Original Paper

Human Heredity

Hum Hered 2012;74:97–108
DOI: 10.1159/000347188

Received: November 20, 2012
Accepted after revision: January 17, 2013
Published online: March 1, 2013

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

Saskia Freytag[a] Heike Bickeböller[a] Christopher I. Amos[c] Thomas Kneib[b] Martin Schlather[d]

[a]Department of Genetic Epidemiology, Medical School, and [b]Department of Statistics and Econometrics, Georg-August University Göttingen, Göttingen, Germany; [c]Department of Community and Family Medicine, Dartmouth College, Geisel School of Medicine, Lebanon, N.H., USA; [d]Institute for Mathematics, University of Mannheim, Mannheim, Germany

Key Words
Logistic kernel machine regression · Size bias · Pathway analysis · GWAS · Rheumatoid arthritis

Abstract
Objectives: The logistic kernel machine test (LKMT) is a testing procedure tailored towards high-dimensional genetic data. Its use in pathway analyses of case-control genome-wide association studies results from its computational efficiency and flexibility in incorporating additional information via the kernel. The kernel can be any positive definite function; unfortunately, its form strongly influences the test's power and bias. Most authors have recommended the use of a simple linear kernel. We demonstrate via a simulation that the probability of rejecting the null hypothesis of no association just by chance increases with the number of SNPs or genes in the pathway when applying a simple linear kernel. ***Methods:*** We propose a novel kernel that includes an appropriate standardization in order to protect against any inflation of false positive results. Moreover, our novel kernel contains information on gene membership of SNPs in the pathway. ***Results:*** When applying the novel kernel to data from the North American Rheumatoid Arthritis Consortium, we find that even this basic genomic structure can improve the ability of the LKMT to identify meaningful associations. We also demonstrate that the standardization effectively eliminates problems of size bias. ***Conclusion:*** We recommend the use of our standardized kernel and urge caution when using non-adjusted kernels in the LKMT to conduct pathway analyses.

Introduction

Conventional genome-wide searches for associations between a single SNP and a complex disease have recently been complemented by pathway-based association analysis [1]. Pathway-based analysis aims to identify associations between networks of genes and the investigated disease. Growing experimental evidence suggests that common diseases are caused by many genes organized in highly interconnected networks rather than individual genes. Such networks typically represent key functions of the biological processes involved in building and sustaining an organism. Thus, pathway-based analyses can help to provide biologically meaningful interpretations [2] and

KARGER

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

0001–5652/12/0742–0097$38.00/0

Saskia Freytag
Department of Genetic Epidemiology, Medical School
Georg-August University Göttingen
DE–37073 Göttingen (Germany)
E-Mail saskia.freytag@med.uni-goettingen.de

may even highlight promising candidates for therapeutic intervention. They take full advantage of the wide opportunities provided by genome-wide association (GWA) studies, while helping to overcome the main challenges of GWA studies. Grouping SNPs in biologically meaningful sets improves otherwise low power. Firstly, the number of tests conducted is reduced, allowing a less stringent significance threshold. Secondly, joint activity of multiple moderately associated SNPs in the same pathway will be detected with a higher probability than in the single marker setting [3]. Finally, pathways might account for genetic heterogeneity, as most SNPs in the same pathway contribute towards one particular biological function.

The field of pathway analysis has seen a marked increase in progress, instigated by sustained methodological research. In this paper, we focus on one of these pathway-based methods – the logistic kernel machine test (LKMT), a semiparametric kernel-based testing procedure [4]. This procedure belongs to a branch of machine learning tools known as kernel methods, which have proven extremely valuable in the analysis of high-dimensional data. Furthermore, they perform well without the need to specify the functional relationship between the effects of SNPs in a pathway and the disease status correctly. It is well-known that genes, and therefore SNPs, in the same pathway do not convey disease risk independently of each other; instead, they are often involved in disease susceptibility and progression through complex networks. Such interactions between SNPs are difficult to accommodate in alternative regression-based methods without incurring sizeable power losses. In the LKMT, such relationships can be easily allowed for through the specification of an appropriate kernel function. Additionally, the LKMT is computationally efficient and permits the seamless integration of covariate effects. Its validity and good performance have been demonstrated in various genetic scenarios [4–6].

In this framework, the kernel acts as the core of the LKMT. We demonstrate here that the commonly used kernels introduce bias. This bias, which manifests itself in an inflation of the type I error, results from differently sized pathways. Size refers to either the number of SNPs or the number of genes that belong to a pathway. This particular type of bias has long been known to exist for many alternative pathway-based methods [1]. It is usually accounted for by computationally costly permutations. Here, we propose a different strategy that requires considerably less computational time than permutation strategies by using novel kernels with a correction for the expected size bias.

We applied the LKMT to real GWA data from the North American Rheumatoid Arthritis Consortium (NARAC) [7]. Rheumatoid arthritis (RA) is one of the few complex diseases in which GWA studies have been able to identify many susceptibility genes [8]. However, only a limited number of links between genes and RA, apart from the pivotal human leukocyte antigen (HLA) region, have been demonstrated convincingly. Furthermore, immune responses, which involve multiple positive and negative genetic regulators, critically determine development and progression of inflammatory diseases such as RA [2]. This makes RA data an interesting data set in the development of pathway-based methods. Our analysis with the LKMT identified many pathways that include already known susceptibility genes. This does not merely confirm previously established results, but has the potential of revealing compelling functional connections through the network structure. Moreover, we identified novel associations with pathways for ATP-binding cassette (ABC) transporters and extracellular matrix (ECM) receptor interaction. We found these associations particularly intriguing, as these pathways are known to be involved in many inflammatory diseases.

The remainder of this paper is organized as follows. First, we give a brief description of the LKMT and outline the proposed new kernel functions. Then, we present simulation results confirming the existence of size bias for p values obtained using the LKMT with existing kernel functions. We conducted simulation studies focusing exclusively on the scenario of no real genetic effect, since simulating scenarios with comparable true genetic effect for different pathway lengths is extremely challenging. Furthermore, there exists little knowledge concerning the exact interaction structures causing disease. Size bias is also demonstrated for the LKMT with some kernel functions based on our real RA data. Later, we compare results from the LKMT to results from two other pathway-based methods. Finally, we discuss our findings and their implications for RA research.

Methods

Logistic Kernel Machine Test

In 2008, Liu et al. [4] were the first to apply the kernel machine framework to genetic pathways in GWA studies. In 2010, Wu et al. [5] demonstrated via a gene-based simulation study that this framework indeed allows for powerful and flexible analysis.

One of the appeals of the kernel machine framework is its ability to deal with high dimensionality, i.e. the number of explanatory variables is much greater than the number of samples. High dimensionality is usually addressed through penalization, namely

by minimizing a loss function plus a positive, increasing penalty function. In the kernel machine framework, this selection is done according to the scalar product of the reproducing kernel Hilbert space, $\mathbf{H}_K$. Such a minimization problem has a solution of the form $f(x) = \tilde{\boldsymbol{\alpha}}^T\mathbf{K}(x)$, where $\tilde{\boldsymbol{\alpha}}$ represents a vector of unknown parameters and $\mathbf{K}(x)$ is a vector of functions produced by the kernel function K (representer theorem) [9]. Here, we consider the function space $\mathbf{H}_K$ as being generated by the kernel function K, in that the $\mathbf{H}_K$ is the closure of all linear combinations $f(x) = \tilde{\boldsymbol{\alpha}}^T\mathbf{K}(x)$. Any positive definite function K is allowed, thus a wide range of statistical models is subsumed without the need to specify their functional forms correctly. Furthermore, such a penalization with a maximum likelihood loss function can be shown to be equivalent to generalized linear mixed models (GLMMs) and thus allows the use of the rapidly computed score test. In the following, we will briefly explain how this reasoning yields the LKMT in the specific scenario of genetic pathway analysis.

