Original Paper

Human Heredity

Hum Hered 2013;76:64–75
DOI: 10.1159/000357567

Received: August 26, 2013
Accepted after revision: November 26, 2013
Published online: January 14, 2014

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

Saskia Freytag[a] Juliane Manitz[b] Martin Schlather[d] Thomas Kneib[b, c] Christopher I. Amos[e] Angela Risch[f, g] Jenny Chang-Claude[h] Joachim Heinrich[i] Heike Bickeböller[a, c]

[a]Institute of Genetic Epidemiology, Medical School, [b]Department of Statistics and Econometrics, and [c]Center for Statistics, Georg-August University Göttingen, Göttingen, and [d]Institute for Mathematics, University of Mannheim, Mannheim, Germany; [e]Department of Community and Family Medicine, Geisel School of Medicine, Dartmouth College, Lebanon, N.H., USA; [f]Division of Epigenomics and Cancer Risk Factors, Translational Lung Research Center Heidelberg, German Cancer Research Center, [g]Translational Lung Research Center Heidelberg, Member of the German Center for Lung Research, and [h]Division of Cancer Epidemiology, German Cancer Research Center, Heidelberg, and [i]Institute of Epidemiology, Helmholtz Center Munich, German Research Center for Environmental Health, Neuherberg, Germany

Key Words

Kernel machine test · Pathway · Network · Gene-gene interaction · Score test · Generalized linear model · Lung cancer · Rheumatoid arthritis · Disease association · Genetic association studies

Abstract

Biological pathways provide rich information and biological context on the genetic causes of complex diseases. The logistic kernel machine test integrates prior knowledge on pathways in order to analyze data from genome-wide association studies (GWAS). In this study, the kernel converts the genomic information of 2 individuals into a quantitative value reflecting their genetic similarity. With the selection of the kernel, one implicitly chooses a genetic effect model. Like many other pathway methods, none of the available kernels accounts for the topological structure of the pathway or gene-gene interaction types. However, evidence indicates that connectivity and neighborhood of genes are crucial in the context of GWAS, because genes associated with a disease often interact. Thus, we propose a novel kernel that incorporates the topology of pathways and information on interactions. Using simulation studies, we demonstrate that the proposed method maintains the type I error correctly and can be more effective in the identification of pathways associated with a disease than non-network-based methods. We apply our approach to genome-wide association case-control data on lung cancer and rheumatoid arthritis. We identify some promising new pathways associated with these diseases, which may improve our current understanding of the genetic mechanisms.

S.F. and J.M. share first co-authorship.

KARGER

E-Mail karger@karger.com
www.karger.com/hhe

0001–5652/14/0762–0064$39.50/0

Saskia Freytag
Institut für Genetische Epidemiologie, UMG
Humboltallee 32
DE–37073 Göttingen (Germany)
E-Mail saskia.freytag@med.uni-goettingen.de

Introduction

Pathway-based analysis can supplement the exploration of data from genome-wide association studies (GWAS) through the integration of prior biological knowledge [e.g. 1–4]. Primarily, the success of pathway-based analysis may be explained by its focus on jointly testing of functionally related SNPs. On the one hand, this allows the identification of pathways via multiple causal low-effect SNPs, which are usually hard to detect with conventional GWAS approaches. Pathway-based analysis also considerably reduces the multiple-testing problem. On the other hand, it has the potential to benefit directly from the knowledge on functional dependencies in the human organism [5]. Results obtained from pathway-based analysis can be interpreted in this context. This allows the easier generation of hypotheses for both diagnostic and prognostic targets [6] and can contribute to the development of novel treatment strategies.

The range of pathway-based analysis approaches is steadily expanding; for an overview of some methods see Wang et al. [7] and Varadan et al. [8]. Gene-set enrichment analysis (GSEA) [9], which was originally developed for gene expression data, has remained the most popular method. Essentially, this method consists of a nonparametric test for the enrichment of SNP-disease associations in a pathway. Like nearly all other pathway-based analysis approaches, it fails to utilize most available knowledge on pathways. In particular, it ignores information on which genes interact in the pathway. Instead, given a pathway, GSEA treats genes and their corresponding SNPs independently from each other.

There is increasing evidence that precisely such information on functional relationships among genes, i.e. the topology of the pathway, is of relevance in the context of GWAS. Several studies demonstrated that disease-causing genes often directly interact with each other as part of larger regulatory or functional systems [10, 11]. For Crohn's disease, Chen et al. [12] demonstrated that 'genes in the same neighborhood within a pathway tend to show similar association status'. In fact, it has been estimated that '80% of the currently missing heritability for Crohn's disease could be due to genetic interactions' [13]. However, not only direct interaction seems to be important. Lee et al. [14] demonstrated that SNP-trait associations are enriched in genes occupying structurally relevant positions in known pathways.

Some researchers have already recognized the potential of incorporating pathway topology, also called network, into the analysis of GWAS data. Chen et al. [12] proposed a Markov Random Field to include topological structures. Pan [15] developed a procedure that reduces the multiple-testing burden according to the average distance between genes in a pathway. Others have coined methods that aim to identify significantly associated subnetworks [16, 17]. However, all these methods are based on p values, which summarize the risk for a disease for whole genes, rather than on raw genotype data.

The integration of networks via kernels is not new, for example Rapaport et al. [18] considered one in a support vector machine analyzing microarray data. In general, kernels are the basis of many powerful statistical methods, such as support vector machines, nonparametric regression and smoothing splines. Thereby, kernels are positive semi-definite functions that reflect the pairwise similarity between observations. The use of such kernel methods rapidly gained popularity in the identification of associations between pathways and complex traits, as they are both powerful and flexible [19, 20]. Schaid [21] speculated that appropriate modifications of the kernel could also allow for the inclusion of networks in GWAS. In this light, we propose sophisticated kernels for the logistic kernel machine test (LKMT) that account for pathway topology. Here, pathway topology includes not only the network, i.e. gene-gene interactions, but also the nature of interactions, which may either constitute activation or inhibition.

We apply the LKMT with our novel network-based kernels to genome-wide case-control data on rheumatoid arthritis (RA) and lung cancer (LC). Both diseases are common in industrialized nations with an enormous social and economic impact. Moreover, generally effective cures or prevention strategies have not been discovered yet. In fact, for the United States, estimated 228,190 new LC cases will occur in 2013, making it the most common type of cancer [22]. Even though exposure to tobacco smoke determines most of the risk of developing LC, many studies also suggest genetic influences. Other than a few rare LC syndromes, only a moderate number of genetic effects, each contributing to only a weak increase in risk, are known. RA is the most common chronic joint disease and affects nearly 1% of the adult population in the United States. Many genetic factors have been firmly established as contributing to RA risk, in particular the human leukocyte antigen (HLA) region on chromosome 6 [23]. Thanks to their different genetic profiles, the study of both these diseases offers an excellent opportunity to evaluate the performance of novel statistical methods whose aim is to detect genetic associations of different strengths. Using kernels that incorporate known network structures of pathways within the LKMT has the potential

to discover previously unknown genetic risk factors. Through its focus on pathways, it also promises to elucidate disease etiology [5].

Next, we briefly outline the framework of the LKMT. We suggest a way to construct kernels that integrate information on a pathway topology obtained from the publicly available 'KEGG: Kyoto encyclopedia of genes and genomes' (KEGG) [24]. We then present simulation results demonstrating the power of our kernels compared to commonly used kernels in simple scenarios. We also demonstrate that our kernels retain the correct type I error level. Moreover, we discuss the results obtained from the analysis of the 2 GWAS on LC and RA. Using concepts from statistical network theory, we verify empirically that the integration of network information does not lead to artifacts or conceal genuine effects. Finally, we conclude with a discussion of the promising results of our research as well as possible further improvements to our method. Most of the analysis was conducted with the statistical software package R [25] unless stated otherwise.

