Simplified Power Calculations for Aggregate-level Association Tests Provide Insights to Challenges for Rare Variant Association Studies

Andriy Derkach[1*], Haoyu Zhang[2] and Nilanjan Chatterjee[2]

[1] Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Rockville MD 20850

[2] Department of Biostatistics, Bloomberg School of Public Health, and Department of Oncology, School of Medicine, John Hopkins University, Baltimore, MD 21205

Corresponding author

Nilanjan Chatterjee, Ph.D.

Department of Biostatistics

Bloomberg School of Public Health

Johns Hopkins University

615 N Wolfe St

Baltimore, MD 21205

E-mail: nilanjan@jhu.edu

Abstract

Genome-wide association studies are now shifting focus from analysis of common to uncommon and rare variants with an anticipation to explain additional heritability of complex traits. As power for association testing for individual rare variants may often be low, various aggregate level association tests have been proposed to detect genetic loci that may contain clusters of susceptibility variants. Typically power calculations for such tests require specification of large number of parameters, including effect sizes and allele frequencies of individual markers, making them difficult to use in practice. In this report, we approximate power to varying degree of accuracy using a smaller number of key parameters, including the total genetic variance explained by multiple variants within a locus. We perform extensive simulation studies to assess the accuracy of the proposed approximation in realistic settings. Using the simplified power calculation methods, we then develop an analytic framework to obtain bounds on genetic architecture of an underlying trait given results from a genome-wide study and observe important implications for lack or limited number of findings in many currently reported studies. Finally, we provide insights into the required quality of annotation/functional information for identification of likely causal variants to make meaningful improvement in power of subsequent association tests. A shiny application in R implementing the methods is made publicly available.

Introduction

Over the last decade genome-wide association studies of common variants of increasingly sample size have been the main driving force for discovery of susceptibility loci associated with complex diseases and traits. While analysis of heritability suggests that common variants have further ability to explain additional variation of these traits [1-3], the focus of the field is inevitably shifting towards studies of less common and rare variants with the rapidly decreasing cost of sequencing technologies and increasing sophistication of imputation algorithms [4-6]. However, limited or lack of findings from early studies [7-21] indicate that effect sizes of rare susceptibility variants in general are likely to be modest and discovery of underlying loci will require large sample size in future studies[22-24].

Testing of association at the level of genetic loci or regions using various aggregate-level statistics have been proposed as a strategy to improve power of discovery in association studies of rare variants[25-28]. Simulation studies have been used under various anticipated genetic architecture of the traits for demonstration of potential power of these procedures [27-29]. In particular, analysis of power for variance component based test, such as the popular SKAT method, can be complex as they require specification of many different parameters including number of genetic variants under study, proportion of causal variants, allele frequency and effect-size distributions. Use of various functional and annotation information to identify likely pathogenic variants a priori has also been proposed as a strategy to improve power of rare variant association tests[30; 31]. To the best of our knowledge, however, there has been no systematic study of the effect of use such extraneous information on power of the association tests.

In this report, we first describe a first-order approximation that allows analytic characterization of power for popular variance component association tests simply based on the degree of phenotypic variance a locus explain and the number of variants under study – thus dramatically reducing the complexity of the power calculations. We perform simulation studies using allele frequency distribution observed in Exome Aggregation Consortium (ExAC)[32] and various models for effect-size distribution to assess the accuracy of the proposed approximation in realistic settings. We then use this simplified framework to characterize power of association tests that may pre-select variants based on prior functional/annotation information. These derivations allow us to study effect of sensitivity and specificity of extraneous information to identify causal variants on the power the association tests.

We assess the power of a number of recently reported association studies of rare variants using the proposed framework and provide insights into the implications for current lack of findings on bounds of genetic architecture of the underlying traits. Our analysis also provides important insights into the required quality of annotation/functional information for identification of likely causal variants to make meaningful improvement in power of subsequent association tests. Finally, to facilitate convenient and rapid power calculations for rare variant association tests, we make a shiny app PCAAT (Power Calculations for Aggregated Association Test) available in R.

Methods:

Existing Power Calculations

Many tests for association at the level of a genetic loci or a region aggregating SNP level association statistic have been proposed [25-29; 33-36]. Multiple authors [22; 33; 37] have shown that existing methods can be

classified as sum-based tests [25; 26; 34; 35], variance component tests [27; 29; 37], and hybrid tests that are functions of both classes [28; 33; 36]. Here, we focus on methods from sum-based and variance component classes. We do not consider hybrid tests because their power is usually close to one of the two components. Sum-based tests aggregate SNP level association statistics by a linear combination

$$T_{ST} = \sum_{j=1}^{J} w_j T_j,$$

where w_js are MAF(p_j)-based weights and T_is are score statistics for J SNPs $(j = 1, \dots, J)$ from linear or logistic regressions [28; 33; 37]. Variance component tests aggregate SNP level association statistics by quadratic combination:

$$T_{VC} = \sum_{j=1}^{J} w_j T_j^2.$$

Existing analytic power formulas for sum-based and variance component tests are complex functions of many parameters including number of genetic variants under study, proportion of causal variants, allele frequencies and effect-size distributions [37]. Here we show that under modest model assumptions such as genetic variants explain small proportion of phenotypic variation and low correlation between rare variants, existing power calculations can be simplified substantially.

We start with observing that analytic power of a single SNP statistic Z_j (e.g. Wald's, Score tests),

$$Z_j^2 = \frac{T_j^2}{Var(T_j)} \sim \chi^2_{1,nc_j}$$

with a non-centrality parameter $nc_j = 2p_j(1-p_j){\beta_j}^2 N = EV_j N$, is a function of three parameters: effective sample size (N), level of the test (α) and proportion of phenotypic variation explained by the j^{th} SNP (EV_j)[38]. Derkach et al.[37] showed analytic power for a sum-based test statistic Z_{ST},

$$Z_{ST}^2 = \frac{T_{ST}^2}{Var(T_{ST})} \sim \chi^2_{1,nc_L},$$

depends on non-centrality parameter $nc_{ST} = N \frac{\left(\sum_{j=1}^{J} w_j sign(\beta_j)\sqrt{p_j(1-p_j)EV_j}\right)^2}{\sum_{j=1}^{J} w_j^2 p_j(1-p_j)}$, a function of a *vector* of SNP level coefficients of explained variation.

Previous studies[27; 37] have shown that a variance component statistic is asymptotically distributed as a linear combination of non-central chi-square random variables,

$$T_{VC} \sim \sum_{j=1}^{J} \lambda_j \chi^2_{1,nc_j},$$

with non-centrality parameters $nc_j = EV_j N$ as a functions of phenotypic variation explained by a single SNP and weights $\lambda_j = w_j p_j(1-p_j)N$. Power calculations for a variance component statistic use method derived in Liu et al. (2009) [39] to approximate asymptotic distribution of T_{VC} by single non-central chi-square distribution. There are also several modifications of this method matching higher moments to improve the tail probability approximation [27; 40]; however, a power difference seems to be marginal and we focus on Liu's approximation here. Non-centrality parameter and degrees of freedom of chi-square distribution are calculated by matching at four cumulants c_k of the test statistic T_{VC},

$$c_k = \sum_{j=1}^{J} \lambda_j^k + kN \sum_{j=1}^{J} \lambda_j^k EV_j \text{ for } k = 1, \dots, 4, \qquad (1)$$

a function of a *vector* of SNP level coefficients of explained variation.

These power calculations for aggregate level tests are implemented in several statistical packages that require specification of MAFs and genetic effects for all SNPs in a locus [27; 40; 41].

Approximate Power Calculations for Aggregate Tests

Following, we describe simple formulae for approximating power for different aggregate level tests using a limited number of key parameters. First, we show that for sum-based test, under an assumption of independence between coefficients of explained variation and MAF, we can roughly estimate non-centrality parameter as

$$nc_{ST} \approx N \frac{|J_D - J_P|}{J} \frac{|J_D - J_P|}{J_D + J_P} EV,$$

where J_D, J_P is a number of deleterious and protective SNPs in a locus and $\sum_{j=1}^{J} EV_j = EV$ is a proportion of variation explained by J SNPs (see Appendix A). Hence power of linear statistic depends on *three parameters*, total variation explained by all causal variants in a locus, proportion of "effective" causal variants in a locus ($\frac{|J_D - J_P|}{J}$) and proportion of "effective" causal variants in a set of all causal variants ($\frac{|J_D - J_P|}{J_D + J_P}$). If all the causal variants in a locus can be assumed to deleterious (or protective), then the non-centrality parameter can be characterized by EV and the proportion of causal variants.

