Original Paper

Human Heredity

Hum Hered 2014;78:164–178
DOI: 10.1159/000368676

Received: June 10, 2014
Accepted after revision: September 15, 2014
Published online: December 10, 2014

Rare Variant Testing of Imputed Data: An Analysis Pipeline Typified

Dmitriy Drichel[a] Christine Herold[a, b] André Lacour[a] Alfredo Ramirez[c] Frank Jessen[c] Wolfgang Maier[c] Markus M. Noethen[d, e] Markus Leber[a] Tatsiana Vaitsiakhovich[b] Tim Becker[a, b]

[a]German Center for Neurodegenerative Diseases (DZNE), [b]Institute for Medical Biometry, Informatics and Epidemiology, [c]Department of Psychiatry and Psychotherapy, and [d]Institute of Human Genetics, University of Bonn, and [e]Department of Genomics, Life and Brain Center, Bonn, Germany

Key Words
Rare variant analysis · Genotype imputation · Variable threshold · Alzheimer's disease

Abstract
Important methodological advancements in rare variant association testing have been made recently, among them collapsing tests, kernel methods and the variable threshold (VT) technique. Typically, rare variants from a region of interest are tested for association as a group ('bin'). Rare variant studies are already routinely performed as whole-exome sequencing studies. As an alternative approach, we propose a pipeline for rare variant analysis of imputed data and develop respective quality control criteria. We provide suggestions for the choice and construction of analysis bins in whole-genome application and support the analysis with implementations of standard burden tests (COLL, CMAT) in our INTERSNP-RARE software. In addition, three rare variant regression tests (REG, FRACREG and COLLREG) are implemented. All tests are accompanied with the VT approach which optimizes the definition of 'rareness'. We integrate kernel tests as implemented in SKAT/SKAT-O into the suggested strategies. Then, we apply our analysis scheme to a genome-wide association study of Alzheimer's disease. Further, we show that our pipeline leads to valid significance testing procedures with controlled type I error rates. Strong association signals surrounding the known *APOE* locus demonstrate statistical power. In addition, we highlight several suggestive rare variant association findings for follow-up studies, including genomic regions overlapping *MCPH1*, *MED18* and *NOTCH3*. In summary, we describe and support a straightforward and cost-efficient rare variant analysis pipeline for imputed data and demonstrate its feasibility and validity. The strategy can complement rare variant studies with next generation sequencing data.

Introduction

A large portion of heritability of human diseases remains unexplained by the 'common disease, common variant' hypothesis [1]. Rare variants continue to generate attention as candidates responsible for a fraction of the so-called 'missing heritability' [2]. Theoretical considerations, large reference data sets generated by the 1,000 Genomes Project [3], for instance, and databases of genome-wide association study (GWAS) findings as well as computational simulations [4, 5] support the hypothesis that rare variants are numerous and potentially more deleterious than common variation.

KARGER

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

0001–5652/14/0784–0164$39.50/0

Tim Becker
German Center for Neurodegenerative Diseases (DZNE)
Sigmund-Freud-Str. 25
DE–53127 Bonn (Germany)
E-Mail tim.becker@dzne.de

The study of methods for detecting association of rare variants with complex diseases has made substantial progress over the last few years [6–12]. This development was accompanied by rapid improvements in technology and reference data availability, which allow to increase genotyping density either directly or by imputation [13–15]. Rare variant association tests have been implemented in various software packages [11, 16–18]. Some methods have explicitly been optimized for data generated from next generation sequencing (NGS). Current NGS studies are typically of moderate sample size and often focus on the exomes rather than the whole genome. Exome sequencing is undoubtedly a cost-efficient strategy, but the motivation for considering non-exomic regions is, nevertheless, manifold. While protein-coding sequences cover only approximately 1.5% of the human genome, comparative studies show that a large portion of genomic regions in the human genome has undergone purifying selection, with a recent estimate of at least 5.5% [19]. Recent data from the ENCODE project suggests that elements interacting with biochemical processes cover at least about 80% of the human genome, with different RNA types covering 62% of the genome [20]. Results from GWAS show that a large proportion of discovered associations lie outside of exomes (88% according to Hindorff et al. [21]), which supports the hypothesis that similar portions might be expected for rare variant associations.

Given the fact that genome-wide coverage in NGS is not yet affordable for all research groups, we introduce an analysis pipeline for rare SNP variants obtained by genotype imputation. Imputation of rare variants from GWAS panels is a viable strategy for moderately rare variants already, although increasingly large reference panels are required for accurate imputation of very rare alleles [22, 23]. Therefore, our analysis will be limited to those SNPs which are imputed at high quality. As a consequence, our approach will focus on rare variants, but will not cover very rare variants which are disease specific and often not contained in public databases. In this sense, our strategy has to be viewed as an intermediate improvement until large-scale NGS studies have become a standard. In the following, we refer to variants with a MAF <0.05 as 'rare variants' and those with a MAF <0.001 as 'very rare variants'.

In contrast to the development of association testing methods, the discussion of two essential aspects of analysis is often neglected. First, there is no consensus on how to define 'rareness' in rare variant analysis, or even how to estimate a suitable frequency threshold given a data set to be analyzed. Often, a MAF threshold (MAF_T) defining 'rare' and 'common' variants is conveniently, but somewhat arbitrarily, chosen to be 0.05 or 0.01, for instance, while other researchers will consider singletons to be the typical rare variant. Differential weighting of individual variants according to their MAF has been proposed to alleviate this problem [9]. The variable threshold (VT) approach [24] is an important methodological improvement that reduces the choice of MAF_T to a choice of the largest MAF_T that shall be considered. All possible threshold frequencies are subsequently tested to determine the optimal MAF. Permutation of affection status is used for multiple testing correction.

The second problem is the proper assignment of sets of SNPs that shall be analyzed together ('binning'). Given NGS exome data, genes as units of analysis are a most evident choice in rare variant testing. However, the power is limited if the chosen bins do not exactly correspond to associated regions. In one scenario, the tested gene could contain a truly associated region, but a bin that extends beyond the limits of the region inevitably captures noise that reduces the power up to the point where no evidence for association can be detected. In another scenario, the selected bin only partially covers the associated region.

Until now, most of the available software relies on the user to provide both the definitions of analysis bin as well as that of MAF_T. In order to address the requirements of rare variant testing and to support genome-wide application, we have developed the analysis software INTERSNP-RARE. It implements various binning strategies and combines existing rare variant methods with the VT approach [24], among them the collapsing test (COLL) [6] and the cumulative minor allele test (CMAT) [7]. Furthermore, we implemented regression tests as well as the fractional and collapsing regression methods (COLLREG and FRACREG) [8], which also can be combined with the VT technique. INTERSNP-RARE is specifically designed to handle large data sets and is able to perform genome-wide rare variant analysis within manageable time frames.

In addition to the VT methods, we are also going to investigate the analysis bins defined by our software with kernel methods. To this purpose, we will use the SKAT and SKAT-O methods and their implementation in the R package SKAT [11].

As a supplementary approach to NGS rare variant studies, we propose a pipeline for rare variant analysis of imputed data. We describe respective quality control (QC) criteria and will provide suggestions for the choice and construction of analysis bins in whole-genome application. We evaluated the practicability of our framework by applying it to a GWAS on Alzheimer's disease (AD).

Methods

In the section Pipeline, we will propose a pipeline for the analysis of imputed rare variants. The approach uses binning strategies and statistical methods implemented in INTERSNP-RARE as well as statistical tests implemented in SKAT [11]. Therefore, we first describe these implementations.

INTERSNP-RARE is an extension of INTERSNP [25], a genome-wide association and interaction analysis software, written in C++. Shared-memory parallelization is implemented using the OpenMP [26] framework for C++. INTERSNP-RARE is compatible with all features of INTERSNP and allows the combination of methods. Oftentimes, it makes sense to define a preselection of variants to be read from the input file, for instance to include only rare variants to improve computational speed. Using a preselection of SNPs also allows for even more advanced binning methods, for example to combine genes from the same pathway or to exclude introns in a gene-based analysis. INTERSNP-RARE is compatible with all PLINK [27] input data formats.

Binning with INTERSNP-RARE

Binning is the partition of available chromosomes into contiguous sequences, each one containing *nRV* rare variants for subsequent analysis. Given a chromosome with *n* rare variants, there are $n\,(n-1)/2$ possible distinct bins. However, hypothetical truly associated regions are expected to be relatively limited in extent ($nRV << n$), so that very large bins can be safely excluded from the analysis. In the following, we assume $nRV \leq 100$. INTERSNP-RARE provides various binning strategies. The software can handle boundaries predefined by external resources, as, for instance, from the Ensembl release 70 [28] which specifies gene boundaries. In the same way, haplotype block boundaries or bin boundaries defined in preceding studies can be used. Analysis bins can automatically be concatenated, in order to analyze neighboring blocks as a unit, or split into bins with a number of rare variants below a predefined threshold. In addition, bins can be defined on a count basis, either by window size in base pairs or by the number of rare variants in a bin.

In the following, we will explore gene binning and two binning strategies based on haplotype blocks in a genome-wide setting.

Association Testing with INTERSNP-RARE

We adapted multiple rare variant association tests from the literature. First, we describe the conventional case: analysis with a fixed MAF_T. In this section, we describe association testing methods assuming a sample size of *N* individuals and a predefined bin containing *n* rare variants.

