Published in final edited form as:

Genet Epidemiol. 2019 June ; 43(4): 365–372. doi:10.1002/gepi.22183.

A Simple and Accurate Method to Determine Genomewide Significance for Association Tests in Sequencing Studies

Dan-Yu Lin

Department of Biostatistics, University of North Carolina, Chapel Hill, NC 27599, USA

Abstract

Whole-exome sequencing (WES) and whole-genome sequencing (WGS) studies are underway to investigate the impact of genetic variants on complex diseases and traits. It is customary to perform single-variant association tests for common variants and region-based association tests for rare variants. The latter may target variants with similar or opposite effects, interrogate variants with different frequencies or different functional annotations, and examine a variety of regions. The large number of tests that are performed necessitates adjustment for multiple testing. The conventional Bonferroni correction is overly conservative as the test statistics are correlated. To address this challenge, we propose a simple and accurate method based on parametric bootstrap to assess genomewide significance. We show that the correlations of the test statistics are determined primarily by the genotypes, such that the same significance threshold can be used in different studies that share a common sequencing platform. We demonstrate the usefulness of the proposed method with WES data from the National Heart, Lung, and Blood Institute (NHLBI) Exome Sequencing Project (ESP) (Auer et al., 2016) and WGS data from the 1000 Genomes Project (The 1000 Genomes Project Consortium, 2015). We recommend the p -value of 5×10^{-9} as the genomewide significance threshold for testing all common and low-frequency variants (MAFs 0.1%) in the human genome.

Keywords

Gene-based association tests; Parametric bootstrap; Sliding windows; SNPs; Whole-exome sequencing studies; Whole-genome sequencing studies

1. INTRODUCTION

Massively parallel sequencing has ushered in a new era in human genetics. A growing number of WES and WGS projects have been launched to understand the genetic architecture of complex diseases and traits. For example, the NHLBI ESP performed WES on ~ 7,000 individuals to identify protein-coding variants associated with heart, lung, and blood disorders. As part of the Precision Medicine Initiative, the Trans-Omics for Precision Medicine (TOPMed) Program currently consists of 40 WGS studies on 120,000 individuals, and the National Human Genome Research Institute's Genome Sequencing Program aims to perform WGS on 99,000 individuals and WES on 127,000 individuals by the end of 2019.

Single-variant association tests are usually performed on common variants in both WES and WGS studies. Gene-based association tests for rare variants are typically performed in WES

studies. The most common gene-based association test is the burden test, which creates a burden score for each study subject by aggregating the variants with minor allele frequencies (MAFs) below a certain cutoff and relates that burden score to the phenotype of interest (Morgenthaler and Thilly, 2007; Madsen and Browning, 2009; Morris and Zeggini, 2010; Lin and Tang, 2011). A second approach is the variance-component test, such as the sequence kernel association test (SKAT), which aims to detect variants with opposite effects within a gene (Tzeng and Zhang, 2007; Neale et al., 2011; Wu et al., 2011). With both approaches, the definition of a rare variant is somewhat arbitrary; thus, one may try different MAF cutoffs (Price et al., 2010; Lin and Tang, 2011). In addition, one may consider different classes of variants (e.g., nonsynonymous, loss of function). The large number of tests that are performed necessitates adjustment for multiple testing. The commonly used Bonferroni correction (i.e., the nominal significance level divided by the total number of tests) is overly conservative because the test statistics are correlated not only between genes (due to LD), but also within genes (due to multiple use of the same variants within a gene).

Gene-based association tests can also be applied to WGS studies. Regulatory annotations may be used to connect a non-coding variant to a gene. A more agnostic approach is to aggregate rare variants within sliding windows (Morrison et al., 2017). The windows may be overlapping or non-overlapping and may contain the same number or a different number of variants; the windows may also be nested. Thus, the correlation structures of the test statistics in WGS studies tend to be more complicated than in the case of WES studies, thereby making proper adjustment for multiple testing even more challenging, especially if there are overlapping subsets of variants among windows.

In this article, we propose a simple and accurate method to assess genomewide significance for association tests in WES and WGS studies. Our method is completely general in that it applies to any sequencing platform and any type of association test, provided that the p -value for each test statistic can be calculated. We account for the correlations of the test statistics through parametric bootstrap (Davison and Hinkley, 1997) such that the resulting genomewide significance threshold provides accurate control of the family-wise error rate. We show that the correlations of the test statistics are determined primarily by the genotypes, regardless of the phenotype and covariates. Thus, the same threshold can be used in different studies that interrogate the same or similar sets of variants, which facilitates the calculation of power in the design stage and the comparison of results across studies. We apply the proposed method to WES data from the NHLBI ESP (Auer et al., 2016) and WGS data from the 1000 Genomes Project (The 1000 Genomes Project Consortium, 2015). We provide practical guidelines for declaring genomewide significance in WGS studies.

