

# Evaluating Strategies for Marker Ranking in Genome-wide Association Studies of Complex Traits

A. Scherag<sup>1</sup>; J. Hebebrand<sup>2</sup>; H.-E. Wichmann<sup>3,4</sup>; K.-H. Jöckel<sup>1</sup>

<sup>1</sup>Institute for Medical Informatics, Biometry and Epidemiology, University Hospital of Essen, University Duisburg-Essen, Essen, Germany;

<sup>2</sup>Department of Child and Adolescent Psychiatry, University of Duisburg-Essen, Essen, Germany;

<sup>3</sup>Helmholtz Zentrum München, German Research Center for Environmental Health, Institute of Epidemiology, Neuherberg, Germany;

<sup>4</sup>Ludwig-Maximilians University Munich, Institute of Medical Data Management, Biometrics and Epidemiology, Chair of Epidemiology, Munich, Germany

## Keywords

Genome-wide association study, p-value, q-value, FPRP, BFDP

## Summary

**Background:** Genome-wide association studies (GWAS) were highly successful in identifying new susceptibility loci of complex traits. Such studies usually start with genotyping fixed arrays of genetic markers in an initial sample. Out of these markers, some are selected which will be further genotyped in independent samples. Due to the very low a priori probability of a true positive association, the vast majority of all marker signals will turn out to be false positive. Thus, several methods to sort marker data have been proposed which will be evaluated here.

**Objectives:** We compared statistical properties of ranking by p-values, q-values, the False Positive Report Probability (FPRP) and the Bayesian False-Discovery Probability (BFDP).

**Methods:** We performed simulation studies for a genomic region derived from GWAS data sets and calculated descriptive statistics as well as mean square errors with regard to the true marker ranking. Additionally, we applied all measures to a GWAS for early onset extreme obesity superimposing a priori information on candidate genes.

**Results:** Despite the known, more extreme probability results for traditional p-values, we observed that both p-values and the BFDP were more precise in reconstructing the “true” order of the markers in a region. In addition, the BFDP was useful to attenuate unexpected effects at a genome-wide scale.

**Conclusions:** For the purpose of selecting markers from an initial GWAS and within the limits of this study, we recommend either ranking by p-values or the application of a full Bayesian approach for which the BFDP is a first approximation.

polymorphism) markers in an initial sample. The best markers are then followed up, i.e., they are genotyped in independent samples. However, due to the very low a priori probability of a true positive genetic association finding, the vast majority of the marker signals followed-up will turn out to be false positive. Increasing the sample size of the GWAS studies by means of meta-analyses allowing for more stringent p-value cut-offs is one solution to this problem. However, for some hard to sample (complex) phenotypes it may be impossible to build up large consortia to perform meta-analyses and a different solution is required. In addition, it has been shown that some of the true associations will not be among those with the smallest nominal p-values of the initial genome-wide step (e.g. [3] for type II diabetes mellitus). Thus, ranking of the genetic markers according to nominal p-values may not be the best strategy to follow-up the initial genome-wide association findings. Consequently, alternative methods to sort genetic markers have been proposed such as the q-value [4–8], the False Positive Report Probability (FPRP [9, 10]), and the Bayesian False-Discovery Probability (BFDP [11–13]).

In this study we evaluate the statistical properties of these methods jointly in order to explore their degree of overlap with regard to ranking of GWAS association signals.

## Correspondence to:

André Scherag  
Institute for Medical Informatics, Biometry and Epidemiology  
University of Duisburg-Essen  
Hufelandstr. 55  
45122 Essen  
Germany  
E-mail: andre.scherag@uk-essen.de

Methods Inf Med 2010; 49: 632–640

doi: 10.3414/ME09-02-0055

received: December 9, 2009

accepted: February 24, 2010

prepublished: August 5, 2010

## 1. Introduction

Genetic association studies have become the most widely used gene mapping tool for the identification of genetic loci related to complex traits or diseases. Due to technical advances, their realization on a genome-wide

scale became possible since 2005. Ever since, these genome-wide association studies (GWAS) have been very successful in discovering new susceptibility loci for complex traits [1, 2]. Technically, these discoveries start with genotyping or imputing a fixed genome-wide array of SNP (single nucleotide

## 2. Methods

### 2.1 Measures to Rank the Results of GWAS

The standard approach to analyze genomic data of GWAS is to derive marker-wise test

statistics and p-values. The current best practice is the ranking of single marker GWAS results by their unadjusted (nominal) p-values. The markers with the lowest unadjusted (nominal) p-values are then followed-up in independent samples.

To address the problem of false positive claims while not being overly conservative, Benjamini and Hochberg [14] introduced the false discovery rate (FDR) as the expected proportion of false discoveries among all discoveries. The q-value introduced by Storey [5] was developed as an analog to the p-value under the FDR perspective. The q-value of an individual hypothesis test is the minimum positive false discovery rate at which the test may be called “significant” (see e.g. [5] for a formal definition). Note that marker-wise ranking by q-values or p-values should be very similar. In practice, however, estimation problems may lead to ties that will have a practical impact on the order of the ranks.

Interestingly, the ideas of Storey provide a link to Bayesian inference if the proportion of true null hypotheses among all hypotheses tested is viewed as an a priori probability. Addressing a priori probabilities more explicitly, Wacholder et al. [9, 10] proposed the False Positive Reporting Probability (FPRP) as a measure to judge the credibility of a genetic association finding. The FPRP was defined as  $[\alpha(1 - \pi)] / [(1 - \beta) \pi + \alpha(1 - \pi)]$  with  $\alpha$  denoting the significance level of a statistical test,  $(1 - \beta)$  denoting the statistical power to detect a true association and  $\pi$  denoting the (a priori) probability of a true association. To determine the FPRP, it is necessary to specify  $\pi$  whereas both  $\alpha$  and  $(1 - \beta)$  have to be estimated from the data which in turn requires the specification of the true genetic effect under the alternative hypothesis. With regard to the ranking of genetic markers, the results for the FPRP ranking should also be rather similar to those obtained for the rankings by q- or p-values for the case where the expected true genetic effect is kept constant for all markers. Differences may result if the marker-wise standard errors of the log-odds ratio estimates vary strongly.