Rare Variant Tests

COLL

COLL dichotomizes *N* individuals according to the presence of at least one rare variant. The corresponding 2 × 2 table contains the counts of affected (*A*) and unaffected (*U*) carriers (*C*) and non-carriers (*NC*), respectively (table 1). It can be evaluated by applying Pearson's χ^2 test with 1 degree of freedom, resulting in the following test statistic:

$$T_{COLL} = \frac{\left(N_A + N_U\right)\left(N_U N_A^C - N_A N_U^C\right)^2}{N_A N_U \left(N_A^C + N_U^C\right)\left(N_A^{NC} + N_U^{NC}\right)}. \quad (1)$$

Table 1. Contingency table for COLL

Number of individuals	Carriers	Noncarriers
Affected	N_A^C	N_A^{NC}
Unaffected	N_U^C	N_U^{NC}

Table 2. Contingency table for CMAT

Number of alleles	Rare/Minor	Common/Major
Affected	m_A	M_A
Unaffected	m_U	M_U

CMAT

CMAT is conducted by pooling all variants, divided into rare and common across all cases and controls, and by taking the sum of the number of major ($M_{A/U}$) and minor ($m_{A/U}$) alleles of affected and unaffected individuals. The resulting contingency table (table 2) leads to the following test statistic:

$$T_{CMAT} = \frac{\left(N_A + N_U\right)\left(m_A M_U - m_U M_A\right)^2}{2\,n\,N_A N_U \left(m_A + m_U\right)\left(M_A + M_U\right)}. \quad (2)$$

Due to LD, the asymptotic p value is not applicable and the p value is obtained via permutation (see Significance Testing by Permutation and Variable Threshold Analysis). CMAT can be extended to a Cochran-Mantel-Haenszel-like test, so that the inclusion of a categorical variable is possible [7].

Regression Tests: REG, FRACREG, COLLREG

REG is a full regression test conducted on a set of rare variants. Logistic and linear regressions are used for quantitative and binary phenotypes, respectively. In the logistic regression, the probability to be a case is modeled as:

$$\text{logit}(p) = \beta_0 + \beta^T b + \gamma^T g, \quad (3)$$

for every individual, with $\text{logit(p)} = \log(\text{p}(1-\text{p})^{-1})$, and the vectors of regression coefficients β and γ to be estimated. The covariate vector $b = \{b_1, \ldots, b_k\}$ encodes covariate quantities of the individual to be included in the model, while *g* is the vector of genotypes of the individual. In the log-additive model, the genotypes are coded as 1, 0 and –1 for the three cases when the minor allele is homozygous, the genotype is heterozygous or when the major allele is homozygous. The vector of covariate coeffi-

cients β consists of k covariate coefficients $\{\beta_1, \ldots, \beta_k\}$, while γ estimates the risk contribution of each genotype. The null hypothesis H_0: $\gamma = 0$ is investigated with a likelihood ratio test that compares the likelihood of the full model (equation 3) to the restricted model

$$\text{logit}(p) = \beta_0 + \beta^T b. \tag{4}$$

The likelihood ratio test statistic is χ^2 distributed with m degrees of freedom, where m is the number of variants contained in the bin.

The regression test is able to detect association when there are both causal and protective rare variants. This property sets them apart from the 'burden' tests COLL and CMAT described before, which test for an excess of rare variants in cases. Two further burden regression tests have been described in Morris and Zeggini [8]. The main idea is to combine rare genotypes from a bin into one variable and to fit a single regression coefficient in a regression model. In this sense, the approach transfers the main ideas of COLL and CMAT into a regression setting. Therefore, we refer to the respective tests as COLLREG and FRACREG in the following.

COLLREG defines a regression model over collapsed variants. An individual is assumed to carry the full genetic burden when at least one rare allele is present. The indicator function $I(g)$ equals 1 if an individual has at least one rare allele in the region of interest and 0 otherwise. One coefficient, λ, is estimated:

$$\text{logit}(p) = \beta^T b + \lambda I(g). \tag{5}$$

FRACREG also estimates a single parameter, coefficient λ with the trait modeled as:

$$\text{logit}(p) = \beta^T b + \lambda \frac{r}{n}. \tag{6}$$

Here, r is the count of genotypes of an individual containing at least one minor allele. The ratio r/n is the accumulated genetic burden with equal contribution from each genotype containing one or two minor alleles.

Both COLLREG and FRACREG test H_0: $\lambda = 0$. The corresponding likelihood ratio test is χ^2 distributed with 1 degree of freedom.

Significance Testing by Permutation and Variable Threshold Analysis

Permutation testing procedures assess the typically unknown distribution of a statistic under H_0 by evaluating replicated data sets which are constructed by random assignment of affection status or quantitative trait value. The statistic obtained for the 'real' (non-permuted) data is compared to the statistic obtained from permutation replicates, the p value is obtained by counting the number of permuted test statistics that are greater or equal to the test statistic of the real data:

$$p = \frac{1}{N_{SIM}} \sum_{j=1}^{N_{SIM}} I\left(T^j \geq T^0\right). \tag{7}$$

The indicator function $I()$ returns 1 if the condition is satisfied and 0 otherwise.

Confidence intervals for p values determined by permutation can be calculated using the Wilson score method [29]:

$$\text{CI}^{\pm} = \left(1 + \frac{z^2_{\alpha/2}}{N_{SIM}}\right)^{-1} \left(p + \frac{z^2_{\alpha/2}}{2N_{SIM}} \pm z_{\alpha/2} \sqrt{\frac{1}{N_{SIM}} p\left(1-p\right) + \frac{z^2_{\alpha/2}}{4N^2_{SIM}}}\right), \tag{8}$$

where $z_{\alpha/2} = 1.96$ for the 95% confidence interval. The Wilson method performs well even for small values and is known to be more accurate for binomial proportions than the commonly used normal approximation interval (Wald interval) [30].

The extension of rare variant tests with the VT approach is motivated by the method described in Price et al. [24]. Instead of using a fixed threshold, an optimal threshold that maximizes the test statistic is computed from all possible MAFs that are smaller or equal to MAF_T.

For COLL and CMAT, the VT method can be described as follows. The maximum test statistic T^j_{max} is determined across all MAF levels $\text{MAF}_\text{i} \leq \text{MAF}_\text{T}$ for each of the N_{SIM} permutations labeled by j. The proportion of those T^j_{max} that are larger or equal to the test statistic T^0_{max} in real non-permuted data is the permutation-based p value (see equation 9).

$$p = \frac{1}{N_{SIM}} \sum_{j=1}^{N_{SIM}} I\left(T^j_{max} \geq T^0_{max}\right). \tag{9}$$

The MAF at which the test statistic has the maximal value T^0_{max} is the 'optimal' threshold. The variable threshold method is computationally more intense than fixed-threshold methods, since test statistics at all possible MAF_T have to be computed.

If the VT approach should be applied to regression tests REG, FRACREG and COLLREG, the T_{max} approach has to be replaced by a p_{min} approach. With varying MAF_T, the number of SNPs to be included in the regression model changes and leads to test statistics with different degrees of freedom. The regression test statistic will be systematically higher for larger MAF_T, since these lead to the inclusion of more model SNPs. To overcome this problem, the test statistics obtained for different MAF_T are replaced by their respective nominal p value, thereby transforming them to a comparable scale. Instead of using equation 9, the overall p value now can be computed according to

$$p = \frac{1}{N_{SIM}} \sum_{j=1}^{N_{SIM}} I\left(p^j_{min} \leq p^0_{min}\right). \tag{10}$$

Since regression tests are computationally intense, their genome-wide application under the VT approach needs high computing power, and it will not always be possible to include it in the standard analysis.

Association Testing with SKAT

The R package SKAT [11] implements two rare variant tests, SKAT and SKAT-O. SKAT, the sequence kernel association test, uses single SNP score test statistics. In more detail, let m be the number of rare SNPs from a bin. Then, the SKAT test statistic can be written as:

$$Q = \sum_{1 \leq j \leq m} w_j^2 S_j^2$$

where S_j is the score statistic for testing the logistic single SNP model for SNP j ($1 \leq j \leq m$). SKAT allows for weighting of variants. The default weighting scheme is based on the family of Beta(x, α, β) $\propto x^{\alpha-1}(1-x)^{\beta-1}$ distributions:

$$w_j = Beta(MAF_j, 1, 25) = 25(1 - MAF_j)^{24}.$$

This weighting scheme assigns significantly higher weights to variants with low MAFs.

SKAT-O is an advanced method that uses a linear combination of a burden and a non-burden test statistic and it is a linear combination of SKAT with a burden test. The p values are computed analytically. An adjustment using bootstrapping is necessary for samples with <2,000 individuals.

Pipeline

We suggest the following analysis pipeline for genome-wide rare variant analysis of imputed data.

(1) We start with a GWAS data set that went through a stringent standard QC protocol. For instance, in order to minimize differential errors caused by different imputation qualities of cases and controls, we require a SNP call rate above 99%, both in cases and controls. Next, the data is imputed using standard imputation software packages [13–15]. As reference panel, we recommend to use the latest release of the 1,000 Genomes Project [3]. Ideally, the complete reference sample should be used, rather than ethnical subgroups fitting the data [13].

(2) We apply postimputation QC and filter out SNPs with a low info score (e.g. below 0.8).

(3) Next, we 'call' SNPs/individual genotypes by assignment of the genotype with the highest imputation probability. Note that there is a practical need for 'calling', since the COLL test, for instance, is not defined on dosage data. Depending on the quality metrics of the resulting data set, further QC procedures can be implemented.

