Bioinformatics, 34(9), 2018, 1506–1513
doi: 10.1093/bioinformatics/btx770
Advance Access Publication Date: 28 November 2017
Original Paper

Genetics and population analysis

Power Analysis for Genetic Association Test (PAGEANT) provides insights to challenges for rare variant association studies

Andriy Derkach[1], **Haoyu Zhang**[2] **and Nilanjan Chatterjee**[2,3,*]

[1]Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Rockville, MD 20850, USA, [2]Department of Biostatistics, Bloomberg School of Public Health, School of Medicine, Johns Hopkins University, Baltimore, MD 21205, USA and [3]Department of Oncology, School of Medicine, Johns Hopkins University, Baltimore, MD 21205, USA

*To whom correspondence should be addressed.

Associate Editor: Oliver Stegle

Received on June 7, 2017; revised on October 2, 2017; editorial decision on November 24, 2017; accepted on November 27, 2017

Abstract

Motivation: Genome-wide association studies are now shifting focus from analysis of common to rare variants. 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. Typically, power calculations for such tests require specification of large number of parameters, including effect sizes and allele frequencies of individual variants, making them difficult to use in practice. We propose to approximate power to a varying degree of accuracy using a smaller number of key parameters, including the total genetic variance explained by multiple variants within a locus.

Results: We perform extensive simulation studies to assess the accuracy of the proposed approximations in realistic settings. Using these simplified power calculations, we develop an analytic framework to obtain bounds on genetic architecture of an underlying trait given results from genome-wide association studies with rare variants. Finally, we provide insights into the required quality of annotation/functional information for identification of likely causal variants to make meaningful improvement in power.

Availability and implementation: A shiny application that allows a variety of Power Analysis of GEnetic AssociatioN Tests (PAGEANT), in R is made publicly available at https://andrewhaoyu.shinyapps.io/PAGEANT/.

Contact: nilanjan@jhu.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Over the last decade, genome-wide association studies (GWAS) of common variants of increasingly large sample sizes 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 (Park *et al.*, 2010; Yang *et al.*, 2010; Lee *et al.*, 2011, 2012; Stahl *et al.*, 2012; Visscher *et al.*, 2012; Dudbridge 2013; Wood *et al.*, 2014; Bulik-Sullivan *et al.*, 2015; Locke *et al.*, 2015; Sampson *et al.*, 2015), 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 (Huang *et al.*, 2015; Kreiner-Moller *et al.*, 2015; Davies *et al.*, 2016). However, limited or lack of findings from early studies (Purcell *et al.*, 2014; Tang *et al.*, 2014; Cheng *et al.*, 2015; UK10K Consortium *et al.*, 2015; Huang *et al.*, 2015; Mahajan *et al.*, 2015; Wessel *et al.*, 2015; Xu *et al.*, 2015; Zheng *et al.*, 2015; Fuchsberger *et al.*, 2016; Ganna *et al.*, 2016; CHARGE Consortium Hematology Working Group, 2016; Haddad *et al.*, 2016; Liu *et al.*, 2016; Luo *et al.*, 2017)

Published by Oxford University Press 2017. This work is written by US Government employees and is in the public domain in the US.

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 (Park *et al.*, 2011; Zuk *et al.*, 2014; Moutsianas *et al.*, 2015).

Testing of associations 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 (Li and Leal, 2008; Madsen and Browning, 2009; Liu and Leal, 2010; Morris and Zegginі, 2010; Price *et al.*, 2010; Ionita-Laza *et al.*, 2011; Lin and Tang, 2011; Neale *et al.*, 2011; Wu *et al.*, 2011; Lee *et al.*, 2012; Derkach *et al.*, 2013; Sun *et al.*, 2013). Simulation studies have been used under various anticipated genetic architectures of the traits for the demonstration of potential power of these procedures (Neale *et al.*, 2011; Wu *et al.*, 2011; Lee *et al.*, 2012; Moutsianas *et al.*, 2015). In particular, analysis of power for variance component-based tests, such as the popular SKAT method, can be complex as they require specification of many different parameters including a 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 the power of rare variant association tests (Kosmicki *et al.*, 2016; Richardson *et al.*, 2016). To the best of our knowledge, however, there has been no systematic study of the effect of the use of such extraneous information on power of the association tests.

In this report, we first describe approximations that allow analytic characterizations of power for popular aggregate-level association tests based on a few key parameters, thus dramatically reducing the complexity of power calculations. We perform simulation studies using allele frequency distribution observed in Exome Aggregation Consortium (ExAC) (Lek *et al.*, 2016) under various models for effect size distributions to assess the accuracy of the proposed approximations in realistic settings. We then develop a framework for genome-wide power calculations based on the underlying genetic architecture of a trait characterized by a number of underlying causal loci and total variability they explain. 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 lack of discoveries on bounds of genetic architecture of the underlying traits.

We also use the proposed framework to characterize power of association tests that may preselect variants based on prior functional/annotation information. These derivations provide important insights into the required quality of annotation/functional information for identification of likely causal variants to make meaningful improvement in the power of subsequent association tests. Finally, to facilitate convenient and rapid power calculations for rare variant association tests, we make a shiny app PAGEANT (Power Analysis for GEnetic AssociatioN Test) available in R.

