bioRxiv preprint doi: https://doi.org/10.1101/020198; this version posted June 2, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

The weighting is the hardest part: on the behavior of the likelihood ratio test and score test under weight misspecification in rare variant association studies

Camelia C. Minică[1@], Giulio Genovese[2,3,4], Dorret I. Boomsma[1], Christina M. Hultman[5], René Pool[1], Jacqueline M. Vink[1], Conor V. Dolan[1*@], Benjamin M. Neale[2,3,6*@]

* These authors contributed equally to this work

[1]Department of Biological Psychology, Vrije Universiteit, Amsterdam, The Netherlands.

[2]The Stanley Center for Psychiatric Research, Broad Institute of the Massachusetts Institute of Technology and Harvard, Cambridge, MA.

[3]The Program in Medical and Population Genetics, Broad Institute of the Massachusetts Institute of Technology and Harvard, Cambridge, MA.

[4]The Department of Genetics, Harvard Medical School, Cambridge, MA.

[5]The Department of Medical Epidemiology and Biostatistics, Karolinska Institute, Stockholm.

[6]The Analytical and Translational Genetics Unit, Department of Medicine, Massachusetts General Hospital and Harvard Medical School, Boston, Massachusetts, USA.

[@]To whom correspondence should be addressed:

Camelia C. Minică, Department of Biological Psychology, Vrije Universiteit Amsterdam, Van der Boechorststraat 1, 1081 BT, Amsterdam, The Netherlands. Telephone number: +31 20 59 83035

Email: camelia.minica@gmail.com

bioRxiv preprint doi: https://doi.org/10.1101/020198; this version posted June 2, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

ABSTRACT

Rare variant association studies are gaining importance in human genetic research with the increasing availability of exome/genome sequence data. One important test of association between a target set of rare variants (RVs) and a given phenotype is the sequence kernel association test (SKAT). Assignment of weights reflecting the hypothesized contribution of the RVs to the trait variance is embedded within any set-based test. Since the true weights are generally unknown, it is important to establish the effect of weight misspecification in SKAT.

We used simulated and real data to characterize the behavior of the likelihood ratio test (LRT) and score test under weight misspecification. Results revealed that LRT is generally more robust to weight misspecification, and more powerful than score test in such a circumstance. For instance, when the rare variants within the target were simulated to have larger betas than the more common ones, incorrect assignment of equal weights reduced the power of the LRT by ~5% while the power of score test dropped by ~30%. Furthermore, LRT was more robust to the inclusion of weighed neutral variation in the test.

To optimize weighting we proposed the use of a data-driven weighting scheme. With this approach and the LRT we detected significant enrichment of case mutations with MAF below 5% (P-value=7E-04) of a set of highly constrained genes in the Swedish schizophrenia case-control cohort of 4,940 individuals with observed exome-sequencing data.

The score test is currently widely used in sequence kernel association studies for both its computational efficiency and power. Indeed, assuming correct specification, in some circumstances the score test is the most powerful test. However, our results showed that LRT has the compelling qualities of being generally more robust and more powerful under weight misspecification. This is a paramount result, given that, arguably, misspecified models are likely to be the rule rather than the exception in the weighting-based approaches.

Keywords: LRT, score, SKAT, variable weighting, robustness, MAF thresholding, schizophrenia.

bioRxiv preprint doi: https://doi.org/10.1101/020198; this version posted June 2, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

INTRODUCTION

With the availability of high-coverage exome/genome sequence data in increasingly large samples, rare variant association studies (RVAS) are gaining importance in human genetic research. One important test of association between a target set of rare variants (RVs) and a given phenotype is the sequence kernel association test (SKAT; (Wu, Lee et al. 2011; Lee, Wu et al. 2012; Chen, Meigs et al. 2013; Ionita-Laza, Lee et al. 2013; Listgarten, Lippert et al. 2013; Lippert, Xiang et al. 2014; Svishcheva, Belonogova et al. 2014)). SKAT is based on a random effects model, in which the effect sizes of the RVs are assumed to be drawn from a zero mean distribution and variance that can be specified by weights. These weights are typically assigned based on meta-information about the RVs, such as allele frequency and functional predictions (Kryukov, Pennacchio et al. 2007; Madsen and Browning 2009; Price, Kryukov et al. 2010; Wu, Lee et al. 2011), with rarer and functional variants expected to have larger effects. Allele frequency, in particular, is an important weighting factor, as the rarer the variant is, the stronger the average purifying selection coefficient (Pritchard 2001; Schork, Murray et al. 2009). If this assumption is true, the effect sizes for rare variants will tend to be larger than for more common variants.

The relationship between effect size, frequency and selection, however, rests on assumptions about the extent of direct selection on the phenotype in question and the demographic history of the population (Eyre-Walker and Keightley 2007; Price, Kryukov et al. 2010; Zuk, Schaffner et al. 2014). Genomic regions under low selection pressures may harbor rare as well as more common so-called ‘goldilocks’ alleles, both with strong functional effects, as simulation studies (Price, Kryukov et al. 2010) and empirical results have demonstrated (e.g., (Cohen, Boerwinkle et al. 2006)). Testing such genomic regions by relying on a weighting scheme which up-weighs rarer variants and puts low or zero weights on the more common ones, may weaken the association signal. Correct weighting is expected to boost

bioRxiv preprint doi: https://doi.org/10.1101/020198; this version posted June 2, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

the power of detection (Wu, Lee et al. 2011). However, as the true weights are generally unknown, it is important to establish the effect of weight misspecification in a kernel-based variance component test. Here we assessed the loss of power associated with incorrect weighting in sequence-based kernel association tests. Because hypothesis testing can be performed by using either the score (Wu, Lee et al. 2011) or the likelihood ratio test (Listgarten, Lippert et al. 2013), we characterize the behavior of both tests within the misspecification space. We considered various weighting schemes and target regions harboring functional variants or mixtures of functional and neutral variants. We show that the choice of the statistical test has an important bearing on power, with the likelihood ratio test being appreciably more robust to weight misspecification. Furthermore, we show that the power loss depends not only on the degree of misspecification, but also on the presence of neutral variants within the target set. As to how to minimize the power loss resulting from misspecification of weights, we examined the efficiency of a data-driven weighting scheme. We propose the use of a set of theoretically defensible weighting schemes, of which, we assume, the one that gives the largest test statistic is likely to capture best the allele frequency-functional effect relationship. The use of alternative weighting schemes is intended to accommodate genomic regions where only very rare variants are likely to be functional, as well as regions under weak selection pressures, harboring both rare and common variants, both (possibly) related to the risk of the disease of interest. Family-wise error rate can be protected either by permutations or by using a Bonferroni correction. We show the power benefits conferred by the use of such a variable data-driven weighting procedure both in simulated and in empirical data.

Below we first formulate the model and briefly consider the two tests of variance components, namely the likelihood ratio test and the score test. Next we explore the behavior of the two tests under (in)correct model specification in a simulation study. We then present and evaluate the use of a data-driven weighting scheme in simulated and empirical data.

bioRxiv preprint doi: https://doi.org/10.1101/020198; this version posted June 2, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Finally, we discuss the robustness of the likelihood ratio test to misspecification and the power advantages conferred by our proposed weighting procedure in SKAT.