Criticizing some of the heuristic properties of the FPRP, Wakefield [11–13] introduced the Bayesian False-Discovery Probability (BFDP) as a measure to assess

whether a study is likely to provide a noteworthy association. While FPRP calculations require specifying one particular alternative/true genetic effect, Wakefield argued that both the alternative and the null hypothesis should be modeled as different a priori distributions. To address computational constraints that would arise when performing a full Bayesian analysis for GWAS, he proposed an approximate Bayes factor (ABF) which depends not only on the observed test statistic but also on the statistical power which is related to the minor-allele frequency and the size of the genetic effect. To enable comparisons to the other measures, the BFDP is given by  $\{ABF [\pi / (1 - \pi)]\} / \{1 + ABF [\pi / (1 - \pi)]\}$  where  $[\pi / (1 - \pi)]$  is the prior odds for a true association. Ordering SNPs of a GWAS by ABF or BFDP will in general provide a different ordering than the ranking by p-values. The reason for these expected different rankings is that the observed data is judged under both the null and the alternative a priori distributions.

## 2.2 Simulation Studies, Application and Evaluation of the Marker Ranking Measures

To assess the performance of the proposed methods for marker ranking we conducted simulation studies based on real GWAS data and applied the different methods to a published real data GWAS.

### 2.2.1 Simulation Studies

Real GWAS data were derived from independent, unrelated individuals of Central European descent. In particular, these were 1644 (831 females) individuals from the KORA S3/F3 cohort genotyped by the Affymetrix GeneChip Human Mapping 500K Array Set, 380 (198 females) individuals from the Heinz Nixdorf Recall study genotyped by Illumina BeadChip Human-Hap550 and 1736 (949 females) independent individuals from the Essen obesity study genotyped by the Affymetrix Genome-Wide Human SNP Array 6.0. All three data sets (3760 individuals; 1978 females) were merged for a randomly selected 187 kb genomic region on chromo-

some 10 for which 29 SNPs of 170 HapMap SNPs in the CEU release 22 panel of 60 individuals (<http://www.hapmap.org>) were available for all three data sets. A central marker was chosen as simulated disease-related SNP and we calculated pairwise linkage disequilibrium (LD;  $r^2$ ) between this SNP and all other 28 other SNPs. We derived three classes “high LD” ( $r^2 > 0.97$ ; 3 SNPs), “intermediate LD” ( $r^2 \leq 0.97$  and  $r^2 > 0.75$ ; 6 SNPs) and “low to no LD” ( $r^2 \leq 0.75$ ; 19 SNPs). Next the combined data set was stratified into three data sets based on the genotype status of an individual at the disease-related SNP. Finally, individuals were drawn from these three data sets with probabilities given by the multinomial distribution based on the underlying disease model scenario [15, 16]. Altogether, we explored  $3 \times 3 \times 2 = 18$  disease model scenarios. We investigated dominant, (log-)additive (multiplicative) or recessive genetic models for the disease-related variant as defined by the (odds ratio) effects  $(\psi_1, \psi_2) = (1.5^2, 1.5^2)$ ,  $(1.5, 1.5^2)$  and  $(1, 1.5^2)$ . In addition, we varied the minor allele frequencies (MAF = 5, 10 and 30%). Finally, the prevalence of the disease was either 1% or 20%. For each simulation replicate, 1000 cases and 1000 controls were generated while 10,000 replicates were created for each scenario.

All simulations were run using R 2.8.1 and the function `glm` to derive p-values from the Wald-type statistic of the logistic regression under a log-additive genetic model. To derive q-values, we used the R-package `q-value 1.1.1` with default settings. For FPRP and BFDP we used the code provided by the authors with an a priori probability for a true association of  $\pi = 0.01$  for a true log-additive (multiplicative) genetic effect of  $(\psi_1, \psi_2) = (1.5, 1.5^2)$ . Note that alternative choices of the a priori probability can be easily assessed by using Bayes’ theorem as they only have an impact on the prior odds for both FPRP and the BFDP. Finally, for BFDP it was assumed that the distribution under the alternative hypothesis centered around this true association signal with a probability mass of 60%.

### 2.2.2 GWAS Application

To demonstrate the impact of the proposed methods on a genome-wide scale, we ap-

**Table 1** Simulation study medians of the distributions<sup>1</sup> for unadjusted, nominal p-values and the q-values. For sake of comparison all values are  $-\log_{10}$ -transformed. Bold printing indicates scenarios with median values  $< 5 \times 10^{-8}$  (a commonly applied threshold used to claim genome-wide significance).

