HHS Public Access

Author manuscript

Hum Hered. Author manuscript; available in PMC 2021 July 28.

Published in final edited form as:

Hum Hered. 2019 ; 84(6): 256–271. doi:10.1159/000508558.

Power and Sample Size Calculations for Genetic Association Studies in the Presence of Genetic Model Mis-Specification

Camille M. Moore[1,*], **Sean A. Jacobso**[1], **Tasha E. Fingerlin**[1]

[1]Center for Genes, Environment, and Health, National Jewish Health, Denver, CO, USA

Abstract

Introduction: When analyzing data from large-scale genetic association studies, such as targeted or genome-wide resequencing studies, it is common to assume a single genetic model, such as dominant or additive, for all tests of association between a given genetic variant and the phenotype. However, for many variants, the chosen model will result in poor model fit and may lack statistical power due to model mis-specification.

Objective: We develop power and sample size calculations for tests of gene and gene by environment interaction, allowing for mis-specification of the true mode of genetic susceptibility.

Methods: The power calculations are based on a likelihood ratio test framework and are implemented in an open-source R package ("genpwr").

Results: We use these methods to develop an analysis plan for a resequencing study in idiopathic pulmonary fibrosis and show that using a 2-degree of freedom test can increase power to detect recessive genetic effects while maintaining power to detect dominant and additive effects.

Conclusions: Understanding the impact of model mis-specification can aid in study design and developing analysis plans that maximize power to detect a range of true underlying genetic effects. In particular, these calculations help identify when a multiple degree of freedom test or other robust test of association may be advantageous.

Keywords

Genetic model misspecification; power analysis; gene-environment interaction; sample size calculation; GWAS

[*]Corresponding Author: Camille M. Moore, Center for Genes, Environment, and Health, National Jewish Health 1400 Jackson St. Denver, CO, 80206, USA, Tel: 303-398-1039, moorec@njhealth.org.

Author Contributions

CMM developed the described power and sample size calculation methods, over-saw the development of the genpwr R package, and wrote the first draft of the manuscript. SAJ developed the genpwr R package and prepared tables and figures and reviewed the manuscript. TEF consulted on the development of the statistical methods, the design of the example power and sample size calculations, and analysis plans and aided in writing the manuscript.

Statement of Ethics

Ethical approval was not required for this research, which describes tools for designing genetic association studies. No data was analyzed as part of this research.

Disclosure Statement

The authors have no conflicts of interest to declare.

Introduction

Genetic association studies remain an important study design for testing hypotheses related to the relationship between DNA polymorphism and biological or clinical traits. Many of these studies are now more focused than a genome-wide study as the results of genome-wide studies are used to design follow-up investigations of the impact of identified genetic risk variants. For example, some studies examine the association between disease-risk variants and the clinical features related to that disease, such as FEV1 in studies of COPD [1–3]. Other study designs use re-sequencing of a small portion of the genome identified in a larger study to further understand how many and which variants are contributing to an association signal identified by relatively sparse genotyping [4–7]. As such, there remains a need to estimate power and/or sample size in the design of genetic association studies.

Several tools for power and sample size calculations for genetic association studies are available [8–14], although many are specialized for specific study designs. For example, GWAPower performs power and sample size calculations for quantitative traits, assuming 1-degree of freedom tests (additive, dominant or recessive genetic effects) and allows users to specify effect sizes in terms of heritability [8]. CaTS was developed specifically to aid in the design 2-stage genetic association studies with a discrete phenotype and allows for multiplicative, recessive, additive and dominant genetic effects [14]. QUANTO is a power calculation tool for both discrete and quantitative phenotypes and for testing gene by environment interactions. QUANTO allows for additive, dominant or recessive genetic effects [9, 10]. ESSPRESSO is a flexible simulation-based approach to power and sample size calculations for both quantitative and discrete phenotypes and gene by environment interactions, while accounting for genotyping errors and phenotype mis-classification [15, 16]. Additive and binary genetic effects are allowed. GPC was originally developed for power calculations for variance component tests for quantitative traits, although other tests are also available[12].

While these tools are all helpful aids for study design, most assume that the model being tested represents the true underlying biological model, which is not known in practice. For example, many studies choose to assume an additive model for testing purposes, based on convenience and/or because it has been shown to be more robust to model misspecification than either a dominant or recessive model [17]. Previous simulation studies have demonstrated the loss of power that can occur when the statistical model is mis-specified (e.g. using an additive model when the true genetic model of susceptibility is recessive)[17–19]. Several robust test statistics have been proposed to address the issue of model mis-specification in the analysis of genetic association studies [20–29]. Popular approaches involve taking the maximum of test statistics assuming several plausible genetic models. For example, the MAX3 or So-Sham test statistic [18] takes a maximum of test statistics assuming additive, recessive and dominant models, with p-values appropriately adjusted to account for these multiple comparisons. While these robust tests have lower power than a correctly specified model, they perform well for additive, dominant and recessive tests and have higher power than an incorrect model [30, 31]. Another simple approach is to use a 2-degree of freedom test, sometimes referred to as a genotypic test, which does not impose any assumptions about the underlying genetic model. Simulation studies have shown that 2-

degree of freedom tests are slightly less powerful than the robust tests described above for truly additive, dominant and recessive genetic effects; however, they are more powerful for other arbitrary genetic effects that do not follow these models, as is the case in overdominance. In addition, 2-degree of freedom tests were found to have the best efficiency robustness when arbitrary genetic effects were considered in addition to the standard genetic models, and have been recommended as a viable alternative to robust test statistics for genome wide scans [30–32].