2 Materials and methods

2.1 Existing power calculations

A variety of statistics have been proposed for testing genetic associations at the levels of genetic loci or regions by aggregating association statistics over multiple genetic variants (Li and Leal, 2008; Madsen and Browning, 2009; Liu and Leal, 2010; Price *et al.*, 2010; Lin and Tang, 2011; Neale *et al.*, 2011; Wu *et al.*, 2011; Lee *et al.*, 2012; Derkach *et al.*, 2013). Multiple studies (Lin and Tang, 2011; Derkach *et al.*, 2014; Moutsianas *et al.*, 2015) have shown that existing methods can be classified as sum-based (Li and Leal, 2008; Madsen and Browning, 2009; Morris and Zeggini, 2010; Price *et al.*, 2010), variance component (Neale *et al.*, 2011; Wu *et al.*, 2011; Derkach *et al.*, 2014), and hybrid tests that are functions of both classes (Lin and Tang, 2011; Lee *et al.*, 2012; Derkach *et al.*, 2013). Here, we focus on sum-based and variance component tests. We do not consider hybrid tests because their power is usually close to one of the two components. Sum-based tests aggregate variant-level association statistics by a linear combination in the forms

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

and variance component tests aggregate by quadratic combination in the forms

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

where w_js are weights that depend on minor allele frequency (MAF) and T_js are score statistics for associations for individual SNPs $(j = 1, \ldots, J)$, the latter of which are typically derived from a regression model, such as the linear or logistic regression (Lin and Tang, 2011; Lee *et al.*, 2012; Derkach *et al.*, 2014).

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 (Derkach *et al.*, 2014). The analytic power of a single-variant statistic

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

can be derived based on one degree-of-freedom chi-square distribution with a non-centrality parameter of the form $nc_j = 2p_j(1-p_j)\beta_j{}^2 N = EV_j N$, which depends on two parameters: effective sample size (N) and proportion of phenotypic variation explained by the j^{th} variant $[EV_j = 2p_j(1-p_j)\beta_j{}^2]$, a function of MAF (p_j) and genetic effect (β_j), measured in the unit of per copy of an allele (Park *et al.*, 2010).

Under assumption of low LD between rare variants and high probabilities of observing the variant in a sample Derkach *et al.* (2014) derived analytical power formulas for rare variant association tests. They showed that 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 the non-centrality parameter

$$nc_{ST} = N \frac{\left(\sum_{j=1}^{J} w_j sign\left(\beta_j\right)\sqrt{p_j(1-p_j)EV_j}\right)^2}{\sum\limits_{j=1}^{J} w_j^2 p_j(1-p_j)},$$

which depends on coefficients of explained variations associated with individual variants. Previous studies (Ionita-Laza *et al.*, 2011; Derkach *et al.*, 2014) 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 $nc_j = EV_j N$ and weights $\lambda_j = w_j p_j(1-p_j)N$. It has been suggested that analytic power calculations for variance component tests be done by approximating the asymptotic distribution of T_{VC} by a single non-central chi-square distribution matched up to four cumulants (Wu and Pankow, 2016). There are also several modifications of this method matching higher moments to improve the tail probability approximation (Wu *et al.*, 2011; Wu and Pankow, 2016); however, power differences seem to be marginal. The cumulants c_k of the test statistic T_{VC} can be written as

$$c_k = \sum_{j=1}^{J} \lambda_j^k + kN \sum_{j=1}^{J} \lambda_j^k EV_j \ \ for \ k = 1, \ldots, \ 4, \tag{1}$$

which require specification of effect sizes (EV_j) and allele frequencies of individual variants. The power calculations for aggregate-level tests requiring specification of MAFs and genetic effects for individual variants have been implemented in several statistical packages (Wu *et al.*, 2011; Wang *et al.*, 2014; Wu and Pankow, 2016).

2.2 Approximate power calculations for aggregate tests

In the 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 the sum-based test, under an assumption of independence between coefficients of explained variations and MAFs [e.g. genetic effect of a variant β_j is inversely proportional to $\sqrt{2p_j(1-p_j)}$], we can roughly estimate non-centrality parameter as

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

where J_D and J_P are numbers of deleterious and protective variants in a locus and $\sum_{j=1}^{J} EV_j = TEV$ is the total proportion of variation explained by all J variants in the locus (see Supplementary Appendix S1). If all of the causal variants in a locus are deleterious (or protective), then the non-centrality parameter can be characterized by TEV and the proportion of causal variants. In Supplementary Appendix S1, we state approximations for other two common relationships between genetic effects and MAFs.

