for ν. A useful default value is setting $\nu = 0.1$ [34].

Data Focus: Boosting algorithms, as mentioned above, focus on the observations most difficult to classify. In traditional classification algorithms, such as *AdaBoost*, the data are re-

weighted in every step. Previously incorrectly classified observations are upweighted, while those correctly classified are downweighted by iteratively assigning more influence to the difficult observations [30].

Gradient descent boosting implicitly shifts the focus on the more challenging measurements by considering the gradient of the loss function instead. This may be regarded as fitting the errors of the previous iteration [34] and can best be seen by looking at an exemplary loss function, such as the commonly used squared error loss $\rho(y, \eta(x)) = (y - \eta(x))^2$. Here, the derived negative gradient is equal to $2(y - \hat{\eta}(x))$, basically leading to re-fitting of the residuals.

3.3 Boosting with Kernel Functions as Base-Learner

We developed a novel kernel boosting approach, integrating a kernel function as base-learner into the functional gradient descent algorithm. We consider an additive predictor incorporating the influence of environmental covariates as well as genetic effects to model the logit of the probability of being a case. Mimicking the setup in the logistic kernel machine approach, we model the effect of n_q environmental covariates in the transposed $n_q \times 1$ vector $\boldsymbol{x}_i^t$ with corresponding coefficient vector $\boldsymbol{\beta}$ parametrically, while we incorporate the $n_s \times 1$ dimensional genotype vector $\boldsymbol{z}_i$ for the n_s SNPs in the SNP set under investigation via a kernel function. Note that $n_q + n_s = q$ (as above). Assuming we have P different SNP sets, each representing a particular pathway, this results in a model

$$logit(P(y_i = 1)) = \eta(\boldsymbol{x}_i, \boldsymbol{z}_i) = \boldsymbol{x}_i^t \boldsymbol{\beta} + \sum_{p=1}^{P} f_p(\boldsymbol{z}_i) \tag{3.11}$$

where y_i denotes the case-control status of individual i. Considering the matrix $\boldsymbol{Z}_p$ of all individuals' genotypes for the SNP set representing pathway p, the kernel function base-learner $f_p(\cdot)$ is equal to

$$f_p(\boldsymbol{Z}_p) = \boldsymbol{K}_p \boldsymbol{\gamma} = \boldsymbol{Z}_p \boldsymbol{A}_p \boldsymbol{N}_p \boldsymbol{A}_p^t \boldsymbol{Z}_p^t \boldsymbol{\gamma}, \tag{3.12}$$

where $\boldsymbol{A}_p$ and $\boldsymbol{N}_p$ denote the adjacency and network matrix as introduced in Chapter 2, respectively, and $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_n)^t$ is the coefficient vector. In practice, we use an additional smoothness constraint and estimate $\boldsymbol{\gamma}$ via penalized least squares with outcome $\boldsymbol{u}^{[m]}$.

By defining a separate kernel base-learner for each pathway, component-wise functional gradient boosting on GWAS data leads to the identification of a set of pathways most influential on disease risk. A prediction model composed of the pathways selected may be used to predict the disease status of further individuals. Note that owing to the usual lack of a separate validation dataset, the prediction accuracy of the model is optimized by cross-validation inside the same sample.

Kernel boosting, unlike testing procedures frequently employed in the investigation of GWAS data, does not compute a mere p value for the investigated effects, but instead creates a prediction model for the trait of interest, while simultaneously reducing the set of candidate pathways. Further information on the kernel boosting approach may be found in Summary 5.3 and [38].

Examples of Application

Three real-world datasets were considered in this work to investigate the performance of both newly developed and existing methods in the analysis of genetic data. Here follows a short description of the three datasets employed.

4.1 Lung Cancer

Cancers are complex diseases which are frequently analysed in the framework of GWAS studies. Lung cancer, one of the most common and severe forms, especially in industrialized nations, is responsible for the greatest proportion of deaths caused by cancer worldwide [39]. Although one of the major risk factors is tobacco exposure, a number of genetic influences have already been revealed by many studies [40]. Nevertheless, the heritability of the disease still remains to be explained fully, as all the genetic factors contributing to the risk of developing lung cancer have not been completely elucidated so far.

The German Lung Cancer GWAS consists of 488 lung cancer patients and 478 controls, resulting from the combination of participants in three individual studies. These studies comprise Lung Cancer in the Young (LUCY), a population-based multicentre study carried out by the University Medical Centre in Göttingen and the Helmholtz Zentrum München. Here, a total of 847 lung cancer patients under the age of 51 and 5,524 family members were recruited in 31 German hospitals until 2011 [41, 42]. The Heidelberg Lung Cancer Case-Control Study was conducted by the Thoraxklinik in Heidelberg and the German Cancer Research Center (DKFZ) [43]. More than 2000 cases and 750 controls have been recruited in an on-going hospital-based study since 1997. The third study, Cooperative Health Research in the Augsburg Region (KORA) [44], is a population-based genome-wide study on more than 18,000 participants. It was carried out by the Helmholtz Zentrum München between 1984 and 2001. A subset of the individuals considered in these studies were genotyped on a HumanHap 550K SNP chip and form the German Lung Cancer GWAS.

4.2 Rheumatoid Arthritis

One of the most common inflammatory diseases of the joints is rheumatoid arthritis. It is one of the major causes of disability and is known to be strongly influenced by genetic factors.

The human leukocyte antigen (HLA) region located on chromosome 6 was revealed as highly associated with rheumatoid arthritis disease susceptibility [45, 46].

We investigated a GWAS study conducted by the North American Rheumatoid Arthritis Consortium (NARAC). Eight-hundred sixty-eight rheumatoid arthritis cases were collected along with 1,194 controls matching the self-reported ethnic background. All the cases were recruited from hospitals located in New York with rheumatoid arthritis being diagnosed according to the criteria of the American College of Rheumatology. The study participants were genotyped with the HumanHap500v1 array [47, 48].

4.3 San Antonio Family Studies

The Genetic Analysis Workshops intend to encourage the development, testing, and discussion of new statistical methods in the analysis of genetic data. The family dataset distributed in the context of Genetic Analysis Workshop 19 (GAW19) was taken from the Type 2 Diabetes Genetic Exploration by Next-generation sequencing in multi-Ethnic Samples (T2D-GENES) Project 2, a pedigree-based study aiming towards the detection of variants influencing the susceptibility of developing type-2 diabetes. The project involves 20 Mexican-American families, recruited in the San Antonio Family Heart Study (SAFHS) and the San Antonio Family Diabetes/Gallbladder Study (SAFDGS). SAFHS examined Mexican Americans aged 40 to 60 years, randomly selected with regard to disease, and their family members for 20 years, beginning in 1991. Starting in the same year, SAFDGS recruited Mexican Americans with diagnosed type-2 diabetes and their relatives for three examinations per person.

Whole-genome sequencing on the 20 pedigrees selected for T2D-GENES Project 2 was performed at Complete Genomics Inc, with 464 individuals passing quality control filters. Sequence data of these 464 family members and imputed GWAS data for an additional 495 individuals from the 20 families were provided for the odd-numbered autosomes in GAW19. GWAS data for the second group of individuals were obtained from several Illumina chips, including: HumanHap550v3, HumanExon510Sv1, Human660W-Quadv1, Human1Mv1, and Human1M-Duov3 [49]. Phenotype data available in GAW19 included systolic and diastolic blood pressure measurements, longitudinally collected at up to four points in time, along with information on age, sex, year of examination, use of antihypertensives, hypertension diagnosis, and smoking behaviour. Simulated blood pressure measurements were based on real sequence data and pedigree information and available for 200 replicates. A total of 245 genes were simulated as having an effect on blood pressure traits, with varying effect strengths [50].

Summaries

5.1 Comparing Strategies for Combined Testing of Rare and Common Variants in Whole Sequence and Genome-Wide Genotype Data

The overall association of a SNP set with a phenotype may be evaluated efficiently with the kernel score test. It is capable of analysing genotype data from rare and common markers alike, making it suitable for the analysis of GWAS as well as sequence data. Multiple kernel functions for the transformation of genotypes into a genetic similarity matrix have been proposed, differing amongst other things in their ability to integrate interactions. Upon calculation of the kernel matrix, optional weights can be assigned to individual SNPs. An upweighting of rare markers, here denoting SNPs with a minor allele frequency below 5%, allows them a greater overall contribution to the analysis results.

We evaluated several strategies for testing the association of a SNP set representing a gene, taking into account marker density, frequency, different weighting schemes, several SNP-set definitions, and two kernel choices. More precisely, we investigated the following questions:

- Does a higher density of available markers lead to an increase in power?
- Can linkage disequilibrium information be used to improve the definition of SNP sets representing genes? What impact does the set definition have on analysis results?
- What is the best strategy for joint investigation of SNPs with differing minor allele frequencies? Are SNP weights capable of facilitating the detection of association signals?
- What influence does the choice of kernel function have?

We investigated these questions using real and simulated systolic blood pressure (SBP) data taken from the family dataset distributed in the context of Genetic Analysis Workshop 19. Considered were 706 individuals in real data and between 740 and 781 individuals in simulations, dependent on the data quality and number of missing phenotypes. We concentrated on the chosen candidate genes, *AGTR1* on chromosome 3 in the real data, known to be influential

on SBP in the family sample, as well as *MAP4*, *TNN*, *LEPR*, *GSN*, and *FLT3* in simulated data, with varying LD patterns.

Four SNP-set definitions, based on differently sized genetic regions, were employed to represent the gene *AGTR1*. More precisely, we considered boundaries defined by first and last exonic position, these limits enlarged by 30 kbp and 500 kbp flanking regions, as well as a SNP set based on LD blocks. These LD blocks were calculated in the software package Haploview [51], with the help of a Hapmap [52] reference sample for Mexican-American ancestry. LD blocks were determined with Haploview's default algorithm, in which a pair of SNPs is defined to be in *strong LD*, if the 95% confidence bounds on D' exceed certain limits [53]. We set greater than 0.8 for the upper and 0.5 for the lower confidence interval limits and declared a region to be an LD block if at least 70% of pairwise comparisons among SNPs were categorized as *strong LD*. If a gene boundary overlapped with an LD block, we enlarged the gene limit to the extent of the corresponding LD block. On analysis of the simulated data, all gene-representing SNP sets were based on the LD-block definition.

The influence of genetic information on the trait was evaluated in the kernel score test for family data [54]. We compared the results obtained for a linear kernel, only modelling additive effects of the included SNPs, and a multiplicative kernel, considering interactions between markers in addition. Environmental covariate information on age and sex was included. It is possible either to use all SNPs in a considered set jointly to calculate a kernel matrix for the score test, or split SNPs according to their MAF into rare and common markers. The first way of proceeding results in a single test statistic and a corresponding p value. The latter approach results in two separate kernel matrices with two corresponding test statistics. These statistics may be combined by the weighted sum approach to form a composite statistic, for which one p value reflecting the effect of all markers can be derived. Alternatively, a p value for each of the two test statistics may be calculated and combined via Fisher's p-value pooling. We further examined the impact of three differing SNP weighting schemes. We considered equal weights for all markers, weights according to the inverse MAF and weights based on the beta distribution. The latter two approaches upweight rare marker alleles, with beta weights distinguishing MAFs more moderately. All analyses were conducted on both sequence and GWAS data.

The most important result is that the LD-block-based SNP-set definition had the highest power to identify associations. It should thus be preferred in SNP-set analysis, as the kernel exploits correlations between markers within the SNP set. The collective evaluation of rare and common SNPs yielded better results than the investigation of only one of the groups. As power of all joint tests was very similar, the approach with which the two types of marker were combined had little effect. In most cases, the analysis of sequence data was more powerful compared to that of GWAS data. Inverse MAF weights can improve performance for

common markers, however, they must be used with caution for rare markers. The kernel choice had almost zero effect on the analysis of single genes. Please refer to [55] for further details on the methods investigated and the results.

5.2 Filtering Genetic Variants and Placing Informative Priors Based on Putative Biological Function

Given the increase in marker density in genetic data currently, even large studies may be underpowered to detect association signals. One possible way of countering this problem is to incorporate additional biological knowledge into the analysis. Nowadays, a multitude of tools and databases for this purpose are available, offering an increasing amount of biologically meaningful information on genetic markers. This article summarizes different approaches to filtering, prioritising, and grouping SNPs which were contributed by the members of the GAW19 discussion group *Filtering variants and placing informative priors*. All analyses were carried out on the genotype data of families or unrelated individuals, represented in minor allele count coding for GWAS and sequence markers. Associations with real and simulated blood pressure traits were evaluated, mostly employing regression approaches. Furthermore, it was demonstrated how improvements in grouping and filtering of markers can noticeably improve power in the evaluation of genotype data, and that the incorporation of additional knowledge can facilitate the detection of associations. Questions addressed include:

- How can additional biological knowledge be integrated into the analysis of genotype data and what impact does it have on power?
- From which sources may additional biological information be obtained and to what extent does it differ between databases?
- Which strategies for filtering and grouping SNPs are beneficial and which weighting schemes for SNPs, test statistics, or p values may be used?
- How can functional or structural information on markers be considered on evaluation of p values?

Today, different databases and software tools are available to annotate both the location as well as the function of a SNP, allowing various filtering and grouping strategies. Investigated SNPs may be restricted to markers with known or supposed biological function, such as regulatory influence, or aggregated to represent a gene, exon, or other genomic unit of interest. Among the contributions of the discussion group, two powerful regression frameworks for SNP-set testing were considered: Firstly, tests of the burden type, which transform minor

allele counts over a set of markers into a score for each individual. Association between this score and the trait is evaluated. Secondly, the sequence kernel association test (SKAT; LKMT is a test of SKAT type in the logistic context, see Section 2.1), that transforms all genotypes into a matrix with entries reflecting the genetic similarity of any pair of individuals in the sample. Association is evaluated based on the variance component of the genetic effect. Within both frameworks, weights reflecting additional knowledge may be assigned to single SNPs. Typical weighting schemes are based on MAFs or functional importance: For example, rare markers may be upweighted to allow them more overall contribution, or several SNPs may be assigned higher scores reflecting their regulatory importance. Alternatively, weights can be incorporated upon combination of different test statistics derived within a SKAT or burden framework. These test statistics may originate from different SNP sets, as well as be calculated by varying methods. For example, separate kernel matrices (optionally incorporating weights on single markers) can be used for rare and common SNPs, or SKAT and burden test statistics may be combined. Moreover, the evaluation of p values may consider functional or structural knowledge on the investigated SNP set. This is possible by weighting p values or by adjusting the significance level according to the number of independent tests, considering LD among investigated markers. Depending on the method employed, correlations between markers may either be exploited in a suitable way or have to be accounted for in order to prevent distortion of the results.

The results of analysis differed in the group, which may be explained to a large extent by varying choices of genetic and phenotypic data. However, the choice of methods was also an influential factor, as the application of varying methods to the same data yielded differing results. The inclusion of biological knowledge generally increased power in the analysis of association studies. Filtering of markers according to functional relevance was particularly useful. However, filtering involves the risk of information loss through the exclusion of influential markers, as the rating of marker functions can vary substantially between different databases. For example, the GAW19 simulations were based on PolyPhen2 functional prediction scores [56], which can differ largely from SIFT [57] and RegulomeDB [58] scores. Thus, the other scores provided non-matching priors for simulated blood pressure traits whenever the database information varied. In real application scenarios, there is no ideal choice of functional annotation, suggesting that one should consider multiple databases jointly. Weights can assist in the detection of associations and had a strong influence on power on SKAT analysis. Furthermore, structural information, as represented in LD patterns on the considered SNPs, had an effect on the results. Kernel methods may benefit from the consideration of LD patterns by exploiting correlations and therefore should be calculated on LD blocks. Correlations also play a role in the calculation of the significance level, where the effective number of independent tests may be determined using a beta distribution. The according adjustment of the significance level leads to a strong reduction in the multiple testing burden. Thus, it is

desirable to involve filtering, grouping, and weighting of SNPs, as well as an adjustment of the significance level in combination, in order to reach the highest possible power in analysis. More details on the results of the discussion group may be found in [59].

5.3 Pathway-Based Kernel Boosting for the Analysis of Genome-Wide Association Studies

The investigation of pathway-representing SNP sets benefits from the advantages of set-evaluation approaches. This includes lowering the multiple testing burden, facilitating the detection of moderate effects, and assisting the biological interpretation of the results. Although a suitable tool for the analysis of single pathways, kernel methods cannot account for correlations between SNP sets and thus lack the ability to discriminate biological mechanisms influencing disease risk from isolated effects included in a pathway resulting from gene overlap. Furthermore, they can only evaluate a pathway's impact and do not offer any trait prediction. With the limitations of single pathway methods and the benefits resulting from the simultaneous analysis of genetic information in mind, we aimed towards the development of a new approach, enabling the mutual analysis of multiple pathway-representing SNP sets. We intended to detect associations, while at the same time create a prediction model for the considered trait, based on the pathways identified as influential. To this purpose, we integrated kernel functions into a boosting algorithm. Our project had the following specific aims:

- Develop a new method to enable the joint analysis of multiple pathways, building upon the kernel-based pathway test and maintaining its beneficial properties.
- Enable the prediction of disease status based on pathways identified as being influential.
- Ensure flexibility in the approach in terms of included data, such that additional genetic information or environmental covariates may be considered.
- Make sure that additional variables can either be subjected to the boosting algorithm or included as mandatory effects in the model.

Our method integrates two existing, powerful approaches: The LKMT and the functional gradient descent boosting algorithm. We chose to include the network kernel function as base-learner in the boosting algorithm, as it allows for interactions between markers in the considered SNP set and may incorporate topological information on the pathway.

We evaluated the performance of the method on simulated genotype data for SNPs representing 50 randomly selected real-world pathways in existence obtained from the Kyoto Encyclopedia of Genes and Genomes (KEGG) database. Six effect scenarios, differing in

sample size and effect strengths, as well as non-informative genetic data (null case) were considered. Each sample included equal numbers of cases and controls and was simulated for 100 replications. Two pathways were chosen to each include three interconnected effect genes, with two influential SNPs per gene. All effects were simulated as additive and equally strong for all associated SNPs. Owing to the adaptation of real pathway structures, influential genes overlapped into six further pathways.

As examples of application, we investigated a lung cancer and a rheumatoid arthritis study. Here, 73 pathways with known connection to human diseases were analysed including relevant environmental covariates. These covariates were incorporated as mandatory variables in a starting model for the algorithm, to prevent them from being subjected to the boosting algorithm and compete with the investigated pathways. All kernel boosting results were compared to the results given by LKMT single-pathway tests on the same data.

Kernel boosting performed reliably in terms of false positive results: The application on non-informative genotype data resulted in few, randomly distributed selections, not suggestive of any falsely detected association signal. The power to identify our causal pathways was very high for both considered methods in scenarios including more individuals and stronger effects. However, while the LKMT also detected pathways including any effect gene with a high probability, the multivariable kernel boosting approach was able to discriminate pathways with highest explanatory power from those pathways including isolated effects owing to gene overlaps. Association signals were less clearly identified by both methods in scenarios with smaller samples or weaker effects, with a more pronounced drop in power for the LKMT. By overcoming the multiple testing burden, kernel boosting may have greater potential to identify associations when the LKMT is underpowered. Analysis of the lung cancer dataset resulted in the selection of one pathway by kernel boosting and no significant result by the LKMT. In the rheumatoid arthritis dataset, p values for 46 pathways were significant in single-pathway tests, while kernel boosting narrowed down the number of identified pathways to 32, providing a better basis for understanding the biological processes involved in disease risk. Overall, we believe that kernel boosting constitutes a valuable and highly flexible approach in GWAS analysis, capable of incorporating various datatypes and predicting clinical outcomes. For more details on this method, please refer to [38].

5.4 Kangar00: Kernel Approaches for Nonlinear Genetic Association Regression

Multiple kernel functions for the analysis of pathway-representing SNP sets were developed, which are able to integrate differing extents of biologically relevant information from databases online. The incorporation of additional data on a pathway usually requires a number of data preparation steps. Processed database information and genotype data for a SNP

set of interest can be used to form a kernel matrix, with which the influence of a considered pathway on disease risk may be evaluated. To our knowledge up until the time of publication, no software package for the kernel-based analysis of genetic data covering all functionalities ranging from the download of required genetic information to the calculation of p values reflecting the effect of a pathway on a trait had been available. We implemented all the necessary functionalities in the R-package kangar00, which furthermore enables the application of our newly developed kernel boosting approach (see Section 5.3). The key functions included in the kangar00 package are the following:

- The downloading and processing of biological information on pathways, genes, and SNPs from online databases, including the annotation of SNPs via genes to pathways.
- Evaluation of the effect of a pathway on disease risk through a logistic model framework via integration of biological knowledge and genotype information.
- A choice of options: Three kernel functions implemented, along with two methods for p-value calculation for the corresponding test statistic.
- Flexibility: Straightforward incorporation of covariates of differing types and easy adaptation to the analysis of genes, LD blocks, or other genomic units.
- Interface to the R-package mboost, providing the ability to run kernel boosting and creating a prediction model for a binary outcome based on the pathways selected.

Kangar00 offers several functions for the extraction and further processing of biological data from online databases. For any chosen pathway of interest, all necessary information needed to evaluate its effect on a binary outcome can be obtained, considering different levels of structural information on the gene-interaction pattern. Information on the pathway itself is downloaded from the Kyoto Encyclopedia of Genes and Genomes (KEGG) [60] database, requiring the pathway to be identified by a KEGG id-number. Information on genes contained in the pathway is obtained from the Ensembl [61] database. Listed are the gene names as well as transcript start and end positions in the latest genomic build and the corresponding chromosomes. SNPs contained in the considered GWAS study are mapped to those genes via base pair positions in order to form pathway-representing SNP sets. In the case of non-matching genomic builds for SNP and gene locations, SNP positions may be optionally updated to the current build used in the Ensembl database, in order to match the considered gene boundaries. Interaction information on the pathway is converted into a quadratic adjacency matrix representing the network structure. In the process, existing connections between genes are categorized as either activating or inhibiting type. Furthermore, effect directions can be optionally represented in the adjacency matrix.

In kangar00, the influence of a pathway on the binary case-control status is evaluated within the LKMT. The corresponding test statistic includes the kernel matrix, interpretable as genetic similarity matrix, derived from the genotypes of the particular pathway-representing SNP set. Our package offers a choice of three kernel functions, the linear kernel and the network kernel (see Section 2.3), as well as a multiplicative, size-adjusted kernel, each differing in their ability to incorporate pathway characteristics upon calculation of the matrix. For more information on the two more advanced kernels, refer to [18] and [20], respectively. The resulting test statistic follows a mixture of χ^2-distributions for which a p value can be derived based on the Satterthwaite approximation or Davies' algorithm [25]. Both options are implemented within Kangar00.

Furthermore, kangar00 provides all the functions necessary to establish an interface to mboost [62] and apply kernel boosting on a GWAS dataset under consideration. In particular, functions relating to prediction of the case-control status based on selected pathways are available. In addition, multiple options are available to visualize data, including a function to plot pathway networks. All the functionalities are described in the package documentation, with an explanation of their usage being given in the package vignette. The package can be downloaded free of charge from the R repository CRAN at the URL given in [63].

Discussion

Kernel methods constitute a well-established approach toward the analysis of biologically meaningful sets of SNPs. They possess many characteristics desirable when investigating genes or pathways. However, room for improvement still remains by considering additional biological knowledge, or integrating kernel methods into a powerful analysis framework. We examined how the performance of kernel-based SNP-set tests, such as the LKMT, is related to the definition of a SNP set. In particular, it has been demonstrated that defining SNP sets based on the boundaries of LD blocks is beneficial, as kernel methods can exploit correlations among markers. This approach resulted in the improved detection of association signals and was therefore implemented in all the following analyses.

In addition to correlations within a SNP set, dependencies may also occur between different sets of SNPs. This can be caused by LD between SNPs in different sets, as well as by the same markers being included in several sets. The latter can for example occur in the analysis of pathway-representing SNP sets, in which case this results from gene overlaps. In contrast to single-pathway testing methods, the joint evaluation of pathways in a selection framework facilitates the consideration of correlations between pathways. From a biological point of view, this allows us to discriminate causal biological mechanisms influential on disease risk from isolated, overlapping effect genes. This constitutes an important step towards the understanding of the genetic effects involved in disease susceptibility. Our newly developed kernel boosting approach is specifically designed to account for correlations between pathways to enable this discrimination. Furthermore, the approach creates a prediction model for the disease status based on the pathways identified as influential on the trait. Our newly developed R-package kangar00 constitutes a convenient way to perform single-pathway testing and, in collaboration with the R-package mboost, provides all the functionalities necessary to run kernel boosting. The automatic pipeline for the download and processing of required pathway, gene, and SNP information simplifies the retrieval of the latest database information, accounting for the fact that databases are always a work in progress and provide ever-changing biological knowledge.

Owing to the mutual evaluation of several pathways and the cross-validation performed by the algorithm, kernel boosting is computationally rather demanding. It is however fea-

sible on current high-performance computing cluster systems for single GWAS studies, but not (yet) possible for larger datasets, as are available nowadays from data-sharing consortia or considered in meta-analysis. However, I am convinced that this limitation will play an increasingly diminishing role in the future, as hardware systems continue to improve in both performance and capacity.

I believe that boosting approaches and their capability of dealing with high-dimensional data resulting in interpretable prediction models will be of great interest in the analysis of genetic data in the future. Their flexibility, especially with respect to the incorporation of various types of data simultaneously, gives me strong reason to be confident that they will be of particular use in a large number of application scenarios. I trust that kernel methods will further assist towards the elucidation of the genetic influence on disease in a general shift from a framework of simply testing single pathways currently toward selection and prediction in the context of powerful boosting approaches.

Bibliography

[1] Laird NM, Lange C. The Fundamentals of Modern Statistical Genetics. Statistics for Biology and Health. Springer-Verlag New York; 2011.

[2] Zheng G, Yang Y, Zhu X, Elston RC. Analysis of Genetic Association Studies. Statistics for Biology and Health. Springer US; 2012.

[3] Ziegler A, König IR, Pahlke F. A Statistical Approach to Genetic Epidemiology: Concepts and Applications. Wiley-Blackwell; 2010. 2nd Edition.

[4] Bickeböller H, Fischer C. Einführung in die Genetische Epidemiologie. Statistik und ihre Anwendungen. Springer Berlin Heidelberg; 2007.

[5] Palmer LJ, Burton PR, Smith GD. An introduction to genetic epidemiology. Health and Society. Policy Press; 2011.

[6] Stram DO. Design, Analysis, and interpretation of Genome-Wide Association Scans. Statistics for Biology and Health. Springer-Verlag New York; 2014.

[7] Lewontin RC. The Interaction of Selection and Linkage. I. General Considerations; Heterotic Models. Genetics. 1964;49(1):49–67.

[8] García-Campos MA, Espinal-Enríquez J, Hernández-Lemus E. Pathway Analysis: State of the Art. Frontiers in Physiology. 2015;6:383.

[9] Cork JM, Purugganan MD. The evolution of molecular genetic pathways and networks. Bioessays. 2004;26(5):479–484.

[10] Wu MC, Kraft P, Epstein MP, Taylor DM, Chanock SJ, Hunter DJ, et al. Powerful SNP-Set Analysis for Case-Control Genome-wide Association Studies. American Journal of Human Genetics. 2010;86(6):929–942.

[11] Ackermann M, Strimmer K. A general modular framework for gene set enrichment analysis. BMC Bioinformatics. 2009;10:47.

[12] Maciejewski H. Gene set analysis methods: statistical models and methodological differences. Briefings in Bioinformatics. 2014;15(4):504–518.

[13] Nam D, Kim SY. Gene-set approach for expression pattern analysis. Briefings in Bioinformatics. 2008;9(3):189–197.

[14] Khatri P, Sirota M, Butte AJ. Ten Years of Pathway Analysis: Current Approaches and Outstanding Challenges. PLoS Computational Biology. 2012;8(2):e1002375.

[15] Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences of the United States of America. 2005;102(43):15545–15550.

[16] Liu D, Lin X, Ghosh D. Semiparametric Regression of Multidimensional Genetic Pathway Data: Least-Squares Kernel Machines and Linear Mixed Models. Biometrics. 2007;63(4):1079–1088.

[17] Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. Rare-Variant Association Testing for Sequencing Data with the Sequence Kernel Association Test. American Journal of Human Genetics. 2011;89(1):82–93.

[18] Freytag S, Manitz J, Schlather M, Kneib T, Amos CI, Risch A, et al. A Network-Based Kernel Machine Test for the Identification of Risk Pathways in Genome-Wide Association Studies. Human Heredity. 2013;76(2):64–75.

[19] De Leeuw CA, Neale BM, Heskes T, Posthuma D. The statistical properties of gene-set analysis. Nature Reviews Genetics. 2016;17(6):353–364.

[20] Freytag S, Bickeböller H, Amos CI, 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.

[21] Liu D, Ghosh D, Lin X. 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. 2008;9:292.

[22] Fahrmeir L, Kneib T, Lang S. Regression: Modelle, Methoden und Anwendungen. Statistik und ihre Anwendungen. Springer Berlin Heidelberg; 2009.

[23] Zhang D, Lin X. Hypothesis testing in semiparametric additive mixed models. Biostatistics. 2003;4(1):57–74.

[24] Satterthwaite FE. An approximate distribution of estimates of variance components. Biometrics Bulletin. 1946;2(6):110–114.

[25] Davies RB. Algorithm AS 155: The distribution of a linear combination of χ 2 random variables. Journal of the Royal Statistical Society Series C (Applied Statistics). 1980;29(3):323–333.

[26] Duchesne P, De Micheaux PL. Computing the distribution of quadratic forms: Further comparisons between the Liu–Tang–Zhang approximation and exact methods. Computational Statistics & Data Analysis. 2010;54(4):858–862.

[27] Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K. KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Research. 2017;45(D1):D353–D361.

[28] Bader GD, Cary MP, Sander C. Pathguide: a pathway resource list. Nucleic Acids Research. 2006;34(suppl_1):D504–D506.

[29] Stobbe MD, Houten SM, Jansen GA, van Kampen AH, Moerland PD. Critical assessment of human metabolic pathway databases: a stepping stone for future integration. BMC Systems Biology. 2011;5(1):165.

[30] Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning. Springer Series in Statistics. Springer-Verlag New York; 2009.

[31] Schapire RE. The strength of weak learnability. Machine Learning. 1990;5(2):197–227.

[32] Freund Y, Schapire RE. Experiments with a new boosting algorithm. In: Icml. vol. 96; 1996. p. 148–156.

[33] Freund Y, Schapire RE. A desicion-theoretic generalization of on-line learning and an application to boosting. In: European conference on computational learning theory. Springer; 1995. p. 23–37.

[34] Mayr A, Binder H, Gefeller O, Schmid M. The evolution of boosting algorithms. From machine learning to statistical modelling. Methods of Information in Medicine. 2014;53(6):419–427.

[35] Hofner B. Boosting in structured additive models; 2011. LMU München; http://nbn-resolving.de/urn:nbn:de:bvb:19-138053.

[36] Friedman JH. Greedy function approximation: A gradient boosting machine. The Annals of Statistics. 2001;29(5):1189–1232.

[37] Bühlmann P, Yu B. Boosting with the L 2 loss: regression and classification. Journal of the American Statistical Association. 2003;98(462):324–339.

[38] Friedrichs S, Manitz J, Burger P, Amos CI, Risch A, Chang-Claude J, et al. Pathway-Based Kernel Boosting for the Analysis of Genome-Wide Association Studies. Computational and Mathematical Methods in Medicine. 2017;2017.

[39] Ferlay J, Soerjomataram I, Dikshit R, Eser S, Mathers C, Rebelo M, et al. Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. International Journal of Cancer. 2015;136(5).

[40] Brennan P, Hainaut P, Boffetta P. Genetics of lung-cancer susceptibility. The Lancet Oncology. 2011;12(4):399–408.

[41] Sauter W, Rosenberger A, Beckmann L, Kropp S, Mittelstrass K, Timofeeva M, et al. Matrix metalloproteinase 1 (MMP1) is associated with early-onset lung cancer. Cancer Epidemiology and Prevention Biomarkers. 2008;17(5):1127–1135.

[42] Rosenberger A, Illig T, Korb K, Klopp N, Zietemann V, Wölke G, et al. Do genetic factors protect for early onset lung cancer? A case control study before the age of 50 years. BMC Cancer. 2008;8(1):60.

[43] Dally H, Gassner K, Jäger B, Schmezer P, Spiegelhalder B, Edler L, et al. Myeloperoxidase (MPO) genotype and lung cancer histologic types: the MPO- 463 A allele is associated with reduced risk for small cell lung cancer in smokers. International Journal of Cancer. 2002;102(5):530–535.

[44] Wichmann HE, Gieger C, Illig T, study group M. KORA-gen-resource for population genetics, controls and a broad spectrum of disease phenotypes. Das Gesundheitswesen. 2005;67(S 01):26–30.

[45] Firestein GS. Evolving concepts of rheumatoid arthritis. Nature. 2003;423(6937):356–361.

[46] Raychaudhuri S. Recent advances in the genetics of rheumatoid arthritis. Current Opinion in Rheumatology. 2010;22(2):109–118.

[47] Amos CI, Chen WV, Seldin MF, Remmers EF, Taylor KE, Criswell LA, et al. Data for Genetic Analysis Workshop 16 Problem 1, association analysis of rheumatoid arthritis data. BMC Proceedings. 2009;3(suppl 7):S2.

[48] Plenge RM, Seielstad M, Padyukov L, Lee AT, Remmers EF, Ding B, et al. TRAF1–C5 as a Risk Locus for Rheumatoid Arthritis - A Genomewide Study. New England Journal of Medicine. 2007;357(12):1199–1209.

[49] Almasy L, Dyer TD, Peralta JM, Jun G, Wood AR, Fuchsberger C, et al.; BioMed Central. Data for Genetic Analysis Workshop 18: human whole genome sequence, blood pressure, and simulated phenotypes in extended pedigrees. BMC Proceedings. 2014;8(Suppl 1):S2.

[50] Blangero J, Teslovich TM, Sim X, Almeida MA, Jun G, Dyer TD, et al. Omics-squared: human genomic, transcriptomic and phenotypic data for genetic analysis workshop 19. BMC Proceedings. 2016;10(Suppl 7):20.

[51] Barrett JC, Fry B, Maller J, Daly MJ. Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics. 2004;21(2):263–265.

[52] Gibbs RA, Belmont JW, Hardenbol P, Willis TD, Yu FL, Yang HM, et al. The international HapMap project. Nature. 2003;426(6968):789–796.

[53] Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, et al. The Structure of Haplotype Blocks in the Human Genome. Science. 2002;296(5576):2225–2229.

[54] Malzahn D, Friedrichs S, Rosenberger A, Bickeböller H. Kernel score statistic for dependent data. BMC Proceedings. 2014;8(Suppl 1):S41.

[55] Malzahn D, Friedrichs S, Bickeböller H. Comparing Strategies for Combined Testing of rare and common variants in whole sequence and genome-wide genotype data. BMC Proceedings. 2016;10(Suppl 7):17.

[56] Adzhubei I, Jordan DM, Sunyaev SR. Predicting functional effect of human missense mutations using PolyPhen-2. Current Protocols in Human Genetics. 2013;0 7:Unit7.20.

[57] Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nature Protocols. 2009;4(7):1073–1081.