Materials and Methods

In this section, we firstly describe the LKMT, followed by details of the network-based kernel and its construction. We secondly introduce the GWAS and pathway data used. Finally, we describe the simulations performed and the analysis of their results.

The Logistic Kernel Machine Test

The LKMT assumes a semi-parametric logistic regression model for the probability of being a case. It models genetic effects nonparametrically and environmental effects parametrically:

$$\text{logit}(P(y_i = 1)) = \mathbf{x}_i^T\boldsymbol{\beta} + h(\mathbf{z}_i), \quad (1)$$

where y_i is the case-control indicator (control: $y_i = 0$, case: $y_i = 1$) for $i = 1, ..., n$ individuals. The vector $\boldsymbol{\beta}$ represents the intercept and regression coefficient terms related to the environmental covariates $\mathbf{x}_i$ for the i-th individual ($i = 1, ..., n$). These typically include gender and other trait-relevant information, which are modeled parametrically as fixed effects. The variable $\mathbf{z}_i$ denotes the genotype vector of some selected or all SNPs, coded in the usual trinary fashion (the number of minor alleles, i.e. $z_{is} \in \{0, 1, 2\}$ for any modeled SNP s in individual i). The nonparametric function $h \in \mathrm{H}_K$ describes how the risk of being affected by the disease depends on the observed genotypes. Here, H_K denotes a reproducing kernel Hilbert space generated by a positive semi-definite and symmetric kernel K. The mathematical properties imply that any function in that space, $h \in \mathrm{H}_K$, can be approximated arbitrarily close by linear combinations of its corresponding kernel [26], i.e.

$$h(\mathbf{z}_i) = \sum_{j=1}^{n} \alpha_j K(\mathbf{z}_j, \mathbf{z}_i), \alpha_j \in \mathbf{R}.$$

The kernel $K(\mathbf{z}_i, \mathbf{z}_j)$, evaluated for individuals i and j, can also be understood as measuring the similarity between the individuals i and j based on their genotypes. Hence, by selecting a different kernel, one specifies a different concept of similarity and, implicitly, a different model for the effect of the SNPs on the risk of developing the investigated disease. One of the most commonly used kernels is the linear kernel (LIN), $K(\mathbf{z}_i, \mathbf{z}_j) = \mathbf{z}_i^T \mathbf{z}_j$, which measures the correlation between pairs of individuals. It assumes each SNP delivers a random independent and additive contribution with the same variance, in fact, specifying a linear multiple marker logistic regression. In case of a squared loss function instead of a log-likelihood, the model implied by the LIN can be shown to be equivalent to ridge regression. Note that this also highlights the close relationship to principle component methods [27]. Obviously, such a model neglects interactions among the considered SNPs [20].

On the basis of the semi-parametric logistic regression model, we test the null hypothesis (H_0) that none of the modeled SNPs is associated with the disease. We can express this mathematically by H_0: $h(\mathbf{z}_i) = 0$ for all $i = 1, ..., n$. Such a H_0 can be tested by constructing a score-type statistic. Score statistics are known from variance component tests or lack-of-fit of fixed effect models. In our case, the score-type statistic used in the LKMT is given by:

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

where $\mathbf{y} = (y_1, ..., y_n)^T$ denotes the vector of all individual case-control outcomes and $\hat{\boldsymbol{\mu}}^{(0)}$ is a vector with elements $\hat{\mu}_i^{(0)} = \text{logit}^{-1}(\mathbf{x}_i\boldsymbol{\beta})$, the maximum likelihood estimate under H_0 for the i individuals. The matrix $\mathbf{K}$ corresponds to the kernel evaluated for all combinations of individuals. Due to its quadratic form, the test statistic Q follows asymptotically an unknown mixture of χ_1^2 distributions. In order to obtain a p value for significance, this distribution is well-approximated by a moment matching method (see Wu et al. [20]). When testing many different pathways, multiple-testing corrections should be applied to the p values. In our analysis, we used the rather conservative but simple Bonferroni correction.

Construction of Network-Based Kernels

In order to accommodate network topologies of pathways, Schaid [21] proposed the kernel matrix $\mathbf{K} = \mathbf{ZSZ}^T$ for genomic information, where the matrix $\mathbf{S}$ scores the similarity of SNPs. The matrix $\mathbf{Z} = (\mathbf{z}_1, ..., \mathbf{z}_n)^T$ denotes the genotype matrix, i.e. the collection of genotype vectors $\mathbf{z}_1$, ..., $\mathbf{z}_n$ of all individuals. However, Schaid does not give a general specification of $\mathbf{S}$, but he reviews different choices for some exemplary genomic applications instead. The kernel, which we develop to take network topologies into account, is motivated by the viewpoint of a kernel as a similarity measure: SNPs located in the same gene or in interacting genes are scored to be more similar than SNPs far apart regarding the network structure. Such a notion of similarity is sometimes also referred to as 'guilt-by association' [28] and has been verified empirically for several complex diseases. More precisely, we define the matrix $\mathbf{S}$ as $\mathbf{ANA}^T$, where matrix $\mathbf{A}$ maps SNPs to genes and matrix $\mathbf{N}$ represents the network (for an illustration of the kernel's construction, see also fig. 1). Altogether, the kernel matrix is defined as $\mathbf{K} = \mathbf{ZANA}^T\mathbf{Z}^T$. Here, the genotype matrix $\mathbf{Z}$ is allowed to contain missing values making imputation necessary.

The elements $a_{sg} \in \{0, 1\}$ of matrix $\mathbf{A}$ represent the membership of SNP s in gene g. Most commonly, SNPs are assigned to genes purely on the basis of their location on the genome, but other annotations are conceivable [7]. In the 2 real GWAS, we assign a SNP to a gene, when it is directly located in a gene or in the 500-kbp