Rank measure	Genetic model $\psi_1, \psi_2$	MAF <sup>2</sup> (%)	Prevalence (%)	Disease-related SNP	High LD	Intermediate LD	Low to no LD	
$-\log_{10}(\text{p-values})$	$1.5^2, 1.5^2$	5	1	<b>8.89</b>	<b>8.48</b>	<b>8.21</b>	1.03	
			20	<b>13.60</b>	<b>13.00</b>	<b>12.60</b>	1.37	
		10	1	<b>13.50</b>	<b>12.80</b>	<b>12.50</b>	2.13	
			20	<b>20.70</b>	<b>19.60</b>	<b>19.20</b>	2.94	
		30	1	<b>14.00</b>	<b>13.10</b>	<b>12.80</b>	3.82	
			20	<b>20.80</b>	<b>19.50</b>	<b>19.10</b>	5.45	
		$1.5, 1.5^2$	5	1	2.67	2.56	2.47	0.43
				20	3.90	3.74	3.61	0.52
			10	1	4.49	4.28	4.19	0.86
				20	6.75	6.44	6.28	1.12
			30	1	<b>8.88</b>	<b>8.37</b>	<b>8.18</b>	2.61
				20	<b>13.30</b>	<b>12.60</b>	<b>12.30</b>	3.64
	$1, 1.5^2$	5	1	0.33	0.33	0.32	0.17	
			20	0.35	0.35	0.35	0.17	
		10	1	0.62	0.60	0.59	0.25	
			20	0.78	0.76	0.74	0.28	
		30	1	5.30	5.06	4.91	1.68	
			20	<b>7.91</b>	<b>7.56</b>	<b>7.34</b>	2.31	
	$-\log_{10}(\text{q-values})$	$1.5^2, 1.5^2$	5	1	<b>8.85</b>	<b>8.55</b>	<b>8.48</b>	1.94
				20	<b>13.60</b>	<b>13.20</b>	<b>13.00</b>	2.34
			10	1	<b>13.60</b>	<b>13.10</b>	<b>12.90</b>	3.11
				20	<b>20.70</b>	<b>19.90</b>	<b>19.60</b>	3.94
			30	1	<b>14.00</b>	<b>13.40</b>	<b>13.30</b>	4.74
				20	<b>20.90</b>	<b>19.90</b>	<b>19.70</b>	6.42
$1.5, 1.5^2$			5	1	2.47	2.40	2.41	0.89
				20	3.69	3.61	3.60	1.13
			10	1	4.39	4.28	4.27	1.58
				20	6.71	6.53	6.50	2.00
			30	1	<b>8.79</b>	<b>8.52</b>	<b>8.45</b>	3.46
				20	<b>13.30</b>	<b>12.90</b>	<b>12.70</b>	4.60
$1, 1.5^2$		5	1	0.13	0.13	0.13	0.03	
			20	0.15	0.14	0.15	0.04	
		10	1	0.50	0.49	0.50	0.27	
			20	0.66	0.65	0.65	0.36	
		30	1	5.30	5.17	5.15	2.47	
			20	<b>7.94</b>	<b>7.72</b>	<b>7.64</b>	3.17	

<sup>1</sup> the smallest values were 0 for the recessive model and under "low to no LD" whereas the maximum values for "low to no LD" were 35.6 (p-values), 36.0 (q-values), 11.9 (FPRP) and 30.8 (BFDP) indicating that none of these SNPs was completely independent of the disease-related SNP; this explanation was supported by performing coalescent-based simulation studies which showed that the correct level  $\alpha$  was met under the global null hypothesis (data not shown).

<sup>2</sup> minor allele frequencies

plied all the methods to a GWAS case-control data set genotyped by Affymetrix Genome-Wide Human SNP Array 6.0. This sample is a subsample from Hinney et al. [17] including 453 extremely obese cases and 435 healthy lean control individuals. Of the 909,622 SNPs which are available on this array, 542,900 autosomal SNPs passed the quality control criteria (as described in [17]) and were subsequently analyzed by logistic regression (log-additive genetic model) as implemented in PLINK v1.05 [18]. In this GWAS analysis, we marked candidate gene regions for obesity to explore the impact of different choices of a priori parameters within or out of these candidate gene regions. The information on candidate genes for obesity was extracted from the “Obesity Gene Map Database” (<http://obesitygene.pbrcc.edu>) and merged with the physical position information from the UCSC browser for all RefSeq genes (06/24/2008 (hg18)). Genes for which no definite physical position was available were rejected. A total of 112 autosomal candidate genes were extracted. Finally, we defined a candidate gene region as coding region  $\pm 100\text{kb}$  and observed that a total of 5256 SNPs were within these regions.

### 2.2.3 Evaluation Methods to Compare Marker-ranking Measures

To compare the performance of the marker-ranking measures we first summarized the simulation distributions descriptively stratified by simulation scenario and marker category (disease-related SNP, high, intermediate and low to no LD SNPs). Secondly, we determined “true” ranks from 1 to 29 for the 29 SNPs of the simulation studies. To derive “true” ranks, we ordered the SNPs according to their LD with the disease-related SNP and for those SNPs with  $r^2 = 0$  the ranking was by physical distance relative to the disease-related SNP. Next, we ranked each simulation replicate by each marker-ranking measure. Afterwards, we calculated the mean squared error to the “true” ranks for each of the 29 SNPs and plotted the mean square error against the “true” ranks for each simulation scenario. Thirdly, for the GWAS application, we explored two scenarios for the

BFDP assuming that the distribution under the alternative hypothesis was centered around the true association with a probability of 60% as in the simulation studies (BFDP I) or 90% (BFDP II). To investigate the impact of these different modifications on the GWAS marker ranking, we ranked the 542,900 autosomal GWAS SNPs by unadjusted p-values as the standard procedure. In addition, we ranked the SNPs based on q-values, FPRP, BFDP I and BFDP II (with the different a priori probabilities  $\pi = 0.01$  in and  $\pi = 0.0001$  out of candidate gene regions) and calculated rank differences in comparison to the p-value ranking. Next, we subtracted  $(542,900 + 1)/2$  from each rank to standardize all ranks. Taking the standardized ranks based on unadjusted p-values and the standardized ranks based on one of the four alternative measures, we then applied the Bland-Altman procedure [19]. Bland-Altman procedure has been proposed as graphical display showing the agreement between two methods assessing the same variable (e.g., diastolic blood pressure in the same individual at the same time measured by two devices). The graph displays means vs. difference of the two measurements by a dot for each individual. Here we replaced the individual by a particular SNP and the measurements by standardized ranks. If the rankings based on q-values, FPRP, BFDP I and BFDP II do not deviate from the p-value ranking, all plotted dots, each representing one SNP, should fall on the x-axis at  $y = 0$ . If there is variance in the rankings which is unrelated to the ranking position, we expect a homoscedastic scattering centered round the x-axis at  $y = 0$ . Otherwise, the Bland-Altman-type plot does indicate that differences in ranking depend on the rank.