[58] Boyle AP, Hong EL, Hariharan M, Cheng Y, Schaub MA, Kasowski M, et al. Annotation of functional variation in personal genomes using RegulomeDB. Genome Research. 2012;22(9):1790–1797.

[59] Friedrichs S, Malzahn D, Pugh EW, Almeida M, Liu XQ, Bailey JN. Filtering genetic variants and placing informative priors based on putative biological function. BMC Genetics. 2016;17(Suppl 2):S8.

[60] Kanehisa M, Goto S. KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Research. 2000;28(1):27–30.

[61] Yates A, Akanni W, Amode MR, Barrell D, Billis K, Carvalho-Silva D, et al. Ensembl 2016. Nucleic Acids Research. 2016;44(D1):D710–D716.

[62] Hothorn T, Buehlmann P, Kneib T, Schmid M, Hofner B. mboost: Model-Based Boosting; 2016. R package version 2.6-0. URL: http://CRAN.R-project.org/package=mboost.

[63] Manitz J, Friedrichs S, Burger P, Hofner B, Ha NT, Freytag S, et al.. kangar00: Kernel Approaches for Nonlinear Genetic Association Regression; 2016. R package version 1.0. URL: https://CRAN.R-project.org/package=kangar00.

A References of Original Work

A.1 Articles

Malzahn D, Friedrichs S, Bickeböller H: **Comparing Strategies for Combined Testing of Rare and Common Variants in Whole Sequence and Genome-wide Genotype Data.** *BMC Proceedings* 2016, 10(Suppl 7):17; doi:10.1186/s12919-016-0042-9.
URL: https://doi.org/10.1186/s12919-016-0042-9

Friedrichs S*, Malzahn D*, Pugh EW, Almeida M, Liu XQ, Bailey JN: **Filtering genetic variants and placing informative priors based on putative biological function.** *BMC Genetics* 2016, 17(Suppl 2):S8; doi:10.1186/s12863-015-0313-x.
URL: https://doi.org/10.1186/s12863-015-0313-x. * these authors share first authorship

Friedrichs S, Manitz J, Burger P, Amos CI, Risch A, Chang-Claude J, Wichmann HE, Kneib T, Bickeböller H, Hofner B: **Pathway-Based Kernel Boosting for the Analysis of Genome-Wide Association Studies.** *Computational and Mathematical Methods in Medicine* 2017, vol. 2017; doi:10.1155/2017/6742763.
URL: https://www.hindawi.com/journals/cmmm/2017/6742763/

A.2 Software

Manitz J, Friedrichs S, Burger P, Hofner B, Ha NT, Freytag S, Bickeböller H: **kangar00: Kernel Approaches for Nonlinear Genetic Association Regression.**
URL: https://CRAN.R-project.org/package=kangar00

The Author(s) *BMC Proceedings* 2016, **10**(Suppl 7):17
DOI 10.1186/s12919-016-0042-9

BMC Proceedings

PROCEEDINGS Open Access

Comparing strategies for combined testing of rare and common variants in whole sequence and genome-wide genotype data

Dörthe Malzahn*, Stefanie Friedrichs and Heike Bickeböller

From Genetic Analysis Workshop 19
Vienna, Austria. 24-26 August 2014

Abstract

We used our extension of the kernel score test to family data to analyze real and simulated baseline systolic blood pressure in extended pedigrees. We compared the power for different kernels and for different weightings of genetic markers. Moreover, we compared the power of rare and common markers with 3 strategies for joint testing and on marker panels with different densities. Marker weights had much greater influence on power than the kernel chosen. Inverse minor allele frequency weights often increased power on common markers but could decrease power on rare markers. Furthermore, defining the gene region based on linkage disequilibrium blocks often yielded robust power of joint tests of rare and common markers.

Background

The kernel score test is a global covariate-adjusted multilocus procedure that tests for overall association of sets of markers (see Schaid [1] for a review). This reduces the multiple-testing burden. Tested marker sets can, for example, belong to a pathway or candidate gene. The kernel score test can be applied to common and rare variants alike, as well as to data of genome-wide association studies (GWAS) or sequence data where it is named SKAT (sequence kernel association test). The kernel score test was developed for independent subjects [1]. Recent contributions by others and ourselves [2–6] extended the kernel score test to family data.

The kernel is chosen to describe genetic correlation among subjects. Different kernels have been suggested for genetic epidemiological applications. These kernels differ in whether marker–marker interactions are modeled and how complex the interaction effects may be. A frequent choice is to apply the kernel function on weighted minor allele dosage data (thus using an additive coding of minor allele effects). The dosage weights increase with decreasing minor allele frequency corresponding to the *a priori* assumption that less-frequent variants may have larger effects. Weighting allows rarer variants to contribute more to the overall test despite of their low frequencies.

With appropriate weighting, rare and common variants may be entered together into the kernel for joint testing. Recently however, Ionita-Laza et al. [7] proposed alternatives that can be more powerful. We explored these alternative joint tests on rare and common variants in the Genetic Analysis Workshop 19 (GAW19) family data. Moreover, we compared the power of different marker weights and kernels on sequence and GWAS panels. As we focused on genes, we also explored how size or positioning of a flanking region affects the test power.

Methods

Data

We analyzed baseline systolic blood pressure (SBP) and dosage data in the extended Mexican American pedigrees of the GAW19 family data, which are identical to the Genetic Analysis Workshop 18 data [8]. As before

* Correspondence: dmalzah@gwdg.de
Department of Genetic Epidemiology, University Medical Center, Georg-August University Göttingen, Humboldtallee 32, 37073 Göttingen, Germany

[6], we considered subjects with known baseline SBP and baseline diastolic blood pressure, sex, and age, who were not on blood pressure medication (real SBP: 706 subjects, excluding the first listed monozygotic twin of 2 observed twin pairs; simulated SBP: 740 to 781 subjects, numbers vary for 200 simulated study replicates because of inclusion criteria). For real SBP, we considered candidate gene *AGTR1* [9] on chromosome (chr) 3 that tends to associate with SBP in the present family sample [6]. For simulated SBP, we selected from the simulation answers 5 strongly associated genes with various linkage disequilibrium (LD) structures: *MAP4* (very homogeneous LD, chr3) and, in the order of increasing variability of LD, *TNN* (chr1), *FLT3* (chr13), *LEPR* (chr1), and *GSN* (chr9). We used NCBI build 37, International Haplotype Map Project (HapMap) [10] reference data for Mexican Americans and the default algorithm in Haploview 4.2 [11] with a required fraction of strong LD of 0.7 and confidence interval limits of 0.5 and 0.8 to determine LD-blocks based on the D' measure. Gene regions were defined as the LD-block(s) that contained the gene. For *AGTR1*, we also considered the region from the first to the last exonic position and flanking regions of 30 kb or 500 kb. For the same subjects, we used 2 single-nucleotide polymorphism (SNP) panels: sequence (allele dosage data) and GWAS (allele dosage data reduced to GWAS SNPs). Biallelic SNPs were included for testing if their Hardy-Weinberg equilibrium test p values were equal to or greater than 10^{-5} (rounding imputed dosages for this purpose only) and if at least 7 observations of the minor allele were present in the sample. The latter parallels minimum data requirements in parametric regression.

Kernel score test for family data

Here we briefly summarize our method introduced in [6], denoting vectors and matrices by bold letters. Baseline SBP is right-skewed distributed and was therefore rank-normalized by Blom transformation [12] to standard normally distributed target variables $\mathbf{Y} = (Y_1,\ldots,Y_n)$. $\mathbf{Y}$ depend on fixed covariate effects $\mathbf{b}$ (intercept, age, sex, age × sex interaction), random effects $\mathbf{c}$ that adjust for familial polygenic background, a semiparametric model $\mathbf{h}(\mathbf{G})$ of genetic markers $\mathbf{G}$, and regression residuals $\mathbf{e} \sim N(0,s^2\mathbf{I})$ with residual variance s^2.

$$\mathbf{Y} = \mathbf{X}\mathbf{b}^{\mathrm{T}} + \mathbf{Z}\mathbf{c}^{\mathrm{T}} + \mathbf{h}(\mathbf{G}) + \mathbf{e} \quad (1)$$

$\mathbf{X}$, $\mathbf{Z}$ are the design matrices for fixed covariate effects and random family effects. $\mathbf{h}(\mathbf{G}) = \mathbf{K}\mathbf{a}^{\mathrm{T}}$ depends on a $n \times n$ dimensional kernel matrix $\mathbf{K}$ of genetic similarities between n subjects on markers $\mathbf{G}$, and multivariate normally distributed random effects $\mathbf{a} \sim N(0,\tau\mathbf{K})$ [1]. One tests for a genetic covariance component τ.

The kernel score test is computed from restricted maximum likelihood parameter estimates of the genetic null model (where $\mathbf{h}(\mathbf{G}) = \mathbf{0}$). Thus, the null model estimates fixed covariate effects $\mathbf{b_o}$, random pedigree effects $\mathbf{c_o}$, the variance s^2_{fam} of the polygenic familial component, and the residual variance s^2_o. The null model was adjusted for polygenic familial background based on the kinship coefficient matrix $\Phi_{\mathrm{kin}} = \mathbf{Z}\mathbf{Z}^{\mathrm{T}}$ using R-packages kinship2 and coxme with R-function lmekin. The kernel score test statistic is.

$$Q = \mathbf{R}^{\mathrm{T}}\mathbf{M}\mathbf{R} \quad (2)$$

$\mathbf{R} = \mathbf{P}_o^{1/2}\mathbf{Y}$ are standard normally distributed residuals and matrix $\mathbf{M} = (\mathbf{P}_o^{1/2}\mathbf{K}\ \mathbf{P}_o^{1/2})/2$ incorporates the kernel [6]. $\mathbf{P_o} = \mathbf{V_o^{-1}} - \mathbf{V_o^{-1}}\mathbf{X}(\mathbf{X^T}\mathbf{V_o^{-1}}\mathbf{X})^{-1}\mathbf{X^T}\mathbf{V_o^{-1}}$ is the null projection matrix with $\mathbf{V}_o = s^2_o\mathbf{I} + s^2_{\mathrm{fam}}\mathbf{Z}\mathbf{Z}^{\mathrm{T}}$. The p values for test statistic (2) were calculated by Davies' exact method [13] with the R package CompQuadForm from sample estimates Q and all eigenvalues of matrix $\mathbf{M}$.

Kernels and single-nucleotide polymorphism weights

We applied all kernel functions on allele dosage data $\mathbf{g}_i$, $\mathbf{g}_j$ (for pairs of subjects i, j) on N_{SNP} biallelic SNP markers. The kernel matrix entries are

$$\text{Linear kernel } \mathbf{K}_{ij} = \mathbf{g}_i^{\mathrm{T}}\mathbf{W}\mathbf{g}_j \quad (3)$$

$$\text{Radial basis function (RBF) kernel } \mathbf{K}_{ij} = \exp\left(-\mu^{-1}\cdot\left(\mathbf{g}_i-\mathbf{g}_j\right)^{\mathrm{T}}\mathbf{W}\left(\mathbf{g}_i-\mathbf{g}_j\right)\right) \quad (4)$$

with diagonal weight matrix $\mathbf{W}$. The linear kernel (3) does not allow for SNP interactions opposed to the RBF kernel (4), which yields polynomial models. Dosage weights are normed $\mathbf{W}_{\mathrm{mm}} = f(\nu_m)/\Sigma_m f(\nu_m)$ for any chosen SNP set $m = 1,\ldots,N_{\mathrm{SNP}}$ and depend on the minor allele frequency (MAF) ν of the respective SNP. We considered: $f(\nu_m) = 1$ (treating SNPs alike), $f(\nu_m) = 1/\nu_m$, as well as $f(\nu_m) = Beta(\nu_m,1,25)$ for ν_m equal to or less than 5 % and $f(\nu_m) = Beta(\nu_m,0.5,0.5)$ for ν_m greater than 5 % as suggested earlier [7]. *Beta*-density weights distinguish MAFs more moderately than 1/ν-weights. For the RBF kernel (4), the scale parameter μ was the average weighted squared genetic difference between subjects $\Sigma_{i,j}((\mathbf{g}_i-\mathbf{g}_j)^{\mathrm{T}}\mathbf{W}(\mathbf{g}_i-\mathbf{g}_j))/n^2$ multiplied by the effective number of independent SNPs in the tested set [14].

Strategies for combined testing of common and rare variants

By default, the kernel score test, Eq. (2), is performed with a kernel matrix $\mathbf{K_{all}}$ computed on all dosages with a weighting of common and rare SNPs.

In contrast, Ionita-Laza et al. [7] recently suggested computing the kernel separately for rare SNPs ($\mathbf{K_{rare}}$)

and for common SNPs ($\mathbf{K_{common}}$), respectively, in a region of interest. Analogous to Eq. (2), this yields matrices $\mathbf{M_{rare}}$, $\mathbf{M_{common}}$, test statistics Q_{rare}, Q_{common}, and p values p_{rare}, p_{common}. The null model, $\mathbf{P_o}$ and $\mathbf{R}$ were always the same. The weighted sum test (WS) on common and rare variants has test statistic [7].

$$Q_{WS} = (1-\varphi)\cdot Q_{rare} + \varphi\cdot Q_{common} \quad (5)$$

Weight $\varphi = (\mathrm{tr}(\mathbf{M_{rare}}\cdot\mathbf{M_{rare}})/(\mathrm{tr}(\mathbf{M_{rare}}\cdot\mathbf{M_{rare}}) + \mathrm{tr}(\mathbf{M_{common}}\cdot\mathbf{M_{common}})))^{1/2}$ may be chosen such that $(1-\varphi)\cdot Q_{rare}$ and $\varphi\cdot Q_{common}$ have the same variance. P values are obtained by Davies' exact method from sample estimates Q_{WS} and all eigenvalues of matrix $((1-\varphi)\cdot\mathbf{M_{rare}} + \varphi\cdot\mathbf{M_{common}})$. Alternatively, Fishers p value pooling can be applied.

$$Q_{FISHER} = -2\ln(p_{rare}) - 2\ln(p_{common}) \quad (6)$$

Under H_0, $Q_{FISHER}/(1 + 0.25\cdot cov)$ is chi-square distributed with $16/(4 + cov)$ degrees of freedom [7]. With $r = \mathrm{tr}(\mathbf{M_{rare}}\cdot\mathbf{M_{common}})/(\mathrm{tr}(\mathbf{M_{rare}}\cdot\mathbf{M_{rare}})\cdot\mathrm{tr}(\mathbf{M_{common}}\cdot\mathbf{M_{common}}))^{1/2}$, the covariance between p_{rare} and p_{common} is $cov \approx r\cdot(3.25 + 0.75\cdot r)$ for $0 \le r \le 1$ and $cov \approx r\cdot(3.27 + 0.71\cdot r)$ for $-0.5 \le r \le 0$. Only test statistic (6) yields approximate p values; all other p values are obtained with Davies' method and are exact.

Results and discussion

Our test extension to families holds the nominal significance level and correctly adjusts for a polygenic familial variance component (as demonstrated in [6]). Table 1 lists the p values obtained for association testing of *AGTR1* on real SBP, considering common SNPs (MAF >5 %) and rare SNPs (MAF ≤5 %) as well as 3 joint tests (default test $\mathbf{K_{all}}$, WS, Fisher). *Beta*-weights (not shown) performed between equal weights and 1/ν-weights. The 1/ν-weight lowered p values particularly on common SNPs. *AGTR1* association is suggested by common as well as rare SNPs. Joint testing of rare and common SNPs was beneficial. In particular, WS and Fisher test p values were often smaller (and otherwise close to) the smallest p value of the separate rare and common SNP tests. When using ad hoc definitions of the *AGTR1* flanking region, Fisher and WS p values remained relatively stable and were also smaller compared to the default test $\mathbf{K_{all}}$. However, on the *AGTR1* containing LD-block all joint tests performed highly similar, p values were the smallest and also relatively stable regardless of SNP weights and SNP density.

Table 1 Analysis of real data: real SBP and candidate gene *AGTR1*

SNP panel	Weight	Common SNPs MAF >5 %		Rare SNPs MAF ≤5 %		Joint tests		
						Default	WS	Fisher
		N_{SNP}	p value	N_{SNP}	p value	p value	p value	p value
AGTR1 with no flanking region, positions 148415571–148460795								
GWAS	equal	11	0.189	7	0.097	0.177	0.102	0.101
	1/ν	11	0.113	7	**0.050**	0.054	**0.044**	**0.043**
SEQ	equal	74	0.203	138	0.060	0.173	0.076	0.076
	1/ν	74	0.160	138	0.098	0.083	0.088	0.090
AGTR1 with 30 kb flanking region, positions 148385571–148490795								
GWAS	equal	30	0.100	12	0.072	0.092	**0.050**	0.052
	1/ν	30	**0.045**	12	0.069	**0.030**	**0.029**	**0.029**
SEQ	equal	198	0.053	300	0.067	**0.047**	**0.030**	**0.032**
	1/ν	198	**0.039**	300	0.172	**0.045**	**0.044**	**0.050**
AGTR1 with 500 kb flanking region, positions 147915571–148960795								
GWAS	equal	277	0.206	51	**0.048**	0.196	0.061	0.065
	1/ν	277	0.151	51	0.064	0.102	0.059	0.066
SEQ	equal	2170	0.192	2244	0.069	0.173	0.080	0.085
	1/ν	2170	0.157	2244	0.051	0.062	0.057	0.060
AGTR1 containing LD-block, positions 148344702–148568958								
GWAS	equal	80	0.058	19	0.076	0.055	**0.035**	**0.036**
	1/ν	80	**0.040**	19	0.114	**0.034**	**0.036**	**0.039**
SEQ	equal	499	**0.029**	592	0.106	**0.027**	**0.027**	**0.030**
	1/ν	499	**0.027**	592	0.112	**0.025**	**0.026**	**0.030**

Association of *AGTR1* with real SBP was tested with a linear kernel on minor allele dosage data for GWAS and sequence (SEQ); $p \le 0.05$ bold. N_{SNP} common and rare SNPs, respectively, were combined into joint tests: kernel $\mathbf{K_{all}}$ (default), weighted sum test (WS), and Fisher's p value pooling for correlated p values

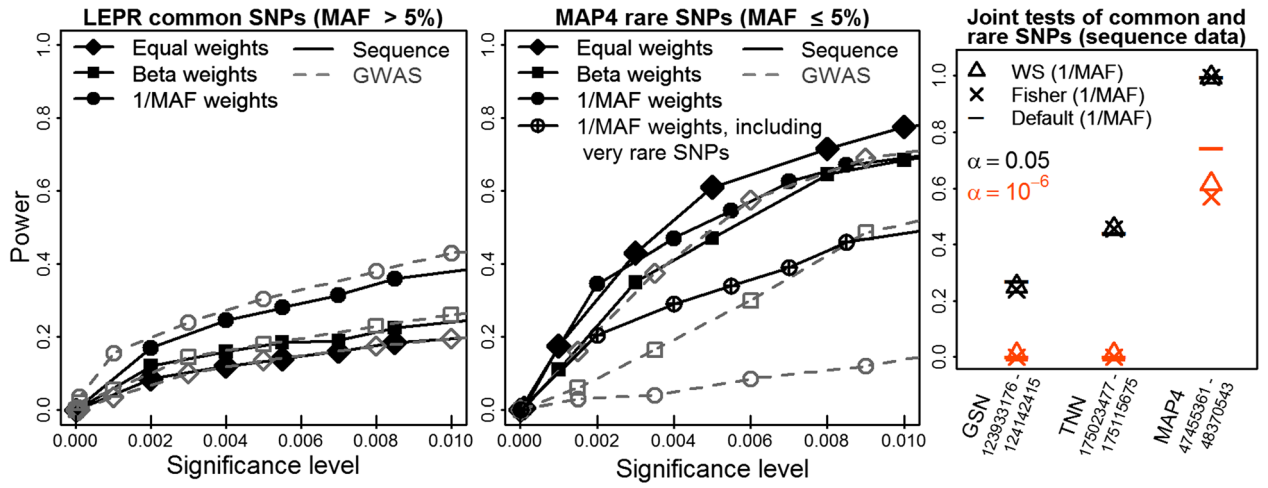

Fig. 1 Test power on simulated SBP may greatly depend on SNP weights. *Left* and *middle panels:* Power of the kernel score test over 200 study replicates of simulated SBP as function of the significance level for different SNP weights and SNP panels. *Right panel:* Power of joint tests of rare and common SNPs at 2 significance levels $\alpha = 0.05, 10^{-6}$ when using 1/v-weights on the sequence of gene-containing LD-blocks. Power estimates for *LEPR* (positions 65743083 to 66106465) and *FLT3* (28490385–28713642) (not shown) were highly similar to *TNN*

Next, we analyzed LD-blocks that contain the genes *MAP4*, *TNN*, *LEPR*, *GSN*, or *FLT3*. Figure 1 displays the average test power on 200 data replicates of simulated SBP. Sequence-derived variants were often more powerful than GWAS with some exceptions (Fig. 1 left and middle panels, black solid lines vs. gray dashed lines). The best were often 1/v-weights (circle), otherwise equal weights (diamond) were favored. Particularly 1/v-weights may be beneficial on common SNPs *(LEPR)* and occasionally detrimental on rare SNPs *(MAP4)*. The latter is an exceptional finding but consistent with Table 1 on candidate gene *AGTR1*. On rare *MAP4* SNPs, 1/v-weights lowered the power, especially when testing also extremely rare SNPs (encircled plus), but less so when testing only MAF equal to or less than 5 % SNPs that had at least 7 observations of the minor allele (filled circle; sequence data). On gene-containing LD-blocks, all joint tests (default test $\mathbf{K}_{\mathbf{all}}$, WS, Fisher) often had similar power (Fig. 1, right panel: *LEPR*, *FLT3*, *TNN* with highly similar results [only *TNN* shown]; *GSN* sequence). However, default test $\mathbf{K}_{\mathbf{all}}$ was the most powerful test on the gene with homogeneous strong LD (*MAP4*: sequence [Fig. 1, right] and GWAS [not shown]) and on the gene with the most variable LD structure (*GSN*: when using GWAS SNPs, not shown). Then, $\mathbf{K}_{\mathbf{all}}$ likely exploited SNP correlations better. When LD-blocks were enlarged by flanking regions, WS and Fisher often were slightly more powerful than $\mathbf{K}_{\mathbf{all}}$ (results not shown). The linear kernel had always similar or better power than the RBF kernel (results not shown).

Conclusions

As the power of kernel methods increases through the exploitation of SNP correlations [2], this ability should be utilized fully by analyzing LD-blocks. SNP weights have a far greater impact on test power than the kernel chosen. Currently, the benefit of 1/v-weights may be underestimated for common SNPs. On rare SNPs, 1/v-weights often improve power, but can also be detrimental. Findings are consistent with both real and simulated data. Our results suggest using 1/v-weights on all SNPs in a single kernel $\mathbf{K}_{\mathbf{all}}$ testing LD-blocks and only SNPs with sufficient minor allele observations. Alternatively, one may use WS with 1/v-weights on common SNPs and equal weights on rare SNPs in the kernels. WS upweights the rare variant contribution globally; see Eq. (5).

Acknowledgements

This work was supported by the Deutsche Forschungsgemeinschaft DFG (grant Klinische Forschergruppe [KFO] 241: TP5, BI 576/5-1; grant Research Training Group "Scaling Problems in Statistics" RTG 1644).

Declarations

This article has been published as part of *BMC Proceedings* Volume 10 Supplement 7, 2016: Genetic Analysis Workshop 19: Sequence, Blood Pressure and Expression Data. Summary articles. The full contents of the supplement are available online at http://bmcproc.biomedcentral.com/articles/supplements/volume-10-supplement-7. Publication of the proceedings of Genetic Analysis Workshop 19 was supported by National Institutes of Health grant R01 GM031575.

Authors' contributions

Authors contributed as follows: study concept, DM and HB; data extraction and analysis, DM and SF; SNP mapping with NCBI build 37 and LD calculations, SF; and writing of the manuscript, DM. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Published: 18 October 2016

References

1. Schaid DJ. Genomic similarity and kernel methods I: advancements by building on mathematical and statistical foundations. Hum Hered. 2010;70(2):109–31.
2. Schifano ED, Epstein MP, Bielak LF, Jhun MA, Kardia SL, Peyser P, Lin X. SNP set association analysis for familial data. Genet Epidemiol. 2012;36(8):797–810.
3. Chen H, Meigs JB, Dupuis J. Sequence kernel association test for quantitative traits in family samples. Genet Epidemiol. 2013;37(2):196–204.
4. Oualkacha K, Dastani Z, Li R, Cingolani PE, Spector TD, Hammond CJ, Richards JB, Ciampi A, Greenwood CM. Adjusted sequence kernel association test for rare variants controlling for cryptic and family relatedness. Genet Epidemiol. 2013;37(4):366–76.
5. Huang J, Chen Y, Swartz MD, Ionita-Laza I. Comparing the power of family-based association test for sequence data with applications in the GAW18 simulated data. BMC Proc. 2014;8 Suppl 1:S27.
6. Malzahn D, Friedrichs S, Rosenberger A, Bickeböller H. Kernel score statistic for dependent data. BMC Proc. 2014;8 Suppl 1:S41.
7. Ionita-Laza I, Lee S, Makarov V, Buxbaum JD, Lin X. Sequence kernel association tests for the combined effect of rare and common variants. Am J Hum Genet. 2013;92(6):841–53.
8. Almasy L, Dyer TD, Peralta JM, Jun G, Wood AR, Fuchsberger C, Almeida MA, Kent Jr JW, Fowler S, Blackwell TW, et al. Data for Genetic Analysis Workshop 18: human whole genome sequence, blood pressure, and simulated phenotypes in extended pedigrees. BMC Proc. 2014;8 Suppl 1:S2.
9. Baudin B. Polymorphism in angiotensin II receptor genes and hypertension. Exp Physiol. 2005;90(3):277–82.
10. The International HapMap Consortium. The International HapMap project. Nature. 2003;426(6968):789–96.
11. Barrett JC, Fry B, Maller J, Daly MJ. Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics. 2005;21(2):263–5.
12. Blom G. Statistical estimates and transformed beta variables. New York: John Wiley & Sons; 1958.
13. Davies RB. Algorithm AS 155: the distribution of a linear combination of chi-2 random variables. J R Stat Soc: Ser C: Appl Stat. 1980;29(3):323–33.
14. Cheverud JM. A simple correction for multiple comparisons in interval mapping genome scans. Heredity (Edinb). 2001;87(Pt 1):52–8.

Friedrichs et al. BMC Genetics 2016, 17(Suppl_2):8
DOI 10.1186/s12863-015-0313-x

BMC Genetics

PROCEEDINGS Open Access

Filtering genetic variants and placing informative *priors* based on putative biological function

Stefanie Friedrichs[1], Dörthe Malzahn[1], Elizabeth W. Pugh[2], Marcio Almeida[3], Xiao Qing Liu[4,5] and Julia N. Bailey[6,7*]

From Genetic Analysis Workshop 19
Vienna, Austria. 24-26 August 2014

Abstract

High-density genetic marker data, especially sequence data, imply an immense multiple testing burden. This can be ameliorated by filtering genetic variants, exploiting or accounting for correlations between variants, jointly testing variants, and by incorporating informative *priors*. *Priors* can be based on biological knowledge or predicted variant function, or even be used to integrate gene expression or other omics data. Based on Genetic Analysis Workshop (GAW) 19 data, this article discusses diversity and usefulness of functional variant scores provided, for example, by PolyPhen2, SIFT, or RegulomeDB annotations. Incorporating functional scores into variant filters or weights and adjusting the significance level for correlations between variants yielded significant associations with blood pressure traits in a large family study of Mexican Americans (GAW19 data set). Marker rs218966 in gene *PHF14* and rs9836027 in *MAP4* significantly associated with hypertension; additionally, rare variants in *SNUPN* significantly associated with systolic blood pressure. Variant weights strongly influenced the power of kernel methods and burden tests. Apart from variant weights in test statistics, *prior* weights may also be used when combining test statistics or to informatively weight *p* values while controlling false discovery rate (FDR). Indeed, power improved when gene expression data for FDR-controlled informative weighting of association test *p* values of genes was used. Finally, approaches exploiting variant correlations included identity-by-descent mapping and the optimal strategy for joint testing rare and common variants, which was observed to depend on linkage disequilibrium structure.

Background

With the availability of very dense genetic marker data sets, such as sequence data, even large association studies can become underpowered. This raises the need to filter, or prioritize, or jointly test genetic variants.

Filters or *priors* on genes may be derived from methylation or expression data if available in the same individuals. Alternatively, one may use external information. Recently, multiple annotation tools have become available using several databases and algorithms that predict functional effects of genetic variants. Commonly used are, for example, ANNOVAR (Annotate Variation) [1], VariantTools [2], PolyPhen [3], SIFT (Sorting Intolerant From Tolerant) [4], ENCODE (Encyclopedia of DNA Elements) [5], RegulomeDB [6], CADD (Combined Annotation-Dependent Depletion) [7], or Gerp++ [8]. Tools like ANNOVAR additionally provide variant annotation to genes and to regions such as conserved regions among species, predicted transcription factor binding sites, and segmental duplication regions. Many of the above-listed tools also provide information on regulatory elements that control gene activity. This article demonstrates that functional scores can contribute to the success of association studies. Simultaneously, functional scores may differ substantially between databases and prediction tools as they can be based on different functional aspects.

* Correspondence: JBailey@mednet.ucla.edu
Stefanie Friedrichs and Dörthe Malzahn share first authorship.
[6]Department of Epidemiology, Fielding School of Public Health, University of California, Los Angeles, Los Angeles, CA, USA
[7]Epilepsy Genetics/Genomics Laboratory, West Los Angeles Veterans Administration, Los Angeles, CA, USA
Full list of author information is available at the end of the article

Additionally, variant annotations to chromosomal positions continue to be updated with the National Center for Biotechnology Information (NCBI) [9] human genome build as standard. Furthermore, variants can be annotated to genes based on different sources, such as ENSEMBL [10], Vega [11], GENCODE [12], and many more. Researchers also use a variety of definitions of flanking regions. Finally, genes may be grouped by function or biological pathway, again with substantial variability between data bases such as KEGG [13], Biocarta [14], or Pathway Interaction Database [15]. This article discusses approaches that filtered or prioritized genetic variants, regions, or genes. Pathway-based approaches, although also incorporating filters or *priors*, are discussed separately by Kent [16].

Many researchers filter genetic variants. The simplest forms of filters are minor allele frequency (MAF), candidate genes or variants, or considering the exome. Filters and statistical models are chosen to increase the power under a hypothetical disease model. The advent of sequencing renewed interest in disease mechanisms less frequent but more penetrant than common single nucleotide polymorphisms (SNPs) of genome-wide association studies (GWAS). This led, for example, to screening for recessive variants by examining runs of homozygosity [17, 18]. When multiple rare causal variants cluster within a gene, identity-by-descent (IBD) mapping may be more powerful than single-locus association testing [19]. IBD mapping can be used in 2-step approaches. For example, Balliu et al [20] identified regions where hypertension cases shared more segments of IBD than controls in one part of the sample. They modeled aggregate effects of each of these regions on blood pressure (BP) in the sample remainder. Aggregation tests are used especially for testing rare single-nucleotide variants (SNVs). Aggregation tests are burden tests, variance-component tests, or a combination of both, such as SKAT-O (optimal unified sequence kernel association test) (see, eg, Lee et al [21] for a review). Kernel-based approaches (see Schaid [22] for a review) such as the sequence kernel association test (SKAT) [23] are variance-component tests. Examples of genetic burden tests are T5, combined multivariate collapsing (CMC) [24], or C-α [25]; see also Santorico et al [26]. Aggregation tests can prioritize SNVs by weighting minor allele dosages in the test statistic. Typical weights account for MAF, but may also incorporate putative functional relevance of SNVs [27, 28]. Moreover, weights may be used to combine aggregation test statistics [21, 29, 30], and one may weight p values while controlling the false discovery rate (FDR) [31, 32]. For example, GWAS p values may be weighted based on functional annotations. For aggregation tests on genes, p value weights can be utilized to integrate gene expression or other omics data [33].

This article summarizes contributions of the Genetic Analysis Workshop (GAW) 19 group on filtering variants and placing informative *priors* (Tables 1 and 2). These investigations found that improving SNV grouping or selection can noticeably increase power. Moreover, including functional scores or gene expression data as filters or weights on variants, genes, or when combining test statistics assisted in detecting associations. Some contributions also exploited SNV correlations to increase power or improved the multiple-testing adjusted significance threshold by accounting for SNV correlations.

Materials

Analyzed data were provided by GAW 19 and included a family sample (n = 959) with extended pedigrees of Mexican Americans from the San Antonio Family Heart Study (SAFHS) and the San Antonio Family Diabetes/ Gallbladder Study (SAFDS/ SAFGS) [34]. The family sample also contained 103 unrelated sequenced subjects; 259 subjects had gene expression data. This study was designed to identify low-frequency or rare variants influencing susceptibility to type 2 diabetes (T2D) as part of the T2D Genetic Exploration by Next-generation sequencing in Ethnic Samples (T2D-GENES) Consortium. Phenotypes included real and simulated longitudinal systolic (SBP) and diastolic blood pressure (DBP) and hypertension (HT) status. Available were sequence for 464 pedigree members and GWAS SNPs for all 959 subjects. Additionally, all subjects were imputed to sequence based on original genotypes and familial relationships [34]. Approaches described herein mostly analyzed imputed dosages to avoid missing genotypes and to maximize sample size. Zhang et al [28] analyzed the GAW19 sample of 1943 independent Hispanic subjects with whole exome sequence. This sample had been ascertained by T2D status. However, GAW19 provided real and simulated cross-sectional BP traits instead [35], using the same trait-simulation model as for the family study.

All approaches described herein are nonlongitudinal analyses of BP traits (SBP, DBP, or HT) in relation to minor allele dosages of sequence SNVs or genome-wide SNPs.

Methods

Statistical methods employed by this group (see Table 1) to incorporate filters or informative *priors* are mostly based on regression models [27, 30, 33, 36, 37]; one is also based on counting methods [28]. Analyses of family data adjusted for familial dependence based on the kinship matrix. They included the familial covariance in a linear mixed model [27, 30, 36] or transformed the trait to a conditionally independent surrogate variable [33]. Analyses of independent subjects accounted for population structure (cryptic relatedness and admixture) [37] by using the programs Eigensoft [38] and Admixture [39].

