Twin Research and Human Genetics
Volume 20 ▌ Number 2 ▌ pp. 108–118 ▌ © The Author(s) 2017 ▌ doi:10.1017/thg.2017.7

The Weighting is the Hardest Part: On the Behavior of the Likelihood Ratio Test and the Score Test Under a Data-Driven Weighting Scheme in Sequenced Samples

Camelia C. Minică,[1,2] Giulio Genovese,[3,4,5] Christina M. Hultman,[6] René Pool,[1,2] Jacqueline M. Vink,[7] Michael C. Neale,[1,8] Conor V. Dolan,[1,2,*] and Benjamin M. Neale[3,4,9,*]

[1]Department of Biological Psychology, Vrije Universiteit, Amsterdam, The Netherlands
[2]The EMGO+ Institute for Health and Care Research, Amsterdam, The Netherlands
[3]The Stanley Center for Psychiatric Research, Broad Institute of the Massachusetts Institute of Technology and Harvard, Cambridge, MA
[4]The Program in Medical and Population Genetics, Broad Institute of the Massachusetts Institute of Technology and Harvard, Cambridge, MA
[5]The Department of Genetics, Harvard Medical School, Cambridge, MA
[6]The Department of Medical Epidemiology and Biostatistics, Karolinska Institute, Stockholm
[7]Behavioural Science Institute, Radboud University, Nijmegen, The Netherlands
[8]Virginia Institute for Psychiatric and Behavioral Genetics, Virginia Commonwealth University, Richmond, USA
[9]The Analytical and Translational Genetics Unit, Department of Medicine, Massachusetts General Hospital and Harvard Medical School, Boston, Massachusetts, USA

Sequence-based association studies are at a critical inflexion point with the increasing availability of exome-sequencing data. A popular test of association is the sequence kernel association test (SKAT). Weights are embedded within SKAT to reflect the hypothesized contribution of the variants to the trait variance. Because the true weights are generally unknown, and so are subject to misspecification, 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. We show that the use of alternative weights obviates the need to impose arbitrary frequency thresholds. As both the score test and the likelihood ratio test (LRT) may be used in this context, and may differ in power, we characterize the behavior of both tests. The two tests have equal power, if the weights in the set included weights resembling the correct ones. However, if the weights are badly specified, the LRT shows superior power (due to its robustness to misspecification). With this data-driven weighting procedure the LRT detected significant signal in genes located in regions already confirmed as associated with schizophrenia — the *PRRC2A* ($p = 1.020e\text{-}06$) and the *VARS2* ($p = 2.383e\text{-}06$) — in the Swedish schizophrenia case-control cohort of 11,040 individuals with exome-sequencing data. The score test is currently preferred for its computational efficiency and power. Indeed, assuming correct specification, in some circumstances, the score test is the most powerful test. However, LRT has the advantageous properties of being generally more robust and more powerful under weight misspecification. This is an important result given that, arguably, misspecified models are likely to be the rule rather than the exception in weighting-based approaches.

■ **Keywords:** SKAT, variable weighting, robustness, MAF thresholding, power, schizophrenia

With the increasing availability of exome/genome sequencing data, rare variant (RV) association studies are gaining importance in human genetic research. One important test of association between a target set of RVs and a given phenotype is the sequence kernel-based association test (SKAT) (Chen et al., 2013; Ionita-Laza et al., 2013; Lee

RECEIVED 15 December 2016; ACCEPTED 20 December 2016. First published online 27 February 2017.

ADDRESS FOR CORRESPONDENCE: Dr Camelia C. Minică, Department of Biological Psychology, Vrije Universiteit Amsterdam, Transitorium 2B-03, Van der Boechorststraat 1, 1081 BT, Amsterdam, the Netherlands. E-mail: camelia.minica@gmail.com
* These authors contributed equally to this work

et al., 2012; Lippert et al., 2014; Listgarten et al., 2013; Svishcheva et al., 2014; Wu et al., 2011). SKAT is based on a random effects model, in which the effect sizes of the RVs are assumed to be drawn from a distribution with a zero mean and a variance. That the effect sizes are characterized by a single variance is a strong assumption that is made plausible by weighting of effect sizes. The required weights are typically assigned based on meta-information about the tested variants, such as allele frequency and functional predictions (Kryukov et al., 2007; Madsen & Browning, 2009; Price et al., 2010; Wu 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 et al., 2009). Accordingly, the effect sizes for RVs will tend to be larger than for more common variants.