Next we consider first- and second-order approximations for power calculations for variance component tests T_{VC}. Exact theoretical power is calculated from single chi-square distribution with degrees of freedom and non-centrality parameter obtained by matching cumulants c_k. The first order approximation uses the same approach but estimates cumulants c_k in (1) as a function of a proportion of a variance explained by a locus EV and number of variants in a locus J. The second order approximation uses, one additional parameter, number of causal variants in a locus $J_C = J_P + J_D$ to improve accuracy of the first-order approximation.

For the first order approximation, we propose to estimate the sum $\sum_{j=1}^{J} \lambda_j^k EV_j$ in c_k as an expectation of the form $JE\left(\lambda_j^k EV_j\right)$. We decompose the expectation as a product of EV and some function of λ_j^k and p_j, $f(\lambda_j^k, p_j)$ measures relationship between EV_j and MAF p_j. For example, if we assume independence between proportion of variation explained by SNP and MAF (e.g. $f(\lambda_j^k, p_j) = 1$), we approximate c_k as

$$c_k = \sum_{j=1}^{J} \lambda_j^k + kN \sum_{j=1}^{J} \lambda_j^k EV_j \approx \sum_{j=1}^{J} \lambda_j^k + \frac{\sum_{j=1}^{J} \lambda_j^k}{J} EV$$

In Appendix B, we derive approximations of c_k for three commonly assumed relationships between genetic effects and MAFs and summarize them in Table 1. Because the first order approximation implicitly treats all variants in a locus as causal which may result in loss of accuracy when signal is very sparse in a locus. To improve accuracy, we propose the second order approximation to estimate the sum

$\sum_{j=1}^{J} \lambda_j^k EV_j$ in (1) as function of EV and J_C by using the same technique (see Appendix B). For example, if we assume the same hypothesis of independence between proportion of variation explained by SNP and MAF and that the first J_C variants are causal, sum $\sum_{j=1}^{J} \lambda_j^k EV_j$ in c_k is approximated as $\frac{1}{J_C}\sum_{j=1}^{J_C} \lambda_j^k \, EV$.

Probability of M Discoveries in Genome-wide Studies

Using the proposed simplified power calculation framework, we further develop a mathematical framework to study bounds on genetic architecture of underlying traits from results reported in a genome-wide association study. We characterize probability of number of discoveries in a given study as a function of sample size N, number of underlying causal loci K and distribution of their effect-sizes (EV), i.e. the total genetic variances the loci explain aggregated over the underlying susceptibility variants. In Appendix C we show that probability of M discoveries in a study is

$$\mathrm{P}(\mathrm{M\ Discoveries}|\mathrm{Genetic\ Model}) \approx \binom{K}{M} E\left(pow(\alpha, N, EV_k, J_k, \boldsymbol{p}_k)\right)^M e^{(K-M)E\left(\log(1-pow(\alpha,N,EV_k,J_k,\boldsymbol{p}_k))\right)}, \quad (2)$$

which is essentially a function of average power of a variance component test calculated over distribution of causal loci characterized by three parameters: a proportion of phenotypic variation explained EV_k, number of SNPs in a locus J_k and vector of MAFs $\boldsymbol{p}_k$.

Now, suppose $\mathrm{M} = \mathrm{m}$ is the number of discoveries reported based on gene-based tests in a given genome-wide association study of sample size N. We can calculate $\mathrm{P}(\mathrm{M} \leq \mathrm{m})$ using the above formula based on known distributions for MAFs ($\boldsymbol{p}_k$), and size of genes (J_k). In our calculations, we generated a class of L-shaped effect-size distribution a two-parameter gamma distribution: Gamma(α, γ) with $\alpha \leq 1$. Under this model, the total heritability explained by causal loci is given $K\mu$ by where $\mu \approx \alpha\gamma$. For various combination of K and μ, we evaluate the maximum value of $\mathrm{P}(\mathrm{M} \leq \mathrm{m})$ over different values of the dispersion parameter (α/γ). When this probability is low (e.g. < 5%), we conclude the underlying model for genetic architecture is unlikely. For example, many recent studies have reported no discoveries based on gene-level aggregated association tests. In these studies, the probability of no discoveries, $\mathrm{m} = 0$ can be used to provide bound on genetic architecture of the underlying trait.

Effects of filtering variants by extraneous information

We use our simplified framework to efficiently study the effects of filtering of variants based on prior functional/annotation information on the power of association tests. Here, power of association tests can be summarized as function of sensitivity and specificity of a filtering method. Sensitivity (Se) is a probability of selecting a SNP given that it is truly causal variant, while specificity (Sp) represents a probability of a filtering a SNP out given that is non-causal variant. If selection/filtering is independent of MAF or proportion of variation explained by a causal variant, then number of remaining variants in a locus is $J_S = Se \cdot J_C + (1 - SP) \cdot (J - J_C)$ and proportion variation explained by them is $EV_S = Se \cdot EV$. Now, with new values of J and EV, we estimate power for a sum based and a variance component tests and compare it with base values.

If all SNPs are selected, then $Se = 100\%$, but specificity is $Sp = 0\%$. If only a small subset of variants is selected within a gene based on functional annotation, then sensitivity may be reduced as some true causal variants could be missed while specificity may improve because of removal of non-causal SNP. If one takes random subset of SNPs, which is not enriched by causal SNP, then $Se = 1 - Sp$ as the casual and non-causal SNPs are selected at the same rate. If the functional/annotation information used for

screening is predictive of whether the SNPs are likely to be causal for the trait of interest, then one would expect specificity > 1- sensitivity. Using the proposed power calculation framework, we explore power of aggregated tests for various combinations of sensitivity and specificity of the underlying screening algorithm for identifying true causal SNP. Further, assuming than an underlying normally distributed continuous score represent the functional/annotation information for the SNPs, we evaluate receiver operating characteristic (ROC) curve generated by combination of sensitivity and specificity at different threshold for SNP selection. We track power of different methods along combinations of sensitivity and specificity that leads to specific value of the area under the curve (AUC), which is an overall summary of the ability of the underlying score to discriminate between causal and non-causal SNP (see Figure 5).

Empirical Studies:

Properties of the first- and second-order approximations

We conduct extensive simulation studies to evaluate accuracy of the proposed power calculations for variance component tests in comparison to exact theoretical method that require specification of effect-sizes of individual markers. Here we focus on SKAT test statistic as representative of the class of variance component tests. For each fixed combination of size of region (J) and total variance explained (EV), the two key parameters that determine the approximate power of the SKAT test, we simulate various possible values of allele frequencies and effect sizes for individual markers EV_j. Then average power of the first- and the second-order approximations calculated over various values of allele frequencies is compared with average power of exact theoretical calculations estimated over various values of allele frequencies and effect sizes for individual markers EV_j.

We consider three types of simulation scenarios: S1 ("MAF-independent EV") assumes that coefficients of explained variation EV_j. is independent of MAF; S2 ('MAF-independent β_j') assumes that size of genetic effect β_j is independent of MAF and S3 ('MAF-log-dependent β_j') assumes that genetic effect is related to MAF through $log10$ function (as defined in Table 1). For each type of simulation scenario, we estimate power for a locus of the size $J = 50, 100, 200 \text{ and } 400$. We additionally consider four values of a number of causal SNP $J_C = 10, 20, 30 \text{ and } 50$ for calculations based on exact theoretical and second-order calculations. In Appendix D, we describe simulation mechanisms in detail and we summarize simulation models and parameters required for each method in Table 2.

Bounds on variation explained by a causal locus

In Table 3, we provide key parameters that summarize recently published association analysis with rare variants[7; 10; 13-15; 17-21]. Typically, studies on Human Exome BeadChip (Exome Chip) had larger sample sizes then studies on sequencing platform; however, they covered a much smaller number of rare variants. These much larger studies typically had at least one significant discovery, while studies with other platforms generally did not report any. We use our mathematical framework presented in the previous section to compare genetic bounds implied by these results. Here, we focus on results from the largest study with each sequencing and exome chip platforms[18; 21] (see 5th and 9th rows of the Table 3).