Annotating genetic variants for location and function

A variety of freely available genetic databases and highly developed software tools support annotation of location

Table 1 Statistical tests and analyzed data

Marker data	Data set	Statistical tests	Covariates	Trait(s)
Almeida et al [36]				
Sequence	Family study	Single-variant regression in SOLAR	Smoking, BP medication, PC1-3, sex, age, age^2, sex*age, sex*age^2	Real SBP and DBP at first time point, own simulated trait for H_0
Liu et al [37]				
Chr3: GWASmp and sequence	Unrelated individuals (from family study)	Regress pairwise DBP residual difference and sum on IBD sharing status; sequence data analyses by SKAT-O	Sex, age, smoking, PC 1-3	Real DBP at first time point
Kim and Wei [27]				
Sequence	Family study	Informative SNV weights in burden test T5 and SKAT; with R: seqMeta	Age, sex, smoking, BP medication	Real SBP at earliest available measurement
Zhang et al [28]				
Exome sequence	Unrelated individuals (large Hispanic sample)	LRT, C-*α*, CMC on informatively weighted SNV burden	None	Simulated HT status; real SBP, DBP with cutoffs for case-control status
Malzahn et al [30]				
Sequence and GWASmp	Family study	SKAT with R (coxme, kinship2, QuadCompForm); strategies for joint testing of rare and common SNVs	Sex, age, sex*age; subjects not on BP medication	Real and simulated SBP at first time point
Ho et al [33]				
Sequence and GWASmp	Family study, including gene expression data	Seq-aSum-VS burden test; regression on gene expression data; gene set enrichment analysis	PC1-3	Average real SBP and DBP

BP blood pressure, *Chr* Chromosome, *CMC* Combined multivariate collapsing, *DBP* diastolic blood pressure, *GWASmp* genome-wide association study marker panel, *HT* hypertension, *IBD* identity-by-descent, *LRT* likelihood ratio test, *PC* principal component, *SBP* systolic blood pressure, *SKAT* sequence kernel association test, *SNV* single nucleotide variant, *Seq-aSum-VS* sequential sum

and biological function of SNVs. In our group, SNV locations were obtained by ANNOVAR [28, 36] or determined based on reference data, for example, from the Genome Reference Consortium [40] or the International Haplotype Map (HapMap) Consortium [41] [30, 37]. Reference data were also used to determine linkage disequilibrium (LD) blocks [30] with Haploview [42].

Kim and Wei [27] and Almeida et al [36] used functional annotations from ENCODE, PolyPhen or PolyPhen2, and SIFT, while Liu et al [37] used CADD. In contrast, Zhang et al [28] annotated putative protein binding sites based on 2 different algorithms using random forest classifiers [43].

Filtering genetic variants

Not all areas of the genome were studied. Some researchers filtered the data prior to analyses. Zhang et al [28] investigated exome sequence and Almeida et al [36] molecularly functional nonsynonymous SNVs predicted by PolyPhen and SIFT. Liu et al [37] examined IBD sharing regions on chromosome 3. Malzahn et al [30] considered gene-containing LD blocks for selected candidate genes. Ho et al [33] analyzed rare SNV burden in genes containing less than 50 and more than 1 rare SNV (MAF <0.01).

Accounting for correlations between genetic variants

An important difference between methods is that variant correlations can either be a nuisance or may be used to increase power. For example, IBD mapping exploits variant correlations. IBD mapping can be more powerful than single-locus association testing when multiple causal rare variants cluster within a gene [19]. Therefore, Liu et al [37] tested the relationship between IBD sharing status and trait differences and sums for pairs of individuals. Moreover, the power of kernel methods such as SKAT may be increased through the exploitation of variant correlations [44]. This ability can be utilized fully by analyzing LD blocks [30]. On the other hand, single-locus methods need to account for variant correlations to appropriately correct the significance level for multiple testing. Hence, Almeida et al [36] determined the effective number of independent tests by extreme value theory based on replicates of a simulated unassociated trait.

Correcting the significance level for the number of independent tests

The significance level used with multiple testing is always an issue as too conservative a correction will cause false negatives and not correcting enough will cause false positives.

Table 2 Filters, *priors*, and findings

Filter	*Prior*	Conclusions	Annotation
Almeida et al [36]			
Functional annotation, LD-corrected effective number of tests	None	LD-correction in WGS reduces multiple-testing burden by 85 %, significant associations: *PFH14* with SBP, *MAP4* with DBP	Location: ANNOVAR; functional annotation: PolyPhen, SIFT
Liu et al [37]			
IBD sharing	None	No significances, *ZPLD1* had strongest evidence	IBD mapping: BEAGLE; functional annotation: CADD
Kim and Wei [27]			
Sliding window on MAF ≤5 % SNVs	*SNV-weights:* based on MAF or regulatory importance	Significant association: *SNUPN*	Functional annotation: ENCODE, RegulomeDB, PolyPhen2
Zhang et al [28]			
Genes, exome-sequence	*SNV-weights:* up-weight protein binding sites, apply direction weights	Top-ranked genes differ between weighted burden tests LRT, C-α, CMC; but good overlap with literature	ANNOVAR, variant tools; random forest classifiers assign SNVs to protein binding sites; DSSP, PSAIA, DOMINO
Malzahn et al [30]			
Gene covering LD-blocks	*SNV-weights:* using MAF *Overall weight:* on rare SNV variance component in SKAT	SKAT: power depends on SNV weights, exploiting LD is very beneficial, optimal strategy for joint testing rare and common SNVs depends on LD structure	Haploview with HapMap data for LD-calculation
Ho et al [33]			
Rare SNVs in genes with >1 and <50 rare SNVs (MAF < 0.01)	*p value weights:* improve gene ranking	Power of burden tests improved by incorporating phenotype associated gene expression into *p* value weights	Genes: hg19; GO biological process categories

CADD combined annotation dependent depletion, *DBP* diastolic blood pressure, *DOMINO* database of domain–peptide interactions, *DSSP* define secondary structure of proteins, *ENCODE* encyclopedia of DNA elements, *GO* gene ontology, *IBD* identity-by-descent, *LD* linkage disequilibrium, *MAF* minor allele frequency, *PSAIA* protein structure and interaction analyzer, *SBP* systolic blood pressure, *SIFT* sorting intolerant from tolerant, *SKAT* sequence kernel association test, *SNV* single nucleotide variant, *WGS* whole genome sequence

Almeida et al [36] adjusted the significance level for single locus analyses by estimating the number of independent tests [45]. A total of 1000 replicates of a quantitative phenotype with no genetic effects were simulated and tested on whole genome sequence data, using linear mixed models in SOLAR (Sequential Oligogenic Linkage Analysis Routines) [46]. The smallest p value per simulation run was extracted. The density of these 1000 extremely small p values was fitted to a theoretical beta distribution $beta(1,n_e)$ where n_e is the effective number of independent tests [47]; yielding the adjusted significance level $a^* = \frac{0.05}{n_e}$. This procedure was applied to both whole genome sequence and functional nonsynonymous SNVs.

Identity-by-descent mapping

IBD mapping aims to detect loci sharing ancestral segments in unrelated individuals. In particular, unrelated subject-pairs with smaller trait differences are expected to share significantly more rare causative variants than pairs with larger trait differences. Liu et al [37] estimated IBD sharing segments with BEAGLE [48]. The squared trait difference (D) and squared trait sum (S) for trait DBP between pairs of unrelated subjects was regressed on IBD sharing status. This yielded parameter estimates for slopes $(\hat{\beta}_S, \hat{\beta}_D)$ and variances (σ_S^2, σ_D^2), which were combined into an overall slope estimate $\hat{\beta} = \left(\frac{\sigma_D^2}{\sigma_S^2+\sigma_D^2}\right)\hat{\beta}_s + \left(\frac{\sigma_S^2}{\sigma_S^2+\sigma_D^2}\right)\hat{\beta}_D$. Linkage was tested with test statistic $t = \frac{\hat{\beta}}{SE(\hat{\beta})}$ under the null hypothesis of an overall slope of zero [37]. The significance threshold for non-independent pairs was estimated by permutation procedure.

Priors on genes and variants

Genetic *priors* can be incorporated by variant weights in aggregation tests such as burden tests or SKAT [21]. Burden tests collapse minor allele dosages x_{ik} of a set of $i = 1, \ldots, m$ variants into a burden score $s_k = \sum_{i=1}^{m} \omega_i x_{ik}$ per individual k using a priori specified variant weights ω_i. One tests trait association with genetic burden s_k. Although burden tests are powerful when causal SNVs have the same effect direction, SKAT is more powerful when effect directions differ or if many noncausal SNVs are included in testing [21, 49]. SKAT is based on an underlying Bayesian model that estimates a random effect per SNV [50]. Specified is a kernel matrix of genetic

between-subject similarity and this kernel constitutes a *prior* on genetic model space [51]. SNV weights are incorporated in the kernel (see, eg, Malzahn et al [30]).

Typically, rarer SNVs get assigned more weight to counterbalance their reduced power compared to more frequent SNVs. Used are, for example, weights $\omega_j = \frac{1}{\mathrm{MAF_j}(1-\mathrm{MAF_j})}$ [52], inverse MAF weights $\omega_j = \frac{1}{\mathrm{MAF_i}}$, or *beta*-weights such as $\omega_j = b(\mathrm{MAF_i})$ [23], where b is the probability density function of a $beta(1, 25)$ random variable. Malzahn et al [30] compared the power of SKAT when using different SNV weights and different kernel functions that either allow or do not allow for SNV interactions in the genetic model. Alternatively, SNV weights may be based on regulatory importance [27] or protein binding effects [28].

Incorporating functional information into variant weights

Kim and Wei [27] categorized SNVs according to RegulomeDB and PolyPhen2 functional relevance scores. SNV weights were defined based on $f(s) = S^2$ where s equaled the reverse order of categories, namely s = 6, 5, 4, 3, 2, 1 for category 1 ("most likely affecting binding and expression") to category 6 ("not functionally relevant"). Kim and Wei [27] tested rare SNVs jointly, in sets defined by sliding windows of 4 kb size, for association with SBP. They compared the power of SNV weighting schemes in SKAT ($\omega_j = \sqrt{f(s_j)}$ versus $\omega_j = b(\mathrm{MAF_j})$), and burden test T5 ($\omega_j = f(s_j)$ versus $\omega_j = \frac{1}{\mathrm{MAF_j}(1-\mathrm{MAF_j})}$). SKAT and T5 provide analytical asymptotically exact p values with good small sample size behavior.

Zhang et al [28] used a likelihood ratio test (LRT) [53] to test if the proportion of subjects with an informatively weighted minor allele burden exceeding a given threshold differed between HT cases and controls. P values were obtained by permutation procedure. SNV weights ω_i accounted for putative effect direction and distinguished between functional SNVs in binding-sites ($|\omega_i| = 10$), not in binding-sites ($|\omega_i| = 5$), and nonfunctional SNVs ($|\omega_i| = 1$). The informatively weighted LRT was compared with C-α and CMC burden tests.

Optimal joint testing of rare and common variants

When not filtering for rare or common SNVs, optimal joint testing of both becomes an issue. Suppose, one computed 2 SKAT statistics, $\mathrm{Q_{rare}}$ and $\mathrm{Q_{common}}$, separately on rare SNVs and common SNVs, in the same region of interest, for the same trait, based on the same genetic null model. As SKAT is a variance-component test, combining $\mathrm{Q_{rare}}$ and $\mathrm{Q_{common}}$ [29]

$$\mathrm{Qws} = (1-\lambda)\cdot \mathrm{Q_{rare}} + \lambda \cdot \mathrm{Q_{common}} \quad (1)$$

weights the rare SNV variance-component by overall a priori weight (1-λ) relative to the common SNV variance-component (see Ionita-Laza et al [29] and Malzahn et al [30] for choices of λ). The weighted sum test (1) is another way of structuring a *prior* in SKAT. Note that $\mathrm{Q_{rare}}$ and $\mathrm{Q_{common}}$ may use different kernel functions or different SNV weights. Malzahn et al [30] compared this form of joint testing of rare and common SNVs with the default choice of entering *all* SNVs with appropriate weights into a *single* kernel. Exact p values for SKAT and weighted sum test (1) were obtained by Davies method [54]. Another investigated alternative was Fisher pooling of the correlated p values resulting from the separate rare SNV and common SNV SKAT statistics. Fisher pooling accounted for correlations by Satterthwaite approximation and Brown's method ([55]; see also [29, 30]).

Note that analogously to equation (1), SKAT-O combines SKAT and burden tests with statistic $Q = (1-\rho)Q_{SKAT} + \rho Q_{burden}$ where $0 \le \rho \le 1$ [56].

Informed *p* value weighting for genes

Ho et al [33] obtained gene-wise p values, p_g, for association of average BP T with rare SNV burden s_g in genes g that had more than 1 and less than 50 rare SNVs (MAF <0.01)

$$T \sim b_{s,g}\cdot s_g \quad (2)$$

Restricting the number of rare SNVs avoids collapsing too many null variants. Ho et al [33] used the sequential sum test [57], which data-adaptively assigned SNV weights $\omega_i = 0, 1, -1$. Earlier, Genovese et al [31] and Roeder and Wasserman [32] had proven that informative weighting of p values $\frac{p_g}{v_g}$ with weights $v_g > 0$, $\bar{v}_g = 1$ maintains proper FDR control; where $\frac{p_g}{v_g} \le \alpha_{FDR}$ means significance. Ho et al [33] determined such weights v_g as follows. They tested if rare minor allele burden s_g^* (with SNV weights $\omega_i = 1$, for simplicity) also associated with gene expression E_g

$$E_g | T \sim b_{E,g}\cdot s_g^* + c\cdot T \quad (3)$$

and further if gene expression E_g associated with trait value T

$$T | s_g^* \sim b_{T,g}\cdot E_g + d\cdot s_g^* \quad (4)$$

Association tests (2) to (4) were made conditionally independent by adjusting test (3) for trait value T and test (4) for rare minor allele burden s_g^* (Fig. 1). P value weights $v_g = v_g^* \bar{v}_g^*$ were derived as $v_g^* = max\left(\left(\frac{\hat{b_{E,g}}}{SE(\hat{b_{E,g}})}\right)^2 \times \left(\frac{\hat{b_{T,g}}}{SE(\hat{b_{T,g}})}\right)^2\right)$ where the maximum was over all gene expression measurements and $\bar{v}_g^*$ was the average of all v_g^*.

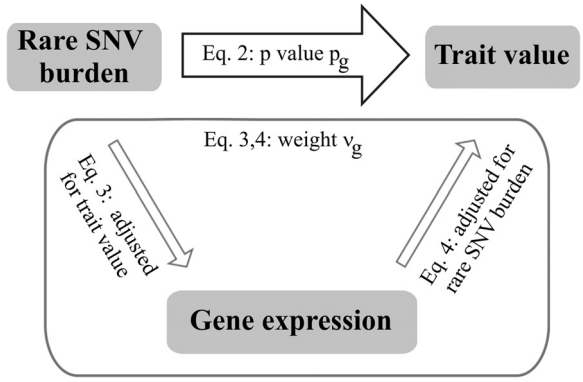

Fig. 1 Informed *p* value weighting for genes based on conditionally independent associations between rare variant burden, gene expression, and trait. The *p* value weight v_g was defined as the product of the association strengths of rare SNV burden with gene expression and gene expression with trait value

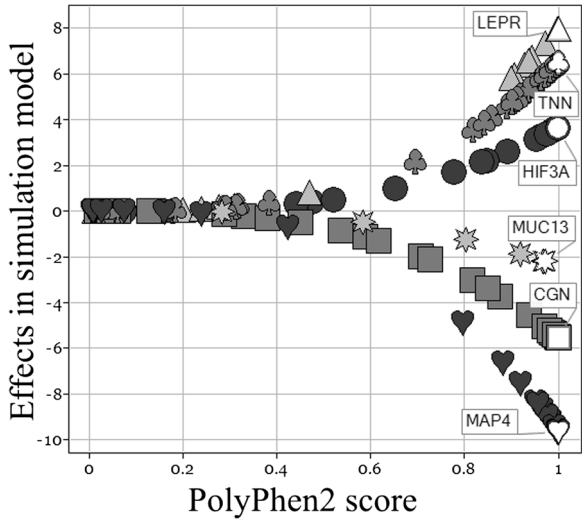

Fig. 2 SNV effect sizes on GAW19 simulated DBP increase with increasing PolyPhen2 scores. Depicted are 6 genes with a range of SNV effect sizes that could be simultaneously displayed. Symbols depict SNVs in the same gene: LEPR (▲), TNN (♣), HIF3A (●), MAP4(♥), MUC13(✳), CGN(■)

Results and discussion

The results for this GAW19 working group varied widely as a result of the different objectives of each contributor. Table 2 provides a brief summary of specific results.

Under H_0, extreme *p* values follow a beta distribution [47]. Almeida et al [36] reported that the beta distribution provided an excellent fit to determine the effective number of independent tests n_e for n single-locus tests. For whole genome sequence, $\frac{n_e}{n} = 15\%$; that is, accounting for LD reduced the multiple-testing burden by 85 %. However, significant associations could only be found when LD-correcting the significance level after a priori reducing sequence data based on functional annotations. Then 2 SNPs were detected: rs218966 in gene *PHF14* associated with SBP and rs9836027 in *MAP4* associated with DBP.

Liu et al [37] scanned chromosome 3 (GWAS data) for IBD sharing segments that associated with DBP. No genome-wide significance was found. However, several risk variants were detected in the region of gene *ZPLD1* by using CADD functional scores and sequence for the most promising region at 3q12.3.

In the GAW19 trait simulation model, SNV effect sizes were based on PolyPhen2 functional prediction scores (Fig. 2) [35]. In Figs. 2 and 3, displayed SNV effects, PolyPhen2 scores, and the assignment to positions and genes (NCBI build37, human genome build 19) came from the simulation answers. To illustrate differences between functional annotations, SIFT scores (and rs-numbers) were added by annotating sequence (variant call format [vcf] files) with ANNOVAR and merging vcf files and simulation answers by chromosome and position. RegulomeDB scores were merged by dbsnp138 rs-identifier. Furthermore, functional scores were transformed to have the same directionality (Fig. 3). Different functional annotations focus on different information about SNVs and only annotate selected SNVs. PolyPhen2 and SIFT both annotate nonsynonymous coding SNVs by a metric score that can be categorized to distinguish benign mutations from damaging ones affecting protein function. Nevertheless, PolyPhen2 and SIFT scores differ to a substantial extent in value and category (Fig. 3a). RegulomeDB annotates regulatory SNVs by an ordinal score ranging from the highest evidence (eQTL, expression quantitative trait locus) to the lowest. Figure 3c illustrates that some SNVs were rated to affect gene expression and transcription factor binding (RegulomeDB scores 1 to 5) but not the protein function (scored "benign" by PolyPhen2). For *simulated* BP, SIFT and RegulomeDB annotations yield mismatched filters or *priors* whenever they deviate from the PolyPhen2 score used to simulate SNV effects. For example, SIFT annotated some SNVs with large effects in gene *TNN* as benign mutations (Fig. 3b) and only few SNVs in associated genes were rated to be of regulatory importance (Fig. 3d). Nevertheless, for *real* SBP, several multiple-testing adjusted significant windows (2 with SKAT, 4 with burden test T5) were only found when including RegulomeDB scores as variant weights for rare SNV analysis [27]. One of these regions contained *SNUPN* [27] which is a novel finding not previously reported to associate with BP. T5 and SKAT maintained the nominal significance level on simulated unassociated trait Q1 also when incorporating RegulomeDB scores into variant weights [27]. Kim and Wei [27] and Zhang et al [28]

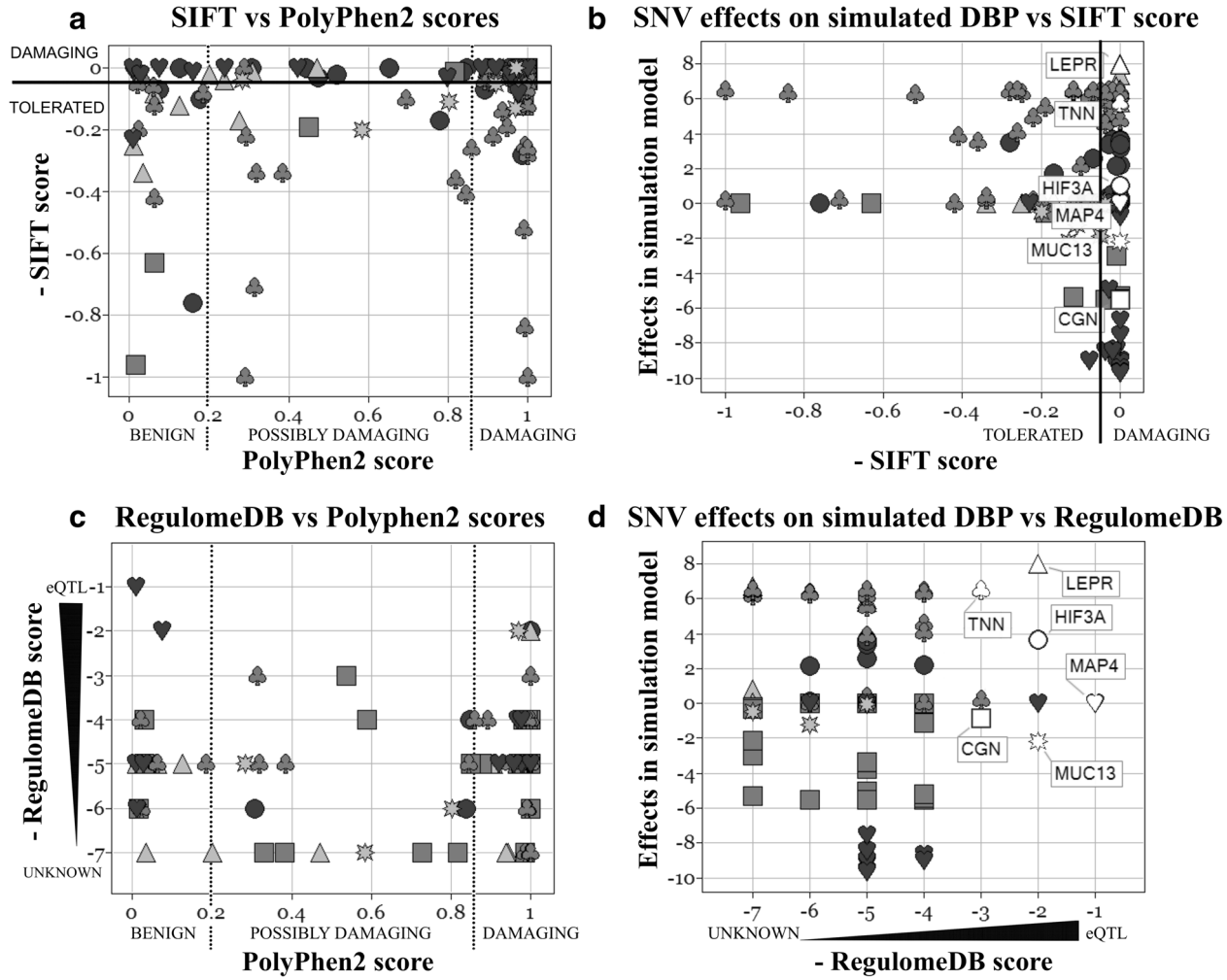

Fig. 3 Comparison between the PolyPhen2, SIFT, and RegulomeDB functional prediction scores. *Left column*: Correlation of PolyPhen2 functional prediction scores with (**a**) SIFT or (**c**) RegulomeDB scores. Functional scores were transformed to have the same directionality. Nonsynonymous coding SNVs that alter the protein function should receive a PolyPhen2 score of 1 and a SIFT score of 0. Scores are metric and can be categorized as displayed. RegulomeDB annotates regulatory SNVs by an ordinal score ranging from the highest evidence (eQTL, expression quantitative trait locus) to the lowest. *Right column*: Filters or *priors* based on (**b**) SIFT or (**d**) RegulomeDB functional scores are partially mismatched on GAW19 simulated DBP. Symbols depict SNVs in the same gene: LEPR (▲), TNN (♣), HIF3A (●), MAP4(♥), MUC13(✳), CGN(■)

both recommended using relatively big differences in SNV weights distinguishing functional from nonfunctional SNVs. Zhang et al [28] observed that different burden tests with functionally informative SNV weights yielded different top ranked genes. Although no gene was significant, many of them had been reported in the BP literature before. For SKAT, Malzahn et al [30] found that variant weights, but not kernel choice, had a strong influence on power, for rare as well as common SNVs. Kernel methods may gain power by exploiting SNV correlations. This can be utilized fully by analyzing LD blocks [30]. LD structure also influenced which strategy yielded the best joint test of rare and common SNVs with SKAT [30].

When using gene expression data to informatively weight gene-wise p values for association of rare SNV burden with BP [33], 153 genes (out of 6118) reached nominal significance (weighted $p \leq 0.05$). P value weights were determined such that evidence for phenotype associated gene expression lowered burden test p values. As no gene reached multiple-testing adjusted significance, Ho et al [33] used gene set enrichment analysis as aggregation test to relate the 153 top genes to biological pathways.

Conclusions

All analyses presented herein used a cross-sectional design by analyzing trait data of the first examination, the first available examination, or longitudinally averaged traits. This mainly contributed to differences in sample

size and trait variability. Furthermore, analyzing trait values at different time points may affect the marginal effect of genes that interact with age.

Including biological knowledge increased the power of association studies performed in our GAW group; especially filtering variants based on putative functional relevance. *Prior* weights can be included at different stages of the testing procedure. They can be incorporated into the test statistic of SKAT or burden tests, used when combining test statistics, or applied to association test p values. Selecting variant-sets also should take genetic structures into consideration, such as LD or IBD sharing. Moreover, the effective number of independent tests can be determined relatively easily by extreme value theory. This enables appropriate adjustment of the significance level for multiple testing to avoid an overly conservative approach. Ideally, variant grouping and selection, inclusion of biological information, and significance level adjustment can be applied simultaneously. Strategies like these are useful in increasing power in analyses of highly dense genetic data sets.

Filtering variants clearly boosted power in the discussed studies. However, filtering might also lose information. Functional scores such as PolyPhen2, SIFT, CADD, or RegulomeDB differ as they focus on different information about SNVs. Moreover, appropriateness of functional scores for a considered trait is a priori unknown. Hence, one is well advised to use and combine multiple functional annotations into a single filter or *prior*. This is feasible as functional annotations yield strong filters that greatly reduce the SNV space.

Competing interests
The authors declare they have no competing interests.

Authors' contributions
SF and DM wrote the manuscript. EWP contributed the comparison between the PolyPhen2, SIFT, and RegulomeDB functional annotation scores. All authors critically reviewed the manuscript for important intellectual content and interpretation of findings. All authors read and approved the final manuscript.

Acknowledgements
The authors would like to thank Zheyang Wu and Peng Wei for their comments and suggestions, as well as the GAW organizers for all their efforts. SF and DM were supported by the Deutsche Forschungsgemeinschaft (DFG, grant Research Training Group "Scaling Problems in Statistics" RTG 1644; grant Klinische Forschergruppe (KFO) 241: TP5, BI 576/5-1). EWP and JNB acknowledge support by National Institutes of Health (NIH) grants (HHSN268201200008I, R01 NS055057). XQL was supported by the University of Manitoba start-up funds. T2D-GENES is supported by NIH grants U01 DK085524, U01 DK085501, U01 DK085526, U01 DK085584 and U01 DK085545, the SAFHS by grant P01 HL045222, the SAFDS by grant R01 DK047482, and the SAFGS by grant R01 DK053889. Genetic Analysis Workshop 19 was supported by NIH grant R01 GM031575.

Declarations
This article has been published as part of *BMC Genetics* Volume 17 Supplement 2, 2016: Genetic Analysis Workshop 19: Sequence, Blood Pressure and Expression Data. Summary articles. The full contents of the supplement are available online at www.biomedcentral.com/bmcgenet/supplements/17/S2. Publication of the proceedings of Genetic Analysis Workshop 19 was supported by National Institutes of Health grant R01 GM031575.

Author details
[1]Department of Genetic Epidemiology, University Medical Center, Georg-August University Göttingen, Göttingen, Germany. [2]Center for Inherited Disease Research, Institute of Genetic Medicine, Johns Hopkins School of Medicine, Baltimore, MD, USA. [3]South Texas Diabetes and Obesity Institute, University of Texas Rio Grande Valley, Brownsville, TX, USA. [4]Department of Obstetrics, Gynecology, and Reproductive Sciences, Department of Biochemistry and Medical Genetics, Faculty of Health Sciences, University of Manitoba, Winnipeg, MB, Canada. [5]Children's Hospital Research Institute of Manitoba, Winnipeg, MB, Canada. [6]Department of Epidemiology, Fielding School of Public Health, University of California, Los Angeles, Los Angeles, CA, USA. [7]Epilepsy Genetics/Genomics Laboratory, West Los Angeles Veterans Administration, Los Angeles, CA, USA.

Published: 3 February 2016

References
1. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 2010;38:e164–e164.
2. San Lucas FA, Wang G, Scheet P, Peng B. Integrated annotation and analysis of genetic variants from next-generation sequencing studies with variant tools. Bioinformatics. 2012;28:421–2.
3. Adzhubei I, Jordan DM, Sunyaev SR. Predicting functional effect of human missense mutations using PolyPhen-2. Curr Protoc Hum Genet. 2013;Chapter 7:Unit 7.20.
4. Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat Protoc. 2009;4:1073–81.
5. Kellis M, Wold B, Snyder MP, Bernstein BE, Kundaje A, Marinov GK, et al. Defining functional DNA elements in the human genome. Proc Natl Acad Sci U S A. 2014;111:6131–8.
6. Boyle AP, Hong EL, Hariharan M, Cheng Y, Schaub MA, Kasowski M, et al. Annotation of functional variation in personal genomes using RegulomeDB. Genome Res. 2012;22:1790–7.
7. Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet. 2014;46:310–5.
8. Davydov EV, Goode DL, Sirota M, Cooper GM, Sidow A, Batzoglou S. Identifying a high fraction of the human genome to be under selective constraint using GERP++. PLoS Comput Biol. 2010;6:e1001025.
9. NCBI: National center for biotechnology information search database. http://www.ncbi.nlm.nih.gov/.
10. Cunningham F, Amode MR, Barrell D, Beal K, Billis K, Brent S, et al. Ensembl 2015. Nucleic Acids Res. 2015;43(Database issue):D662–9.
11. Harrow JL, Steward CA, Frankish A, Gilbert JG, Gonzalez JM, Loveland JE, et al. The vertebrate genome annotation browser 10 years on. Nucleic Acids Res. 2014;42:D771–9.
12. Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, et al. GENCODE: the reference human genome annotation for the ENCODE project. Genome Res. 2012;22:1760–74.
13. Kanehisa M, Goto S. KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 2000;28:27–30.
14. Nishimura D. BioCarta. Biotech Softw Internet Rep. 2001;2:117–20.
15. Schaefer CF, Anthony K, Krupa S, Buchoff J, Day M, Hannay T, et al. PID: the pathway interaction database. Nucleic Acids Res. 2009;37:D674–9.
16. Kent Jr JW. Pathway-based analyses. BMC Genet. 2015;16 Suppl 3:S5.
17. Gibson J, Morton NE, Collins A. Extended tracts of homozygosity in outbred human populations. Hum Mol Genet. 2006;15:789–95.
18. Hildebrandt F, Heeringa SF, Rüschendorf F, Attanasio M, Nürnberg G, Becker C, et al. A systematic approach to mapping recessive disease genes in individuals from outbred populations. PLoS Genet. 2009;5:e1000353.
19. Browning SR, Thompson EA. Detecting rare variant associations by identity-by-descent mapping in case-control studies. Genetics. 2012;190:1521–31.
20. Balliu B, Uh HW, Tsonaka R, Boehringer S, Helmer Q, Houwing-Duistermaat JJ. Combining information from linkage and association mapping for next-generation sequencing longitudinal family data. BMC Proc. 2014;8 Suppl 1:S34.
21. Lee S, Abecasis GR, Boehnke M, Lin X. Rare-variant association analysis: study designs and statistical tests. Am J Hum Genet. 2014;95:5–23.

22. Schaid DJ. Genomic similarity and kernel methods I: advancements by building on mathematical and statistical foundations. Hum Hered. 2010;70:109–31.
23. Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. Rare variant association testing for sequencing data using the sequence kernel association test (SKAT). Am J Hum Genet. 2011;89:82–93.
24. Li B, Leal SM. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. Am J Hum Genet. 2008;83:311–21.
25. Neale BM, Rivas MA, Voight BF, Altshuler D, Devlin B, Orho-Melander M, et al. Testing for an unusual distribution of rare variants. PLoS Genet. 2011;7:e1001322.
26. Santorico SA, Hendricks AE. Progress in methods for rare variant association. BMC Genet. 2015;16 Suppl 3:S7.
27. Kim T, Wei P. Incorporating ENCODE information into association analysis of whole genome sequencing data. BMC Proc. 2015;9 Suppl 8:S34.
28. Zhang D, Cui H, Korkin D, Wu Z. Incorporation of protein binding effects into likelihood ratio test for exome sequencing data. BMC Proc. 2015;9 Suppl 8:S37.
29. Ionita-Laza I, Lee S, Makarov V, Buxbaum JD, Lin X. Sequence kernel association tests for the combined effect of rare and common variants. Am J Hum Genet. 2013;92:841–53.
30. Malzahn D, Friedrichs S, Bickeböller H. Comparing strategies for combined testing of rare and common variants in whole sequence and genome-wide genotype data. BMC Proc. 2015;9 Suppl 8:S36.
31. Genovese CR, Roeder K, Wasserman L. False discovery control with p-value weighting. Biometrika. 2006;93:509–24.
32. Roeder K, Wasserman L. Genome-wide significance levels and weighted hypothesis testing. Stat Sci. 2009;24:398–413.
33. Ho YY, Guan W, Basu S. Powerful association test combining rare variant and gene expression using family data from genetic analysis workshop 19. BMC Proc. 2015;9 Suppl 8:S33.
34. Almasy L, Dyer TD, Peralta JM, Jun G, Wood AR, Fuchsberger C, et al. Data for Genetic Analysis Workshop 18: human whole genome sequence, blood pressure, and simulated phenotypes in extended pedigrees. BMC Proc. 2014;8 Suppl 1:S2.
35. Blangero J, Teslovich TM, Sim X, Almeida MA, Jun G, Dyer TD, et al. Omics-squared: human genomic, transcriptomic and phenotypic data for Genetic Analysis Workshop 19. BMC Proc. 2015;9 Suppl 8:S2.
36. Almeida M, Blondell L, Peralta J, Kent JW, Jun G, Teslovich TM, et al. Independent test assessment using the extreme value distribution theory. BMC Proc. 2015;9 Suppl 8:S32.
37. Liu X-Q, Fazio J, Hu PZ, Paterson AD. Identity-by-descent mapping for diastolic blood pressure in unrelated Mexican Americans. BMC Proc. 2015;9 Suppl 8:S35.
38. Patterson N, Price AL, Reich D. Population structure and eigenanalysis. PLoS Genet. 2006;2:e190.
39. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. Genome Res. 2009;19:1655–64.
40. GRC: The Genome Reference Consortium. http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/.
41. The International HapMap Consortium. The international HapMap project. Nature. 2003;426:789–96.
42. Barrett JC, Fry B, Maller J, Daly MJ. Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics. 2005;21:263–5.
43. Sikić M, Tomić S, Vlahovicek K. Prediction of protein-protein interaction sites in sequences and 3D structures by random forests. PLoS Comput Biol. 2009;5(1):e1000278.
44. Schifano ED, Epstein MP, Bielak LF, Jhun MA, Kardia SL, Peyser P, et al. SNP set association analysis for familial data. Genet Epidemiol. 2012;36:797–810.
45. Dudbridge F, Gusnanto A. Estimation of significance thresholds for genome wide association scans. Genet Epidemiol. 2008;32:227–34.
46. Almasy L, Blangero J. Multipoint quantitative-trait linkage analysis in general pedigrees. Am J Hum Genet. 1998;62:1198–211.
47. Sidak Z. Rectangular confidence regions from means of multivariate normal distributions. J Am Stat Assoc. 1967;62:626–33.
48. Browning SR, Browning BL. Rapid and accurate haplotype phasing and missing data inference for whole genome association studies by use of localized haplotype clustering. Am J Hum Genet. 2007;81:1084–97.
49. Chen H, Malzahn D, Balliu B, Li C, Bailey JN. Testing genetic association with rare and common variants in family data. Genet Epidemiol. 2014;38 Suppl 1:S37–43.
50. Liu D, Lin X, Ghosh G. Semiparametric regression of multidimensional genetic pathway data: least-squares kernel machines and linear mixed models. Biometrics. 2007;63:1079–88.
51. Rasmussen CE, Williams CKI. Gaussian processes for machine learning. Cambridge: MIT Press; 2006.
52. Madsen BE, Browning SR. A groupwise association test for rare mutations using a weighted sum statistic. PLoS Genet. 2009;5:e1000384.
53. Chen Y-C, Carter H, Parla J, Kramer M, Goes FS, Pirooznia M, et al. A hybrid likelihood model for sequence-based disease association studies. PLoS Genet. 2013;9:e1003224.
54. Davies RB. Algorithm as 155: the distribution of a linear combination of chi-2 random variables. J R Stat Soc: Ser C: Appl Stat. 1980;29:323–33.
55. Brown MB. A method for combining non-independent, one-sided tests of significance. Biometrics. 1975;31:987–92.
56. Lee S, Wu MC, Lin X. Optimal tests for rare variant effects in sequencing association studies. Biostatistics. 2012;13:762–75.
57. Basu S, Pan W. Comparison of statistical tests for disease association with rare variants. Genet Epidemiol. 2011;35:606–19.