When designing a genetic association study, it is important that the power calculations match the statistical analysis plan as closely as possible, as data analysis choices can influence statistical power [33]. However, most power and sample size calculation tools do not allow for model mis-specification or for the use of robust test statistics or 2-degree of freedom tests [19]. We know that using an incorrect model, a robust test statistic or a 2-degree of freedom test will result in lower power than a correctly specified model. Assuming that a single model is correct for all variants when performing sample size calculations may lead to an underpowered study. Understanding how model misspecification influences power and sample size calculations is essential to appropriately power genetic association studies and to develop robust analysis plans. Given the wide range of potential study parameters, including sample size, type of outcome, and expected strength of the associations, no one model is most powerful or appropriate for all studies.

To address this need, we have developed a power and sample size calculation tool, GENPWR, that allows calculation of power and sample size under model misspecification. Importantly, we include power calculations for a 2-degree of freedom test. As simulation studies have shown that these tests are slightly less powerful than robust test statistics for the most common genetic models, these calculations can serve as conservative estimates or lower bounds of power (or upper bounds on required sample size) for studies that plan to use robust test statistics in their analyses. GENPWR allows for both discrete and continuous phenotypes, as well as gene by environment interactions.

Materials and Methods

Notation and Models for Genetic Association

In this paper, we extend power and sample size calculations proposed by Gauderman (2002) to account for genetic model misspecification. For consistency, we follow Gauderman's notation where possible [10]. Let Y be the phenotype of interest, either a binary disease trait or a continuous, normally distributed phenotype. If Y is a binary disease trait, define p_d as the prevalence of disease in the study population. If Y is a continuous, normally distributed measurement, let σ_Y be the standard deviation of the measurement in the study population.

Let G be the genotype at a candidate locus with risk allele A and alternative allele a and let q_A represent the risk allele frequency (RAF). Assuming Hardy-Weinberg equilibrium, the distribution of the genotypes is $P(g = AA \mid q_A) = q_A^2$, $P(g = aA \mid q_A) = 2q_A(1 - q_A)$ and $P(g = aa \mid q_A) = (1 - q_A)^2$. In order to perform statistical testing using a linear or logistic regression model, we must choose a genetic model and code the three possible genotypes into a genetic covariate or covariates, X

(Table 1). However, the model that we assume for the pattern of inheritance may be incorrect, which can impact power to detect associations between phenotype and genotype. To allow greater flexibility, we might use two covariates to represent genotype, with X_1 as an indicator of genotype aA and X_2 as an indicator of gentoype AA. We refer to this as "genotypic" or "2 degree of freedom" (2df) coding.

For categorical Y, we consider the logistic regression model:

$$P(Y = 1 \mid X) = \frac{e^{\beta_0 + X\beta_g}}{1 + e^{\beta_0 + X\beta_g}}$$

where $\frac{e^{\beta_0}}{1 + e^{\beta_0}}$ is the probability of disease when X = 0 (or when $X_1 = 0$ and $X_2 = 0$ for 2df coding) and $OR_g = e^{\beta_g}$ is the genetic odds ratio comparing X = 1 to X = 0 for dominant, recessive and additive coding. For 2df coding, $\beta_g = (\beta_{g1}, _{g2})$, with $e^{\beta_{g1}}$ representing the odds ratio comparing genotype aA to aa and $e^{\beta_{g2}}$ comparing AA to aa. For continuous Y, we use a linear regression model:

$$f(Y \mid X) = \beta_0 + X\beta_g + \varepsilon$$

where β_0 is the baseline mean of Y, β_g is the genetic effect, and ε is a normally distributed error term with mean 0 and standard deviation $\sigma_{\in}$. Again, for 2df coding, $\beta_g = (\beta_{g1}, _{g2})$, with β_{g1} representing the difference in means between aA and aa and β_{g2} the difference between AA and aa. To perform power and sample size calculations, the parameters in the above models need to be specified. For the logistic regression model, OR_g and p_d are sufficient to define all model parameters, as β_0 can be calculated from these two quantities. For the linear model, β_g and σ_Y must be specified; from this, $\sigma_{\in}$ can be determined.

True vs. Test Models

As mentioned in Section 2.1, the coding of the genotype covariates for inclusion in the statistical model relies on assumptions about the pattern of inheritance for genetic susceptibility, which may or may not hold for any given candidate gene. For example, we may use dominant coding in our statistical model when in reality the effect of the risk allele on the outcome is additive. Define X^* as the true or correct coding of genetic susceptibility for a gene, and X^M as the coding in the model used to perform statistical testing. Define $\beta^* = \left(\beta_0^*, \beta_g^*\right)'$ as the true values of β_0 and β_g when genotype is correctly included in the model as X^* and $\beta^M = \left(\beta_0^M, \beta_g^M\right)'$ as the values of β_0 and β_g when genotype is included in the model as X^M. Similarly, let σ_ϵ^* be the value of $\sigma_{\in}$ when genotype is coded as X^* and σ_ϵ^M be the value when genotype is coded as X^M.

Calculation of Power and Sample Size

Following Gauderman (2002), we base our power and sample size calculations on likelihood ratio tests. The likelihood ratio test statistic is $\Lambda = 2(L^1 - L^0)$, where L^1 is the log likelihood

of the full model and L^0 is the log likelihood of the reduced model, where β_g is constrained to its null value of 0. Under the null hypothesis, $\Lambda \sim \chi^2_{\mathrm{p}}$, where p is the difference in the number of parameters between the full and reduced models; for additive, dominant, and recessive tests p = 1, while for the 2df test p=2. Under the alternative hypothesis, Λ follows a non-central chi-square distribution with the non-centrality parameter equal to Λ [34].