The relationship between effect size, frequency, and selection, however, rests on directional assumptions about the extent of selection on the phenotype in question and the demographic history of the population (Eyre-Walker & Keightley, 2007; Price et al., 2010; Zuk et al., 2014). Specifically, there are several conditions that have to hold for the frequency to be genuinely informative about the functional effect that a genetic variant has on a trait, namely: (a) the population under study has not experienced recent severe bottlenecks; (b) the selection on the trait of interest is direct; (c) strong (i.e., selection coefficient $s \geq 10^{-2.5}$); and (d) it acts uniformly across the associated genes. Yet, for the reasons detailed below, the circumstances in which these conditions are expected to hold are rather special. First, population genetics theory predicts that the frequency of deleterious variants will vary with the size of the effect the associated trait has on fitness. For instance, risk variants implicated in early-onset diseases (e.g., autism) will be mostly rare, that is, kept at low frequencies by selection pressures because of the high impact these diseases have on reproductive fitness (Manolio et al., 2009). In contrast, variants associated with a trait having a negligible effect on fitness (e.g., Alzheimer's disease) will likely escape selection and so may occur at relatively high frequencies in the population (Zuk et al., 2014). Second, it should be noted that even if the trait of interest is under strong selection pressure, variants across the whole frequency spectrum may jointly contribute to disease risk, as simulation studies (Price et al., 2010) and empirical results (e.g., Cohen et al., 2006; Teslovich et al., 2010) have demonstrated. Third, allele frequency distribution is expected to vary as a function of the demographic history of the population. Using population genetics simulations, Zuk et al. (2014) showed that given the same selection coefficient s, the frequency of deleterious alleles influencing a trait will depend on mutation rate and on whether the population under study has encountered recent severe bottlenecks. For example, given strong selection pressures (i.e., $s > 10^{-2.5}$) acting directly on the phenotype, the median frequency of the associated alleles may vary from as high as 0.0377 in recently bottlenecked populations (e.g., Finland), to as low as 9.36E-005 in a large population with simple exponential expansion. Finally, the strength of selection is expected to vary across genes, and so will the allele frequency–functional effect relationship (Price et al., 2010; Zuk et al., 2014). Genes under weak selection will harbor both common and RVs, both with functional effects, whereas functional variants within genes under strong selective constraints will mainly be rare. The examples above indicate that testing genomic regions by relying on a weighting scheme which up-weights rarer variants and puts low or zero weights on the more common ones is optimal only in specific circumstances.

Because the true weights are generally unknown and, therefore, subject to misspecification, 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 set of alternative weighting schemes will accommodate genomic regions where only very RVs 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. As such, this adaptive weighting procedure renders the (arbitrary) MAF thresholding unnecessary. Family-wise error rate can be protected by using a multiple testing correction method (e.g., the Bonferroni method) or by using permutation. Using simulations, we demonstrate that the use of alternative (incorrect) weights does not inflate the type I error rate. We show the power benefits conferred by the use of such a data-driven weighting procedure in both simulated and empirical data. As both the score test (Wu et al., 2011) and the likelihood ratio test (LRT) (Listgarten et al., 2013) may be used in this context, and may differ in power (Zeng et al., 2014), we characterize the behavior of both tests.

Below, we first formulate the model and briefly describe the LRT and the score test. We then present and evaluate the use of a data-driven weighting scheme in simulated and empirical data. Specifically, we evaluate the efficiency of the two tests under (a) the data-driven weighting scheme, relative to their efficiency under (b) incorrect, and (c) correct weighting. Finally, we discuss the robustness of the two tests 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 phenotypic scores obtained in a sample of n individuals. Let X be the $n \times p$ design matrix containing covariates. Let G be the $n \times m$ matrix of genotype values, with the g_{ij} element denoting

the genotype value of the individual i ($i = 1\ldots n$) at locus j ($j = 1\ldots m$). Genotypes are coded as additive-codominant, that is, $g_{ij} = (0,1,2)$. The association between the phenotype and the set of m variants is modeled within the linear mixed model framework as follows:

$$y = X\beta + Gb + e, \tag{1}$$

with $\beta^t = (\beta_1, \ldots, \beta_p)$ being the p-dimensional vector of fixed effects of covariates, $b^t = (b_1, \ldots, b_m)$ being the $m\times 1$ 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 $m\times m$ diagonal matrix containing the weights used to weigh the contribution to the test statistic of the variants in the set. The normally distributed phenotype y has expected mean $E[y] = X\beta$ and variance–covariance matrix:

$$\sum\nolimits_y = E[(y - E(y))(y - E(y))^t] = GWG^t\frac{\sigma_b^2}{m} + I\sigma_e^2, \tag{2}$$

with GWG^t being the weighted kernel or genetic relationship matrix. As implemented in the SKAT (Wu et al., 2011), the diagonal elements of the W matrix, diag($\omega_1 \ldots \omega_m$) are related to the minor allele frequency of the j^{th} 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 j^{th} variant effect and its minor allele frequency (MAF; see the 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 LRT or a score test. The LRT 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:

$$\begin{aligned}\text{LogL}\left(\sigma_b^2, \sigma_e^2\right) &= 1/2\log|\Sigma_y| - 1/2\log\left|X^t\Sigma_y^{-1}X\right| \\ &- 1/2r^t\Sigma_y^{-1}r - 1/2(n-p)\log(2\pi),\end{aligned} \tag{3}$$

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

In evaluating the statistical significance of the restricted LRT, we note the null distribution of the test statistic is a $\pi\chi_0^2 : (1-\pi)a\chi_d^2$ mixture of distributions, with the mixture parameter π, the scale parameter a, and the degrees of freedom d on the second component estimated using the computationally efficient permutation-based approach developed by Listgarten et al. (2013).

The score test is computed as follows:

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

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 in Hardy–Weinberg equilibrium 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. We set Σ_{LD} to equal an identity matrix, as we considered sets of RVs expected to be in linkage equilibrium (see e.g., Daye et al., 2012); but see the Supplementary material for results based on rare, and rare and common variants in linkage disequilibrium simulated using a coalescent model (Shlyakhter et al., 2014). The multivariate normally distributed variables were then discretized given chosen thresholds based on the MAF at each locus. We considered, MAFs varying randomly between 0.005 and 0.05, 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 $n\times 1$ vector of phenotypes, y, was generated as follows:

$$y_i = \sum_{j=1}^{m} g_{ij}b_j * \sqrt{\sigma_b^2} + e_i * \sqrt{\sigma_e^2}, \tag{5}$$

b_j, the regression weight of the variant at the j^{th} locus was computed as a function of MAF_j and of its contribution to the standardized variance of the polygenic scores (Mather & 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, 0.5, 0.5) weights (Wu et al., 2011), with weights increasing with decreasing MAF. The variance σ_b^2 equaled 0.01 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 $N(0, 1)$.