Hindawi
Computational and Mathematical Methods in Medicine
Volume 2017, Article ID 6742763, 17 pages
https://doi.org/10.1155/2017/6742763

Research Article

Pathway-Based Kernel Boosting for the Analysis of Genome-Wide Association Studies

Stefanie Friedrichs,[1] Juliane Manitz,[2,3] Patricia Burger,[1] Christopher I. Amos,[4] Angela Risch,[5,6,7] Jenny Chang-Claude,[8] Heinz-Erich Wichmann,[9,10,11] Thomas Kneib,[2] Heike Bickeböller,[1] and Benjamin Hofner[12,13]

[1] *Institute of Genetic Epidemiology, University Medical Centre, Georg-August University Göttingen, Göttingen, Germany*
[2] *Department of Statistics and Econometrics, Georg-August University Göttingen, Göttingen, Germany*
[3] *Department of Mathematics and Statistics, Boston University, Boston, MA, USA*
[4] *Department of Community and Family Medicine, Geisel School of Medicine, Dartmouth College, Lebanon, NH, USA*
[5] *Division of Molecular Biology, University of Salzburg, Salzburg, Austria*
[6] *Translational Lung Research Center Heidelberg (TLRC-H), Member of the German Center for Lung Research (DZL), Heidelberg, Germany*
[7] *Division of Epigenomics and Cancer Risk Factors, German Cancer Research Center (DKFZ), Heidelberg, Germany*
[8] *Division of Cancer Epidemiology, German Cancer Research Center (DKFZ), Heidelberg, Germany*
[9] *Institute of Medical Informatics, Biometry and Epidemiology, Chair of Epidemiology, Ludwig-Maximilians University, Munich, Germany*
[10] *Helmholtz Center Munich, Institute of Epidemiology II, Munich, Germany*
[11] *Institute of Medical Statistics and Epidemiology, Technical University Munich, Munich, Germany*
[12] *Department of Medical Informatics, Biometry and Epidemiology, Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany*
[13] *Section Biostatistics, Paul-Ehrlich-Institut, Langen, Germany*

Correspondence should be addressed to Stefanie Friedrichs; sfriedr2@gwdg.de

Received 10 February 2017; Revised 15 April 2017; Accepted 10 May 2017; Published 13 July 2017

Academic Editor: Angelo Facchiano

The analysis of genome-wide association studies (GWAS) benefits from the investigation of biologically meaningful gene sets, such as gene-interaction networks (pathways). We propose an extension to a successful kernel-based pathway analysis approach by integrating kernel functions into a powerful algorithmic framework for variable selection, to enable investigation of multiple pathways simultaneously. We employ genetic similarity kernels from the logistic kernel machine test (LKMT) as base-learners in a boosting algorithm. A model to explain case-control status is created iteratively by selecting pathways that improve its prediction ability. We evaluated our method in simulation studies adopting 50 pathways for different sample sizes and genetic effect strengths. Additionally, we included an exemplary application of kernel boosting to a rheumatoid arthritis and a lung cancer dataset. Simulations indicate that kernel boosting outperforms the LKMT in certain genetic scenarios. Applications to GWAS data on rheumatoid arthritis and lung cancer resulted in sparse models which were based on pathways interpretable in a clinical sense. Kernel boosting is highly flexible in terms of considered variables and overcomes the problem of multiple testing. Additionally, it enables the prediction of clinical outcomes. Thus, kernel boosting constitutes a new, powerful tool in the analysis of GWAS data and towards the understanding of biological processes involved in disease susceptibility.

1. Introduction

Many human diseases are complex in nature. They are caused by an interplay of several, often moderate genetic effects and environmental factors (i.e., demographic, clinical, and other nongenetic data [1]). Their genetic architecture is often analyzed in genome-wide association studies (GWAS). Herein, genetic information is represented by the genotypes of a multitude of single-nucleotide polymorphisms (SNPs) located across the whole genome. Numerous SNPs associated with various diseases have already been discovered in GWAS analyses; however they cannot account for the full heritability of the corresponding disease [2]. Different methods to approach this problem of *missing heritability* have been proposed, including the joint analysis of several SNPs representing a particular part of the genetic information, such as a gene or gene set.

Gene-set analysis methods facilitate the detection of associations between an individual's genetic information and a phenotype of interest, for example, disease status. The joint analysis of several genes often leads to increased power, as it reduces the overall number of conducted tests and assists in the detection of moderate associations [3]. Furthermore, the results are usually more meaningful, as they are based on functional units rather than on single SNPs. One form of gene-set analysis is the investigation of pathways, such as networks of interacting genes responsible for a specific cell function or regulation [4]. The proteins coded by genes within a pathway can enhance or reduce the expression of other genes, to which we refer as activation or inhibition. Thus, genes interact directly as well as indirectly in a series of interconnected steps within pathways. Different types of biological pathway exist, for example, involved in metabolism or signal transduction. Faults in function can occur and such malfunction of biological pathways may lead to disease onset and development.

Large sample sizes are required to detect weak genetic effects influencing disease risk. Thanks to technical advances and the formation of data-sharing consortia in particular, larger GWAS datasets have become available over recent years. However, genotyping and participant recruitment are still cost and work intensive. Especially in rare diseases, taking as an example the analysis of histological subtypes of a disease, it is very challenging to achieve sample sizes that result in adequate power in analyses [5]. Another challenge we face is to understand the biological meaning of detected associations. It is often difficult to interpret the results of GWAS analysis in the elucidation of the precise biological processes and corresponding functional units influencing disease susceptibility. Single-pathway analysis methods are often successful in the identification of genetic effects influencing disease susceptibility. However, they usually can not discriminate causal biological processes from isolated effects included in pathways due to gene overlap [6, 7]. Another limitation of many pathway analysis approaches is the lacking ability to predict the disease state, or other outcomes of interest, based on the identified genetic effects.

Kernel methods in statistics have already been demonstrated as dealing well with the challenges faced when analyzing GWAS data [8, 9]. They are capable of handling high-dimensional data, without requiring any direct specification of the functional relationship between genetic effects. Furthermore, kernel methods are computationally efficient and allow the straightforward incorporation of environmental covariates [9–11]. Kernels are used to calculate a quantitative value from genotype data, which may be interpreted as reflecting the genetic similarity between each pair of individuals. Different kernels have been proposed in the analysis of pathways [9, 12, 13]. While some kernels only evaluate SNP membership in genes, others can also adjust for differing gene numbers and sizes or even include gene interaction structures or other information (please refer to Materials and Methods and [13] for an overview). We focussed on the network-based kernel, as it allows us to include interaction structures and has been demonstrated as being superior in performance for interconnected effects [13].

We extend kernel-based analysis of GWAS data by integrating a network-based kernel function into a boosting framework, in order to identify genetic variation modulating disease susceptibility. Boosting emerged from the field of machine learning and was later transferred to statistical modelling. It implements an ensemble of many weak learners (so-called base-learners, simple models that are slightly improved over random guessing) to optimize the predictive accuracy of a model [14]. Since it is able to combine the power from several predictors with weak signals into a strong prediction set [15, 16], it may prove to be a powerful tool in the analysis of GWAS. Component-wise boosting enforces variable selection and includes additional effect regularization, which makes it especially useful for high-dimensional data [17]. Model-based boosting can be seen as an extension of classic boosting approaches (see, e.g., [18, 19]). Diverse base-learners, which represent special effect types, may be chosen and combined arbitrarily [20]. Thus, boosting allows the simultaneous inclusion of genetic information and demographic or other environmental data. This joint investigation of multiple variables allows taking into account correlations between different pathways and will likely facilitate discrimination of causal biological processes from effects included in pathways only due to gene overlap. The derived models can be assessed and interpreted directly. Our kernel boosting approach overcomes the problem of multiple testing thanks to its inherent variable selection property [21]. Thereby the overall gain in power in the analysis of GWAS supports the analysis of smaller samples and moderate-to-weak genetic effects. Of note, the main focus of boosting (as well as of other machine learning methods) is not on hypothesis testing but on the development of a multivariable prediction model.

We applied our approach to two GWAS datasets, one on lung cancer and one on rheumatoid arthritis. Lung cancer is one of the most common forms of cancer, especially in industrialized nations. It is responsible for the greatest proportion of deaths caused by cancer worldwide [22]. Although the exposure to tobacco is known to be the major risk factor for lung cancer susceptibility, a number of genetic influences have been revealed by many studies [23]. The actual number of known genetic influences, excepting some specific lung

cancer syndromes, is still limited, and each only accounts for a minor increase in disease risk. Rheumatoid arthritis is the most frequently occurring inflammatory disease of the joints, predominantly affecting the hands and feet. It is one of the major causes of disability and is strongly influenced by genetic factors in the human leukocyte antigen (HLA) region located on chromosome 6 [24, 25]. The investigation into these two diseases with different genetic architectures provides the ideal platform to evaluate the performance of our novel method.

In Section 2, we introduce the model structure utilized and describe the construction of network-based kernel functions. We provide a short introduction to boosting and derive the novel boosting algorithm with kernel-based base-learners. Section 3 comprises a description of the simulation study used to evaluate the method's performance and an overview of the application to rheumatoid arthritis and lung cancer GWAS datasets. The results of the simulation study and GWAS analyses are summarized in Section 4. Finally, we end the paper with a discussion and an outlook.

1.1. Software. We used the statistical software environment `R` [26] to perform all analyses unless stated otherwise. The methodological developments were implemented in the `R` packages `kangar00` [27] and `mboost` [28]. An exemplary application of the kernel boosting method to a simulated data set is given in Supplementary Material 2, available online at https://doi.org/10.1155/2017/6742763.

2. Materials and Methods

We aim to model the disease status of an individual, based on environmental covariates and genetic information obtained from GWAS. The genetic information given by the genotypes of different SNPs is mapped via genes to pathways. For each pathway, we compute a kernel matrix transforming the genotype vectors of each two individuals into a numeric value, which may be interpreted as the genetic similarity of the two individuals. Based on these matrices, we fit a kernel-based boosting model to identify relevant pathways and to find a prediction model for disease status. In the following paragraphs, we define all the relevant parts to this approach.

2.1. Model Definition and Notation. We assume an additive logistic regression model for the conditional probability of being a case for individual i, $i = 1, \ldots, n$:

$$\operatorname{logit}\left[P\left(y_i = 1 \mid \mathbf{x}_i, \mathbf{z}_i\right)\right] = \eta\left(\mathbf{x}_i, \mathbf{z}_i\right), \tag{1}$$

with additive predictor

$$\eta\left(\mathbf{x}_i, \mathbf{z}_i\right) = \mathbf{x}_i\boldsymbol{\beta} + f_1\left(\mathbf{z}_i\right) + \cdots + f_P\left(\mathbf{z}_i\right), \tag{2}$$

where y_i is the case-control indicator ($y_i = 0$ control; $y_i = 1$ case), $\mathbf{x}_i = (x_{i,1}, \ldots, x_{i,n_c})$ is the n_c dimensional environmental covariate vector, and $\mathbf{z}_i$ denotes the genotype vector of the n_s SNPs of the ith individual. Note that the non- or semiparametrically modelled genetic effects $f_p(\mathbf{z}_i)$ usually only depend on a pathway specific subset of SNPs, $\mathbf{z}_i^{(p)}$. However, for the sake of notational convenience we dropped the pathway index (p).

The vector $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_{n_c})^\top$ represents the regression coefficients (including an intercept β_0) related to the environmental covariates. They typically include information on age, sex, or other traits relevant to the disease investigated. The genotype variables $\mathbf{z}_i$ are coded as number of minor alleles, resulting in $z_{i,s} \in \{0, 1, 2\}$ for any SNP s and individual i. The nonparametric functions f_p, $p = 1, \ldots, P$, describe how the risk of being affected by the disease depends on the observed genotypes. Here, we aggregate the genotype information according to SNP membership in P different gene interaction pathways.

2.2. Network-Based Kernels. Liu et al. [10] introduced the kernel machine framework to the field of pathway analysis. Since genes in pathways can include complex interactions, nonparametric approaches are advisable. The logistic kernel machine test (LKMT) can model the effect of a pathway on a binary outcome nonparametrically, while including parametrically modelled covariates. In the resulting logistic regression model, the genetic influence is incorporated by a function from the reproducing kernel Hilbert space generated by a positive definite kernel function K.

In a genetic application, this kernel function is evaluated for the genotypes of each two individuals i and j, whereby the kernel matrix element $K_{ij} = K(\mathbf{z}_i, \mathbf{z}_j)$ is obtained. This value can be understood as the genetic similarity between the two individuals. To embed this definition into the mathematically well-defined framework of a reproducing kernel Hilbert space, the kernel matrix has to fulfill some requirements: it has to be quadratic, symmetric, and positive semidefinite. A variety of kernel functions are available. In the pathway-based analysis of GWAS data, a network-based kernel can be used, which is able to incorporate the pathway topology [13].

Assume $\mathbf{Z} = (\mathbf{z}_1, \ldots, \mathbf{z}_n)^\top$ denotes the $n \times n_s$ pathway specific genotype matrix consisting of the genotype vectors $\mathbf{z}_i$, which include only the SNPs relevant for pathway p, for all $i = 1, \ldots, n$ individuals. Then, the network-based kernel is defined by

$$\mathbf{K} = \mathbf{ZANA}^\top\mathbf{Z}^\top, \tag{3}$$

where $\mathbf{A}$ is an $n_s \times n_g$ matrix mapping all SNPs to the n_g investigated genes (including an adjustment to account for differing sizes of genes) and $\mathbf{N}$ represents the (modified) $n_g \times n_g$ matrix network adjacency matrix of gene interactions. To ensure positive semidefiniteness of the kernel, the network adjacency matrix is processed in a number of preparatory steps: if a gene is not represented by any SNPs in the investigated GWAS dataset, it cannot be considered in the analysis. To prevent loss of information about interactions in the network, genes which have previously been connected via the omitted gene will be linked directly. The new link's weight is determined in a multiplicative fashion, based on the weights of the two omitted links. For a graphical representation refer to Figure 1. The resulting matrix is further mirrored along

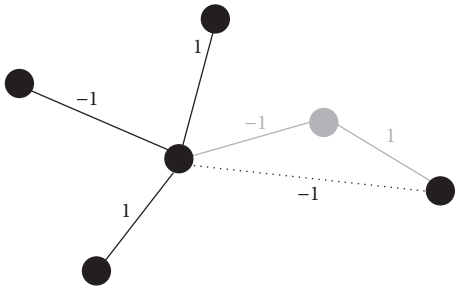

Figure 1: Graphical representation of rewiring step in data preparation. Nodes are representing genes in the pathway, while edges indicate interactions between the corresponding genes. Assume the gene depicted in grey is not represented by any genetic markers in the considered study and thus cannot be analyzed. To retain information about the (indirect) interaction of the two genes previously linked to the omitted gene, a new direct link is established between them. Its interaction type is determined by multiplication of the weights inherent to the two dropped links.

its diagonal and transformed to obtain positive semidefiniteness. The applied transformation is given by

$$\rho\mathbf{N} + (1-\rho)\,\mathbf{I}, \tag{4}$$

where $\mathbf{I}$ denotes the identity matrix and ρ is a weight based on the smallest eigenvalue of $\mathbf{N}$. For more details, see [13].

2.3. Model-Based Boosting. Model fitting in general aims to minimize the loss when relating observed responses y_i to an estimated model characterized by the additive predictor $\eta_i := \eta(\mathbf{x}_i, \mathbf{z}_i)$ as defined in (2). Thus, boosting minimizes the empirical risk

$$\frac{1}{n}\sum_{i=1}^{n} -l(y_i, \eta_i), \tag{5}$$

where $-l(\cdot)$ denotes a suitable loss function. Here, we use the negative binomial log-likelihood as loss function, which results in additive logistic regression models in analogy to the LKMT. In general, the loss function characterizes the model and can be defined in terms of a suitable negative log-likelihood or other appropriate loss functions, for example, the quadratic error loss for Gaussian regression or the absolute error loss for quantile regression. For an overview on loss functions see Hofner et al. [20]. Boosting solves this optimization problem via functional gradient descent by moving in the direction of the loss function's steepest descent along the additive effects of predictor (2). This can be seen in the following (simplified) algorithm:

(1) Initialize the additive predictor with $\widehat{\eta}_i^{[0]} = \overline{y}$, $i = 1, \ldots, n$, and all function estimates with $\widehat{f}_p^{[0]} = 0$, $p = 1, \ldots, P^+$. Note that P^+ includes all P kernels and possibly additional effects for environmental covariates.

(2) For $m = 1, \ldots, m_{\text{stop}}$ do the following:

(a) Compute the negative gradient of the loss function evaluated at the estimates of the previous iteration:

$$u_i^{[m]} = -\left.\frac{\partial(-l(y_i, \eta_i))}{\partial\eta}\right|_{\eta_i = \widehat{\eta}_i^{[m-1]}(\mathbf{x}_i, \mathbf{z}_i)}, \quad i = 1, \ldots, n. \tag{6}$$

(b) Estimate the negative gradient vector $\mathbf{u}^{[m]} = (u_1^{[m]}, \ldots, u_n^{[m]})$ *separately* for each effect in the additive predictor (2) by base-learners $\widehat{\mathbf{u}}^{[m]} = \widehat{\mathbf{f}}_p$, $p = 1, \ldots, P^+$, with $\widehat{\mathbf{f}}_p := (\widehat{f}_p(\mathbf{x}_i, \mathbf{z}_i))_{i=1,\ldots,n}$ by fitting simple regression models via (penalized) least squares. Thus, each base-learner regresses the negative gradient vector $\mathbf{u}^{[m]}$ separately on each of the predictors.

(c) Choose the best-fitting base-learner $\widehat{\mathbf{f}}_{p^\star}$ with the minimal residual sum of squares.

(d) Compute the update for the additive predictor by adding the best-fitting base-learner with a step-length factor $0 < \nu \leq 1$:

$$\widehat{\boldsymbol{\eta}}^{[m]} = \widehat{\boldsymbol{\eta}}^{[m-1]} + \nu \cdot \widehat{\mathbf{f}}_{p^\star}. \tag{7}$$

The corresponding update of function estimate $\widehat{\mathbf{f}}_{p^\star}$ is given by

$$\widehat{\mathbf{f}}_{p^\star}^{[m]} = \widehat{\mathbf{f}}_{p^\star}^{[m-1]} + \nu \cdot \widehat{\mathbf{f}}_{p^\star}, \tag{8}$$

while

$$\widehat{\mathbf{f}}_p^{[m]} = \widehat{\mathbf{f}}_p^{[m-1]}, \tag{9}$$

for all $p \neq p^\star$.

Note that each base-learner $\widehat{\mathbf{f}}_p$ usually depends on only one environmental covariate or one pathway based on a suitable subset of the genotypes of $\mathbf{z}$. However, other dependencies are also possible. For details on the algorithm, see [20]. A graphical display of the main features of the kernel boosting algorithm is given in Figure 2.

2.4. Model Tuning. The major tuning parameter of the functional gradient descent boosting algorithm is the number of iterations m_{stop}. We usually choose m_{stop} via cross-validation methods (such as bootstrap, k-fold cross-validation, or subsampling) in order to avoid overfitting: one fits the model on the selected subset of the data and chooses m_{stop} such that it minimizes the empirical risk on the data that were not used to estimate the model. Subsampling is recommended to avoid overly complex models [29]. The step-length ν is another tuning parameter. In general it is of minor importance as long as it is relatively small. It determines the trade-off between speed of convergence and variable selection ability and is typically set to 0.1 [30].

The current estimate $\widehat{\boldsymbol{\eta}}^{[m]}$ of the additive predictor $\boldsymbol{\eta}$ usually depends on only a subset of the possible predictors: as we select the best-fitting base-learner in each step and choose m_{stop} such that it maximizes prediction accuracy

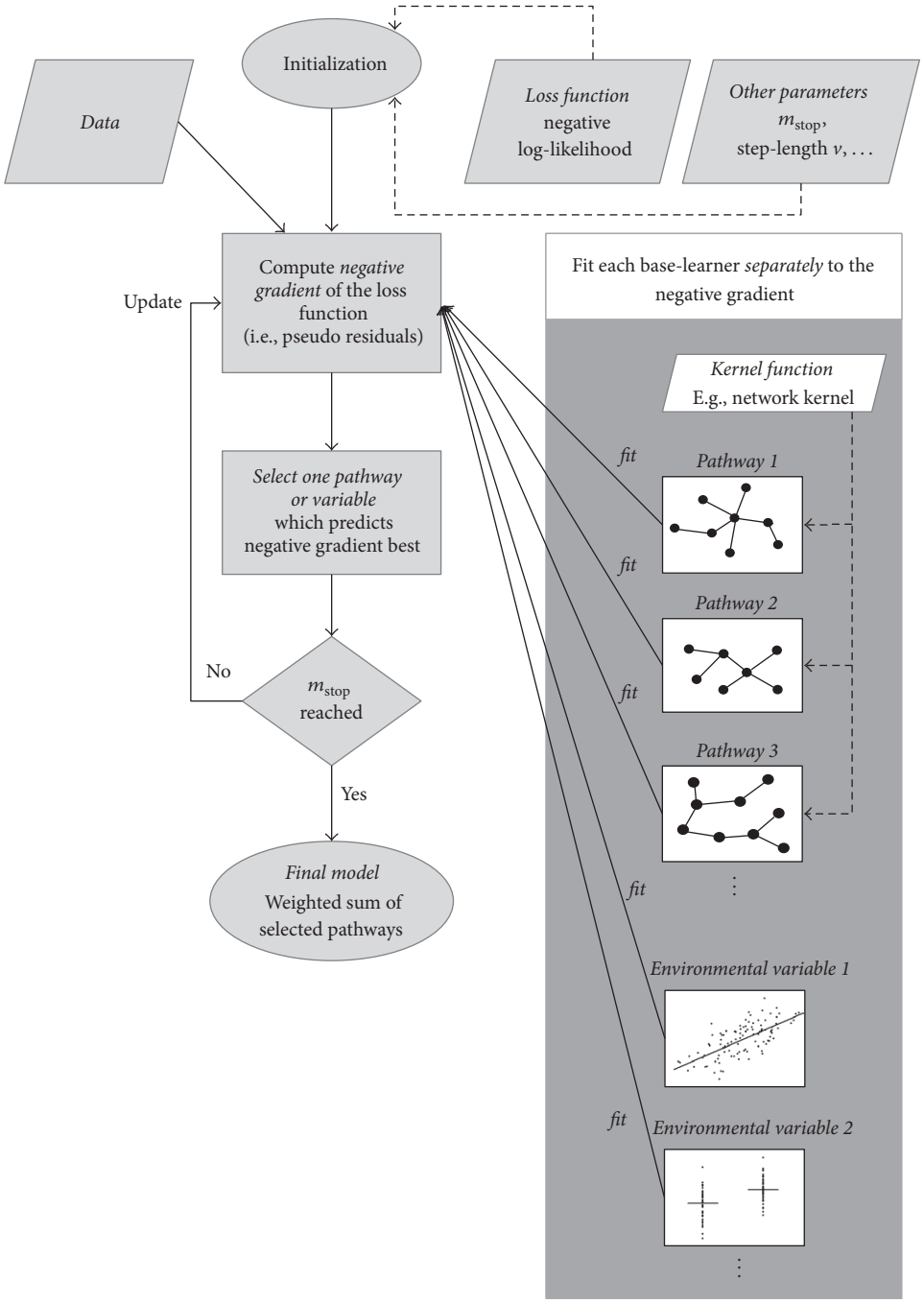

Figure 2: Graphical representation of the main features of the kernel boosting algorithm.

(i.e., usually relatively small so that not all base-learners are selected), boosting selects base-learners and thus variables. In our approach, we exploit this behaviour to identify genetic associations. Note that a base-learner can be selected multiple times. Hence, its function estimate $\widehat{\mathbf{f}}_p$, $p \in 1, \ldots, P^+$, is the weighted sum with weights ν of the individual estimates over all iterations in which the base-learner was selected (see (8)).

2.5. Boosting with Network-Based Kernel as Base-Learner. To incorporate genotype data, aggregated to represent a particular pathway, we utilize kernel-based base-learners. Using a kernel function K, we transform the definition of the genotypic information of all pairs of individuals to $K_{ij} = K(\mathbf{z}_i, \mathbf{z}_j)$, $i, j = 1, \ldots, n$, as mentioned before, and collect them in the kernel matrix $\mathbf{K}$. With this matrix, we can estimate

$$f(\mathbf{Z}) = \mathbf{K}\boldsymbol{\gamma} = \mathbf{ZANA}^\top \mathbf{Z}^\top \boldsymbol{\gamma}, \tag{10}$$

The function $f(\mathbf{Z})$ is used to map the influence of SNP profiles to the clinical outcome (see (2)). As we expect

patients with similar SNP profiles to have similar outcomes, we aim to discourage large differences in $f(\mathbf{Z})$ for genetically similar individuals. According to the standard penalization approaches in the boosting context, we thus introduce an additional smoothness constraint on the coefficient vector $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_n)^\top$ based on the kernel distances:

$$\mathcal{J}(\boldsymbol{\gamma}) = \boldsymbol{\gamma}^\top \mathbf{K} \boldsymbol{\gamma}. \tag{11}$$

Thus, we define a separate kernel base-learner for each pathway in the boosting framework. Using the negative gradient vector $\mathbf{u}^{[m]} = (u_1^{[m]}, \ldots, u_n^{[m]})$ from the mth boosting iteration, we can estimate the coefficient vector $\boldsymbol{\gamma}$ of each base-learner (see step 2b of the algorithm) via penalized least squares

$$\widehat{\boldsymbol{\gamma}}^{[m]} = \left(\mathbf{K}^\top \mathbf{K} + \lambda \mathbf{K}\right)^{-1} \mathbf{K}^\top \mathbf{u}^{[m]}, \tag{12}$$

where we dropped the function index p for the sake of notational convenience. Note that kernel matrix $\mathbf{K}$ plays the role of design matrix as well as the role of penalty matrix with penalty parameter λ, which governs the smoothness of the estimate. Usually, the penalty parameter λ is chosen such that all base-learners have equal degrees of freedom to allow an unbiased selection. A common choice is four degrees of freedom if only smooth effects are used or one degree of freedom if linear effects are to be included; see Hofner et al. [21] for details.

In some rare cases, the derived kernel matrix $\mathbf{K}$ is numerically not positive semidefinite (i.e., minimal deviations might occur), even though this should theoretically always be the case. To ensure a numerically positive semidefinite matrix $\mathbf{K}$, we apply transformation (4) not only to $\mathbf{N}$ but also on the resulting kernel matrix $\mathbf{K}$. The proposed approach is very fast and results in smaller absolute differences in the matrix elements than alternatives such as the procedure suggested by Higham [31] (results not shown).

For numerical reasons, we reformulate the estimation problem from (12) by multiplying the design matrix with the inverse of the square root of the penalty matrix [32]. Thus, we obtain the design matrix

$$\widetilde{\mathbf{K}} = \mathbf{K}\mathbf{K}^{-1/2}, \tag{13}$$

while the penalty matrix simplifies to the identity matrix $\mathbf{I}$. Now, we can equivalently write

$$\widehat{\boldsymbol{\gamma}}^{[m]} = \left(\widetilde{\mathbf{K}}^\top \widetilde{\mathbf{K}} + \lambda \mathbf{I}\right)^{-1} \widetilde{\mathbf{K}}^\top \mathbf{u}^{[m]}. \tag{14}$$

A similar approach based on radial basis functions, which, for example, uses correlation functions to measure distances, was introduced to the boosting framework by Hofner [33].

2.6. Model Prediction Using Kernels. Boosting specifically aims to optimize prediction accuracy. As in all regression models, we can use the estimated coefficients to predict the outcome for new observations. However, some extra work is required to set up the kernel, that is, the design matrix, with new genotype data $\mathbf{Z}^* = (\mathbf{z}_1^*, \ldots, \mathbf{z}_{n^*}^*)^\top$. In this context, the kernel can be understood to compute the similarity between genotype information of individuals to be predicted and the observations used to fit the model, the training data $\mathbf{Z}$ itself. Thus,

$$\mathbf{K}^\star = \left(K\left(\mathbf{z}_i^\star, \mathbf{z}_j\right)\right)_{i=1,\ldots,n^*,\ j=1,\ldots,n} = \mathbf{Z}^\star \mathbf{A} \mathbf{N} \mathbf{A}^\top \mathbf{Z}^\top. \tag{15}$$

The resulting kernel $\mathbf{K}^*$ has the dimension $n^* \times n$, with n^* being new and n previously used observations. Note that kernel matrix $\mathbf{K}^*$ must no longer be of full rank nor be positive semidefinite. Using $\mathbf{K}^*$, we can predict the effect of a pathway on the outcome as

$$\widehat{f}(\mathbf{Z}^*) = \mathbf{K}^* \widehat{\boldsymbol{\gamma}}, \tag{16}$$

where $\widehat{\boldsymbol{\gamma}}$ is obtained as the weighted sum with weights ν over the estimates from (14) for all iterations in which the pth base-learner was selected (see (8)).

2.7. Incorporation of Environmental Covariates. To incorporate environmental variables into the boosting model, we can choose different base-learners suited to different types of effect. Linear effect base-learners are suited to a continuous covariate x such as patient age, while categorical effect base-learners facilitate the incorporation of categorical environmental variables such as gender. For details on inclusion of environmental variables, refer to [20].

With the inclusion of environmental variables as base-learners, these are also subject to the selection process inherent to boosting and compete with the pathway-based genetic effects. However, one usually wishes to consider only the added effect of genetic pathways. To ascertain that the model is corrected for environmental variables, one may include them as mandatory effects. This can be done by fitting a standard logistic regression model for the effect of the environmental variables on the clinical outcome and using the estimates as a start model (offset) for the boosting algorithm (see [34, 35]). This approach is very similar to the LKMT procedure, which tests if the logistic regression model can be improved via addition of a nonparametric effect incorporating a particular pathway.

3. Simulations and Applications

3.1. Simulation Study. To evaluate the performance of kernel boosting, we conducted a simulation study based on simulated SNP data in combination with gene networks from existing biological pathways. Pathway information was extracted from the Kyoto Encyclopedia of Genes and Genomes (KEGG) [36]. For simulation purposes, we considered a sample of 50 networks, randomly chosen from the total of 284 pathways available in January 2015. Please refer to Figure 3 for a list of these pathways and refer to Table 1 for their network topology characteristics. The primary aim of this study was to determine whether kernel boosting can detect associated pathways and is able to distinguish them from noninfluential pathways. Thus, we investigated the method's performance on data without genetic effects (null

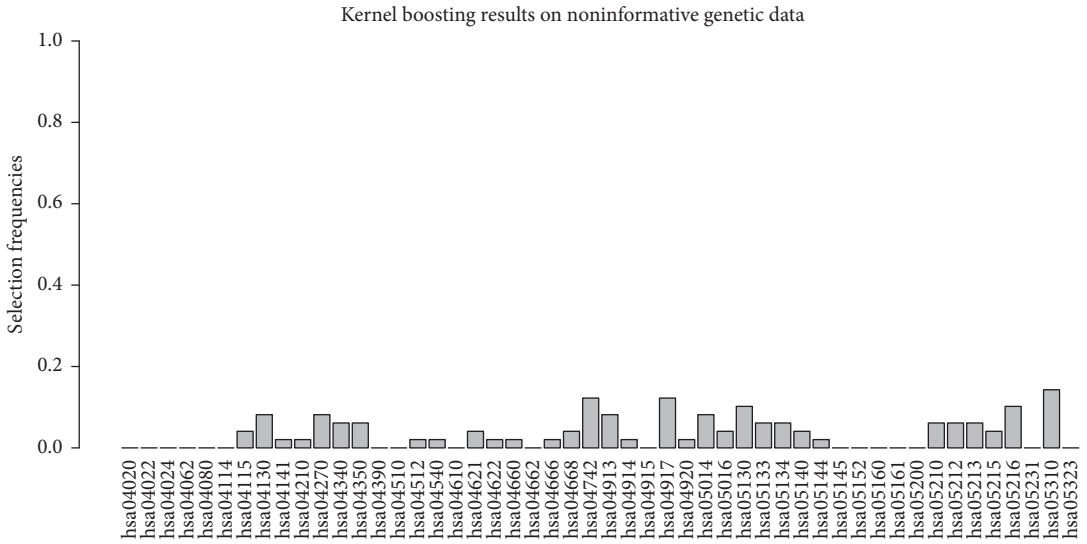

Figure 3: Relative frequency of datasets in which a pathway was selected for 50 pathways in the noninformative simulation scenario.

Table 1: Description of network properties for pathway topology of pathways used in simulations, compared to the properties of the two effect pathways hsa04020 and hsa04022. Nodes equal the number of included genes, links give the number of interactions, inhibition links the count of interactions of inhibiting type, the average degree of a node is the mean number of adjacent edges, density is the ratio between numbers of existing links and possible links, diameter denotes the distance to the farthest node in the graph, transitivity (also called cluster coefficient) calculates the probability of adjacent vertices of a vertex being connected, and signed transitivity considers the type of interaction in this calculation.

	Min	Mean	Median	Max	hsa04020	hsa04022
Nodes	29.00	103.60	86.5	398.00	180.00	167.00
Links	1.00	197.81	87.5	1493.00	297.00	372.00
Inhibition links	0.00	27.08	10.50	148.00	7.00	67.00
Average degree	0.07	3.18	2.36	15.62	3.30	4.46
Density	0.00	0.03	0.03	0.16	0.02	0.03
Inhibition degree	0.00	0.52	0.24	2.62	0.08	0.80
Diameter	1.00	7.36	7.00	18.00	6.00	7.00
Transitivity	0.00	0.02	0.00	0.14	0.00	0.03
Signed transitivity	−0.02	0.01	0.00	0.10	0.00	0.03

case) including 1000 individuals and in six effect scenarios, differing in effect strengths (relative risk of 1.1 and 1.5 per allele) and sample sizes ($n \in \{500, 1000, 2000\}$ with a 1 : 1 ratio of cases to controls). Datasets for all scenarios were simulated for 100 replications. Note that these scenarios are small compared to typically available sample sizes nowadays. The reason can be found in the computational demands of the method for an insightful number of replications. Accordingly, comparably strong effects of markers were chosen to match the sample sizes used in our simulations.