For a single observation, the log likelihood for the logistic regression model is:

$$\log\left[L^1\left(\beta^M\right)\right] = Y\log\left(\frac{e^{\beta_0^M + X^M\beta_g^M}}{1+e^{\beta_0^M + X^M\beta_g^M}}\right) + (1-Y)\log\left(\frac{1}{1+e^{\beta_0^M + X^M\beta_g^M}}\right)$$

For the linear regression model, the log likelihood is:

$$\log\left[L^1\left(\beta^M\right)\right] = -0.5\log\left(2\pi\sigma_\epsilon^{2\mathrm{M}}\right) - \frac{\left(Y-\beta_0^M - X^M\beta_g^M\right)^2}{2\sigma_\epsilon^{2\mathrm{M}}}$$

The likelihoods depend on the value of Y. Therefore, we calculate the expected log likelihood, given the true parameter values, β^*, X^*, q_A, p_d and σ_ϵ^*. This differs from the method proposed by Gauderman et al [10], since we allow the coding of the genetic covariate to differ between the true pattern of genetic susceptibility and the coding used in statistical tests of association. In other words, we allow for $X^M \neq X^*$ and $\beta^M \neq \beta^*$.

For the logistic regression model the expected log likelihood is:

$$\begin{aligned} E\left(\log\left[L\left(\beta^M\right)\right]\right) &= \sum_G P(g \mid q_A)\left[E(Y \mid G=g)\log\left(\frac{e^{\eta_g^M}}{1+e^{\eta_g^M}}\right) + E(1-Y \mid G=g)\log\left(\frac{1}{1+e^{\eta_g^M}}\right)\right] \\ &= \sum_G P(g \mid q_A)\left[\frac{e^{\eta_g^*}}{1+e^{\eta_g^*}}\log\left(\frac{e^{\eta_g^M}}{1+e^{\eta_g^M}}\right) + \frac{1}{1+e^{\eta_g^*}}\log\left(\frac{1}{1+e^{\eta_g^M}}\right)\right] \end{aligned}$$

where $\eta_g^* = \beta_0^* + X_g^*\beta_g^*$ and $\eta_g^M = \beta_0^M + X_g^M\beta_g^M$ under the alternative hypothesis and $\eta_g^M = \beta_0^M$ under the null hypothesis. X_g^* is the value of X^* for genotype g and X_g^M is the value of X^M for genotype g. For the linear regression model:

$$\begin{aligned} E\left(\log\left[L\left(\beta^M\right)\right]\right) &= \sum_G P(g \mid q_A)\left\{-0.5\log\left(2\pi\sigma_\epsilon^{2M}\right) - E\left[\frac{\left(Y-\eta_g^M\right)^2}{2\sigma_\epsilon^{2M}}\right]\right\} \\ &= \sum_G P(g \mid q_A)\left\{-0.5\log\left(2\pi\sigma_\epsilon^{2M}\right) - \frac{\sigma_\epsilon^{2\,*} + \left(\eta_g^*\right)^2 - 2\eta_g^*\eta_g^M + \left(\eta_g^M\right)^2}{2\sigma_\epsilon^{2M}}\right\} \end{aligned}$$

Power and/or sample size can be calculated using these expected likelihoods as described by Gauderman et al [10].

Extensions for Tests of Gene × Environment Interactions

The above framework can easily be extended to tests of gene × environment interaction. Let E be an exposure or environmental factor, which can be either categorical or continuous. For continuous E, we assume E is normally distributed with standard deviation σ_e. For categorical E, we define p_e as the probability of exposure in the study population. We assume independence of the exposure and genotype in the population.

Again, we consider two basic models to describe the association of genotype, environment and the outcome, a logistic regression for categorical Y and a linear regression for continuous Y. For the logistic regression model:

$$P(Y = 1 \mid X, E) = \frac{e^{\beta_0 + X}\beta_g + E\beta_e + XE\beta_{ge}}{1 + e^{\beta_0 + X\beta_g + E\beta_e + XE\beta_{ge}}}$$

where $\frac{e^{\beta_0}}{1+e^{\beta_0}}$ is the probability of disease when X = E = 0, $OR_g = e^{\beta_g}$ is the genetic odds ratio when $E = 0$, $OR_e = e^{\beta_e}$ is the odds ratio for the environment when $G = 0$, and $OR_{ge} = e^{\beta_{ge}}$ is the interaction odds ratio. For the linear regression model:

$$f(Y \mid X) = \beta_0 + X\beta_g + E\beta_e + XE\beta_{ge} + \varepsilon$$

where β_0 is the baseline mean of Y, β_g is the genetic effect, β_e is the environment effect, β_{ge} is the interaction effect, and ε is a normally distributed error term with mean 0 and standard deviation σ_ϵ. To perform power and sample size calculations for the logistic regression model, OR_g, OR_e, OR_{ge}, p_d, and either p_e or σ_e are sufficient to define all model parameters, as β_0 can be calculated from these quantities. For the linear model, β_g, β_e, β_{ge}, σ_Y, and either p_e or σ_e must be specified; from this, σ_ϵ can be determined.

For a categorical environmental factor, the expected log likelihood for the logistic regression model is:

$$E\left(\log\left[L\left(\beta^M\right)\right]\right) = \sum_E \sum_G P(X_e \mid p_e)P(g \mid q_A)\left[\frac{e^{\eta^*_{ge}}}{1+e^{\eta^x_{ge}}}\log\left(\frac{e^{\eta^M_{ge}}}{1+e^{\eta^M_{ge}}}\right) + \frac{1}{1+e^{\eta^*_{ge}}}\log\left(\frac{1}{1+e^{\eta^M_{ge}}}\right)\right]$$

where $\eta^*_{ge} = \beta^*_0 + X^*_g\beta^*_g + X_e\beta^*_e + X^*_g X_e\beta^*_{ge}$, $\eta^M_{ge} = \beta^M_0 + X^M_g\beta^M_g + X_e\beta^M_e + X^M_g X_e\beta^M_{ge}$ under the alternative hypothesis and $\eta^M_{ge} = \beta^M_0 + X^M_g\beta^M_g + X_e\beta^M_e$ under the null hypothesis. For the linear regression model:

$$E\left(\log\left[L\left(\beta^M\right)\right]\right) = \sum_E \sum_G P(X_e \mid p_e)P(g \mid q_A)\left\{-0.5\log\left(2\pi\sigma^2_\epsilon{}^M\right) - \frac{\sigma^2_\epsilon{}^* + \left(\eta^*_{ge}\right)^2 - 2\eta^*_{ge}\eta^M_{ge} + \left(\eta^M_{ge}\right)^2}{2\sigma^2_\epsilon{}^M}\right\}$$