Data-Driven Search for Optimal Weights: Exploring the Misspecification Space

Because the strength and effectiveness of selection pressures vary across the genome, committing to a single weighting scheme when testing thousands of genes may only capture signal from genes under selection pressures matching the chosen weighting scheme. An optimal weighting scheme should be allowed to vary across the tested genes, to match

variable selection pressures. To this end, we evaluated the efficiency of a data-driven search for optimal weights. We carried out simulations to evaluate the efficiency of the LRT and the score test under (a) the variable data-driven weighting scheme, relative to their efficiency under (b) incorrect, and (c) correct weighting.

The m-dimensional vector of weights w was computed using the beta density function, with the j^{th} element calculated as $w_j = \text{dbeta}(\text{MAF}_j; a_1, a_2)$ given the MAF of the j^{th} variant and the shape parameters a_1 and a_2. As described in the previous section, data were simulated according to dbeta (0.5, 0.5) weights (i.e., the true weights increase with decreasing MAF). Next, in computing the tests statistic we (mis)specified the weights as: (a) dbeta (1,1); (b) dbeta (0.5, 0.5); (c) dbeta (1,25), and (d) dbeta(1,50). 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 et al., 2011) and in FaST-LMM-set (Listgarten et al., 2013). Standardization and assignment of weights dbeta (0.5, 0.5) are equivalent weighting schemes (Wu 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 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 these are the default weights in SKAT (Wu et al., 2011). Finally, we considered the effect of a more extreme weighting scheme, dbeta (1,50), including weights that overlook common variants and favor the contribution to the test statistic of rarer ones. This weighting scheme pertains to the hypothesis that only ultra-RVs contribute to the phenotypic variance.

We performed association tests by using the set of three incorrect weighting schemes, that is, (a) dbeta (1,1), (b) dbeta (1,25), and (c) dbeta (1,50). The p value for the gene equaled the minimum Bonferroni corrected p value minP_{LRT} ($\text{minP}_{\text{score}}$) out of the three p values obtained, given the genotypes transformed according to each of the weighting schemes enumerated above. We also report the power of the tests under each of these misspecified weighting schemes, as it is of interest to assess whether our procedure confers power gains relative to a test that uses a single (misspecified) weighting scheme (i.e., three tests vs. one test). We assessed the behavior of the two tests under the above weighting schemes by considering target regions harboring both deleterious and beneficial variants.

Evaluating the Type I Error Rates and Power

We evaluated the type I error rate by generating 1,000 datasets under the null hypothesis of no phenotypic variance explained by the variants 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 $\alpha = 0.01$. We refer to Listgarten et al. (2013) for an exhaustive evaluation of the type I error at more stringent alpha levels.

Power was assessed based on 1,000 simulated datasets, an effect size of 1% explained phenotypic variance and 7 alpha thresholds. Given the 7 alpha thresholds, power was computed using the permutation-based procedure implemented in FaST-LMM-Set (Listgarten et al., 2013). Estimation of the free parameters π, a, and d of the null distribution $\pi\chi_0^2 : (1-\pi)a\chi_d^2$ used 1,000 permutations. As a validity check of our simulation program, we also report the power and the type I error rates of the true (i.e., correct) model.

Software

The R-package MASS (Venables & Ripley, 2002) was used for data generation. Model fitting was performed in FaST-LMM-set (Listgarten et al., 2013). The software is readily available for use on Github. For the sake of comparison, we analyzed one simulated sample of 5,000 individuals by using four independent programs implementing genetic similarity/kernel-based variance component tests: the nlme R-package, the software Genome-wide Complex Trait Analysis (GCTA; Yang et al., 2011), the software FaST-LMM-set (Listgarten et al., 2013), and the R-package OpenMx (Neale et al., 2016). The values for the LRT and the estimates for the variance component obtained by the four programs were almost identical (see Table S1 for details), indicating that these implement equivalent approaches. Having established the equivalence, the empirical analyses were conducted using the software FaST-LMM-set. Analyses were carried out on the Broad Institute Gold Compute cluster and on the Lisa cluster (https://www.surf.nl/en).

Empirical Analysis: Evaluating the Importance of Thresholding and Variable Weighting

We compared the performance of the LRT and of the score test under our proposed data-driven weighting scheme in a real dataset. For this illustration, we used the Swedish schizophrenia case-control cohort of 11,040 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 (Dalman et al., 2002; Kristjansson et al., 1987). 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 Purcell et al. (2014),

and Ripke et al. (2013), for a detailed description of the sample). There were 175 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 4,867 cases and 6,173 controls; 6052 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. Data are available through dbGAP.

Exome-sequencing was performed in 12 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. For samples in the waves 2– 12, 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 & 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 et al., 2010). Selected singletons were validated using Sanger sequencing (see Purcell 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 et al., 2011; Van der Auwera et al., 2013). There were 1,584,195 variants meeting all our quality control criteria.

For this empirical illustration, we focused on two partially overlapping sets of genes (1,435 genes) likely relevant to schizophrenia. The first set consisted of 941 genes that are part of the list identified by Samocha 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 favoring the hypothesis that schizophrenia and ASD share genetic etiology (Fromer et al., 2014; Schizophrenia Working Group of the Psychiatric Genomics Consortium, 2014), this set of genes is likely to be relevant also to schizophrenia. The second set consisted of 768 genes targeted by the fragile-X mental retardation protein (FMRP). This set is part of the list of genes derived by Darnell 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 non-synonymous case mutations in both ASD (Iossifov et al., 2012) and schizophrenia (Fromer et al., 2014). Purcell et al. (2014) also tested the FMRP set for enrichment of RVs in half of the current sample, and their analysis yielded nominally significant results.