We assume that a population-based case-control GWA study with n individuals was conducted. Let pathway p contain l_p SNPs with genotype values $z_{i1}, ..., z_{il_p}$ for the i-th subject, which are coded in a ternary fashion corresponding to the number of minor alleles. The case-control status for the i-th individual is denoted by y_i. Furthermore, for every individual, an additional set of m informative covariates $x_{i1}, ..., x_{im}$ was collected. The LKMT assumes a semiparametric model given by $\text{logit}(P(y_i = 1)) = \beta_0 + \beta_1 x_{i1} + ... + \beta_m x_{im} + h(z_{i1}, ..., z_{il_p})$, where $\beta_0, \beta_1, ..., \beta_m$ are intercept and regression coefficient terms, also summarized as vector β. The function $h \in \mathbf{H}_K$ describes the influence of the SNPs on the logit of the probability of being a case. Omitting mathematical details, the above model can be shown to be equivalent to a hierarchically expressed GLMM of the following form [4]:

$$\text{logit}\,(P(y_i = 1)) = \beta_0 + \beta_1 x_{i1} + ... + \beta_m x_{im} + b_i \quad (1)$$
$$\mathbf{b} \sim \mathcal{N}(\mathbf{0}, \sigma_K^2\mathbf{K}).$$

Random regression coefficient vector $\mathbf{b}$ has a multivariate normal distribution with mean vector $\mathbf{0}$ and variance matrix $\sigma_K^2\mathbf{K}$. Matrix $\mathbf{K}$ results from applying kernel function K to every combination of individual pairs in the data set. This hierarchical expression, common in the Bayesian perspective, allows the kernel matrix to be interpreted as a prior covariance structure [9].

In the context of genetic epidemiology, we are interested as to whether or not there is an overall genetic pathway effect, i.e. the null hypothesis H_0: $h(\mathbf{z}_p) = 0$, where $\mathbf{z}_p$ is the genotype matrix for pathway p. Taking advantage of the connection to GLMMs, such a null hypothesis is equivalent to testing no variance component, i.e. H_0: $\sigma_K^2 = 0$. This can be done via a score test, which has been demonstrated to be quite powerful in such situations [4]. Moreover, we only need to estimate β under H_0, saving considerable computational effort. The test statistics can be expressed by

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

where $\mathbf{y}$ is the vector of the n responses, $\mathbf{y} = (y_1, ..., y_n)$, and $\hat{\mu}$ is a vector with elements $\hat{\mu}_i = \text{logit}^{-1}(\beta_0 + \beta_1 x_{i1} + ... + \beta_m x_{im})$, the maximum likelihood estimate under H_0 for the i-th individual. The distribution of Q is a complicated mixture of distributions. Fortunately, it can be well approximated using a Satterthwaite procedure, which allows the approximate calculation of the effective degrees of freedom of a linear combination of χ_1^2 distributions (for more details, see Wu et al. [5]).

Existing Kernels for Genomic Information in Pathways

In genetic epidemiology, kernel functions are constructed to convert genomic information from two individuals to a quantitative value reflecting their genetic similarity or dissimilarity [10]. Beside the requirement that the kernel must be a positive definite function, no mathematical guidelines for the creation of kernels exist. This immense flexibility makes the construction of meaningful kernels challenging. Since the kernel defines the whole set of correlation coefficients between individuals, its choice strongly influences the power of the method (see Schaid [9], p. 118). So a kernel function that does not anticipate the nature of the true effect will lose efficiency.

Authors implementing the LKMT for GWA studies have advocated the use of either the linear (LIN) or identity-by-state (IBS) kernel [4, 5]. The LIN kernel is given by

$$\mathbf{K}\left(\mathbf{z}_i, \mathbf{z}_j\right) = \sum_{k=1}^{l_p} z_{ik} z_{jk}$$

for subject i and j. The function space H defined by this kernel implies the usual multiple marker logistic regression model. The IBS kernel uses the numbers of alleles shared between two individuals as a similarity metric:

$$\mathbf{K}\left(\mathbf{z}_i, \mathbf{z}_j\right) = \sum_{k=1}^{l_p} \frac{2\mathbb{I}\left(z_{ik} = z_{jk}\right) + \mathbb{I}\left(|z_{ik} - z_{jk}| = 1\right)}{2l_p}.$$

Although the set of corresponding functions for this kernel is unknown, Wu et al. [5] demonstrated in their simulation study that this kernel has adequate power for more complex genetic relationships.

Despite their frequent use, both kernels suffer from deflation of p values due to size bias. The bias results from the tendency of larger pathways to have more variability among their kernel matrix entries. When calculated by either the IBS or LIN kernel, the probability of two values of the similarity metric differing increases with the dimension of l_p. Since SNPs cumulatively add to the value of the metric, two similarity metrics are more likely to be different when more SNPs contribute. Thus, using both the LIN and IBS kernel, the probability that we reject H_0 by chance increases. Note that the numerical constraint of the IBS kernel does not serve as a correction because it represents a constant scale factor. Since a constant scale factor would be subsumed under σ_K^2, it has no effect on the magnitude of the resulting p value. Schaid et al. [11] further noted that '[t]he magnitude of this bias depends not only on the number of SNPs in a [pathway] but also on the correlations between the SNPs'.

Moreover, neither of the two kernel functions considers all available knowledge of the genetic architecture of pathways. Up to now, the prior genomic information is confined to pure SNP membership in pathways. However, it is conceivable that the incorporation of even the simplest genomic structure could improve power significantly.

Construction of Novel Kernels for Genomic Information in Pathways

The construction of a novel kernel allows to simultaneously address all shortcomings mentioned in the previous section. In particular, we aimed to construct kernels that produce p values invariant to the size of the pathway. We also decided to account for membership of SNPs in genes. To obtain kernels with these properties, we built on the experience with positive definite functions in the

field of spatial statistics. In recent years, methods originating from this field, such as kriging, have already been successfully applied in animal genetics [12].

We suggest the variogram value between two individuals *i* and *j* given by

$$\gamma_{i,j}^{g} = \left(\frac{||\mathbf{z}_i^g - \mathbf{z}_j^g||^2}{s_g} \right)^{\delta_g}$$

as a measure for the genetic distance between individuals; for a particular gene *g*. Vectors $\mathbf{z}_i^g$ and $\mathbf{z}_j^g$ in the squared norm denote the vectors of genotype values of SNPs contained in gene *g*; δ_g and s_g are parameters with restrictions $0 < \delta_g \leq 1$ and $s_g > 0$. Schoenberg's relation now nicely suggests a kernel, which enables us to correct for the number of SNPs in the gene [13]: the kernel related to the variogram $\gamma_{i,j}^g$ is $C(\gamma_{i,j}^g) = \exp(-\gamma_{i,j}^g)$, which is called the stable kernel or the powered exponential kernel. Choosing

$$\delta_g = \sqrt{\frac{k_g}{\max_{g \in p} k_g}} \in [0,1] \text{ and} \tag{3}$$

$$s_g = \hat{\mu}_g k_g \tag{4}$$

for each gene *g*, we can keep the mean and variance of $C(\gamma_{i,j}^g)$ pretty much constant across all genes in a pathway (please also refer to Section A of the online supplementary material, see www.karger.com/doi/10.1159/000347188). Here, scalar k_g equals the number of SNPs contained in gene *g*; $\hat{\mu}_g$ denotes the empirical expectation of the squared norm calculated from two genotype vectors of size one SNP.