(4) In a genome-wide setting, the assignment of variants to bins is not self-evident. In the next section, we present the analysis of empirical data (see AD Study). Here, we applied three example binning strategies: first, the conventional approach is based on the locations of protein-coding genes ('gene binning'). We used gene positions as defined in Ensembl release 70 [28]. For the second strategy ('block binning'), LD blocks were determined using PLINK [27]. We computed LD blocks from the genotype data set and used the default block size limit of 200 kb. The set of block intervals was not contiguous and covered only a fraction of rare variants in the imputed data set. To achieve full coverage, the intervals at each gap were extended by an equal number of base pair positions until the gap was closed. The resulting set of contiguous intervals was used. For the third strategy, three consecutive contiguous block intervals were concatenated into a single interval ('3-block binning' or 'concatenated LD block binning'). We also address computational problems caused by 'overly' small or large bins. Bins with >100 variants were split into multiple bins of equal size with ≤100 rare variants. Small bins with <4 variants were excluded from the analysis.

(5) Since we wish to investigate the performance of the pipeline in general, we apply all statistical tests described before (COLL, CMAT, REG, FRACREG, COLLREG, SKAT and SKAT-O). For practical analysis, specific choices might be advisable. The test statistics of COLL and CMAT are analyzed with VT MAF optimization. We recommend to use 10,000 permutation replicates in an initial run, with one job per chromosome. Top results are followed up with an increasing number of permutations until two significant digits of the p value are obtained, up to 10^9. For a large number of permutations, the parallel version of INTERSNP-RARE can be used to reduce computational time. Regression tests are conducted without permutation testing using a fixed MAF of 0.05 and population covariate parameters to adjust for potential stratification. SKAT and SKAT-O [11] are analyzed with beta-weighting and population covariates.

(6) Candidate regions that provide evidence for possible association can be investigated in a refinement analysis, the extent of which is decided on a case-by-case basis. Functional characteristics, known association signals and additional available information on variants contained in promising regions can be used to decide if the candidate regions are worth following up in independent samples. It can be examined if the association is a result of LD with a common variant. For this purpose, the common variant can be used as a covariate parameter.

AD Study

We analyzed a previously unpublished late-onset AD genome-wide case-control study, genotyped on the Illumina®Omni1M microarray. AD patients were recruited within the German Dementia Competence Network, DCN (http://www.kompetenznetz-demenzen.de), and the interdisciplinary memory clinic of the Department of Psychiatry and the Department of Neurology at the University Hospital in Bonn, Germany. Diagnosis of AD dementia was established according to the NINCDA-ADRDA criteria [31]. All patients gave written informed consent for participation in the entire study.

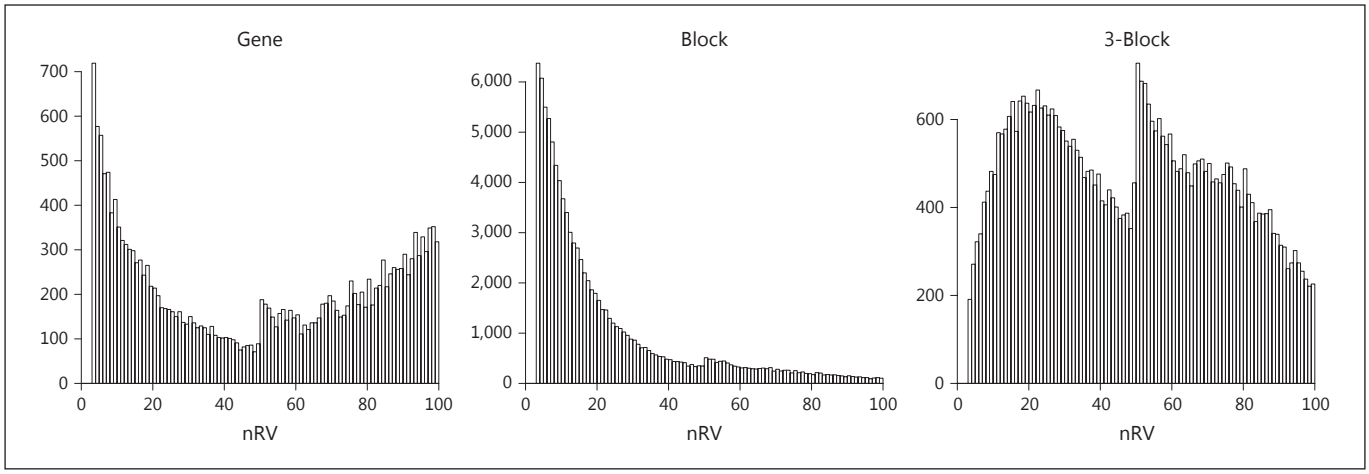


Fig. 1. Distribution of bin sizes for all three analysis-ready binning strategies, obtained from raw interval files by closing gaps (block and 3-block), concatenating (3-block) and bin size control. The discontinuities at $nRV = 50$ result from splitting of bins with $nRV > 100$.

First, we applied standard GWAS QC procedures. We detected 7 pairs of individuals with evidence for relatedness based on an elevated genome-wide genotype identity-by-state status and removed that individual from each of the pairs with the lower genotyping rate. SNPs and individuals were filtered for a genotyping rate of at least 99%. In addition, SNPs with p values indicating a deviation from the Hardy-Weinberg equilibrium at $p = 5 \times 10^{-4}$ were excluded. Finally, variants for which the test for association between missingness and affection status (PLINK's test missing) resulted in p values ≤0.05 were removed from the analysis. After QC, 850,612 genotypes from 531 cases and 1,096 controls remained. We performed multi-dimensional scaling implemented in PLINK [27] and retained the 6 leading dimensions as covariate parameters to adjust for population stratification.

Next, we used the software IMPUTE2 [13] to impute the data sets into the May 2012 release of the 1,000 Genomes Project [3]. Following suggestions from the literature [13], we used the complete reference sample, i.e. individuals from all ethnical groups. After imputation, we removed SNPs with an info score (a quality metric provided by the imputation software) smaller than 0.8. Next, we 'called' individual genotypes by assigning that genotype to each SNP/individual with the highest imputation probability. In total, 7,664,433 variants remained for analysis, of which 2,234,422 were rare (MAF <0.05).

Binning Strategies

The distributions of bin sizes for all three strategies described above are shown in figure 1. Although the strategies resulted in distributions with different characteristics (table 3), all of them led to valid analysis results as will be shown below. Note that bins with <4 rare variants were excluded. The numbers of resulting bins for the gene, block and 3-block strategies were used to determine the Bonferroni-corrected 'binning-wide' significance levels by dividing the nominal significance threshold (0.05) by the number of bins, resulting in 2.44×10^{-6}, 5.14×10^{-7} and 1.09×10^{-6}, respectively. The 'strategy-wide' significance levels were further adjusted for the number of conducted tests, resulting in 3.49×10^{-7}, 7.34×10^{-8} and 1.55×10^{-7}, respectively. The additional correction for the

Table 3. Comparison of the three binning strategies

	Gene	Block	3-Block
Bins	20,475	97,283	46,002
Variants	968,326	2,196,256	2,233,593
Variants included	43.3%	98.3%	99.6%
Median variants/bin	51	14	49
Binning-wide α	2.44×10^{-6}	5.14×10^{-7}	1.09×10^{-6}
Strategy-wide α	3.49×10^{-7}	7.34×10^{-8}	1.55×10^{-7}
Experiment-wide α	1.16×10^{-7}	2.45×10^{-8}	5.18×10^{-8}

The two block-binning strategies include <100% of the rare variants due to the filtering of small bins. The binning-wide significance level α was Bonferroni-corrected and determined by dividing 0.05 by the number of bins. The strategy-wide significance level α was additionally corrected for the 7 statistical tests. The experiment-wide significance level α was obtained from the strategy-wide significance level α by Bonferroni correction for the three binning strategies.

number of binning strategies determined the 'experiment-wide' significance levels of 1.16×10^{-7}, 2.45×10^{-8} and 5.18×10^{-8}, respectively. Please note that the correction is somewhat conservative for several reasons. Variants from different bins are locally correlated due to LD. Apart from that, there is an overlap between bins defined by different binning strategies. Finally, there is correlation between the results from different tests performed on the same bin.

Running Time

VT analysis was conducted in two steps. First, a genome-wide analysis was performed with 10^4 permutation replicates. Next, bins with a p value $\leq 10^{-4}$ were re-evaluated with 10^7 permutation replicates. To obtain two significant digits of the p value, up to 10^9 permutation replicates were performed. The analysis was conducted on a high-performance cluster of the University of Bonn, Bonn, Germany. It comprises 32 main nodes with 50 GB working memory and 24 processors with 3 GHz each. For the analysis presented here, we used one dedicated node (24 processors).

The regression tests REG, FRACREG and COLLREG were performed in joint runs, one for each chromosome. The running time of each job was <2 min for all chromosomes and binning strategies. The VT analysis of COLL and CMAT was also performed in joint runs, one per chromosome. The running time of the step 1 analysis with 10^4 permutation replicates varied from 33 min for chromosome 21 (gene-based strategy) to 509 min for chromosome 2 (concatenated LD block strategy). The follow-up analysis of 183 top-ranking bins with 10^7 permutation replicates was conducted in single runs of the parallel version of INTERSNP-RARE using 24 processors and took 18 h.