We performed sequence-based kernel association analyses using the LRT and score tests with variable weights. The analyses were carried out using the FaST-LMM-Set software (Listgarten et al., 2013). To adjust for ancestry, we included into analysis the first two principal components explaining the largest amount of variance in the sample and reflecting the Finish and Northern/Southern Swedish ancestry (see Extended Data Figure 1 in Purcell et al. (2014); see also Genovese et al. (2016)). Principal components were computed from genotypes at variants shared with the 1,000 Genomes Project phase 1 dataset. To accommodate the scenario in which only RVs are likely to be functional, as well as the scenario in which the targeted region is 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 (0.5, 0.5), and dbeta (1,1). The use of alternative weights obviates the need for choosing arbitrary frequency thresholds to select the target set. However, for the sake of illustration, we also report the results obtained in the analyses stratified based on allele counts thresholds (i.e., we selected variants with a minor allele count (MAC) up to 10 and a MAC up to 50). For each of the tested genes, we selected the Bonferroni corrected p value corresponding to the weighting scheme that yields the largest test statistic (i.e., the p value was adjusted for multiple hypothesis testing of 1,435 genes and three weighting schemes). An alpha of 0.05 was used as the significance threshold. For computational ease, we used a linear model (Listgarten 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 et al., 2014).

Results

Type I Error

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

Both the restricted LRT and the score test yield correct type I error rates, regardless of whether the weights used are correctly specified or misspecified. The two tests show good control of the type I error rate also under our proposed Bonferroni data-driven weighting procedure. Note that these conclusions generalize to scenarios in which the target set includes common and RVs in linkage equilibrium/disequilibrium (see Table S2).

Power

Figure 1 displays the results relating to power to detect a target set of 50 functional variants (but see Figure S1 for

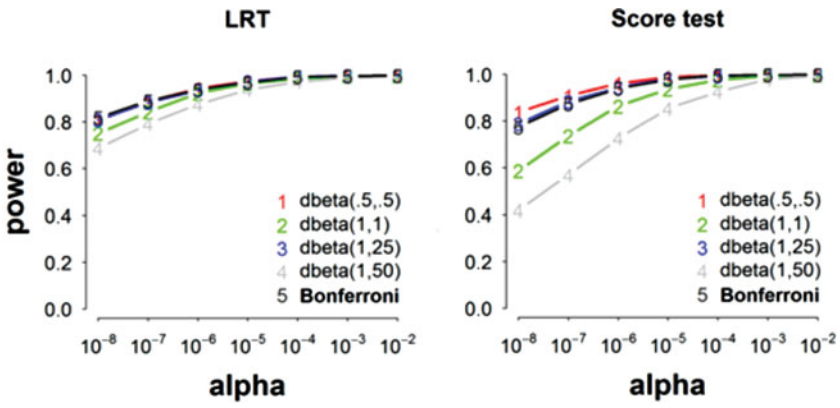

FIGURE 1

(Colour online) The power of the likelihood ratio test (LRT) and the score test to detect a gene harboring 50 functional variants, jointly explaining 1% of the phenotypic variance (minor allele frequency 0.5–5%). Data were simulated according to weights dbeta (0.5, 0.5). Power was evaluated in 1,000 datasets consisting of 10,000 individuals each.

simulation results involving a region harboring a mixture of functional and phenotypically neutral variants).

Four important conclusions follow from our simulation results. First, the restricted LRT and the score test have equal power when the weights are correctly specified. This is expected, as the two tests are asymptotically equivalent when the model is true, that is, correctly specified (e.g., Greene, 2003). The powers of the two tests are similar when the assigned weights correspond to the true weights, that is, dbeta (0.5, 0.5).

Second, misspecification of weights always reduces power. This is shown in Figure 1 as the departure of the power under model misspecification from the power of the true model, that is, dbeta (0.5, 0.5). 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; for example, when the assigned weights dbeta (1,25) resemble the true weights dbeta (0.5, 0.5), as illustrated in Figure 1. However, the power may suffer dramatically with increasing misspecification. For instance, using a dbeta (1,50) weighting scheme — which acts as a frequency threshold, removing from the test the more common variants — results in a loss in power of up to ∼10% and ∼34% (given an alpha of 10^{-7}), for the restricted LRT and for the score test, respectively.

Third, relative to the score test, we note that the restricted LRT is consistently more robust to weight misspecification. These results are consistent with those reported by Zeng et al. (2014) and by Lippert et al. (2014), who found the 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.

TABLE 1

The Empirical 95% Confidence Intervals around the Type I Error for the Restricted Likelihood Ratio Test (LRT) and the Score Test, given data simulated under the null model of no association between the target region and the phenotype

Weights dbeta	LRT	Score test
(0.5, 0.5)	[0.0043, 0.0176]	[0.0030, 0.0150]
(1,1)	[0.0054, 0.0183]	[0.0037, 0.0163]
(1,25)	[0.0043, 0.0176]	[0.0030, 0.0150]
(1,50)	[0.0054, 0.0183]	[0.0037, 0.0163]
Bonferroni	[0.0018, 0.0123]	[0.0024, 0.0137]

Note: Type I error was evaluated at alpha = 0.01. The tests were computed for five weighting schemes in each of the 1,000 simulated samples of 10,000 individuals with genotypes at 50 variants in linkage equilibrium (minor allele frequency 0.5–5%).