To estimate genetic bounds from the results of the first study on educational attainment with sequencing platform we use mathematical framework presented in previous section. We estimate probability of no discoveries (*m=0*) given a genetic model from formula (2) by using following parameters. We set effective sample size to $N = 15{,}000$ to match number of whole genome and

exome sequenced individuals in this study (see 5th row of the Table 3). We assume that analysis was conducted by the SKAT statistic with a gene as a unit. We use publicly available EXaC database[32] to obtain empirical distributions for number of rare variants in a gene (J_k) and vector of MAFs $\boldsymbol{p}_k$. To ensure validity of asymptotic power formulas, we assume that MAF of rare variant ranges between 0.0001 and 0.01 (e.g. no singletons and doubletons). Lastly, we set Type 1 error threshold $\alpha = \frac{0.05}{20{,}000} = 2.5 \cdot 10^{-6}$.

To estimate genetic bounds for the second study on blood pressure outcomes with Exome Chip platform, we use following parameters to estimate probability of at most three discoveries(m=3). We set effective sample size to $N = 140{,}000$ to match number of individuals genotyped by Human Exome BeadChip platform (see 9th row of the Table 3). Similarly, to the first study we assume that analysis was conducted by the SKAT statistic with a gene as a unit. Empirical distributions for number of rare variants in a gene (J_k) and vector of MAFs $\boldsymbol{p}_k$ are obtained from EXaC database. Compared to the first study, we only assume that MAFs for rare variants are smaller than 0.01. Lastly, we use the same Type 1 error threshold $\alpha = 2.5 \cdot 10^{-6}$.

For each combination of a number underlying of causal loci K and L-shaped effect size distribution, we calculate probability *m* discoveries using formula (2) with study specific parameters. Expectations in (2) are estimated using 100,000 Monte Carlo simulations. For each K and L-shaped effect size distribution, we recalculate power of the SKAT statistic using generated coefficient of variation explained EV_k, the number of variants in a locus J_k and a vector of MAFs p_k. Average values are then plugged in (2). In this report, we calculate power of the SKAT test statistic under assumption of independence between proportion of variation explained by a SNP and MAF. Results for other genetic architecture are also discussed and presented in the Supplementary Materials.

Effects of filtering variants by extraneous information on power

We consider two values of a size of a locus $J = 50, 100$ and two values of a number of causal variants in a locus, $J_C = 10, 20$. Initial value of *EV* is selected so that power of the SKAT test is equal to 40% at Type 1 error $\alpha = 0.05/20{,}000 = 2.5 \cdot 10^{-6}$ and sample size $N = 10{,}000$. For every combination of sensitivity and specificity, we estimate average power for the SKAT test with recalculated parameters J_S and EV_S,. With the same values of the coefficient of variation and the proportion of causal variants, we additionally investigate properties of burden test statistic assuming all causal variants are deleterious. For this empirical study, we also assume independence between proportion of variation explained by a SNP and MAF. Results for other genetic architectures are presented in the Supplementary Materials.

Results:

Properties of the First- and Second-Order Approximation

We evaluate accuracy of first- and second-order approximations compared to exact power calculation of variance component test under variety of genetic models (see Table 2). As expected, the first order approximation matches exact calculation better as number of causal variants in a locus J_C increases (*see Figure 1*). Particularly, we observe that with more than 20 causal variants in a locus ($J_C \geq 20$) difference in power between those two methods is small regardless of the total number of variants in a locus $J = 50, 100$ *(*see *Figure 1 B and D* and *Supplementary Figure S1)*. Similar conclusion holds also for very large loci $J = 200, 400$ and other relationships between genetic effect and MAF (see *Supplementary Figures S2, S3* and *S4*). With lower number of causal variants in a locus (e.g. $J_C = 10$), we observe upward bias

in first-order estimates when exact power is high and downward bias when exact power is low (see *Figure 1 A and C*).

We observe that the second-order approximation is more accurate at estimating the exact power (see Figure 2). Now, difference in power between approximate and exact calculations is small even when number of causal variants in a locus is small, $J_C = 10$ (*see Supplementary Figures* S5-S7). Overall, our simulations demonstrate that the first-order approximation accurately estimates exact power when number of causal variants in a locus is not too small. However, if a number of causal variants in a locus is small then the first order approximation may produce biased results. On the other hand, the second-order approximation estimates exact power more accurately regardless of underlying generic architecture but it requires specification of additional parameter, number of causal variants in a locus J_C.

Estimation of Upper Bounds on Effect Size Distribution

We use publicly available summary statistics from ExAC's database to obtain empirical distributions for size of a gene (J) and MAF ($\boldsymbol{p}$) for both studies on sequencing and Exome Chip platforms. We provide key parameters of these empirical distributions in Supplementary Figures S8-S11. As expected, study with Exome Chip platform has on average smaller number of variants in a gene than a study on whole exome sequencing (WES) platform. We observe that average number of rare variants per gene (J) are 35.5 and 13 in the studies on sequencing and Exome Chip platforms. We also note that the first study with 15,000 whole exome sequenced individuals observes 743,094 rare variants and the second study with 140,000 Exome Chip genotyped individuals observes 215,674 rare variants.

In Figure 3, we plot maximum probability of no discoveries in the study on educational attainment with sequencing platform[18] for various combinations of number of underlying causal loci *K* and total phenotypic variation explained by causal loci. We observe that models with small number of underlying causal loci and large values of total phenotypic variability are very unlikely (e.g. maximum probity is smaller than 0.05). For example, if underlying loci explain 20% of phenotypic variation of a trait then it is very unlikely to have less than 250 underlying causal loci. If we assume independence between genetic effect of a SNP and MAF (e.g. the SKAT test has more power), then large number of loci should be present to explain the same total heritability explained by underlying loci (*see Supplementary Figures S12*). For example, if underlying loci explain 20% of phenotypic variation, then it is unlikely to have less than 700 underlying causal loci.

In Figure 4, we plot maximum probability of observing three or less discoveries in much larger study on blood pressure with Exome Chip platform[21]. We observe that limited number of findings in the study with very large sample size and smaller number of variants per gene yields very sharp bound on relationship between number of underlying causal loci and total heritability explained by the causal loci. For example, if underlying causal loci explain 20% of phenotypic variation, then there should be at least 6,500 underlying casual loci. Identically to results for the WES study, genetic bound is even sharper if independence between MAF and genetic effect is assumed (*see Supplementary Figures S13*).

In conclusion, limited number of discoveries in two types of studies indicates substantial percentage of heritability of the underlying traits can be explained by rare variants under study only under highly polygenic model involving many causal loci with very small effects. It is also possible that rare variants included in these studies contributes very minimal to the heritability of the underlying traits.

Effects of SNP Selection on Power of Aggregated Test

Lastly, we investigated potential effects of pre-selecting variants by functional/annotation information on power of the gene based tests (Figure 5). We consider a setting where if all SNPs are selected within a gene the variance component and sum-test have moderate and comparable power (40% and 36%, respectively). Apriori SNP selection does not improve the power of variance-component test substantially (e.g. by 10%) unless the underlying algorithm has very high accuracy to discriminate between causal and non-causal SNPs (AUC between 80-90%). On the other hand, power for sum-based test can improve substantially with more modest discriminatory accuracy of the SNP selection algorithm (AUC between 70-80%). Further, we observe the role of sensitivity and specificity is not symmetric on power of the tests. For both tests, substantial improvement of power is possibly only if sensitivity is at the minimal 30-40%. On the other hand, substantial improvement in power is possible with fairly poor specificity (e.g. about 20%) as along as sensitivity is high (e.g. 90%). We observe the similar results in a studies with different genetic architecture and large number of SNPs in a locus (*Supplementary Figures S14-S16*).