Next, we consider simplified power calculations for variance component tests T_{VC} by approximating cumulants c_k in (1) as a function of the total proportion of variance explained by all variants within a locus (TEV). For example, if we assume independence between MAF and proportion of variation explained across individual variants, we can obtain a first-order approximation in the form

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

In Supplementary Appendix S2, we derive approximations of c_k for three commonly assumed relationships between genetic effects and MAFs and summarize them in Supplementary Table S1. The first-order approximations, which implicitly treat all variants in a locus to be causal, can be inaccurate when the number of true causal variants is small. To improve accuracy, we propose second-order approximations to estimate the sum $\sum_{j=1}^{J} \lambda_j^k EV_j$ in (1) as function of TEV and number of underlying casual variants (J_C) (see Supplementary Appendix S2 and Table S1). For example, if we assume the same hypothesis of independence between proportion of variation explained and MAF across individual variants, then we approximate $\sum_{j=1}^{J} \lambda_j^k EV_j$ in c_k as $\frac{1}{J_C}\sum_{j=1}^{J_C} \lambda_j^k$.

2.3 Genome-wide power calculations and bounds on genetic architecture

Using the proposed first-order power calculation framework, we further develop a mathematical framework to study bounds on genetic architecture of underlying traits from limited findings reported in a GWAS. We first characterize the probability of a number of discoveries in a given study as a function of sample size N, the number of underlying causal loci K and the distribution of their effect sizes, $TEVs$. Let $P(TEV, J, p)$ be a power of a test to detect a locus explaining TEV of phenotypic variation. Then in Supplementary Appendix S3, we show that probability of M discoveries in a study of sample size N is

$$\begin{aligned} &P(\text{M Discoveries} \,|\, \text{Genetic Model}) \approx \\ &\qquad \approx \binom{K}{M} E[P(TEV, J, p)]^M e^{(K-M)E\{\log[1-P(TEV,J,p)]\}}, \end{aligned} \tag{2}$$

which depends on average power of the underlying association tests over the different causal loci in the genome and can be calculated by specifying number of underlying causal loci and distributions of proportion of phenotypic variations they explain (TEV), number of SNPs within a locus (J) and MAF p across the causal loci (see Supplementary Appendix S3). This formula can be extended using the second-order approximations by specifying distribution for a number of causal variants per locus. Because of limited information about this parameter, we do not focus on it.

Now, if $M = m$ is the number of discoveries reported based on a given GWAS of sample size N, we can calculate $P(M \leq m)$ using the above formula based on empirical distributions for MAFs (p_k), size of genes (J_k) observed in real data and various hypothesized values for number of causal loci (K) and parameters for underlying effect size distributions for TEV. Specifically, we generated a class of L-shaped effect size distribution using a two-parameter gamma distribution: Gamma(α, γ) with $\alpha \leq 1$ and restriction that the total variance of a trait explained by causal loci, GEV is smaller than 50%. Under this model, GEV is given by $K\mu$ where $\mu \approx \alpha\gamma$. For various combination of K and μ, we evaluate the maximum value of $P(\mathrm{M} \leq \mathrm{m})$ over different values of the dispersion parameter ($\alpha/\gamma = \alpha^2/\mu$) determined from wide range of possible values of α.

When this probability is low (e.g. <5%), we conclude that the underlying model for genetic architecture is unlikely. For example, many recent studies have reported no discoveries based on gene-level association tests. In these studies, the probability of no discoveries, $m = 0$, can be used to provide bound on genetic architecture of the underlying trait.

2.4 Effects of filtering variants by extraneous information

We use the proposed framework to 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 a function of sensitivity and specificity of the underlying filtering method. Sensitivity (Se) is the probability of selecting a variant given that it is truly causal, while specificity (Sp) is the probability of filtering a variant out given that is non-causal. If selection/filtering is independent of MAFs and proportions of variations explained, then the number of remaining variants after filtering in a locus is $J_S = Se \cdot J_C + (1 - Sp) \cdot (J - J_C)$ and the proportion of variation explained by them is $TEV_S = Se \cdot TEV$. Now, with new values of J and TEV, we estimate power for aggregated tests and compare them to corresponding base values if no filtering was applied (e.g. $Se = 100\%$ but specificity is $Sp = 0\%$).

If only a small subset of variants is selected, then sensitivity may be reduced as some true causal variants could be missed while specificity may improve because of removal of non-causal variants. If one takes a random subset of the variants, then $Se = 1 - Sp$ as the casual and non-causal variants 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 framework, we explore the power of aggregated tests for various combinations of sensitivity and specificity of the underlying filtering algorithm. Furthermore, we assume that the functional/annotation information for the variants is measured by underlying normally distributed continuous score and variants are included in association test if their score is above a specified threshold. For example, several functional/annotation tools (Wang *et al.*, 2010; Grant *et al.*, 2011; Asmann *et al.*, 2012) summarize multiple sources into one continuous score. Under these assumptions, we evaluate receiver operating characteristic curves generated by the combination of sensitivity and specificity at different thresholds for SNP selection. We track power of different methods along different combinations of sensitivity and specificity parameters that lead to specific values 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 variants (see Fig. 3).