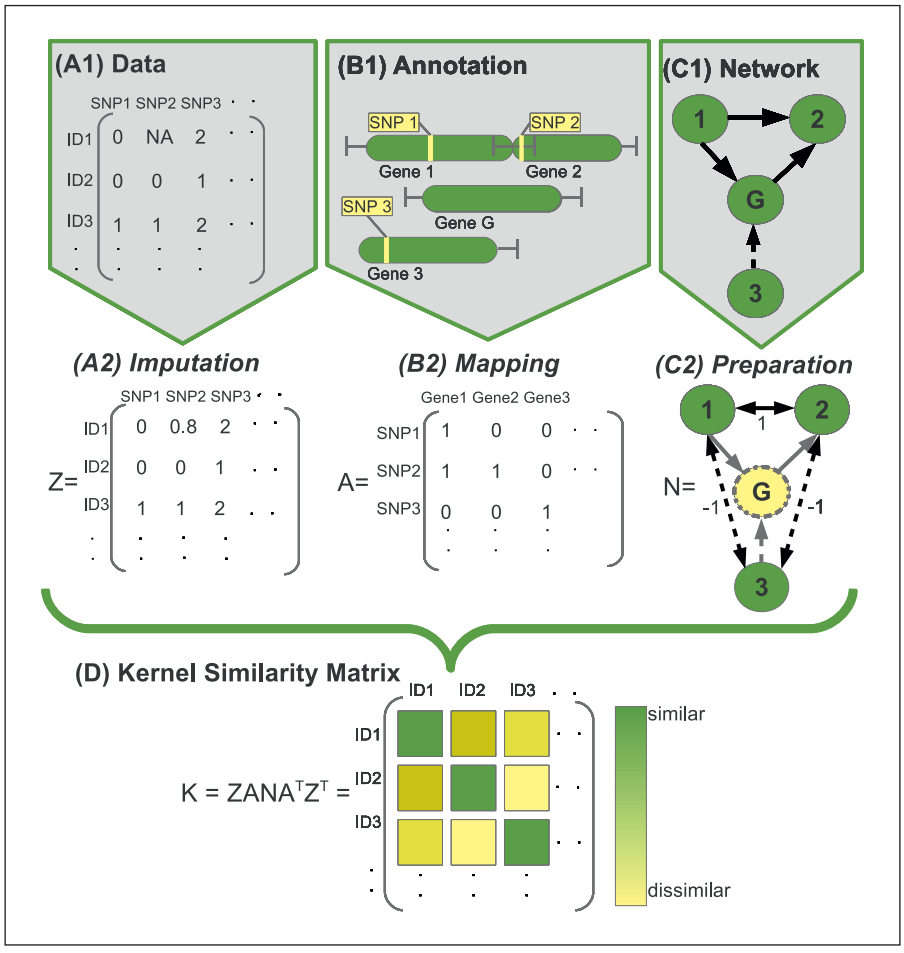

Fig. 1. Pipeline of the construction of the network-based kernel matrix $\mathbf{K} = \mathbf{ZANA}^T\mathbf{Z}^T$. (**A1**) Genotype data (SNP No.) coded in trinary fashion for cases and controls (ID No.) are presented in a matrix. (**B1**) SNP-gene annotation is mapping all SNPs to pathway genes, as long as they are located in the gene or in the 500-kbp windows around the gene. (**C1**) The pathway network with activating (solid arrows) and inhibiting (dashed arrow) interactions between genes. (**A2**) Imputation of missing genotype values via BEAGLE [33] and deletion of SNPs that cannot be mapped to a pathway resulting in genotype matrix **Z**. (**B2**) Representation of the SNP-gene annotation as matrix (**A**), where 1 indicates membership. (**C2**) The network structure is modified so that genes without any genotyped SNPs (yellow node) and their corresponding links (grey arrows) are deleted, but their directed interactions with their next neighbors are retained (black arrows); the network structure is then converted to an undirected adjacency matrix **N**, where 1 represents activation and –1 represents inhibition. (**D**) Calculation of the network-based kernel similarity matrix by $\mathbf{K} = \mathbf{ZANA}^T\mathbf{Z}^T$.

windows on either side. Note that, a SNP can be assigned to more than one gene due to some overlap of genes. Further, we adjust for different gene sizes by re-weighting the impact of a gene effect. This ensures an equal treatment despite different numbers of genotyped SNPs in the genes. We denote the modified **A** by **A*** with elements

$$a^*_{sg} = \frac{a_{sg}}{\sqrt{r_g}},$$

where r_g equals the number of SNPs in gene g. In the following, we refer to network-based kernels using the unadjusted gene-SNP annotation as NET and ANET under utilization of the size-adjusted gene-SNP matrix **A***.

The matrix **N** denotes the quadratic adjacency matrix of the neighborhood structure of the genes in the pathway. Its dimension equals the number of genes in the pathway. We consider selfinteractions, i.e. that every gene interacts with itself, by setting all diagonal elements of matrix **N** = 1. Unlike other network-based methods, we distinguish between activating and inhibiting gene-gene interactions. Thus, an element $n_{g,g'}$ of **N** equals 1 or –1, if genes g and g' interact in an activating or inhibiting fashion, respectively. In the following, we refer to the use of adjacency matrices that distinguish between inhibition and activation as *signed* and networks with unspecified interaction types as *unsigned*.

This basic network structure must be further modified to ensure a well-defined kernel, which should be complete, symmetric and positive semi-definite. Firstly, to ensure completeness of the pathway topology, we rewire certain interactions, which are associated to genes without genotyped SNPs. During mapping computation, $\mathbf{S} = \mathbf{ANA}^T$ such genes and their interactions would be removed from the analysis automatically. Firstly, to preserve full information on interactions in the pathway, we project links of genes without genotyped SNPs to their immediate neighbors. This means, we include additional links, where earlier 2 interactions existed and which would otherwise have been removed entirely. Thereby, the link sign of the newly created interaction is determined in a multiplicative fashion, e.g. the combination of a former inhibition and activation results in a new inhibition. Secondly, we transform the directed pathway structure into an undirected network via mirroring along the diagonal. Finally, kernels are required to be positive semi-definite, while undirected adjacency matrices **N** are not necessarily positive semi-definite. Thus, we introduce a new procedure to find the closest matrix **N*** by superimposing as much noise as necessary to render the new matrix positive semi-definite without introducing additional interactions to the network. If **N** is not positive semi-definite, we replace the original matrix **N** in the kernel equation by the weighted $\mathbf{N}^* = \rho\mathbf{N} + (1 + \rho)\mathbf{I}$, where **I** is the identity matrix. It can be

easily verified that $\mathbf{N}^*$ is a positive semi-definite matrix if $\rho \in [0, \rho_{max}]$, where

$$\rho_{max} = \frac{1}{1 - \lambda_{min}} \quad (3)$$

and λ_{min} is the smallest eigenvalue of $\mathbf{N}$. Our approach of approximating the symmetric matrix $\mathbf{N}$ by a positive semi-definite one has the advantage that the original network topology is exactly preserved although the link weights are eased. It also allows for an interpretation of the identity matrix as a noise component. We suggest using $\rho = \rho_{max}$ since $\mathbf{N}^*$ is the closest to the original matrix $\mathbf{N}$ but is positive semi-definite and has the minimum eigenvalue zero. We also tested normalized and ordinary Laplacian matrices [29] as well as an algorithm by Higham [30] to find the nearest positive semi-definite approximation of the network matrix, but found them to have inferior performances (data not shown) when compared to $\mathbf{N}$ and its replacement described above. Moreover, the alternative methods change the network topology by including additional interactions, while our method preserves the structure of network.

Data

RA and LC GWAS

The German Lung Cancer Study (GLCS) examines the role of genetic polymorphisms on the risk of developing LC at a relatively early age, specifically LC diagnosed prior to the age of 50 years [31]. Cases for this study, which comprise both small-cell LC as well as non-small-cell LC, were sampled from 31 German hospitals, while controls are from the KORA epidemiological survey of individuals living near the southern German city of Augsburg. The second study, which was conducted by the North American Rheumatoid Arthritis Consortium (NARAC), aims to identify genetic risk factors for RA [32]. Thereby, the criterion of being a RA case was set by the American College of Rheumatology and cases were procured from New York hospitals. Informed consent was obtained from all participants of both studies; the studies were conducted according to the Declaration of Helsinki.

We applied stringent quality control (QC) measures, notably the exclusion of possibly related individuals. Furthermore, SNPs with a call rate <90% were eliminated. For all remaining SNPs, missing genotypes were imputed using the standard software BEAGLE [33]. The number of cases, controls and genotyped SNPs can be found in table 1. Since some SNPs could not be assigned to any genes, not all genotyped SNPs were used in the analysis. We included sex as an additional environmental covariate. In GLCS, we also considered age at LC diagnosis (cases) or exam (controls) and the cigarette consumption in pack-years, i.e. the number of cigarettes smoked per day multiplied by the years of exposure through active smoking.

While participants in the LC study are fairly homogeneous with regards to ethnicity, the ancestries of the participants in the RA study ranged from Northern to Southern European. Despite this, we did not correct explicitly for population stratification in either study. There is cumulative evidence that multiple marker methods used in high-dimensional settings inherently capture cryptic relatedness, rendering additional corrections obsolete [34, 35]. For multiple regression models which do not include population structure explicitly, Setakis et al. [36] were able to demonstrate their robustness for population stratification effects via simulation studies. Thus, it stands to reason that additional correction for population stratification in the LKMT, which is similar to such a model, would lead to overcorrection and in turn loss of power.