2. METHODS

Suppose that we are interested in m variants within a genomic region, with genotypes $G = (g_1, \dots, g_m)^T$, where g_j is the number of minor alleles at the j th variant site. Let Y denote the phenotype of interest, which may be continuous or discrete, and let X denote a set of covariates (e.g., principal components for ancestry, demographic variables, environmental factors) plus the unit component. We relate Y to G and X through the generalized linear model with the conditional density function

$$\exp \left\{ \frac{y(\boldsymbol{\beta}^T \mathbf{G} + \boldsymbol{\gamma}^T \mathbf{X}) - b(\boldsymbol{\beta}^T \mathbf{G} + \boldsymbol{\gamma}^T \mathbf{X})}{a(\phi)} + c(y, \phi) \right\}, \quad (1)$$

where $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are regression parameters, ϕ is a dispersion parameter, and a , b , and c are specific functions. We use the first and second derivatives of b , which are denoted by b' and b'' . For the linear regression model, $a(\phi) = \sigma^2$, $b(z) = (1/2)z^2$, $b'(z) = z$, and $b''(z) = 1$. For the logistic regression model, $a(\phi) = 1$, $b(z) = \log(1 + e^z)$, $b'(z) = e^z/(1 + e^z)$, and $b''(z) = e^z/(1 + e^z)^2$.

For a study with n unrelated individuals, the data consist of $(Y_i, \mathbf{G}_i, \mathbf{X}_i)$ ($i = 1, \dots, n$). The score statistic for testing the null hypothesis $H_0: \boldsymbol{\beta} = \mathbf{0}$ takes the form

$$\mathbf{U} = \frac{1}{a(\hat{\phi})} \sum_{i=1}^n \{Y_i - b'(\hat{\boldsymbol{\gamma}}^T \mathbf{X}_i)\} \mathbf{G}_i,$$

where $\hat{\boldsymbol{\gamma}}$ and $\hat{\phi}$ are the restricted maximum likelihood estimators of $\boldsymbol{\gamma}$ and ϕ under H_0 . Under H_0 , $\mathbf{U}$ is asymptotically m -variate normal with mean $\mathbf{0}$ and covariance matrix

$$\mathbf{V} = \frac{1}{a(\hat{\phi})} \left[\sum_{i=1}^n b''(\hat{\boldsymbol{\gamma}}^T \mathbf{X}_i) \mathbf{G}_i \mathbf{G}_i^T - \left\{ \sum_{i=1}^n b''(\hat{\boldsymbol{\gamma}}^T \mathbf{X}_i) \mathbf{G}_i \mathbf{X}_i^T \right\} \times \left\{ \sum_{i=1}^n b''(\hat{\boldsymbol{\gamma}}^T \mathbf{X}_i) \mathbf{X}_i \mathbf{X}_i^T \right\}^{-1} \left\{ \sum_{i=1}^n b''(\hat{\boldsymbol{\gamma}}^T \mathbf{X}_i) \mathbf{X}_i \mathbf{G}_i^T \right\} \right],$$

which is the information matrix for $\boldsymbol{\beta}$ evaluated at $\boldsymbol{\beta} = \mathbf{0}$, $\boldsymbol{\gamma} = \hat{\boldsymbol{\gamma}}$, and $\phi = \hat{\phi}$.

For common variants, we adopt the quadratic form

$$\mathbf{U}^T \mathbf{V}^{-1} \mathbf{U},$$

whose null distribution is χ_m^2 . In the case of $m = 1$, this is the conventional single-variant test. For rare variants, it is common to use the burden test

$$(\mathbf{c}^T \mathbf{U})^2 / \mathbf{c}^T \mathbf{V} \mathbf{c},$$

where $\mathbf{c}$ is an m -vector of weights depending on the MAFs (Lin and Tang, 2011). By default, $\mathbf{c} = \mathbf{1}$. The null distribution of this test statistic is χ_1^2 . To detect variants with opposite effects, we use SKAT

$$\mathbf{U}^T \mathbf{W} \mathbf{U},$$

where $\mathbf{W}$ is a diagonal weight matrix that depends on the MAFs through a beta function (Wu et al., 2011). The null distribution of this test is $\sum_{j=1}^m \lambda_j \chi_{1,j}^2$, where $\lambda_1, \dots, \lambda_m$ are eigenvalues of $\mathbf{V}^{1/2} \mathbf{W} \mathbf{V}^{1/2}$, and $\chi_{1,1}^2, \dots, \chi_{1,m}^2$ are independent χ_1^2 random variables.