2.5 Empirical investigations

2.5.1 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 methods that require specification of effect sizes of individual variants. Here, we focus on the SKAT test statistics as a representative of variance component tests, and in Supplementary Material, we present results for the burden and simple sum test statistics (Li and Leal, 2008; Madsen and Browning, 2009) as a representatives of sum-based tests and C-alpha (Neale *et al.*, 2011) as other variance component test. For each fixed combination of the size of a region (J) and the total variance explained (TEV), 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 the power based on the first- and second-order approximations averaged over empirically derived distribution of allele frequencies is compared with power based on exact theoretical calculations averaged over distribution of allele frequencies and distribution of effect sizes for individual markers EV_j (see Fig. 1 and Supplementary Appendix S4 for more detail). Additional to evaluation of accuracy of the proposed power calculations, in Supplementary Material, we evaluate accuracy of estimation of sample size required to achieve 80% of power for variance component tests in comparison to exact theoretical method (see Supplementary Appendix S4). We consider three types of simulation scenarios: S1 ('MAF-independent EV') assumes that coefficients of explained variations (EV) is independent of MAF; S2 ('MAF-independent β_j') assumes that size of genetic effect (β), measured in the unit of per copy of an allele [$\beta^2 = EV/2MAF(1 - MAF)$] is independent of MAF and S3 ['MAF-log-dependent β_j'] assumes that genetic effect is related to MAF through $\log_{10}$ function (as defined in Supplementary Table S1). For each type of simulation scenario, we estimate the power for a locus of size $J = 50,\ 100,\ 200$ and 400 with the number of underlying causal variants $J_C = 10,\ 20,\ 30$ and 50. In Supplementary Appendix S4, we describe simulation mechanisms in detail and we summarize simulation models and parameters required for each method in Supplementary Table S2.

2.5.2 Bounds on variation explained by a causal locus

In Supplementary Table S3, we provide key parameters that summarize a variety of recently published association studies of rare variants (Purcell *et al.*, 2014; UK10K Consortium *et al.*, 2015; Wessel *et al.*, 2015; Xu *et al.*, 2015; Zheng *et al.*, 2015; Fuchsberger *et al.*, 2016; Ganna *et al.*, 2016; CHARGE Consortium Hematology Working Group, 2016; Liu *et al.*, 2016; Luo *et al.*, 2017). Typically, studies on Human Exome BeadChip (Exome Chip) had larger sample sizes than studies on sequencing platform; however, the latter covered a much smaller number of rare variants. We use our mathematical framework to obtain bounds on genetic architecture implied by the results from two of the largest studies, one of which studied educational attainment with exome sequencing and the other studied blood pressure employing exome chip (Ganna *et al.*, 2016; Liu *et al.*, 2016) (see 5th and 9th rows of the Supplementary Table S3).

For ease of illustration, we assume that analysis in these studies was conducted by the SKAT statistic with a gene as a unit, although the study of educational attainment (Ganna *et al.*, 2016) did not explicitly report results from gene-based analysis. We used the publicly available ExAC database (Lek *et al.*, 2016) to obtain empirical distributions for number of rare variants in a gene (J_k) and vector of MAFs p_k across exome. The average numbers of rare variants per gene (J) were 35.5 and 13 in the studies of educational attainment and blood pressure, which used exome sequencing and Exome Chip platforms, respectively. We provide empirical distributions and other key parameters in Supplementary Figures S11 and S12. Finally, we set Type 1 error threshold $T1 = \frac{0.05}{20,000} = 2.5 \times \cdot 10^{-6}$. To estimate genetic bounds from the results of study on educational attainment, we estimate the probability of no discoveries ($m = 0$) given a genetic model from (2) by using the sample size $N = 14,\ 000$, the number of sequenced individuals. 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).

To estimate genetic bounds from the study of blood pressure outcome, we calculate probability of at most three discoveries ($m = 3$) under a sample size of $N = 140,\ 000$ to match the number of discoveries reported and number of individuals genotyped in this study (see 9th row of Supplementary Table S3). In contrast to the previous study, here we did not put any lower bound for MAFs.

For each combination of the number underlying of causal loci (K) and parameters of effect size distributions, we calculate the probability of less than or equal to m discoveries using (2) with study-specific parameters. Expectations in (2) are estimated using 100 000 Monte Carlo simulations. In this report, we calculate power of the SKAT test statistic under the assumption of independence between proportion of variations explained and MAF of individual variants. Results for other genetic architecture are also discussed and presented in the Supplementary Material.

2.5.3 Effects of variant filtering on power of aggregated tests

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 values of TEV are selected so that power of the SKAT test is equal to 40% at Type 1 error $T1 = 0.05/20,\ 000 = 2.5 \cdot 10^{-6}$ and sample size $N = 10,\ 000$. For every combination of sensitivity and specificity, we estimate average power for the SKAT and burden tests. For this study, we also assume independence between proportion of variations explained and MAF for the individual variants. Results for other genetic architectures are presented in the Supplementary Material.

3 Results

3.1 Properties of the first- and second-order approximations

We evaluate the accuracy of first- and second-order approximations compared to exact power calculations of the variance component test under variety of genetic models (see Supplementary Table S2). The first-order approximations match exact calculations better as the number of causal variants in a locus J_C increases (Fig. 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 Fig. 1B and D and Supplementary Fig. S1). Similar conclusion holds also for very large loci $J = 200,\ 400$ and other relationships between genetic effect and MAF (see Supplementary Figs S2–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 Fig. 1A and C).