METHODS

Model formulation

Let y be the n-dimensional vector of continuous phenotypes measured in a sample consisting of n individuals. Let **X** be the n×p design matrix containing the relevant covariates. Let **G** be the n×m matrix of genotype values, with the g_{ij} element being the genotype value of the individual *i* (*i*=1…n) at locus *j* (*j*=1…m). Genotypes are coded as additive-codominant, i.e., g_{ij}=(0,1,2). The association between the phenotype and the set of **m** SNPs is modeled within the linear mixed model framework as:

$$y = X\beta + Gb + e \qquad \text{(Equation 1)}$$

with $\beta^t = (\beta_1,...,\beta_p)$ being the p-dimensional vector of fixed effects of covariates, $b^t = (b_1,...,b_m)$ being the m-dimensional vector of regression coefficients in the regression of the phenotype on the m genetic variants within the target set, and e being the n-dimensional vector of random residuals. The random vectors b and e are assumed to be normally distributed: $b \sim N(0, I\sigma_b^2)$ and $e \sim N(0, I\sigma_e^2)$, with I being the identity matrix of appropriate dimension.

Let **W** be the m×m diagonal matrix containing the weights used to weigh the contribution to the test statistic of the SNPs in the set. The normally distributed phenotype y has expected mean $E[y] = X\beta$ and variance-covariance matrix:

$$\Sigma_y = E[(y - E(y))(y - E(y))^t] = GWG^t \frac{\sigma_b^2}{m} + I\sigma_e^2 \qquad \text{(Equation 2)}$$

bioRxiv preprint doi: https://doi.org/10.1101/020198; this version posted June 2, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

with GWG^t being the weighted kernel or genetic relationship matrix. As implemented in the SKAT (Wu, Lee et al. 2011), the diagonal elements of the W matrix, $diag(w_1 ... w_m)$, are related to the minor allele frequency of the jth variant by means of the beta density distribution function (dbeta), which is characterized by two shape parameters. The specification of the two shape parameters is informed by the hypothesized relationship between the jth variant effect and its minor allele frequency (MAF; see section on 'Weighting' below).

Tests of variance components

To test whether the parameter of interest σ_b^2 deviates significantly from zero, one can employ a likelihood ratio test (LRT) or a score test. The likelihood ratio test is computed as two times the difference between the log-likelihoods of the null model (σ_b^2 constrained to equal 0) and the alternative model (σ_b^2 estimated freely). Parameter estimation can be performed by restricted/residual maximum likelihood (REML):

$$\mathrm{LogL}\left(\sigma_b^2, \sigma_e^2\right) = 1/2\log\left|\Sigma_y\right| - 1/2\log\left|X^t\Sigma_y^{-1}X\right| - 1/2 r^t\Sigma_y^{-1}r - 1/2(n-p)\log(2\pi)$$

(Equation 3)

where $\mathrm{r} = y - X(X^t\Sigma_y^{-1}X)^{-}X^t\Sigma_y^{-1}y$ with superscript '-' denoting a generalized inverse (Basilevsky 1983).

In evaluating the statistical significance of the restricted LRT, we note that the null distribution of the test statistic is an equally weighted .5:.5 mixture of a χ_0^2 and a χ_1^2 distributions (see e.g., (Stoel, Garre et al. 2006; Visscher 2006; Wu and Neale 2013)). Alternatively, the null distribution can be constructed empirically by using a permutation-based approach (e.g., (Listgarten, Lippert et al. 2013)), or a parametric bootstrap (e.g., (Crainiceanu and Ruppert 2004)).

bioRxiv preprint doi: https://doi.org/10.1101/020198; this version posted June 2, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

The score test is computed as:

$$Q_{SKAT} = (y - X\hat{\beta})^t GWG^t (y - X\hat{\beta}) \qquad \text{(Equation 4)}$$

with its expected null distribution following a mixture of chi-square distribution and statistical significance assessed by means of the Davies exact method (Davies 1980).

Data simulation

Phenotypes and genotypes were generated in samples of n = 10,000 unrelated individuals. Specifically, we simulated two m-dimensional random vectors of continuous variables representing alleles at m equidistant loci for each individual *i* from the sample. The vectors were drawn from a multivariate distribution with zero mean and Σ_{LD} correlation matrix. The SNPs were in linkage equilibrium, i.e., we set Σ_{LD} to equal an identity matrix. The multivariate normally distributed variables were then discretized given chosen thresholds based on the MAF at each locus. We considered MAFs varying randomly between .005 and .05, sampled from a uniform distribution. Given the vectors of alleles, we then created the m vectors of genotypes, g_{ij}. Based on the genotypes, the n×1 vector of phenotypes, **y**, was generated as:

$$y_i = \sum_{j=1}^{m} g_{ij} b_j * \sqrt{\sigma_b^2} + e_i * \sqrt{\sigma_e^2} \qquad \text{(Equation 5)}$$

b_j, the regression weight of the SNP at the jth locus, was computed as a function of MAF_j and of its contribution to the standardized variance of the polygenic scores (Mather and Jinks 1977). Namely, the regression weights varied with MAF, while their contribution to the genetic variance was equal. Simulating data in this fashion is equivalent to simulation according to dbeta(MAF, .5,.5) weights (Wu, Lee et al. 2011), with weights increasing with decreasing

bioRxiv preprint doi: https://doi.org/10.1101/020198; this version posted June 2, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

MAF. We also simulated data according to dbeta (1,1) weights (second simulation scenario), where SNPs had equal weights regardless of MAF. This scenario is illustrative for situations where the tested region harbors both common and rare variants, both having functional effects on the trait (i.e., where there is no relationship between allele frequency and effect size). The variance σ_b^2 equaled .01 across all scenarios we considered, and $\sigma_e^2 = 1 - \sigma_b^2$. The n-dimensional vector of environmental scores e was drawn from a standard normal distribution $N(0,1)$.

Exploring the misspecification space: Weighting

To explore the effect of weight misspecification on the power and type I error rates of the LRT and the score test we carried out simulations. The m-dimensional vector **w** of SNP weights was computed using the beta density function, with the jth element calculated as $w_j = dbeta(MAF_j; a1, a2)$ given the MAF of the jth variant and the shape parameters a1 and a2. As described in the previous section, data were simulated according to: a) dbeta(.5,.5) weights (i.e., the true weights increase with decreasing MAF); and b) dbeta(1,1) weights (i.e., the SNPs have equal weights, regardless of MAF). Next, in computing the tests statistic we (mis)specified the weights as: a) dbeta(1,1); b) dbeta(.5,.5); c) dbeta(1,25). The first weighting scheme pertains to the hypothesis that there is no relationship between the regression weight and the frequency of the variant (hence, the more common variants contribute on average more to variation in the phenotype). In this scenario the association test is carried out with raw additive-codominant coding of the genotypes. The use of the second weighting scheme is equivalent to standardization of the genotypic values prior to the analysis. We considered the effect of this weighting scheme as this treatment of the genotypes is default in GCTA (Yang, Lee et al. 2011) and in FaST-LMM-set (Listgarten, Lippert et al. 2013). Standardization and

bioRxiv preprint doi: https://doi.org/10.1101/020198; this version posted June 2, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

assignment of weights dbeta(.5,.5) are equivalent weighting schemes (Wu, Lee et al. 2011) in which the contribution to the test of rarer variants is up-weighed relative to that of the more common ones (Speed, Hemani et al. 2012), and hence the variants contribute on average equally to the variance in the phenotype, regardless of frequency. We also considered the effects of the third weighting scheme (dbeta(1,25)) as weights computed as such are the default weights in SKAT (Wu, Lee et al. 2011).

We assessed the behavior of the two tests under weight misspecification by considering: a) target regions harboring solely functional variants with opposite effects on the phenotypic mean, and b) regions harboring a mixture of protective, deleterious and neutral effects.

Evaluating the type I error rates and power

We evaluated the type I error rates by generating 1,000,000 datasets under the null hypothesis of no phenotypic variance explained by the SNPs within the target set. The type I error rate was computed as the proportion of datasets in which the tests incorrectly rejected the null hypothesis and it was evaluated given α=.01 and .001.