Using the principle that the set of positive definite functions is closed under addition and multiplication, we obtain the kernels

$$K(\mathbf{z}_i^g, \mathbf{z}_j^g) = \sum_{g \in p} C(\gamma_{i,j}^g) = \sum_{g \in p} \exp(-\gamma_{i,j}^g) \text{ and} \tag{5}$$

$$K(\mathbf{z}_i^g, \mathbf{z}_j^g) = \prod_{g \in p} C(\gamma_{i,j}^g) = \prod_{g \in p} \exp(-\gamma_{i,j}^g), \tag{6}$$

which we call the additive powered exponential (ADD) and multiplicative powered exponential (MULT) kernels. The application of an additive kernel to all genes in the same pathway implicitly assumes a linear influence of the genes on the phenotype of interest. The application of a multiplicative kernel to all genes, on the other hand, assumes interactions.

Despite correcting for size bias on the gene level, neither the ADD nor the MULT kernel accounts for the number of genes in the pathway. Again, this is a potential source of type I error inflation. Using the relationship between bounded variograms and kernels [14], we can find a kernel function for additive gene effects that includes a correction regarding the number of contributing genes. The resulting standardized ADD (STAND.ADD) kernel is defined as

$$K(\mathbf{z}_i^g, \mathbf{z}_j^g) = 1 - \left[\frac{\sum_{g \in p} \{1 - C(\gamma_{i,j}^g)\}}{r_p \hat{\mu}_p} \right]^{\delta_p}, \tag{7}$$

where r_p denotes the number of genes in pathway *p* and $\hat{\mu}_p = \hat{E}\{C(\gamma_{i,j})\}$. Similar to the standardization within a gene, we choose

$$\delta_p = \sqrt{\frac{r_p}{\max_p r_p}} \in [0,1].$$

Details on this standardization approach can be found in Section B of the supplementary material. We correct the MULT kernel by recognizing that equation 6 can easily be rewritten. Without changing the positive definite nature of this function, we can then introduce a correction factor

$$\sqrt{\frac{1}{r_p}}$$

and obtain the standardized MULT (STAND.MULT) kernel:

$$K(\mathbf{z}_i^g, \mathbf{z}_j^g) = \exp\left(-\sqrt{\frac{1}{r_p}} \sum_{g \in p} \gamma_{i,j}^g \right). \tag{8}$$

Details on this standardization approach can be found in Section C of the supplementary material.

We know that a large percentage of SNPs located in genes is in strong pairwise LD. Consequently, our assumption of independence is violated. Depending on the kernel, this could result in an increase in false positives or an increase in false negatives. We use the concept of effective number of independent SNPs in a region, introduced by Cheverud [15], to account for such dependencies in our standardization. He suggested measuring the total amount of correlation between SNPs by the variance of the eigenvalues obtained from their correlation matrix. If the SNPs are perfectly correlated, the variance of the eigenvalues will be maximized. When there is no correlation, the variance equals 1. This means one can calculate the proportional reduction in the number of independent elements and thus an effective number of SNPs, k_g^{eff}. The LD-scaled versions of the STAND.MULT (LD.STAND.MULT) and of the STAND.ADD (LD.STAND.ADD) kernel use this k_g^{eff} instead of k_g.

NARAC RA Data

We applied the LKMT with our novel kernel functions to real GWA data and compared our results with those achieved when using either the LIN or the IBS kernel. To demonstrate consistency and robustness of our method, we compared our results to those obtained by Sohns et al. [16] on the same data with two different pathway analysis methods.

The NARAC conducted a GWA study on 868 independent cases and 1,194 independent controls to identify genetic markers associated with RA risk [7]. All participants in this study gave IRB-approved informed consent to have their genetic predictors of risk for RA evaluated. All samples were genotyped at 545,080 loci with the HumanHap500v1 array. After stringent quality control, 866 cases, 1,189 controls and 492,209 additively coded SNPs on chromosomes 1–22 remained. Missing genotypes in loci with a genotype missing rate below 10% (other loci were previously excluded) were imputed using the software BEAGLE [17]. Imputed genotype dosages make the calculation of genomic similarities possible. Imputed genotype dosage can take any value between 0 and 2, reflecting the uncertainty of imputation.

NARAC's study design hints at some level of population stratification. Cases were predominately of Northern European ancestry. Controls were sampled in the New York Metropolitan area. Therefore, in accordance with the New York ethnicity mix, controls were somewhat enriched for Southern European or Ashkenazi Jewish ancestry compared to cases. Population stratification can introduce confounding and should therefore be accounted for in the analysis. Despite this, we did not specifically correct for population stratification, assuming instead that all analyzed pathways

included a sufficient number of null markers, i.e. markers without true genetic effect. Setakis et al. [18] demonstrated in a simulation study that the inclusion of such markers in a logistic regression model provides good protection against possible ramifications of population stratification; they stated '... each of the [null markers] soaks up some of the effect of population stratification, but because this effect is shared across many markers, none of them is individually significant'. Since the logistic regression model with a random effect can be shown to be equivalent to the LKMT with the LIN kernel, we argue that population stratification is inherently taken care of.

We considered all pathways in the KEGG: Kyoto Encyclopedia of Genes and Genomes [19]. A SNP is mapped to a pathway only when it is located within one or more genes that belong to this particular pathway. Under these criteria, we were able to assemble 62,892 SNPs in 244 pathways. Summary statistics concerning this particular annotation can be found in table 1. (We also tried more liberal assignments, but the results are not shown as they were qualitatively similar.) We tested each of the pathways under the LIN, IBS, STAND.MULT, STAND.ADD, LD.STAND.MULT and LD.STAND.ADD kernel. (The MULT and the ADD kernels were not analyzed, as they are clearly inferior.)

Comparing Different Pathway-Based Approaches Using the NARAC Data

To evaluate the performance of the LKMT with our novel kernels, we followed the analysis protocol by Sohns et al. [16] for the same data. They restricted the number of pathways to 100 candidate pathways. These were selected to include all top associations from a previously conducted single marker analysis. SNPs were assigned to genes if they were located either inside the gene or in a 500-kbp window around the transcription start or end. Sohns et al. [16] applied hierarchical Bayes prioritization (HBP) [20] and Gene Set Enrichment Analysis (GSEA) [21] to the data. HBP is an empirical Bayes approach aiming at re-ranking markers using prior covariates. GSEA assesses whether associated genes in a pathway are overrepresented. The approaches are conceptually as well as methodically different, thus serving as diverse benchmarks. Since the HLA complex has an unusually strong association with RA, the authors reanalyzed the data, excluding all genes in the region between MOG and KIFC1. This allows for an assessment of performance beyond the implication of HLA. We analyzed the 100 candidate pathways: once when the HLA complex was included and once when it was excluded using the LD.STAND.MULT kernel function (as this turned out to be the most size-invariant kernel). In particular, we were interested in how similar the ranked lists obtained by the different methods are. In order to visualize similarities, we used a method introduced by Antosh et al. [22] in 2011. First, we ranked the results from each method from most significant to least significant. We then plotted the fraction of overlap between the lists produced by GSEA and the LKMT as well as HBP and the LKMT against the proportion of selected entries in these lists. Furthermore, we evaluated whether or not these overlaps are statistically significant.