For each simulated dataset, we fitted a boosting model with pathway kernels. In order to tune the model, that is, to derive the optimal number of boosting steps m_{stop}, we used 20-fold subsampling for each model on each of the datasets with a maximum number of 200 iterations. Using the network-based kernel function in both methods, we compared the results from our kernel boosting approach on multiple pathways to those obtained from the single-pathway LKMT [9–11]. Additional simulations with cross-validated models and a maximum number of up to 1000 iterations were conducted to gain more insight into the proposed algorithm and are presented in Supplementary Material 1, Section A.

All genotypes were simulated with the help of a reference dataset from the International HapMap Consortium [37]. The reference data include 1,184 individuals of European descent (CEU) and a total of 1,440,616 SNPs, of which 116,565 are located on chromosome one. For each gene included in at least one of the 50 selected pathways, we defined a *pseudogene* to represent the gene within our simulations. Such a *pseudogene* was a randomly selected DNA segment on chromosome one of the reference data including five different SNPs. Between each two sampled regions, we ensured a distance of at least 100 kilo base pairs to prevent distortive LD correlations between them [38]. The location of *pseudogenes* was left unchanged for all simulations, resulting in a realistic correlation structure for all simulation scenarios. In each of

Table 2: Counts of included influential genes within pathways used for simulation purposes. Pathways without simulated causal genes are not displayed.

KEGG id	Name of pathway	Effect genes included
hsa04020	Calcium signaling pathway	4
hsa04022	cGMP-PKG signaling pathway	5
hsa04024	cAMP signaling pathway	1
hsa04080	Neuroactive ligand-receptor interaction	2
hsa04270	Vascular smooth muscle contraction	2
hsa04540	Gap junction	2
hsa04610	Complement and coagulation cascades	1
hsa05200	Pathways in cancer	2

the 100 simulation runs, new genotype data for a total of 11,665 SNPs in 2,333 *pseudogenes* were simulated using the HAPGEN2 software. This software generates new haplotype data by combining a given set of reference haplotypes with previously simulated data. The detailed procedure is described in [39].

In the null case, noninformative genetic data were simulated for 1000 individuals. In each replication, new genotypes without association signals were generated for 11,665 SNPs. The disease status was assigned at random with 0.5 binomial probability of being a case, completely independent of genotype information. In each of the six effect scenarios, genotype data for a previously chosen equal number of cases and controls were simulated such that two pathways affected disease status. Association signals were included in three genes per causal pathway. In each of the resulting six genes, two randomly selected SNPs were chosen to be influential on the binary clinical outcome. Within one simulation scenario, all associated SNPs had the same effect strength and for each SNP the minor allele was influential. All effects were simulated as additive. To simplify the evaluation, we decided not to include environmental variables in these settings.

We chose two typical pathways (KEGG ids *hsa04020* and *hsa04022*) to include causal genes. In accordance with the findings in [13], the influential genes in the two causal pathways were chosen to be interconnected within the corresponding pathway. Here, we additionally sampled one effect gene in each pathway, with the probability of being selected set to its betweenness centrality. Betweenness centrality measures the amount of shortest connections between each two genes in the network passing through the gene. Different studies have indicated that genes in topologically relevant positions of a pathway are more likely to be involved in disease association [40]. Two neighbouring genes of the sampled gene were randomly chosen to complete the connected scenario. In hsa04020, the genes GNA11, TACR1, and BDKRB2 were simulated to include SNPs influencing disease susceptibility. For hsa04022, genetic effects were placed on the genes PRKG2, ATP2B2, and KCNU1. For each of these genes, two SNPs were simulated as being influential on disease status. Note that existing biological pathways can have genes in common. Thus, beside our two pathways chosen to include influential effects, six additional pathways contain association signals. Refer to Table 2 for an overview of influential genes included in simulation pathways.

Application: GWAS for Rheumatoid Arthritis and Lung Cancer. We considered the German Lung Cancer study (GLC) with 488 cases and 478 controls, based on the data of participants taken from the following three individual studies: Lung Cancer in the Young (LUCY), a population-based multicentre study run by the Helmholtz Zentrum Munich, and the University Medical Centre of the Georg-August-University in Goettingen. This study includes data of lung cancer patients under the age of 51 and family members recruited in German hospitals [41, 42]. The Heidelberg lung cancer case-control study, conducted by the German Cancer Research Centre (DKFZ) and the Thoraxklinik in Heidelberg, Germany, recruited cases and controls in a hospital-based study [43]. Additional controls were provided by Cooperative Health Research in the Augsburg Region (KORA), a population-based genome-wide study carried out by the Helmholtz Zentrum Munich [44]. A subset of the study participants of these three studies was chosen to form the German Lung Cancer GWAS. These individuals were genotyped on a HumanHap 550K SNP chip.

The second GWAS is a rheumatoid arthritis study of the North American Rheumatoid Arthritis Consortium (NARAC). It includes 868 cases from New York hospitals, in which rheumatoid arthritis was diagnosed based on the criteria of the American College of Rheumatology. Additionally, 1,194 controls matching in self-reported ethnic background were collected. All individuals were genotyped with the HumanHap500v1 array [45, 46].

For the rheumatoid arthritis study, we utilized gender as environmental covariate. In the lung cancer study, age and smoking exposure, measured in pack years, were also considered. To determine the pack year, one multiplies the number of packs of cigarettes smoked per day by the number of years an individual has smoked.

All GWAS data were subjected to strict quality control. Only individuals with a genotype call rate of at least 95% were considered. SNPs with more than 10% missing values or with a minor allele frequency (MAF) below 0.1% were excluded from further analysis. Missing values in remaining markers were imputed with BEAGLE [47]. No SNPs beyond

Table 3: Characteristics of analyzed GWAS datasets. Numbers of case and control individuals after quality control and SNP numbers for several analysis stages are displayed. Preprocessing of SNPs included quality control of genotype data, as well as updating genomic SNP positions according to the latest information (genomic build 38). The last column indicates the total number of all SNPs annotated to a pathway under investigation.

Study	Cases/controls	SNPs genotyped	SNPs after preprocessing	SNPs in analysis
Lung cancer	467/468	561,466	533,062	148,938
Rheumatoid arthritis	866/1189	545,080	491,695	137,839

the original chip were imputed. The base pair positions of all SNPs were updated to NCBI build 38 using the Ensembl database [48], which was accessed using the `R` package `biomaRt` [49, 50]. Gene start and end positions were extracted from the same database, also using NCBI build 38. SNPs with no unique position were excluded. Refer to Table 3 for an overview of the study characteristics. Note that, during analysis, only SNPs mapped to genes within pathways were considered. The assignment of SNPs to genes was based on their base pair location and gene boundaries. SNPs closely located to each other are often in linkage disequilibrium (LD). For SNP annotation, we specified gene regions including LD-blocks extending beyond gene boundaries, as recommended in [51].

The KEGG database groups pathways in disjoint subsets according to their biological functionality. In the analysis of the rheumatoid arthritis and lung cancer data, we used a subgroup of 73 pathways connected to human diseases (see Table 4). The information on this group of pathways was downloaded in April 2016. An offset model containing only the environmental covariates was fitted for each of the studies to serve as start model for the kernel boosting of pathways.

For each pathway analyzed, the network-based kernel function with 4 degrees of freedom served as base-learner. The optimal number of iterations m_{stop} was derived via 20-fold subsampling and the default step length of 0.1 was used. For the purpose of comparison, each of the pathways considered in GWAS data analysis was also tested individually on the corresponding data using the LKMT. The same environmental variables that were used in the offset model for boosting were also considered for the LKMT. Prediction accuracy was measured by the misclassification rate and the area under the ROC curve (AUC) for both datasets. Of note, prediction accuracy is influenced by the applied model but also by the dataset at hand, that is, the amount of information contained in the data. Additionally, we provided the cross-validation results, that is, the (average) negative binomial likelihood on the data that was not used for model fitting (see Supplementary Material 1, Section B, for these results).

4. Results

4.1. Simulation Results. We compared the number of pathways each approach identified as associated with disease risk and considered the respective overlap in the results. The noninformative genetic data simulation comprised genotype data for 50 pathways and 1,000 individuals. Figure 3 displays the percentage of runs in which a pathway was selected. We can observe that the application of kernel boosting to these data does not lead to a high selection frequency for any pathway. Selection of pathways appears to be distributed randomly across all networks, not suggesting any clearly recognizable association with disease status. Note that, in kernel boosting, we do not conduct tests to evaluate the pathways' influence but select pathways based on their predictive performance. Thus, we cannot calculate a type I error to evaluate our method's performance. However, we can quantify the empirical type I error. Within 100 simulation runs on 50 pathways, a total number of 88 false selections occurred. Thus, a pathway was falsely selected in 1.76% of all possible cases. In 51 out of the 100 simulation runs, no single pathway was chosen by the algorithm. Hence, we conclude that kernel boosting can be trusted to reliably avoid false positive selections in noninformative data.

Figures 4 and 5 compare the results of effect simulations with a relative risk of 1.5 per allele for 1,000 cases and 1,000 controls to those for 250 cases and 250 controls. (a) in each figure contains barplots indicating selection frequencies of the 50 pathways across all simulation runs when applying kernel boosting to the corresponding simulation scenario. (b) compares these results with the selection frequencies using the LKMT. Here, both the percentages of results with a p value below 0.05 (lighter grey bars) and those with p values below the Bonferroni-corrected significance level of 0.001 (darker grey bars) are indicated. Pathways containing influential genes are additionally highlighted in italics.

The results of kernel boosting in the sample of 2,000 individuals (Figure 4(a)) display three pathways clearly identified as influential on the clinical outcome, as their selection frequency is close to 100%. These are the pathways originally chosen to include genetic effects, *hsa04020* and *hsa04022*, and the pathway *hsa04610*. It seems that the latter pathway is able to depict some of the information of the influential gene more effectively than the causal pathway for which it was originally simulated. This can be explained, as hsa04610 has the highest transitivity (0.14), also known as global clustering coefficient, of all simulation pathways and contains an effect gene. As the network kernel was designed to work especially well in detection of interconnected genetic effects, the causal gene is identified very well in the pathway when using this base-learner. Note that the same pathway did not stand out in the noninformative simulation scenario. Thus, we conclude that high transitivity facilitates the detection of causal effects when using the network-based kernel but does not lead to false positives (i.e., here, pathways which do not contain any effect gene). Several other pathways were also selected, but only with very low frequencies. In the same simulation scenario, the LKMT had very high

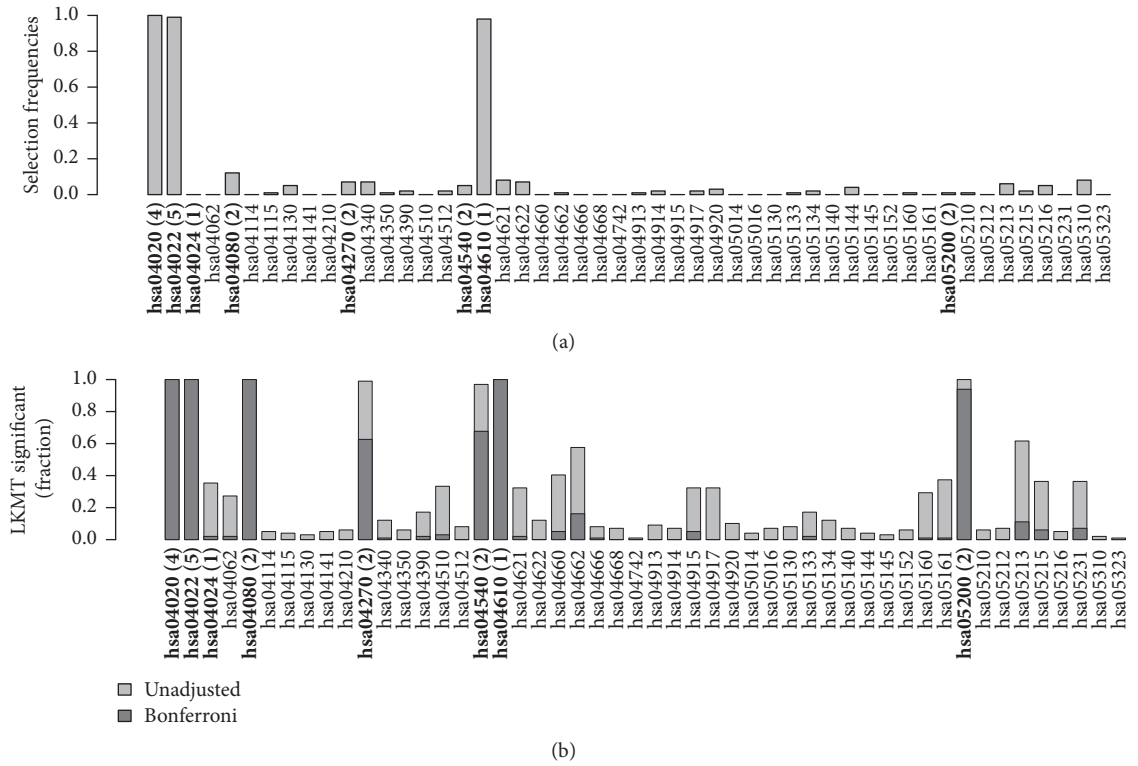

Figure 4: Relative frequency of datasets in which a pathway was selected using (a) kernel boosting (n = 2000, RR = 1.5) and (b) LKMT (n = 2000, RR = 1.5) for a sample size of 2000 individuals. Pathways including effect genes are labeled in bold; numbers in brackets denote the count of included influential genes within the pathway. All effects were simulated with a relative risk of 1.5 per allele.

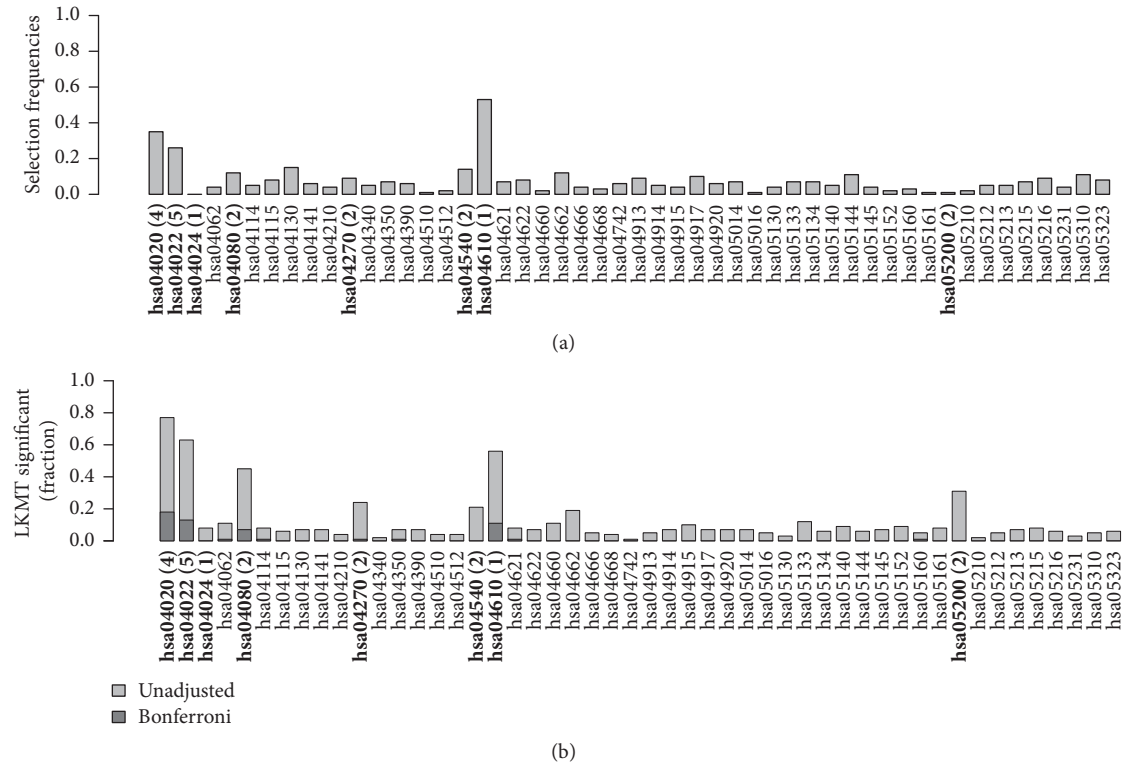

Figure 5: Relative frequency of datasets in which a pathway was selected using (a) kernel boosting (n = 500, RR = 1.5) and (b) LKMT (n = 500, RR = 1.5) for a sample size of 500 individuals. Pathways including effect genes are labeled in bold; numbers in brackets denote the count of included influential genes within the pathway. All effects were simulated with a relative risk of 1.5 per allele.

Table 4: KEGG pathways in the human diseases class as downloaded in April 2016. Pathways are sorted according to p value, derived from LKMT application on the rheumatoid arthritis dataset, in ascending order. p values for pathways significantly associated after Bonferroni correction are listed. Pathways selected by kernel boosting on the same dataset are marked in italics. Pathways containing one or several genes belonging to the HLA complex are marked with an asterisk behind the id number.

KEGG id	Name of pathway	p value
hsa05133	Pertussis	1.562×10^{-32}
*hsa05150**	*Staphylococcus aureus infection*	1.029×10^{-30}
hsa04933	AGE-RAGE signaling pathway in diabetic complications	3.877×10^{-17}
*hsa05169**	*Epstein-Barr virus infection*	2.651×10^{-16}
hsa05144	*Malaria*	3.087×10^{-15}
hsa05206	*MicroRNAs in cancer*	3.969×10^{-15}
*hsa05330**	*Allograft rejection*	4.131×10^{-12}
hsa05200	*Pathways in cancer*	7.695×10^{-11}
*hsa05166**	*HTLV-I infection*	1.344×10^{-11}
hsa05030	Cocaine addiction	1.353×10^{-11}
*hsa05323**	*Rheumatoid arthritis*	1.466×10^{-11}
hsa05310*	Asthma	2.268×10^{-11}
hsa05134	Legionellosis	1.699×10^{-05}
*hsa04940**	*Type I diabetes mellitus*	3.591×10^{-10}
hsa05031	Amphetamine addiction	3.735×10^{-10}
hsa05145*	Toxoplasmosis	4.555×10^{-10}
*hsa05203**	*Viral carcinogenesis*	1.814×10^{-09}
hsa05332*	Graft-versus-host disease	5.940×10^{-09}
hsa05020	*Prion diseases*	1.530×10^{-07}
hsa05143	African trypanosomiasis	2.114×10^{-07}
hsa05222	Small-cell lung cancer	3.782×10^{-07}
hsa05205	Proteoglycans in cancer	1.236×10^{-06}
*hsa05322**	*Systemic lupus erythematosus*	1.702×10^{-06}
hsa05161	*Hepatitis B*	1.757×10^{-06}
hsa05410	*Hypertrophic cardiomyopathy (HCM)*	1.980×10^{-06}
hsa05010	Alzheimer's disease	7.234×10^{-06}
hsa05142	Chagas disease (American trypanosomiasis)	1.048×10^{-05}
*hsa05168**	*Herpes simplex infection*	1.109×10^{-05}
hsa05012	*Parkinson's disease*	1.368×10^{-05}
hsa04932	Nonalcoholic fatty liver disease (NAFLD)	1.823×10^{-05}
hsa05321*	Inflammatory bowel disease (IBD)	2.124×10^{-05}
hsa04931	*Insulin resistance*	3.625×10^{-05}
hsa05219	*Bladder cancer*	4.133×10^{-05}
hsa05215	*Prostate cancer*	4.220×10^{-05}
hsa05202	Transcriptional misregulation in cancer	7.697×10^{-05}
hsa05220	Chronic myeloid leukemia	8.464×10^{-05}
hsa05146	Amoebiasis	1.003×10^{-04}
hsa05414	Dilated cardiomyopathy	1.014×10^{-04}
hsa05231	Choline metabolism in cancer	1.504×10^{-04}
hsa05032	*Morphine addiction*	1.672×10^{-04}
hsa05162	*Measles*	2.390×10^{-04}
hsa05214	Glioma	2.506×10^{-04}
hsa05164*	Influenza A	2.720×10^{-04}
*hsa05416**	*Viral myocarditis*	3.384×10^{-04}
hsa05132	*Salmonella infection*	5.147×10^{-04}
hsa05014	Amyotrophic lateral sclerosis (ALS)	5.568×10^{-04}
hsa04930	Type II diabetes mellitus	Not significant
hsa05218	Melanoma	Not significant
hsa05140*	Leishmaniasis	Not significant

Table 4: Continued.

KEGG id	Name of pathway	p value
hsa05213	*Endometrial cancer*	Not significant
hsa05211	*Renal cell carcinoma*	Not significant
hsa05340	*Primary immunodeficiency*	Not significant
hsa05160	*Hepatitis C*	Not significant
hsa05212	Pancreatic cancer	Not significant
hsa05016	Huntington's disease	Not significant
hsa05221	Acute myeloid leukemia	Not significant
hsa04950	*Maturity onset diabetes of the young*	Not significant
hsa05412	*Arrhythmogenic right ventricular cardiomyopathy (ARVC)*	Not significant
hsa05223	Non-small-cell lung cancer	Not significant
hsa05034	Alcoholism	Not significant
hsa05130	Pathogenic Escherichia coli infection	Not significant
hsa05120	Epithelial cell signaling in Helicobacter pylori infection	Not significant
hsa05131	*Shigellosis*	Not significant
hsa05204	*Chemical carcinogenesis*	Not significant
hsa05100	Bacterial invasion of epithelial cells	Not significant
hsa05216	Thyroid cancer	Not significant
hsa05152*	Tuberculosis	Not significant
hsa05210	Colorectal cancer	Not significant
hsa05230	Central carbon metabolism in cancer	Not significant
hsa05217	*Basal cell carcinoma*	Not significant
hsa05320*	Autoimmune thyroid disease	Not significant
hsa05033	Nicotine addiction	Not significant
hsa05110	Vibrio cholerae infection	Not significant

power to detect the two pathways simulated to affect disease risk, however, also detected other pathways including any of the causal genes on the Bonferroni-adjusted significance level (Figure 4(b)). Three of the six other effect-containing pathways were selected in almost 100% of the replications and two of the remaining ones in more than 60% and one other pathway which contained an effect gene was hardly selected.

Overall, this indicates that kernel boosting can identify the pathways with the most explanatory power with respect to disease status and is less likely than LKMT to select pathways due to overlapping effect genes (see [6] for a discussion). The reason can be found in the multivariate nature of the kernel boosting approach, in which pathways are not tested separately for their influence, but a multivariate model is fitted to incorporate multiple influential predictors at the same time.

Figure 5(a) reveals that the selection frequencies of associated pathways drop noticeably when sample size decreases. The same three pathways as in the larger sample reached the highest selection frequencies but here only between 20% and 60%. Simultaneously, the number of selections across nonassociated pathways increased slightly compared to the larger sample. This indicates that a reduction in sample size leads to less clear identification of the main influential pathways by kernel boosting. In Figure 5(b), we notice a similar behaviour of the selection frequency in LKMT analysis. Here again, the power to identify pathways, previously well detected in the larger sample, drops clearly with the smaller dataset. Regarding the percentage of detected pathways on the Bonferroni-corrected significance level, the drop is even more pronounced in the LKMT than for kernel boosting. This indicates that kernel boosting is less strongly influenced by sample size and may have greater potential in the identification of causal effects in smaller datasets for which the LKMT is underpowered.

Figures 6 and 7 compare the results of kernel boosting and the LKMT for differing effect sizes in equally sized samples of 1,000 individuals. The graphics are structured as Figures 4 and 5, with kernel boosting selection frequencies plotted in (a) and LKMT selection frequencies in (b). Figure 6 contains a simulation scenario with relative risk of 1.5 per causal allele and Figure 7 the results for a relative risk of 1.1 per allele. Again, pathways containing influential genes are additionally highlighted.

In the kernel boosting plot in Figure 6(a), the three pathways standing out in Figure 4 again reached very high selection frequencies. All three bars decreased slightly in size compared to the scenario with 2,000 individuals but still illustrate selections in more than 80% of simulation runs. Selection frequencies of the other effect pathways increased compared to the scenarios in Figure 4. However, as selections across noninfluential pathways occurred more frequently here, they cannot clearly be identified as influential based on their selection frequencies alone. In the LKMT analysis of this sample, the power to detect causal effects noticeably drops compared to the 2,000 individuals' sample

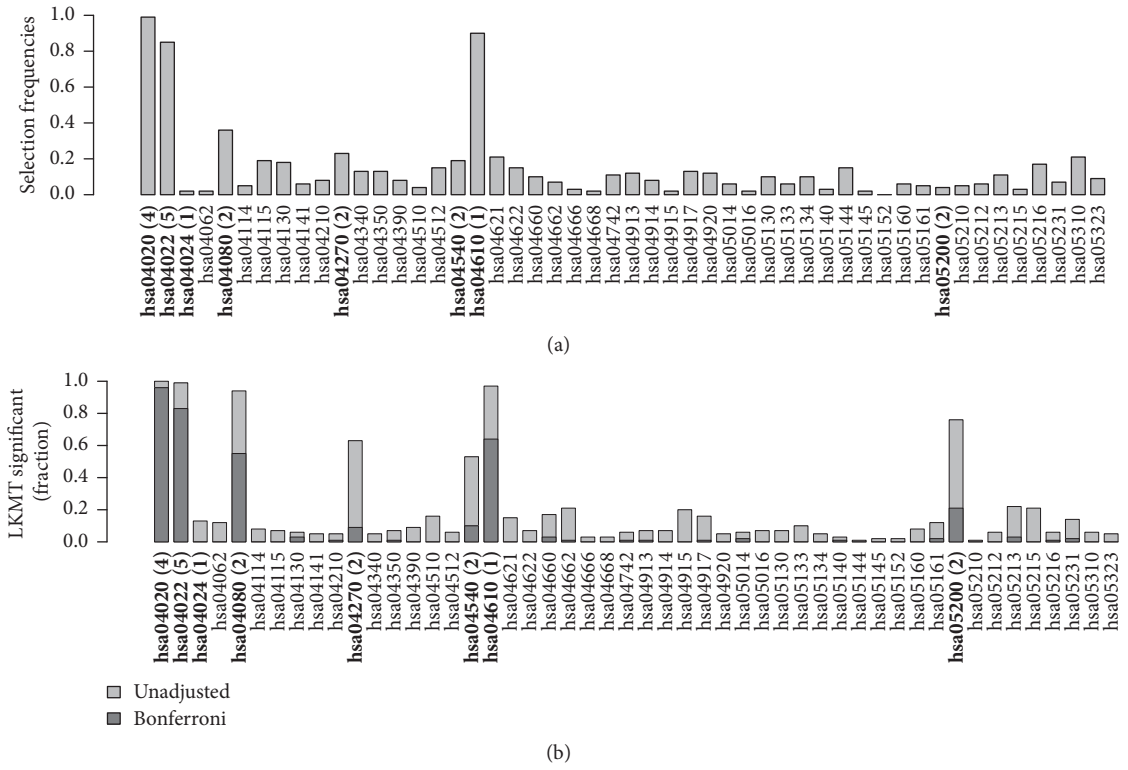

Figure 6: Relative frequency of datasets in which a pathway was selected using (a) kernel boosting (n = 1000, RR = 1.5) and (b) LKMT (n = 1000, RR = 1.5) for sample sizes of 1000 individuals. Effect strength was set to relative risks of 1.5 per allele. Pathways including effect genes are labeled in bold; numbers in brackets denote the count of included influential genes within the pathway.

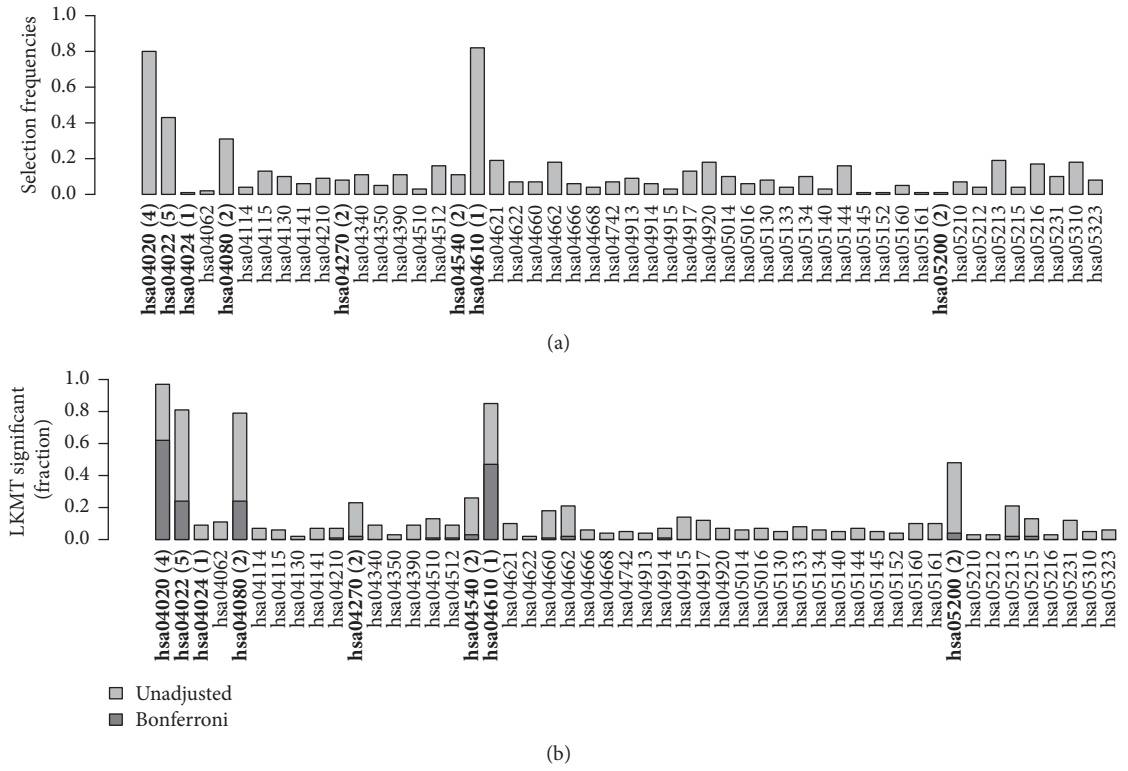

Figure 7: Relative frequency of datasets in which a pathway was selected using (a) kernel boosting (n = 1000, RR = 1.1) and (b) LKMT (n = 1000, RR = 1.1) for sample sizes of 1000 individuals. Effect strength was set to relative risks of 1.1 per allele. Pathways including effect genes are labeled in bold; numbers in brackets denote the count of included influential genes within the pathway.

illustrated in Figure 4(b). Comparing Figures 6 with 7, we can see a drop in selection frequencies as well as in power to detect associated pathways. In Figure 6, the two chosen effect pathways were detected in almost 100% and around 80% of simulation runs for both methods. In Figure 7, we observe that kernel boosting reaches selection frequencies of around 80% and 40%, while the LKMT with Bonferroni correction only achieves selection frequencies slightly greater than 60% and 20%, respectively. In a similar fashion to the results of the scenarios compared in Figures 4 and 5, both methods have higher power to detect associations for stronger effects; however the drop in power is less pronounced for kernel boosting. We conclude that kernel boosting firstly has no inferior performance in terms of power compared to the LKMT. It may even prove more likely to identify influential pathways with smaller genetic effects as it overcomes the multiple testing problem. Secondly, we infer that, in contrast to single-pathway testing approaches, kernel boosting has the ability to discriminate crucial biological processes associated with disease risk from effects included in pathways only due to overlapping genes.

4.2. GWAS Analysis Results. Kernel boosting on the human disease pathways in the lung cancer dataset resulted in selection of only the prion diseases pathway (KEGG id hsa05020). No other pathway was selected. The misclassification error of the tuned boosting model for lung cancer (evaluated at the optimal cut point as defined by the minimal Youden index) was 24.5% and the AUC was 0.785. The ROC curve and the cross-validation results are presented in the Supplementary Material 1, Section B. The LKMT with network-based kernel did not detect any associated pathway on the Bonferroni-corrected significance level. The prion diseases pathway appears ranked 20 out of 73 pathways, when sorting pathways according to ascending Bonferroni-corrected p values. Prions are misfolded proteins capable of changing the structure of other, properly folded proteins into their own incorrect prion structure. They have mostly been reported in connection with neurodegenerative diseases [52]. Nevertheless, a connection with different forms of cancer has also previously been suspected [53, 54]. A full table of results from the analysis of the lung cancer dataset can be found in Supplementary Material 1, Section B.

As expected, analysis of the rheumatoid arthritis dataset discovered a variety of pathways (compare results in [13]). Kernel boosting constructed an explanatory model for disease status based on 32 selected pathways (see pathways written in italics in Table 4). It is well known that genes belonging to the human leukocyte antigen (HLA) complex are highly correlated with rheumatoid arthritis [55]. The HLA family, located on the short arm of chromosome 6, is a highly polymorphic genetic system mainly responsible for the regulation of the immune system [56]. In the human disease class, 18 pathways contain at least one of the HLA genes. These pathways are marked with an asterisk in Table 4. Between the 18 pathways containing HLA genes and the 32 pathways selected by kernel boosting, there is an overlap of 10 pathways. This may be explained by the multivariate nature of the method, in which only the pathway most clearly representing a particular genetic effect will be selected, conditionally on previously selected effects. Testing the human disease pathways' influence on disease status with the LKMT resulted in a large number of 46 significantly associated pathways out of 73 pathways after Bonferroni correction (see pathways with p values in Table 4). These included almost all HLA pathways (15 out of 18). The more specific identification of influential pathways by kernel boosting provides a more complete basis to the understanding of the crucial biological processes involved in disease susceptibility. The misclassification error of the tuned boosting model for rheumatoid arthritis (evaluated at the optimal cut point as defined by the minimal Youden index) was 22.7% and the AUC was 0.850. The ROC curve and the cross-validation results are presented in Supplementary Material 1, Section B.

5. Discussion

We extend a successful method for single-pathway tests to a multivariate selection approach for simultaneous analysis of several pathways. The resulting kernel boosting method benefits from the advantages of a kernel-based analysis, while at the same time overcomes some of the limitations inherent to testing procedures.

Moreover, our multivariable approach to GWAS data analysis does not provide p values, which only provide limited information on the relevance of a genetic effect. A more meaningful result would be an effect measure for the investigated trait or better still the ability to predict an outcome. Kernel boosting facilitates prediction, based on the selected influential variables, as was elucidated in the application where the overall prediction accuracy of each of the models was reported. Thus, it is also possible to interpret the influence of a specific genetic alteration by comparing the change in the predicted outcomes. A high degree of prediction accuracy for the model is ensured through the convenient evaluation of its performance on subsamples of the investigated dataset. This procedure usually results in good prediction accuracy and a sparse model.

Owing to the built-in shrinkage, our boosting approach is capable of dealing with correlated effects. Hence, correlated pathways, which partly include the same genes, can be handled within this framework. Thanks to the multivariable nature of the approach, only the best-fitting pathways, evaluated in terms of prediction accuracy, will be chosen to enter the model. Thus, only the pathway most clearly representing a particular genetic effect will be selected, depending on those pathways selected previously. Our observations support the statement by de Leeuw et al. [57] that competitive gene-set analysis methods (multivariate approach, pathways in competition), in contrast to self-contained approaches (univariate approach, one pathway at a time), can potentially differentiate widely spread heritability of polygenetic outcomes from causal biological processes. This property can be very helpful in the identification and understanding of specific biological functions involved in disease susceptibility.