Power was assessed based on 1000 simulated datasets, an effect size of 1% explained phenotypic variance and 7 alpha thresholds. Given the 7 alpha thresholds, power equaled the proportion of datasets in which the effect was detected. As a validity check of our program, for all the scenarios considered we also report the power and the type I error rates of the true (i.e., correct) model.

Variant weighting schemes: data-driven search for optimal weights

Because the application of a single weighting scheme might not be accurate when testing thousands of genes scattered across the whole exome (possibly subjected to selection pressures of varying intensities) we also considered the efficiency of a data-driven search for

bioRxiv preprint doi: https://doi.org/10.1101/020198; this version posted June 2, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

the optimal weights. We generated 1,000 samples according to weights dbeta(.5,.5) as described above. Each generated sample comprised N=3,000 individuals with phenotypes and genotypes observed at 50 variant sites. As above, the variants were either all functional (deleterious or protective, first scenario) or a mixture of functional and neutrals (second scenario). We performed association tests by using a set of 7 weighting schemes: a) dbeta(1,75) ; b) dbeta(1,50); c) dbeta(1,35); d) dbeta(1,25); e) dbeta(1,5); f) dbeta(1,1) and g) dbeta(.5,.5). Statistical significance was assessed by means of permutations on phenotypes. Specifically, we computed a maximum test statistic max_{LRT} (max_{score}) as the largest out of the seven tests obtained given the genotypes transformed according to each of the weighting schemes enumerated earlier. We then repeated this step in 1,000 permuted datasets obtained by shuffling the phenotypes. We computed the p-value as the proportion of datasets in which the max_{LRT} (max_{score}) was larger in the permuted than in unpermuted data. Power equalled the proportion of datasets yielding a p-value smaller than .01.

Software

The R-package MASS (Venables and Ripley 2002) was used for data generation. Model fitting was performed in R-nlme (Pinheiro, Bates et al. 2014), and SKAT (Lee and Lee 2014). We used the anova function in R to obtain the restricted likelihood ratio test, with the p-value computed by halving the supplied p-value (Pinheiro and Bates 2000). To check our model fitting approach, we analyzed one simulated sample of 10,000 individuals by using 3 independent programs implementing genetic similarity/kernel-based variance component tests: the nlme R-package, the software Genome-wide Complex Trait Analysis (GCTA; (Yang, Lee et al. 2011)) and the software FaST-LMM-set (Listgarten, Lippert et al. 2013). The values for the restricted LRT and the estimate for the variance component obtained by the 3 programs were almost identical (see Table 1 Supplementary Material for details:

bioRxiv preprint doi: https://doi.org/10.1101/020198; this version posted June 2, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

https://www.dropbox.com/s/yk5xe6caeoc4cml/camelia_minica_et_al_Supplementary_Material_Weighting.docx?dl=0), indicating that these are equivalent approaches. Having established the equivalence, all the simulations were next conducted using the nlme program. Simulations were carried out on the Broad Institute Gold Compute cluster.

Empirical analysis: testing the constrained and the FMRP gene sets for rare case mutations enrichment

We compared the performance of the likelihood ratio test and of the score test under alternative weights in a real dataset. For this illustration we used the Swedish schizophrenia case-control cohort of 4,940 individuals with exome-sequencing data from blood DNA. Cases had a clinical diagnosis of schizophrenia and at least two hospitalizations as determined by expert review based on the Hospital Discharge Register (Kristjansson, Allebeck et al. 1987; Dalman, Broms et al. 2002). Controls, without a diagnosis of schizophrenia or bipolar disorder, were randomly selected from population registries. Both cases and controls are of Scandinavian ancestry, aged 18 or older (see (Ripke, O'Dushlaine et al. 2013; Purcell, Moran et al. 2014) for a detailed description of the sample). There were 169 individuals with unreliable samples (i.e., duplicates, ethnic outliers or having a genotype missing rate higher than 10%) whom we removed from the analysis. This left for the analysis 2461 cases and 2479 controls. 2732 of these were males. Written informed consent was obtained from all participants (or legal guardian consent and subject assent). All procedures were approved by the ethical committees in Sweden and in the United States.

Exome-sequencing was performed in seven waves at the Broad Institute of MIT and Harvard. For samples in the first wave, hybrid capture was performed using the Agilent SureSelect Human All Exon Kit method. In this version, the method targets ~28 million base-pairs partitioned in ~160 000 regions. Sequencing was done using Illumina GAII instruments.

bioRxiv preprint doi: https://doi.org/10.1101/020198; this version posted June 2, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

For samples in the waves two to seven, hybrid capture was done by using the newer version of the Agilent SureSelect Human All Exon v.2 Kit method, which targets ~32 million base-pairs partitioned in ~190,000 regions. Sequencing was performed using the Illumina HiSeq 2000 and HiSeq 2500 instruments. We used BWA ALN version 0.5.9 (Li and Durbin 2009) to align the reads to the GRCh37 human genome reference and we applied Picard/GATK to process the sequence data and to call variants (http://broadinstitute.github.io/picard/; (McKenna, Hanna et al. 2010)). Selected singletons were validated using Sanger sequencing (see (Purcell, Moran et al. 2014) for details).

Variants out of Hardy-Weinberg equilibrium (P-value < 5E-8) and showing excess heterozygosity, or variants showing excessive correlation (P-value < 5E-8) with the covariates (that could not be explained by principal components) were excluded from the analysis. In addition, we excluded variants that did not pass the GATK default filters (DePristo, Banks et al. 2011; Auwera, Carneiro et al. 2013). There were 892,306 variants with MAF < 5% meeting all our quality control criteria.

For this empirical illustration we considered the gene-sets rather than the genes as the unit of analysis. The reason that we extended the targeted region is that the current sample sizes afford insufficient power for gene-based tests (see Purcell et al., 2014) but are more adequate for gene-set enrichment analyses which consider jointly a larger number of weak effects. This type of analysis has the added benefit of reducing substantially the burden of multiple testing. By extending the targeted region, the number of tested variants is large, and hence the effects of (possible) weight misspecification are expected to be large. In addition, as we do not focus on a specific class of alleles but rather lump together all observed variants with frequency below specific thresholds, a large amount of variation contributing to the test statistic will possibly be neutral. This makes the example a near optimal situation for illustrating the

bioRxiv preprint doi: https://doi.org/10.1101/020198; this version posted June 2, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

difference in robustness to both model misspecification and neutral variation of the LRT and the score test.

We tested for enrichment of case mutations two partially overlapping gene-sets likely relevant to schizophrenia. The first set consisted of 899 genes which are part of the list identified by Samocha et al. (Samocha, Robinson et al. 2014) as highly constrained. These constrained genes were proposed as candidates in autism spectrum disorder (ASD) given their enrichment for de novo loss of function case mutations. Given evidence favouring the hypothesis that schizophrenia and ASD share genetic aetiology (Consortium 2014; Fromer, Pocklington et al. 2014), this set of genes is likely to be relevant also to schizophrenia. The second set consisted of 749 genes targeted by the Fragile-X mental retardation protein (FMRP). This set is part of the list of genes derived by Darnell et al. (Darnell, Van Driesche et al. 2011) from mouse brain as likely implicated in regulating synaptic plasticity. Genes targeted by FMRP were found to be enriched for de novo nonsynonimous case mutations in both ASD (Iossifov, Ronemus et al. 2012) and schizophrenia (Fromer, Pocklington et al. 2014). Purcell et al. (2014) also tested the FMRP set for enrichment of rare variants in the current sample, and their analysis yielded nominally significant results. Note that the strategy we adopted here is however, different. That is, rather than using gene-based statistic, our procedure tests for the joint effect (variance explained) of rare variants with MAF lower than 5% and 1% within the gene-set (note that the MAF thresholds are, however, arbitrary: variants defined as rare in one sample might feature as common in another sample).