## 3. Results

### 3.1 Simulation Studies

All methods for marker ranking allow for detecting the disease-related SNP in most of the replicates in the best case scenario (dominant genetic model with a common variant and a common trait). On the other hand, none of the methods can cope with

the worst case scenario (recessive genetic model with a low frequency variant and a less frequent trait) if only 1000 case-control pairs are sampled. Between these extreme scenarios, however, the ranking methods differed. As expected we observed the largest  $-\log_{10}$ -transformed values among the p-values followed by the q-values and the BFDP and FPRP (despite the large a priori probability of  $\pi = 0.01$ ).

For observations under the alternative hypothesis, we took median values larger than  $-\log_{10} = 7.3$  corresponding to a p-value  $< 5 \times 10^{-8}$  as an arbitrary but theoretically justified (e.g. [20]) genome-wide cut-off to compare the ranking methods. We observed that all the scenarios simulated under the dominant model were detectable by p- and q-values whereas the case of a low-frequency variant (MAF = 5%) and a rare trait (prevalence = 1%) under a dominant model was undetectable by FPRP and BFDP even for a large a priori probability of  $\pi = 0.01$ . Moreover, lowering the median cut-off to e.g.  $-\log_{10} = 5$  did not alter this result and of course detectability declined with weaker LD to the disease-related SNP. Interestingly, the FPRP was also worse than the other three measures for the common trait scenario and a low-frequency variant (MAF = 5%) with dominant genetic effect. For the simulations under the (log-)additive and the recessive genetic model, the highest association signals were observable for the common disease (prevalence = 20%) and common disease-related SNP (MAF = 30%) scenario. For the recessive genetic model, only the median of the  $-\log_{10}$ -transformed p- and q-values were larger than 7.3, followed by the FPRP and the weakest signals for the BFDP. A similar phenomenon was observable for the (log-)additive genetic model.

To describe the accuracy with which the methods are able to pick up the “true” ordering of the 29 SNPs, we calculated mean squared errors (MSEs; ► Fig 1). In general, the MSEs tended to be larger for larger ranks, i.e., SNPs that are in decreasingly smaller LD with the disease-related SNP. Overall larger MSEs were observed for the q-values and FPRP rankings relative to the rankings by p-values or BFDP. Interestingly, the BFDP ranking resulted in MSEs

**Table 2** Simulation study medians of the distributions<sup>1</sup> for the False Positive Report Probability (FPRP) and the Bayesian False-Discovery Probability (BFDP). For sake of comparison all values are  $-\log_{10}$ -transformed. Bold printing indicates scenarios with median values  $< 5 \times 10^{-8}$  (a commonly applied threshold used to claim genome-wide significance).

Rank measure	Genetic model $\psi_1, \psi_2$	MAF <sup>2</sup> (%)	Prevalence (%)	Disease-related SNP	High LD	Intermediate LD	Low to no LD
$-\log_{10}(\text{FPRP})$ $\pi = 0.01$	1.5 <sup>2</sup> , 1.5 <sup>2</sup>	5	1	4.08	3.80	3.72	0.05
			20	5.32	5.02	5.01	0.09
		10	1	<b>7.49</b>	7.11	7.02	0.33
			20	<b>8.25</b>	<b>7.86</b>	<b>7.89</b>	0.59
		30	1	<b>9.54</b>	<b>9.01</b>	<b>8.91</b>	1.68
			20	<b>10.80</b>	<b>10.40</b>	<b>10.40</b>	2.55
	1.5, 1.5 <sup>2</sup>	5	1	0.52	0.44	0.41	0.01
			20	1.18	1.05	0.99	0.01
		10	1	2.18	1.98	1.91	0.03
			20	3.77	3.52	3.40	0.05
		30	1	6.53	6.08	5.90	0.71
			20	<b>9.40</b>	<b>8.95</b>	<b>8.80</b>	1.55
	1, 1.5 <sup>2</sup>	5	1	0.01	0.01	0.01	0.01
			20	0.01	0.01	0.01	0.01
		10	1	0.02	0.02	0.02	0.01
			20	0.03	0.02	0.02	0.01
		30	1	3.28	3.05	2.90	0.17
			20	5.74	5.40	5.19	0.48
$-\log_{10}(\text{BFDP})$ $\pi = 0.01$	1.5 <sup>2</sup> , 1.5 <sup>2</sup>	5	1	4.88	4.50	4.25	0.00
			20	<b>9.06</b>	<b>8.46</b>	<b>8.14</b>	0.00
		10	1	<b>9.32</b>	<b>8.68</b>	<b>8.38</b>	0.02
			20	<b>16.00</b>	<b>15.10</b>	<b>14.60</b>	0.09
		30	1	<b>9.86</b>	<b>9.02</b>	<b>8.75</b>	0.42
			20	<b>16.50</b>	<b>15.20</b>	<b>14.80</b>	1.73
	1.5, 1.5 <sup>2</sup>	5	1	0.08	0.07	0.06	0.00
			20	0.54	0.45	0.38	0.00
		10	1	0.96	0.80	0.73	0.00
			20	2.99	2.70	2.55	0.00
		30	1	4.98	4.50	4.31	0.05
			20	<b>9.21</b>	<b>8.51</b>	<b>8.25</b>	0.32
	1, 1.5 <sup>2</sup>	5	1	0.00	0.00	0.00	0.00
			20	0.00	0.00	0.00	0.00
		10	1	0.00	0.00	0.00	0.00
			20	0.00	0.00	0.00	0.00
		30	1	1.57	1.37	1.24	0.01
			20	4.05	3.72	3.51	0.03