We also observe that the second-order approximation is more accurate at estimating the exact power (Fig. 1). Now, difference in power between approximate and exact calculations is small even when the number of causal variants in a locus is small, (see also Supplementary Figs S1–S4). Overall, our simulations demonstrate that the first-order approximation accurately estimates exact power when the number of causal variants in a locus is not too small. However, if the 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 an additional parameter, namely the number of causal variants in a locus (J_C). Very similar results hold for another type of variance component test, C-alpha (see Supplementary Fig. S5).

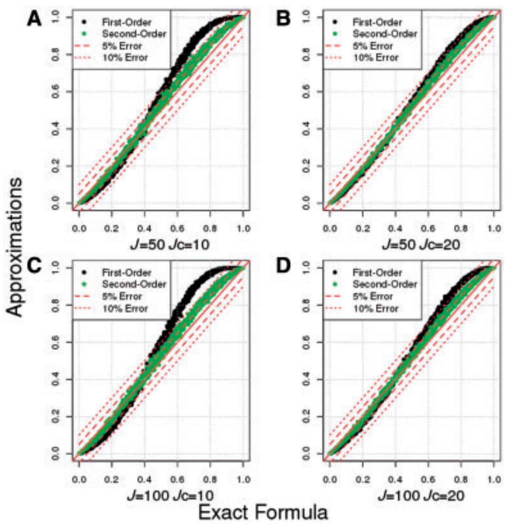

Fig. 1. Evaluation of the accuracy of the first-and second-order approximations to the power of SKAT test under simulation scenario S1 (MAF-independent EV). Exact Formula represents estimated average power over empirical distribution of MAFs and genetic effect sizes using exact theoretical formulas for the SKAT test statistic. First-order represents estimated average power over empirical distribution of MAFs using the first-order approximation for the SKAT test statistic. Second-order represents estimated average power over empirical distribution of MAFs using the second-order approximation for the SKAT test statistic. 5% and 10% error represents 5% and 10% error bounds over exact power calculations. Number of variants in a locus (J) and number of causal variants (Jc): A) $J = 50$, $Jc = 10$; B) $J = 50$, $Jc = 10$; C) $J = 100$, $Jc = 10$ and D) $J = 100$, $Jc = 20$.

We also observe that the accuracy of the proposed approximations for sample size calculations is consistent with what is expected from power calculations (see Supplementary Figs S8–S10). The second-order approximation provides unbiased estimates for average sample size, while accuracy of the first-order approximation improves as the number of causal variants in a locus J_C increases.

As for sum-based statistic, we observe that the accuracy of the proposed approximations does not drastically depend on number of causal variants in a locus (Supplementary Figs S6 and S7). However, the accuracy depends on the variation in SNP-specific coefficients of variations as we take square inside of expectation (see Supplementary Appendix S1). Particularly accuracy should improve if causal variants have similar generic effects.

3.2 Bounds on effect size distribution

Genome-wide power analysis of the educational attainment study (Ganna *et al.*, 2016) (see Fig. 2A), which implemented whole-exome sequencing, shows implausibility of models that correspond to a small number of underlying causal loci explaining significant total phenotypic variance. For example, the probability of observing no discovery is less than 5% under genetic models that involve less than 250 loci to explain a total of 20% or more phenotypic variation. If we assume independence between genetic effects and MAFs across variants, a scenario under which SKAT test has higher power, then even larger number of loci will be needed to explain the same total variance (see Supplementary Fig. S13A).

Genome-wide power analysis of the study of blood pressure (Liu *et al.*, 2016) (see Fig. 2B), which is much larger in sample size but implements the Exome Chip platform, provides a very sharp bound on the relationship between number of underlying causal loci and

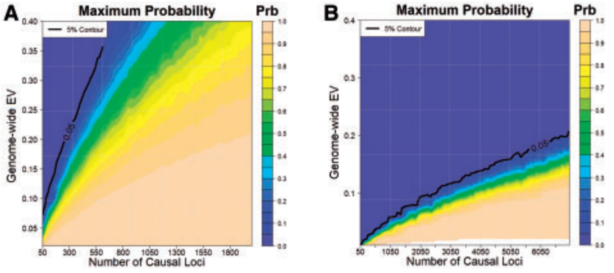

Fig. 2. Bounds for genetic architecture based on results reported in studies of education attainment (EA) and blood pressure (BP). (**A**) Maximum probability of observing no discoveries in the EA study, which used whole-exome sequencing platform, as a function of the number of underlying causal loci K and the total variation explained by them with a sample size of 14 000. (**B**) The maximum probability of observing three statistical significant discoveries in the BP study, which used exome chip, as a function of the number of underlying causal loci K and the total variation explained by them with a sample size of 140 000. In both cases, it is assumed that gene-based tests have been performed using the SKAT test statistics at the level of $T1 = 2.5 \times 10^{-6}$. Probabilities are estimated by (2) and assumption of independence between minor allele frequency and explained variations. Maximum probability was calculated over a set of possible effect size distributions. The black line shows approximate contours (bounds) corresponding to the probability of 5%