We performed sequence-based kernel association analyses using the likelihood ratio and score tests with variable weights. For this empirical analysis we used the FaST-LMM-Set software (Listgarten, Lippert et al. 2013). To adjust for ancestry we included into analysis the first two principal components. Principal components were computed from genotypes at variants shared with the 1000 Genomes Project phase 1 dataset. To accommodate genomic

bioRxiv preprint doi: https://doi.org/10.1101/020198; this version posted June 2, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

regions where only very rare variants are likely to be functional, as well as regions under weak selection pressures, harboring both rare and more common variants, both (possibly) related to the risk of disease (regardless of frequency), we used three alternative weighting schemes: dbeta(1,25), dbeta(.5,.5) and dbeta(1,1). To reduce the computational burden, we chosen to adapt our alpha for multiple testing rather than to rely on permutations to compute the P-value. Hence for each tested pathway, we chose the P-value corresponding to the weighting scheme that yields the largest test statistic. An alpha of 0.05/12= 0.004 was used, corrected for multiple hypothesis testing of 2 gene-sets, 2 frequency thresholds and 3 weighting schemes. For computational ease we used a linear model (Listgarten, Lippert et al. 2013). The linear LRT (and the linear score test) shows good control of the type I error rate and has performed as well as a generalized linear model in case-control samples (see (Lippert, Xiang et al. 2014)).

RESULTS

Type I error

Tables 1 and 2 contain the results pertaining to the type I error rates of the two tests, given correct and incorrect model specification.

-- Table 1 & Table 2 --

Across all conditions evaluated here, the score test shows good control of the type I error rate. The likelihood ratio test appears slightly conservative, regardless of whether the weights are correctly specified or misspecified. A similar result was reported by Listgarten et al. (Listgarten, Lippert et al. 2013) who suggested that relying on a .5:.5 mixture of a χ_0^2 and a χ_1^2 distribution to assess statistical significance of the one variance component LRT might

bioRxiv preprint doi: https://doi.org/10.1101/020198; this version posted June 2, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

be conservative. We used this approach in the simulations as this is default in most statistical software (e.g., in GCTA, (Yang, Lee et al. 2011)). Alternatively, Listgarten et al. (2013) proposed a permutation based approach to construct the null distribution of the test statistic, approach that maintains the type I error rate of the restricted LRT closer to the expectation. This approach, however, is computationally demanding especially when the number of tested variants within the target and the sample is large.

Power

Figure 1 displays the results relating to power.

-- Figure 1 --

Four important conclusions follow from our simulation results. First, the restricted LRT and the score test have equal power under correct weight specification. This is expected, as the two tests are asymptotically equivalent when the model is true, i.e., correctly specified (e.g., (Greene 2003)). The powers of the two tests - displayed in grey in the power figures - are indistinguishable when the assigned weights correspond to the true weights.

Second, misspecification of weights always reduces power. This is shown in Figure 1 and in Figure 2, as the departure of the power under model misspecification (the black lines) from the power of the true models (the grey lines). The exact loss in power depends on the degree of weight misspecification and on the statistical test employed. We note that the power loss is relatively small given mild misspecification of weights. This result is illustrated in Figure 1A, where the assigned weights dbeta(1,25) resemble the true weights dbeta(.5,.5). In this circumstance, it is mainly the presence of neutral SNPs in the target that dilutes the power (see Figure 1C). However, the power may suffer dramatically with increasing misspecification.

bioRxiv preprint doi: https://doi.org/10.1101/020198; this version posted June 2, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

For instance, when data were simulated according to the dbeta(.5,.5) weights, using a dbeta (1,1) weighting scheme (equal weights assigned to all variants) results in a loss in power of up to ~ 5% and ~30% for the restricted LRT and for the score test, respectively (see Figures 1B and 1D). This result is informative for RVASs in which the raw genotypes (unweighted) are used in the test of association. A more dramatic power loss is illustrated in Figure 2D where we consider the reverse situation: weights dbeta (.5,.5) are assigned to SNPs simulated under flat weights. That is, in this scenario, the allele frequency is incorrectly used to inform on the weights assignment. With this misspecification the drop in power relative to the true model is ~17% and ~80% for the restricted LRT and for the score test, respectively.

Third, the inclusion of neutral SNPs dilutes the power of both tests. In our examples, with 40% neutral SNPs the power drops are in the range of ~10%-~17% relative to the power of the true model, regardless of the degree of weight misspecification. Clearly, discarding neutral variation present within the target is beneficial to improve power to detect significant association.

Forth, relative to the score test, we note that the restricted LRT is consistently more robust, both to weight misspecification and to the presence of neutral variation in the target region. These results are consistent with those reported by Lippert et al. (Lippert, Xiang et al. 2014), who found their proposed LRT to be generally more powerful than the score test across their simulated settings. Although Lippert et al. did not consider the behavior of the two tests under misspecified weights, they reported the same pattern of results in real data analysis, where the LRT yielded consistently more associations than the score test. As the real weights are in all likelihood not known, the superior power of the restricted LRT in real data might be explained as well by its robustness to weight misspecification and to the inclusion of weighed neutral variation in the computation of the test statistic.

bioRxiv preprint doi: https://doi.org/10.1101/020198; this version posted June 2, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Variable weighting schemes: data-driven search for optimal weights

Table 3 contains the simulation results relating to the power of the two tests under alternative weighting schemes.

-- Table 3 --

On one hand, the likelihood ratio test appears to benefit from the use of variable weights. When the variants within the target are all functional, the use of alternative weights increases its power relative to the incorrect weighting setting (power goes up from 48.3% to 52.9%). This increase is to be larger (about 7% given alpha of 0.01) when neutral variants are also present in the target set. On the other hand, the score test performs worst in the variable weighting setting. Power goes down about 10% relative to the incorrect weighting setting, both when the target set contains only functional variants or a mixture of functional and neutral variants.

It should be noted, however, that there is a price to pay in terms of power by using a variable weighting scheme in contrast to correct weighting. The price is largest for regions containing mixtures of functional and neutral variants (e.g., the power of the LRT decreases from 70.9% given correct weights to 56.4% with the variable weighting approach) and relatively small for the (less realistic) scenarios in which the target set contains only functional variants (i.e., with the LRT, the power drops about 2%) .

As typically the true weights are unknown, conjecturing the correct ones by employing alternative weights and using the likelihood ratio test appears to be the strategy likely to maintain the power close to that of the true model. This strategy appears to be advantageous especially when the target set contains also neutral variants. However, by being based on

bioRxiv preprint doi: https://doi.org/10.1101/020198; this version posted June 2, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

permutations, the variable weighting approach is likely to be computationally too complex when the number of tests and the sample is large.

Empirical analysis: testing the constrained and the FMRP gene sets for rare case mutations enrichment

We also looked at the behavior of the score test and of the likelihood ratio test (Listgarten, Lippert et al. 2013) under variable weights in the empirical dataset. Table 4 displays results pertaining to the enrichment tests in the gene-set-based analyses.

-- Table 4 --

From Table 4 we note that the likelihood ratio test appears more powerful than the score test across all conditions evaluated here. It is likely the combination of weight misspecification coupled with the presence of neutral variation in the target set that yielded the difference in power between the two tests. With the current sample and the likelihood ratio test with weights dbeta(1,1), the set of constrained genes showed significant enrichment for disruptive case mutations with MAF below 5% (i.e., P-value = 7E-04; see Table 4 A). The score test under flat weights (i.e., dbeta(1,1)) with its associated p-value also passed the significance threshold, providing support for enrichment for disruptive rare case mutations of the constrained gene-set, although the evidence was weaker (P-value=0.0031).