In WES studies, it is customary to perform single-variant tests for common variants and gene-based tests, typically both burden and SKAT tests, for rare variants. In WGS studies, we may apply burden and SKAT tests to genes or sliding windows. The windows may have the same size or different sizes, and they may be overlapping or non-overlapping. The window size may be determined by the total number of rare variants or the total number of base pairs. A prevailing approach is to use relatively large windows, such as 5kb or 50kb windows, with skip length of 2.5 kb or 25 kb, respectively (Morrison et al., 2017). A contrasting approach is to consider all possible contiguous windows with varying sizes within each region. In any case, the association tests among different windows will be correlated because of LD and possible overlap of variants.

Suppose that we perform a total of K association tests for the global null hypothesis over the genome, with p -values $p_1, \dots, p_K$. For $k = 1, \dots, K$, we declare the k th test to be statistically significant if the corresponding p -value p_k is less than a certain threshold p_0 . To control the family-wise error rate at α , we choose p_0 such that

$$\Pr(\min_{k=1, \dots, K} p_k \leq p_0) = \alpha. \quad (1)$$

We estimate p_0 by parametric bootstrap (Davison and Hinkley, 1997). Specifically, we simulate the phenotype value $\tilde{Y}_i$ from the estimated null density function

$$\exp \left\{ \frac{y(\hat{\gamma}^T \mathbf{X}_i) - b(\hat{\gamma}^T \mathbf{X}_i)}{a(\hat{\phi})} + c(y, \hat{\phi}) \right\}.$$

We then perform the K association tests on the bootstrap sample $(\tilde{Y}_i, \mathbf{G}_i, \mathbf{X}_i) (i = 1, \dots, n)$ to obtain the p -values $\tilde{p}_1, \dots, \tilde{p}_K$. Because $(\tilde{p}_1, \dots, \tilde{p}_K)$ and $(p_1, \dots, p_K)$ have approximately the same joint distribution, we approximate equation (1) by

$$\Pr(\min_{k=1, \dots, K} \tilde{p}_k \leq p_0) = \alpha.$$

We solve this equation empirically by bootstrapping the data R times. For $r = 1, \dots, R$, let p_r^* denote the smallest value of $\tilde{p}_1, \dots, \tilde{p}_K$ on the r th bootstrap sample. We sort the p_r^* from the smallest to the largest and set p_0 to the (αR) th value of the sorted p_r^* . It is customary to adopt $\alpha = 0.05$. We recommend $R = 10,000$, such that $\alpha R = 500$. (With $R = 10,000$ and $\alpha = 0.05$, the standard error for the estimated family-wise error rate is approximately 0.002, which is very small relative to the target level of 0.05.)

We wish to avoid calculations of the p -values whenever possible. If the test statistics have the same null distribution, then it is not necessary to calculate the p -values of individual tests because the largest test statistic has the smallest p -value. Specifically, let Q_r^* be the largest value of the K test statistics on the r th bootstrap sample. We sort the Q_r^* from the largest to the smallest and set q_0 to the (αR) th value of the sorted Q_r^* . We then convert q_0 to p_0 by referring q_0 to the appropriate null distribution. If the test statistics have different null distributions, then we group them according to their null distributions. For the r th bootstrap sample, we identify the largest test statistic in each group of tests and calculate the corresponding p -value; the smallest p -value among all the groups is equal to the smallest p -value among the K tests, i.e., p_r^* .

The p -value threshold p_0 depends on the number of tests and their correlations. The correlations of the test statistics are determined by the correlations of the score statistics. Let U_1 and U_2 be two score statistics with genotypes G_1 and G_2 , respectively. The covariance matrix between U_1 and U_2 is

$$\frac{1}{a(\hat{\phi})} \left[\sum_{i=1}^n b''(\hat{\gamma}^T X_i) G_{1i} G_{2i}^T - \left\{ \sum_{i=1}^n b''(\hat{\gamma}^T X_i) G_{1i} X_i^T \right\} \times \left\{ \sum_{i=1}^n b''(\hat{\gamma}^T X_i) X_i X_i^T \right\}^{-1} \left\{ \sum_{i=1}^n b''(\hat{\gamma}^T X_i) X_i G_{2i}^T \right\} \right].$$