We consider pathways as analysis units; however various other options exist. Single SNPs in transcribed or untranscribed regions, and SNP sets aggregated to represent a

specific genomic region, environmental variables, or other variables, may be investigated and even combined arbitrarily within one model. For example, the application of our method to the genes comprising a pathway may help to identify key influential genes within the network (for gene boosting, see also the work of Ma et al. [58]; for good overviews of feature selection methods and machine learning tools in bioinformatics refer to [59, 60]). Known influential factors may be embedded in an initial model prior to the selection procedure to adjust for environmental or genetic effects. Furthermore, the considered effects can be incorporated into the model via a multitude of possible base-learners.

The choice of a base-learner can influence effect selections. We observed this behaviour during the simulations, in which the highly connected pathway containing only one effect gene was identified owing to the network-based kernel's high power on interconnected effects. Thus, the well-considered selection of base-learners to be utilized is advisable. We account for the high complexity of possible gene interactions in pathways via the use of a kernel function, which accounts for additive and interaction effects. Such a kernel function will likely lead to a higher degree of prediction accuracy than a simple linear kernel. The application of our method to GWAS datasets on rheumatoid arthritis and lung cancer returned biologically plausible results. Particularly with the rheumatoid arthritis dataset, the number of identified pathways could be reduced considerably compared to single-pathway tests. While the LKMT resulted in 46 significantly associated pathways, kernel boosting narrowed the selection down to 32 pathways. Genes within the HLA region are known to have a strong influence on rheumatoid arthritis. Their effects can reach far across pathways, such that the LKMT detects many pathways including HLA genes as significantly associated. Boosting seems to help to pinpoint down signals even among those pathways and reduces the number of identified pathways to a more reasonable level.

Our results indicate that kernel boosting outperforms single kernel machine tests, as exemplified by the LKMT, in certain genetic scenarios. It may help to discriminate causal biological processes from isolated effects included in pathways only due to gene overlap and facilitate discovering weak signals, especially in studies of limited size. This is of particular interest in the investigation of rare diseases and disease subtypes, in which established methods often fail to find any significantly associated pathways owing to a lack of power.

Datasets of the size investigated here can be analyzed with kernel boosting quite efficiently on current high-performance cluster computing (HPCC) systems. However, such analysis of very large datasets places a rather high demand even on the most powerful HPCC systems to date. Usually, our kernel base-learners are based on the pairwise similarities of all observations. This leads to $n \times n$ similarity matrices as design matrices and hence to parameter vectors $\boldsymbol{\gamma}$ of size n. Instead of using all pairwise similarities, it is possible to compute the similarities only to a representative subset of the observations, or so-called knots. These knots can be chosen as subset of the observations which covers the complete observation space (space-filling algorithm; see [33, 61, 62]). Consequently, we obtain reduced-rank design matrices of dimension $n \times \tilde{n}$, where $\tilde{n}$ is the number of knots, and a parameter vector of size $\tilde{n}$. This reduces the computational burden for the construction of the kernel base-learners and effect estimation and makes kernel-based methods even feasible in situations with many observations. The exact number of observations that can be processed depends, among others, on the considered number of individuals, SNPs, base-learners chosen, and the available hardware.

Kernel boosting constitutes a new and potentially powerful tool in the analysis of GWAS data. It offers a highly flexible and extensible framework, suitable for a wide range of application scenarios. We account for the high complexity of possible gene interactions via the use of kernel functions, while reducing the complexity of the resulting model with the built-in shrinkage of the boosting approach. The resulting model enables us to predict traits and returns more meaningful results than a testing procedure. We conclude that kernel boosting is a suitable methodological addition for the analysis of GWAS, which supports the detection and interpretation of genetic risk factors influencing disease susceptibility.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This work was supported by the German Research Foundation, Research Training Group 1644 "Scaling Problems in Statistics." The rheumatoid arthritis data considered in this article were provided by the National Institutes of Health (Grant AR44422). The analyzed lung cancer study was made available through TRICL Grant no. U19CA148127. The authors would like to thank Andrew Entwistle for his critical review of the manuscript.

References

[1] J. Craig, "Complex diseases: Research and applications," *Nature Education*, vol. 1, article 184, no. 1, 2008.

[2] T. A. Manolio, F. S. Collins, N. J. Cox et al., "Finding the missing heritability of complex diseases," *Nature*, vol. 461, no. 7265, pp. 747–753, 2009.

[3] G. Fehringer, G. Liu, L. Briollais et al., "Comparison of pathway analysis approaches using lung cancer GWAS data sets," *PLoS ONE*, vol. 7, no. 2, Article ID e31816, 2012.

[4] R. M. Cantor, K. Lange, and J. S. Sinsheimer, "Prioritizing gwas results: a review of statistical methods and recommendations for their application," *The American Journal of Human Genetics*, vol. 86, no. 1, pp. 6–22, 2010.

[5] E. P. Hong and J. W. Park, "Sample size and statistical power calculation in genetic association studies," *Genomics & Informatics*, vol. 10, no. 2, pp. 117–122, 2012.

[6] P. Khatri, M. Sirota, and A. Butte, "Ten years of pathway analysis: current approaches and outstanding challenges," *PLoS Computational Biology*, vol. 8, no. 2, Article ID e1002375, 2012.

[7] M. García-Campos, J. Espinal-Enríquez, and E. Hernández-Lemus, "Pathway analysis: state of the art," *Frontiers in Physiology*, vol. 6, 2015.

[8] W. Pan, "Network-based model weighting to detect multiple loci influencing complex diseases," *Human Genetics*, vol. 124, no. 3, pp. 225–234, 2008.

[9] M. C. Wu, P. Kraft, M. P. Epstein et al., "Powerful snp-set analysis for case-control genome-wide association studies," *The American Journal of Human Genetics*, vol. 86, no. 6, pp. 929–942, 2010.

[10] D. Liu, D. Ghosh, and X. Lin, "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*, vol. 9, article no. 292, 2008.

[11] S. Basu, W. Pan, and W. S. Oetting, "A dimension reduction approach for modeling multi-locus interaction in case-control studies," *Human Heredity*, vol. 71, no. 4, pp. 234–245, 2011.

[12] S. Freytag, H. Bickeböller, C. I. Amos, T. Kneib, and M. Schlather, "A novel kernel for correcting size bias in the logistic kernel machine test with an application to rheumatoid arthritis," *Human Heredity*, vol. 74, no. 2, pp. 97–108, 2013.

[13] S. Freytag, J. Manitz, M. Schlather et al., "A network-based kernel machine test for the identification of risk pathways in genome-wide association studies," *Human Heredity*, vol. 76, no. 2, pp. 64–75, 2014.

[14] A. Mayr, H. Binder, O. Gefeller, and M. Schmid, "The evolution of boosting algorithms: From machine learning to statistical modelling," *Methods of Information in Medicine*, vol. 53, no. 6, pp. 419–427, 2014.

[15] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," *The Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, 2001.

[16] J. Friedman, T. Hastie, and R. Tibshirani, "Additive logistic regression: a statistical view of boosting," *The Annals of Statistics*, vol. 28, no. 2, pp. 337–407, 2000.

[17] P. Bühlmann and B. Yu, "Boosting with the L 2 Loss: Regression and Classification," *Journal of the American Statistical Association*, vol. 98, no. 462, pp. 324–339, 2003.

[18] T. Kneib, T. Hothorn, and G. Tutz, "Variable selection and model choice in geoadditive regression models," *Biometrics. Journal of the International Biometric Society*, vol. 65, no. 2, pp. 626–634, 2009.

[19] T. Hothorn, P. Bühlmann, T. Kneib, M. Schmid, and B. Hofner, "Model-based boosting 2.0," *Journal of Machine Learning Research (JMLR)*, vol. 11, pp. 2109–2113, 2010.

[20] B. Hofner, A. Mayr, N. Robinzonov, and M. Schmid, "Model-based boosting in R: a hands-on tutorial using the R package **mboost**," *Computational Statistics*, vol. 29, no. 1-2, pp. 3–35, 2014.

[21] B. Hofner, T. Hothorn, T. Kneib, and M. Schmid, "A framework for unbiased model selection based on boosting," *Journal of Computational and Graphical Statistics*, vol. 20, no. 4, pp. 956–971, 2011.

[22] J. Ferlay, I. Soerjomataram, R. Dikshit et al., "Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012," *International Journal of Cancer*, 2014.

[23] P. Brennan, P. Hainaut, and P. Boffetta, "Genetics of lung-cancer susceptibility," *The Lancet Oncology*, vol. 12, no. 4, pp. 399–408, 2011.

[24] G. S. Firestein, "Evolving concepts of rheumatoid arthritis," *Nature*, vol. 423, no. 6937, pp. 356–361, 2003.

[25] S. Raychaudhuri, "Recent advances in the genetics of rheumatoid arthritis," *Current Opinion in Rheumatology*, vol. 22, no. 2, pp. 109–118, 2010.

[26] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2011, https://www.R-project.org/.

[27] J. Manitz, S. Friedrichs, P. Burger et al., "kangar00: Kernel Approaches for Nonlinear Genetic Association Regression," R package version 1.0, 2017, https://CRAN.R-project.org/package=kangar00.

[28] T. Hothorn, P. Buehlmann, T. Kneib, M. Schmid, and B. Hofner, "mboost: Model-Based Boosting," R package version 2.8-0, 2017, http://CRAN.R-project.org/package=mboost.

[29] A. Mayr, B. Hofner, and M. Schmid, "The importance of knowing when to stop—a sequential stopping rule for component-wise gradient boosting," *Methods of Information in Medicine*, vol. 51, no. 2, pp. 178–186, 2012.

[30] M. Schmid and T. Hothorn, "Boosting additive models using component-wise P-splines," *Computational Statistics and Data Analysis*, vol. 53, no. 2, pp. 298–311, 2008.

[31] N. J. Higham, "Computing the nearest correlation matrix—a problem from finance," *IMA Journal of Numerical Analysis*, vol. 22, no. 3, pp. 329–343, 2002.

[32] E. E. Kammann and M. P. Wand, "Geoadditive models," *Journal of the Royal Statistical Society. Series C. Applied Statistics*, vol. 52, no. 1, pp. 1–18, 2003.

[33] B. Hofner, *Boosting in Structured Additive Models*, LMU München, 2011, http://nbn-resolving.de/urn:nbn:de:bvb:19-138053.

[34] A. L. Boulesteix and T. Hothorn, "Testing the additional predictive value of high-dimensional molecular data," *BMC Bioinformatics*, vol. 11, article 78, 2010.

[35] R. De Bin, "Boosting in Cox regression: a comparison between the likelihood-based and the model-based approaches with focus on the R-packages CoxBoost and mboost," *Computational Statistics*, vol. 31, no. 2, pp. 513–531, 2016.

[36] M. Kanehisa, S. Goto, Y. Sato, M. Kawashima, M. Furumichi, and M. Tanabe, "Data, information, knowledge and principle: back to metabolism in KEGG," *Nucleic Acids Research*, vol. 42, no. 1, pp. D199–D205, 2014.

[37] International HapMap Consortium, "The international hapmap project," *Nature*, vol. 426, pp. 789–796, 2003.

[38] D. E. Reich, M. Cargili, S. Boik et al., "Linkage disequilibrium in the human genome," *Nature*, vol. 411, no. 6834, pp. 199–204, 2001.

[39] Z. Su, J. Marchini, and P. Donnelly, "HAPGEN2: simulation of multiple disease SNPs," *Bioinformatics*, vol. 27, no. 16, Article ID btr341, pp. 2304-2305, 2011.

[40] Y. Lee, H. Li, J. Li et al., "Network models of genome-wide association studies uncover the topological centrality of protein interactions in complex diseases," *Nature*, vol. 20, no. 4, pp. 619–629, 2013.

[41] W. Sauter, A. Rosenberger, L. Beckmann et al., "Matrix metalloproteinase 1 (MMP1) is associated with early-onset lung cancer," *Cancer Epidemiology Biomarkers and Prevention*, vol. 17, no. 5, pp. 1127–1135, 2008.

[42] A. Rosenberger, T. Illig, K. Korb et al., "Do genetic factors protect for early onset lung cancer? A case control study before the age of 50 years," *BMC Cancer*, vol. 8, no. 1, article 60, 2008.

[43] H. Dally, K. Gassner, B. Jäger et al., "Myeloperoxidase (MPO) genotype and lung cancer histologic types: The MPO-463 a

allele is associated with reduced risk for small cell lung cancer in smokers," *International Journal of Cancer*, vol. 102, no. 5, pp. 530–535, 2002.

[44] H.-E. Wichmann, C. Gieger, and T. Illig, "KORA-gen—resource for population genetics, controls and a broad spectrum of disease phenotypes," *Gesundheitswesen*, vol. 67, Supplement 1, pp. S26–S30, 2005.

[45] C. I. Amos, W. Chen, M. F. Seldin et al., "Data for genetic analysis workshop 16 problem 1, association analysis of rheumatoid arthritis data," *BMC Proceedings*, vol. 3, Supplement 7, 2009.

[46] R. M. Plenge, M. Seielstad, L. Padyukov et al., "TRAF1-C5 as a risk locus for rheumatoid arthritis—a genomewide study," *The New England Journal of Medicine*, vol. 357, no. 12, pp. 1199–1209, 2007.

[47] S. R. Browning and B. L. Browning, "Rapid and accurate haplotype phasing and missing data inference for whole genome association studies by use of localized haplotype clustering," *The American Journal of Human Genetics*, vol. 81, no. 5, pp. 1084–1097, 2007.

[48] A. Yates, W. Akanni, M. R. Amode et al., "Ensembl 2016," *Nucleic Acids Research*, vol. 44, no. D1, pp. D710–D716, 2016.

[49] S. Durinck, P. T. Spellman, E. Birney, and W. Huber, "Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt," *Nature Protocols*, vol. 4, no. 8, pp. 1184–1191, 2009.

[50] S. Durinck, Y. Moreau, A. Kasprzyk et al., "BioMart and bioconductor: a powerful link between biological databases and microarray data analysis," *Bioinformatics*, vol. 21, no. 16, pp. 3439-3440, 2005.

[51] D. Malzahn, S. Friedrichs, and H. Bickeböller, "Comparing strategies for combined testing of rare and common variants in whole sequence and genome-wide genotype data," *BMC Proceedings*, vol. 10, Supplement 7, pp. 269–273, 2016.

[52] P. L. A. Leighton and W. Ted Allison, "Protein misfolding in prion and prion-like diseases: Reconsidering a required role for protein loss-of-function," *Journal of Alzheimer's Disease*, vol. 54, no. 1, pp. 3–29, 2016.

[53] H. Antony, A. P. Wiegmans, M. Q. Wei, Y. O. Chernoff, K. K. Khanna, and A. L. Munn, "Potential roles for prions and protein-only inheritance in cancer.," *Cancer metastasis reviews*, vol. 31, no. 1-2, pp. 1–19, 2012.

[54] J. L. Silva, L. P. Rangel, D. C. F. Costa, Y. Cordeiro, and C. V. De Moura Gallo, "Expanding the prion concept to cancer biology: dominant-negative effect of aggregates of mutant p53 tumour suppressor," *Bioscience Reports*, vol. 33, no. 4, pp. 593–603, 2013.

[55] C. M. Weyand and J. J. Goronzy, "Association of MHC and rheumatoid arthritis. HLA polymorphisms in phenotypic variants of rheumatoid arthritis," *Arthritis Research*, vol. 2, no. 3, pp. 212–216, 2000.

[56] S. Y. Choo, "The HLA system: genetics, immunology, clinical testing, and clinical implications," *Yonsei Medical Journal*, vol. 48, no. 1, pp. 11–23, 2007.

[57] C. A. de Leeuw, B. M. Neale, T. Heskes, and D. Posthuma, "The statistical properties of gene-set analysis," *Nature Reviews Genetics*, vol. 17, no. 6, pp. 353–364, 2016.

[58] S. Ma, Y. Huang, J. Huang, and K. Fang, "Gene network-based cancer prognosis analysis with sparse boosting," *Genetics Research*, vol. 94, no. 4, pp. 205–221, 2012.

[59] Y. Saeys, I. Inza, and P. Larrañaga, "A review of feature selection techniques in bioinformatics," *Bioinformatics*, vol. 23, no. 19, pp. 2507–2517, 2007.

[60] P. Larrañaga, B. Calvo, R. Santana et al., "Machine learning in bioinformatics," *Briefings in Bioinformatics*, vol. 7, no. 1, pp. 86–112, 2006.

[61] M. E. Johnson, L. M. Moore, and D. Ylvisaker, "Minimax and maximin distance designs," *Journal of Statistical Planning and Inference*, vol. 26, no. 2, pp. 131–148, 1990.

[62] D. Nychka and N. Saltzman, "Design of air-quality monitoring networks," in *Case Studies in Environmental Statistics*, D. Nychka, W. W. Piegorsch, and L. H. Cox, Eds., vol. 132, pp. 51–76, Springer US, New York, NY, USA, 1998.

Supplemental Material 1

for

Pathway-Based Kernel Boosting for the Analysis of Genome-Wide Association Studies

Stefanie Friedrichs[1,*], Juliane Manitz[2,3], Patricia Burger[1], Christopher I. Amos[4], Angela Risch[5,6,7], Jenny Chang-Claude[8], H.-Erich Wichmann[9,10,11], Thomas Kneib[2], Heike Bickeböller[1], and Benjamin Hofner[12, 13]

[1]Institute of Genetic Epidemiology, University Medical Centre, Georg-August University Göttingen, Göttingen, Germany.
[2]Department of Statistics and Econometrics, Georg-August University Göttingen, Göttingen, Germany
[3]Department of Mathematics and Statistics Boston University, Bosten, USA
[4]Department of Community and Family Medicine, Geisel School of Medicine, Dartmouth College, Lebanon, NH, United States of America
[5]Division of Molecular Biology, University of Salzburg, Austria
[6]Translational Lung Research Center Heidelberg (TLRC-H), Member of the German Center for Lung Research (DZL), Heidelberg, Germany
[7]Division of Epigenomics and Cancer Risk Factors, DKFZ German Cancer Research Center, Heidelberg, Germany
[8]Division of Cancer Epidemiology, German Cancer Research Center (DKFZ), Heidelberg, Germany
[9]Institute of Medical Informatics, Biometry and Epidemiology, Chair of Epidemiology, Ludwig-Maximilians University, Munich, Germany
[10]Helmholtz Center Munich, Institute of Epidemiology 2, Germany
[11]Institute of Medical Statistics and Epidemiology, Technical University Munich, Germany
[12]Department of Medical Informatics, Biometry and Epidemiology, Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany
[13]Section Biostatistics, Paul-Ehrlich-Institut, Langen, Germany

A Additional Analysis of Simulation Study

A.1 Choice and distribution of m_{stop}

In the primary analysis of the simulation study, we tried to convey a clear picture of the selection properties of the boosting algorithm, which can be easily related to the selection of pathways based on LKMT tests. As such we chose a relatively small number of boosting iterations to check if the influential pathways are selected early on and if they can be clearly distinguished from non-influential pathways. Hence, in the analysis of simulation results reported in the manuscript, the ideal number of iterations m_{stop} was determined within a search range of 0 to 200. Specifying a (relatively small) maximum number of possible iterations might force an early stopping of the algorithm in some simulation runs.

To investigate this issue, we re-analysed all simulation scenarios with a larger number of maximal iterations permitted, in order to allow the algorithm to reach the optimal boosting iteration, i.e., to find an iteration m_{stop} such that the out-of-bag risk is minimal. The number of iterations needed usually depends on the strength of the signal (effect size), the number of informative base-learners and the number of observations. In our simulation study, the number of iterations was mainly influenced by the number of observations (but also, though to a lesser extend) by the effect size. For simulation scenarios up to 1000 individuals, we considered a maximum of 500 iterations, while for samples of 2000 individuals, the algorithm was allowed to perform up to 1000 iterations.

In Figure 1 we display the observed number of iterations required for each simulation scenario to reach the optimal prediction accuracy as measured by the cross-validated out-of-bag Binomial log-likelihood.

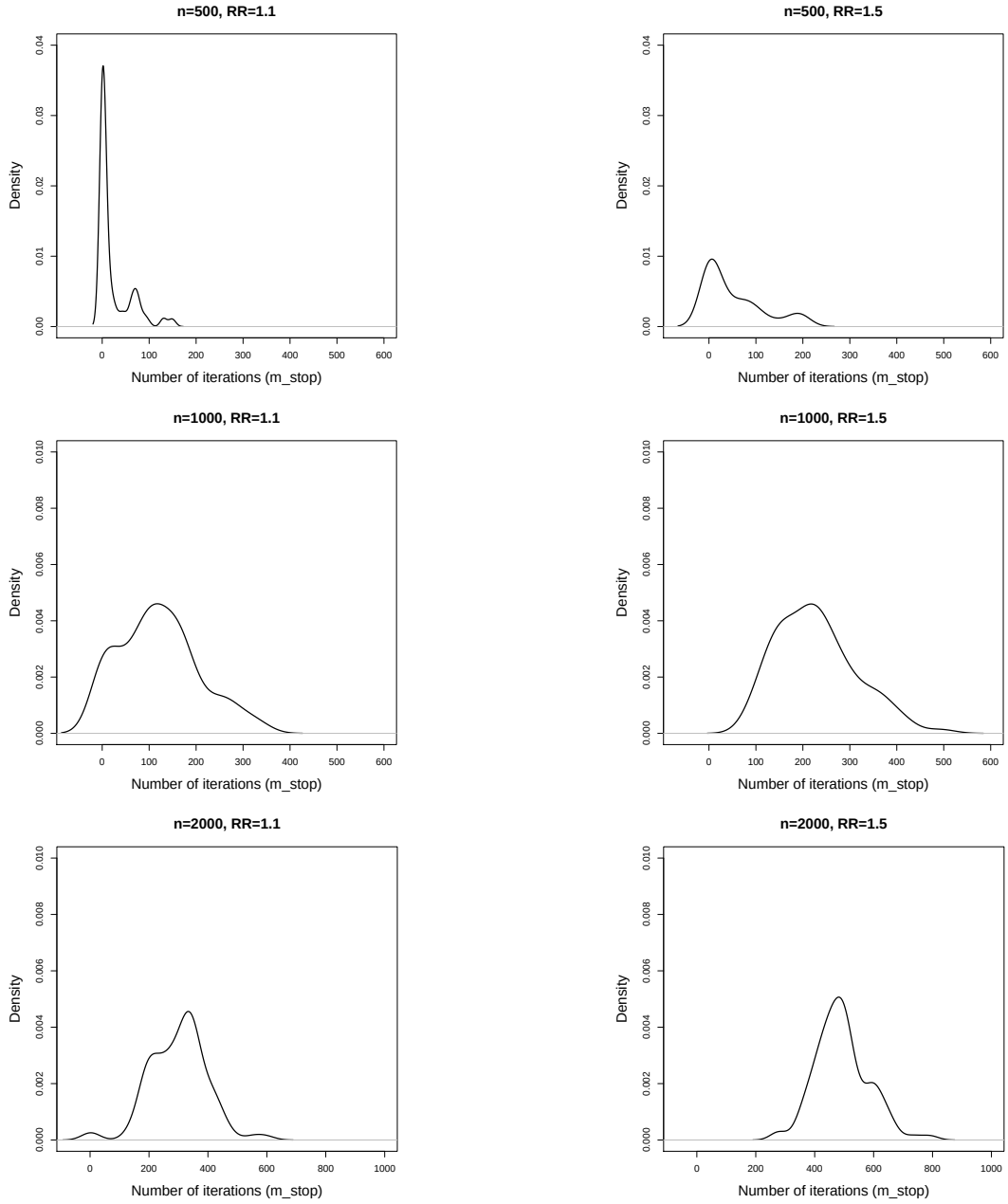

Figure 1: Kernel density estimates of the number of iterations (m_{stop}) in the 100 simulation runs for the different simulation scenarios.

A.2 Selection of Pathways

Increasing of the number of iterations, as discussed in the previous section, leads to an increase in runtime and likely results in the selections of additional pathways. Even though boosting tends to have a slow overfitting behavior [1, 2], at a certain point, non-influential effects are selected as well. This is more pronounced for data sets with many observations compared to the number of base-learners (i.e., "$n > p$"). Especially in later boosting iterations, it might happen that non-informative pathways are selected. However, these

pathways are usually selected infrequently and with a small effect on the predicted outcome. Pathways selected early and often will have much more influence on the prediction.

The additional selections of causal and also non-causal pathways results in a less clear discrimination of influential biological processes. This disadvantage can be compensated for, however, by evaluating the results of kernel boosting in more detail. As the boosting algorithm can not only select a pathway once, but will usually select the same effect variable multiple times, if it is highly influential on the outcome, we can interpret the selection frequency of each pathway for a single simulated data set. This is one means to take the clinical relevance into account. Alternatively, one could consider the effect size, i.e., the size of the coefficient for linear base-learners or the norm of the coefficient vector for pathway kernel base-learners.

In the following paragraph, we assess the selection properties of the boosting algorithm when run until convergence. The upper panels of Figures 2 to 7 depict the relative selection frequencies of each base-learner averaged over all 100 simulation runs per scenario. Here, we firstly count how often each pathway has been selected in a single simulation run. This number is then transformed into a proportion of selections by deviding it with the chosen m_{stop} in the corresponding run. Secondly, these proportions per pathway are averaged across all 100 simulation runs. In this way, we are taking into account the relative importance of that effect. For comparisons the lower panel in each of the figures shows the relative frequency of simulation runs in which a base-learner was selected at least once. The latter plots are equal in structure to those in the paper, they merely show results for larger values of m_{stop}.

We can see, that for the simulation scenarios of 500 and 1000 individuals, no remarkable change was detected when increasing the maximum number of iterations. Especially in the simulation scenarios with 500 individuals, hardly any difference between top and lower barplots is visible (Figures 2 and 3). In simulation scenarios of 1000 individuals, depicted in Figures 4 and 5, we can see that the influential biological processes, represent by the two simulated effect pathways, are more precisely distinguished from non-causal pathways when also taking into account relative selection frequencies. For the scenario with 2000 individuals (Figure 7) we can see that considering relative selection frequencies has more impact in larger samples. Here a clear difference between the upper and lower barplot is visible. When only considering if a pathway was ever selected (lower row), influential and non-influential pathways can less clearly be discriminated. Additional evaluation of the relative selection frequency (top row) gives a much clearer picture and facilitates identification of the causal pathways. Note, that the top barplot for the scenario with 2000 individuals and a relative risk of 1.5 per allele (Figure 7) looks similar to Figure 4 in the Paper, which evaluated selections only on the same data for a smaller number of iterations. This means, that we can identify the influential pathways in a dataset with a noticeably reduction in computation time using early stopping.

We conclude that he discrimination of biologically relevant processes from gene overlaps is possible by letting the algorithm run until the optimal m_{stop} when taking not only into account if a pathway was selected, but also considering the relative selection frequencies. Using this approach, causal pathways were even more precisely distinguished from non-causal pathways than in the case of evaluating only if a pathway was seletced at least once or not.

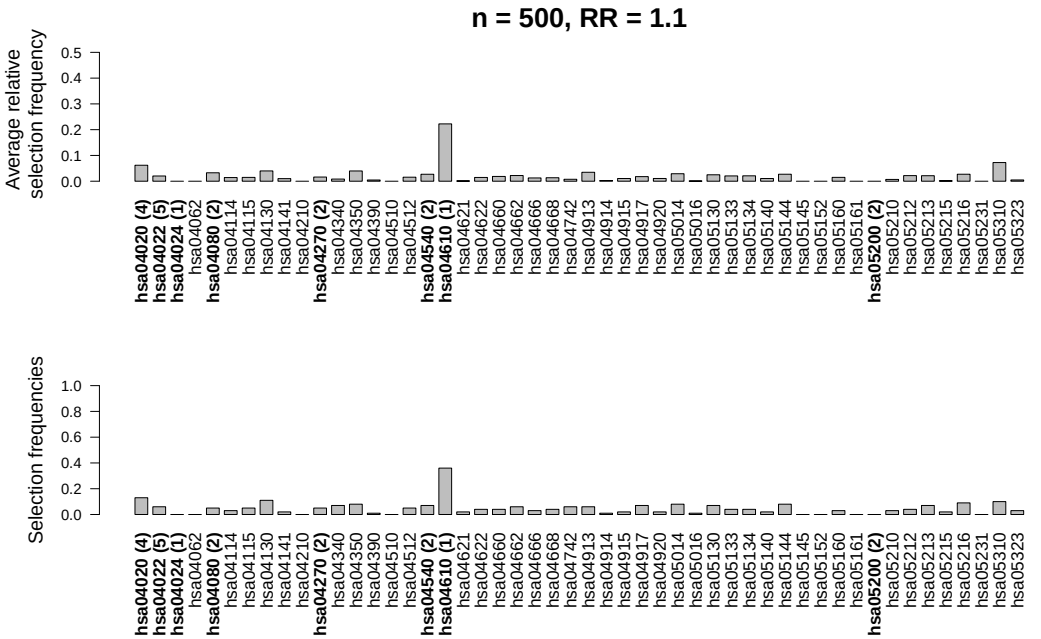

Figure 2: Barplots for the relative selection frequencies of each base-learner in a single run averaged over all 100 simulation runs (top) and relative frequencies of simulation runs in which a base-learner was selected at least once (bottom). The sample comprised 250 cases and 250 controls and the effect strength was set to relative risks of 1.1 per allele. Pathways including effect genes are labeled in bold; numbers in brackets denote the count of included influential genes within the pathway.

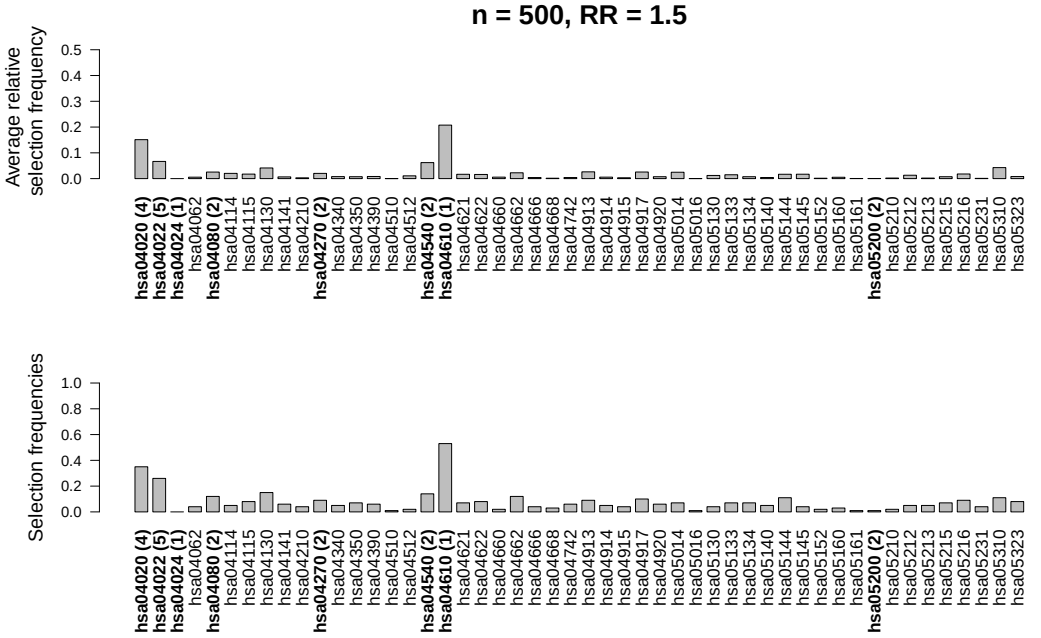

Figure 3: Barplots for the relative selection frequencies of each base-learner in a single run averaged over all 100 simulation runs (top) and relative frequencies of simulation runs in which a base-learner was selected at least once (bottom). The sample comprised 250 cases and 250 controls and the effect strength was set to relative risks of 1.5 per allele. Pathways including effect genes are labeled in bold; numbers in brackets denote the count of included influential genes within the pathway.

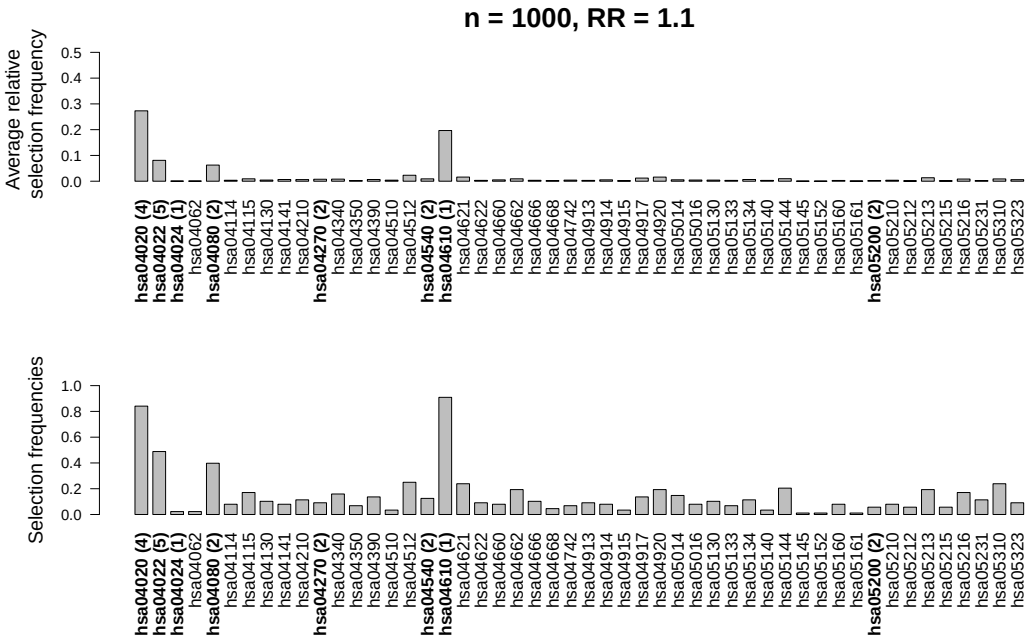

Figure 4: Barplots for the relative selection frequencies of each base-learner in a single run averaged over all 100 simulation runs (top) and relative frequencies of simulation runs in which a base-learner was selected at least once (bottom). The sample comprised 500 cases and 500 controls and the effect strength was set to relative risks of 1.1 per allele. Pathways including effect genes are labeled in bold; numbers in brackets denote the count of included influential genes within the pathway.