For a continuous environment variable, the summations over E can be replaced with an integral:

$$E\left(\log\left[L\left(\beta^M\right)\right]\right) = \frac{1}{\sqrt{(2\pi\sigma_e^2)}}\int e^{\frac{X_e^2}{2\sigma_e^2}}\left[\sum_G P(g \mid q_A)E\left(\log\left[L\left(\beta^M\right)\right] \mid G = g, E = X_e\right)\right]dX_e$$

Power and sample size can be computed as described in Gauderman to test for $\beta_{ge}^M = 0$, using the expected log likelihoods $E\left(\log\left[L\left(\beta_0^M, \beta_g^M, \beta_e^M, \beta_{ge}^M\right)\right]\right)$ and $E\left(\log\left[L\left(\beta_0^M, \beta_g^M, \beta_e^M\right)\right]\right)$ for the alternative and null hypotheses, respectively.

Comparison of Power Across Study Designs

In Figures 1 and 2, we depict how power to detect an odds ratio of 1.5 in a case control GWA study changes as a result of differing sample size, significance level, RAF and proportion of cases in the study population. In Figures 3 and 4 we focus on variants with relatively low RAF's ranging from 0.005 to 0.05. As expected, higher sample sizes and less stringent significance thresholds result in increased power for both RAF > 0.05 and < 0.05. In Figures 2 and 4 we see that a balanced study design, with a 1:1 ratio of cases and controls results in the higher power than study designs with either 25% cases or 75% cases for both RAF > 0.05 and < 0.05.

As expected, the correctly specified model always has the highest power to detect an *OR* of 1.5 across the range of RAFs; however, in practice, the correct model is unknown. The relative performance of incorrectly specified models and the 2df model depends on the RAF, the sample size, the significance level and the odds ratio to detect. For variants that act in a dominant manner, the additive model has higher power than the 2df model for lower RAF, but lower power for higher RAF. A similar pattern is seen when using 2df and dominant coding to test truly additive genetic effects. The exact RAF at which the power curves for the 2df test and the mis-specified additive or dominant test cross depends on the sample size, significance threshold, ratio of cases to controls and OR to detect. With increasing sample size, the RAF beyond which 2df tests have higher power decreases. For example, assuming a 5×10^{-8} significance level and a 1:1 case control design, with a total sample size of 5,000, the 2df test has higher power to detect an additive genetic effect than the dominant test for all RAF greater than 14.5%, while with 20,000 subjects, the 2df test has higher power for all RAF greater than 8.5%. This RAF also decreases with less stringent significance thresholds and larger detectable OR's.

For truly recessive effects, the 2df test has higher power than either the additive or dominant models, particularly at lower RAF. Notably, for low RAF there is very low power to detect recessive effects, even when using a correctly specified model, as the probability of observing a subject homozygous for the risk allele becomes extremely low. For example, with a RAF of 0.01 and sample size of 5,000, we would expect < 1 subject homozygous for the risk allele. In the absence of subjects with the aa genotype, the additive, dominant and genotypic tests become equivalent. We see that power for these tests becomes increasingly similar as RAF declines, regardless of the true genetic effect.

Example Modeling Strategies for GWA Study with N = 5,000

We illustrate how these results could be used to develop analysis plans for a 1:1 case control GWA study design with total sample size of 5,000, assuming a 5×10^{-8} level is used to determine statistical significance. With a sample size of 5,000, additive odds ratios of 1.5 can be detected with 80% power with either the dominant, additive or 2df models for RAF ≥≈ 12% (12%, 11.5%, and 12.5%, respectively, Table 2). For truly additive effects, the 2df test has higher power than the dominant test for all RAF greater than 14.5%.

For variants that act in a dominant manner, using the correctly specified dominant model coding results in greater than 80% power to detect an *OR* of 1.5 for all RAF ≥ 14.5%, while for additive and 2df coding, greater than 80% power is achieved for RAF ≥ 17% and 16.5% respectively (Table 2). For truly dominant effects, the 2df test has higher power than the additive test for all RAF greater than 15.5%.

For recessive variants, no testing model has power greater than 80% to detect an *OR* of 1.5 for RAF < 50%. However, moderate size recessive *OR*'s can be detected at a RAF of 30% using a 2df or recessive model (Table 3. Given these power curves, we would recommend using the 2df test or another robust test of association for all variants with RAF greater than 15.5%, as this model has higher power than an incorrectly specified dominant or additive model. For variants with RAF below 15.5%, we would suggest using either the additive or dominant test, since there is little power to detect recessive odds ratios and these tests have higher power than the 2df test to test for additive or dominant effects in this RAF range (Figure 1, second row). However, it should be noted that power drops below 50% for all tests for RAF < 8%; we would suggest considering alternative testing strategies, such as a grouped variant testing method for these low RAFs.