Table 1. Number of individuals, SNPs and genes in the 2 GWAS of LC and RA

		GLCS	NARAC
Cases	Before QC	506	868
	After QC	467	866
	Males	286	226
	Females	181	640
Controls	Before QC	480	1,194
	After QC	468	1,189
	Males	237	341
	Females	231	848
SNPs	Before QC	561,466	545,080
	After QC	529,637	492,209
	In the analysis	255,241	243,096
Genes	In the analysis	2,808	2,807

Table 2. Network characteristics for the 182 investigated pathways

Network characteristic	Mean	Median	Range
Dimension	22.85	14.00	[2.00, 316.00]
Density	0.24	0.16	[0.00, 1.00]
Average degree	4.22	2.00	[0.00, 303.19]
Inhibition degree	0.14	0.00	[0.00, 3.07]
Diameter	3.57	3.00	[0.00, 15.00]
Transitivity	0.50	0.50	[0.00, 1.00]
Signed transitivity	0.32	0.31	[–0.20, 1.00]

Besides applying the LKMT with our network-based kernels and the LIN kernel, we analyzed both data sets using GSEA. Unlike the LKMT, GSEA tests competitive hypotheses, i.e. whether a particular pathway tends to be more associated with the disease than all other investigated pathways. As a direct result of this fundamental difference between the LKMT and GSEA, comparisons of their results are of particular interest. Here, we use the publicly available GenGen software [9] to implement GSEA.

Pathway Data

We decided to use the popular database KEGG due to its manual curation. Moreover, it offers a selected range of pathways including experimentally verified metabolic pathways, information and cellular processing pathways as well as those related to organismal system information and human diseases. We did not access KEGG directly, but extracted the adjacency matrices by means of the R package rBioPaxParser [37], which allows the use of the standardized Biological Pathway Exchange (BioPAX) language. Viswanathan et al. [38] called BioPAX the 'currently [...] best-suited format for mathematical modeling and simulations'. Our analysis in-

cluded the topology of 182 pathways, which have sufficient network information. After preparation, 38 adjacency matrices **N** were already positive semi-definite. For the remaining networks, we found the closest positive semi-definite counterpart with the aforementioned procedure (ρ_{max} computed by equ. 3 has a mean value of 0.48).

We found the structures of the different networks to be very diverse, which is supported by common descriptive network statistics (see table 2). We considered:

Dimension counting the total number of genes in the pathway.
Density denoting the ratio of existing interactions to the possible number of interactions in a fully connected pathway.
Average degree referring to the mean number of interactions from or to a gene.
Diameter measuring the maximum length of the shortest path between all pairwise combinations of genes.
Transitivity denoting the probability of triangles, i.e. the interaction between 2 neighbors of a gene.

For transitivity and degree, we also distinguished between signed and unsigned networks. In the case of average degree, we also looked at the average degree of inhibitions only. Its low mean highlights that there are only very few inhibiting interactions in the data base. Furthermore, we used the extension of transitivity introduced by Kunegis et al. [39], which is able to take the interaction type into account. In general, examination of the means and medians of all descriptive statistics revealed strongly left-skewed distributions for all introduced network characteristics (see fig. 2).

Simulation Study

To evaluate the performance of the LKMT with our network-based kernels, we studied empirical type I error rates and power in different genetic settings. Note that null simulations for testing the type I error rate are equivalent to the scenarios for testing power without genetic effects. Empirical power or empirical type I error rates are determined as the proportion of simulations for which a p value < the ordinary 0.05 threshold is obtained. Ideally, the empirical type I error rate should be exactly 0.05, while conservative approaches are acceptable; whereas power should be as high as possible. We compared type I error rates and power of the LKMT with our network-based kernels (NET) with the performance of the LKMT with the LIN kernel and the minimum p value approach (minP). In the latter method, the minimum p value from single-marker tests applied to every SNP in the pathway represents the association of the entire pathway. Since larger pathways are more likely to generate low p values by random chance [7], we used a conservative Bonferroni correction to adjust the obtained p value by the size of the simulated pathway.

A comprehensive pathway disease model that explains how interactions between genes with susceptibility variants lead to the development of a disease connecting biological and statistical thinking has not been developed so far. Even if such a model were to exist, its necessary complexity would render it extremely challenging to simulate. Our network-based kernels have been developed with such a degree of complexity in mind, but we use a simpler simulation model. This model meets many assumptions of the LKMT with the LIN kernel, and therefore we expect the LIN kernel to be favored. Roughly, our method of simulation can be divided into 4 parts:

(1) Choosing the genetic setting with respect to a known network structure and corresponding genetic effects.
(2) Simulating genetic variants and corresponding case-control status for all individuals.
(3) Creating a structure of a pathway by mapping genetic variants to 'genes' and 'genes' to 'pathways'.
(4) Applying the pathway analysis approaches to the simulated data.

As pathways we choose to investigate network structures of 2 real KEGG pathways; *path:hsa04950* with 22 genes and *path:hsa05218* with 9 genes (compare fig. 3). Values of dimension, density, average degree, and average negative degree of *path:hsa04950* are very close to the mean values of these network characteristics obtained from all investigated KEGG pathways. In contrast, the network characteristics of *path:hsa05218* are more extreme compared to the KEGG pathway averages. In order to examine empirical power, we simulated 2 different genetic settings each at different strengths. In the 'connected' setting, 3 'genes', each of which contains 3 causal genetic variants, were selected in a way that they directly interact in the network. In the 'apart' setting, 3 'genes', each including 3 causal genetic variants, were far away from each other with respect to the given network structures (see fig. 3). We expected our network-based kernel to perform better in the 'connected' setting than in the 'apart' setting, as our network-based kernel was developed with the aim of exploiting connections explicitly. In both settings, detection should be aided by the presence of strong linkage disequilibrium (LD) between causal genetic variants and simulated noncausal variants (compare online suppl. fig. S1; for all online suppl. material, see www.karger.com/doi/10.1159/000357567; Barrett et al. [40]). The effect strength was varied by increasing heterozygous risk from 1.05 to 1.20 and the homozygous risk accordingly from 1.10 to 1.40 for each causal variant.

Given the causal variants and their effect sizes, we simulated genetic variants and corresponding case-control status for 1,000 individuals using the HAPGEN2 [41] and the CEU sample of the International HapMap Project [42]. HAPGEN2 is considered to mimic real genetic studies due to its reliance on reference populations and observed fine-scale recombination rates. Thus, it preserves natural LD structures in the human genome. We simulated 1,100 genetic variants in the region between 1,054 and 11,657 kbp on chromosome 1 for 500 cases and 500 controls. For each scenario, we repeated the simulations 1,000 times. Note that we did not use the pathway topology directly when simulating data.

To apply our network-based kernel, we require genetic variants to be assigned to genes, which are in turn mapped to a network topology. Since for reasons of feasibility we simulate genetic variants in one genomic region, we worked with local regions acting as substitutes for real genes. We selected 22 or 9 local regions each with 50 genetic variants separated by 500 kbp to prevent LD between 'genes'. By restricting our analysis to same size 'genes', there was no difference between results obtained with either the NET or the ANET kernel. In the situation of equally sized genes, the adjustment for ANET reduces to a constant scale factor, which vanishes during the moment matching procedure.

Finally, we could apply all 3 investigated methods to the different simulations. For the LKMT with the NET kernel, we utilized the signed as well as unsigned versions of the pathways. Note that only the NET kernel uses the created structure of the pathway. Neither the LIN kernel nor the minP approach even takes into account which genetic variants belong to the same 'gene'.