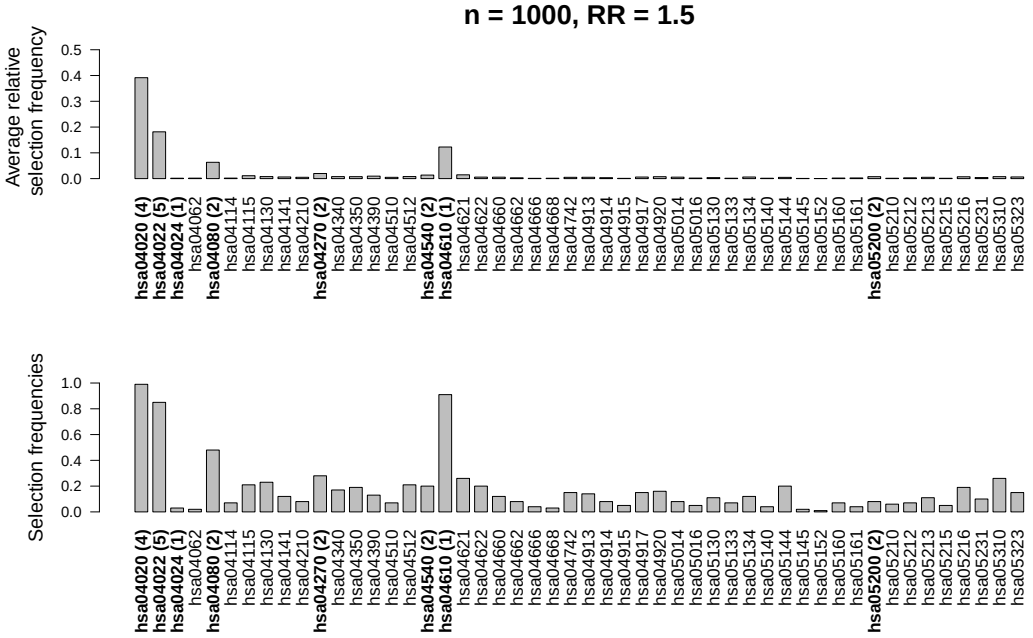

Figure 5: Barplots for the relative selection frequencies of each base-learner in a single run averaged over all 100 simulation runs (top) and relative frequencies of simulation runs in which a base-learner was selected at least once (bottom). The sample comprised 500 cases and 500 controls and the effect strength was set to relative risks of 1.5 per allele. Pathways including effect genes are labeled in bold; numbers in brackets denote the count of included influential genes within the pathway.

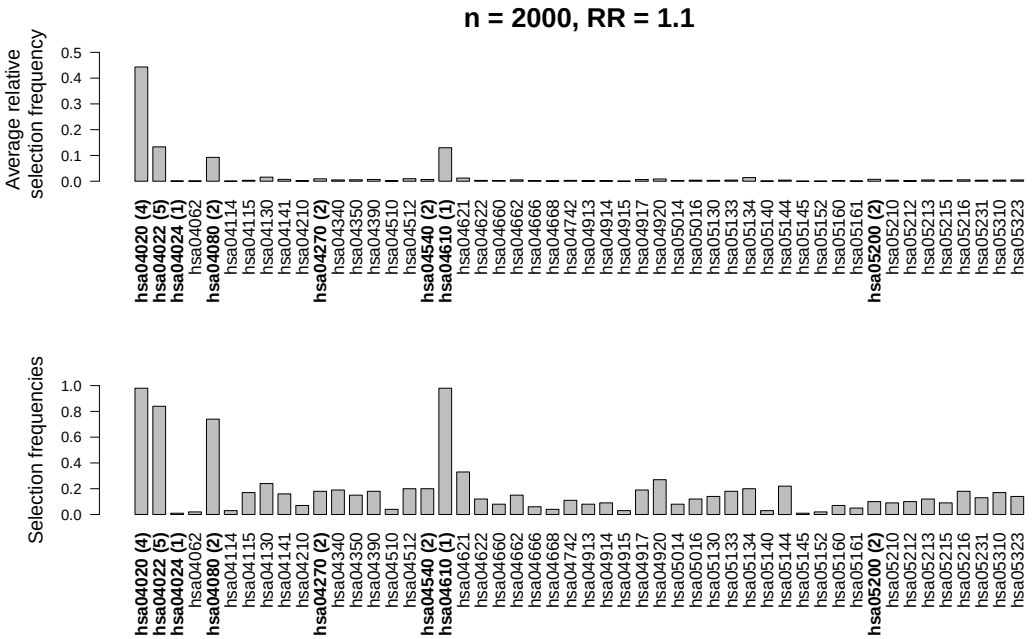

Figure 6: Barplots for the relative selection frequencies of each base-learner in a single run averaged over 100 simulation runs (top) and relative frequencies of simulation runs in which a base-learner was selected at least once (bottom). The sample comprised 1000 cases and 1000 controls and the effect strength was set to relative risks of 1.1 per allele. Pathways including effect genes are labeled in bold; numbers in brackets denote the count of included influential genes within the pathway.

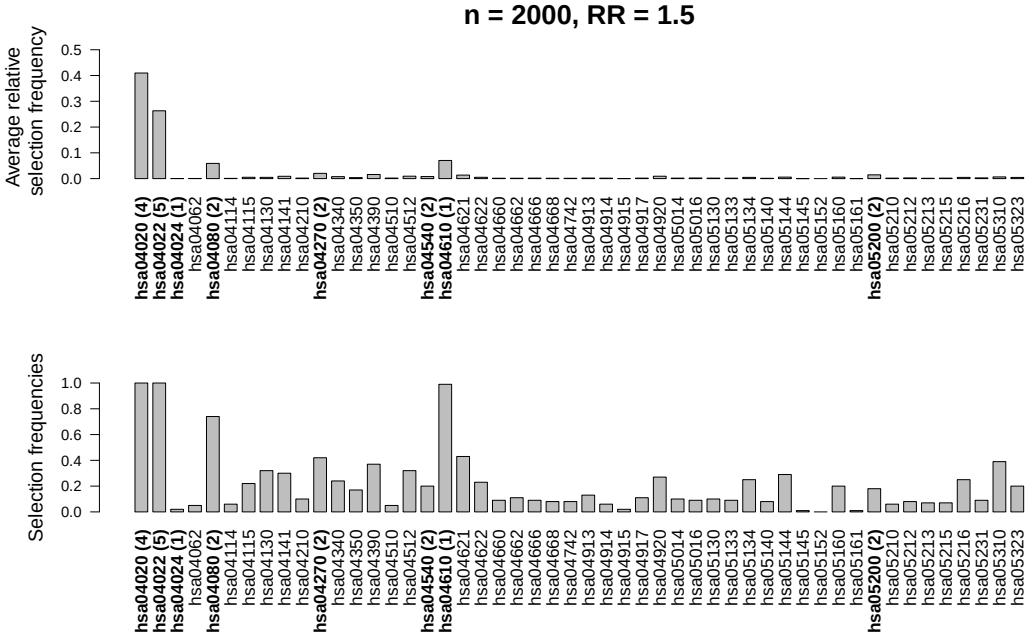

Figure 7: Barplots for the relative selection frequencies of each base-learner in a single run averaged over all 100 simulation runs (top) and relative frequencies of simulation runs in which a base-learner was selected at least once (bottom). The sample comprised 1000 cases and 1000 controls and the effect strength was set to relative risks of 1.5 per allele. Pathways including effect genes are labeled in bold; numbers in brackets denote the count of included influential genes within the pathway.

A.3 Computational Requirements

In the following, we provide run times and memory requirements for exemplary simulation runs. The measurements include the model fitting with 50 simulated pathways and 20-fold cross-validation to determine the optimal m_{stop}. Cross-validation was run in parallel on 20 cores. We report the runtime (time actually needed for the process), the CPU time (sum of run time over all CPUs used; approximates the runtime if the process was run sequentially) and maximum memory allocation:

- Kernel boosting for the simulation scenario with 500 individuals required a runtime of 12.8 minutes (corresponding CPU time 3.5 hours) as well as a maximum memory use of 11.6 GB to determine the optimal m_{stop} between 0 and 500.
- Analysis of the simulation scenario including 1000 individuals resulted in a runtime of 1.9 hours, equalling a CPU time of 24.9 hours, for the same search range of m_{stop}. The maximum memory use was approximately 40 GB.
- The simulation scenario with 2000 individuals needed a runtime of 23.3 hours (CPU time 340.6 hours), and utilized a maximum memory of 132 GB. Here, the ideal number of iterations was to be determined between 0 and 1000.

Note, that the actual runtime can vary (e.g. depending on the system, the CPU and the memory available). In practice, the runtime is significantly smaller than the CPU time, as can be seen above, as it is very easy to run the cross-validation in parallel. Of course, parallelization also requires a higher amount of memory. Hence, running the cross-validation sequentially will require less memory, but will take longer.

A.4 Details on Effect Pathways

A graphical display of the two networks that were simulated to contain effect genes is given in Figures 9 and 8.

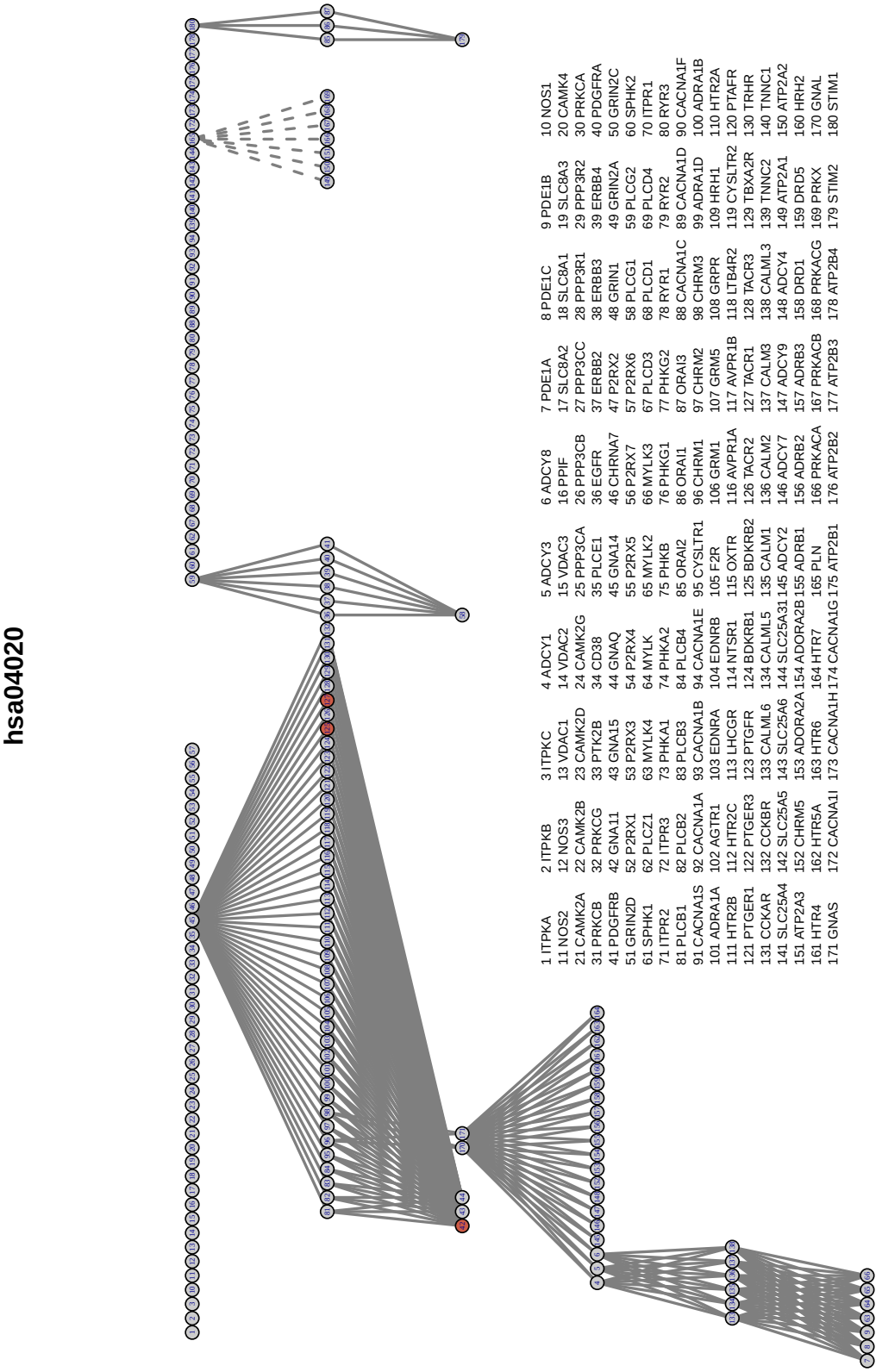

Figure 8: Network structure and placement of effect genes (red nodes) in the pathway hsa04020 used in simulations.

Figure 9: Network structure and placement of effect genes (red nodes) in the pathway hsa04022 used in simulations.

B Additional Results of Data Analyses

Figure 10 shows the out-of-bag risk for the 20-fold subsampling: The model is fitted 20 times on random subsets of the data and the (negative) Binomial likelihood is computed for the derived model on the new data (for each value of m_{stop}). Each of the gray lines is the out-of-bag risk for one model. The black line is the averaged risk for all 20 models. This estimates the goodness of fit, as measured by the likelihood, or better said the risk as measured by the negative likelihood. Essentially, we see how well the model would perform to predict the outcome for new data. The vertical dotted line indicates the optimal m_{stop} chosen on the dataset. The cross-validated risk for the lung cancer data shows that this data set seems to contain very little information as the risk almost imediately starts to increase. The optimal boosting iteration was chosen as $m_{\text{stop}} = 4$. The cross-validated risk for the rheumatoid athritis data shows that many updates were required to find the optimal model ($m_{\text{stop}} = 993$). It seems that this GWAS data set contains much more information on the disease status. The Receiver operating characteristic (ROC) curves of the two model for lung cancer and rheumatoid arthritis are depcited in Figure 11. These graphs display the overall prediction accuracy of the derived models.

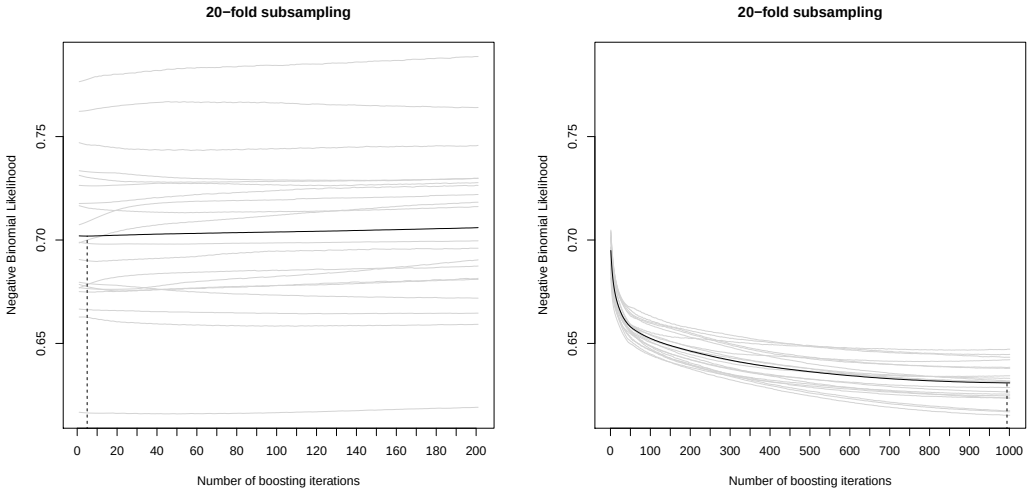

Figure 10: Cross-validated out-of-bag prediction accuracy for the lung cancer (left) and rheumatoid arthritis dataset (right).

Table 1 gives an overview the pathways used for the lung cancer data set together with the p-values derived via LKMT.

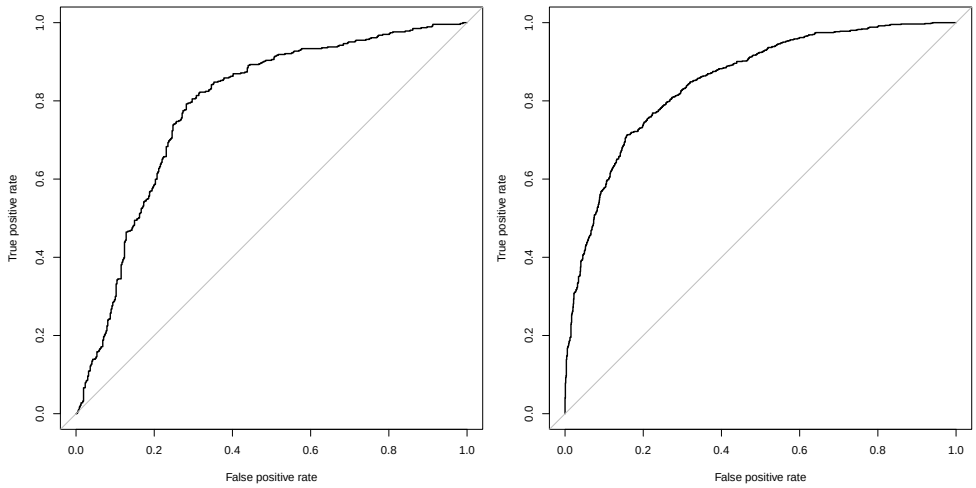

Figure 11: Receiver operating characteristic (ROC) curve depcting the prediction accuraccy of the boosted model for lung cancer (left) and for rheumatoid arthritis (right).

KEGG id	Name of Pathway	P-value
hsa05134	Legionellosis	0.0389
hsa05016	Huntington's disease	0.0446
hsa05323	Rheumatoid arthritis	0.0986
hsa05231	Choline metabolism in cancer	0.1232
hsa05210	Colorectal cancer	0.1421
hsa05169	Epstein-Barr virus infection	0.1464
hsa05220	Chronic myeloid leukemia	0.1698
hsa04940	Type I diabetes mellitus	0.1754
hsa05143	African trypanosomiasis	0.1758
hsa05014	Amyotrophic lateral sclerosis (ALS)	0.1800
hsa05205	Proteoglycans in cancer	0.1933
hsa05223	Non-small cell lung cancer	0.1991
hsa05144	Malaria	0.2080
hsa05211	Renal cell carcinoma	0.2274
hsa05332	Graft-versus-host disease	0.2590
hsa05214	Glioma	0.2653
hsa05212	Pancreatic cancer	0.3032
hsa05010	Alzheimer's disease	0.3177
hsa05031	Amphetamine addiction	0.3185
hsa05020	**Prion diseases**	0.3286
hsa05340	Primary immunodeficiency	0.3478
hsa05166	HTLV-I infection	0.3656
hsa05213	Endometrial cancer	0.4011
hsa04932	Non-alcoholic fatty liver disease (NAFLD)	0.4029
hsa05145	Toxoplasmosis	0.4054
hsa05218	Melanoma	0.4109
hsa05230	Central carbon metabolism in cancer	0.4262
hsa05330	Allograft rejection	0.4288
hsa04933	AGE-RAGE signaling pathway in diabetic complications	0.4297
hsa05206	MicroRNAs in cancer	0.4305
hsa05221	Acute myeloid leukemia	0.4315
hsa05219	Bladder cancer	0.4322
hsa05032	Morphine addiction	0.4411
hsa05133	Pertussis	0.4637
hsa05012	Parkinson's disease	0.4690
hsa05310	Asthma	0.4709
hsa05033	Nicotine addiction	0.4756

hsa05150	Staphylococcus aureus infection	0.4834
hsa05416	Viral myocarditis	0.5194
hsa05120	Epithelial cell signaling in Helicobacter pylori infection	0.5271
hsa05110	Vibrio cholerae infection	0.5287
hsa05161	Hepatitis B	0.5366
hsa05200	Pathways in cancer	0.5648
hsa04931	Insulin resistance	0.5697
hsa05217	Basal cell carcinoma	0.5736
hsa05030	Cocaine addiction	0.5852
hsa05215	Prostate cancer	0.5860
hsa05130	Pathogenic Escherichia coli infection	0.6437
hsa05204	Chemical carcinogenesis	0.6518
hsa05203	Viral carcinogenesis	0.6630
hsa05216	Thyroid cancer	0.6693
hsa05202	Transcriptional misregulation in cancer	0.6722
hsa05168	Herpes simplex infection	0.7000
hsa05131	Shigellosis	0.7154
hsa05100	Bacterial invasion of epithelial cells	0.7165
hsa05132	Salmonella infection	0.7292
hsa05320	Autoimmune thyroid disease	0.7341
hsa05152	Tuberculosis	0.7453
hsa05162	Measles	0.7702
hsa05222	Small-cell lung cancer	0.7793
hsa05140	Leishmaniasis	0.7971
hsa05142	Chagas disease (American trypanosomiasis)	0.8150
hsa05164	Influenza A	0.8419
hsa05322	Systemic lupus erythematosus	0.8594
hsa05146	Amoebiasis	0.8903
hsa05034	Alcoholism	0.8912
hsa04930	Type II diabetes mellitus	0.8960
hsa04950	Maturity onset diabetes of the young	0.9191
hsa05321	Inflammatory bowel disease (IBD)	0.9214
hsa05414	Dilated cardiomyopathy	0.9664
hsa05410	Hypertrophic cardiomyopathy (HCM)	0.9732
hsa05412	Arrhythmogenic right ventricular cardiomyopathy (ARVC)	0.9858
hsa05160	Hepatitis C	0.9863

Table 1: KEGG pathways in the Human Diseases class as downloaded in April 2016. Pathways are sorted according to p-value, derived from LKMT application on the lung cancer dataset, in ascending order. No pathways reached a significant p-value after Bonferroni correction are listed. The pathway selected by kernel boosting on this same dataset is marked in bold.

References

[1] Bühlmann P, Hothorn T. Boosting Algorithms: Regularization, Prediction and Model Fitting. Statistical Science. 2007;22:477–505.

[2] Mayr A, Binder H, Gefeller O, Schmid M. The Evolution of Boosting Algorithms - From Machine Learning to Statistical Modelling. Methods of Information in Medicine. 2014;53(6):419–427.

Kangar00: Kernel Approaches for Nonlinear Genetic Association Regression

Stefanie Friedrichs

2017-04-26

Introduction

The genetic information collected in genome-wide association studies (GWAS) is represented by the genotypes of various single-nucleotid polymorphisms (SNPS). Testing biological meaningful SNP Sets is a successful strategy for the evaluation of GWAS data, as it may increase power as well as interpretation of results. Via mapping of SNPs to genes forming a network, association between pathways and disease risk can be investigated.
Kernel methods are particularly well suited to cope with the challenges connected to the analysis of large SNP sets from GWAS data. They do not require to model a direct functional relationship between SNPs and effects, while at the same time can deal with high-dimensional data and allow for straightforward incorporation of covariates. The model for a logistic kernel machine regression of a pathway on a binary outcome is given by

$$logit(P(y_i = 1|x_i, z_i)) = x_i^t\beta + h(z_i)(1)$$

where y_i denotes the case or control status of individual i, x_i is the vector including informative covariates (such as age, sex, etc.) and z_i represents the genotypes of individual i. β is the regression coefficients for the parametric part of the model, while $h(.)$ denotes an unknown function, non-parametrically incorporating the pathway's influence. The intercept is assumed to be included in x_i. For more details see Liu et al (2008).

Different kernels have been proposed that convert the genomic information of two individuals into a quantitative value reflecting their genetic similarity. This package includes the linear kernel as well as two more advanced kernels, adjusting for size bias in the number of SNPs and genes in a pathway or incorporating the network structure of genes within the pathway, respectively. The kernel functions are described in more detail in the instructions below.

A variance component test, constructed around the similarity matrix, can be used to evaluate a pathway's influence on disease risk. In `kangar00` p-values can be calculated with the Satterthwaite approximation or Davies method as described in Schaid (2010) and Davies (1980), respectively.

Data extraction and preparation

Pathways

The kangar00 package offers several functions for data extraction from internet databases. In the following they will be explained using the Circadian rhythm pathway as an example.

- In the KEGG database (Kanehisa et al 2014) this pathway is identified with the id *hsa04710*.
- The function `pathway_info()` can use this id to create a table listing all genes included in Circadian rhythm. For each gene the startpoint, endpoint and the chromosome are listed.
- Gene membership is obtained directly from KEGG, while startpoints, endpoints and chromosome information is extracted from Ensembl (Cunningham et al 2015). The database is accessed via the

function `getBM()` in the biomaRt package. This means that the gene boundaries given will equal the current build used in Ensemble. An internet connection is required for this step.

```
pathway_info('hsa04710')
```

will return a `pathway_info` object containing a data frame of the form

pathway	gene_start	gene_end	chr	gene
hsa04710	13276652	13387266	11	ARNTL
hsa04710	4979116	4985323	3	BHLHE40
hsa04710	26120026	26125127	12	BHLHE41
hsa04710	...	...	...	...

listing information on all genes KEGG assigned to Circadian rhythm.

Pathway object

In `kangar00` all information on a specific pathway is combined in a `pathway` object. It includes

- The pathway's ID as used in KEGG.
- The adjacency `matrix`, which equals the network matrix without signs.
- A `vector` giving the signs for the interactions.

The following example creates a new pathway object, to which gene-interaction information has yet to be added

```
pathw <- pathway(id='hsa04710', adj=matrix(0), sign=as.vector(matrix(0)[matrix(0)!=0]))
```

Networkmatrix

The gene-gene interactions within pathways are represented by a network matrix. This quadratic matrix is of dimension equal to the number of genes in the corresponding pathway. It includes entries equal to 1 (representing an activation interaction), -1 (denoting an inhibiting interaction) or 0 (no interaction).

A network matrix can be created using the function `get_network_matrix()`. Gene interaction information for a specific pathway is extracted from the KEGG database. It is accessed via the function `retrieveKGML()` from the `KEGGgraph` package. An internet connection is required for this step.

```
pathw_complete <- get_network_matrix(pathw, directed=FALSE)
```

will download the KEGG XML file for the pathway with ID 'hsa04710' and save it in the working directory. The function will convert the data into a network matrix and add it to the given pathway object. The expanded pathway object will be returned. The user can specify whether the gene-interaction matrix should be given directed (directed=TRUE) or undirected (directed=FALSE).

SNP positions

Kangar00 offers a function to download positions of the SNPs available in your GWAS dataset from the Ensembl database.

- `snp_info()` will take a vector of rs-numbers and give the corresponding base pair positions.

- Positions are extracted from the Ensembl database and thus equal the current build used on the website. The database is accessed via the function `getBM()` from the package `biomaRt`. This requires an internet connection.

```
snp_info("rs234")
```

will return a `snp_info` object containing the data frame

chr	position	rsnumber
7	105920689	rs234

Pathway Annotation

To define SNP sets representing a pathway, the function `get_anno()` can be used.

- Input arguments are a `pathway_info` as well as a `snp_info` object.
- If you do not want to change positions in your SNP file using the `snp_info()` function, you will have to transform it into a `snp_info` object including a data frame listing all SNPs to be annotated. This data frame must include the columns 'chr', 'position' and 'rsnumber', giving for each SNP the chromosome it lies on, its base pair position on the chromosome and the rs-numbers identifier, respectively. See also the output description of `snp_info()`.
- For annotation the package `sqldf` is used.

```
get_anno(snp_info, pathway_info)
```

will return a data frame listing all SNPs that lie inside the boundaries of one or more genes in the pathway. That means that genes can appear several times, depending on the number of SNPs mapped to them. A SNP can and will be mapped to multiple genes if they overlap. The data frame will have the following format

pathway	gene	chr	snp	position
hsa04710	CSNK1E	22	rs11089885	38413480
hsa04710	CSNK1E	22	rs13054361	38336819
hsa04710	CSNK1E	22	rs135757	38307648
hsa04710	...	...	...	...

GWAS data

Data from a case control study is needed to test a pathways influence on disease risk with the logistic kernel machine test in `kangar00`. Here, GWAS data is represented by the `GWASdata` object. It includes

- Genotype data for each individual.
  - Genotype data needs to be a matrix with one line per individual and one column for each SNP.
  - Rownames give ID numbers for the individuals while columnames give the rs-numbers corresponding to the SNPs genotyped in the study.
  - Note that missing values are not allowed and SNPs with missing genotypes have to be imputed or excluded from the sample prior to creation of the `GWASdata` object.
- Phenotype data for each individual.
  - Phenotypes need to be given in a `data frame` with the first column including the individual IDs as in the genotype sample.

  - Further columns can contain informative covariates (such as age, sex, ...) to be used in the logistic regression model.
- Annotation of study SNPs to pathways created by `get_anno`.
  - This `data frame` defines the SNP set representing a specific pathway. It can be created using the function `get_anno()`.
- A `character` describing the data can be added to the `GWASdata` object. This could for instance be the name of the study.

A `GWASdata` object can be constructed as

```
my_gwas <- GWASdata(pheno=pheno, geno=geno, anno=anno, desc="study xy")
```

Calculation of Kernel Matrices

Once a `GWASdata` object is created, we can start to calculate kernel matrices to test a pathways influence on disease risk. `Kangar00` offers three different kernel functions to compute a similarity matrix for the individuals in analysis. They will be explained in the following.

Linear Kernel (Lin)

The linear kernel assumes additive SNP effects. It is calculated as

$$ZZ^t(2)$$

where Z denotes the genotype matrix (See also Liu et al, 2010). In `kangar00` a linear kernel can be created using the function `kernel_lin()`. It requires as arguments

- A `GWASdata` object containing the genotype information.
- A `pathway` object specifying the pathway to be tested.
- A value for argument `calculation` to decide how the kernel should be calculated. Options are `cpu` for calculation on cpu and `gpu` for gpu calculation.

```
K_lin <- lin_kernel(gwas, p, calculation='cpu')
```

will return a quadratic `matrix` of dimension equal to the number of individuals in the `GWASdata` object.

Size-adjusted Kernel (Sia)

The size-adjusted kernel takes into consideration the numbers of SNPs and genes in a pathway to correct for size bias. It is calculated as

$$K_{i,j} = exp(-\sqrt{\frac{1}{r_p}\sum_g(\frac{||z_i^g - z_j^g||}{\mu_g k_g^{eff}})_g^\delta})(3)$$

Here z_i^g is the vector of individual i 's genotypes in gene g and r_p the number of genes in pathway p. Scaling parameters k_g^{eff}, μ^g and δ_g adjust for the number of genes in the pathway and the number of SNPs within these genes (for more details refer to Freytag et al. 2012).

A kernel of this type can be calculated using the function `kernel_sia()` with
the following arguments

- A `GWASdata` object containing the genotype information.
- A `pathway` object specifying the pathway to be tested.
- A value for argument `calculation` to decide how the kernel should be calculated. Currently only `cpu` for cpu calculation is available.

```
K_sia <- sia_kernel(gwas, p, calculation='cpu')
```

will return a quadratic `matrix` of dimension equal to the number of individuals in the `GWASdata` object.

Network Kernel (Net)

The network kernel incorporates information about gene-gene interactions into the model. It is defined as

$$K = ZANA^tZ^t (4)$$

where matrix A maps SNPs to genes, N represents the underlying network structure, and Z is the genotype matrix. The network based kernel matrix for a pathway can be calculated with the function `kernel_net()`. Following arguments are needed

- A `GWASdata` object containing the genotype information.
- A `pathway` object specifying the pathway to be tested.
- A value for argument 'calculation' to decide how the kernel should be calculated.

```
K_net <- net_kernel(gwas, p, calculation='cpu')
```

will return a quadratic `matrix` of dimension equal to the number of individuals in the `GWASdata` object.

Alternatively, kernel matrices can be calculated using the function `calc_kernel()`. Here the kernel type is specified via an additional argument `type`. It can be set to lin, sia or net.

```
K <- calc_kernel(gwas, p, type='lin', parallel='none')
```

This function will simply call the suitable kernel function as described above and therefore has the same output.

Variance Component Test

A pathways influence on the probability of being a case is evaluated in a variance component test. The test statistic is

$$Q = \frac{1}{2}(y-\mu)^t K (y-\mu) (5)$$

with μ the vector of null model estimators given by $\mu_i = logit^{-1}(x_i^t\beta)$ for an individual i and K a kernel matrix of the pathway to be tested. Q follows a mixture of X^2 distributions which can be approximated using the Satterthwaite procedure (Schaid 2012) or Davies method as implemented in the R package `QuadCompForm` (Davies 1980). More details on the test can be found in Wu et al (2010).

In `kangar00` the logistic kernel machine test can be applied to a SNP set defining a pathway with the function `lkmt`. It needs the following arguments

- A formula specifying the null model to be used in the test. The dependent variable is the case control status of the individual (in the example denoted as 'pheno') and is explained by an intercept and optional covariates.
- A linear, size-adjusted or network kernel matrix calculated by one of the kernel functions `kernel_lin()`, `kernel_sia()` or `kernel_net()`.

- A `GWASdata` object including the genotype based on which the test should be performed.
- A `character` specifying which method should be used to calculate the p-value. Available are 'satt' for the Satterthwaite approximation (Schaid

2010) or 'davies' for Davies method (Davies 1980).

```
pval_net <- lkmt(pheno ~ 1+sex+age, K_mat, my_gwas, method='satt')
```

will return an object of type `lkmt` giving the test result for the pathway on which the kernel matrix 'K_mat' was calculated. The `GWASdata` object 'my_gwas' has to be the same as used to calculate the kernel matrix. The formula above would for example fit for a phenotype file of the following format (IDs in first column are always required in phenotype file)

ID	pheno	sex	age	smoker
ind1	1	1	41	1
ind2	0	0	38	0
ind3	1	1	56	1
...	...	...	...	...

note, that the columns to be used in the model are specified in the formula given to the `lkmt()` function and not all covariates have to be used.

References

- Liu D, Ghosh D, Lin X. 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 2008 9:292.
- Schaid DJ: Genomic similarity and kernel methods I: advancements by building on mathematical and statistical foundations. Hum Hered 2010, 70:109-131.
- Davies R: Algorithm as 155: the distribution of a linear combination of chi-2 random variables. J R Stat Soc Ser C 1980, 29:323-333.
- Wu MC, Kraft P, Epstein MP, Taylor DM, Chanock SJ, Hunter DJ, Lin X: Powerful SNP-Set Analysis for Case-Control Genome-Wide Association Studies. Am J Hum Genet 2010, 86:929-42
- Cunningham F, Amode MR, Barrell D et al. Ensembl 2015. Nucleic Acids Research 2015 43 Database issue:D662-D669
- Kanehisa, M., Goto, S., Sato, Y., Kawashima, M., Furumichi, M., and Tanabe, M.; Data, information, knowledge and principle: back to metabolism in KEGG. Nucleic Acids Res. 42, D199-D205 (2014).
- Freytag S, Bickeboeller H, Amos CI, Kneib T, Schlather M: A Novel Kernel for Correcting Size Bias in the Logistic Kernel Machine Test with an Application to Rheumatoid Arthritis. Hum Hered. 2012, 74(2):97-108.
- Freytag S, Manitz J, Schlather M, Kneib T, Amos CI, Risch A, Chang-Claude J, Heinrich J, Bickeboeller H: A network-based kernel machine test for the identification of risk pathways in genome-wide association studies. Hum Hered. 2013, 76(2):64-75.

Package 'kangar00'

April 27, 2017

Type Package

Title Kernel Approaches for Nonlinear Genetic Association Regression

Version 1.0

Date 2017-04-26

Author Juliane Manitz [aut], Stefanie Friedrichs [aut], Patricia Burger [aut], Benjamin Hofner [aut], Ngoc Thuy Ha [aut], Saskia Freytag [ctb], Heike Bickeboeller [ctb]

Maintainer Juliane Manitz <r@manitz.org>

Description Methods to extract information on pathways, genes and SNPs from online databases. It provides functions for data preparation and evaluation of genetic influence on a binary outcome using the logistic kernel machine test (LKMT). Three different kernel functions are offered to analyze genotype information in this variance component test: A linear kernel, a size-adjusted kernel and a network based kernel.

License GPL-2

Collate 'pathway.r' 'GWASdata.r' 'data.R' 'kernel.r' 'lkmt.r'

Depends R (>= 3.1.0)

Imports methods, KEGGgraph, biomaRt, bigmemory, sqldf, CompQuadForm, data.table, lattice, igraph

LazyData true

Suggests knitr, rmarkdown

VignetteBuilder knitr

NeedsCompilation no

RoxygenNote 6.0.1

Repository CRAN

Date/Publication 2017-04-27 11:41:53 UTC

R topics documented:

`kangar00-package` *kangar00 package*

Description

This package includes methods to extract information on pathways, genes and SNPs from online databases and to evaluate these data using the logistic kernel machine test (LKMT) (Liu et al. 2008).