If the covariates are independent of or weakly correlated with the genotypes, then this covariance matrix is approximately

$$\frac{1}{a(\hat{\phi})} \left[n^{-1} \sum_{i=1}^n b''(\hat{\gamma}^T X_i) \sum_{i=1}^n G_{1i} G_{2i}^T - \left\{ n^{-1} \sum_{i=1}^n G_{1i} \sum_{i=1}^n b''(\hat{\gamma}^T X_i) X_i^T \right\} \times \left\{ \sum_{i=1}^n b''(\hat{\gamma}^T X_i) X_i X_i^T \right\}^{-1} \left\{ n^{-1} \sum_{i=1}^n b''(\hat{\gamma}^T X_i) X_i \sum_{i=1}^n G_{2i}^T \right\} \right],$$

which reduces to

$$\frac{1}{a(\hat{\phi})} n^{-1} \sum_{i=1}^n b''(\hat{\gamma}^T X_i) \sum_{i=1}^n G_{1i} G_{2i}^T$$

after centering the genotype values. Thus, the correlation matrix between U_1 and U_2 is approximately the same as the correlation matrix between G_1 and G_2 . This result, which was used by Lee et al. (2013) and Pasniuc et al. (2014) without rigorous proofs, implies that, given the association tests to be performed, the p -value threshold p_0 is determined by the genotypes and not by the phenotype or covariates. Indeed, we can simply simulate $\tilde{Y}_1, \dots, \tilde{Y}_n$ from the standard normal distribution in the bootstrap procedure, such that the actual phenotype and covariate data are not used at all. Thus, the same threshold can be used in different studies, which facilitates the calculation of power in the design stage and the comparison of results across studies.

The above simplification hinges on the assumption of independence (or weak correlation) between covariates and genotypes. Most covariates, such as age and environment, are independent of genotypes. Although principal components for ancestry are constructed from genotypes, the correlation between a principal component and a particular set of variants is small. The only covariate that may be strongly correlated with genotypes is race. We recommend to calculate score statistics separately for each race group, as is commonly done in practice, such that no covariate is strongly correlated with genotypes within each race group. Both race-specific and race-combined thresholds can be obtained.

3. RESULTS

3.1. NHLBI ESP

The NHLBI ESP was designed to identify genetic variants in all protein-coding regions of the human genome that are associated with heart, lung, and blood disorders (Auer et al., 2016). Approximately 7,000 individuals were selected from seven population-based cohorts: Atherosclerosis Risk in Communities, Coronary Artery Risk Development in Young Adults, the Cardiovascular Health Study, the Framingham Heart Study, the Jackson Heart Study, the Multi-Ethnic Study of Atherosclerosis, and the Womens Health Initiative. The DNA samples were sequenced on the Roche NimbleGen SeqCap EZ or the Agilent SureSelect Human All Exon 50 MB at the University of Washington or the Broad Institute. The variants were called at the University of Michigan. We removed all variants with call rates $< 90\%$. We used ANNOVAR with GENCODE genes (v.7; UCSC Genome Browser, hg19) to annotate variants as nonsense, splice, read-through, missense, synonymous, untranslated region, or noncoding, and we selected the most deleterious annotation for each variant.

For this illustration, we aim to identify rare variants that influence low-density lipoprotein cholesterol (LDL-C). After removing individuals with missing LDL-C measurements or quality-control issues (i.e., sex mismatch, close relatives), we had a total of 3,665 individuals. We adjusted the LDL-C values by medication, age, and gender and included race and top two principal components for ancestry as covariates in a linear regression model. We performed eight types of gene-based association tests: burden tests for non-synonymous variants with MAFs $< 5\%$, 1% , 0.5% , and 0.1% ; SKAT for non-synonymous variants with MAFs $< 5\%$ and 1% ; and burden test and SKAT for loss-of-function (LoF) variants with MAFs $< 5\%$. For all eight tests, we excluded genes with minor allele counts < 10 . For non-synonymous variants, the number of genes with MAFs $< 5\%$, 1% , 0.5% , or 0.1% is 13,812, 13,587, 13,414, or 12,360, respectively. For LoF variants, the number of genes with MAFs $< 5\%$ is 852.

We used the proposed bootstrap method to calculate the genomewide significance threshold for each of the eight types of tests, which is the threshold for declaring genomewide significance if only one type of test is performed. The eight thresholds are shown in the middle column of Table 1. These thresholds are only slightly less stringent than their Bonferroni counterparts, which reflects the fact that LD among rare variants is weak.