Fourth, we note that both tests benefit from the use of variable weights. The data-driven search for optimal weights confers power advantages over a model that uses misspecified weights, and maintains the power close to that afforded by a correctly specified model. It should be noted, however, that there is a price to pay in terms of power by using this data-driven weighting scheme in contrast to correct weighting (i.e., using alternative weights increases the burden of multiple testing). The two tests have equal powers with the Bonferroni corrected data-driven weighting procedure; this is due to the fact that the weights resembling the correct ones were included in the procedure (the more weights one tries, the largest the price in terms of power one has to pay). Had the procedure included weights misspecified to a greater extent, the power of the score test would have decreased relative to that of the LRT (which appears to be more robust to misspecification). As the true weights are

TABLE 2

Results of the gene based analysis run in the Swedish sample ($N = 11,040$)

Chromosome (position range)	Gene (autosome variants)	Weights dbeta	LRT	Score test
9 (135762714–135804294)	*TSC1* (142)	(1,1)	**0.0001** *(0.5596)*	0.0027 *(1)*
		(0.5,0.5)	0.0064 *(1)*	0.0256 *(1)*
		(1,25)	0.0014 *(1)*	0.0026 *(1)*
15 (52058632–52100672)	*TMOD2* (52)	(1,1)	**0.0004** *(1)*	0.0039 *(1)*
		(0.5,0.5)	0.0069 *(1)*	0.015 *(1)*
		(1,25)	0.0034 *(1)*	0.0038 *(1)*
4 (62363001–62935992)	*LPHN3* (131)	(1,1)	**0.0006** *(1)*	0.0029 *(1)*
		(0.5, 0.5)	0.0146 *(1)*	0.0563 *(1)*
		(1,25)	0.0041 *(1)*	0.0029 *(1)*

Note: The gene-based analysis was restricted to variants with minor allele count below 10. Bonferroni corrected *p* values are given in italics. For each gene, the lowest *p* value is given in bold type.

typically unknown, conjecturing the correct ones by employing the proposed Bonferroni scheme with alternative weights and using the LRT appears to be the strategy most likely to maintain the power close to that of the true model.

Empirical Analysis: Evaluating the Importance of Thresholding and Variable Weighting

We next looked at the behavior of the score test and of the LRT (Listgarten et al., 2013) under variable weights in the empirical dataset. Tables 2 and 3 display results pertaining to the association tests in the analyses stratified based on arbitrary MAC thresholds.

From Table 2, we note that the LRT appears to be more powerful than the score test. The two tests seem to agree in selecting the top association signals, as both ranked in the top three the same genes. All three weighting schemes tend to pick up nominally significant association signals. Of these, the dbeta (1,1) weighting scheme yields the lowest *p* value for all three genes. Similar trends in the results were observed when we restricted the analyses to variants with a MAC below 50 (see Table 3).

The use of alternative weights obviates the need of thresholding to prioritize the contribution of the variants to the test statistic (the thresholds are, however, arbitrary: variants defined as rare in one sample might feature as common in another sample). We conducted the analysis using our proposed data-driven weighting scheme, without imposing any frequency threshold. Table 4 contains the results.

For the top three genes, Table 4 shows that the dbeta (1,1) weighting scheme appears to best capture the allele frequency–functional effect relationship. This weighting scheme yields the largest test statistic and singles out the *PRRC2A* and the *VARS2* as significantly associated with schizophrenia disease status given our chosen alpha threshold (i.e., p value = 1.020e-06 and p value = 2.383e-06, respectively). The third top gene is the *AKT3* gene (p value = 2.825e-05). All three genes belong to the Samocha et al. (2014) list of genes under selection constraints. Had one relied on a weighting scheme that up-weights rarer variants and down-weights the more common ones, these association signals would have been missed. As these genes did not pass the significance threshold in the analyses stratified by MAC, the results suggest that arbitrary thresholding might remove from the target causal variants and in doing so might weaken the association signal. We observed similar trends in power when we simulated sets of common and rare functional variants, where — similar to a frequency threshold — the dbeta (1,25) weighting scheme discarded from the target set causal variants (see Figure S2).

Importantly, association signals in all three genes have been previously reported (e.g., Ripke et al., 2013) and replicated (e.g., Aberg et al., 2013), suggesting that these results are unlikely to be false positives. Without thresholding, common variants might also be included in the analysis. In our sample, of the 43 (*AKT3*), 238 (*PRRC2A*), and 408 (*VARS2*) tested variants, 1, 29, and 15 variants, respectively, had a MAC greater than 50. The question remains whether the test was dominated by these common variants. We checked in our sample whether the common variants, if tested with a univariate test, do yield genome-wide significant association signals. Results showed that none of them would be detected in an ordinary genome-wide association study (GWAS) (see Tables S3–S5). Hence, either thresholding or relying on a default weighting scheme would result in missing true association signals. We elaborate on these results in the Discussion.

Discussion

We considered the issue of optimizing weighting in association studies based on the sequence kernel test. Consistent with empirical (Lippert et al., 2014) and simulation (Zeng et al., 2014) results, we found that the LRT is generally more robust to weight misspecification, and more powerful than the score test in such a circumstance. The principal finding of this study is that using a weighting scheme that includes alternative weights is likely to boost statistical power. 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.

In the literature, weighting is mostly informed by allele frequency; frequency is taken as indicative of the strength of the purifying selection coefficient (Kryukov et al., 2007).