Analyzing chromosome 2 under the concatenated LD block strategy took 115 min with SKAT and 253 min with SKAT-O. The running time was longer than with logistic regression, since the sample size is <2,000 and the bootstrap feature of SKAT had to be used.

QQ Plots

QQ plots for all 7 statistical tests under all binning strategies are presented in figures 2–4, together with the respective significance levels (table 3) and inflation factors obtained by the software package GenABEL [32].

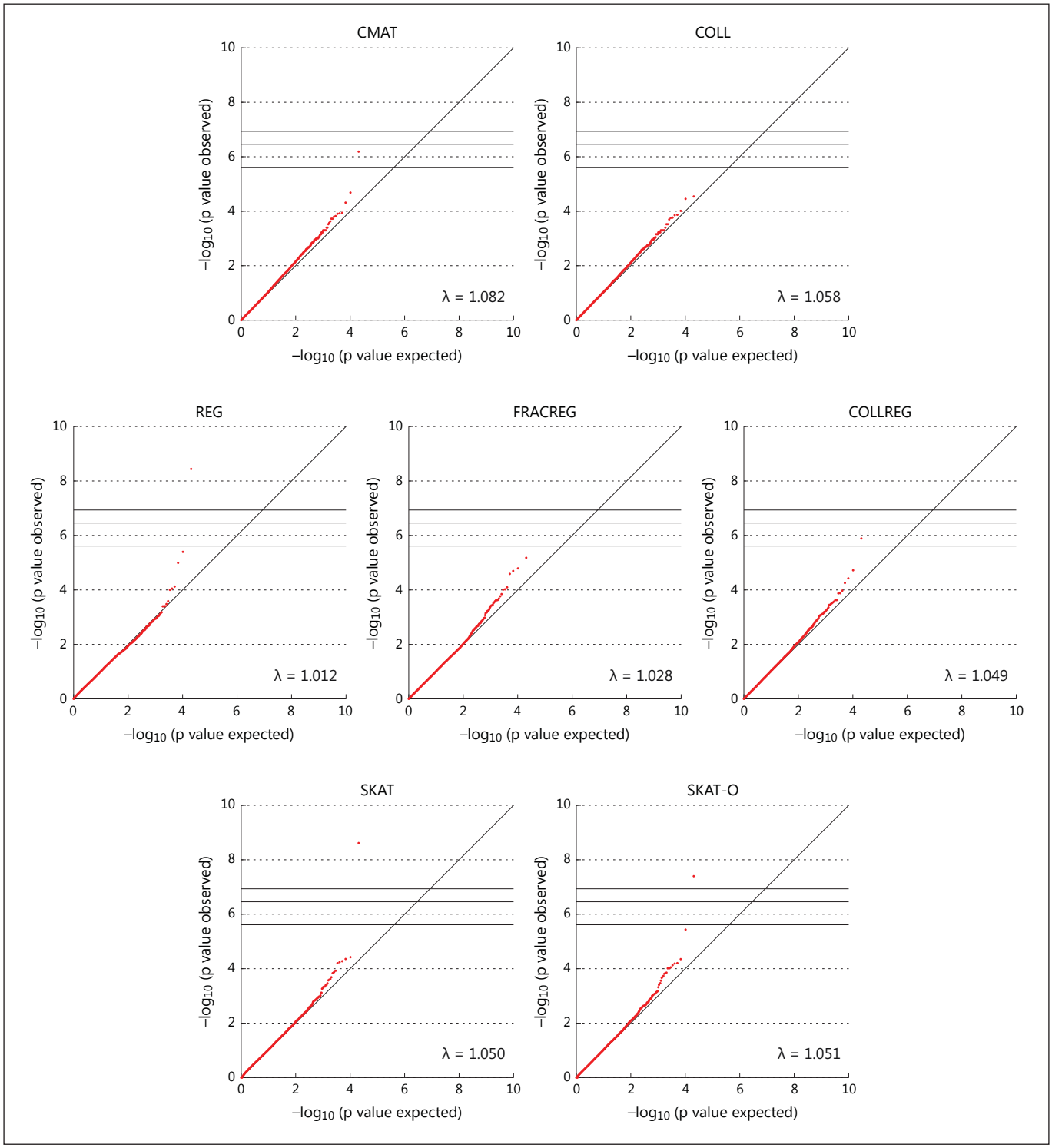
Under the gene-based strategy, we did not observe substantial deviations of the p value distribution for any of the tests considered. Under the block-based strategy, COLL and CMAT showed some tendency of inflation. This has to some extent been expected, since these tests do not adjust for population stratification. The analogous regression tests COLLREG and FRACREG, which were conducted with population covariates, however, did not show strong signs of inflation. The QQ plot of the SKAT statistic appears to be slightly inflated for p values with –log(p) >3.5, while the regression test fits the expectation perfectly well for p values with –log(p) <4. Overdispersion in the lower p value range can be explained by multiple hits that reflect true associations with the *APOE* locus (see Analysis of the *APOE* LD Region). Under the 3-block strategy, we observe in general the same tendencies as before. COLL and CMAT, which do not account for population covariates, produced slightly worse results, while the other tests exhibited QQ plots that fitted the expectation almost as good as with the simple block strategy. Overall, we can state that the QQ plots of the p value distributions confirm the validity of the suggested approach. This conclusion is also supported by the reported λ inflation. In general, most strategies showed inflation factors below or around 1.05. The already observed tendency for inflation of COLL and CMAT is also reflected in their λ values. In particular under the three-block strategy, COLL ($\lambda = 1.084$) and CMAT ($\lambda = 1.11$) show substantial inflation. A validation of the results using covariates adjusting for population stratification is, therefore, obligatory.

Analysis of the APOE LD Region

Two common SNP polymorphisms, rs429358 and rs7412, define the well-known *APOE* ε2, ε3, ε4 polymorphism which is strongly associated with AD susceptibility [33]. None of the rare SNPs available in our study is located within the boundaries of *APOE* (chromosome 19, 45409011–45412650 bp), which explains that there is no rare variant association signal for *APOE* itself (table 4). However, table 4 contains multiple significant bins in the LD region surrounding *APOE*.

With the gene-binning strategy, the bins corresponding to the *TOMM40* and *PVRL2* genes achieved experiment-wide significance (table 4). SKAT ($p = 2.4 \times 10^{-9}$) identified *TOMM40*, which contains 22 rare variants, at experiment-wide significance. In addition, the regression test REG ($p = 4.0 \times 10^{-6}$) and CMAT showed suggestive evidence ($p = 6.5 \times 10^{-7}$) for *TOMM40*. The *PVRL2* gene contains 37 rare variants and reached strongest evidence for association with the regression test REG ($p = 3.6 \times 10^{-9}$), while the p value reached with SKAT was much less impressive ($p = 1.2 \times 10^{-3}$).

With the block-binning strategy, two bins achieved strategy-wide significance (table 4). These blocks are part of, or overlap with, *TOMM40* and *APOC1*, respectively. The strongest statistical evidence was achieved for *APOC1* which contains 8 rare variants. SKAT ($p = 2.7 \times 10^{-14}$) and



Color version available online

Fig. 2. QQ plots for the gene-binning strategy. Two *APOE*-associated genes reach experiment-wide significance, *TOMM40* with SKAT and SKAT-O ($p = 2.4 \times 10^{-9}$ and $p = 4.0 \times 10^{-8}$, respectively) and *PVRL2* with REG ($p = 3.6 \times 10^{-9}$). *TOMM40* also reaches binning-wide significance with CMAT ($p = 6.5 \times 10^{-7}$) and is slightly above binning-wide significance with REG ($p = 4.0 \times 10^{-6}$). *MED18* reaches binning-wide significance with COLLREG ($p = 1.3 \times 10^{-6}$). *MED18* is also the third-highest scoring result in REG ($p = 1.0 \times 10^{-5}$), top-scoring result in FRACREG ($p = 6.6 \times 10^{-6}$) and the second-highest scoring result in SKAT-O ($p = 3.7 \times 10^{-6}$), although none of these cases achieves binning-wide significance.