Simulations to Assess Size Bias

In order to verify our theories regarding size bias and the LIN kernel, we studied the LKMT by considering its empirical type I error for pathways of different sizes. For feasibility of implementation, we did not simulate a real pathway, considering instead the region between 239,000 and 247,200 kbp on chromosome 1. Genotype data at 3,500 loci (all with a MAF >5%) as well as case-control status were simulated using HAPGEN2 [23] and the CEU sample of the International HapMap Project. Since HAPGEN2 simulates haplotypes based on reference data and fine-scale recombination rates, it preserves LD structures from the reference population. Thus, it closely mimics real genetic studies.

We considered 1,000 simulations with 1,000 cases and 1,000 controls. For each simulation, we applied the LKMT with the LIN kernel for different numbers of SNPs. We started with the first 250 SNPs and then increased this number at each turn by 250 until reaching 3,500 SNPs. For each configuration, we found the proportion of Bonferroni-adjusted p values less than $\alpha = 0.01, 0.05$ and 0.1.

We also investigated the impact different pathway sizes have on the size of p values calculated by the LKMT under all described kernels. Doing this via a simulation study would have proven to be computationally infeasible. Therefore, we randomly formed 'pseudo-pathways' including different numbers of SNPs or genes from the real NARAC RA data. We repeated this 50 times for each pathway size. We varied the number of randomly selected SNPs to be members of our pseudo-pathways from 100 to 800 (step size = 100). In the gene-based analysis, we examined pseudo-pathways containing between 10 and 50 genes (step size = 10). In order to be able to separate gene size and SNP size, we excluded genes with more than 200 SNPs in this scenario. For the sake of computational ease, we used the STAND kernels as a proxy for their LD-scaled versions.

Table 1. Summary statistics for all pathways in the analysis, when the assignment of SNPs to pathways is strict (i.e. SNPs need to be located in a gene of the pathway)

	SNPs in pathway	Genes in pathway	Genes with >10 SNPs in pathway
1st quantile	199	21	6
Median	487	40	13
Mean	900	54.3	20.6
3rd quantile	1,227	68	31
Maximum	10,850	839	272

Results

Inflation of Empirical Type I Error with Pathway Size

The empirical type I error rates for the LKMT with the LIN kernel are presented in figure 1. On the basis of our simulation, the test has a correct error rate for $\alpha = 0.1$. The test appears to be too liberal for smaller values of α, e.g. $\alpha = 0.01$. As suspected, we can observe an inflation of the empirical type I error rate with an increasing number of SNPs located in the considered region. This phenomenon exists for all investigated type I error rates. Considering

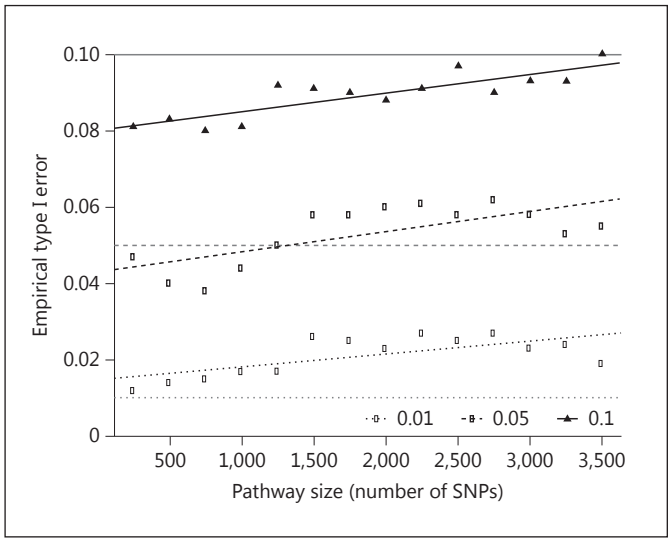

Fig. 1. Empirical type I error rates at α = 0.01, 0.05, 0.1 (grey lines) for the LKMT with the LIN kernel function when applied to regions of varying sizes based on 1,000 simulations. The black lines are the corresponding regression lines estimated on the points.

the frequently used p value of 0.05, the type I error rate is not protected as soon as more than 1,250 SNPs are considered. For significantly fewer SNPs, the test even appears conservative. Note that this means that unless pathway analysis is performed, where the number of SNPs in a pathway easily exceeds 1,000, the LIN kernel produces valid results. Thus, the application of the LKMT with the LIN or IBS kernel as a method for single gene analysis poses no problem.

Assessment of Size Bias Based on Randomly Sampled Pseudo-Pathways

The results from resampling pseudo-pathways are summarized in figure 2. If the number of SNPs is increased, the magnitude of the p values for all investigated kernels decreases. However, the different kernel functions are not affected equally. The trend is most profound for the IBS kernel, while the ADD kernel seems to be the least affected. For all kernels, we observed an increase in the variability of the p values when the number of SNPs in the pseudo-pathway rises. Since we did not remove SNPs that have any genuine influence on an individual's RA risk, such an effect is to be expected. On the one hand, raising the number of contributing SNPs increases the probability that a causal SNP will be sampled. On the other hand, many null markers are capable of diluting genuine genetic effects, as observed in simple linear regression. For linear regression, the p value associated with the coefficient of a genuinely correlated variable increases when further uninformative variables are added to the model. Overall, these results suggest that accounting for the number of SNPs in a pathway does not offer sufficient protection against size bias. Figure 2a verifies this statement. p values obtained using the LKMT with the ADD kernel decrease with an increasing number of genes. It needs to be noted that the scale of this effect is small compared to that observed, owing to an increase in the number of SNPs. For the STAND.ADD kernel, such an increase can also be observed when there are more than 30 genes in the pathways. In fact, when there are a large number of genes in the pathway, the STAND.ADD kernel seems to perform worse than the ADD kernel. The STAND.MULT kernel displays the opposite trend. The standardization works well up to 30 genes and then becomes overly conservative. For the strict SNP pathway assignment chosen here, around 25% of pathways have more than 30 genes and are thus affected (compare table 1). If the assignment were chosen differently though, this could render the LKMT with this particular kernel structure ineffective. The multiplicative nature of the STAND.MULT kernel implies that genuine effects are diluted with a greater probability whenever more genes are added. Furthermore, it is possible that these trends can be explained partly by LD patterns that we fail to account for in this permutation study.