TABLE 3

Results of the gene-based analysis run in the Swedish sample ($N = 11,040$)

Chromosome (position range)	Gene (autosome variants)	Weights dbeta	LRT	Score test
9 (109685651–109773313)	*ZNF462* (224)	(1,1)	**0.0001** *(0.4735)*	0.0032 *(1)*
		(0.5,0.5)	0.0078 *(1)*	(0.0547) *(1)*
		(1,25)	0.0001 *(0.4735)*	0.003 *(1)*
15 (52058615–52100672)	*TMOD2* (54)	(1,1)	0.0002 *(1)*	0.0091 *(1)*
		(0.5,0.5)	0.0054 *(1)*	0.0063 *(1)*
		(1,25)	**0.0002** *(1)*	0.0083 *(1)*
8 (141669548–141900779)	*PTK2* (139)	(1,1)	0.0008 *(1)*	0.0034 *(1)*
		(0.5,0.5)	0.0136 *(1)*	0.0429 *(1)*
		(1,25)	**0.0008** *(1)*	0.0033 *(1)*

Note: The gene-based analysis was restricted to variants with minor allele count below 50. Bonferroni corrected *p* values are given in italics. For each gene, the lowest *p* value is given in bold type.

TABLE 4

Results of the gene-based analysis run in the Swedish schizophrenia case-control sample ($N = 11,040$)

Chromosome (position range)	Gene (autosome variants)	Weights dbeta	LRT	Score test
6 (31584304–31607461)	*PRRC2A* (408)	(1,1)	**1.020e-06** *(0.0043)*	2.556e-06 *(0.011)*
		(0.5, 0.5)	5.8e-04 *(1)*	9.886e-05 *(0.4255)*
		(1,25)	0.055 *(1)*	0.057 *(1)*
6 (30877202–30894026)	*VARS2* (238)	(1,1)	**2.383e-06** *(0.0102)*	0.0043 *(1)*
		(0.5, 0.5)	0.0031 *(1)*	0.0048 *(1)*
		(1,25)	1 *(1)*	0.534 *(1)*
1 (243668558–244006487)	*AKT3* (43)	(1,1)	**2.825e-05** *(0.1216)*	7e-04 *(0.7533)*
		(0.5, 0.5)	0.0036 *(1)*	0.0063 *(1)*
		(1,25)	1.6e-04 *(0.6888)*	7.586e-05 *(0.3265)*

Note: The gene-based analysis was conducted by relying on the data-driven weighting procedure, without imposing a-priori a frequency threshold. Bonferroni corrected *p* values are given in italics. For each gene, the lowest *p* value is given in bold type.

Accordingly, rarer variants are typically being assigned larger weights/contribution to the test statistic (e.g., Wu 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 & Keightley, 2007; Price et al., 2010; Zuk et al., 2014). Genes under weak selection may harbor rare as well as more common variants with disruptive effects (Zuk 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 et al., 2010).

Achieving maximal power when testing such regions requires adapting the weighting scheme to match the hypothesized selection types. 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 locate previously reported genes conferring risk to schizophrenia (e.g., Aberg et al., 2013; Ripke et al., 2013), 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 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 et al., 2011). Alternative weighting schemes that incorporate probabilities of a variant being damaging, as estimated by annotation tools such as, for example, Polyphen-2 (Adzhubei et al., 2010) or SIFT (Ng & Henikoff, 2003), may also be considered.

We emphasize that our data-driven weighting approach renders thresholding unnecessary. Thresholding (either based on counts or on allele frequency) has been initially used in burden tests (e.g., Li & Leal, 2008; Madsen & Browning, 2009; Price et al., 2010; see also Franić et al., 2015, for an overview on burden tests), but it has been employed also in sequence-based variance component tests (Lohmueller et al., 2013; Xu et al., 2014) for the purpose of removing neutral variation (see, e.g., Kryukov et al., 2007). Yet, in our empirical analysis this practice was counterproductive: imposing the (arbitrarily chosen) MAC thresholds muted the signal in genes located in regions already confirmed as associated with schizophrenia (i.e., the *PRRC2A* and the *VARS2* genes; e.g., Aberg et al., 2013; and the *AKT3* gene; e.g., Ripke et al., 2013). 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. That is, we note that none of these common variants would be detected in an ordinary GWAS (see Tables S3–S5). 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 harbor rare as well as common variants, both with functional effects. To locate genes under stronger selection pressures, larger samples (see Zuk et al., 2014) and the inclusion of more extreme weights (i.e., weights that overlook common variants and favor rare ones) will probably be required.

The LRT and the score test had equal power under the data-driven weighting approach. Note, however, that this equivalence hinged upon the inclusion of weights that closely resemble the true ones among the alternatives. The powers of the two tests will likely diverge when the weights in the set are all badly specified; in such a circumstance, the LRT is expected to show superior power (due to its robustness to assumption violation). This is likely illustrated in the empirical analysis where the LRT has always yielded lower *p* values. Yet, despite these differences in power, currently the score test is the dominant association test with RVs, involving single studies and also in meta-analyses (see, e.g., Tang & Lin, 2015). Integrating LRT into meta-analytic techniques for rare-variant association testing is desirable — to ensure maximal power of detection — and will likely boost its application.

Both in the simulations and in the empirical analysis, we chose to correct alpha by using the Bonferroni method. We chose this method for the sake of simplicity. Although one may argue that the method is slightly conservative as the tests are correlated, it is important to note that the Bonferroni corrected weighing procedure confers more power than a badly specified weighting scheme would do; *p* value correction for larger number of tests can be easily obtained using the *p.adjust* function implemented in the *stats* R-package. Permutation may also be used to compute the *p* value. However, 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 genes) and the sample are 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 algorithm for optimal search for the 'true' weights (e.g., Neale & Cardon, 1992) or limiting the choice of weights based on knowledge on theorized selection on each gene (Zuk et al., 2014) would decrease the burden of multiple testing, and further increase power.