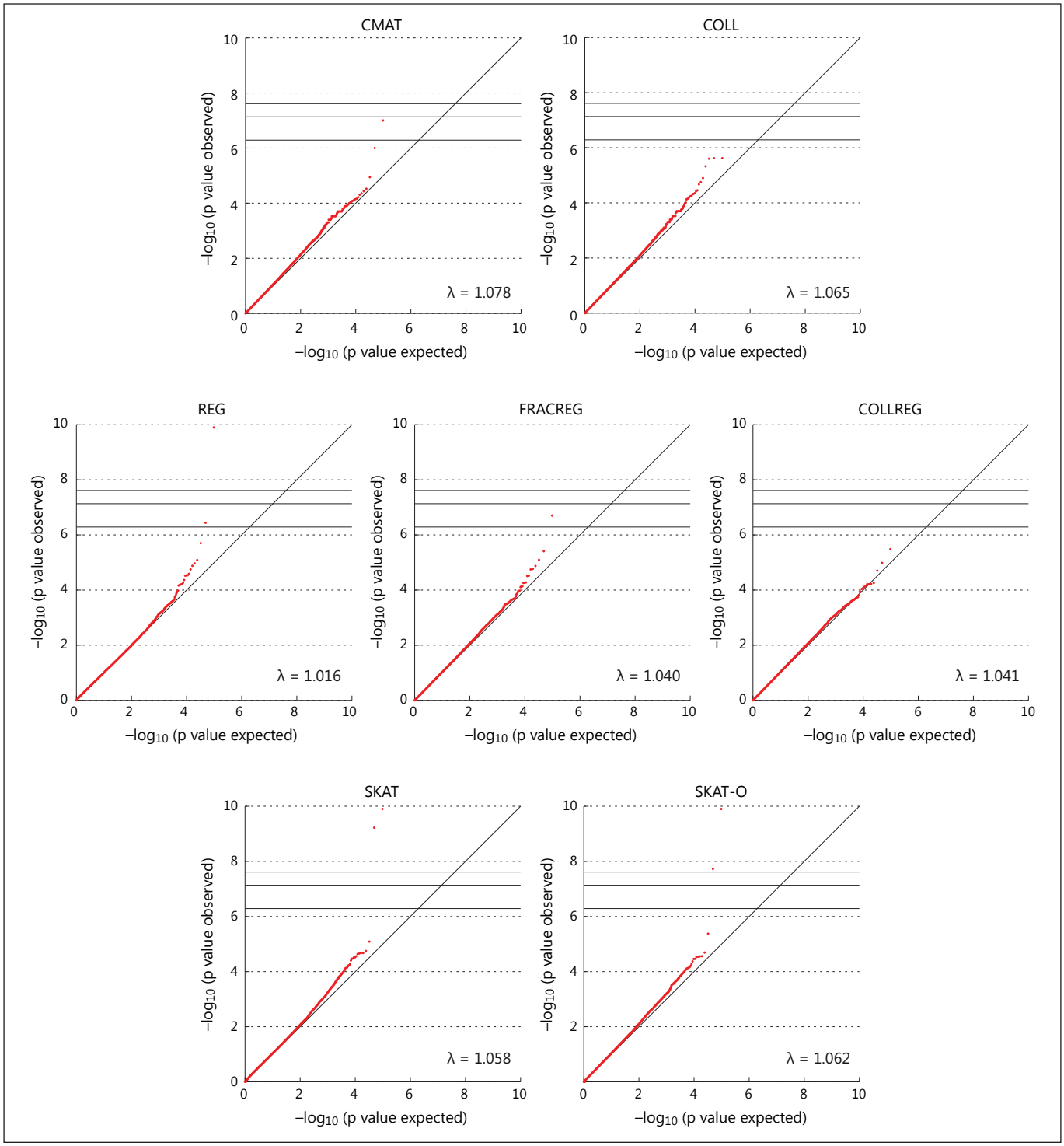


Fig. 3. QQ plots for the block-binning strategy. A bin overlapping with the *APOC1* gene has highly significant p values in three of the tests: REG ($p = 3.2 \times 10^{-14}$), SKAT ($p = 2.7 \times 10^{-14}$) and SKAT-O ($p = 1.7 \times 10^{-12}$). Note that all three p values are beyond the plotted range and are symbolically shown at $\approx 10^{-10}$. *APOC1* also achieves binning-wide significance with CMAT ($p = 4.6 \times 10^{-8}$) and FRACREG ($p = 2.0 \times 10^{-7}$). The second bin with experiment-wide significance overlaps with *TOMM40* (SKAT with $p = 6.0 \times 10^{-10}$ and SKAT-O with $p = 1.9 \times 10^{-8}$). It is also the second-highest scoring result with CMAT ($p = 6.6 \times 10^{-7}$) and the third-highest scoring result in REG ($p = 2.0 \times 10^{-6}$). In both cases, the hit is below the binning significance threshold. The second-highest scoring bin in REG overlaps with *PVRL2* and achieves binning-wide significance ($p = 3.6 \times 10^{-7}$).

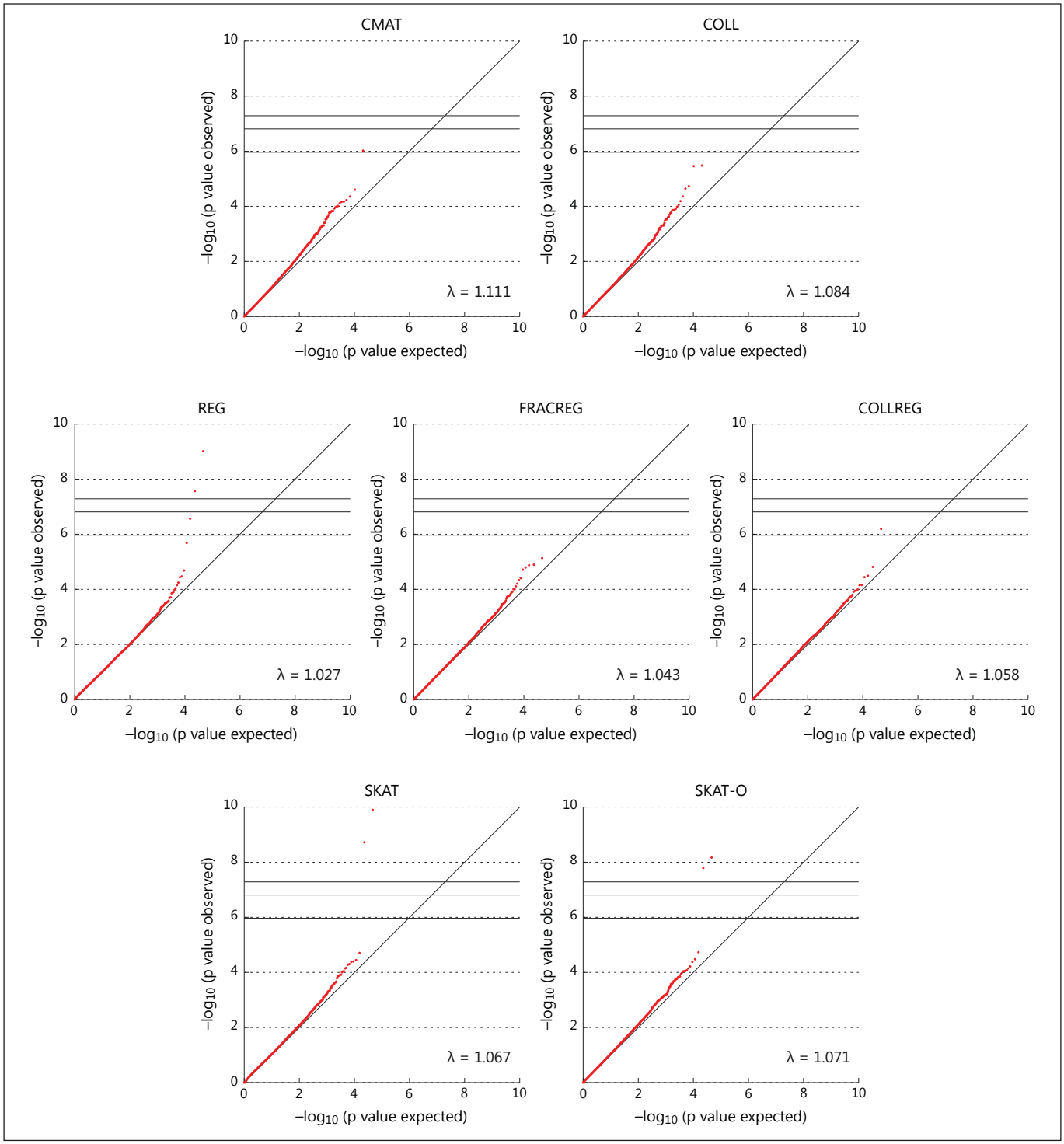


Fig. 4. QQ plots for the 3-block-binning strategy. Three *APOE*-associated bins achieve experiment-wide significance. The most significant result is a bin containing the genes *APOC1*, *APOC2* and *APOC4* (SKAT with $p = 7.6 \times 10^{-11}$, REG with $p = 9.7 \times 10^{-10}$ and SKAT-O with $p = 6.7 \times 10^{-9}$). Note that the SKAT p value is shown symbolically at $\approx 10^{-10}$, since the true value is beyond the plotted range. The second experiment-wide significant result is the *TOMM40*, *APOE* bin (SKAT with $p = 1.9 \times 10^{-9}$ and SKAT-O with $p = 1.6 \times 10^{-8}$). The bin achieves binning-wide significance with REG ($p = 2.7 \times 10^{-7}$). Finally, *PVRL2* is experiment-wide significant with REG ($p = 2.7 \times 10^{-8}$). This bin does not come close to significance with any other test. In addition, a non-*APOE*-associated bin containing *MCPH1* is detected with COLLREG below the binning-wide significance threshold ($p = 6.5 \times 10^{-7}$).