NARAC RA Data Analysis Results

Empirical Evidence for Size Bias

In figure 3, we plotted the p values of pathways failing to reach the Bonferroni-corrected significance threshold of 0.05 against the pathways' respective sizes as represented by number of SNPs. We excluded significant pathways from our analysis, in order to prevent distortion due to genuine associations. We observed obvious correlations between size and magnitude of p value for the LIN and the IBS kernel. Since these observations are in agreement with the previously discussed results, this is further confirmation of the existence of size bias. Unfortunately, there is also some evidence for such a trend in the STAND.ADD and the LD.STAND.ADD kernel. This is in accordance with the previously discussed permutation results. The STAND.MULT kernel, on the other hand, seems to overcompensate for large pathways. There is no evidence of bias in any direction for the LD.STAND.MULT kernel. A simple linear regression, modeling the p value of a pathway with this kernel by its respective size

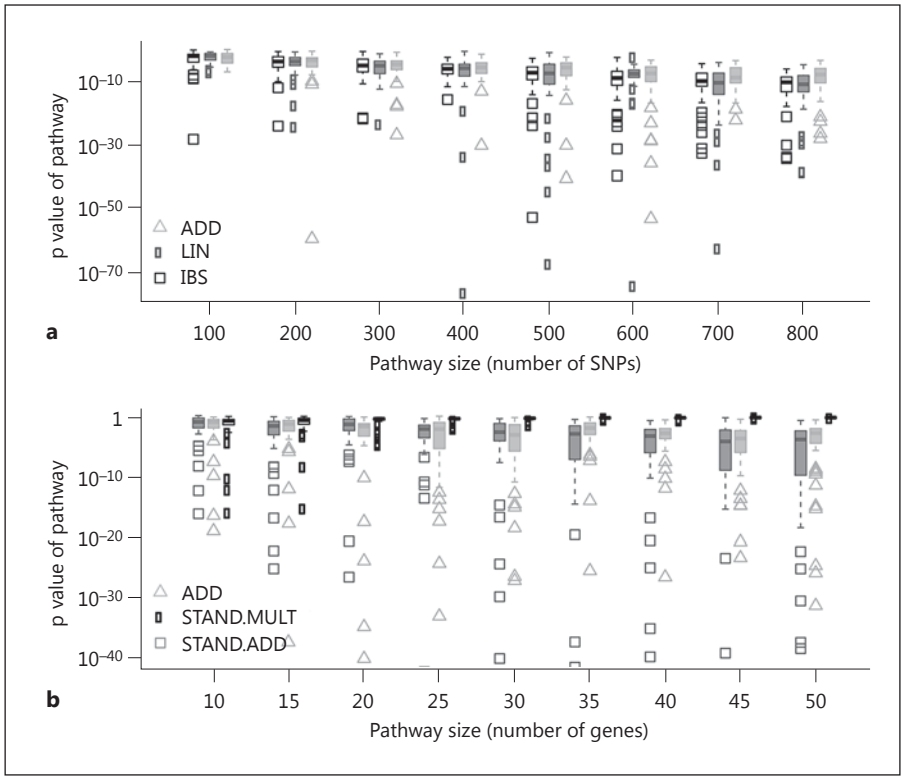

Fig. 2. a Boxplots of p values obtained by the LKMT with the LIN, IBS and modified ADD kernel function against the size of respective pseudo-pathways as measured by number of SNPs. **b** Boxplots of p values obtained by the LKMT with the ADD, STAND.ADD and STAND.MULT kernel function against the size of pseudo-pathways as measured by number of genes. For each investigated size, the pseudo-pathways were randomly put together 50 times.

(measured by the number of SNPs in the pathways), reveals a regression coefficient close to 0. Furthermore, the explained variation of the model amounts to no more than 7%. Thus, we decided to carry out further analyses primarily with this kernel.

Pathway Association Results from the LKMT

Table 2 summarizes the number of significant results from the analysis with the LKMT using different kernels. The significance thresholds 0.05 and 0.1 in the table are both Bonferroni-corrected. As expected, the LIN and IBS kernels identify the most pathways as associated with RA. Furthermore, almost all previously identified susceptibility pathways are replicated, i.e. pathways which include at least one gene, with more than 10 SNPs genotyped in the NARAC data, that has also been significantly associated with RA in at least one scientific publication. (For a complete list of all these susceptibility genes, please refer to Section D of the online suppl. material.) However, we obtained the greatest ratio of significant previously identified susceptibility pathways to all significant pathways for the LD.STAND.MULT kernel. This yet again confirms the robustness of this kernel. In the following, all further analyses will therefore be focused on this kernel.

Table 2. The number of significant pathways and the number of previously implicated pathways identified by the LKMT with different kernel functions at the Bonferroni-corrected thresholds 0.05 and 0.1

LKMT	All KEGG pathways		Previously identified susceptibility pathways	
	0.05	0.10	0.05	0.10
LIN	112	120	48	50
IBS	103	112	46	46
STAND.MULT	26	27	20	20
LD.STAND.MULT	32	33	22	22
STAND.ADD	64	70	32	33
LD.STAND.ADD	41	45	27	28

Previously implicated pathways refer to pathways including at least one gene shown to be significantly associated with RA in a scientific publication (compare table 1).

The Venn diagram in figure 4 indicates that 10 novel pathways are detected using the LD.STAND.MULT kernel. Interestingly, three of these would have been missed using the LIN kernel. This leads us to believe that we do