Discussion:

Although large genome wide association studies of low frequency and rare variants are now becoming increasingly feasible due to technological advances, the likely yield of such studies in future remain uncertain as studies conducted to date have only yielded limited number of findings[7-22]. For studies of common variants, which have mostly relied on association testing at the level of individual variants, we and other have shown that yield of genome-wide association studies critically depend on distribution of phenotypic variances explained by individual variants across the genome. For studies of rare variants, it has been suggested that tests for genetic associations be performed at an aggregated level by combining signals across multiple variants for powerful detection of underlying susceptibility loci[22; 25-27; 29]. In this report, we show that how power for some of these more complex tests critically relates to total genetic variances explained by multiple variants within a locus. Based on such power calculations, we assess bounds on distributions of locus-level genetic variances that are consistent with limited findings reported in current studies. Further, based on these simplified power calculations, we evaluate potential for improving power for aggregated tests by pre-selection of likely causal variants based on functional/annotation information.

Power analysis of a number of current studies of large sample sizes provides important bounds on genetic architecture of the underlying traits. In particular, our analysis suggests that rare variants investigated in these studies could explain significant fraction of heritability of the underlying traits only under highly polygenic models in which causal variants are distributed over hundreds or even thousands of different genetic loci. For example, lack of findings from gene-based analysis in the study of educational attainment involving 15,000 individuals[18] indicate that at the minimum of 250 loci needs to be involved to explain 20% heritability for the underlying traits based on rare variants that are included in this sequencing platform (e.g. WES+WGS). Similarly, a very few findings ($m=3$) from a much larger study of blood pressure involving 140,000 individuals[21] indicate that at the minimum of loci needs to be involved to explain 20% heritability of 6500 based on Exome Chip platform. These results are intuitive given that if a relatively small number, e.g. a few dozens, of genetic loci could explain a substantial fraction of heritability of these traits, then at least some of these loci will be detected by the sample size achieved so far in the current studies.

A number of rare variant studies that have conducted both individual-variant and aggregated tests have detected more genetic loci using the former than the later approach[10; 15; 20; 21] (see Table 3). Such finding is consistent with genetic architecture models where number of causal variants within susceptibility loci is sparse. In such scenario, aggregated tests will have low power and more informative bounds on

genetic architecture could be obtained by analysis of power for individual variant tests. The analytic formula we propose for calculating probability of certain number of discoveries under various models for genetic architecture can also be applied for single-variant tests. Applications in several real studies suggest that rare variants included in the studies can explain significant fraction of heritability of the trait only under highly polygenic disease architecture.

A variety of studies have studied genetic architecture of common variants by characterization of underlying heritability, number of susceptibility variants and effect-size distributions[24; 38]. All of these studies consistently point toward a highly polygenic model where disease etiology may involve thousands or even tens of thousands common susceptibility variants, each conferring only a modest risk, but in combination they can explain substantial variation in risk. Some recent studies have reported that low frequency and rare variant studies have the potential to explain significant fraction of heritability for selected traits[10; 42; 43]. Further insights into genetic architecture of these traits can be obtained by comparing observed number discoveries in these studies with those from simulated studies under different models for genetic architecture[23; 44]. The proposed analytic framework provides an alternative fast and simple way of evaluating expected discoveries for a large variety of genetic models and quantification of their plausibility given results from a given study.

Power calculations for aggregated tests with selected subset of SNPs point towards some challenges for use of functional and annotation information for discovery of susceptibility locus. Overall, it appears that pre-selection of SNPs can significantly improve the power of aggregated test only if the underlying functional/annotation information have fairly high accuracy to discriminate (AUC > 70-80%) between causal and non-causal variants for the underlying disease of interest. In particular, the algorithm should be highly sensitive to capture the underlying causal variants for a disease. Use of too stringent criterion for SNP selection may increase specificity, but will lead to decreased sensitivity and hence could lead to loss of power in aggregated tests. More empirical studies are needed to assess the impact of SNP selection on power of aggregated tests.

Sophisticated imputation algorithms[4; 5] and increasing sample size of reference datasets[4] allowing imputation of low frequency and rare variants with increasing accuracy. Many association studies are now being conducted based on imputation in existing large genome-wide association studies. A limitation of our method is that it currently cannot account for imputation accuracy, which is expected to reduce with decreasing allele frequency. At the level of individual variants, it is possible to characterize reduction of power based on formula for effect-size attenuation due to imputation[45]. Further studies are needed to understand impact of imputation on aggregated tests encompassing variants of different allele frequency spectra. In this report, we have illustrated application of the framework in exome-based analysis where aggregated tests can be applied across largely non-overlapping genes. For whole genome sequencing studies, where aggregated test may be applied in a sliding window fashion[10; 14], more work is needed for genome-wide power calculations in terms of underlying models for genetic architecture.

In conclusion, in this report we provide simple analytic approaches to power calculations for rare variants association tests at the individual locus level and at the whole genome level in terms of a few key parameters of underlying models for genetic architecture. These methods together, which we implement in a Shiny application in R, will provide useful design tools for planning next generation genome-wide association studies.

Bibliography

1. Wood, A.R., Esko, T., Yang, J., Vedantam, S., Pers, T.H., Gustafsson, S., Chu, A.Y., Estrada, K., Luan, J., Kutalik, Z., et al. (2014). Defining the role of common variation in the genomic and biological architecture of adult human height. Nat Genet 46, 1173-1186.
2. Locke, A.E., Kahali, B., Berndt, S.I., Justice, A.E., Pers, T.H., Day, F.R., Powell, C., Vedantam, S., Buchkovich, M.L., Yang, J., et al. (2015). Genetic studies of body mass index yield new insights for obesity biology. Nature 518, 197-206.
3. Sampson, J.N., Wheeler, W.A., Yeager, M., Panagiotou, O., Wang, Z., Berndt, S.I., Lan, Q., Abnet, C.C., Amundadottir, L.T., Figueroa, J.D., et al. (2015). Analysis of Heritability and Shared Heritability Based on Genome-Wide Association Studies for Thirteen Cancer Types. J Natl Cancer Inst 107, djv279.
4. Huang, J., Howie, B., McCarthy, S., Memari, Y., Walter, K., Min, J.L., Danecek, P., Malerba, G., Trabetti, E., Zheng, H.F., et al. (2015). Improved imputation of low-frequency and rare variants using the UK10K haplotype reference panel. Nat Commun 6, 8111.
5. Kreiner-Moller, E., Medina-Gomez, C., Uitterlinden, A.G., Rivadeneira, F., and Estrada, K. (2015). Improving accuracy of rare variant imputation with a two-step imputation approach. Eur J Hum Genet 23, 395-400.
6. Davies, R.W., Flint, J., Myers, S., and Mott, R. (2016). Rapid genotype imputation from sequence without reference panels. Nat Genet 48, 965-969.
7. Purcell, S.M., Moran, J.L., Fromer, M., Ruderfer, D., Solovieff, N., Roussos, P., O'Dushlaine, C., Chambert, K., Bergen, S.E., Kahler, A., et al. (2014). A polygenic burden of rare disruptive mutations in schizophrenia. Nature 506, 185-+.
8. Tang, H.Y., Jin, X., Li, Y., Jiang, H., Tang, X.F., Yang, X., Cheng, H., Qiu, Y., Chen, G., Mei, J.P., et al. (2014). A large-scale screen for coding variants predisposing to psoriasis. Nature Genetics 46, 45-+.
9. Cheng, C.Y., Yamashiro, K., Chen, L.J., Ahn, J., Huang, L., Huang, L., Cheung, C.M., Miyake, M., Cackett, P.D., Yeo, I.Y., et al. (2015). New loci and coding variants confer risk for age-related macular degeneration in East Asians. Nat Commun 6, 6063.
10. Consortium, U.K., Walter, K., Min, J.L., Huang, J., Crooks, L., Memari, Y., McCarthy, S., Perry, J.R., Xu, C., Futema, M., et al. (2015). The UK10K project identifies rare variants in health and disease. Nature 526, 82-90.
11. Huang, L.Z., Li, Y.J., Xie, X.F., Zhang, J.J., Cheng, C.Y., Yamashiro, K., Chen, L.J., Ma, X.Y., Cheung, C.M.G., Wang, Y.S., et al. (2015). Whole-exome sequencing implicates UBE3D in age-related macular degeneration in East Asian populations. Nature Communications 6.
12. Mahajan, A., Sim, X., Ng, H.J., Manning, A., Rivas, M.A., Highland, H.M., Locke, A.E., Grarup, N., Im, H.K., Cingolani, P., et al. (2015). Identification and functional characterization of G6PC2 coding variants influencing glycemic traits define an effector transcript at the G6PC2-ABCB11 locus. PLoS Genet 11, e1004876.
13. Xu, H., Zhang, H., Yang, W.J., Yadav, R., Morrison, A.C., Qian, M.X., Devidas, M., Liu, Y., Perez-Andreu, V., Zhao, X.J., et al. (2015). Inherited coding variants at the CDKN2A locus influence susceptibility to acute lymphoblastic leukaemia in children. Nature Communications 6.
14. Zheng, H.F., Forgetta, V., Hsu, Y.H., Estrada, K., Rosello-Diez, A., Leo, P.J., Dahia, C.L., Park-Min, K.H., Tobias, J.H., Kooperberg, C., et al. (2015). Whole-genome sequencing identifies EN1 as a determinant of bone density and fracture. Nature 526, 112-117.
15. Fuchsberger, C., Flannick, J., Teslovich, T.M., Mahajan, A., Agarwala, V., Gaulton, K.J., Ma, C., Fontanillas, P., Moutsianas, L., McCarthy, D.J., et al. (2016). The genetic architecture of type 2 diabetes. Nature 536, 41-47.
16. Haddad, S.A., Ruiz-Narvaez, E.A., Haiman, C.A., Sucheston-Campbell, L.E., Bensen, J.T., Zhu, Q., Liu, S., Yao, S., Bandera, E.V., Rosenberg, L., et al. (2016). An exome-wide analysis of low frequency