Application to a Re-Sequencing Study of Idiopathic Pulmonary Fibrosis

We use these methods to develop an analysis plan for a study of gene by environment interaction. Idiopathic pulmonary fibrosis (IPF) is a progressive fibrotic lung disease with a median survival of 3 years [35]. Studies of both familial and sporadic disease have identified rare mutations in telomerase (TERT, TERC, RTEL1, and PARN) and surfactant protein (SFTPC and SFTPA2) genes [36–39] and common variants in 12 genetic loci [40–43]. The strongest known risk factor for both familial and sporadic IPF is a polymorphism in the distal promoter region of MUC5B gene, rs35705950. The risk variant is common (10% frequency) among individuals of European ancestry and acts in an additive fashion, with each additional risk allele resulting in approximately a 5 times increase in the odds of IPF. In comparison, other polymorphisms associated with IPF have odds ratios ranging from approximately 1.25 to 1.5 [44].

In addition to these genetic polymorphisms, other potential risk factors for the development of IPF have been identified, including environmental and occupational exposures, tobacco smoking, and comorbidities such as gastroesophageal reflux disease. Meta-analyses have found that odds of IPF are 1.58 times higher in ever-smokers compared to never smokers (95% CI 1.27–1.97) [45]. The prevalence of ever-smoking in IPF cases has been reported at

72% vs. 63% in age, sex and geographically matched controls [46] and there may be a dose response relationship, with those with longer smoking histories having increased odds of disease [47].

An important question is whether the effect of smoking on the odds of IPF is modified by genetic risk factors. For example, genetic variants that lead to over-expression of the MUC5B mucin may impair mucociliary function [48]. This may in turn cause excess retention of inhaled substances, such as cigarette smoke, and could increase the potential for inhaled substances to damage the lungs. We develop an analysis plan to test for gene by smoking interactions in a cohort of 3,624 cases with IPF and 4,442 controls. Subjects in the cohort have previously had targeted, deep sequencing in regions with GWAS signals associated with IPF. Our analysis decisions are informed by considering power to detect a gene by smoking interaction odds ratio of 1.5, 1.75 and 2, assuming a 5×10^{-5} significance level, chosen based on the number of variants available for analysis. We perform power calculations assuming genetic odds ratios of 1.25, 1.5, and 5, based on the range of effects seen in previous studies of IPF and assume a 1.6 odds ratio for ever-smoking and an ever-smoking prevalence of 67%.

Power curves to detect gene by smoking interaction odds ratios of 1.5, 1.75 and 2 are presented in Figures 5, 6, and 7. First we note that results for OR_g of 1.25 and 1.5 are very similar to one another and that these results largely follow the same patterns as the tests for genetic association seen in Section 3, with the correct test being the most powerful. For OR_g of 1.25 and 1.5, we again see that for dominant (additive) variants, a mis-specified additive (dominant) test is more powerful than the 2df test only for lower RAF. For recessive variants, the 2df test is more powerful than additive or dominant tests.

In the case of OR_g of 5, we see that the power for most tests is reduced compared to when OR_g is smaller. This may be due to the extreme odds ratio resulting in few subjects in certain disease status and genotype combinations. For example, with such a large odds ratio, we would expect few controls to be homozygous for the risk allele. While results for additive and recessive tests follow the same patterns as described above, for dominant variants, the additive test somewhat unexpectedly has the highest power. While the expected log likelihood for the correctly specified dominant model is higher than that of the mis-specified additive model, the difference in expected log likelihoods between the null and full models is larger for the additive model. This larger change indicates a bigger improvement in model fit when the gene by smoking interaction term is included in the additive model, resulting in higher power.

As we believe that the majority of variants will have odds ratios much less than 5, we focus on the results for genetic OR's of 1.25 and 1.5. In addition, we see that there is limited power to detect interaction odds ratios of 1.5, and therefore focus on OR_{ge} of 1.75 and 2. For dominant variants, the 2df model has higher power to detect the GxE effect than the additive model for all RAF > 0.26 when OR_g= 1.25 and for RAF > 0.335 when OR_g= 1.5. However, the power for the 2df model and the incorrectly specified additive model remain relatively similar even at RAF of 0.2. For additive variants, the 2df model has higher power to detect the GxE effect than the dominant model for all RAF > 0.16 when OR_g= 1.25 and for RAF >

0.155 when OR_g= 1.5. Since compared to the additive and dominant models, the 2df model has substantially increased power to detect interactions for recessive variants, we plan to test all variants with RAF > 0.2 with the 2df model and all variants with RAF ≤ 0.2 with the additive model.

Discussion/Conclusion

We have described and implemented a power and sample size approach that allows misspecification of the statistical model used for testing compared to the true underlying biological model. The ability to assess the impact of model misspecification on power is important when considering the trade-offs between analysis approaches in terms of type I error, power, and simplicity of implementation. We have shown that our approach is helpful for making both study design and analysis approach decisions based on the specifics of the goals and parameters of the study of interest.

Using these power calculation tools, we show when using a 2df test or another robust test of association is advantageous compared to mis-specified additive or dominant models. 2df tests have higher power to detect recessive genetic effects across the range of RAF. For additive and dominant effects, 2df tests have higher power than an incorrectly specified additive or dominant model for higher RAF and lower power for low RAF. The RAF where these power curves cross depends on several factors, including the sample size, the ratio of cases to controls, the OR to detect, and the significance level. In general, as sample size, significance level and OR to detect increase, the RAF at which a 2df test becomes more powerful decreases. While it is clear that a 2df or robust tests of association should be used at higher RAF due to higher power than other mis-specified tests for additive, dominant and recessive variants, an additive or dominant test may be preferable at lower RAF where there is little power to detect recessive effects and mis-specified additive (or dominant) models have higher power than 2df tests for truly dominant (or additive) genetic effects. Notably, as RAF decreases, the power of additive and dominant tests become more similar since the probability of observing a subject homozygous for the risk allele becomes smaller.