Note the difference in the strength of association of the two tests under variant weighting schemes. For instance, in the 5% MAF threshold analyses, the enrichment signal in the constrained gene-set was rendered non-significant when the dbeta(1,25) weights were used with the score test (p-value = 0.0331), and yet it reached statistical significance when the likelihood ratio test was employed instead (p-value=0.0037). Had one relied on the score test

bioRxiv preprint doi: https://doi.org/10.1101/020198; this version posted June 2, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

and a default weighting scheme, the association signals in this pathway would have been missed.

The FMRP-Darnell gene-set showed no significant enrichment for rare case mutations, regardless of the test, MAF threshold and weighting schemes used. This result does not rule out the possibility that rarer variants (e.g., singletons) within the pathway play a role in the liability to schizophrenia phenotype. To implicate such variants, however, testing approaches other than those exploiting genetic similarity among the individuals are required.

The 1% MAF threshold yielded similar differences among the two tests (see table 1B). Note that the signal in the constrained gene-set no longer reached statistical significance. This result suggests that imposing this threshold probably removed from the target causal variants and so, weakened the association signal.

Summarizing, the empirical analysis showed that the choice of the test and of the weighting scheme is no trivial matter. The LRT always yielded smaller p-values than the score test, probably due to the greater sensitivity the latter has to weighed neutral variation and to model misspecification (as we found in the simulated data). We also found that either thresholding or relying on default weights would trick one into missing association signals. We elaborate on these results in the Discussion.

DISCUSSION

We characterized the behavior of the likelihood ratio test and of the score test under weight misspecification in association studies based on the rare variant sequence kernel test. The principal finding of this study is that the likelihood ratio test is generally more robust to weight misspecification, and more powerful than the score test in such a circumstance. Our results are of interest because weight assignment is embedded within any set-based test and the true weights of the variants within the target are generally unknown.

bioRxiv preprint doi: https://doi.org/10.1101/020198; this version posted June 2, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

As we found the power to be maximal under correct model specification, we next considered the issue of optimizing weighting. In the literature, weighting is mostly informed by allele frequency; frequency is taken as indicative of the strength of the purifying selection coefficient (Kryukov, Pennacchio et al. 2007). Accordingly, rarer variants are typically being assigned larger weights/contribution to the test statistic (e.g., (Wu, Lee et al. 2011)). This relationship between effect size, frequency and selection is not always straightforward, however, because it relies on assumptions about the extent of direct selection on the phenotype in question and the demographic history of the population (Eyre-Walker and Keightley 2007; Price, Kryukov et al. 2010; Zuk, Schaffner et al. 2014). Genes under weak selection may harbor rare as well as more common variants with disruptive effects (Zuk, Schaffner et al. 2014). Such variants with deleterious effects, escaping selection and occurring at relatively high frequencies in the population, are plausible also under strong purifying selection, as simulation studies have demonstrated (Price, Kryukov et al. 2010). Achieving maximal power when testing such regions requires adapting the weighting scheme to match the hypothesized selection. To this end, we proposed the use of a data-driven weighting approach. Our simulation results showed that such an approach maintains the power close to that of the true (i.e., correctly specified) model. When applied to real data, this approach allowed us to capture significant enrichment signal coming from variants with MAF below 5% within the constrained pathway (Samocha, Robinson et al. 2014); P-value = 7E-04), lending support to the conclusion that such a variable weighting approach is likely to boost statistical power. Such adaptive approaches were also recommended by Zuk et al. (2014) and by Price et al. (2010) as being optimal for gene-based tests. Deriving weights based on allele frequency is but one of the possible ways of prioritizing the contribution to the test statistic of the variants within the target set (Wu, Lee et al. 2011). Alternative weighting schemes that incorporate probabilities of a variant being damaging (as

bioRxiv preprint doi: https://doi.org/10.1101/020198; this version posted June 2, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

estimated by annotation tools such as e.g., Polyphen-2 (Adzhubei, Schmidt et al. 2010) or SIFT (Ng and Henikoff 2003) may also be considered.

It should be emphasized that our variable weighting approach renders thresholding unnecessary. Thresholding (either based on counts or on allele frequency) has been initially used in burden tests (e.g., (Li and Leal 2008; Madsen and Browning 2009; Price, Kryukov et al. 2010); see also (Franić, Dolan et al. 2015) for an overview on burden tests), but it has been employed also in sequence-based variance component tests (e.g., (Lohmueller, Sparsø et al. 2013; Xu, Tachmazidou et al. 2014)) for the purpose of removing neutral variation (see e.g., (Kryukov, Pennacchio et al. 2007)). Yet, in our empirical analysis this practice was counterproductive: imposing the (arbitrarily chosen) 1% MAF threshold reduced the association signal in the constrained gene-set below the significance threshold. Considering common variants along with the rare ones in sequence-based kernel association tests appears to be justified for three main reasons. First, the use of variable weighting schemes is equivalent to applying variable frequency thresholds: the weights are removing from the test or favoring the contribution to the test statistic of the variants within the target set based on their frequency. Second, only the joint signal – coming from rare and more common variants - enabled us to detect significant enrichment. And third, importantly, with the current samples, our tests are mostly powered to locate regions under relatively weak selection pressures, and such regions are expected to harbour rare as well as common variants both with functional effects. To locate genes under stronger selection pressures, larger samples (see (Zuk, Schaffner et al. 2014)) and the inclusion of more extreme weights (i.e., weights that overlook common variants and favour the rarer ones) will probably be required.

In the empirical analysis, we chose to correct out alpha in place of using permutations to compute the p-value. The data-driven weighting approach based on permutations is prohibitively slow when the number of tested variants within the target set (or the number of

bioRxiv preprint doi: https://doi.org/10.1101/020198; this version posted June 2, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

genes) and the sample is large. The Bonferroni correction though easier computationally, comes at a price in terms of power: the more weighting schemes one 'tries', the more stringent the significance threshold correction. An optimization algorithm for an optimal search for the 'true' weights (e.g., (Neale and Cardon 1992)) or limiting the choice of weights based on knowledge on theorized selection on each gene (Zuk, Schaffner et al. 2014) would decrease the burden of multiple testing, and further increase power.

The score test is currently widely used in sequence-based association studies (e.g., (Huyghe, Jackson et al. 2013; Zhan, Larson et al. 2013; Cruchaga, Karch et al. 2014; Peloso, Auer et al. 2014)) for both its computational efficiency and power (Wu, Lee et al. 2011). Indeed, assuming correct specification, in some circumstances the score test is the most powerful test (Wu, Lee et al. 2011; Lippert, Xiang et al. 2014). However, the results provided herein showed that the likelihood ratio test has the compelling qualities of being generally more robust and more powerful under weight misspecification. This is a paramount result, given that, arguably, misspecified models are likely to be the rule rather than the exception in the weighting-based approaches.

bioRxiv preprint doi: https://doi.org/10.1101/020198; this version posted June 2, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Table 1: Type I error for the restricted likelihood ratio test (RLRT) and the score test, given genotypic data simulated under the null model of no association between the target region and the phenotype. The sample consisted of 10,000 individuals with genotypes at 50 SNPs having minor allele frequencies (MAFs) sampled from the uniform distribution and ranging from .5% to 5%. The restricted LRT and the score tests were computed for three sets of weights beta in each of the 1,000,000 simulated samples. Type I error equals the proportion of datasets in which the null hypothesis has been incorrectly rejected given the three significance thresholds.