We defined SNP sets representing genes and whole pathways using knowledge on gene membership and interaction from the Kyoto Encyclopedia of Genes and Genomes (KEGG) database (Kanehisa et al. 2014). SNPs are mapped to genes via base pair positions of SNPs and transcript start and end points of genes as documented in the Ensemble database (Cunningham et al. 2015).

In the LKMT, we employed the linear kernel (Wu et al. 2010) as well as two more advanced kernels, adjusting for size bias in the number of SNPs and genes in a pathway (size-adjusted kernels), and incorporating the network structure of genes within the pathway (pathway kernels), respectively (Freytag et al. 2012, 2014). P-values are derived in a variance component test using a moment matching methode (Schaid, 2010) or Davies' algorithm (Davies, 1980).

Details

Package:	kangar00
Version:	1.0
Date:	2017-04-26
License:	GPL-2

Author(s)

Juliane Manitz [aut], Stefanie Friedrichs [aut], Patricia Burger [aut], Benjamin Hofner [aut], Ngoc Thuy Ha [aut], Saskia Freytag [ctb], Heike Bickeboeller [ctb]
Maintainer: Juliane Manitz <r@manitz.de>

References

- Cunningham F., M. Ridwan Amode, Daniel Barrell et al. Ensembl 2015. Nucleic Acids Research 2015 43 Database issue:D662-D669

- Davies R: Algorithm as 155: the distribution of a linear combination of chi-2 random variables. J R Stat Soc Ser C 1980, 29:323-333.

- Freytag S, Bickeboeller H, Amos CI, Kneib T, Schlather M: A Novel Kernel for Correcting Size Bias in the Logistic Kernel Machine Test with an Application to Rheumatoid Arthritis. Hum Hered. 2012, 74(2):97-108.

- Freytag S, Manitz J, Schlather M, Kneib T, Amos CI, Risch A, Chang-Claude J, Heinrich J, Bickeboeller H: A network-based kernel machine test for the identification of risk pathways in genome-wide association studies. Hum Hered. 2013, 76(2):64-75.

- Friedrichs S., J. Manitz, P. Burger, C.I. Amos, A. Risch, J.C. Chang-Claude, H.E. Wichmann, T. Kneib, H. Bickeboeller, B. Hofner: Pathway-Based Kernel Boosting for the Analysis of Genome-Wide Association Studies. 2017. Submitted to Computational and Mathematical Methods in Medicine.

- Kanehisa, M., Goto, S., Sato, Y., Kawashima, M., Furumichi, M., and Tanabe, M.; Data, information, knowledge and principle: back to metabolism in KEGG. Nucleic Acids Res. 42, D199-D205 (2014).

- Liu D, Ghosh D, Lin X. 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. 2008 9:292.

- Schaid DJ: Genomic similarity and kernel methods I: advancements by building on mathematical and statistical foundations. Hum Hered 2010, 70:109-131.

- Wu MC, Kraft P, Epstein MP, Taylor DM, Chanock SJ, Hunter DJ, Lin X: Powerful SNP-Set Analysis for Case-Control Genome-Wide Association Studies. Am J Hum Genet 2010, 86:929-42

`anno` *Example annotation file for three pathways.*

Description

A dataset containing an annotation example for 4056 SNPs in three different pathways.

Usage

```
data(anno)
```

Format

A `data frame` with 4056 rows and 5 variables:

pathway includes KEGG identifiers of three example pathways

gene names of genes in the pathways

chr specifies the chromosome

snp includes rs-numbers of example SNPs

position gives positions of example SNPs

Source

simulated data

`calc_kernel` *Calculates the kernel-matrix for a pathway*

Description

Uses individuals' genotypes to create a `kernel` object including the calculated kernel `matrix` for a specific `pathway`. Each numeric value within this `matrix` is calculated from two individuals' genotypevectors of the SNPs within the `pathway` by a kernel function. It can be interpreted as the genetic similiarity of the individuals. Association between the `pathway` and a binary phenotype (case-control status) can be evaluated in the logistic kernel machine test, based on the `kernel` object. Three kernel functions are available.

Usage

```
## S4 method for signature 'GWASdata'
calc_kernel(object, pathway, knots = NULL,
  type = c("lin", "sia", "net"), calculation = c("cpu", "gpu"), ...)

## S4 method for signature 'GWASdata'
lin_kernel(object, pathway, knots = NULL,
  calculation = c("cpu", "gpu"), ...)

## S4 method for signature 'GWASdata'
sia_kernel(object, pathway, knots = NULL,
  calculation = c("cpu", "gpu"), ...)

## S4 method for signature 'GWASdata'
net_kernel(object, pathway, knots = NULL,
  calculation = c("cpu", "gpu"), ...)
```

Arguments

`object`	`GWASdata` object containing the genotypes of the individuals for which a `kernel` will be calculated.
`pathway`	object of the class `pathway` specifying the SNP set for which a `kernel` will be calculated.
`knots`	`GWASdata` object, if specified a `kernel` will be computed.
`type`	`character` indicating the `kernel` type: Use `'lin'` to specify the linear kernel, `'sia'` for the size-adjusted or `'net'` for the network-based kernel.
`calculation`	`character` specifying if the kernel `matrix` is computed on CPU or GPU.
`...`	further arguments to be passed to `kernel` computations.

Details

Different types of kernels can be constructed:

- `type='lin'` creates the linear kernel assuming additive SNP effects to be evaluated in the logistic kernel machine test.
- `type='sia'` calculates the size-adjusted kernel which takes into consideration the numbers of SNPs and genes in a `pathway` to correct for size bias.
- `type='net'` calculates the network-based kernel. Here not only information on gene membership and gene/pathway size in number of SNPs is incorporated, but also the interaction structure of genes in the `pathway`.

For more details, check the references.

Value

Returns an object of class `kernel`, including the similarity `matrix` of the `pathway` for the considered individuals.

If knots are specified low-rank kernel of class a lowrank_kernel will be returned, which is not necessarily quadratic and symmetric.

Methods (by class)

- GWASdata: Calculates a linear kernel
- GWASdata: Calculates a size adjusted-kernel
- GWASdata: Calculates a network-based kernel

Author(s)

Stefanie Friedrichs, Juliane Manitz

References

- Wu MC, Kraft P, Epstein MP, Taylor DM, Chanock SJ, Hunter DJ, Lin X Powerful SNP-Set Analysis for Case-Control Genome-Wide Association Studies. Am J Hum Genet 2010, 86:929-42
- Freytag S, Bickeboeller H, Amos CI, Kneib T, Schlather M: A Novel Kernel for Correcting Size Bias in the Logistic Kernel Machine Test with an Application to Rheumatoid Arthritis. Hum Hered. 2012, 74(2):97-108.
- Freytag S, Manitz J, Schlather M, Kneib T, Amos CI, Risch A, Chang-Claude J, Heinrich J, Bickeboeller H: A network-based kernel machine test for the identification of risk pathways in genome-wide association studies. Hum Hered. 2013, 76(2):64-75.

See Also

kernel-class,pathway

Examples

```
data(gwas)
data(hsa04020)
calc_kernel(gwas, hsa04020, knots = NULL, type='net', calculation='cpu')
```

---

geno *Example genotypes for 50 individuals.*

Description

A matrix containing example genotypes for 4056 SNPs of 50 individuals. Column names give the rs-numbers of 4056 example SNPs, row names the identifiers of 50 example individuals.

Usage

```
data(geno)
```

Format

A matrix with 5 rows and 4056 columns:

each entry in the matrix represents a simulated minor allele count for the corresponding SNP and individual.

Source

simulated data

---

get_anno,snp_info,pathway_info-method

Annotates SNPs via genes to pathways

Description

A function to create the annotation for a GWASdata object. It combines a snp_info and a pathway_info object into an annotation data.frame used for pathway analysis on GWAS. SNPs are assigned to pathways via gene membership.

Usage

```
## S4 method for signature 'snp_info,pathway_info'
get_anno(object1, object2, ...)
```

Arguments

object1	A snp_info object with SNP information as returned by the snp_info function. The included data frame contains the columns 'chr', 'position' and 'snp'.
object2	A pathway_info object with information on genes contained in pathways. It is created by the pathway_info function and contains a data frame with columns 'pathway', 'gene_start', 'gene_end', 'chr', 'gene'.
...	further arguments can be added.

Value

A data.frame mapping SNPs to genes and genes to pathways. It includes the columns 'pathway', 'gene', 'chr', 'snp' and 'position'.

Author(s)

Stefanie Friedrichs, Saskia Freytag, Ngoc-Thuy Ha

See Also

snp_info, pathway_info

Examples

```
data(hsa04022_info)
data(rs10243170_info)
get_anno(rs10243170_info, hsa04022_info)
```

`get_network_matrix,pathway-method`
Function to calculate the network `matrix` *for a* `pathway` *object*

Description

This function creates the network matrix representing the gene-gene interaction structure within a particular `pathway`. In this process a KEGG kgml file is downloaded and saved in the working directory.

Usage

```
## S4 method for signature 'pathway'
get_network_matrix(object, directed = TRUE)
```

Arguments

`object`	A `pathway` object identifying the pathway for which gene interaction infomation should be extracted. Here, KEGG IDs of format 'hsa00100' are used and information is downloaded from the KEGG database.
`directed`	A `logic` argument, stating whether the network matrix should be returned directed (`TRUE`) or undirected (`FALSE`).

Value

The altered `pathway` object, in which the slots `'adj'` and `'sign'` have been changed according to the downloaded information on the `pathway`.

Author(s)

Stefanie Friedrichs, Patricia Burger

gwas *Example* GWASdata *object.*

Description

An object of type GWASdata containing the example files for annotation, phenotypes and genotypes.

Usage

```
data(gwas)
```

Format

An object of class GWASdata:

geno contains example genotypes

anno example annotation for three pathways

pheno exemplary phenotypes for all 'genotyped' individuals

desc a description of the GWAS study, here 'example study'

Source

simulated data

GWASdata *S4 class for an object representing a Genome-wide Assocaition Study.*

Description

S4 class for an object representing a Genome-wide Assocaition Study.

'GWASdata' is a GWASdata object constructor.

show displays basic information on GWASdata object

summary summarizes the content of a GWASdata object and gives an overview about the information included in a GWASdata object. Summary statistics for phenotype and genotype data are calculated.

GeneSNPsize creates a data.frame of pathway names with numbers of snps and genes in each pathway.

Usage

```
GWASdata(object, ...)

## S4 method for signature 'ANY'
GWASdata(geno, anno, pheno = NULL, desc = "")

## S4 method for signature 'GWASdata'
show(object)

## S4 method for signature 'GWASdata'
summary(object)

## S4 method for signature 'GWASdata'
GeneSNPsize(object)
```

Arguments

object	A GWASdata object.
...	Further arguments can be added to the function.
geno	An object of any type, including the genotype information.
anno	A data.frame containing the annotation file for the GWASdata object.
pheno	A data.frame specifying individual IDs, phenotypes and covariates to be included in the regression model.
desc	A character giving the GWAS description, e.g. name of study.

Methods (by generic)

- GeneSNPsize: creates a data.frame of pathway names with numbers of snps and genes in each pathway.

Slots

geno An object of any type, including genotype information. The format needs to be one line per individual and on colum per SNP in minor-allele coding (0,1,2). Other values between 0 and 2, as from impute dosages, are allowed. Missing values must be imputed prior to creation of a GWASdata object.

anno A data.frame mapping SNPs to genes and genes to pathways. Needs to include the columns 'pathway' (pathway ID, e.g. hsa number from KEGG database), 'gene' (gene name (hgnc_symbol)), 'chr' (chromosome), 'snp' (rsnumber) and 'position' (base pair position of SNP).

pheno A data.frame specifying individual IDs, phenotypes and covariates to be included in the regression model e.g. ID, pheno, sex, pack.years. Note: IDs have to be in the first column!

desc A character giving the GWAS description, e.g. name of study.

Author(s)

Juliane Manitz, Stefanie Friedrichs

Examples

```
data(pheno)
data(geno)
data(anno)
gwas <- new('GWASdata', pheno=pheno, geno=geno, anno=anno, desc="some study")
# show method
data(gwas)
gwas
# summary method
data(gwas)
summary(gwas)

# SNPs and genes in pathway
data(gwas)
GeneSNPsize(gwas)
```

`hsa04020` *Example* `pathway` *object for pathway hsa04020.*

Description

An object of class `pathway` for the pathway with KEGG identifier hsa04020.

Usage

```
data(hsa04020)
```

Format

A `pathway` object including 180 genes.

id KEGG identifier of the example pathways

adj gives the quadratic adjacency `matrix` for the pathway and with that the network topology. Matrix dimensions equal the number of genes in the pathway

sign includes a `vector` of signs to distinguish activations and inhibitions in the adjacency `matrix`

Source

simulated data and Ensembl extract

hsa04022_info — *Example* pathway_info *object for* pathway *hsa04022.*

Description

An object of class pathway_info for the pathway with KEGG identifier hsa04020.

Usage

```
data(hsa04022_info)
```

Format

A pathway_info object including information on 163 genes.

info a data frame including information on genes included in pathway. Has columns 'pathway', 'gene_start', 'gene_end', 'chr', and 'gene'

Source

Ensembl extract

kernel-class — *An S4 class representing a kernel matrix calculated for a pathway*

Description

An S4 class representing a kernel matrix calculated for a pathway

show displays the kernel object briefly

summary generates a kernel object summary including the number of individuals and genes for the pathway

plot creates an image plot of a kernel object

Usage

```
## S4 method for signature 'kernel'
show(object)

## S4 method for signature 'kernel'
summary(object)

## S4 method for signature 'kernel,missing'
plot(x, y = NA, hclust = FALSE, ...)
```

Arguments

object	An object of class kernel
x	the kernel object to be plotted.
y	missing (placeholder).
hclust	logical, indicating whether a dendrogram should be added.
...	further arguments to be passed to the function.

Slots

type A character representing the kernel type: Use 'lin' for linear kernel, 'sia' for the size-adjusted or 'net' for the network-based kernel.

kernel A kernel matrix of dimension equal to the number of individuals

pathway A pathway object

Author(s)

Juliane Manitz

Examples

```
data(net.kernel.hsa04020)
show(net.kernel.hsa04020)
summary(net.kernel.hsa04020)
plot(net.kernel.hsa04020)
```

---

lkmt-class *An S4 class to represent the variance component test.*

Description

An S4 class to represent the variance component test.

show Shows basic information on lkmt object

summary Summarizes information on lkmt object

Usage

```
## S4 method for signature 'lkmt'
show(object)

## S4 method for signature 'lkmt'
summary(object)
```

Arguments

object	An object of class lkmt.
...	Further arguments can be added to the function

Value

`show` Basic information on `lkmt` object.

`summary` Summarized information on `lkmt` object.

Slots

`formula` A formula stating the regression nullmodel that will be used in the variance component test.

`kernel` An object of class `kernel` representing the similarity matrix of the individuals based on which the pathways influence is evaluated.

`GWASdata` An object of class `GWASdata` including the data on which the test is conducted.

`statistic` A `vector` giving the value of the variance component test statistic.

`df` A `vector` containing the number of degrees of freedom.

`p.value` A `vector` giving the p-value calculated for the \ codepathway object considered in the variance component test.

For details on the variance component test see the references.

Author(s)

Juliane Manitz, Stefanie Friedrichs

References

- Liu D, Lin X, Ghosh D: Semiparametric regression of multidimensional genetic pathway data: least-squares kernel machines and linear mixed models. Biometrics 2007, 63(4):1079-88.
- Wu MC, Kraft P, Epstein MP, Taylor DM, Chanock SJ, Hunter DJ, Lin X: Powerful SNP-Set Analysis for Case-Control Genome-Wide Association Studies. Am J Hum Genet 2010, 86:929-42

Examples

```
# show method
data(lkmt.net.kernel.hsa04020)
lkmt.net.kernel.hsa04020
# summary method
summary(lkmt.net.kernel.hsa04020)
```

---

lkmt.net.kernel.hsa04020

Example test result for the network-based `kernel` for `pathway` hsa04020.

Description

An object of class `lkmt` containing exemplary test results for an application of the logistic kernel machine test, derived from the example data.

Usage

```
data(lkmt.net.kernel.hsa04020)
```

Format

An object of class `lkmt` for the network-based `kernel` and the `pathway` hsa04020.

formular gives a formular defining the nullmodel used in the logistic kernel machine test

kernel includes the `kernel` object of the `pathway` to be evaluated

GWASdata gives the `GWASdata` object including the study data considered in testing

statistic gives the value of the test statistic

df specifies the degrees of freedom

p.value includes teh p-value resulting from the test

Source

simulated data and Ensembl extract

---

lkmt_test *A function to calculate the p-values for kernel matrices.*

Description

A function to calculate the p-values for kernel matrices.

For parameter `'satt'` a pathways influence on the probability of beeing a case is evaluated in the logistic kernel machine test and p-values are determined using a Sattherthwaite Approximation as described by Dan Schaid.

For parameter `'davies'` a pathways influence on the probability of beeing a case is evaluated using the p-value calculation method described by Davies. Here the function `davies` from package **CompQuadForm** is used.

Usage

```
lkmt_test(formula, kernel, GWASdata, method = c("satt", "davies"), ...)

## S4 method for signature 'matrix'
score_test(x1, x2)

## S4 method for signature 'matrix'
davies_test(x1, x2)
```

Arguments

formula	The formula to be used for the regression nullmodel.
kernel	An object of class kernel including the pathway representing kernel-matrix based on which the test statistic will be calculated.
GWASdata	A GWASdata object stating the data used in analysis.
method	A character specifying which method will be used for p-value calculation. Available are 'satt' for the Satterthwaite approximation and 'davies' for Davies' algorithm. For more details see the references.
...	Further arguments can be given to the function.
x1	A matrix which is the similarity matrix calculated for the pathway to be tested.
x2	An lm or glm object of the nullmodel with fixed effects covariates included, but no genetic random effects.

Value

An lkmt object including the following test results

- The formula of the regression nullmodel used in the variance component test.
- An object of class kernel including the similarity matrix of the individuals based on which the pathways influence is evaluated.
- An object of class GWASdata stating the data on which the test was conducted.
- statistic A vector giving the value of the variance component test statistic.
- df A vector giving the number of degrees of freedom.
- p.value A vector giving the p-value calculated for the pathway in the variance component test.

Author(s)

Stefanie Friedrichs, Juliane Manitz

References

For details on the variance component test

- Wu MC, Kraft P, Epstein MP, Taylor DM, Chanock SJ, Hunter DJ, Lin X: Powerful SNP-Set Analysis for Case-Control Genome-Wide Association Studies. Am J Hum Genet 2010, 86:929-42
- Liu D, Lin X, Ghosh D: Semiparametric regression of multidimensional genetic pathway data: least-squares kernel machines and linear mixed models. Biometrics 2007, 63(4):1079-88.

For details on the p-value calculation see

- Schaid DJ: Genomic Similarity and Kernel Methods I: Advancements by Building on Mathematical and Statistical Foundations. Hum Hered 2010, 70:109-31
- Davies R: Algorithm as 155: the distribution of a linear combination of chi-2 random variables. J R Stat Soc Ser C 1980, 29:323-333.

Examples

```
data(net.kernel.hsa04020)
data(gwas)
lkmt_test(pheno ~ sex + age, net.kernel.hsa04020, gwas, method='satt')
```

`lowrank_kernel-class` *An S4 class to represent a low-rank kernel for a SNPset at specified knots*

Description

An S4 class to represent a low-rank kernel for a SNPset at specified knots

Details

This kernel is used for predictions. If observations and knots are equal, better construct a full-rank kernel of class `kernel`.

Slots

`type` character, kernel type: Use `'lin'` for the linear kernel, `'sia'` for the size-adjusted or `'net'` for the network-based kernel.

`kernel` kernel `matrix` of dimension equal to individuals

`pathway` `pathway` object

Author(s)

Juliane Manitz

Examples

```
data(gwas)
calc_kernel(gwas, hsa04020, knots=gwas, type='lin', calculation='cpu')
## Not run:
gwas2 <- new('GWASdata', pheno=pheno[1:10,], geno=geno[1:10,], anno=anno, desc=" study 2")
calc_kernel(gwas, hsa04020, knots = gwas2, type='net', calculation='cpu')

## End(Not run)
```

`make_psd,matrix-method`
Adjust network `matrix` *to be positive semi-definite*

Description

Adjust network `matrix` to be positive semi-definite

Usage

```
## S4 method for signature 'matrix'
make_psd(x, eps = sqrt(.Machine$double.eps))
```

Arguments

`x`	A `matrix` specifying the network adjacency matrix.
`eps`	A `numeric` value, setting the tolance for smallest eigenvalue adjustment

Details

For a `matrix` N, the closest positive semi-definite `matrix` is calculated as N* = rho*N + (1+rho)*I, where I is the identity `matrix` and rho = 1/(1 - lambda) with lambda the smallest eigenvalue of N. For more details check the references.

Value

The `matrix` `x`, if it is positive definite and the closest positive semi-definite `matrix` if `x` is not positive semi-definite.

Author(s)

Juliane Manitz, Saskia Freytag, Stefanie Friedrichs

References

- Freytag S, Manitz J, Schlather M, Kneib T, Amos CI, Risch A, Chang-Claude J, Heinrich J, Bickeboeller H: A network-based kernel machine test for the identification of risk pathways in genome-wide association studies. Hum Hered. 2013, 76(2):64-75.

net.kernel.hsa04020 *Example network-based kernel matrix for pathway hsa04020.*

Description

An example of a kernel object.

Usage

```
data(net.kernel.hsa04020)
```

Format

An object of class `kernel` and type 'network' for the pathway hsa04020.

type specifies which kernel function was used to calculate the kernel

kernel includes the kernel matrix calculated for the pathway

pathway includes the `pathway` object of the pathway, for which the kernel matrix was calculated

Source

simulated data and Ensembl extract

pathway *An S4 class to represent a gene-gene interaction network*

Description

An S4 class to represent a gene-gene interaction network

`'pathway'` is the `pathway` object constructor.

`show` displays the `pathway` object briefly

`summary` generates a `pathway` object summary including basic network properties.

`pathway2igraph` converts a `pathway` object into an `igraph` object with edge attribute `sign`

analyze `pathway` network properties

`get_genes` is a helper function that extracts the gene names in a `pathway` and returns a `vector` containing `character` elements of gene names

`plot` plots `pathway` as `igraph` object

`sample_genes` function randomly selects effect genes in a `pathway` and returns a `vector` of length `no` with vertex id's of sampled genes

Usage

```
pathway(object, ...)

## S4 method for signature 'ANY'
pathway(id, adj, sign)

## S4 method for signature 'pathway'
show(object)

## S4 method for signature 'pathway'
summary(object)

## S4 method for signature 'pathway'
pathway2igraph(object)

## S4 method for signature 'pathway'
analyze(object, ...)

## S4 method for signature 'pathway'
get_genes(object)

## S4 method for signature 'pathway,missing'
plot(x, y = NA, highlight.genes = NULL,
  gene.names = c("legend", "nodes", NA), main = NULL, asp = 0.95,
  vertex.size = 11, vertex.color = "khaki1", vertex.label.cex = 0.8,
  edge.width = 2, edge.color = "olivedrab4", ...)

## S4 method for signature 'pathway'
sample_genes(object, no = 3)
```

Arguments

object	An object of class pathway-class
...	further arguments specifying plotting options in plot.igraph
id	A character repesenting the pathway id.
adj	A matrix respresenting the network adjacency matrix of dimension equaling the number of genes (1 interaction, 0 otherwise)
sign	A numeric vector indicating the interaction type for each link (1 activation, -1 inhibition) in the interaction network for the pathway.
x	pathway object
y	missing (placeholder)
highlight.genes	vector of gene names or node id's, which should be highlighted in a different color, default is NULL so that no genes are highlighted
gene.names	character indicating whether the genes names should appear in a legend ('legend'), as vertex label ('nodes'), or should be omitted (NA)

main	optional overall main title, default is NULL, which uses the pathway id
asp	a numeric constant, which gives the aspect ratio parameter for plot, default is 0.95
vertex.size	a numeric constant specifying the vertex size, default is 11
vertex.color	a character or numeric constant specifying the vertex color, default is 'khaki1'
vertex.label.cex	a numeric constant specifying the the vertex label size, default is 0.8,
edge.width	a numeric constant specifying the edge width, default is 2
edge.color	a character or numeric constant specifying the edge color, default is 'olivedrab4'
no	a numeric constant specifying the number of genes to be sampled, default is 3

Value

analyze returns a data.frame consisting of

id pathway id,

vcount number of genes,

ecount number of links,

inh_ecount number of inhibition links,

density network density,

av_deg average degree,

inh_deg average degree of inhibition links,

diam network diamter,

trans transitivity, and

s_trans signed transitivity (Kunegis et al., 2009).

Methods (by generic)

- analyze:
- get_genes:
- sample_genes:

Slots

id A character repesenting the pathway id, e.g. hsa00100 as used in the KEGG database.

adj A matrix respresenting the network adjacency matrix of dimension equaling the number of genes (1 interaction, 0 otherwise)

sign A numeric vector indicating the interaction type for each link (1 activation, -1 inhibition) in the interaction network for the pathway.

Author(s)

Juliane Manitz

References

Details to the computation and interpretation can be found in:

- Kolaczyk, E. D. (2009). Statistical analysis of network data: methods and models. Springer series in statistics. Springer.
- Kunegis, J., A. Lommatzsch, and C. Bauckhage (2009). The slashdot zoo: Mining a social network with negative egdes. In Proceedings of the 18th international conference on World wide web, pp. 741-750. ACM Press.

Examples

```
pathway(id="hsa04022", adj=matrix(0), sign=as.vector(matrix(0)[matrix(0)!=0]))

#show method
data(hsa04020)
hsa04020
#summary method
data(hsa04020)
summary(hsa04020)
# convert to \code{\link[igraph]{igraph}} object
data(hsa04020)
str(hsa04020)
g <- pathway2igraph(hsa04020)
str(g)
# analyse \code{\link{pathway}} network properties
data(hsa04020)
summary(hsa04020)
analyze(hsa04020)
# extract gene names from \code{\link{pathway}}
get_genes(hsa04020)
# plot \code{\link{pathway}} as \code{\link[igraph]{igraph}} object
plot(hsa04020)
sample3 <- sample_genes(hsa04020, no = 3)
plot(hsa04020, highlight.genes = sample3)

# sample effect genes
sample3 <- sample_genes(hsa04020, no = 3)
plot(hsa04020, highlight.genes = sample3)
sample5 <- sample_genes(hsa04020, no = 5)
plot(hsa04020, highlight.genes = sample5)
```

---

pathway_info *An S4 class for an object assigning genes to pathways*

Description

An S4 class for an object assigning genes to pathways

This function lists all genes formig a particular pathway. Start and end positions of these genes are extracted from the Ensemble database. The database is accessed via the R-package **biomaRt**.

show Shows basic information on pathway_info object

summary Summarizes information on pathway_info object

Usage

```
pathway_info(x)

## S4 method for signature 'character'
pathway_info(x)

## S4 method for signature 'pathway_info'
show(object)

## S4 method for signature 'pathway_info'
summary(object)
```

Arguments

x	A character identifying the pathway for which gene infomation should be extracted. Here KEGG IDs (format: 'hsa00100') are used.
object	An object of class pathway_info.

Value

A data.frame including as many rows as genes appear in the pathway. for each gene its name, the start and end point and the chromosome it lies on are given.

show Basic information on pathway_info object.

summary Summarized information on pathway_info object.

Slots

info A data.frame including information on genes contained in pathways with columns 'pathway', 'gene_start', 'gene_end', 'chr' and 'gene'.

Author(s)

Stefanie Friedrichs

Examples

```
pathway_info("hsa04022")

# show method
data(hsa04022_info)
hsa04022_info
# summary method
data(hsa04022_info)
summary(hsa04022_info)
```

pheno *Example phenotype file for 50 individuals.*

Description

A dataset containing simulated example phenotypes for 50 individuals row names include the identifiers of 50 example individuals.

Usage

```
data(pheno)
```

Format

A `data frame` with 50 rows and 3 variables:

pheno includes the case-control status for each individual, coded as 1(case) or 0 (control)

sex includes gender information for the 50 individuals, coded as 1 (male) or 0 (female)

age numerical value giving the persons age

Source

simulated data

`read_geno,character-method`

read genotype data from file to one of several available objects, which can be passed to a GWASdata object `GWASdata`.

Description

read genotype data from file to one of several available objects, which can be passed to a GWASdata object `GWASdata`.

Usage

```
## S4 method for signature 'character'
read_geno(file.path, save.path = NULL, sep = " ",
  header = TRUE, use.fread = TRUE, use.big = FALSE, row.names = FALSE,
  ...)
```

Arguments

file.path	character giving the path to the data file to be read
save.path	character containing the path for the backingfile
sep	character. A field delimeter. See read.big.matrix for details.
header	logical. Does the data set contain column names?
use.fread	logical. Should the dataset be read using the function fread fread from package **data.table**?
use.big	logical. Should the dataset be read using the function read.big.matrix from package **bigmemory**?
row.names	logical. Does the dataset include rownames?
...	further arguments to be passed to read_geno.

Details

If the data set contains rownames specified, set option has.row.names = TRUE.

Examples

```
## Not run:
path <- system.file("extdata", "geno.txt", package = "kangar00")
geno <- read_geno(path, save.path = getwd(), sep = " ", use.fread = FALSE, row.names = FALSE)

## End(Not run)
```

---

rewire_network *Rewires interactions in a pathway, which go through a gene not represented by any SNPs in the considered GWASdata dataset.*

Description

Rewires interactions in a pathway, which go through a gene not represented by any SNPs in the considered GWASdata dataset.

Usage

```
## S4 method for signature 'pathway'
rewire_network(object, x)
```

Arguments

object	pathway object which's network matrix will be rewired
x	A vector of gene names, indicating which genes are not represented by SNPs in the considered GWASdata object and will be removed

Value

A pathway object including the rewired network matrix

Author(s)

Stefanie Friedrichs, Juliane Manitz

Examples

```
## Not run:
data(hsa04020)
rewire_network(hsa04020, c("PHKB", "ORAI2"))

## End(Not run)
```

---

rs10243170_info *Example snp_info object for SNP rs10243170.*

Description

An object of class snp_info for rs10243170.

Usage

```
data(rs10243170_info)
```

Format

A snp_info object including information on the SNP as extracted from the Ensembl database.

info a data frame including the extracted information on the SNP. Columns given are 'chr', 'position', and 'rsnumber'

Source

Ensembl extract

snp_info *An S4 class for an object assigning SNP positions to rs-numbers (for internal use)*

Description

An S4 class for an object assigning SNP positions to rs-numbers (for internal use)

This function gives for a `vector` of SNP identifiers the position of each SNP as extracted from the Ensemble database. The database is accessed via the R-package **biomaRt**.

`show` Shows basic information on `snp_info` object

`summary` Summarizes information on `snp_info` object

Usage

```
snp_info(x, ...)

## S4 method for signature 'character'
snp_info(x)

## S4 method for signature 'snp_info'
show(object)

## S4 method for signature 'snp_info'
summary(object)
```

Arguments

`x`	A `character vector` of SNP rsnumbers for which positions will be extracted.
`...`	further arguments can be added.
`object`	An `object` of class `snp_info`.

Value

A `data.frame` including the SNP positions with columns 'chromosome', 'position' and 'snp'. SNPs not found in the Ensemble database will not be listed in the returned `snp_info` object, SNPs with multiple positions will appear several times.

`show` Basic information on `snp_info` object.

`summary` Summarized information on `snp_info` object.

Slots

`info` A `data.frame` including information on SNP positions

Author(s)

Stefanie Friedrichs

Examples

```
# snp_info
data(rs10243170_info)
snp_info(c("rs234"))

# show
data(rs10243170_info)
rs10243170_info
# summary
data(rs10243170_info)
summary(rs10243170_info)
```

Index

B Curriculum Vitae

Stefanie Friedrichs

Date of birth	15 December, 1987
Place of birth	Kassel
Nationality	German

Education

Since 10/2013	PhD position at the Department of Genetic Epidemiology, University Medical Centre, Georg-August-University Göttingen, supervised by Prof. H. Bickeböller
Since 10/2013	Member of the Research Training Group "Scaling Problems in Statistics" (GRK 1644) funded by the Deutsche Forschungsgemeinschaft (DFG)
04/2013 - 09/2013	Employee at the Department of Genetic Epidemiology, University Medical Centre, Georg-August-University Göttingen
04/2011 - 03/2013	Master of Science in Mathematics Georg-August-University Göttingen
10/2007 - 03/2011	Bachelor of Science in Mathematics Georg-August-University Göttingen
06/2007 - 09/2007	Employee at documenta 12 documenta und Museum Fridericianum Veranstaltungs-GmbH, Kassel
08/1998 - 06/2007	Secondary school education, Abitur
08/1994 - 06/1998	Primary school education

Publications

2017 Yasmeen S, Burger P, Friedrichs S, Papiol S, Bickeböller H: **Relating Drug Response to Epigenetic and Genetic Markers Using a Region-Based Kernel Score Test** *accepted for publication in BMC Proceedings*

2017 Friedrichs S, Manitz J, Burger P, Amos CI, Risch A, Chang-Claude J, Wichmann HE, Kneib T, Bickeböller H, Hofner B: **Pathway-Based Kernel Boosting for the Analysis of Genome-Wide Association Studies.** *Computational and Mathematical Methods in Medicine* 2017; vol. 2017. doi:10.1155/2017/6742763.

2017 Rosenberger A, Sohns M, Friedrichs S, Hung RJ, Fehringer G, McLaughlin J, Amos CI, Brennan P, Risch A, Brüske I, Caporaso NE, Landi MT, Christiani DC, Wei Y, Bickeböller H: **Gene-set meta-analysis of lung cancer identifies pathway related to systemic lupus erythematosus.** *PLoS One* 2017; 12(3):e0173339. doi: 10.1371/journal.pone.0173339.

2016 Malzahn D, Friedrichs S, Bickeböller H: **Comparing Strategies for Combined Testing of Rare and Common Variants in Whole Sequence and Genome-wide Genotype Data.** *BMC Proceedings* 2016; 10(Suppl 7):17. doi:10.1186/s12919-016-0042-9

2015 Friedrichs S*, Malzahn D*, Pugh EW, Almeida M, Liu XQ, Bailey JN: **Filtering genetic variants and placing informative priors based on putative biological function**; * these authors share first authorship. *BMC Genetics* 2016; 17(Suppl 2):S8. doi: 10.1186/s12863-015-0313-x.

2015 Rosenberger A, Friedrichs S, Amos CI, Brennan P, Fehringer G, Heinrich J, Hung RJ, Muley T, Müller-Nurasyid M, Risch A, Bickeböller H: **META-GSA: Combining Findings from Gene-Set Analyses across Several Genome-Wide Association Studies.** *PLoS One* 2015; 10(10):e0140179. doi: 10.1371/journal.pone.0140179.

2014 Malzahn D, Friedrichs S, Rosenberger A, Bickeböller H: **Kernel score statistic for dependent data.** *BMC Proceedings* 2014; 8(Suppl 1):S41. doi: 10.1186/1753-6561-8-S1-S41.