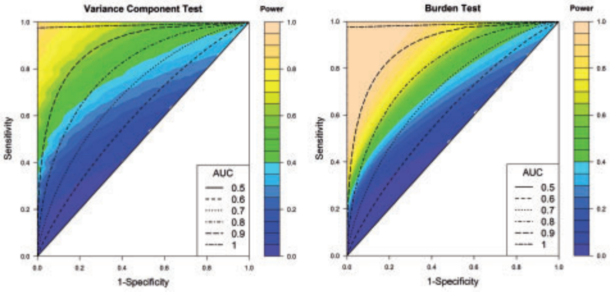

Fig. 3. Effects of sensitivity and specificity for *a priori* variant screening on power of aggregate-level tests. Power of variance component test (SKAT) and burden test are studied under simulation Scenario S1 (MAF-independent EV). Number of variants in a locus is set to $J = 50$ and number of causal variants to $J_C = 10$. The setting corresponds to a baseline power (i.e. if all variants were included in the study) of 40% and 36% for variance component and sum-based tests, respectively

total heritability explained by the underlying variants. We estimate, for example, at least 6000 causal loci will need to be involved if the variants included in the study could explain 20% of phenotypic variance of blood pressure. Identical to results for the WES study, genetic bound is even sharper if independence between MAFs and genetic effects is assumed (see Supplementary Fig. S13B).

3.3 Effects of *a priori* SNP screening on power of aggregated test

A priori 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% and 90%) (see Fig. 3). In contrast, power for sum-based test can improve substantially with more modest discriminatory accuracy of the SNP selection algorithm (AUC between 70% and 80%). Furthermore, we observe that the roles of sensitivity and specificity are not symmetric on power of these tests. For both tests, substantial improvement of power is possible 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 similar results in studies with different genetic architecture and large number of SNPs in a locus (see Supplementary Figs S14 and S15).

4 Discussion

Although large GWAS of low frequency and rare variants are now becoming increasingly feasible due to technological advances, the likely yield of such studies in the future remains uncertain as studies conducted to date have only reported limited number of findings (Purcell *et al.*, 2014; Tang *et al.*, 2014; Cheng *et al.*, 2015; UK10K Consortium *et al.*, 2015; Huang *et al.*, 2015; Mahajan *et al.*, 2015; Wessel *et al.*, 2015; Xu *et al.*, 2015; Zheng *et al.*, 2015; Fuchsberger *et al.*, 2016; Ganna *et al.*, 2016; CHARGE Consortium Hematology Working Group, 2016; Haddad *et al.*, 2016; Liu *et al.*, 2016; Luo *et al.*, 2017). For studies of common variants, which have mostly relied on association testing at the level of individual variants, we and others have shown that the yield of GWAScritically depend on distribution of phenotypic variances explained by individual variants across the genome (Park *et al.*, 2010; Chatterjee *et al.*, 2013; Dudbridge, 2013). 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 (Li and Leal, 2008; Madsen and Browning, 2009; Neale *et al.*, 2011; Wu *et al.*, 2011; Moutsianas *et al.*, 2015). 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. Furthermore, based on these simplified power calculations, we evaluate the potential for improving power for aggregated tests by preselection of likely causal variants based on functional/annotation information.

Power analysis of current studies of large sample sizes may provide important bounds on genetic architecture of the underlying traits. Our analysis suggests that rare variants investigated in current 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. 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 latter approach (UK10K Consortium *et al.*, 2015; Fuchsberger *et al.*, 2016; CHARGE Consortium Hematology Working Group, 2016; Liu *et al.*, 2016) (see Supplementary Table S3). The analytic formula we propose for calculating probability of a certain number of discoveries under various models for genetic architecture can also be applied for single-variant tests. Genetic bounds based on the results from single SNP analysis for the same two studies also show that only under highly polygenic architecture the variants included in these studies can explain a substantial fraction of heritability of the underlying traits (see Supplementary Fig. S16). These genetic bounds based on single SNP analysis are also consistent with corresponding genetic bounds from gene-based analysis and assumption of clustering of multiple causal variants per causal locus. Large studies with more accurate estimates of genetic bounds will provide additional information on the degrees of clustering of multiple rare variants within causal locus.

A variety of studies have studied genetic architecture of common variants by characterization of underlying heritability, number of susceptibility variants and effect size distributions (International Schizophrenia *et al.*, 2009; Park *et al.*, 2010; Yang *et al.*, 2010; Lee *et al.*, 2011, 2012; Park *et al.*, 2011; Stahl *et al.*, 2012; Visscher *et al.*, 2012; Dudbridge, 2013; Zhou *et al.*, 2013; Huang *et al.*, 2015; Loh *et al.*, 2015; Speed *et al.*, 2017). All of these studies consistently point toward a highly polygenic model where disease etiology may involve thousands or even tens of thousands of common susceptibility variants, each conferring only a modest association, but in combinations they can explain substantial phenotypic variations. Some recent studies have reported that low frequency and rare variant studies have the potential to explain significant fraction of heritability for selected traits (UK10K Consortium *et al.*, 2015; Mancuso *et al.*, 2016; Speed *et al.*, 2017). 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 (Price *et al.*, 2010; Zuk *et al.*, 2014). 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 a selected subset of variants point towards challenges for use of functional and annotation information for pre-screening. Overall, it appears that preselection of variants can significantly improve the power of aggregated tests only if the underlying functional/annotation information have a 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 of a disease. Use of a too stringent criterion for variant 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 variant selection on power of aggregated tests.