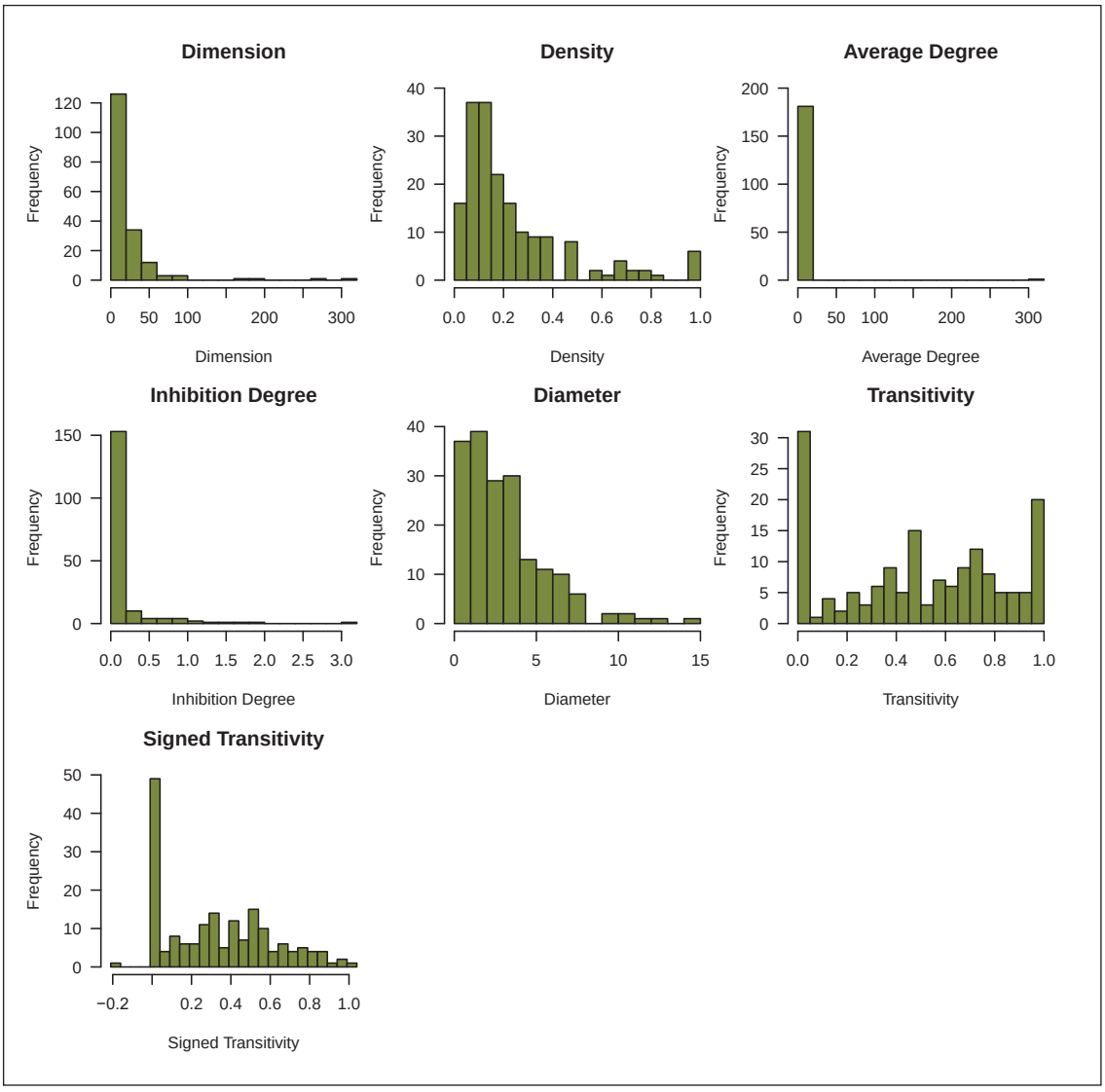

Fig. 2. Histograms for all network properties of the 182 KEGG pathways. The network characteristics include dimension, density, average degree, inhibition degree, diameter, transitivity, and signed transitivity.

Results

Simulation Study

We demonstrate here that the type I error rate is maintained for the LKMT with both the LIN and NET kernel as well as the minP approach in all studied genetic settings (see table 3). Of all investigated pathway analysis approaches, minP is the most conservative possibly due to the utilization of the Bonferroni correction. The type I error rate for all methods was closer to the expected level for the pathway with only 9 genes. Even so, if we were to simulate larger pathways we would observe size bias for the LIN kernel. Size bias refers to the inflation of the type I error rate with increasing numbers of SNPs contained in the pathways. This phenomenon was demonstrated conclusively for the LKMT with the LIN kernel via a simulation study by Freytag et al. [43].

Power simulations indicate that the LKMT with our network-based kernels is indeed superior in performance compared to other pathway analysis approaches for some genetic settings (see fig. 4). In particular, the NET kernel has up to 10% more power than the LIN kernel in the

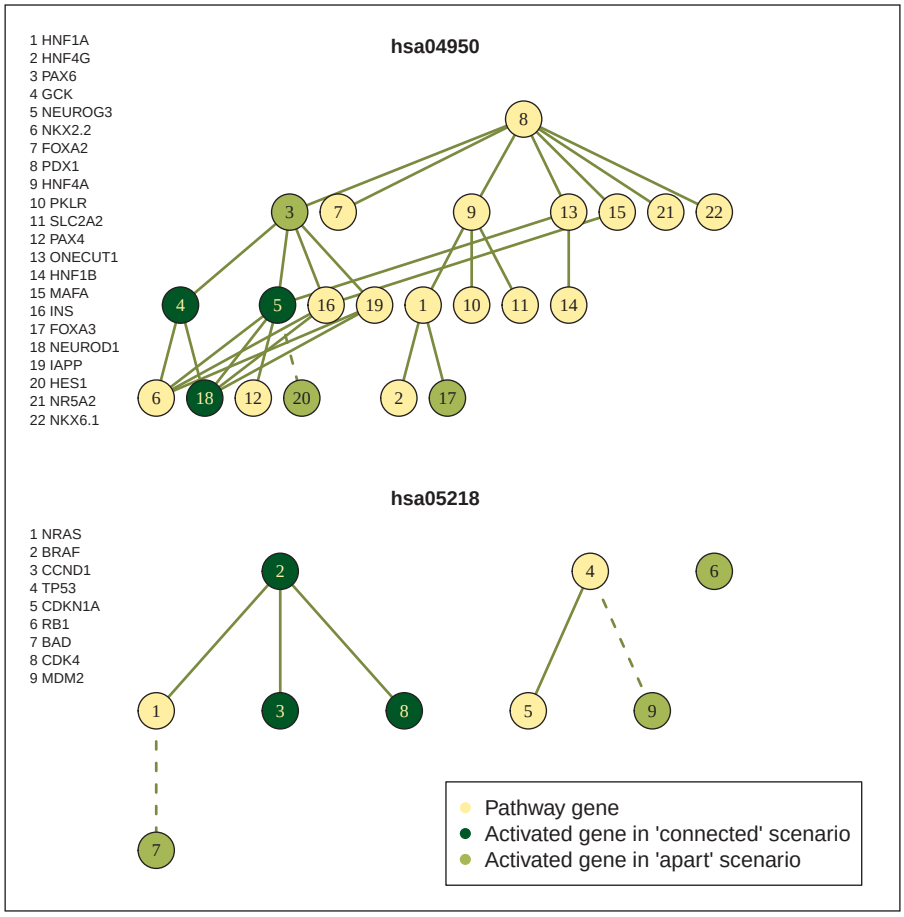

Fig. 3. Pathway network examples: 'maturity onset diabetes of the young' pathway *(path:hsa04950)* and 'melanoma skin cancer' pathway *(path:hsa05218)*. The corresponding HUGO gene identifiers for each node are given in the box at the bottom. Solid lines correspond to activations and dashed lines to inhibitions. The 'connected' scenario refers to the simulations where genes with causal SNPs are close to each other, while in the 'apart' scenario the genes with causal SNPs are far apart.