We also illustrate how to utilize these calculations to develop analysis plans for testing gene by environment interactions in a study of idiopathic pulmonary fibrosis. For moderate genetic odds ratios, the results for power to detect a gene by smoking interaction largely followed the same patterns as the power for tests of single variants. However, when considering genes with high odds ratios, such as MUC5B, some results were unexpected. Power to detect a gene by environment interaction in the presence of a large genetic odds ratio was reduced compared to more moderate genetic odds ratios, likely because the larger odds ratio resulted in fewer subjects in some disease status and genotype combinations. With this larger genetic odds ratio, we also found that for dominant variants, the mis-specified additive test had the highest power to detect a gene by environment interaction, due to a bigger improvement in model fit when the gene by smoking interaction term is included in the additive model. This illustrates the importance of considering the full range of possible parameter values when performing power and sample size calculations, particularly for gene by environment interactions.

The calculations we have described are implemented in the genpwr R package, which is available on the Comprehensive R Archive Network (CRAN). This software performs power and sample size calculations assuming either dichotomous or continuous outcomes, while also allowing for optional dichotomous or continuous environmental exposures and can provide information for tests of both main and interaction effects. In the R package, the user specifies one or more “true” models, such as additive, dominant, and/or recessive, which represent the biological relationships between genotype and the outcome, as well as one or more “test” models (additive, dominant, recessive and/or 2df), which indicate how the genetic effect will be coded for statistical testing. The genpwr package can calculate sample size, power, or detectable effect size, given that the other two of these variables are specified by the user. In addition, the user must also specify a range of minor allele frequencies and type 1 error rate. For binary outcomes, either the disease prevalence in the study population or the number of controls per case must be specified. For continuous outcomes, the standard deviation of the outcome in the study population is required. For calculations involving a gene by environment interaction, depending on the type of exposure measurement, either the prevalence or the standard deviation of the exposure in the study population must be specified.

Funding Sources

CMM, SAJ and TEF were supported by NIH/NHLBI 5R01HL097163.

References

1. Li X, Ortega VE, Ampleford EJ, Graham Barr R, Christenson SA, Cooper CB, et al. Genome-wide association study of lung function and clinical implication in heavy smokers. BMC Med Genet. 2018;19(1):134. [PubMed: 30068317]
2. Sadeghnejad A, Ohar JA, Zheng SL, Sterling DA, Hawkins GA, Meyers DA, et al. Adam33 polymorphisms are associated with COPD and lung function in long-term tobacco smokers. Respir Res. 2009;10:21. [PubMed: 19284602]
3. Wain LV, Shrine N, Miller S, Jackson VE, Ntalla I, Soler Artigas M, et al. Novel insights into the genetics of smoking behaviour, lung function, and chronic obstructive pulmonary disease (UK BiLEVE): a genetic association study in UK Biobank. Lancet Respir Med. 2015;3(10):769–81. [PubMed: 26423011]
4. Diogo D, Kurreeman F, Stahl EA, Liao KP, Gupta N, Greenberg JD, et al. Rare, low-frequency, and common variants in the protein-coding sequence of biological candidate genes from GWASs contribute to risk of rheumatoid arthritis. Am J Hum Genet. 2013;92(1):15–27. [PubMed: 23261300]
5. Johansen CT, Wang J, Lanktree MB, Cao H, McIntyre AD, Ban MR, et al. Excess of rare variants in genes identified by genome-wide association study of hypertriglyceridemia. Nat Genet. 2010;42(8):684–7. [PubMed: 20657596]
6. Momozawa Y, Mni M, Nakamura K, Coppieters W, Almer S, Amininejad L, et al. Resequencing of positional candidates identifies low frequency IL23R coding variants protecting against inflammatory bowel disease. Nat Genet. 2011;43(1):43–7. [PubMed: 21151126]
7. Raychaudhuri S, Iartchouk O, Chin K, Tan PL, Tai AK, Ripke S, et al. A rare penetrant mutation in CFH confers high risk of age-related macular degeneration. Nat Genet. 2011;43(12):1232–6. [PubMed: 22019782]
8. Feng S, Wang S, Chen CC, Lan L. GWAPower: a statistical power calculation software for genome-wide association studies with quantitative traits. BMC Genet. 2011;12:12. [PubMed: 21255436]
9. Gauderman WJ. Sample size requirements for association studies of gene-gene interaction. Am J Epidemiol. 2002;155(5):478–84. [PubMed: 11867360]