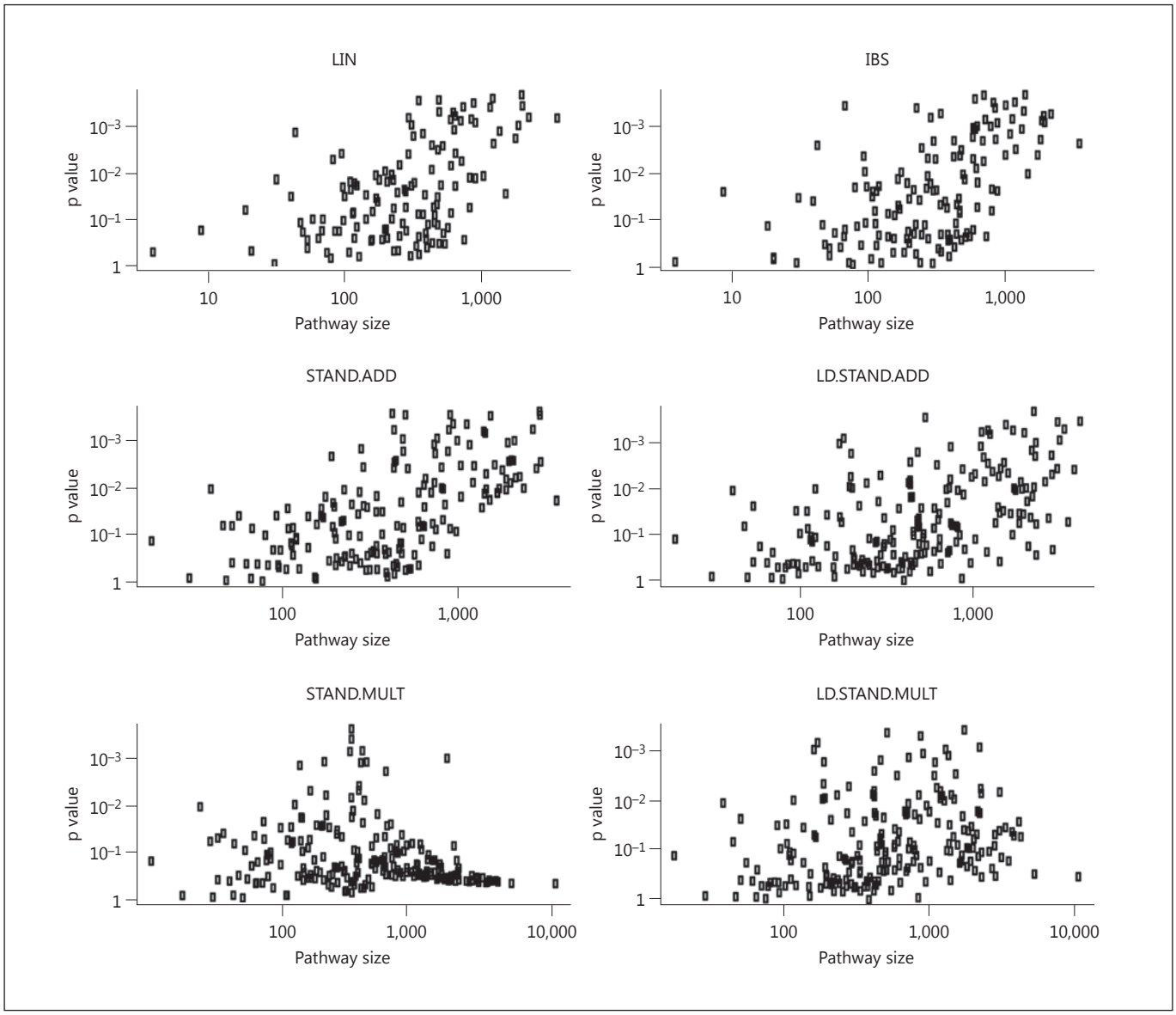

Fig. 3. Plots of nonsignificant p values against the size of respective pathways as measured by number of SNPs. The different panels correspond to different kernel functions used in the LKMT to calculate the p values. Note that both axes are logarithmic and the y-axis is reversed.

indeed add valuable information by incorporating information on gene membership. Table 3 lists all significant findings. As expected, pathways for inflammatory diseases, such as systemic lupus, are detected as being associated. Owing to the strong influence of the HLA region, these have highly significant p values. Newly implicated pathways for RA include pathways for ABC transporters, ECM receptor, nicotinate and nicotinamide metabolism, and focal adhesion. It is interesting to note that of these novel identifications, the pathways for ABC transporters and ECM receptor have the smallest p values. The pathway for vitamin B_6 metabolism, despite being known to be involved in the pathogenesis of RA, is counted as a novel finding. Susceptibility genes, which have been previously identified as being associated with RA and are located in this particular pathway, are not included in our analysis. This once more illustrates the power of pathway-based approaches.

Comparison of Results from the LKMT and Other Pathway-Based Approaches

In figure 5, we compare results from the pathway-based approaches used by Sohns et al. [16] with results obtained by the LKMT with the LD.STAND.MULT kernel. For data including SNPs in the HLA region, there is a considerable significant overlap between the LKMT and GSEA. On the contrary, we did not find any significant overlap between the LKMT and HBP. Sohns et al. [16] demonstrated that the results from GSEA and HBP indeed usually differ. Interestingly, once the HLA region was removed from the data, both HBP and GSEA show some significant overlap.

The LKMT identifies 31 pathways using a Bonferroni-corrected threshold, while GSEA finds 47 significant pathways at a false discovery rate of 0.05. (HBP does not allow significances to be determined for individual pathways.) However, when we excluded SNPs located in the HLA region, the results changed dramatically. No pathways can reach significance in GSEA, although the LKMT identifies 6 susceptibility pathways. This indicates that the LKMT is more powerful and at the same time more robust than GSEA. We think that the difference in robustness stems from the ability of the LKMT to consider all markers in the pathway, whereas GSEA relies on the most significant marker in the genes located in the pathway. Therefore, the p value determined by the LKMT is not solely driven by the strongest association in every gene.

Even though this is not a thorough performance analysis, the results indicate that the LKMT is superior to both methods. Unlike GSEA, it does not rely on permutations. Other than HBP, it has the advantage of producing significance values, while retaining the flexibility of HBP.

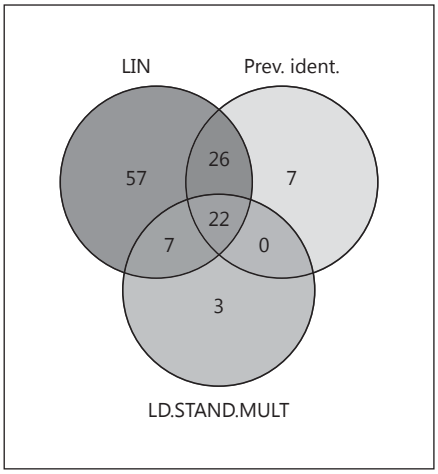

Fig. 4. Venn diagram of results obtained by the LKMT with different kernels (LIN and LD.STAND.MULT) and previous studies on RA, demonstrating the overlap between significant results. We did not include results from any of the alternative kernel functions to keep the Venn diagram clear. Also shown is the overlap with the previously implicated pathways (Prev. ident.).

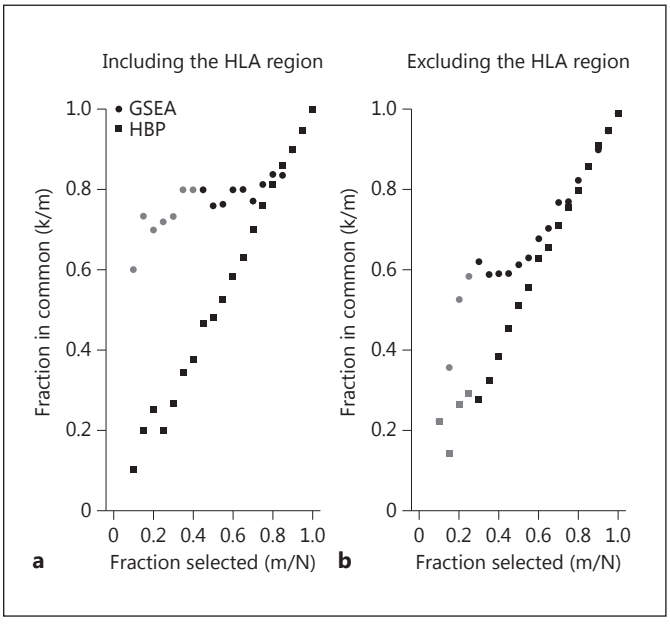

Fig. 5. Proportion of selected pathways (m) from total number (N) in ranked lists against proportion of pathways in common (k) from the selection between the LKMT with the LD.STAND.MULT kernel function and GSEA and HBP, respectively. **a** Results obtained analyzing unmodified pathways. **b** Results obtained when the SNPs in the HLA region were excluded. Overlaps that are statistically significant are in grey.

Discussion

For the conventional kernels in logistic kernel machine-based pathway analysis, we demonstrate deflation of p values with increasing pathway size, in terms of number of genes as well as number of SNPs. In order to expunge this bias, we propose a novel pathway-size-invariant kernel. Its application to a case-control study on RA empirically illustrates the effectiveness of our novel kernel compared with the LIN or IBS kernel. Furthermore, its use reveals new functional connections between biological pathways and RA progression and development.