Conclusion

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

Acknowledgments

We thank the Swedish cohort participants whose data we analyzed in this study. Camelia C. Minică and Michael C. Neale are supported by the National Institute on Drug Abuse Grant DA-018673. Jacqueline M. Vink is supported by the ERC starting grant 284167.

Supplementary Material

To view supplementary material for this article, please visit https://doi.org/10.1017/thg.2017.7.

References

Aberg, K. A., Liu, Y., Bukszár, J., McClay, J. L., Khachane, A. N., Andreassen, O. A., … Gurling, H. (2013). A comprehensive family-based replication study of schizophrenia genes. *JAMA Psychiatry, 70*, 573–581.

Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., … Sunyaev, S. R. (2010). A method and server for predicting damaging missense mutations. *Nature Methods, 7*, 248–249.

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

Chen, H., Meigs, J. B., & Dupuis, J. (2013). Sequence kernel association test for quantitative traits in family samples. *Genetic Epidemiology, 37*, 196–204.

Cohen, J. C., Boerwinkle, E., Mosley Jr, T. H., & Hobbs, H. H. (2006). Sequence variations in PCSK9, low LDL, and protection against coronary heart disease. *New England Journal of Medicine, 354*, 1264–1272.

Cruchaga, C., Karch, C. M., Jin, S. C., Benitez, B. A., Cai, Y., Guerreiro, R., … Bertelsen, S. (2014). Rare coding variants in the phospholipase D3 gene confer risk for Alzheimer/'s disease. *Nature, 505*, 550–554.

Dalman, C., Broms, J., Cullberg, J., & Allebeck, P. (2002). Young cases of schizophrenia identified in a national inpatient register. *Social Psychiatry and Psychiatric Epidemiology, 37*, 527–531.

Darnell, J. C., Van Driesche, S. J., Zhang, C., Hung, K. Y. S., Mele, A., Fraser, C. E., … Chi, S. W. (2011). FMRP stalls ribosomal translocation on mRNAs linked to synaptic function and autism. *Cell, 146*, 247–261.

Davies, R. (1980). The distribution of a linear combination of chi-square random variables. *Journal of the Royal Statistical Society, Series C (Applied Statistics), 29*, 323–333.

Daye, Z. J., Li, H., & Wei, Z. (2012). A powerful test for multiple rare variants association studies that incorporates sequencing qualities. *Nucleic Acids Research, 40*, e60–e60.

DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., … Hanna, M. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics, 43*, 491–498.

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

Franić, S., Dolan, C. V., Broxholme, J., Hu, H., Zemojtel, T., Davies, G. E., … Hottenga, J.-J. (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., Pocklington, A. J., Kavanagh, D. H., Williams, H. J., Dwyer, S., Gormley, P., … Ruderfer, D. M. (2014). De novo mutations in schizophrenia implicate synaptic networks. *Nature, 506*, 179–184.

Genovese, G., Fromer, M., Stahl, E. A., Ruderfer, D. M., Chambert, K., Landén, M., … Sullivan, P. F. (2016). Increased burden of ultra-rare protein-altering variants among 4,877 individuals with schizophrenia. *Nature Neuroscience*, *19*, 1433–1441.

Greene, W. H. (2003). *Econometric analysis*. Upper Saddle River, NJ: Prentice Hall.

Huyghe, J. R., Jackson, A. U., Fogarty, M. P., Buchkovich, M. L., Stančáková, A., Stringham, H. M., … Cederberg, H. (2013). Exome array analysis identifies new loci and low-frequency variants influencing insulin processing and secretion. *Nature Genetics, 45*, 197–201.

Ionita-Laza, I., Lee, S., Makarov, V., Buxbaum, J. D., & Lin, X. (2013). Sequence kernel association tests for the combined effect of rare and common variants. *The American Journal of Human Genetics, 92*, 841–853.

Iossifov, I., Ronemus, M., Levy, D., Wang, Z., Hakker, I., Rosenbaum, J., … Leotta, A. (2012). De novo gene disruptions in children on the autistic spectrum. *Neuron, 74*, 285–299.

Kristjansson, E., Allebeck, P., & Wistedt, B. (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*, 229–234.

Kryukov, G. V., Pennacchio, L. A., & Sunyaev, S. R. (2007). Most rare missense alleles are deleterious in humans: Implications for complex disease and association studies. *The American Journal of Human Genetics, 80*, 727–739.

Lee, S., Wu, M. C., & Lin, X. (2012). Optimal tests for rare variant effects in sequencing association studies. *Biostatistics, 13*, 762–775.

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

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with burrows–wheeler transform. *Bioinformatics, 25*, 1754–1760.

Lippert, C., Xiang, J., Horta, D., Widmer, C., Kadie, C., Heckerman, D., & Listgarten, J. (2014). Greater power and computational efficiency for kernel-based association testing of sets of genetic variants. *Bioinformatics, 30*, 3206–3214.

Listgarten, J., Lippert, C., Kang, E. Y., Xiang, J., Kadie, C. M., & Heckerman, D. (2013). A powerful and efficient set test for genetic markers that handles confounders. *Bioinformatics, 29*, 1526–1533.

Lohmueller, K. E., Sparsø, T., Li, Q., Andersson, E., Korneliussen, T., Albrechtsen, A., … Kiil, K. (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*, 1072–1086.