We also used the proposed bootstrap method to calculate the overall genomewide significance threshold for the eight types of tests, which is the threshold for declaring

genomewide significance if all eight types of tests are performed. The overall genomewide significance threshold is 1.03×10^{-6} . By contrast, the Bonferroni correction for the total number of tests is 6.08×10^{-7} . The former is appreciably less stringent than the latter because different types of tests for the same gene are correlated.

Figure 1 displays the p -values of the SKAT test for non-synonymous variants with MAFs $< 1\%$, together with the bootstrap and Bonferroni thresholds for performing all eight types of tests. The top 3 genes are *PCSK9*, *MKNK2*, and *KIF4A*, with p -values of 2.38×10^{-8} , 1.29×10^{-7} , and 7.19×10^{-7} , respectively. The last p -value lies between the bootstrap and Bonferroni thresholds. Thus, we would be able to declare genomewide significance for *KIF4A* with the bootstrap threshold, but we would not be able to do so with the Bonferroni correction.

For comparison, we applied the proposed bootstrap method without using phenotype or covariate data. The genomewide significance threshold for each type of test and the overall genomewide significance threshold are displayed in the last column of Table 1. These thresholds are highly similar to their counterparts using phenotype and covariate data. This finding supports the theoretical result that genomewide significance thresholds are determined primarily by the genotype data.

For further illustration, we used the proposed bootstrap method without phenotype or covariate data to calculate the genomewide significance thresholds for single-variant tests with MAF cutoffs of 5%, 1%, 0.5%, and 0.1%. We included all 5,108 individuals who passed the quality control criteria, and we considered African Americans (2,023 individuals) and European Americans (3,085 individuals) separately. The results are summarized in Table 2. Because of LD, the bootstrap thresholds are less stringent than their Bonferroni counterparts, especially for common SNPs. The differences between the two thresholds are more profound for European Americans than for African Americans since LD is weaker in African populations than in European populations.

3.2. 1000 Genomes Project

The 1000 Genomes Project set out to provide a comprehensive description of human genetic variation by applying WGS to a diverse set of individuals from multiple populations (The 1000 Genomes Project Consortium, 2015). The final phase of the project reconstructed the genomes for 2,504 individuals from 26 populations in Africa, East Asia, Europe, South Asia, and the Americas. The database contains over 88 million variants, including 84.7 million SNPs, 3.6 million short indels, and 60,000 structural variants. The individuals in the project are anonymous and have no associated phenotype or covariate data.

We used the proposed bootstrap method without phenotype or covariate data to calculate genomewide significance thresholds for various types of association tests. We first considered single-variant tests for all common and low-frequency biallelic SNPs. There are 7,110,201, 12,512,618, 16,011,569, and 27,128,318 SNPs with MAFs 5%, 1%, 0.5%, and 0.1%, respectively. Table 3 displays the genomewide significance thresholds for the single-variant tests with these four MAF cutoffs. The thresholds generated by the proposed bootstrap method are roughly three times higher than their Bonferroni counterparts, the

difference being 3.5 times for MAFs $\geq 5\%$. The bootstrap results suggest that 5×10^{-9} is an appropriate genomewide significance threshold for testing around 30 million common and low-frequency SNPs.

Next, we considered 5-kb sliding windows with skip length of 2.5 kb for rare variants (Morrison et al., 2017); there are a total of 33,236 such windows. We considered burden tests with MAF cutoffs of 5%, 1%, 0.5%, and 0.1%. The corresponding genomewide significance thresholds are shown in Table 4. The thresholds generated by the proposed bootstrap method are considerably less stringent than their Bonferroni counterparts, especially for higher MAF cutoffs. If all four tests are performed, the overall genomewide significance threshold by the proposed method is 7.88×10^{-7} , whereas the Bonferroni threshold is 3.76×10^{-7} .

4. DISCUSSION

We presented a simple and accurate method to evaluate genomewide significance for sequencing studies. We illustrated the usefulness of this method with gene-based and sliding-window association tests, as well as with single-variant tests. There is currently no consensus for how to perform association tests in WGS studies, and new strategies may emerge and gain acceptance. Our method is universal and can be applied to any testing strategies.

Our description has been focused on generalized linear models for unrelated individuals; however, the proposed approach has broad applications. In particular, we can show that the correlations of the test statistics for potentially censored age at onset (Lin and Tang, 2011) are approximately the same as the correlations of the genotypes, such that the same thresholds can be used for non-censored and censored phenotypes. In addition, the parametric bootstrap method is applicable to family studies. Finally, although we have implicitly assumed the additive model in the Methods and Results sections, the proposed method is applicable to any mode of inheritance as long as the p -value for each test statistic can be calculated.