	weights dbeta	alpha=.01 *[99%CI]*	alpha=.001 *[99%CI]*
LRT	(.5,.5)	0.00849 *[0.00826, 0.00873]*	0.00083 *[0.00076, 0.00091]*
	(1,1)	0.00834 *[0.00811, 0.00858]*	0.0008 *[0.00073, 0.00088]*
	(1,25)	0.00847 *[0.00824, 0.00871]*	0.00082 *[0.00075, 0.0009]*
Score test	(.5,.5)	0.00991 *[0.00966, 0.01017]*	0.00098 *[0.0009, 0.00107]*
	(1,1)	0.00992 *[0.00967, 0.01018]*	0.00099 *[0.00091, 0.00107]*
	(1,25)	0.00988 *[0.00962, 0.01013]*	0.00098 *[0.0009, 0.00106]*

bioRxiv preprint doi: https://doi.org/10.1101/020198; this version posted June 2, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Table 2: Type I error for the restricted likelihood ratio test (LRT) and the score test, given genotypic data simulated under the null model of no association between the target region and the phenotype. The sample consisted of 10,000 individuals with genotypes at 50 SNPs having equal beta weights and minor allele frequencies (MAFs) sampled from the uniform distribution and ranging from .5% to 5%. The LRT and the score tests were computed for three sets of weights beta in each of the 1,000,000 simulated samples. Type I error equals the percent of datasets in which the null hypothesis has been incorrectly rejected given the three significance thresholds.

	weights dbeta	alpha=.01 *[99%CI]*	alpha=.001 *[99%CI]*
LRT	(.5,.5)	0.00844 *[0.00821, 0.00868]*	0.0008 *[0.00073, 0.00088]*
	(1,1)	0.00844 *[0.00821, 0.00868]*	0.0008 *[0.00073, 0.00088]*
	(1,25)	0.00817 *[0.00794, 0.00841]*	0.00074 *[0.00067, 0.00082]*
Score test	(.5,.5)	0.00989 *[0.00964, 0.01015]*	0.00099 *[0.00091, 0.00107]*
	(1,1)	0.0098 *[0.00954, 0.01005]*	0.00098 *[0.00090, 0.00106]*
	(1,25)	0.00993 *[0.00968, 0.01019]*	0.00094 *[0.00086, 0.00102]*

bioRxiv preprint doi: https://doi.org/10.1101/020198; this version posted June 2, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Table 3: Power results for the variable threshold approach given all variants being functional in the target set (A) or a mixture of protective, deleterious and neutral variants (B). Power was evaluated given alpha of 0.01 in 1000 simulated samples, each sample comprinsing 3000 individuals. The minor allele frequency of the 50 variants within the target set ranged from .5% to 5%.

A.

weights dbeta	LRT	Score test
(1,1) (Incorrect)	0.483	0.469
Variable weights	0.529	0.364
(.5,.5) (True)	0.547	0.559

B.

weights dbeta	LRT	Score test
(1,1) (Incorrect)	0.498	0.471
Variable weights	0.564	0.382
(.5,.5) (True)	0.709	0.734

bioRxiv preprint doi: https://doi.org/10.1101/020198; this version posted June 2, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Table 4: Results of the gene-set enrichment analysis run in the Swedish sample (N=4940; prevalence in the sample = 0.49). The 2 gene-sets included variants with MAF below 5% (A) or below 1% (B).

A.

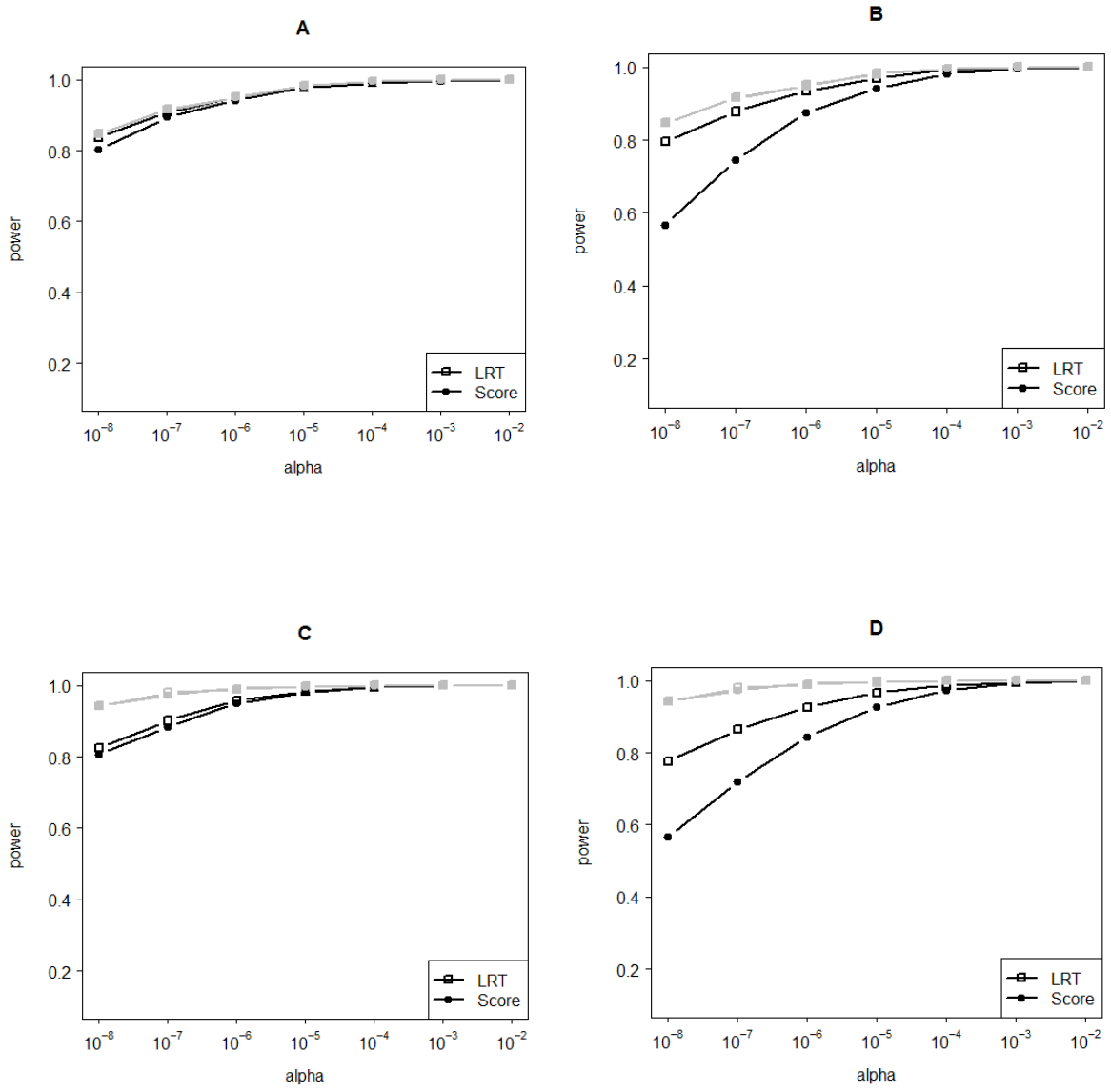
Gene-set (Autosome variants in set)	Weights dbeta	LRT	Score
CONSTRAINED (63,492)	(1,1)	**7E-04**	0.0031
	(.5,.5)	0.1240	0.3444
	(1,25)	0.0037	0.0331
FMRP-Darnell (72,161)	(1,1)	**0.0339**	0.0577
	(.5,.5)	0.1062	0.3384
	(1,25)	0.0434	0.1319

B.