Sophisticated imputation algorithms (Huang *et al.*, 2015; Kreiner-Moller *et al.*, 2015) and increasing sample size of reference datasets (Huang *et al.*, 2015) are allowing imputation of low frequency and rare variants with increasing accuracy. Many association studies are now being conducted based on imputation in existing large GWAS. 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 (Huang *et al.*, 2009). Further studies are needed to understand the impact of imputation on aggregated tests encompassing variants of different allele frequency spectra.

Our simulation results demonstrated that analytical power formulas presented in Section 2.1 match empirical power very well for quantitative trait (see Supplementary Fig. S17). Under case–control settings, exact power calculations are biased due to reliance on asymptotic normal distribution (see Supplementary Fig. S18). As a result, these calculations may not be adequate for low Type 1 thresholds or low MAFs. Further development in an improvement of the accuracy by incorporating of LD structure and sparsity is needed.

In this report, we have illustrated the 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 tests may be applied in a sliding window fashions (UK10K Consortium *et al.*, 2015; Zheng *et al.*, 2015), 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 variant association tests at the levels of individual loci and whole genome in terms of a few key parameters of the underlying models for genetic architecture. These methods, which we implement in a shiny application in R, will provide useful design tools for planning next-generation GWAS.

Acknowledgements

This study utilized the computational resources of the NIH HPC Biowulf cluster.

Funding

A.D. was supported by the National Cancer Institute Intramural Research Program. The research of H.Z. and N.C. was supported by fund available through Bloomberg Distinguished Professorship Endowment at Johns Hopkins University.

Conflict of Interest: none declared.

References

Asmann,Y.W. *et al.* (2012) TREAT: a bioinformatics tool for variant annotations and visualizations in targeted and exome sequencing data. *Bioinformatics*, **28**, 277–278.

Bulik-Sullivan,B.K. *et al.* (2015) LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.*, **47**, 291–295.

CHARGE Consortium Hematology Working Group. (2016) Meta-analysis of rare and common exome chip variants identifies S1PR4 and other loci influencing blood cell traits. *Nat. Genet.*, **48**, 867–876.

Chatterjee,N. *et al.* (2013) Projecting the performance of risk prediction based on polygenic analyses of genome-wide association studies. *Nat. Genet.*, **45**, 400–405. 405e401-403.

Cheng,C.Y. *et al.* (2015) New loci and coding variants confer risk for age-related macular degeneration in East Asians. *Nat. Commun.*, **6**, 6063.

Davies,R.W. *et al.* (2016) Rapid genotype imputation from sequence without reference panels. *Nat. Genet.*, **48**, 965–969.

Derkach,A. *et al.* (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.

Derkach,A. *et al.* (2014) Pooled association tests for rare genetic variants: a review and some new results. *Stat. Sci.*, **29**, 302–321.

Dudbridge,F. (2013) Power and predictive accuracy of polygenic risk scores. *PLoS Genet.*, **9**, e1003348.

Fuchsberger,C. *et al.* (2016) The genetic architecture of type 2 diabetes. *Nature*, **536**, 41–47.

Ganna,A. *et al.* (2016) Ultra-rare disruptive and damaging mutations influence educational attainment in the general population. *Nat. Neurosci.*, **19**, 1563–1565.

Grant,J.R. *et al.* (2011) In-depth annotation of SNPs arising from resequencing projects using NGS-SNP. *Bioinformatics*, **27**, 2300–2301.

Haddad,S.A. *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.

Huang,J. *et al.* (2015) Improved imputation of low-frequency and rare variants using the UK10K haplotype reference panel. *Nat. Commun.*, **6**, 8111.

Huang,L. *et al.* (2009) The relationship between imputation error and statistical power in genetic association studies in diverse populations. *Am. J. Hum. Genet.*, **85**, 692–698.

Huang,L.Z. *et al.* (2015) Whole-exome sequencing implicates UBE3D in age-related macular degeneration in East Asian populations. *Nat. Commun.*, **6**, 6687. doi: 10.1038/ncomms7687.

International Schizophrenia Consortium *et al.* (2009) Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature*, **460**, 748–752.

Ionita-Laza,I. *et al.* (2011) A new testing strategy to identify rare variants with either risk or protective effect on disease. *PLoS Genet.*, **7**, e1001289.

Kosmicki,J.A. *et al.* (2016) Discovery of rare variants for complex phenotypes. *Hum. Genet.*, **135**, 625–634.

Kreiner-Moller,E. *et al.* (2015) Improving accuracy of rare variant imputation with a two-step imputation approach. *Eur. J. Hum. Genet.*, **23**, 395–400.