Table 4. Bins achieving experiment-wide significance

Strategy	Position (chromosome:bp)	Feature	MAF_T	nRV	Test	p	p (*APOE* covariates)
Gene	19:45395193–45405499	*TOMM40*	0.0181	8	COLL	5.1e-02 [4.65e-02; 5.51e-02]	–
			0.0396	17	CMAT	6.5e-07 [5.10e-07; 8.28e-07]	0.24
			0.0412	22	REG	4.0e-06	0.32
			0.0412	22	FRACREG	1.6e-05	0.85
			0.0412	22	COLLREG	9.3e-01	1.00
			0.0412	22	SKAT	2.4e-09	0.83
			0.0412	22	SKAT-O	4.0e-08	1.00
	19:45351516–45391505	*PVRL2*	0.0495	37	COLL	7.0e-04 [3.39e-04; 1.44e-04]	–
			0.0080	9	CMAT	4.0e-01 [3.88e-01; 4.07e-01]	0.97
			0.0495	37	REG	3.6e-09	0.12
			0.0495	37	FRACREG	6.8e-01	0.70
			0.0495	37	COLLREG	1.1e-04	0.30
			0.0495	37	SKAT	1.2e-03	0.17
			0.0495	37	SKAT-O	2.5e-03	0.25
Block	19:45417200–45427779	*APOC1*	0.0298	4	COLL	9.0e-06 [4.74e-06; 1.71e-05]	–
			0.0409	8	CMAT	4.6e-08 [3.67e-08; 5.92e-08]	0.06
			0.0409	8	REG	3.2e-14	0.68
			0.0409	8	FRACREG	2.0e-07	0.38
			0.0409	8	COLLREG	2.1e-03	0.14
			0.0409	8	SKAT	2.7e-14	0.67
			0.0409	8	SKAT-O	1.7e-12	0.39
	19:45396973–45407655	*TOMM40*	0.0181	7	COLL	2.5e-01 [4.56e-01; 6.99e-01]	–
			0.0396	16	CMAT	6.6e-07 [5.19e-07; 8.40e-07]	0.47
			0.0412	21	REG	2.0e-06	0.30
			0.0412	21	FRACREG	1.7e-05	0.91
			0.0412	21	COLLREG	7.0e-01	0.98
			0.0412	21	SKAT	6.0e-10	0.80
			0.0412	21	SKAT-O	1.9e-08	0.39
3-Block	19:45417200–45458466	*APOC1*, *APOC2*, *APOC4*	0.0298	18	COLL	3.9e-02 [3.55e-02; 4.31e-02]	–
			0.0041	22	CMAT	4.1e-03 [3.02e-03; 5.56e-03]	0.73
			0.0498	25	REG	9.7e-10	0.92
			0.0498	25	FRACREG	8.2e-03	0.74
			0.0498	25	COLLREG	1.1e-01	0.47
			0.0498	25	SKAT	7.6e-11	0.99
			0.0498	25	SKAT-O	6.7e-09	0.81
	19:45396973–45414392	*TOMM40*, *APOE*	0.0181	8	COLL	2.8e-02 [2.53e-02; 3.18e-02]	–
			0.0396	16	CMAT	9.5e-07 [8.92e-07; 1.02e-06]	0.25
			0.0415	23	REG	2.7e-07	0.24
			0.0415	23	FRACREG	1.3e-05	0.96
			0.0415	23	COLLREG	5.3e-01	0.85
			0.0415	23	SKAT	1.9e-09	0.80
			0.0415	23	SKAT-O	1.6e-08	1.00
	19:45337918–45374983	*PVRL2*	0.0492	30	COLL	1.2e-01 [1.15e-01; 1.28e-01]	–
			0.0092	10	CMAT	3.1e-02 [2.77e-02; 3.45e-02]	0.72
			0.0492	30	REG	2.7e-08	0.04
			0.0492	30	FRACREG	7.7e-02	0.50
			0.0492	30	COLLREG	1.5e-02	0.57
			0.0492	30	SKAT	5.3e-05	0.20
			0.0492	30	SKAT-O	1.4e-04	0.31

For each of the three strategies, the true association with the *APOE* locus was detected after the most stringent Bonferroni correction. The feature column contains protein-coding genes that overlap with the bin. The MAF_T column contains either the largest MAF (≤0.05) for fixed threshold tests (REG, FRACREG, COLLREG, SKAT and SKAT-O) or the optimal threshold determined by the VT method for permutation tests (CMAT and COLL). The number of rare variants (nRV) is determined by the bin coordinates together with the bin inclusion criterion $MAFi \leq MAF_T$. For p values obtained via permutation, the 95% Wilson confidence interval is given in brackets. The last column contains the p values under inclusion of the *APOE* haplotype status encoded in the covariates. For the CMAT test with covariates, the Cochran-Mantel-Haenszel-like version of CMAT was used, with a categorical variable encoding the *APOE* haplotype. For the regression and kernel tests, the encoded alleles of rs429358 and rs7412 were used as covariates in addition to the six multi-dimensional scaling covariates. After correction for multiple testing, none of the *APOE*-adjusted p values remains significant, leading to the conclusion that rare tests can be sensitive to the well-known association tagged by two common variants.

Table 5. Top 3 non-*APOE* results for each strategy across all tests

Strategy	Position (chromosome:bp)	Feature	CMAT			COLL			Rest						
			MAF_T	nRV	p	MAF_T	nRV	p	MAF_T	nRV	P_{REG}	$P_{FRACREG}$	$P_{COLLREG}$	P_{SKAT}	P_{SKAT-O}
Gene	1:28656262–28660052	*MED18*	0.0132	5	2.1e-05	0.0132	5	2.9e-05	0.0132	5	1.0e-05	6.6e-06	1.3e-06	3.8e-05	3.7e-06
	4:44626103–44679950	*YIPF7*	0.0138	37	7.6e-01	0.0443	58	1.4e-03	0.0443	58	7.1e-03	6.5e-01	1.9e-05	3.2e-01	3.9e-01
	19:15270636–15307070	*NOTCH3*	0.0310	31	5.0e-04	0.0310	31	1.0e-04	0.0310	31	4.7e-02	2.0e-05	5.5e-05	1.6e-02	1.9e-04
Block	3:103703359–103718867	129307bp to *COLL11A1*	0.0433	6	1.6e-02	0.0433	6	2.4e-06	0.0482	12	3.4e-04	5.2e-02	3.3e-06	3.9e-02	4.8e-02
	6:29800142–29810740	MHC region	0.0144	10	4.8e-03	0.0132	8	2.4e-06	0.0283	31	4.7e-04	7.9e-01	2.9e-02	5.0e-02	8.2e-02
	6:29818165–29832926	MHC region	0.0089	40	1.0e-04	0.0089	40	2.8e-06	0.0230	73	9.5e-03	4.8e-01	4.6e-01	4.6e-02	7.0e-02
3-Block	8:6253932–6278000	*MCPH1*	0.0482	59	1.5e-02	0.0455	57	3.3e-06	0.0482	59	3.3e-03	2.8e-03	6.5e-07	6.1e-02	1.9e-02
	21:18738520–18760122	124578 bp to *CXADR*	0.0381	27	7.0e-04	0.0086	3	5.5e-03	0.0452	38	2.1e-06	1.9e-02	2.2e-02	1.4e-04	1.7e-04
	6:29817923–29828276	MHC region	0.0089	31	1.5e-04	0.0089	31	3.5e-06	0.0230	53	9.0e-03	3.4e-01	4.9e-01	3.6e-02	5.5e-02

For the COLL and CMAT tests, optimal MAF_T as found by the VT method are given with the number of rare variants (nRV) with MAF ≤ MAF_T. The remaining tests were conducted under a fixed MAF_T, therefore they share the same MAF_T, which corresponds to the highest MAF ≤0.05 available in the bin.

The most significant non-*APOE* bin in the gene-binning strategy is *MED18* with COLLREG resulting in $p = 1.3 \times 10^{-6}$, which achieved binning-wide significance. The other 6 tests are consistent with this result, with p values ranging from 3.7×10^{-6} (SKAT-O) to 3.8×10^{-5} (SKAT). The second- and third-most significant results (*YIPF7*, $p = 1.9 \times 10^{-5}$ with COLLREG and *NOTCH3* with FRACREG $p = 2.0 \times 10^{-5}$, respectively) did not reach binning-wide significance and were not consistently supported across different tests.

In the block-binning strategy, none of the non-*APOE* bins achieved a significant result. Moreover, the top three results were found by the COLL test. Although the first result (chromosome 3, 103703359–103718867 bp, $p = 2.4 \times 10^{-6}$ with COLL) is somewhat supported by COLLREG ($p = 3.3 \times 10^{-6}$), there is little evidence that any of the bins contained a true positive signal.

The 3-block-binning strategy contains a binning-wide significant result overlapping with the *MCPH1* gene ($p = 6.5 \times 10^{-7}$ with COLLREG, $p = 3.3 \times 10^{-6}$ with COLL). The evidence for association is not supported by non-collapsing tests. The second- and third-most significant non-*APOE* bins ($p = 2.1 \times 10^{-6}$ with REG on chromosome 21, 18738520–18760112 bp and $p = 3.5 \times 10^{-6}$ with COLL on chromosome 6, 29817923–29828276 bp) were not significant at any level and the results were not consistent across different tests.

REG ($p = 3.2 \times 10^{-14}$) performed equally well in detecting the *APOC1* block. CMAT showed suggestive evidence ($p = 4.6 \times 10^{-8}$) at a MAF_T of 0.041. The corresponding signal with FRACREG, which accounts for population covariates, was somewhat less significant ($p = 2.0 \times 10^{-7}$), but still suggestive. The performance of the different tests with the *TOMM40* block was similar to that with the entire gene and is therefore not discussed any further.

Under the 3-block strategy, we also obtained similar performance with *TOMM40* and *PVRL2* as before: *TOMM40* was implied by SKAT, while *PVRL2* was identified with the regression test REG. In contrast to the single-block strategy, the signals for *APOC1* drop markedly. *APOC1* was now merged with blocks from *APOC2* and *APOC4* into a single analysis bin containing 25 variants, instead of 8. As a consequence, SKAT ($p = 7.6 \times 10^{-11}$) and REG ($p = 9.7 \times 10^{-10}$) performed worse than under the block strategy, although they still reached experiment-wide significance.