and rare variants in relation to risk of breast cancer in African American Women: the AMBER Consortium. Carcinogenesis 37, 870-877.

17. Luo, Y., de Lange, K.M., Jostins, L., Moutsianas, L., Randall, J., Kennedy, N.A., Lamb, C.A., McCarthy, S., Ahmad, T., Edwards, C., et al. (2016). Exploring the genetic architecture of inflammatory bowel disease by whole genome sequencing identifies association at ADCY7. bioRxiv.
18. Ganna, A., Genovese, G., Howrigan, D.P., Byrnes, A., Kurki, M.I., Zekavat, S.M., Whelan, C.W., Kals, M., Nivard, M.G., Bloemendal, A., et al. (2016). Ultra-rare disruptive and damaging mutations influence educational attainment in the general population. Nat Neurosci 19, 1563-1565.
19. Wessel, J., Chu, A.Y., Willems, S.M., Wang, S., Yaghootkar, H., Brody, J.A., Dauriz, M., Hivert, M.F., Raghavan, S., Lipovich, L., et al. (2015). Low-frequency and rare exome chip variants associate with fasting glucose and type 2 diabetes susceptibility. Nat Commun 6, 5897.
20. Group, C.C.H.W. (2016). Meta-analysis of rare and common exome chip variants identifies S1PR4 and other loci influencing blood cell traits. Nat Genet 48, 867-876.
21. Liu, C., Kraja, A.T., Smith, J.A., Brody, J.A., Franceschini, N., Bis, J.C., Rice, K., Morrison, A.C., Lu, Y., Weiss, S., et al. (2016). Meta-analysis identifies common and rare variants influencing blood pressure and overlapping with metabolic trait loci. Nat Genet 48, 1162-1170.
22. Moutsianas, L., Agarwala, V., Fuchsberger, C., Flannick, J., Rivas, M.A., Gaulton, K.J., Albers, P.K., Go, T.D.C., McVean, G., Boehnke, M., et al. (2015). The power of gene-based rare variant methods to detect disease-associated variation and test hypotheses about complex disease. PLoS Genet 11, e1005165.
23. Zuk, O., Schaffner, S.F., Samocha, K., Do, R., Hechter, E., Kathiresan, S., Daly, M.J., Neale, B.M., Sunyaev, S.R., and Lander, E.S. (2014). Searching for missing heritability: designing rare variant association studies. Proc Natl Acad Sci U S A 111, E455-464.
24. Park, J.H., Gail, M.H., Weinberg, C.R., Carroll, R.J., Chung, C.C., Wang, Z., Chanock, S.J., Fraumeni, J.F., Jr., and Chatterjee, N. (2011). Distribution of allele frequencies and effect sizes and their interrelationships for common genetic susceptibility variants. Proc Natl Acad Sci U S A 108, 18026-18031.
25. Li, B., and Leal, S.M. (2008). Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. Am J Hum Genet 83, 311-321.
26. Madsen, B.E., and Browning, S.R. (2009). A groupwise association test for rare mutations using a weighted sum statistic. PLoS Genet 5, e1000384.
27. Wu, M.C., 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.
28. Lee, S., Wu, M.C., and Lin, X. (2012). Optimal tests for rare variant effects in sequencing association studies. Biostatistics 13, 762-775.
29. Neale, B.M., Rivas, M.A., Voight, B.F., Altshuler, D., Devlin, B., Orho-Melander, M., Kathiresan, S., Purcell, S.M., Roeder, K., and Daly, M.J. (2011). Testing for an unusual distribution of rare variants. PLoS Genet 7, e1001322.
30. Kosmicki, J.A., Churchhouse, C.L., Rivas, M.A., and Neale, B.M. (2016). Discovery of rare variants for complex phenotypes. Human Genetics 135, 625-634.
31. Richardson, T.G., Campbell, C., Timpson, N.J., and Gaunt, T.R. (2016). Incorporating Non-Coding Annotations into Rare Variant Analysis. PLoS One 11, e0154181.
32. Lek, M., Karczewski, K.J., Minikel, E.V., Samocha, K.E., Banks, E., Fennell, T., O'Donnell-Luria, A.H., Ware, J.S., Hill, A.J., Cummings, B.B., et al. (2016). Analysis of protein-coding genetic variation in 60,706 humans. Nature 536, 285-291.
33. Lin, D.Y., and Tang, Z.Z. (2011). A general framework for detecting disease associations with rare variants in sequencing studies. Am J Hum Genet 89, 354-367.

34. Morris, A.P., and Zeggini, E. (2010). An evaluation of statistical approaches to rare variant analysis in genetic association studies. Genet Epidemiol 34, 188-193.
35. Price, A.L., Kryukov, G.V., de Bakker, P.I., Purcell, S.M., Staples, J., Wei, L.J., and Sunyaev, S.R. (2010). Pooled association tests for rare variants in exon-resequencing studies. Am J Hum Genet 86, 832-838.
36. Derkach, A., Lawless, J.F., and Sun, L. (2013). Robust and powerful tests for rare variants using Fisher's method to combine evidence of association from two or more complementary tests. Genet Epidemiol 37, 110-121.
37. Derkach, A., Lawless, J.F., and Sun, L. (2014). Pooled Association Tests for Rare Genetic Variants: A Review and Some New Results. Stat Sci 29, 302-321.
38. Park, J.H., Wacholder, S., Gail, M.H., Peters, U., Jacobs, K.B., Chanock, S.J., and Chatterjee, N. (2010). Estimation of effect size distribution from genome-wide association studies and implications for future discoveries. Nat Genet 42, 570-575.
39. Liu, H., Tang, Y.Q., and Zhang, H.H. (2009). A new chi-square approximation to the distribution of non-negative definite quadratic forms in non-central normal variables. Comput Stat Data An 53, 853-856.
40. Wu, B., and Pankow, J.S. (2016). On Sample Size and Power Calculation for Variant Set-Based Association Tests. Ann Hum Genet 80, 136-143.
41. Wang, G.T., Li, B., Santos-Cortez, R.P., Peng, B., and Leal, S.M. (2014). Power analysis and sample size estimation for sequence-based association studies. Bioinformatics 30, 2377-2378.
42. Speed, D., Cai, N., Johnson, M., Nejentsev, S., and Balding, D. (2016). Re-evaluation of SNP heritability in complex human traits. bioRxiv.
43. Mancuso, N., Rohland, N., Rand, K.A., Tandon, A., Allen, A., Quinque, D., Mallick, S., Li, H., Stram, A., Sheng, X., et al. (2016). The contribution of rare variation to prostate cancer heritability. Nat Genet 48, 30-35.
44. Agarwala, V., Flannick, J., Sunyaev, S., Go, T.D.C., and Altshuler, D. (2013). Evaluating empirical bounds on complex disease genetic architecture. Nat Genet 45, 1418-1427.
45. Huang, L., Wang, C., and Rosenberg, N.A. (2009). The relationship between imputation error and statistical power in genetic association studies in diverse populations. Am J Hum Genet 85, 692-698.