Lee,S. *et al.* (2012) Optimal tests for rare variant effects in sequencing association studies. *Biostatistics*, **13**, 762–775.

Lee,S.H. *et al.* (2012) Estimating the proportion of variation in susceptibility to schizophrenia captured by common SNPs. *Nat. Genet.*, **44**, 247–250.

Lee,S.H. *et al.* (2011) Estimating missing heritability for disease from genome-wide association studies. *Am. J. Hum. Genet.*, **88**, 294–305.

Lek,M. *et al.* (2016) Analysis of protein-coding genetic variation in 60, 706 humans. *Nature*, **536**, 285–291.

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.

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.

Liu,C. *et al*. (2016) Meta-analysis identifies common and rare variants influencing blood pressure and overlapping with metabolic trait loci. *Nat. Genet*., **48**, 1162–1170.

Liu,D.J. and Leal,S.M. (2010) A novel adaptive method for the analysis of next-generation sequencing data to detect complex trait associations with rare variants due to gene main effects and interactions. *PLoS Genet*., **6**, e1001156.

Locke,A.E. *et al*. (2015) Genetic studies of body mass index yield new insights for obesity biology. *Nature*, **518**, 197–206.

Loh,P.R. *et al*. (2015) Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nat. Genet*., **47**, 284–290.

Luo,Y. *et al*. (2017) Exploring the genetic architecture of inflammatory bowel disease by whole-genome sequencing identifies association at ADCY7. *Nat. Genet*., **49**, 186–192.

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

Mahajan,A. *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.

Mancuso,N. *et al*. (2016) The contribution of rare variation to prostate cancer heritability. *Nat. Genet*., **48**, 30–35.

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.

Moutsianas,L. *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.

Neale,B.M. *et al*. (2011) Testing for an unusual distribution of rare variants. *PLoS Genet*., **7**, e1001322.

Park,J.H. *et al*. (2011) Distribution of allele frequencies and effect sizes and their interrelationships for common genetic susceptibility variants. *Proc. Natl. Acad. Sci. USA*, **108**, 18026–18031.

Park,J.H. *et al*. (2010) Estimation of effect size distribution from genome-wide association studies and implications for future discoveries. *Nat. Genet*., **42**, 570–575.

Price,A.L. *et al*. (2010) Pooled association tests for rare variants in exon-resequencing studies. *Am. J. Hum. Genet*., **86**, 832–838.

Purcell,S.M. *et al*. (2014) A polygenic burden of rare disruptive mutations in schizophrenia. *Nature*, **506**, 185.

Richardson,T.G. *et al*. (2016) Incorporating non-coding annotations into rare variant analysis. *PLoS One*, **11**, e0154181.

Sampson,J.N. *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.

Speed,D. *et al*. (2017) Reevaluation of SNP heritability in complex human traits. *Nat. Genet*., **49**, 986–992.

Stahl,E.A. *et al*. (2012) Bayesian inference analyses of the polygenic architecture of rheumatoid arthritis. *Nat. Genet*., **44**, 483–489.

Sun,J. *et al*. (2013) A unified mixed-effects model for rare-variant association in sequencing studies. *Genet. Epidemiol*., **37**, 334–344.

Tang,H.Y. *et al*. (2014) A large-scale screen for coding variants predisposing to psoriasis. *Nat. Genet*., **46**, 45.

UK10K Consortium *et al*. (2015) The UK10K project identifies rare variants in health and disease. *Nature*, **526**, 82–90.

Visscher,P.M. *et al*. (2012) Five years of GWAS discovery. *Am. J. Hum. Genet*., **90**, 7–24.

Wang,G.T. *et al*. (2014) Power analysis and sample size estimation for sequence-based association studies. *Bioinformatics*, **30**, 2377–2378.

Wang,K. *et al*. (2010) ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res*., **38**, e164.

Wessel,J. *et al*. (2015) Low-frequency and rare exome chip variants associate with fasting glucose and type 2 diabetes susceptibility. *Nat. Commun*., **6**, 5897.

Wood,A.R. *et al*. (2014) Defining the role of common variation in the genomic and biological architecture of adult human height. *Nat. Genet*., **46**, 1173–1186.

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.

Wu,M.C. *et al*. (2011) Rare-variant association testing for sequencing data with the sequence kernel association test. *Am. J. Hum. Genet*., **89**, 82–93.

Xu,H. *et al*. (2015) Inherited coding variants at the CDKN2A locus influence susceptibility to acute lymphoblastic leukaemia in children. *Nat. Commun*., **6**, 7553.

Yang,J. *et al*. (2010) Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet*., **42**, 565–569.

Zheng,H.F. *et al*. (2015) Whole-genome sequencing identifies EN1 as a determinant of bone density and fracture. *Nature*, **526**, 112–117.

Zhou,X. *et al*. (2013) Polygenic modeling with bayesian sparse linear mixed models. *PLoS Genet*., **9**, e1003264.

Zuk,O. *et al*. (2014) Searching for missing heritability: designing rare variant association studies. *Proc. Natl. Acad. Sci. USA*, **111**, E455–E464.