Madsen, B. E., & Browning, S. R. (2009). A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genetics, 5*, e1000384.

Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., … Visscher, P. M. (2009). Finding the missing heritability of complex diseases. *Nature, 461*, 747–753.

Mather, K., & Jinks, J. L. (1977). *Introduction to biometrical genetics*. Ithaca, NY: Cornell University Press.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., … Daly, M. (2010). The genome analysis toolkit: A mapreduce framework for analyzing next-generation DNA sequencing data. *Genome Research, 20*, 1297–1303.

Neale, M. C., Hunter, M. D., Pritikin, J. N., Zahery, M., Brick, T. R., Kirkpatrick, R. M., … Boker, S. M. (2016). OpenMx 2.0: Extended structural equation and statistical modeling. *Psychometrika*, *81*, 535–549.

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

Ng, P. C., & Henikoff, S. (2003). SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Research, 31*, 3812–3814.

Peloso, G. M., Auer, P. L., Bis, J. C., Voorman, A., Morrison, A. C., Stitziel, N. O., … Fornage, M. (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*, 223–232.

Price, A. L., Kryukov, G. V., de Bakker, P. I., Purcell, S. M., Staples, J., Wei, L.-J., & Sunyaev, S. R. (2010). Pooled association tests for rare variants in exon-resequencing studies. *The American Journal of Human Genetics, 86*, 832–838.

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

Purcell, S. M., Moran, J. L., Fromer, M., Ruderfer, D., Solovieff, N., Roussos, P., … Kähler, A. (2014). A polygenic burden of rare disruptive mutations in schizophrenia. *Nature, 506*, 185–190.

Ripke, S., O'Dushlaine, C., Chambert, K., Moran, J. L., Kähler, A. K., Akterin, S., … Fromer, M. (2013). Genome-wide association analysis identifies 13 new risk loci for schizophrenia. *Nature Genetics, 45*, 1150–1159.

Samocha, K. E., Robinson, E. B., Sanders, S. J., Stevens, C., Sabo, A., McGrath, L. M., … Kirby, A. (2014). A framework for the interpretation of de novo mutation in human disease. *Nature Genetics, 46*, 944–950.

Schizophrenia Working Group of the Psychiatric Genomics Consortium. (2014). Biological insights from 108 schizophrenia-associated genetic loci. *Nature, 511*, 421–427.

Schork, N. J., Murray, S. S., Frazer, K. A., & Topol, E. J. (2009). Common vs. rare allele hypotheses for complex diseases. *Current Opinion in Genetics & Development, 19*, 212–219.

Shlyakhter, I., Sabeti, P. C., & Schaffner, S. F. (2014). Cosi2: An efficient simulator of exact and approximate coalescent with selection. *Bioinformatics, 30*, 3427–3429.

Speed, D., Hemani, G., Johnson, M. R., & Balding, D. J. (2012). Improved heritability estimation from genome-wide SNPs. *The American Journal of Human Genetics, 91*, 1011–1021.

Svishcheva, G. R., Belonogova, N. M., & Axenovich, T. I. (2014). FFBSKAT: Fast family-based sequence kernel association test. *PloS One, 9*, e99407.

Tang, Z.-Z., & Lin, D.-Y. (2015). Meta-analysis for discovering rare-variant associations: Statistical methods and software programs. *The American Journal of Human Genetics, 97*, 35–53.

Teslovich, T. M., Musunuru, K., Smith, A. V., Edmondson, A. C., Stylianou, I. M., Koseki, M., … Willer, C. J. (2010). Biological, clinical and population relevance of 95 loci for blood lipids. *Nature, 466*, 707–713.

Van der Auwera, G. A., Carneiro, M. O., Hartl, C., Poplin, R., del Angel, G., Levy-Moonshine, A., … Thibault, J. (2013). From fastq data to high confidence variant calls: The genome analysis toolkit best practices pipeline. *Current Protocols in Bioinformatics, 43*, 11.10.1–11.10.33.

Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with S*. New York, NY: Springer Science & Business Media.

Wu, M. C., Lee, S., Cai, T., Li, Y., Boehnke, M., & Lin, X. (2011). Rare-variant association testing for sequencing data with the sequence kernel association test. *The American Journal of Human Genetics, 89*, 82–93.

Xu, C., Tachmazidou, I., Walter, K., Ciampi, A., Zeggini, E., & Greenwood, C. M. (2014). Estimating genome-wide significance for whole-genome sequencing studies. *Genetic Epidemiology, 38*, 281–290.

Yang, J., Lee, S. H., Goddard, M. E., & Visscher, P. M. (2011). GCTA: A tool for genome-wide complex trait analysis. *The American Journal of Human Genetics, 88*, 76–82.

Zeng, P., Zhao, Y., Liu, J., Liu, L., Zhang, L., Wang, T., … Chen, F. (2014). Likelihood ratio tests in rare variant detection for continuous phenotypes. *Annals of Human Genetics, 78*, 320–332.

Zhan, X., Larson, D. E., Wang, C., Koboldt, D. C., Sergeev, Y. V., Fulton, R. S., … Bragg-Gresham, J. (2013). Identification of a rare coding variant in complement 3 associated with age-related macular degeneration. *Nature Genetics, 45*, 1375–1379.

Zuk, O., Schaffner, S. F., Samocha, K., Do, R., Hechter, E., Kathiresan, S., … Lander, , E., S. (2014). Searching for missing heritability: Designing rare variant association studies. *Proceedings of the National Academy of Sciences, 111*, e455–e464.