Gene-set (Autosome variants in set)	Weights dbeta	LRT	Score
CONSTRAINED (61,269)	(1,1)	0.0373	0.1139
	(.5,.5)	0.2341	0.3988
	(1,25)	**0.0357**	0.1293
FMRP-Darnell (69,668)	(1,1)	0.0723	0.1679
	(.5,.5)	0.1467	0.3621
	(1,25)	**0.0556**	0.1668

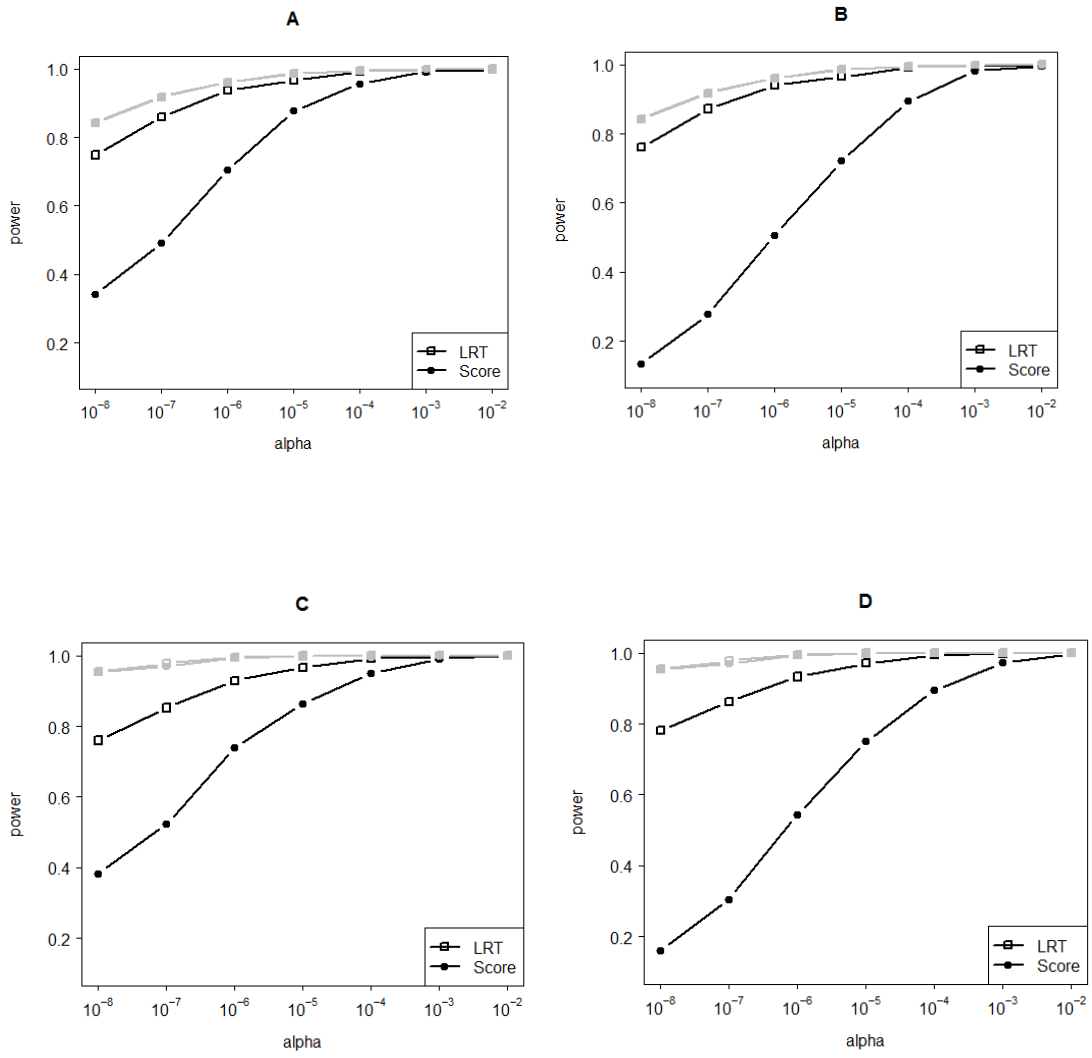
bioRxiv preprint doi: https://doi.org/10.1101/020198; this version posted June 2, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Figure 1: The power of the likelihood ratio test (LRT) and the score test to detect a gene harboring 50 low-frequency SNPs: all functional (A and B) or a mixture of 30 functional and 20 neutral SNPs (C and D). We randomly sampled MAFs ranging from .5% to 5% from the uniform distribution. The gene explains 1% of the phenotypic variance. Genotypic data were simulated according to weights dbeta(.5,.5), models were fitted according to weights dbeta(1,25) (A and C) and dbeta(1,1) (B and D). The power of the true models (i.e., with correct weights) is displayed in grey. Power was evaluated in 1000 datasets consisting of 4000 individuals. Note that while the SNP-set explain the same amount of phenotypic variance (i.e., 1%) across all scenarios considered, the true individual SNP weights increase as the proportion of functional SNPs in the set decreases.



bioRxiv preprint doi: https://doi.org/10.1101/020198; this version posted June 2, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Figure 2: The power of the likelihood ratio test (LRT) and the score test to detect a gene harboring 50 low-frequency SNPs: all functional (A and B) or a mixture of 30 functional and neutral SNPs (C and D). We randomly sampled MAFs ranging from .5% to 5% from the uniform distribution. The gene explains 1% of the phenotypic variance. Genotypic data were simulated according to weights dbeta(1,1), models were fitted according to weights dbeta(1,25) (A and C) and dbeta(.5,.5) (B and D). The power of the true models (i.e., with correct weights) is displayed in grey. Power was evaluated in 1000 datasets consisting of 4000 individuals. Note that while the SNP-set explain the same amount of phenotypic variance (i.e., 1%) across all scenarios considered, the true individual SNP weights increase as the proportion of functional SNPs in the set decreases.



bioRxiv preprint doi: https://doi.org/10.1101/020198; this version posted June 2, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Acknowledgements

We thank to the Swedish cohort participants whose data we analyzed in this study. This research was done while Camelia Minica visited Benjamin Neale at the Broad Institute of MIT and Harvard. Camelia Minica and Jacqueline Vink are supported by the ERC starting grant 284167 (PI-JMV). For this research visit, Camelia Minica was supported also by the Talent Grant FPPT1404 offered by the Scientific and Ethical Review Board and the Faculty Board Vrije Universiteit Amsterdam. The authors wish to acknowledge the generous use of the Broad Institute compute cluster in this work.

bioRxiv preprint doi: https://doi.org/10.1101/020198; this version posted June 2, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

REFERENCES

Adzhubei, I. A., S. Schmidt, et al. (2010). "A method and server for predicting damaging missense mutations." Nature methods **7**(4): 248-249.

Auwera, G. A., M. O. Carneiro, et al. (2013). "From FastQ Data to High Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline." Current protocols in bioinformatics **doi: 10.1002/0471250953.bi1110s43**.

Basilevsky, A. (1983). Applied matrix algebra in the statistical sciences. New York, Elsevier Science Publishing.

Chen, H., J. B. Meigs, et al. (2013). "Sequence kernel association test for quantitative traits in family samples." Genetic epidemiology **37**(2): 196-204.

Cohen, J. C., E. Boerwinkle, et al. (2006). "Sequence variations in PCSK9, low LDL, and protection against coronary heart disease." New England Journal of Medicine **354**(12): 1264-1272.

Consortium, S. W. G. o. t. P. G. (2014). "Biological insights from 108 schizophrenia-associated genetic loci." Nature **511**(7510): 421-427.

Crainiceanu, C. M. and D. Ruppert (2004). "Likelihood ratio tests in linear mixed models with one variance component." Journal of the Royal Statistical Society: Series B (Statistical Methodology) **66**(1): 165-185.

Cruchaga, C., C. M. Karch, et al. (2014). "Rare coding variants in the phospholipase D3 gene confer risk for Alzheimer/'s disease." Nature **505**(7484): 550-554.