<sup>1</sup> the smallest values were 0 for the recessive model and under "low to no LD" whereas the maximum values for "low to no LD" were 35.6 (p-values), 36.0 (q-values), 11.9 (FPRP) and 30.8 (BFDP) indicating that none of these SNPs was completely independent of the disease-related SNP; this explanation was supported by performing coalescent-based simulation studies which showed that the correct level  $\alpha$  was met under the global null hypothesis (data not shown)

<sup>2</sup> minor allele frequencies

comparable or even smaller than those observed for the p-value ranking. In particular, this was true for the scenarios under the recessive genetic model and minor allele frequencies of 5% and 10% for the disease-related SNP. In these scenarios the p-value ranking led to MSEs in the range of 115 to 173 for the disease-related SNP. For the same scenarios, the range of the MSEs for the BFDP ranking was 74 to 84, i.e., about half of the size as those for the p-value rankings.

### 3.2 GWAS Application

To illustrate the impact of each method on a genome-wide scale, we illustrated the differences of ranking by Bland-Altman-type plots using standardized ranks. In all these plots, SNPs in candidate gene regions are again highlighted in black and were given a larger a priori weight ( $\pi = 0.01$  in contrast to  $\pi = 0.0001$ ) for FPRP and BFDP.

The rankings according to q-values led to rank differences in comparison to the p-value rankings in the range of  $\pm 500$  (► Fig 2). These differences were largely independent of the rank. For the rankings according to FPRP values we observed much larger rank differences in comparison to the p-value rankings ( $\pm 50,000$ ) which were largely independent of the rank as well (► Fig 2). Finally, for both rankings according to BFDP I or BFDP II we observed the largest rank differences in comparison to the p-value rankings so that the y-axis of the Bland-Altman-type plots had to be truncated to the range of  $\pm 50,000$  (► Fig. 3). Moreover, larger differences were observable for higher ranks if the SNPs were ranked by BFDP.

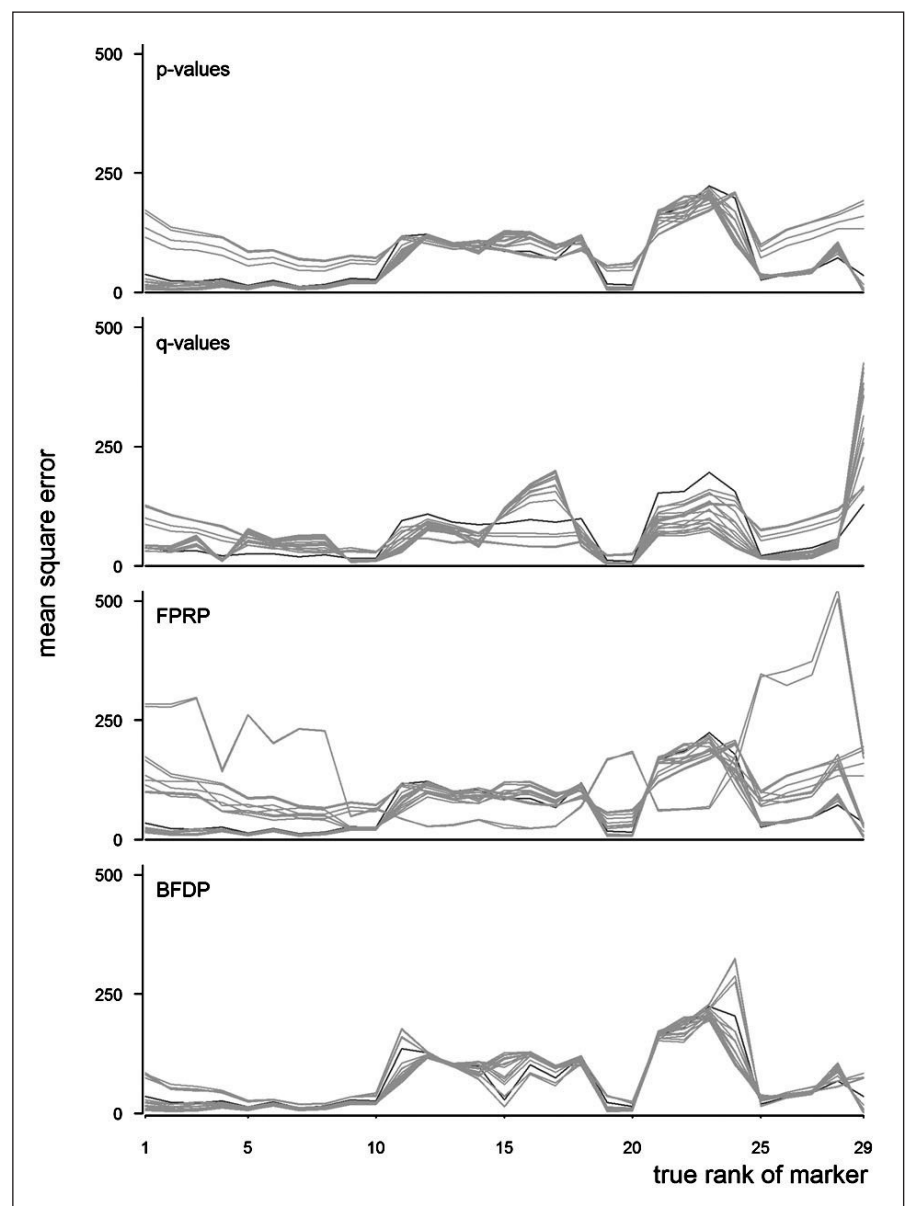
With regard to the candidate gene regions, no clear pattern emerged for q-value ranking. The lines which are observable in the plot result from ties due to the same q-values for multiple SNPs. For FPRP and BFDP the differential a priori weights led to major differences in comparison to the ranking by p-values. Furthermore, we observed a tendency for better p-values ranks of SNPs with the lowest mean ranks in the Bland-Altman-type plots of BFDP I, BFDP II and FPRP. An examination of these SNPs revealed that they are character-