10. Gauderman WJ. Sample size requirements for matched case-control studies of gene-environment interaction. Stat Med. 2002;21(1):35–50. [PubMed: 11782049]
11. Gordon D, Finch SJ, Nothnagel M, Ott J. Power and sample size calculations for case-control genetic association tests when errors are present: application to single nucleotide polymorphisms. Hum Hered. 2002;54(1):22–33. [PubMed: 12446984]
12. Purcell S, Cherny SS, Sham PC. Genetic Power Calculator: design of linkage and association genetic mapping studies of complex traits. Bioinformatics. 2003;19(1):149–50. [PubMed: 12499305]
13. Schork NJ. Power calculations for genetic association studies using estimated probability distributions. Am J Hum Genet. 2002;70(6):1480–9. [PubMed: 11992254]
14. Skol AD, Scott LJ, Abecasis GR, Boehnke M. Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies. Nat Genet. 2006;38(2):209–13. [PubMed: 16415888]
15. Burton PR, Hansell AL, Fortier I, Manolio TA, Khoury MJ, Little J, et al. Size matters: just how big is BIG?: Quantifying realistic sample size requirements for human genome epidemiology. Int J Epidemiol. 2009;38(1):263–73. [PubMed: 18676414]
16. Gaye A, Burton TW, Burton PR. ESPRESSO: taking into account assessment errors on outcome and exposures in power analysis for association studies. Bioinformatics. 2015;31(16):2691–6. [PubMed: 25908791]
17. Lettre G, Lange C, Hirschhorn JN. Genetic model testing and statistical power in population-based association studies of quantitative traits. Genet Epidemiol. 2007;31(4):358–62. [PubMed: 17352422]
18. So HC, Sham PC. Robust association tests under different genetic models, allowing for binary or quantitative traits and covariates. Behav Genet. 2011;41(5):768–75. [PubMed: 21305351]
19. Gaye A, Davis SK. Genetic model misspecification in genetic association studies. BMC Res Notes. 2017;10(1):569. [PubMed: 29115983]
20. Chen Z A new association test based on Chi-square partition for case-control GWA studies. Genet Epidemiol. 2011;35(7):658–63. [PubMed: 22009790]
21. Chen Z Association tests through combining p-values for case control genome-wide association studies. Statistics & Probability Letters. 2013;83(8):1854–62.
22. Chen Z, Ng HK. A robust method for testing association in genome-wide association studies. Hum Hered. 2012;73(1):26–34. [PubMed: 22212363]
23. Gonzalez JR, Carrasco JL, Dudbridge F, Armengol L, Estivill X, Moreno V. Maximizing association statistics over genetic models. Genet Epidemiol. 2008;32(3):246–54. [PubMed: 18228557]
24. Kwak M, Joo J, Zheng G. A robust test for two-stage design in genome-wide association studies. Biometrics. 2009;65(4):1288–95. [PubMed: 19432785]
25. Song K, Elston RC. A powerful method of combining measures of association and Hardy-Weinberg disequilibrium for fine-mapping in case-control studies. Stat Med. 2006;25(1):105–26. [PubMed: 16220513]
26. Talluri R, Wang J, Shete S. Calculation of exact p-values when SNPs are tested using multiple genetic models. BMC Genet. 2014;15(75).
27. Wang K, Sheffield VC. A constrained-likelihood approach to marker-trait association studies. Am J Hum Genet. 2005;77(5):768–80. [PubMed: 16252237]
28. Vukcevic D, Hechter E, Spencer C, Donnelly P. Disease model distortion in association studies. Genet Epidemiol. 2011;35(4):278–90. [PubMed: 21416505]
29. Zheng G, Ng HK. Genetic model selection in two-phase analysis for case-control association studies. Biostatistics. 2008;9(3):391–9. [PubMed: 18003629]
30. Joo J, Kwak M, Ahn K, Zheng G. A robust genome-wide scan statistic of the Wellcome Trust Case-Control Consortium. Biometrics. 2009;65(4):1115–22. [PubMed: 19432787]
31. Zheng G, Freidlin B, Gastwirth JL. Comparison of robust tests for genetic association using case-control studies. Institute of Mathematical Statistics Lecture Notes - Monograph Series. 2006;49:253–65.

32. Joo J, Kwak M, Chen Z, Zheng G. Efficiency robust statistics for genetic linkage and association studies under genetic model uncertainty. Stat Med. 2010;29(1):158–80. [PubMed: 19918942]
33. Sham PC, Purcell SM. Statistical power and significance testing in large-scale genetic studies. Nat Rev Genet. 2014;15(5):335–46. [PubMed: 24739678]
34. Steven G Self RHMaJO. Power Calculations for Likelihood Ratio Tests in Generalized Linear Models International Biometric Society (3, 1992);48(No. 1):31–9
35. Martinez FJ, Collard HR, Pardo A, Raghu G, Richeldi L, Selman M, et al. Idiopathic pulmonary fibrosis. Nat Rev Dis Primers. 2017;3:17074. [PubMed: 29052582]
36. Armanios MY, Chen JJ, Cogan JD, Alder JK, Ingersoll RG, Markin C, et al. Telomerase mutations in families with idiopathic pulmonary fibrosis. N Engl J Med. 2007;356(13):1317–26. [PubMed: 17392301]
37. Nogee LM, Dunbar AE 3rd, Wert SE, Askin F, Hamvas A, Whitsett JA. A mutation in the surfactant protein C gene associated with familial interstitial lung disease. N Engl J Med. 2001;344(8):573–9. [PubMed: 11207353]
38. Stuart BD, Choi J, Zaidi S, Xing C, Holohan B, Chen R, et al. Exome sequencing links mutations in PARN and RTEL1 with familial pulmonary fibrosis and telomere shortening. Nat Genet. 2015;47(5):512–7. [PubMed: 25848748]
39. Wang Y, Kuan PJ, Xing C, Cronkhite JT, Torres F, Rosenblatt RL, et al. Genetic defects in surfactant protein A2 are associated with pulmonary fibrosis and lung cancer. Am J Hum Genet. 2009;84(1):52–9. [PubMed: 19100526]
40. Allen RJ, Porte J, Braybrooke R, Flores C, Fingerlin TE, Oldham JM, et al. Genetic variants associated with susceptibility to idiopathic pulmonary fibrosis in people of European ancestry: a genome-wide association study. Lancet Respir Med. 2017;5(11):869–80. [PubMed: 29066090]
41. Fingerlin TE, Murphy E, Zhang W, Peljto AL, Brown KK, Steele MP, et al. Genome-wide association study identifies multiple susceptibility loci for pulmonary fibrosis. Nat Genet. 2013;45(6):613–20. [PubMed: 23583980]
42. Fingerlin TE, Zhang W, Yang IV, Ainsworth HC, Russell PH, Blumhagen RZ, et al. Genome-wide imputation study identifies novel HLA locus for pulmonary fibrosis and potential role for auto-immunity in fibrotic idiopathic interstitial pneumonia. BMC Genet. 2016;17(1):74. [PubMed: 27266705]
43. Noth I, Zhang Y, Ma SF, Flores C, Barber M, Huang Y, et al. Genetic variants associated with idiopathic pulmonary fibrosis susceptibility and mortality: a genome-wide association study. Lancet Respir Med. 2013;1(4):309–17. [PubMed: 24429156]
44. Moore C, Blumhagen RZ, Yang IV, Walts A, Powers J, Walker T, et al. Resequencing Study Confirms That Host Defense and Cell Senescence Gene Variants Contribute to the Risk of Idiopathic Pulmonary Fibrosis. Am J Respir Crit Care Med. 2019;200(2):199–208. [PubMed: 31034279]
45. Taskar VS, Coultas DB. Is idiopathic pulmonary fibrosis an environmental disease? Proc Am Thorac Soc. 2006;3(4):293–8. [PubMed: 16738192]
46. Baumgartner KB, Samet JM, Stidley CA, Colby TV, Waldron JA. Cigarette smoking: a risk factor for idiopathic pulmonary fibrosis. Am J Respir Crit Care Med. 1997;155(1):242–8. [PubMed: 9001319]
47. Baumgartner KB, Samet JM, Coultas DB, Stidley CA, Hunt WC, Colby TV, et al. Occupational and environmental risk factors for idiopathic pulmonary fibrosis: a multicenter case-control study. Collaborating Centers. Am J Epidemiol. 2000;152(4):307–15. [PubMed: 10968375]
48. Button B, Cai LH, Ehre C, Kesimer M, Hill DB, Sheehan JK, et al. A periciliary brush promotes the lung health by separating the mucus layer from airway epithelia. Science. 2012;337(6097):937–41. [PubMed: 22923574]