Table 1. ***The first-order approximations of c_k as the function of phenotypic variation explained by a locus under three genetic models.***

Underlying Genetic Architecture	Mathematical Representation	First-order Approximation	Second-order Approximation
Proportion of variation explained by a variant EV_j is independent of its MAF p_j	$\beta_j \sim 1/\sqrt{p_j(1-p_j)}$	$c_k \approx \sum_{j=1}^{J} \lambda_j^k \,(1+\frac{kNEV}{J})$	$c_k \approx \sum_{j=1}^{J} \lambda_j^k + kNEV \sum_{j=1}^{J_C} \frac{\lambda_j^k}{J_C}$
Genetic effect β_j is independent of MAF p_j	$EV_j \sim \sqrt{p_j(1-p_j)}$	$c_k \approx \sum_{j=1}^{J} \lambda_j^k (1 + kNEV \frac{\sum_{j=1}^{J} \lambda_j^k p_j}{\sum_{j=1}^{J} p_j \sum_{j=1}^{J} \lambda_j^k})$	$c_k \approx \sum_{j=1}^{J} \lambda_j^k + kNEV \frac{\sum_{j=1}^{J_C} \lambda_j^k p_j}{\sum_{j=1}^{J_C} p_j}$
Genetic effect β_j is function of MAF p_j	$\beta_j \sim \lvert log_{10}(p_j)\rvert$	$c_k \approx \sum_{j=1}^{J} \lambda_j^k + kNEV \frac{\sum_{j=1}^{J} \lambda_j^k}{J} + k\nu(\sum_{j=1}^{J} \lambda_j^k p_j - \frac{\sum_{j=1}^{J} \lambda_j^k \sum_{j=1}^{J} p_j}{J})$, where ν is an average change in EV_j due to one-unit change in p_j.	$c_k \approx \sum_{j=1}^{J} \lambda_j^k + kNEV \frac{\sum_{j=1}^{J_C} \lambda_j^k}{J_C} + k\nu(\sum_{j=1}^{J_C} \lambda_j^k p_j - \frac{\sum_{j=1}^{J_C} \lambda_j^k \sum_{j=1}^{J_C} p_j}{J_C})$, where ν is an average change in EV_j due to one-unit change in p_j.

Table 2: ***Parameters and parameter values for assessing accuracy of the first-order approximations.***
Scenario S1 ('MAF-independent EV') assumes MAFs and EVs are mutually independent. Scenario S2 ("MAF-independent β_j") assumes MAFs and effect sizes are mutually independent. Scenario S3 ("MAF-log-dependent β_j") assumes MAFs and effect sizes are dependent through log10 function. For the exact calculations for the Scenario S1, we directly generate EV_js from specific value EV and for Scenarios S2 and S3, we first generate β_js then calculate corresponding EV_js.

	Parameters	Parameter Values	Parameters used in		
			First-order Approximation	Second-order Approximation	Exact-Calculations
N	Effective sample size	10,000	*Yes*	*Yes*	*Yes*
J	Total number of SNPs	50, 100, 200, 400	*Yes*	*Yes*	*Yes*
EV	Coefficient of explained phenotypic variation by a locus	Ranges between 0.001 and 0.01	*Yes*	*Yes*	*Yes*
p_j	MAF of SNP j	Gamma (1,300) with minimum and maximum values at 0.0002 and 0.01	*Yes*	*Yes*	*Yes*
J_c	Number of causal SNPs	10, 20, 30, 50	*No*	*Yes*	*Yes*
	Scenario S1 ("MAF-independent EV")				
EV_j	Coefficient of explained variation by SNP j	Randomly selected for each causal SNP under the constrain: $\sum_{j=1}^{J_c} EV_j = EV$	*No*	*No*	*Yes*
	Scenario S2 ("MAF-independent β_j")				
β	MAF adjusted average effect of SNP j	MAF adjusted average effect of rare variant	*No*	*No*	*Yes*
β_j	Genetic effect of j^{th} variant	$\beta_j^2 \sim N\left(\beta^2, \left(\beta^2/2\right)^2\right)$, then coefficients of explained variations are scaled by the constant so that $\sum_{j=1}^{J_c} EV_j = EV$	*No*	*No*	*Yes*
	Scenario S3 ("MAF-log-dependent β_j")				
C	Adjustment	$C = \frac{EV}{J_c E\left(2p_j(1-p_j)\log_{10}(p_j)^2\right)}$	*No*	*No*	*Yes*
β_j	Genetic effect of j^{th} variant	$\beta_j = C\log_{10}(p_j)$, then coefficients of explained variations are scaled by the constant so that $\sum_{j=1}^{J_c} EV_j = EV$	*No*	*No*	*Yes*

Table 2. ***Summary of recently published association studies with rare variants.***

Study	Genetic Platform Sample Size	Trait	Association analysis with rare variants[1]		
			Gene based tests	# of significant findings with gene based test	# of significant rare variant findings with single variant test
A polygenic burden of rare disruptive mutations in schizophrenia[7]	WES: 5000	Case/Control ~ 1/1	SKAT and Burden Tests with a gene as a unit	No findings	NA
Whole-genome sequencing identifies EN1 as determinant of bone density and fracture[14]	WGS: ~2,900 WES: ~3,500 Imputation[2]: ~26,500	Multiple QTs	SKAT with sliding window with 30 SNP	0-1 gene[3]	0-1 rare variant
The UK10K project identifies rare variants in health and disease[10]	WGS: ~3,500 Imputation: ~9,200	Multiple QTs	SKAT with sliding window with 50 SNP	0-1 genes	0-1 rare variants
The genetic architecture of type 2 diabetes[15]	WGS: ~2,600 WES: ~13,000 Exome Array: 80,000	Case/Control ~ 1/1 Case/Control ~ 1/1 Case/Control ~ 1/2	SKAT with a gene as a unit	No findings	6 rare variants[4]
Ultra-rare disruptive and damaging mutations influence educational attainment in general population[18]	WGS: ~2,700 WES: ~11,300	QT	Burden Tests with a gene as a unit	No findings	NA
Inherited coding variants at the CDKN2A locus influence susceptibility to acute lymphoblastic leukemia in children[13]	Exome Chip: ~12,000	Case/Control ~ 1/5	SKAT with a gene as a unit	No findings	1 rare variant
Meta-analysis of rare and common exome chip variants identifies *S1PR4* and other loci influencing blood cell traits[20]	Exome Chip: ~52,000	Multiple QTs	SKAT and Burden Tests with a gene as a unit	1-2 genes	1-3 rare variants
Low-frequency and rare exome chip variants associate with fasting glucose and type 2 diabetes susceptibility[19]	Exome Chip: ~61,000	QT	SKAT with a gene as a unit	1 gene	No findings
Meta-analysis identifies common and rare variants influencing blood pressure and overlapping with metabolic trait loci[21]	Exome Chip: ~145,000	Multiple QT	SKAT and Burden Tests with a gene as a unit	1-2 genes	1-2 rare variants

[1] Variants with MAF<1%.
[2] Variants with MAF>0.1%
[3] Identified by single SNP analysis
[4] Variants with MAF<5%

Figure 1: ***Evaluation of the accuracy of the first order approximation under simulation scenario S1 (MAF-independent EV).*** *Exact Formula represents estimated average power using exact theoretical formulas for the SKAT test statistic. The First Order Approximation represents estimated average power using the first order approximation for the SKAT test statistic.*