Table 3. Type I error rates for null simulations differentiated by the tested pathways

Method	Inhibition	Estimated type I error rate	
		path:hsa04950 (1,100 SNPs)	*path:hsa05281* (450 SNPs)
NET	Not considered	0.039	0.050
NET	Considered	0.042	0.050
LIN	–	0.049	0.048
minP	–	0.019	0.023

The type I error rate is based on 1,000 null simulations each with 500 cases and 500 controls.

'connected' setting. However, if the causal variants are distributed more randomly with respect to the network, the LIN kernel does generally better than the NET kernel. Even though, for lower risk the differences between the LIN and NET kernel in the 'apart' setting are not as pronounced. The minP approach was inferior to all other methods for both simulated pathways. Generally, all methods have uniformly higher power for the smaller simulated pathway. Furthermore, differences in power between the signed and unsigned version of the NET kernel existed only for the larger pathway. The equivalence of the signed and unsigned version in the small pathway probably stems from the fact that it only contains one inhibition. Given the simplicity of our simulation study, which favors the LIN kernel, our network-based kernels (NET) performed very well.

Application

GWAS Findings

Previous GWAS revealed many associations for RA, but they detected only a few for LC [23, 31]. The results from our analysis of the RA and LC GWAS confirm these observations. The LKMT with the signed ANET, unsigned ANET, signed NET and unsigned NET detects 26, 27, 25 and 26 pathways, respectively, to be significantly associ-

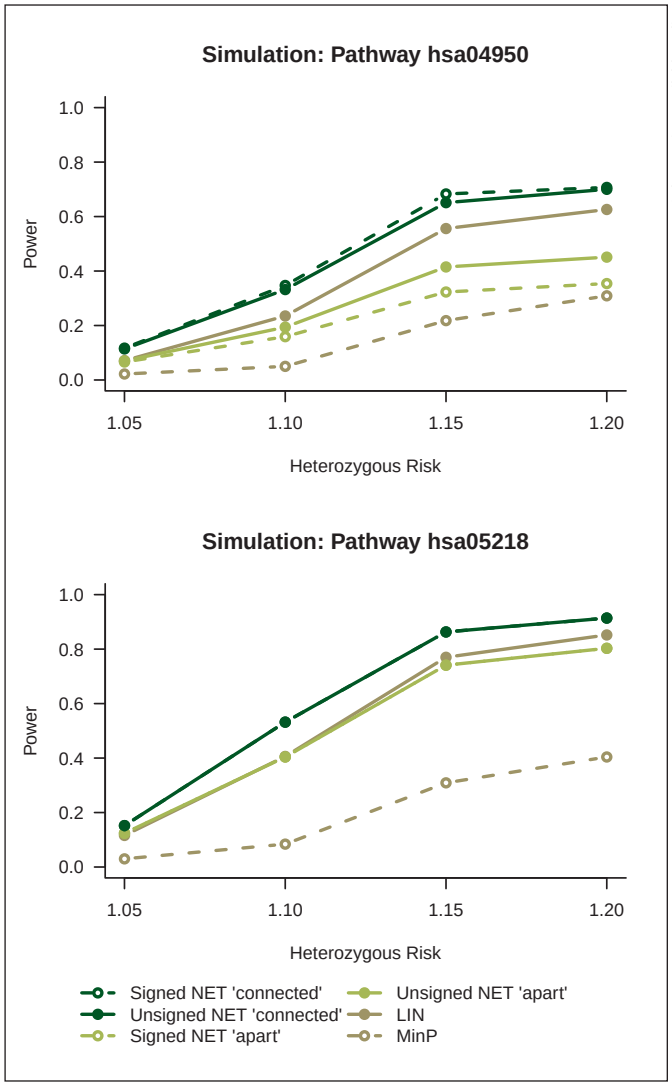

Fig. 4. Results from power simulations. The power in the 'connected' and 'apart' scenario of the network-based kernels is plotted against the heterozygous risk common for all causal SNPs. The results are shown for 2 different network topologies *(path:hsa04950* and *path:hsa05218)*. Note that the results for the signed and unsigned network-based kernel are identical in the second pathway.

ated with RA. In contrast, we were unable to detect any significant pathway associations for LC. Another possible explanation for the lack of significant LC associations could also lie in the small sample size of the GLCS GWAS.

Similar to previous studies on LC, we cannot find any significant pathways either. Thus, we rank the pathways according to their p values in order to capture potential important effects on the disease. The top 5 ranked pathways are largely similar for the different network-based kernels. As an example, we depict the results for signed ANET in online suppl. table S1 as this is the most sophisticated version of our kernels. The smallest p value belongs to the pyruvate pathway *(path:hsa00620)*. The pyruvate pathway converts glucose to pyruvate, which supplies energy to living cells when oxygen is present. When oxygen is lacking, it converts pyruvate to lactate. In cancer cells, this second process takes place regardless of the presence of oxygen, otherwise known as the Warburg effect [44]. Today, the Warburg effect is recognized as one of the important characteristics of cancer-causing mutations.

For RA, most of the identified susceptibility pathways contain genes which have been shown to be associated with the development and progression of RA in at least one scientific publication (for significant results of signed ANET, see online suppl. table S2). Furthermore, genes located in the HLA region were present in the majority of identified pathways. The results obtained using different network-based kernels hardly differ. Results between the signed and the unsigned version only differ by one pathway for the adjusted and unadjusted versions of the network-based kernel probably owing to the lack of inhibitions in the investigated pathways. Interestingly, there are 2 pathways identified by the signed ANET but not by signed NET, and 1 vice versa. This indicates that differences in the weighting of genes can alter the results. For all network-based kernels, the steroid hormone biosynthesis pathway *(path:hsa00140)* is among the pathways with the smallest p values. Steroids are known to influence the immune system heavily. They can, in fact, reduce inflammation, which is the reason that they are still sometimes used in RA treatment. Moreover, we identify 1 novel association with the drug metabolism pathway *path:hsa00983*. This pathway is responsible for processing drugs involved in the inhibition of DNA replication, such as fluorouracil and azathioprine. Interestingly, azathioprine is widely used as an immunosuppressive in the treatment of chronic inflammatory diseases, such as RA. Its efficacy in this area is attributed to its role in the control of T cell apoptosis by modulation of RAC1 activation upon CD28 co-stimulation [45].