ized by estimated genetic effect sizes which are stronger than the a priori expected effect sizes particular if their estimated standard error was relatively small.

## 4. Discussion

In this report we investigated statistical methods for ranking genetic markers in GWAS designed to identify genetic loci related to or causal for complex traits or dis-

eases. These studies typically face the situation of a very low a priori probability for a true positive genetic association finding – in contrast to e.g. gene expression experiments (e.g. [21]). As current best practice, the single marker GWAS results are ranked by their unadjusted (nominal) p-values and those with the lowest unadjusted p-values are followed-up in independent samples. Alternative methods to sort genetic markers, however, may lead to other or even superior rankings. The results of



**Fig. 1** Mean squared errors from each rank measure relative to the “true” ranks for each of the 29 SNPs from the simulation studies for each of 18 simulation scenarios with each scenario displayed by one line. The simulated disease-related SNP can be found at the left side of the figure.

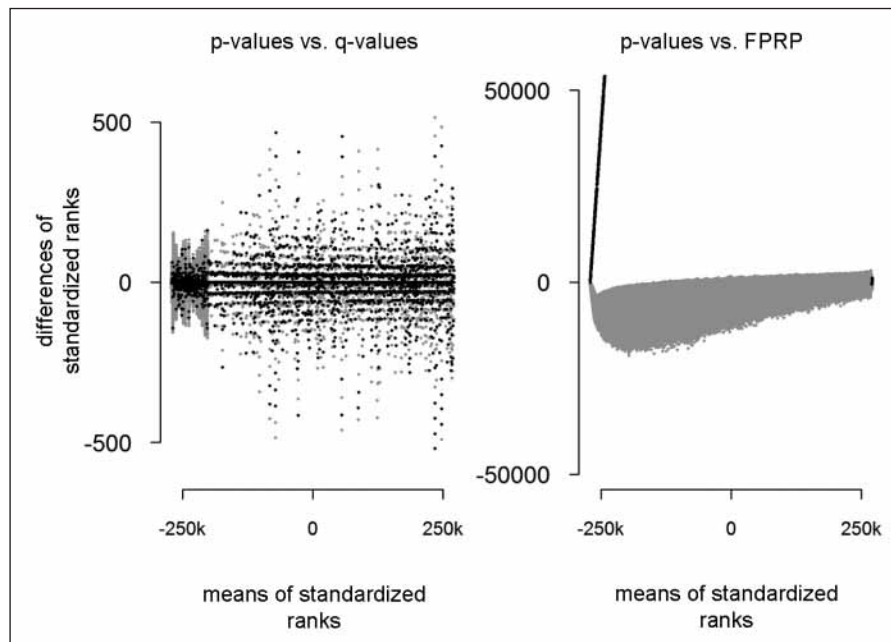
our simulation studies based on real GWAS data for a genomic region showed that p-value association signals were more extreme than all other probability measures. However, one has to keep in mind that the

p-value is the conditional probability of the observed data given that the null hypothesis is true and that the alternative measures explored aim at more directly assessing the probability of the null hypothesis itself

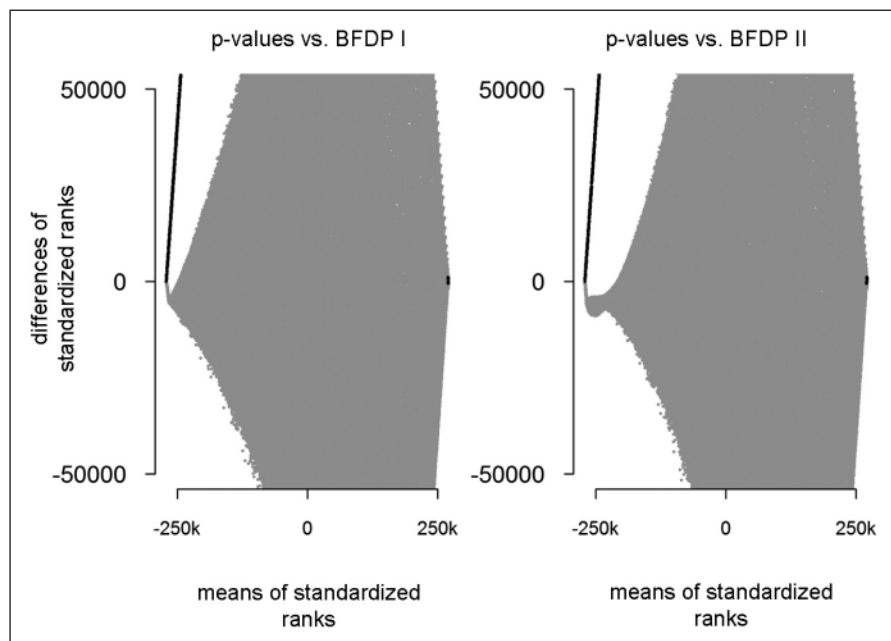
(FPRP and BFDP) or the probability of false associations among all hypotheses called “significant” in the case of the q-value. Thus, they measure the association signal on a different scale with a different purpose. Thus, we omitted formal comparisons of type I error rates and powers as these properties would require a definition as base for a comparison of the new marker ranking measures first. Consequently, we shifted to a descriptive level and observed that the association signals for the alternative measures were in general less extreme than those observed using unadjusted nominal p-values which was expected by construction. Nevertheless, it is a useful property of the alternative measures as it does lower the expectations with regard to the outcomes of a GWAS study. Moreover, with a simulated sample of 1000 case-control pairs, all methods were unable to detect recessive genetic effects in particular if a low-frequency variant and a rare disease were simulated. For the simulated dominant genetic model none of the methods had problems to discover them at least if the minor allele frequency was >10%. For these scenarios, however, the genetic effect sizes may be regarded as too large for a complex trait. For the (log-)additive genetic model again only the more common variants were detectable.