The main idea behind the construction of the new invariant kernel was to standardize the similarity metric by its empirical expectation and variance across all individu-

Table 3. Significant pathways associated with RA as discovered by the LKMT with the LD.STAND.MULT kernel function

KEGG pathway identifier	Description	p value of association	KEGG pathway classification
path:**hsa00240**	Anemia due to nucleotide metabolism disorders	7.539×10^{-6}	Metabolism
path:**hsa00250**	Alanine, aspartate and glutamate metabolism, Canavan disease	9.614×10^{-5}	Metabolism
path:**hsa00280**	Fatty-acid oxidation disorder	2.296×10^{-5}	Metabolism
path:**hsa00512**	Mucin-type O-glycan biosynthesis	5.282×10^{-5}	Metabolism
path:**hsa00750**	Vitamin B_6 metabolism	2.281×10^{-5}	Metabolism
path:hsa00760	Nicotinate and nicotinamide metabolism	4.778×10^{-7}	Metabolism
path:hsa04145	Phagosome	1.075×10^{-296}	Cellular processes
path:**hsa04510**	Focal adhesion	1.991×10^{-4}	Cellular processes
path:**hsa02010**	ABC transporter	3.546×10^{-31}	Environmental information processing
path:**hsa04512**	EMC receptor interaction	8.194×10^{-17}	Environmental information processing
path:hsa04514	Cell adhesion molecules	1.076×10^{-99}	Environmental information processing
path:hsa04612	Antigen processing and presentation	1.046×10^{-171}	Organismal systems
path:hsa04640	Hematopoietic cell lineage	4.191×10^{-230}	Organismal systems
path:hsa04650	Natural killer cell-mediated cytotoxicity	5.982×10^{-5}	Organismal systems
path:hsa04672	Intestinal immune network for IgA production	6.634×10^{-154}	Organismal systems
path:**hsa04974**	Protein digestion and absorption	7.865×10^{-8}	Organismal systems
path:hsa04940	Type I diabetes mellitus, antigen-activated Th1 cells produce IL-2 and IFN-γ	2.319×10^{-295}	Human diseases
path:hsa05140	Leishmaniasis	4.874×10^{-250}	Human diseases
path:hsa05145	Toxoplasmosis	4.351×10^{-179}	Human diseases
path:hsa05150	*Staphylococcus aureus* infection	6.474×10^{-182}	Human diseases
path:hsa05152	Tuberculosis	1.532×10^{-164}	Human diseases
path:hsa05164	Influenza A	1.017×10^{-168}	Human diseases
path:hsa05166	HTLV-I infection	1.230×10^{-96}	Human diseases
path:hsa05168	Herpes simplex infection	2.782×10^{-286}	Human diseases
path:hsa05310	Asthma	7.002×10^{-104}	Human diseases
path:hsa05320	Autoimmune thyroid disease	1.408×10^{-180}	Human diseases
path:hsa05322	Systemic lupus erythematosus	8.970×10^{-237}	Human diseases
path:hsa05323	Rheumatoid arthritis	5.337×10^{-224}	Human diseases
path:hsa05330	Allograft rejection	1.634×10^{-103}	Human diseases
path:hsa05332	Graft-versus-host disease	2.138×10^{-103}	Human diseases
path:hsa05340	Primary immunodeficiency	2.125×10^{-44}	Human diseases
path:hsa05416	Viral myocarditis	$<10^{-300}$	Human diseases

Highlighted in bold are all pathways that do not include previously implicated genes, for which >10 SNPs are genotyped in the data set. The pathway classification, as found in the KEGG database, is also included.

als. The chosen standardization parameters are rather unusual in order to fulfill the requirement that a kernel must be positive definite. Although these parameters are not easily interpreted, they have several advantages compared to other corrections. Most importantly, unlike for permutations, the required additional computational cost is minimal. Another approach, suggested by Schaid et al. [11], capitalizes on the multivariate normal distribution of the score statistics from logistic regression to avoid permutations. However, their approach requires many separate standardization and scoring steps, whereas our new kernel integrates all necessary corrections in an elegant fashion.

Our kernel emphasizes the concept of a pathway. Conventional kernels can share information between SNPs in the same pathway to improve power. However, these kernels fail to harness the rich network structure of pathways, as they do not incorporate information beyond the pathway membership. In our novel kernel, we attempt to utilize this concept additionally. We integrate information on gene membership and model all possible interactions between the genes by choosing a multiplicative kernel

function. Still, such information comprises only a small fraction of what is available. In particular, we fail to include prior information on known direct connections between genes in the same pathway. Chen et al. [24] revealed that associated genes in the same pathway tended to be neighbors as defined by the topology of the network graph in an example considering Crohn's disease. Explicitly specifying direct interactions of genes in the kernel function may prove a good starting point for the LKMT to exploit the vast information provided by pathway databases.

We recommend the kernel machine approach with a size-invariant kernel owing to its clear benefits and superior results compared to GSEA and HBP. The LKMT with a size-invariant kernel is more robust than other tested pathway approaches and adequate in many different scenarios. Furthermore, unlike GSEA and HBP, it is fast to compute moderate sample sizes. We advise against using kernels that offer no adjustment for the number of SNPs in a pathway. However, we see no reason to discourage the use of these kernels for gene analysis with much fewer SNPs. For small significance levels, the LKMT with such kernels fails to control the type I error rate adequately. Generally, this effect can be compared to multiple testing, in which appropriate adjustment is required for the number of tests conducted. Multiple testing corrections are either known to be conservative or liberal, depending on the adjustment and dependencies between the tests. It seems likely that our novel kernel produces conservative results in the case of large pathways. However, more research is needed to establish the effects of dependencies between SNPs, such as LD, or interaction of genes in the same pathways.

Finally, our application to the NARAC data provides promising unknown associations between pathways and RA. In particular, we believe that associations with pathways for ECM molecules and cellular receptors as well as ABC transporters are of interest in the pursuit of the genetic causes of RA. Of course, future research is necessary to study these connections more thoroughly as well as replicate these association results. ECM molecules assemble cartilage proteins. In animal models, antibodies binding to these proteins have been shown to precede arthritis induction [25]. P-glycoprotein, a member of the superfamily of ABC transporters, is thought to play a major role in mechanisms of resistance to the systemic administration of disease-modifying antirheumatic drugs and low-dose glucocorticoids [26]. This type of drug resistance is common in RA patients, who rely on these drugs for the prevention and control of joint damage. Thus, our results could shed more light onto effective treatments of systemic inflammation in RA patients.

Software

Software in the form of R code, together with a sample input data set and complete documentation, is available on request from the corresponding author.

Appendices

Appendix 1 – ADD and MULT Kernels

Let $\psi^g_{i,j}$ equal $||\mathbf{z}^g_i - z^g_j||^2$. Assuming that the minor allele count at an arbitrary SNP in gene g is an independent and identically distributed random quantity, then $\psi^g \sim D(k_g\mu_g, k_g\sigma^2_g)$ has mean $k_g\mu_g$ and variance $k_g\sigma^2_g$. The variables μ_g and σ^2_g refer to the expectation and variance of this distribution based on one SNP, respectively. Simulation studies (not shown) revealed that a minimum of 10 SNPs per gene is needed in order for this standardization to work appropriately.