We presumed that the associations in the *APOE* region (table 4) are largely due to the LD with the common *APOE* ε4/ε3/ε2 polymorphism. To investigate this assumption, the genotypes of rs429358 and rs7412 were included as further covariates in a follow-up analysis of the bins described above. Only one nominally significant association result remained, under the 3-block strategy, REG suggested a bin overlapping with *PVRL2* at p = 0.04. In view of the multiple tests performed, however, we have to state that no compelling evidence for residual rare variant association remains after *APOE* is included as covariate. In other words, it can be stated that the rare variant association detected by us is entirely explained by the LD with common variants. Of course, we cannot rule out the possibility that a truly independent rare variant association close to *APOE* exists. Larger samples are needed to ultimately clarify whether there is an independent rare variant association in the *APOE* region.

While the rare variant association of the *APOE* region is entirely driven by the LD with common variation, virtually all methods produced strong signals in at least one analysis bin. In this sense, these primary results are a valuable demonstration of the usefulness and power of the suggested methods in general. Moreover, the *APOE* example suggests that the choice of the analysis bin has a strong impact on the performance of the tests and that a block-based analysis strategy is a potentially powerful supplement to a gene-based analysis. Concatenating of three blocks into a single analysis bin, however, seems to blur association signals by noise caused by the increased number of rare variants that contribute to the test statistic. In general, the non-burden

Table 6. Regression tests with the VT method for the top 3 non-*APOE* results (see table 5)

Strategy	Position (chromosome:bp)	Feature	REG			FRACREG			COLLREG		
			MAF_T	nRV	p	MAF_T	nRV	p	MAF_T	nRV	p
Gene	1:28656262–28660052	*MED18*	0.0132	5	2.0e-05	0.0132	5	1.6e-05	0.0132	5	4.8e-06
	4:44626103–44679950	*YIPF7*	0.0443	58	3.0e-02	0.0123	13	5.5e-01	0.0443	58	2.6e-04
	19:15270636–15307070	*NOTCH3*	0.0280	28	1.3e-01	0.0310	31	1.6e-04	0.0310	31	6.7e-04
Block	3:103703359–103718867	129 kbp to *COLL11A1*	0.0427	5	2.0e-04	0.0427	5	2.4e-06	0.0427	5	1.4e-06
	6:29800142–29810740	MHC region	0.0132	8	9.5e-04	0.0144	10	1.8e-04	0.0132	8	7.3e-05
	6:29818165–29832926	MHC region	0.0083	25	1.0e-02	0.0083	25	1.3e-04	0.0083	25	4.0e-05
3-Block	8:6253932–6278000	*MCPH1*	0.0482	59	1.1e-02	0.0482	59	1.7e-02	0.0455	57	3.5e-06
	21:18738520–18760122	124 kbp to *CXADR*	0.0452	38	1.4e-05	0.0086	3	2.4e-03	0.0086	3	3.1e-03
	6:29817923–29828276	MHC region	0.0083	20	1.4e-02	0.0083	20	4.4e-04	0.0083	20	1.2e-04

For the three candidate bins from the gene-binning strategy, the same tests returned the most significant p values as in the fixed threshold analysis (COLLREG for *MED18* and *YIPF7* with $p = 4.8 \times 10^{-6}$ and $p = 2.9 \times 10^{-4}$, respectively, and FRACREG for *NOTCH3*). In all three cases, the MAF closest to 0.05 was identified as 'optimal', and the p values were larger in the VT analysis. This is not surprising, since in the VT analysis, the p value of the non-permuted data set at the 'optimal' MAF_T is compared to the p values of the permuted data sets at all MAF_T, instead of just one.

In the block-binning strategy, the COLLREG test returned the most significant values out of the three VT regression tests for all three candidate bins ($p = 1.4 \times 10^{-6}$, $p = 7.3 \times 10^{-5}$ and $p = 4.0 \times 10^{-5}$, respectively), with lower 'optimal' MAF_T than in the fixed threshold case. This is consistent with the genome-wide VT permutation analysis, where COLL identified the most significant MAF_T at values close to, and in case of the bin on chromosome 6, 29800142–29810740 bp identical to, the thresholds identified by the COLLREG VT analysis.

In the 3-block-binning strategy, COLLREG was the most significant test ($p = 3.5 \times 10^{-6}$) in case of *MCPH1*, although at a slightly lower MAF, with two SNPs being excluded compared to the fixed threshold analysis. The bin on chromosome 21, 18738520–18760122 bp was most significant with REG ($p = 1.4 \times 10^{-5}$), at the same MAF as in the fixed threshold case. All three VT regression tests identified the optimal MAF for the bin on chromosome 6, 29817923–29828276 bp at MAF = 0.0083, with COLLREG returning the most significant p value ($p = 1.2 \times 10^{-4}$). Overall, while consistent with the results laid out in table 5, none of the regression tests with VT improved the p value enough to come close to experiment-wide significance.

tests REG and SKAT, without the extension to SKAT-O, strongly outperform burden tests in the *APOE* region. Any generalization concerning the comparison of burden and non-burden tests from the *APOE* example, however, has to be treated with caution, since different allelic architectures might drive associated regions with primary causal rare variants. Nevertheless, it is noteworthy that the classic regression test REG performed equally well as the more sophisticated kernel method of SKAT in case of the *APOC1* under the block strategy, which has a low number of rare variants (*nRV* = 8). For a higher number of rare variants, the relative performances of SKAT and REG varied. While only SKAT identified *TOMM40* (*nRV* = 22) at experiment-wide significance, *PVRL2* (*nRV* = 37) was only detected with the regression test REG.

Non-APOE Results

The top 3 non-*APOE* results from each binning strategy are presented in table 5. With the gene-binning strategy, *MED18*, *NOTCH3* and *YIPF7* showed the lowest p values apart from *APOE*, but none of the genes reached experiment-wide significance. All three genes are detected by burden tests. Binning-wide significance is observed for *MED18* with COLLREG ($p = 1.3 \times 10^{-6}$, $MAF_T = 0.013$). The signal is also supported by SKAT-O ($p = 3.7 \times 10^{-6}$). *YIPF7* is also highlighted by COLLREG, while *NOTCH3* is suggested by the FRACREG approach ($p = 2.0 \times 10^{-5}$, $MAF_T = 0.031$).

With the block strategy, COLL ($p = 2.4 \times 10^{-6}$, $MAF_T = 0.043$) identifies an LD block in an intergenic region, about 130 kb from *COLL11A1*. The association signal stayed stable when population covariates were included ($p_{COLLREG} = 3.3 \times 10^{-6}$). In contrast, two blocks from the MHC region, detected by COLL, mostly disappear when they were analyzed with population covariates ($p_{COLLREG} \geq 0.01$).

With the 3-block-binning strategy, COLLREG identified an excess of rare variants ($p = 6.5 \times 10^{-7}$, $MAF_T = 0.048$) in a bin overlapping *MCPH1*. In addition, the non-burden regression test REG ($p = 2.1 \times 10^{-6}$) suggested an intergenic region about 125 kb apart from *CXADR*. The result was to some extent supported by SKAT ($p = 1.4 \times 10^{-4}$). Finally, COLL suggested a region within the MHC region overlapping with the region identified with the block strategy. However, the signal again disappeared when population covariates were included ($p = 4.9 \times 10^{-1}$).

We applied the regression tests with the VT analysis to the 9 non-*APOE* candidates (table 6). While the application of the VT method on the genome-wide level is still

barely practical due to long running times, the computation of select bins took only up to 27 h (for *MCPH1* with COLLREG and 10^7 permutations at 10-fold parallelization.) Although the bin on chromosome 3, 103703359–103718867 bp resulted in an improved p value across all tests (COLLREG, $p = 1.4 \times 10^{-6}$, $MAF_T = 0.0427$), no additional evidence for a true association was obtained as no significance was achieved.

Summing up, we can state that various strategies and tests suggest interesting regions for follow-up. Although none of the association signals outside the *APOE* region withstood multiple testing, some of the implicated regions might trigger replication in independent studies due to their functional characteristics. *MCPH1*, for instance, is associated with microcephaly and brain size [34]. It was found to be highly polymorphic and contains variants under strong positive selection in humans [35]. *NOTCH3* has been implicated in small vessel disease [36]; and a pathogenic variant (not present in our data) has recently been found in a consanguineous family with a high prevalence of AD [37]. *MED18* is a component of the Mediator complex, which is a co-activator for DNA-binding factors that activate transcription via RNA polymerase II [38]. The gene has not been implied in AD or neurodegeneration previously.

Discussion

We typified an analysis pipeline for rare variants obtained from GWAS data via imputation. We demonstrated by application to a real data set that the suggested pipeline results in an overall valid strategy with controlled type I error rates. Computational feasibility is efficiently supported by the INTERSNP-RARE tool for parallelized permutation-based rare variant association testing. The integration of SKAT and SKAT-O into the pipeline is straightforward and also results in valid procedures. The outlined strategy is intended as a supplement to NGS studies. A disadvantage of the suggested pipeline is that it is naturally restricted to variants contained in public databases. As a consequence, very rare variants with high phenotype specificity will be, if at all, only partially captured variants in LD. Furthermore, the quality of the imputed rare variants is still an issue, both concerning type I and type II error rates. Applying stringent QC criteria, we were able to control the type I error rate in the present study. Nevertheless, follow-up of detected regions by direct genotyping or sequencing is generally recommended. The major advantage of the pipeline is that it can be applied to already existing GWAS data. This directly implies coverage of the whole genome and typically the availability of large samples. In view of these advantages, the analysis scheme presented here can be used to guide focused sequencing studies.