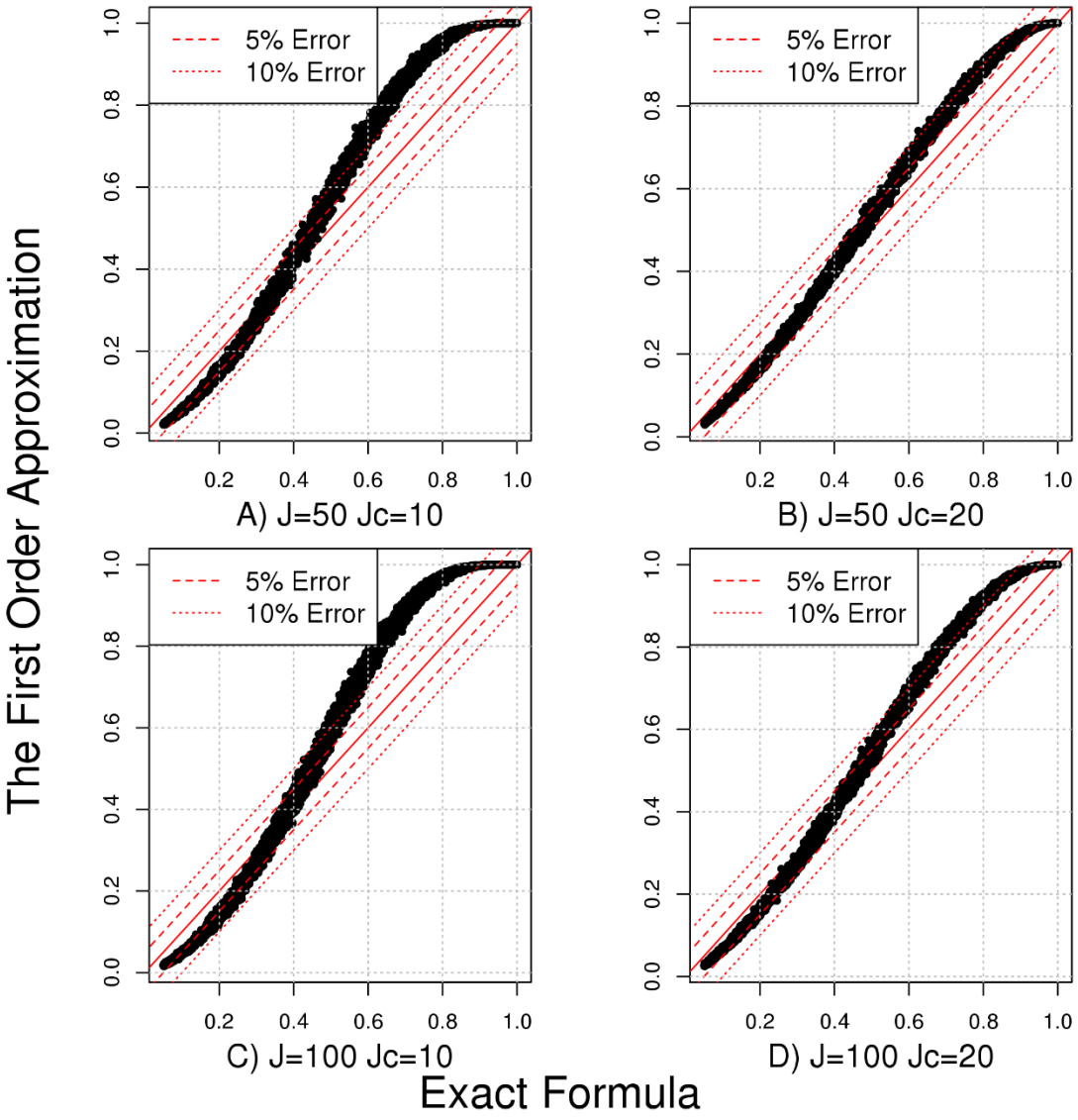

Figure 2: ***Evaluation of the accuracy of the second order approximation under simulation scenario S1 (MAF-independent EV).*** *Exact Formula represents estimated average power using exact theoretical formulas for the SKAT test statistic. The Second Order Approximation represents estimated average power using the second order approximation for the SKAT test statistic.*

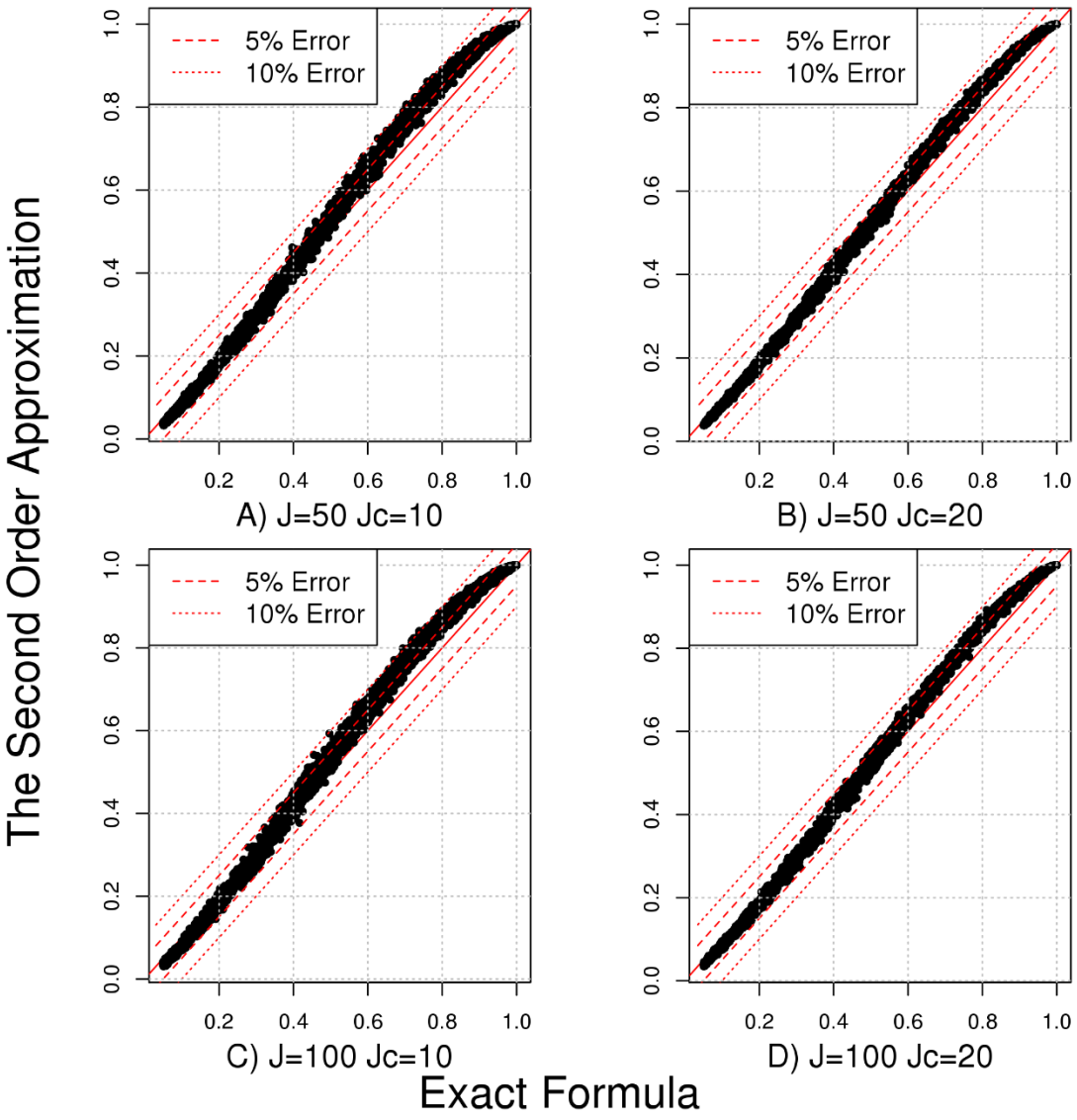

Figure 3: ***Maximum probability of observing no discoveries in the study on whole exome sequencing platform as function of the number of underlying causal loci K and the total variation explained by them, Total EV.*** *Effective sample size is set to 15,000 and level of the test is $2.5 \cdot 10^{-6}$. Probabilities are estimated by (2) and assumption of independence between MAF and EV. We provide approximate contours (bounds) for probability of observing no discoveries at 5%.*