Comparison of the Results by Different Pathway-Based Methods

In addition to our novel signed ANET kernel, we also applied the established GSEA approach and the LKMT with the simpler LIN kernel. For LC, none of the methods detected any significant pathway association. In contrast, the number of identified RA susceptibility pathways differed greatly, but they had a large common subset. The

Table 4. Correlations of network characteristics and p values for the investigated GWAS

Network characteristics	LC GWAS			RA GWAS		
	LIN	ANET	NET	LIN	ANET	NET
Dimension	0.13	–0.11	–0.12	**–0.58**	**–0.33**	**–0.29**
Density	–0.11	0.00	–0.01	**0.38**	**0.32**	**0.28**
Average degree	0.02	**–0.16**	**–0.17**	**–0.23**	–0.05	–0.04
Inhibition degree	0.13	0.06	0.06	**–0.28**	**–0.19**	**–0.17**
Diameter	0.05	–0.11	–0.12	**–0.36**	**–0.25**	**–0.23**
Transitivity	0.04	–0.10	–0.15	–0.07	0.07	0.06
Signed transitivity	0.05	**–0.15**	**–0.12**	**–0.19**	0.00	0.01

The values are nonlinear correlation coefficients (according to Kendall). Values in bold indicate a correlation that substantially differs from zero.

conventional GSEA approach identified only 14 pathways with significant effects, possibly due to the comparative nature of the hypothesis. All of them were detected as well with the LKMT using the signed ANET kernel, which found 26 associated pathways. This might indicate a higher sensitivity of the LKMT with network-based kernels. Instead, the results obtained by using the LIN kernel were less specific, as 130 pathways were determined to be associated with RA. This large proportion of significant results seems to be unlikely. Instead, we believe that size bias in combination with the HLA region is responsible for this oversensitivity. Thus, in our applications the LKMT with the network-based kernel was powerful, generated reasonable results and thus represents the happy medium between sensitivity and specificity.

We also examined the p values of the different methods. We observed that the distribution of the LIN kernel results seem to be anomalously extreme. In contrast, the p value distribution obtained with our network-based kernel, which was fairly close to the one of GSEA, did not exhibit any such anomalies (see online suppl. fig. S2).

Impact of Network Characteristics

Associations of LKMT results and network topology indicate that the effects of the genotypes are concealed by the effects generated by network structures. Thus, we correlated network structure with obtained p values according to Kendall's rank coefficients (see table 4). Here, network topology is described by various network characteristics ranging from the average degree to clustering coefficient.

Apparently, there is some correlation in the RA GWAS between the p values and properties of underlying networks, whereas the LC GWAS results reveal quite low degrees of correlations. We observed correlations between RA p values and pathway dimension for all kernels. This indicates the aforementioned presence of size bias. However, the bias is strongly reduced for our network-based kernels. We believe that further investigations of this issue will lead to better size corrections. Density, which measures the connectivity of the network, also seems to influence the magnitude of the p values. Since this influence is even higher for the LIN kernel, which does not incorporate network information, we assume some spurious correlation. The effective size of the pathway is reflected by the diameter; its correlation therefore depends on size as well as the degree of connectivity. The inhibition degree displays negative correlations, but these were even stronger for the LIN kernel, so that we again assume some spurious correlation. We cannot notice any effect for the extent of clustering in the pathways which is quantified by (signed) transitivity. Altogether, the differences between networks with regard to their non-disease-causing characteristics do not seem to introduce bias.

Discussion

The topology of pathways contains information relevant to our understanding of the functional connections between biological pathways and complex disease progression and development. We developed a network-based kernel for the logistic kernel machine to make use of pathway information when analyzing GWAS. Altogether, this presents a sophisticated and elegant statistical framework, which allows the seamless integration of additional knowledge on biological mechanisms. We demonstrated that our procedure maintains the correct type I error rate and often has more power to detect genuine

associations than 2 conventional pathway analysis methods.

The applications to case-control studies for LC and RA demonstrated the ease of implementation and efficiency of our method. Furthermore, the disease studies revealed its ability to generate plausible results under extremely different genetic profiles. For LC, the most promising result, though not significant, was the suggestion of a relationship with pyruvate metabolism. An immunohistochemical analysis conducted by Koukourakis et al. [44] provided evidence that the pyruvate pathway is repressed in 73% of non-small-cell lung carcinoma. Therefore, it is possible that the attempt to replicate our results in a bigger study may well shed further light on the question as to whether there exists a genuine genetic association or not. In the case of RA, several promising pathways, most of them involving the HLA region, were identified using our network-based procedure. Besides the pathway for drug deactivation, the notch-signaling pathway is of considerable interest in finding the cause of RA. Notch signaling may be responsible for further exacerbating the inflammatory response and joint destruction in RA patients through the formation of dysfunctional microvessels in the papillary dermis of the skin [46].

Currently, there is little knowledge of how the increased occurrence of genetic variation in a pathway affects the functionality of the human system. This lack of a reasonable biological effects model not only severely hampers method development, but it also makes informative simulation studies impossible. For our new kernel in particular, it would be of tremendous interest to investigate power using meaningful pathway-disease scenarios. Since such simulation scenarios would feature interactions between causal variants, we are confident that our network-based kernels would then be by far superior in comparison with commonly used kernels. Such kernels, in particular the LIN kernel, typically assume linearity of effects and thus fail under such conditions. Furthermore, these simulation models would allow us to investigate the effect of incorrectly specified networks. We expect that the network-based kernels can handle some missing links with some power decrease. In the application, we already demonstrated that our approach found a happy medium between sensitivity and specificity, even though the used pathway data are known to be incomplete. Thus, given the extent of our knowledge we will have to rely on the good performance of our kernels in the 2 applications as well as the greatly simplified simulation study.

Our method constitutes a promising foundation for further advances in network-based analysis of GWAS data. In particular, the procedure to generate positive semi-definite network matrices, which can include negative interactions, may find applications in diverse fields of research. As one area of improvement, we see the inclusion of interaction directionality between genes. An adjacency matrix also tracking the direction of the interaction would no longer be symmetric, thus violating the requirement of positive semi-definite kernels. The restriction to undirected adjacency matrices is a common simplification but presents a considerable loss of information. Another improvement would lie in the explicit consideration of link uncertainty via incorporating link prediction approaches or Bayesian methods in the construction of the kernel.

More importantly, the inaccurate and incomplete nature of regulatory models remains the biggest challenge to network-based analysis. Collaborative research by laboratories and institutes has improved our understanding of biological processes greatly, but much work still remains to be done. The true value of network-based methods will only be realized, when network models leverage additional information particular to the investigated disease [5]. In particular, models should account for the cell-specific context and the dynamic nature of the regulation of biological mechanisms dependent on time [47].

Acknowledgements

This work was supported by the Deutsche Forschungsgemeinschaft (DFG) research training group 'Scaling Problems in Statistics' (RTG 1644). The rheumatoid arthritis study analyzed in this article was generously made available by Peter Gregerson and the National Institute of Health [AR44422]. The lung cancer data were kindly provided through TRICL grant U19CA148127.

Disclosure Statement

The authors declared no conflicts of interest.

References

1 Kar SP, Seldin MF, Chen W, Lu E, Hirschfield GM, Invernizzi P, Heathcote J, Cusi D; the Italian PBC Genetics Study Group; Gershwin ME, Siminovitch KA, Amos CI: Pathway-based analysis of primary biliary cirrhosis genome-wide association studies. Genes Immun 2013;14:179–186.

2 Chen QR, Braun R, Hu Y, Yan C, Brunt EM, Meerzaman D, Sanyal AJ, Buetow K: Multi-SNP analysis of GWAS data identifies pathways associated with nonalcoholic fatty liver disease. PLoS One 2013;8:e65982.