The proposed method is easy to implement, especially for the case of no phenotype and no covariates. We showed, both theoretically and empirically, that phenotype and covariate data can be generally disregarded even when they are available. The computation is very fast for single-variant tests and burden tests. For SKAT, the computational burden increases rapidly as the number of variants increases. We have posted our software at <http://dlin.web.unc.edu/software/threshold/>.

Instead of simulating phenotype values, one may randomly shuffle the observed phenotype values. Unlike the parametric bootstrap method, the permutation approach requires access to phenotype data. If the phenotype is discrete or continuous but non-normal, then the p -values for individual test statistics may be inaccurate. By contrast, our method can always simulate normally distributed phenotypes to ensure the accuracy of the p -values.

An alternative approach is to evaluate the empirical correlations of the test statistics (Lin, 2004). This approach also requires access to phenotype data. In addition, it relies on the

multivariate normality of the test statistics, which may be problematic at the extreme tails of the null distribution, especially for rare variants.

A simulation study using data from the HapMap ENCyclopedia Of DNA Elements regions to emulate an infinitely dense map yielded a genomewide significance threshold of 5×10^{-8} for association tests with common SNPs in Europeans (Pe'er et al., 2008). In addition, a genomewide significance threshold of 7.2×10^{-8} was obtained by subsampling genotypes from the Wellcome Trust Case-Control Consortium at increasing density and extrapolating to infinite density (Dudbridge and Gusnanto, 2008). Another approach using sequence simulation under various demographic and evolutionary models found a genomewide significance threshold of 3.1×10^{-8} for a European population (Hoggart et al., 2008). Thus, a genomewide significance threshold of 5×10^{-8} has been widely adopted for studies on European populations, regardless of the actual density of the genotyping array (Sham and Purcell, 2014). Our work, however, showed that this threshold is too liberal for single-variant tests with common and low-frequency SNPs in WGS studies. The threshold of 5×10^{-9} should be used for testing approximately 30 million SNPs with MAF $\geq 0.1\%$; see Table 3 for a range of thresholds conditional on the minimum MAF analyzed.

For single-variant association tests in WGS studies, the proposed bootstrap thresholds, which properly account for LD, are 3–4 times higher than the Bonferroni corrections. The ratios are smaller for WES data. There are modest differences across ancestry groups due to different degrees of LD.

Applying an auto-correlation-based approach to the WGS data from the 1000 Genomes Project, Sobota et al. (2015) obtained the ratio of the number of markers tested to the effective number of independent tests in the range of 12–37, which is much larger than ours. By randomly generating case-control phenotypes for the 1000 Genomes data, Kanai et al. (2016) also obtained genomewide significance thresholds that are more liberal than ours. On the other hand, by using the WGS data from chromosome 3 in the UK10K project and randomly assigning normally distributed phenotypes and then extrapolating from the length of chromosome 3 to the length of the whole genome, Xu et al. (2014) obtained genomewide significance thresholds that are slightly more stringent than ours.

Although our method covers any tests, we did not consider SKAT-O (Lee et al., 2012) in our numerical examples for several reasons. First, SKAT-O is not as interpretable as the burden and SKAT tests. Second, there is no need to perform SKAT-O if the burden and SKAT tests are already performed. Third, SKAT-O is computationally demanding, especially for WGS studies. Indeed, only the burden and SKAT tests have been used in recent publications on WGS studies (e.g., Natarjan et al., 2018; Zekavat et al., 2018).

TOPMed has sequenced ~ 55,000 individuals thus far, and the next data freeze will contain > 100,000 individuals. A large number of working groups have been formed to investigate the genetic architecture of complex diseases and traits using these sequencing data. Currently, 5×10^{-8} is used as the genomewide significance threshold for single-variant tests, and the Bonferroni correction is used for region-based association tests. The former is too liberal whereas the latter is overly conservative, as demonstrated in Section 3. The thresholds we

calculated for the 1000 Genome Project should be good approximations for the TOPMed sequencing studies with similar sets of variants, and the proposed method can be readily applied to the TOPMed data to obtain accurate genomewide significance thresholds for any types of tests.

Acknowledgments

This work was supported by NIH awards R01HG009974, R01GM047845, and P01CA142538. The author thank Christopher Sheldahl for programming assistance and two reviewers for helpful comments.