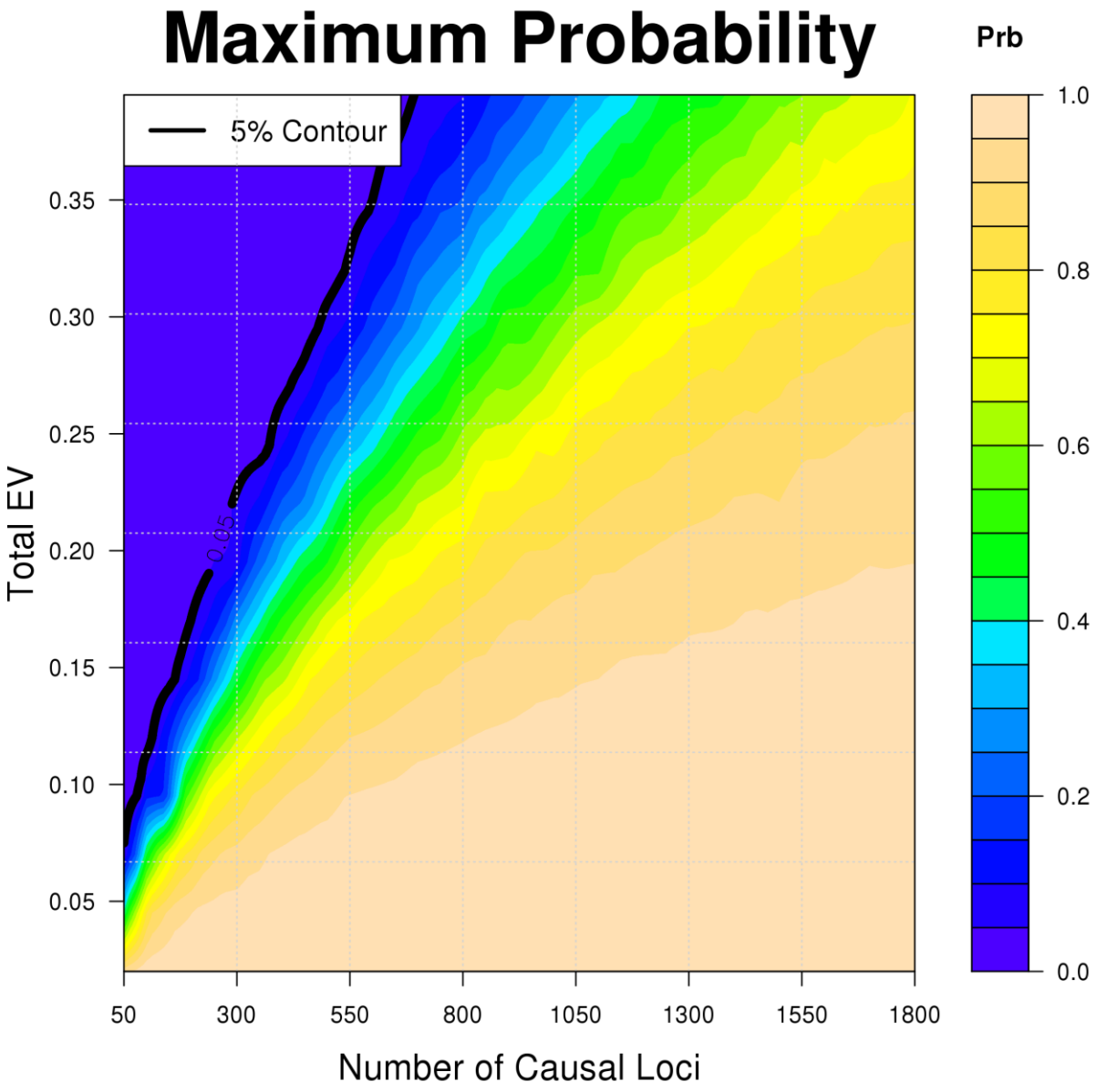

Figure 4: ***Maximum probability of observing three statistical significant discoveries in the Exome Chip study as function of the number of underlying causal loci K and the total variation explained by them, Total EV.*** *Effective sample size is set to 140,000 and level of the test is* $2.5 \cdot 10^{-6}$*. Probabilities are estimated by (2) and assumption of independence between MAF and EV. We provide approximate contours (bounds) for probability of observing no discoveries at 5%.*

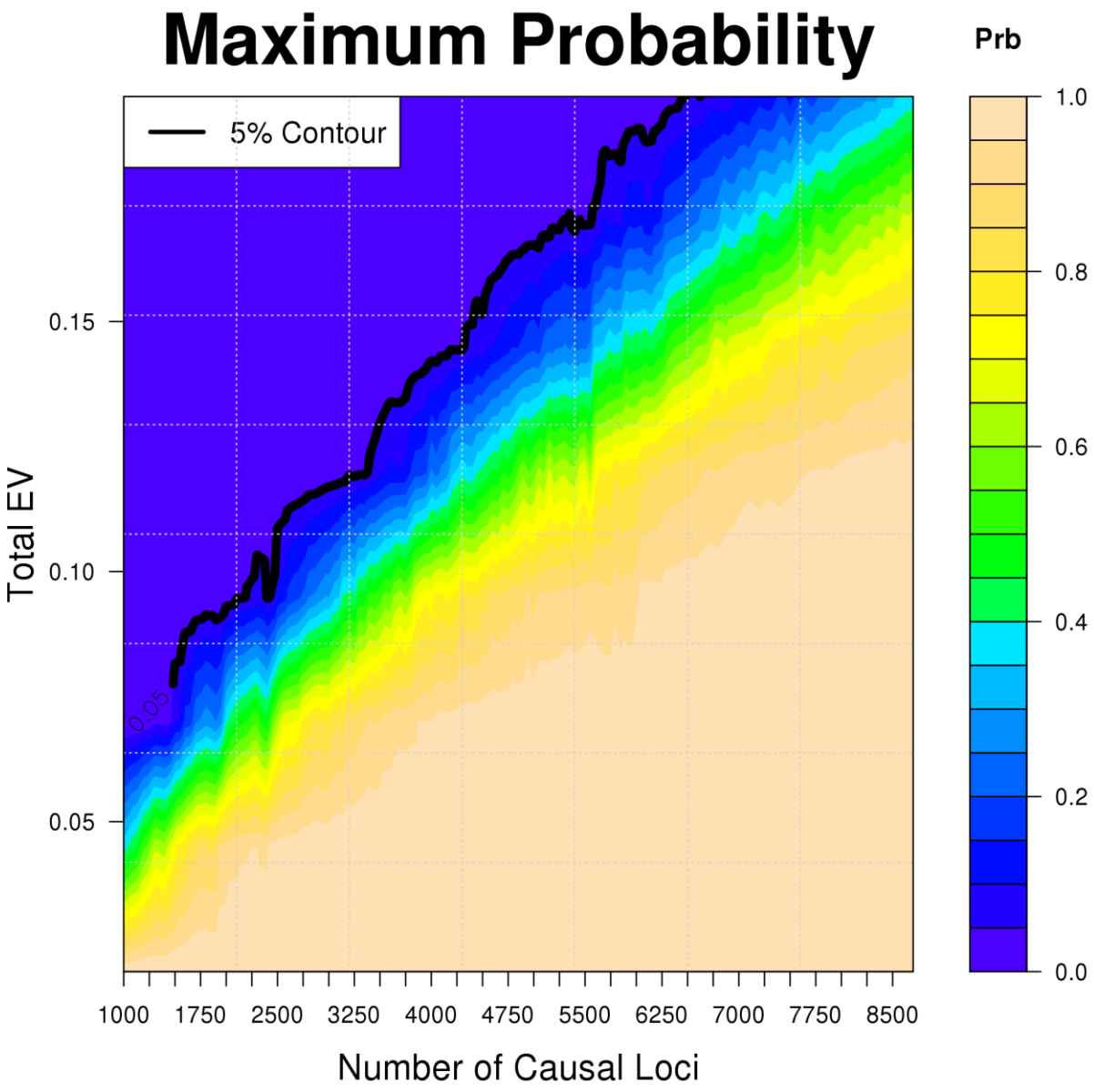

Figure 5: ***Effects of sensitivity and specificity on the power of variance component and burden tests under simulation scenario S1 (MAF-independent EV).*** Initial power for variance component test is set to ***40%,*** number of variance in a locus is set to ***J=50*** and number of causal variants to J_C***=10.*** Initial power of burden test statistic as result corresponds to ***0.36.***