We addressed the question of an appropriate partition of genome-wide data into units of analysis by devising three different binning strategies. Seven different rare variant tests were applied to each of the strategies, resulting in 21 distinct genome-wide rare variant substudies. The type I error rate was well controlled under a variety of different settings, involving different binning strategies and statistical tests, which supports the validity of the approach in general. Indeed, we could demonstrate that the genome-wide p value distribution was in concordance with the expectation under the H_0 under all scenarios. Strong evidence for an association with the *APOE* locus was confirmed by each binning strategy at experiment-wide significance, although no independent evidence for an association remained after the two known AD susceptibility SNPs rs429358 and rs7412 were included as covariates. The primary significance of the rare variant tests, however, confirms that rare variant tests have power to capture common variant association via LD. We could show that block-based testing allows a focused analysis and produces stronger association signals than gene-based analysis in several examples. Block-based analysis is thus a valuable additional analysis option. In the *APOE* example, SKAT and the standard logistic regression test REG performed best, while the relative performance of these two methods depended on the choice of the analysis bin. The example of *TOMM40*, detectable with SKAT, and *PVRL2*, detectable with REG, suggests that the application of multiple analysis schemes increases the chances of picking up association signals, even in view of the increased multiple testing burden.

Summing up, we described a straightforward rare variant analysis pipeline for imputed data. The choice of analysis bin is supported and can be combined with tests implemented in SKAT [11]. In addition, various variable threshold tests are provided by our implementation in INTERSNP-RARE. We suggested an analysis procedure including QC criteria and demonstrated its validity in different settings via its application to a real data set. Due to its cost-efficiency, the pipeline constitutes an interesting supplement to NGS rare variant studies.

Acknowledgements

This work was supported by the DFG project BE 3828/3-2 (T.V.). T.B. is an associate member of the ImmunoSensation Cluster of Excellence.

References

1 Slatkin M: Epigenetic inheritance and the missing heritability problem. Genetics 2009; 182:845–850.
2 Maher B: Personal genomes: the case of the missing heritability. Nature 2008;456:18–21.
3 1000 Genomes Project Consortium, Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, et al: An integrated map of genetic variation from 1,092 human genomes. Nature 2012;491:56–65.
4 Kryukov GV, Pennacchio LA, Sunyaev SR: Most rare missense alleles are deleterious in humans: implications for complex disease and association studies. Am J Hum Genet 2007;80:727–739.
5 Gorlov IP, Gorlova OY, Frazier ML, Spitz MR, Amos CI: Evolutionary evidence of the effect of rare variants on disease etiology. Clin Genet 2011;79:199–206.
6 Li B, Leal SM: Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. Am J Hum Genet 2008;83:311–321.
7 Zawistowski M, Gopalakrishnan S, Ding J, Li Y, Grimm S, Zöllner S: Extending rare-variant testing strategies: analysis of noncoding sequence and imputed genotypes. Am J Hum Genet 2010;87:604–617.
8 Morris AP, Zeggini E: An evaluation of statistical approaches to rare variant analysis in genetic association studies. Genet Epidemiol 2010;34.2:188–193.
9 Madsen BE, Browning SR: A groupwise association test for rare mutations using a weighted sum statistic. PLoS Genet 2009;5:e1000384.
10 Mukhopadhyay I, Feingold E, Weeks DE, Thalamuthu A: Association tests using kernel-based measures of multi-locus genotype similarity between individuals. Genet Epidemiol 2010;34:213–221.
11 Lee S, Emond MJ, Bamshad MJ, Barnes KC, Rieder MJ, Nickerson DA; NHLBI GO Exome Sequencing Project – ESP Lung Project Team, Christiani DC, Wurfel MM, Lin X: Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. Am J Hum Genet 2012;91:224–237.
12 Ladouceur M, Dastani Z, Aulchenko YS, Greenwood CMT, Richards JB: The empirical power of rare variant association methods: results from sanger sequencing in 1,998 individuals. PLoS Genet 2012;8:e1002496.
13 Howie BN, Donnelly P, Marchini J: A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genet 2009; 5:e1000529.
14 Li Y, Willer CJ, Ding J, Scheet P, Abecasis GR: MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. Genet Epidemiol 2010;34:816–834.
15 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.
16 Bhatia G, Bansal V, Harismendy O, Schork NJ, Topol EJ, Frazer K, Bafna V: A covering method for detecting genetic associations between rare variants and common phenotypes. PLoS Comput Biol 2010;6:e1000954.
17 Mägi R, Kumar A, Morris AP: Assessing the impact of missing genotype data in rare variant association analysis. BMC Proc 2011; 5(suppl 9):S107.
18 Li Y, Byrnes AE, Li M: To identify associations with rare variants, just WHaIT: weighted haplotype and imputation-based tests. Am J Hum Genet 2010;87:728–735.
19 Lindblad-Toh K, Garber M, Zuk O, Lin MF, Parker BJ, Washietl S, et al: A high-resolution map of human evolutionary constraint using 29 mammals. Nature 2011;478:476–482.
20 ENCODE Project Consortium; Dunham I, Kundaje A, Aldred SF, Collins PJ, Davis CA, Doyle F, et al: An integrated encyclopedia of DNA elements in the human genome. Nature 2012;489:57–74.
21 Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA: Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci USA 2009;106:9362–9367.
22 Li L, Li Y, Browning SR, Browning BL, Slater AJ, Kong X, Aponte JL, Mooser VE, Chissoe SL, Whittaker JC, Nelson MR, Ehm MG: Performance of genotype imputation for rare variants identified in exons and flanking regions of genes. PLoS One 2011;6.9:e24945.
23 Zaitlen N, Eskin E: Imputation aware meta-analysis of genome-wide association studies. Genet Epidemiol 2010;34:537–542.
24 Price AL, Kryukov GV, de Bakker PI, Purcell SM, Staples J, Wei LJ, Sunyaev SR: Pooled association tests for rare variants in exon-resequencing studies. Am J Hum Genet 2010;86: 832–838.
25 Herold C, Steffens M, Brockschmidt FF, Baur MP, Becker T: INTERSNP: genome-wide interaction analysis guided by a priori information. Bioinformatics 2009;25:3275–3281.
26 OpenMP Architecture Review Board. OpenMP Application Program Interface, version 3.1. July 2011.
27 Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, Sham PC: PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 2007;81:559–575.
28 Flicek P, Amode MR, Barrell D, Beal K, Brent S, Carvalho-Silva D, et al: Ensembl 2012. Nucleic Acids Res 2012;40:D84–D90.
29 Wilson EB: Probable inference, the law of succession, and statistical inference. J Am Stat Assoc 1927;22:209–212.
30 Agresti A, Coull BA: Approximate is better than 'exact' for interval estimation of binomial proportions. Am Stat 1998;52:119–126.
31 McKhann G, Drachman D, Folstein M, Katzman R, Price D, Stadlan EM: Clinical diagnosis of Alzheimer's disease: report of the NINCDS-ADRDA Work Group under the auspices of Department of Health and Human Services Task Force on Alzheimer's Disease. Neurology 1984;34:939–944.
32 Aulchenko YS, Ripke S, Isaacs A, van Duijn CM: GenABEL: an R library for genome-wide association analysis. Bioinformatics 2007;23: 1294–1296.
33 Fullerton SM, Clark AG, Weiss KM, Nickerson DA, Taylor SL, Stengård JH, Salomaa V, Vartiainen E, Perola M, Boerwinkle E, Sing CF: Apolipoprotein E variation at the sequence haplotype level: implications for the origin and maintenance of a major human polymorphism. Am J Hum Genet 2000;67: 881–900.
34 Farooq M, Baig S, Tommerup N, Kjaer KW: Craniosynostosis-microcephaly with chromosomal breakage and other abnormalities is caused by a truncating MCPH1 mutation and is allelic to premature chromosomal condensation syndrome and primary autosomal recessive microcephaly type 1. Am J Med Genet A 2010;152A:495–497.
35 Wang Y-Q, Su B: Molecular evolution of microcephalin, a gene determining human brain size. Hum Mol Genet 2004;13:1131–1137.
36 Schmidt H, Zeginigg M, Wiltgen M, Freudenberger P, Petrovic K, Cavalieri M, Gider P, Enzinger C, Fornage M, Debette S, Rotter JI, Ikram MA, Launer LJ, Schmidt R; CHARGE consortium Neurology working group: Genetic variants of the NOTCH3 gene in the elderly and magnetic resonance imaging correlates of age-related cerebral small vessel disease. Brain 2011;134:3384–3397.
37 Guerreiro RJ, Lohmann E, Kinsella E, Brás JM, Luu N, Gurunlian N, Dursun B, Bilgic B, Santana I, Hanagasi H, Gurvit H, Gibbs JR, Oliveira C, Emre M, Singleton A: Exome sequencing reveals an unexpected genetic cause of disease: NOTCH3 mutation in a Turkish family with Alzheimer's disease. Neurobiol Aging 2012;33:1008.e17–1008.e23.
38 Sato S, Tomomori-Sato C, Banks CA, Sorokina I, Parmely TJ, Kong SE, Jin J, Cai Y, Lane WS, Brower CS, Conaway RC, Conaway JW: Identification of mammalian Mediator subunits with similarities to yeast Mediator subunits Srb5, Srb6, Med11, and Rox3. J Biol Chem 2003;278:15123–15127.