Taking detectability to a practical level, assume that an investigator is restricted to a follow-up of the best 40 SNPs of a GWAS. The number 40 is roughly the limit for one plex for multiplex genotyping in a single well (<http://www.sequenom.com>). In this extreme but rather frequent scenario, one SNP from the candidate gene regions would have been selected by all ranking measures. With a larger a priori weight given to the candidate gene regions, 18 SNPs were selected by FPRP and 17 or 25 SNPs for BFDP (more SNPs were selected for a more peaked alternative distribution).

To generate more general insights on this selection behavior, we investigated differences in marker rankings and observed that the p-value ranking and the BFDP ranking were better suited to capture the “true” ranking within the genomic region. Rather unexpected, the rankings by q-value differed from those of the p-values. The reason



**Fig. 2** Bland-Altman-type plots for q-values and FPRP. The difference between standardized ranks based on q-value or FPRP and standardized p-value ranks (y-axis) is plotted against the mean of the two standardized ranking measures (x-axis) for each SNP.



**Fig. 3** Bland-Altman-type plots for BFDP I and BFDP II. The difference between standardized ranks based on BFDP I (with a more flat alternative distribution) or BFDP II (with a more peaked alternative distribution) and standardized p-value ranks (y-axis) is plotted against the mean of the two standardized ranking measures (x-axis) for each SNP.

for this observation is the generation of the same  $q$ -values for very similar  $p$ -values which was also observable at the genome-wide scale. This property combined with the theoretical limitations of  $q$ -value for correlated data and data sets with a low a priori probability for true association signals, we recommend not to use this method for GWAS SNP ranking. Furthermore, we cannot recommend the use of the FPRP either. Its performance to recover the “true” ranking within simulated genomic region was relatively poor. In comparison to the conceptually closer BFDP the FPRP was inferior to detect the true association signals at a similar cut-off level. Moreover, apart from the arguments based on this work, the FPRP has theoretical weaknesses [11, 22]. Finally, regarding the BFDP, a definite recommendation based on this work alone is not possible. In our simulation studies, the BFDP results were similar to the  $p$ -value results though being conceptually different. In the GWAS application, it became clear that the marker rankings may differ strongly between  $p$ -values and BFDP especially the higher the mean rank. Moreover, the GWAS application also indicated that the BFDP seems to be a valuable tool to attenuate some of the signals while other signals can be amplified using larger a priori probabilities. Basically, this reflects the challenges and chances of all Bayesian approaches (e.g. [12, 23]) which become more important the more detailed knowledge on the genetics of complex traits and high-performance computation facilities are available. Of course, this work also has limitations and weaknesses. With regard to the simulation studies it would have been better to generate replicates at a genome-wide scale and at a higher marker density. Due to computational constraints, however, we decided to investigate a genomic region only. In addition, we decided to use called genotypes instead of imputed genotypes as we did not want to confuse properties of any imputation algorithm with properties of the methods for ranking. To highlight the effects of the methods for ranking at a genome-wide scale, we decided to present a GWAS application. Moreover, genome-wide samples should be drawn from a relatively large population which requires well-organized national and international

biobanks (e.g. [24]) to adequately cover the normal human genetic diversity. Here, we only had access to GWAS data of 3760 individuals of Central European descent genotyped for relatively common variants on different genotyping platforms.

In sum, we have compared alternative measures to rank the single marker results of a GWAS using simulation studies and a real data application. The purpose of this ranking is to derive a list of markers which is enriched for true associations. Based on this study, none of the methods is suitable for fulfilling this purpose if the marker list is limited to a few hundred markers only. Thus, we either recommend the traditional sorting by the smallest nominal  $p$ -values or the application of a full Bayesian approach for which the explored Bayesian False-Discovery Probability is a first approximation. The False Positive Report Probability and the  $q$ -values, however, are the inferior choices compared to the other two methods. In either case and particular for given budget constraints, a formal and cost-optimized but flexible statistical procedure appropriately addressing the sequential testing (e.g. [25, 26]) should be preferred.

### Acknowledgments

We thank all the participants of this study. We also thank A. Hinney (for the Essen obesity study), S. Moebus, R. Erbel (for the Heinz Nixdorf Recall Study) and T. Illig (for the KORA study).

### IRB Approval

The study, including the protocols for subject recruitment and assessment, the informed consent for participants, was reviewed and approved by all local IRB boards and was carried out in accordance with the Declaration of Helsinki.