For the standardization, let

$$\gamma^g = \left(\frac{\psi^g}{s_g}\right)^{\delta_g}.$$

The scaling factor s_g and the exponent δ_g denote the standardization parameters for gene g. As long as $\delta_g \leq 1$ and $s_g > 0$, the kernel function $\exp(-\gamma^g)$ is positive definite. Now using the Taylor expansion with respect to ψ_g around $k_g\mu_g$, we obtain:

$$\gamma^g \approx \left(\frac{\psi^g}{s_g}\right)^{\delta_g}\Bigg|_{\gamma^g = k_g\mu_g} + \delta_g \frac{1}{s_g}\left(\frac{\psi^g}{s_g}\right)^{\delta_g - 1}\Bigg|_{\gamma^g = k_g\mu_g} \left(\psi^g - k_g\mu_g\right). \quad (9)$$

The expectation and variance of γ_g become approximately:

$$E(\gamma_g) = \left(\frac{k_g\mu_g}{s_g}\right)^{\delta_g} = 1,$$

$$Var(\gamma_g) = \delta^2_g \frac{1}{s^2_g}\left(\frac{k_g\mu_g}{s_g}\right)^{2\delta_g - 2} k_g\sigma^2_g = \delta^2_g \frac{1}{k_g}\frac{\sigma^2_g}{\mu^2_g}. \quad (10)$$

Consequently, we get expressions for s_g and δ_g that ensure invariance with regard to the number of SNPs in the pathway. Since we also require $\delta_g \leq 1$ for all genes in the pathway, we choose $\delta_g = 1$ for $k_g = \max_{g\in p} k_g$. This automatically produces

$$\delta_g = \sqrt{\frac{k_g}{\max_{g\in p} k_g}}$$

and $s_g = \mu_g k_g$. Furthermore, we also tested whether our results were independent with regard to the length of the largest gene in the pathway (results not shown). This seemed to be the case for a moderate increase in pathway size. Results should therefore remain consistent even when a new larger non-significant gene is added to the pathway.

Appendix 2 – STAND.ADD Kernel

If one assumes that the central limit theorem holds, $\Sigma_{g\in p} \exp(-\gamma_g)$ has mean $r_p\mu_0$ and variance $r_p\sigma^2_p$. Here, r_p denotes the number of genes in the pathways, while μ_p equals $E(C(\gamma_{i,j}))$ and σ^2_p equals $Var(C(\gamma_{i,j}))$. Similar to the standardization for the gene, one

can now apply a Taylor expansion and find the appropriate standardization parameters.

Appendix 3 – STAND.MULT Kernel

Let $\omega_p = \zeta_p \sum_{g \in p} \gamma_g$ be a random variable for any arbitrary pair of individuals, where $\zeta_p > 0$ is the additional scaling factor regarding the number of genes in the pathway. Let r_p be the number of genes in pathway p. Assuming that the genes in this pathway are independent, we approximately have:

$$E(\omega_p) = \zeta_p r_p E(\gamma_g) \text{ and}$$
$$Var(\omega_p) = \zeta_p^2 r_p Var(\gamma_g), \quad (11)$$

where $E(\gamma_g)$ and $Var(\gamma_g)$ are defined as given in the previous section (see eq. 10). One can easily see that ζ_p has to be selected proportionally to

$$\sqrt{\frac{1}{r_p}}$$

in order to achieve constant variance with regard to the number of genes in the pathway. Further standardization with regard to the expectation, like

$$K(\mathbf{z}_i, \mathbf{z}_j) = \exp\left(-\sqrt{\frac{1}{r_p}} \sum_{g \in p} \left\{\gamma_{i,j}^g - 1\right\}\right),$$

is not necessary. Such standardization represents nothing more than a constant scale factor, and it would therefore merge with σ_K^2.

Acknowledgements

This work was supported by the Deutsche Forschungsgemeinschaft (DFG) research training group 'Scaling Problems in Statistics' (RTG 1644). The data used in this study were made available by the National Institutes of Health grant number AR44422.

References

1 Wang K, Li M, Hakonarson H: Analysing biological pathways in genome-wide association studies. Nat Rev Genet 2010;11:843–854.
2 Eleftherohorinou H, Wright V, Hoggart C, Hartikainen A, Jarvelin M, Balding D, Coin L, Levin M: Pathway analysis of GWAS provides new insights into genetic susceptibility to 3 inflammatory diseases. PLoS One 2009;4:e8068.
3 Fehringer G, Liu G, Briollais L, Brennan P, Amos CI, Spitz MR, Bickebller H, Wichmann HE, Risch A, Hung RJ: Comparison of pathway analysis approaches using lung cancer GWAS data sets. PLoS One 2012;7:e31816.
4 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.
5 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.
6 Basu S, Pan W, Oetting WS: A dimension reduction approach for modeling multi-locus interaction in case-control studies. Hum Hered 2011;71:234–245.
7 Amos CI, Chen W, 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 Proceedings 2009;3(suppl 7):S2.
8 Raychaudhuri S: Recent advances in the genetics of rheumatoid arthritis. Curr Opin Rheumatol 2010;22:109–118.
9 Schaid DJ: Genomic similarity and kernel methods I: advancements by building on mathematical and statistical foundations. Hum Hered 2010;70:109–131.
10 Schaid DJ: Genomic similarity and kernel methods II: methods for genomic information. Hum Hered 2010;70:132–140.
11 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.
12 Ober U, Erbe M, Long N, Porcu E, Schlather M, Simianer H: Predicting genetic values: a kernel-based best linear unbiased prediction with genomic data. Genetics 2011;188:695–708.
13 Wackernagel H: Multivariate Geostatistics. New York, Springer, 2003.
14 Gneiting T, Sasvári Z, Schlather M: Analogies and correspondences between variograms and covariance functions. Adv Appl Probab 2001;33:617–630.
15 Cheverud JM: A simple correction for multiple comparisons in interval mapping genome scans. Heredity 2001;87:52–58.
16 Sohns M, Rosenberger A, Bickeböller H: Integration of a priori gene set information into genome-wide association studies. BMC Proceedings 2009;3(suppl 7):S95.
17 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.
18 Setakis E, Stirnadel H, Balding DJ: Logistic regression protects against population structure in genetic association studies. Genome Res 2005;16:290–296.
19 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.
20 Lewinger JP, Conti DV, Baurley JW, Triche TJ, Thomas DC: Hierarchical Bayes prioritization of marker associations from a genome-wide association scan for further investigation. Genet Epidemiol 2007;31:871–882.
21 Wang K, Li M, Bucan M: Pathway-based approaches for analysis of genomewide association studies. Am J Hum Genet 2007;81:1278–1283.
22 Antosh M, Fox D, Helfand SL, Cooper LN, Neretti N: New comparative genomics approach reveals a conserved health span signature across species. Aging 2011;3:576–583.
23 Su Z, Marchini J, Donnelly P: HAPGEN2: simulation of multiple disease SNPs. Bioinformatics 2011;27:2304–2305.
24 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.
25 Bruckner-Tuderman L, Mark K, Pihlajaniemi T, Unsicker K: Cell interactions with the extracellular matrix. Cell Tissue Res 2009;339:1–5.
26 Honjo K, Takahashi KA, Mazda O, Kishida T, Shinya M, Tokunaga D, Arai Y, Inoue A, Hiraoka N, Imanishi J, Kubo T: MDR1a/1b gene silencing enhances drug sensitivity in rat fibroblast-like synoviocytes. J Gene Med 2010;12:219–227.