Dalman, C., J. Broms, et al. (2002). "Young cases of schizophrenia identified in a national inpatient register." Social psychiatry and psychiatric epidemiology **37**(11): 527-531.

Darnell, J. C., S. J. Van Driesche, et al. (2011). "FMRP stalls ribosomal translocation on mRNAs linked to synaptic function and autism." Cell **146**(2): 247-261.

Davies, R. (1980). "The distribution of a linear combination of chi-square random variables " J. R. Stat. Soc. Ser. C Appl. Stat. **29**: 323–333.

DePristo, M. A., E. Banks, et al. (2011). "A framework for variation discovery and genotyping using next-generation DNA sequencing data." Nature genetics **43**(5): 491-498.

Eyre-Walker, A. and P. D. Keightley (2007). "The distribution of fitness effects of new mutations." Nature Reviews Genetics **8**(8): 610-618.

Franić, S., C. V. Dolan, et al. (2015). "Mendelian and polygenic inheritance of intelligence: A common set of causal genes? Using next-generation sequencing to examine the effects of 168 intellectual disability genes on normal-range intelligence." Intelligence **49**: 10-22.

Fromer, M., A. J. Pocklington, et al. (2014). "De novo mutations in schizophrenia implicate synaptic networks." Nature **506**(7487): 179-184.

bioRxiv preprint doi: https://doi.org/10.1101/020198; this version posted June 2, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Greene, W. H. (2003). "Econometric Analysis." New Jersey: Prentice Hall.

Huyghe, J. R., A. U. Jackson, et al. (2013). "Exome array analysis identifies new loci and low-frequency variants influencing insulin processing and secretion." Nature genetics **45**(2): 197-201.

Ionita-Laza, I., S. Lee, et al. (2013). "Sequence kernel association tests for the combined effect of rare and common variants." The American Journal of Human Genetics **92**(6): 841-853.

Iossifov, I., M. Ronemus, et al. (2012). "De novo gene disruptions in children on the autistic spectrum." Neuron **74**(2): 285-299.

Kristjansson, E., P. Allebeck, et al. (1987). "Validity of the diagnosis schizophrenia in a psychiatric inpatient register: a retrospective application of DSM-III criteria on ICD-8 diagnoses in Stockholm county." Nordic Journal of Psychiatry **41**(3): 229-234.

Kryukov, G. V., L. A. Pennacchio, et al. (2007). "Most rare missense alleles are deleterious in humans: implications for complex disease and association studies." The American Journal of Human Genetics **80**(4): 727-739.

Lee, S., M. C. Wu, et al. (2012). "Optimal tests for rare variant effects in sequencing association studies." Biostatistics **13**(4): 762-775.

Lee, S. S. and M. S. S. Lee (2014). Package 'SKAT'.

Li, B. and S. M. Leal (2008). "Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data." The American Journal of Human Genetics **83**(3): 311-321.

Li, H. and R. Durbin (2009). "Fast and accurate short read alignment with Burrows–Wheeler transform." Bioinformatics **25**(14): 1754-1760.

Lippert, C., J. Xiang, et al. (2014). "Greater power and computational efficiency for kernel-based association testing of sets of genetic variants." Bioinformatics **30**(22): 3206–3214.

Listgarten, J., C. Lippert, et al. (2013). "A powerful and efficient set test for genetic markers that handles confounders." Bioinformatics **29**(12): 1526-1533.

Lohmueller, K. E., T. Sparsø, et al. (2013). "Whole-exome sequencing of 2,000 Danish individuals and the role of rare coding variants in type 2 diabetes." The American Journal of Human Genetics **93**(6): 1072-1086.

Madsen, B. E. and S. R. Browning (2009). "A groupwise association test for rare mutations using a weighted sum statistic." PLoS genetics **5**(2): e1000384.

Mather, K. and J. L. Jinks (1977). Introduction to Biometrical Genetics, Ithaca, NY: Cornell University Press.

bioRxiv preprint doi: https://doi.org/10.1101/020198; this version posted June 2, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

McKenna, A., M. Hanna, et al. (2010). "The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data." Genome research **20**(9): 1297-1303.

Neale, M. and L. Cardon (1992). Methodology for genetic studies of twins and families, Springer Science & Business Media.

Ng, P. C. and S. Henikoff (2003). "SIFT: Predicting amino acid changes that affect protein function." Nucleic acids research **31**(13): 3812-3814.

Peloso, G. M., P. L. Auer, et al. (2014). "Association of low-frequency and rare coding-sequence variants with blood lipids and coronary heart disease in 56,000 whites and blacks." The American Journal of Human Genetics **94**(2): 223-232.

Pinheiro, J., D. Bates, et al. (2014). "nlme: Linear and Nonlinear Mixed Effects Models." R package version **3**: 118.

Pinheiro, J. C. and D. M. Bates (2000). Mixed-effects models in S and S-PLUS, Springer Science & Business Media.

Price, A. L., G. V. Kryukov, et al. (2010). "Pooled association tests for rare variants in exon-resequencing studies." The American Journal of Human Genetics **86**(6): 832-838.

Pritchard, J. K. (2001). "Are rare variants responsible for susceptibility to complex diseases?" The American Journal of Human Genetics **69**(1): 124-137.

Purcell, S. M., J. L. Moran, et al. (2014). "A polygenic burden of rare disruptive mutations in schizophrenia." Nature **506**(7487): 185-190.

Ripke, S., C. O'Dushlaine, et al. (2013). "Genome-wide association analysis identifies 13 new risk loci for schizophrenia." Nature genetics **45**(10): 1150-1159.

Samocha, K. E., E. B. Robinson, et al. (2014). "A framework for the interpretation of de novo mutation in human disease." Nature genetics **46**(9): 944-950.

Schork, N. J., S. S. Murray, et al. (2009). "Common vs. rare allele hypotheses for complex diseases." Current opinion in genetics & development **19**(3): 212-219.

Speed, D., G. Hemani, et al. (2012). "Improved heritability estimation from genome-wide SNPs." The American Journal of Human Genetics **91**(6): 1011-1021.

Stoel, R. D., F. G. Garre, et al. (2006). "On the likelihood ratio test in structural equation modeling when parameters are subject to boundary constraints." Psychological Methods **11**(4): 439.

Svishcheva, G. R., N. M. Belonogova, et al. (2014). "FFBSKAT: fast family-based sequence kernel association test." PloS one **9**(6): e99407.

Venables, W. N. and B. D. Ripley (2002). Modern applied statistics with S, Springer Science & Business Media.

bioRxiv preprint doi: https://doi.org/10.1101/020198; this version posted June 2, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Visscher, P. M. (2006). "A note on the asymptotic distribution of likelihood ratio tests to test variance components." Twin research and human genetics **9**(04): 490-495.

Wu, H. and M. C. Neale (2013). "On the likelihood ratio tests in bivariate ACDE models." Psychometrika **78**(3): 441-463.

Wu, M. C., S. Lee, et al. (2011). "Rare-variant association testing for sequencing data with the sequence kernel association test." The American Journal of Human Genetics **89**(1): 82-93.

Xu, C., I. Tachmazidou, et al. (2014). "Estimating Genome-Wide Significance for Whole-Genome Sequencing Studies." Genetic epidemiology **38**(4): 281-290.

Yang, J., S. H. Lee, et al. (2011). "GCTA: a tool for genome-wide complex trait analysis." The American Journal of Human Genetics **88**(1): 76-82.

Zhan, X., D. E. Larson, et al. (2013). "Identification of a rare coding variant in complement 3 associated with age-related macular degeneration." Nature genetics **45**(11): 1375-1379.

Zuk, O., S. F. Schaffner, et al. (2014). "Searching for missing heritability: designing rare variant association studies." Proceedings of the National Academy of Sciences **111**(4): E455-E464.