### Funding

This work was supported by grants from the German Ministry of Education and Research (BMBF: 01GI0823 and NGFN<sup>plus</sup>: 01GS0820). We thank the Heinz Nixdorf Foundation (chairman: G. Schmidt) for the generous support of the Heinz Nixdorf Recall Study. The KORA Augsburg studies were financed by the Helmholtz Zentrum München – Research Center for Environ-

ment and Health, Neuherberg, Germany and supported by grants from the BMBF and the Munich Center of Health Sciences (MC Health) as part of LMU innovative. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

## References

1. Altshuler D, Daly MJ, Lander ES. Genetic Mapping in Human Disease. *Science* 2008; 322 (5903): 881–888.
2. Hirschhorn JN. Genomewide association studies – illuminating biologic pathways. *N Engl J Med* 2009; 360 (17): 1699–1701.
3. Frayling TM. Genome-wide association studies provide new insights into type 2 diabetes aetiology. *Nat Rev Genet* 2007; 8 (9): 657–662.
4. Storey JD. A direct approach to false discovery rates. *Journal of the Royal Statistical Society Series B-Statistical Methodology* 2002; 64: 479–498.
5. Storey JD. The positive false discovery rate: A Bayesian interpretation and the  $q$ -value. *Annals of Statistics* 2003; 31 (6): 2013–2035.
6. Storey JD, Tibshirani R. Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences of the United States of America* 2003; 100 (16): 9440–9445.
7. Storey JD. The optimal discovery procedure: a new approach to simultaneous significance testing. *Journal of the Royal Statistical Society Series B – Statistical Methodology* 2007; 69: 347–368.
8. Storey JD, Taylor JE, Siegmund D. Strong Control, Conservative Point Estimation and Simultaneous Conservative Consistency of False Discovery Rates: A Unified Approach. *Journal of the Royal Statistical Society Series B – Methodological* 2004; 66 (1): 187–205.
9. Wacholder S, Chanock S, Garcia-Closas M, Katki HA, El Ghormli L, Rothman N. Re: Assessing the probability that a positive report is false: An approach for molecular epidemiology studies – Response. *Journal of the National Cancer Institute* 2004; 96 (22): 1722–1723.
10. Wacholder S, Chanock S, Garcia-Closas M, El Ghormli L, Rothman N. Assessing the probability that a positive report is false: An approach for molecular epidemiology studies. *Journal of the National Cancer Institute* 2004; 96 (6): 434–442.
11. Wakefield J. A Bayesian measure of the probability of false discovery in genetic epidemiology studies. *Am J Hum Genet* 2007; 81 (2): 208–227. Erratum in: *Am J Hum Genet* 2008; 83 (3): 424.
12. Wakefield J. Reporting and interpretation in genome-wide association studies. *International Journal of Epidemiology* 2008; 37 (3): 641–653.
13. Wakefield J. Bayes factors for genome-wide association studies: comparison with  $P$ -values. *Genet Epidemiol* 2009; 33 (1): 79–86.
14. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate – A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society Series B – Methodological* 1995; 57 (1): 289–300.

15. Freidlin B, Zheng G, Li Z, Gastwirth JL. Trend tests for case-control studies of genetic markers: power, sample size and robustness. *Hum Hered* 2002; 53 (3): 146–152.
16. Slager SL, Schaid DJ. Case-control studies of genetic markers: power and sample size approximations for Armitage's test for trend. *Hum Hered* 2001; 52 (3): 149–153.
17. Scherag A, Dina C, Hinney A, Vatin V, Scherag S, Vogel CI, Müller TD, Grallert H, Wichmann HE, Balkau B, Heude B, Jarvelin MR, Hartikainen AL, Levy-Marchal C, Weill J, Delplanque J, Körner A, Kiess W, Kovacs P, Rayner NW, Prokopenko I, McCarthy MI, Schäfer H, Jarick I, Boeing H, Fisher E, Reinehr T, Heinrich J, Rzehak P, Berdel D, Borte M, Biebertmann H, Krude H, Roszkopf D, Rimmbach C, Rief W, Fromme T, Klingenspor M, Schürmann A, Schulz N, Nöthen MM, Mühleisen TW, Erbel R, Jöckel KH, Moebus S, Boes T, Illig T, Froguel P, Hebebrand J, Meyre D. Two new loci for body-weight regulation identified in a joint analysis of genome-wide association studies for early-onset extreme obesity in French and German study groups. *PLoS Genet* 2010; 6 (4): e1000916.
18. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 2007; 81 (3): 559–575.
19. Bland JM, Altman DG. Measuring agreement in method comparison studies. *Statistical Methods in Medical Research* 1999; 8 (2): 135–160.
20. Hoggart CJ, Clark TG, De IM, Whittaker JC, Balding DJ. Genome-wide significance for dense SNP and resequencing data. *Genet Epidemiol* 2008; 32 (2): 179–185.
21. Reipsilber D, Mansmann U, Brunner E, Ziegler A. Tutorial on microarray gene expression experiments. An introduction. *Methods Inf Med* 2005; 44 (3): 392–399.
22. Lucke JF. A critique of the false-positive report probability. *Genet Epidemiol* 2009; 33 (2): 145–150.
23. Stephens M, Balding DJ. Bayesian statistical methods for genetic association studies. *Nat Rev Genet* 2009; 10 (10): 681–690.
24. Wichmann HE. Genetic epidemiology in Germany – from biobanking to genetic statistics. *Methods Inf Med* 2005; 44 (4): 584–589.
25. Pahl R, Schäfer H, Müller HH. Optimal multistage designs – a general framework for efficient genome-wide association studies. *Biostatistics* 2009; 10 (2): 297–309.
26. Scherag A, Hebebrand J, Schäfer H, Müller HH. Flexible designs for genomewide association studies. *Biometrics* 2009; 65 (3): 815–821.