REFERENCES

- Auer PL, Reiner AP, Wang G, Kang HM, Abecasis GR, Altshuler D, Bamshad MJ, Nickerson DA, Tracy RP, Rich SS, et al. (2016). Guidelines for large-scale sequence-based complex trait association studies: lessons learned from the NHLBI Exome Sequencing Project. *Am J Hum Genet* 99, 791–801. [PubMed: 27666372]
- Davison AC and Hinkley DV (1997). *Bootstrap Methods and Their Application*. Cambridge University Press.
- Dudbridge F and Gusnanto A (2008). Estimation of significance thresholds for genomewide association scans. *Genet Epidemiol* 32, 227–234. [PubMed: 18300295]
- Hoggart CJ, Clark TG, De Iorio M, Whittaker JC, and Balding DJ (2008). Genome-wide significance for dense SNP and resequencing data. *Genet Epidemiol* 32, 179–185. [PubMed: 18200594]
- Kanai M, Tanaka T, and Okada Y (2016). Empirical estimation of genome-wide significance thresholds based on the 1000 Genomes Project data set. *J Hum Genet* 61, 861–866. [PubMed: 27305981]
- Lee D, Bigdeli B, Riley BP, Fanous AH, and Bacanu S-A (2013). DIST: direct imputation of summary statistics for unmeasured SNPs. *Bioinformatics* 29, 2925–2927. [PubMed: 23990413]
- Lee S, Emond MJ, Bamshad MJ, Barnes KC, Rieder MJ, Nickerson DA, NHLBI GO Exome Sequencing Project ESP Lung Project Team, Christiani DC, Wurfel MM, and Lin X (2012). Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. *Am J Hum Genet* 91, 224–237. [PubMed: 22863193]
- Lin DY (2004). An efficient Monte Carlo approach to assessing statistical significance in genomic studies. *Bioinformatics* 21, 781–787. [PubMed: 15454414]
- Lin DY and Tang ZZ (2011). A general framework for detecting disease associations with rare variants in sequencing studies. *Am J Hum Genet* 89, 354–367. [PubMed: 21885029]
- Madsen BE and Browning SR (2009). A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genet* 5, e1000384. [PubMed: 19214210]
- Morgenthaler S and Thilly WG (2007). A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: a cohort allelic sums test (CAST). *Mutat Res* 615, 28–56. [PubMed: 17101154]
- Morris AP and Zeggini E (2010). An evaluation of statistical approaches to rare variant analysis in genetic association studies. *Genet Epidemiol* 34, 188–193. [PubMed: 19810025]
- Morrison AC, Huang Z, Yu B, Metcalf G, Liu X, Ballantyne C, Coresh J, Yu F, Muzny D, Feofanova E, et al. (2017). Practical approaches for whole-genome sequence analysis of heart- and blood-related traits. *Am J Hum Genet* 100, 205–215. [PubMed: 28089252]
- Neale BM, Rivas MA, Voight BF, Altshuler D, Devlin B, Orho-Melander M, Kathiresan S, Purcell SM, Roeder K, and Daly MJ (2011). Testing for an unusual distribution of rare variants. *PLoS Genet* 7, e1001322. [PubMed: 21408211]
- Natarajan P, Peloso GM, Zekavat SM, Montasser M, Ganna A, Chaffin M, Khera AV, Zhou W, Bloom JM, Engreitz JM, et al. (2018). Deep-coverage whole genome sequences and blood lipids among 16,324 individuals. *Nature Communications* 9, 3391.

- Pasaniuc B, Zaitlen N, Shi, Bhatia G, Gusev A, Pickrell J, Hirschhorn J, Strachan DP, Patterson N, and Price AL (2014). Fast and accurate imputation of summary statistics enhances evidence of functional enrichment. *Bioinformatics* 30, 2906–2914. [PubMed: 24990607]
- Pe'er I, Yelensky R, Altshuler D, and Daly MJ (2008). Estimation of the multiple testing burden for genomewide association studies of nearly all common variants. *Genet Epidemiol* 32, 381–385. [PubMed: 18348202]
- Price AL, Kryukov GV, de Bakker PIW, Purcell SM, Staples J, Wei LJ, and Sunyaev SR (2010). Pooled association tests for rare variants in exon-resequencing studies. *Am J Hum Genet* 86, 832–838. [PubMed: 20471002]
- Sham PC and Purcell SM (2014). Statistical power and significance testing in large-scale genetic studies. *Nature Rev Genet* 15, 335–346. [PubMed: 24739678]
- Sobota RS, Shriner D, Kodaman N, Goodloe R, Zheng W, Gao Y-T, Edwards TL, Amos CI, and Williams SM (2015). Addressing population-specific multiple testing burdens in genetic association studies. *Ann Hum Genet* 79, 136–147. [PubMed: 25644736]
- The 1000 Genomes Project Consortium. (2015). A global reference for human genetic variation. *Nature* 526, 68–74. [PubMed: 26432245]
- Tzeng JY and Zhang D (2007). Haplotype-based association analysis via variance-components score test. *Am J Hum Genet* 81, 927–938. [PubMed: 17924336]
- Wu MC, Lee S, Cai T, Li Y, Boehnke M, and Lin X (2011). Rare-variant association testing for sequencing data with the sequence kernel association test. *Am J Hum Genet* 89, 82–93. [PubMed: 21737059]
- Xu C, Tachmazidou I, Walter K, Ciampi A, Zeggini E, Greenwood CMT, and the UK10K Consortium (2014). Estimating genome-wide significance for whole-genome sequencing studies. *Genet Epidemiol* 38, 281–290. [PubMed: 24676807]
- Zekavat SM, Ruotsalainen S, Handsaker RE, Alver M, Bloom J, Poterba T, Seed C, Ernst J, Chaffin M, Engreitz J, et al. (2018). Deep coverage whole genome sequences and plasma lipoprotein(a) in individuals of European and African ancestries. *Nature Communications* 9, 2606.