3 Chuang LC, Kao CF, Shih WL, Kuo PH: Pathway analysis using information from allele-specific gene methylation in genome-wide association studies for bipolar disorder. PLoS One 2013;8:e53092.
4 Song GG, Choi SJ, Ji JD, Lee YH: Genome-wide pathway analysis of a genome-wide association study on multiple sclerosis. Mol Biol Rep 2013;40:2557–2564.
5 Califano A, Butte AJ, Friend S, Ideker T, Schadt E: Leveraging models of cell regulation and GWAS data in integrative network-based association studies. Nat Genet 2012;44:841–847.
6 Schadt EE: Molecular networks as sensors and drivers of common human diseases. Nature 2009;461:218–223.
7 Wang K, Li M, Hakonarson H: Analysing biological pathways in genome-wide association studies. Nat Rev Genet 2010;11:843–854.
8 Varadan V, Mittal P, Vaske CJ, Benz SC: The integration of biological pathway knowledge in cancer genomics: a review of existing computational approaches. IEEE Signal Process Mag 2012;29:35–50.
9 Wang K, Li M, Bucan M: Pathway-based approaches for analysis of genomewide association studies. Am J Hum Genet 2007;81:1278–1283.
10 Lin J, Gan CM, Zhang X, Jones S, Sjöblom T, Wood LD, Parsons DW, Papadopoulos N, Kinzler KW, Vogelstein B, Parmigiani G, Velculescu VE: A multidimensional analysis of genes mutated in breast and colorectal cancers. Genome Res 2007;17:1304–1318.
11 Lim J, Hao T, Shaw C, Patel AJ, Szabó G, Rual JF, Fisk CJ, Li N, Smolyar A, Hill DE, Barabási AL, Vidal M, Zoghbi HY: A protein-protein interaction network for human inherited ataxias and disorders of Purkinje cell degeneration. Cell 2006;125:801–814.
12 Chen M, Cho J, Zhao H: Incorporating biological pathways via a Markov random field model in genome-wide association studies. PLoS Genet 2011;7:e1001353.
13 Zuk O, Hechter E, Sunyaev SR, Lander ES: The mystery of missing heritability: genetic interactions create phantom heritability. Proc Natl Sci USA 2012;109:1193–1198.
14 Lee Y, Li H, Li J, Rebman E, Achour I, Regan KE, Gamazon ER, Chen JL, Yang XH, Cox NJ, Lussier YA: Network models of genome-wide association studies uncover the topological centrality of protein interactions in complex diseases. J Am Med Inform Assoc 2013;20:619–629.
15 Pan W: Network-based model weighting to detect multiple loci influencing complex diseases. Hum Genet 2008;124:225–234.
16 Schaid DJ, Sinnwell JP, Jenkins GD, McDonnell SK, Ingle JN, Kubo M, Goss PE, Costantino JP, Wickerham DL, Weinshilboum RM: Using the gene ontology to scan multilevel gene sets for associations in genome wide association studies. Genet Epidemiol 2012;36:3–16.
17 International Multiple Sclerosis Genetics Consortium: Network-based multiple sclerosis pathway analysis with GWAS data from 15,000 cases and 30,000 controls. Am J Hum Genet 2013;92:845–865.
18 Rapaport F, Zinovyev A, Dutreix M, Barillot E, Vert JP: Classification of microarray data using gene networks. BMC Bioinformatics 2007;8:35.
19 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.
20 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;86:929–942.
21 Schaid DJ: Genomic dimilarity and kernel methods II: methods for genomic information. Hum Hered 2010;70:132–140.
22 National Cancer Institute. Lung Cancer. http://www.cancer.gov/cancertopics/types/lung (last access: June 10, 2013).
23 Raychaudhuri S: Recent advances in the genetics of rheumatoid arthritis. Curr Opin Rheumatol 2010;22:109–118.
24 Ogata H, Goto S, Sato K, Fujibuchi W, Bono H, Kanehisa M: KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res 1999;27:29–34.
25 R Core Team: R: A Language and Environment for Statistical Computing. Vienna, R Foundation for Statistical Computing, 2013. ISBN 3–900051–07–0.
26 Hofmann T, Schölkopf B, Smola AJ: Kernel methods in machine learning. Ann Stat 2008;36:1171–1220.
27 Hastie T, Tibshirani R, Friedman JJH: The Elements of Statistical Learning. New York, Springer, 2001, vol 1.
28 Kolaczyk ED: Statistical Analysis of Network Data: Methods and Models. Springer Series in Statistics. New York, Springer, 2009.
29 Smola AJ, Kondor R: Kernels and regularization on graphs; in Schölkopf B, Warmuth MK (eds): Learning Theory and Kernel Machines. Lecture Notes in Artificial Intelligence, vol 2777. Berlin, Springer, 2003, pp 144–158.
30 Higham NJ: Computing the nearest correlation matrix – a problem from finance. IMA J Numer Anal 2002;22:329–343.
31 Sauter W, Rosenberger A, Beckmann L, et al: Matrix metalloproteinase 1 (MMP1) is associated with early-onset lung cancer. Cancer Epidemiol Biomarkers Prev 2008;17:1127–1135.
32 Amos CI, Chen WV, Seldin MF, Remmers EF, Taylor KE, Criswell LA, Lee AT, Plenge RM, Kastner DL, Gregersen PK: Data for Genetic Analysis Workshop 16 Problem 1, association analysis of rheumatoid arthritis data. BMC Proc 2009;3(suppl 7):S2.
33 Browning BL, Browning SR: A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. Am J Hum Genet 2009;84:210–223.
34 Kärkkäinen HP, Sillanpää MJ: Robustness of Bayesian multilocus association models to cryptic relatedness. Ann Hum Genet 2012;76:510–523.
35 Habier D, Fernando R, Dekkers J: The impact of genetic relationship information on genome-assisted breeding values. Genetics 2007;177:2389–2397.
36 Setakis E, Stirnadel H, Balding DJ: Logistic regression protects against population structure in genetic association studies. Genome Res 2005;16:290–296.
37 Kramer F, Bayerlová M, Klemm F, Bleckmann A, Beissbarth T: rBiopaxParser – an R package to parse, modify and visualize BioPAX data. Bioinformatics 2012;29:520–522.
38 Viswanathan GA, Seto J, Patil S, Nudelman G, Sealfon SC: Getting started in biological pathway construction and analysis. PLoS Comput Biol 2008;4:e16.
39 Kunegis J, Lommatzsch A, Bauckhage C: The slashdot zoo: mining a social network with negative edges; in: WWW '09: Proceedings of the 18th International Conference on World Wide Web. New York, ACM, 2009, pp 741–750.
40 Barrett JC, Fry B, Maller J, Daly MJ: Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics 2005;21:263–265.
41 Su Z, Marchini J, Donnelly P: HAPGEN2: simulation of multiple disease SNPs. Bioinformatics 2011;27:2304–2305.
42 International HapMap Consortium, Frazer KA, Ballinger DG, Cox DR, et al: A second generation human haplotype map of over 3.1 million SNPs. Nature 2007;449:851–861.
43 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. Hum Hered 2012;74:97–108.
44 Koukourakis MI, Giatromanolaki A, Sivridis E, Gatter KC, Harris AL; and the Tumor and Angiogenesis Research Group: Pyruvate dehydrogenase and pyruvate dehydrogenase kinase expression in non small cell lung cancer and tumor-associated stroma. Neoplasia 2005;7:1–6.
45 Tiede I, Fritz G, Strand S, Poppe D, Dvorsky R, Strand D, Lehr HA, Wirtz S, Becker C, Atreya R, Mudter J, Hildner K, Bartsch B, Holtmann M, Blumberg R, Walczak H, Iven H, Galle PR, Ahmadian MR, Neurath MF: CD28-dependent Rac1 activation is the molecular target of azathioprine in primary human CD4+ T lymphocytes. J Clin Invest 2003;111:1133–1145.
46 Gao W, Sweeney C, Walsh C, Rooney P, McCormick J, Veale DJ, Fearon U: Notch signalling pathways mediate synovial angiogenesis in response to vascular endothelial growth factor and angiopoietin 2. Ann Rheum Dis 2012;72:1080–1088.
47 Khatri P, Sirota M, Butte AJ: Ten years of pathway analysis: current approaches and outstanding challenges. PLoS Comput Biol 2012;8:e1002375.