Figure 1.
Results of the SKAT test for non-synonymous variants with MAFs < 1% in the NHLBI ESP.

Table 1.

Genomewide Significance Thresholds for Gene-Based Association Tests in the NHLBI ESP

Bootstrap threshold				
Association tests	Number of tests	Bonferroni threshold	With phenotype and covariates	Without phenotype and covariates
Non-synonymous variants				
Burden: MAF < 5%	13,812	3.62×10^{-6}	3.91×10^{-6}	3.77×10^{-6}
Burden: MAF < 1%	13,587	3.68×10^{-6}	4.03×10^{-6}	4.04×10^{-6}
Burden: MAF < 0.5%	13,414	3.73×10^{-6}	3.98×10^{-6}	4.02×10^{-6}
Burden: MAF < 0.1%	12,360	4.04×10^{-6}	4.45×10^{-6}	4.44×10^{-6}
SKAT: MAF < 5%	13,812	3.62×10^{-6}	3.88×10^{-6}	3.89×10^{-6}
SKAT: MAF < 1%	13,587	3.68×10^{-6}	3.94×10^{-6}	3.83×10^{-6}
Loss-of-function variants				
Burden: MAF < 5%	852	5.87×10^{-5}	6.10×10^{-5}	5.99×10^{-5}
SKAT: MAF < 5%	852	5.87×10^{-5}	5.92×10^{-5}	5.76×10^{-5}
All eight tests	82,276	6.08×10^{-7}	1.03×10^{-6}	1.03×10^{-6}

Table 2.

Genomewide Significance Thresholds for Single-Variant Association Tests in the NHLBI ESP

MAF	African Americans			European Americans		
	Number of variants	Threshold		Number of variants	Threshold	
		Bonferroni	Bootstrap		Bonferroni	Bootstrap
5%	53,088	9.42×10^{-7}	2.24×10^{-6}	39,422	1.27×10^{-6}	3.70×10^{-6}
1%	103,924	4.81×10^{-7}	7.51×10^{-7}	54,243	9.22×10^{-7}	1.77×10^{-6}
0.5%	130,369	3.84×10^{-7}	5.37×10^{-7}	72,285	6.92×10^{-7}	1.15×10^{-6}
0.1%	214,937	2.33×10^{-7}	2.96×10^{-7}	107,447	4.65×10^{-7}	6.21×10^{-7}

Table 3.

Genomewide Significance Thresholds for Single-Variant Association Tests in the 1000 Genomes Project

MAF	Number of variants	Threshold	
		Bonferroni	Bootstrap
5%	7,110,201	7.03×10^{-9}	2.47×10^{-8}
1%	12,512,618	4.00×10^{-9}	1.27×10^{-8}
0.5%	16,011,569	3.12×10^{-9}	9.04×10^{-9}
0.1%	27,128,318	1.84×10^{-9}	5.61×10^{-9}

Table 4.

Genomewide Significance Thresholds for Rare-Variant Association Tests With 33,236 Sliding Windows in the 1000 Genomes Project

Association tests	Bonferroni	Bootstrap
Burden: MAF < 5%	1.50×10^{-6}	5.99×10^{-6}
Burden: MAF < 1%	1.50×10^{-6}	3.61×10^{-6}
Burden: MAF < 0.5%	1.50×10^{-6}	2.44×10^{-6}
Burden: MAF < 0.1%	1.50×10^{-6}	1.80×10^{-6}
All four tests	3.76×10^{-7}	7.88×10^{-7}