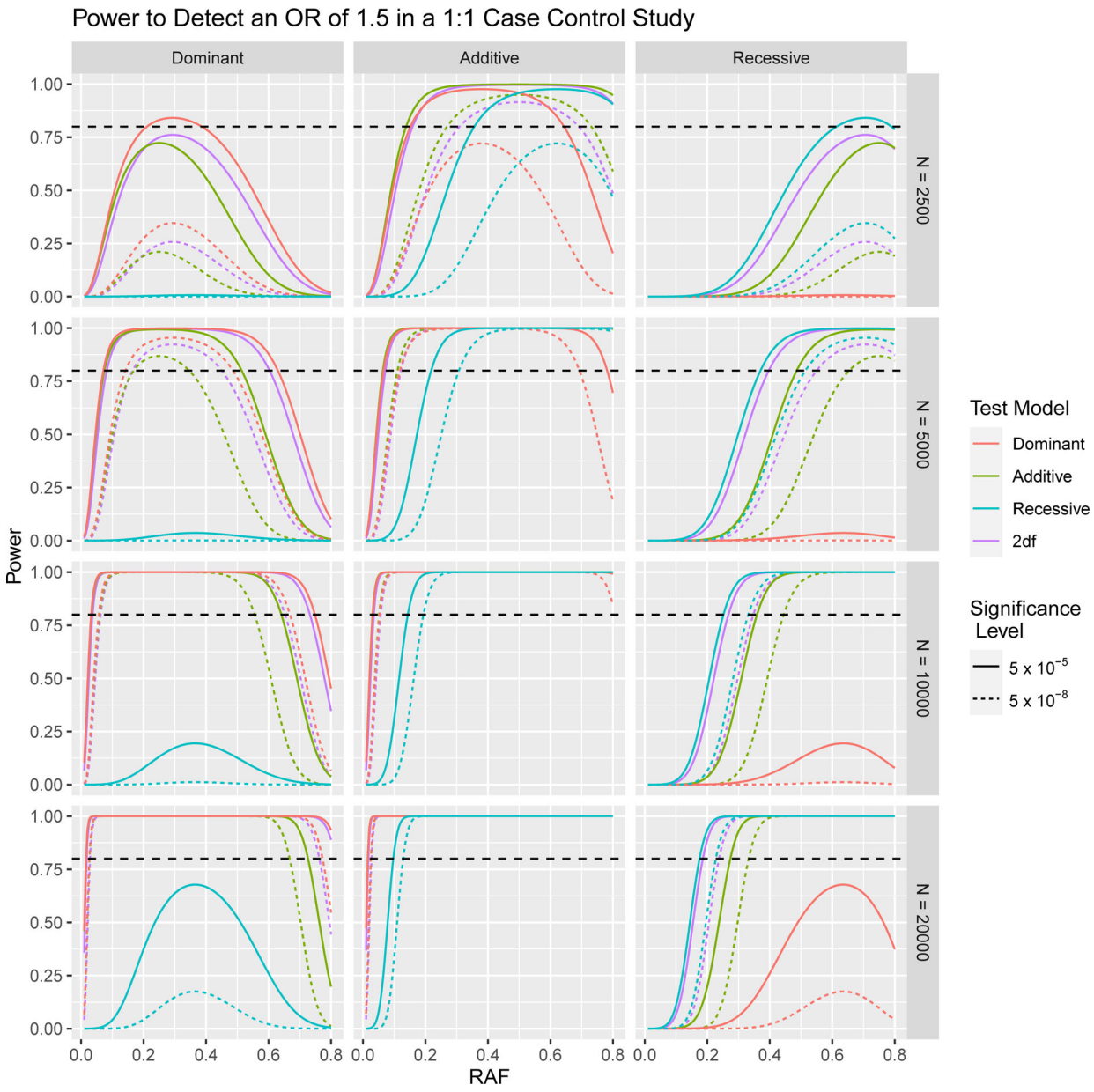

Fig. 1.

Power to Detect an Odds Ratio of 1.5 in 1:1 Case Control Study with a Total of N Subjects, at 5×10^{-5} and 5×10^{-8} Significance Levels. The horizontal dashed line indicates 80% power.

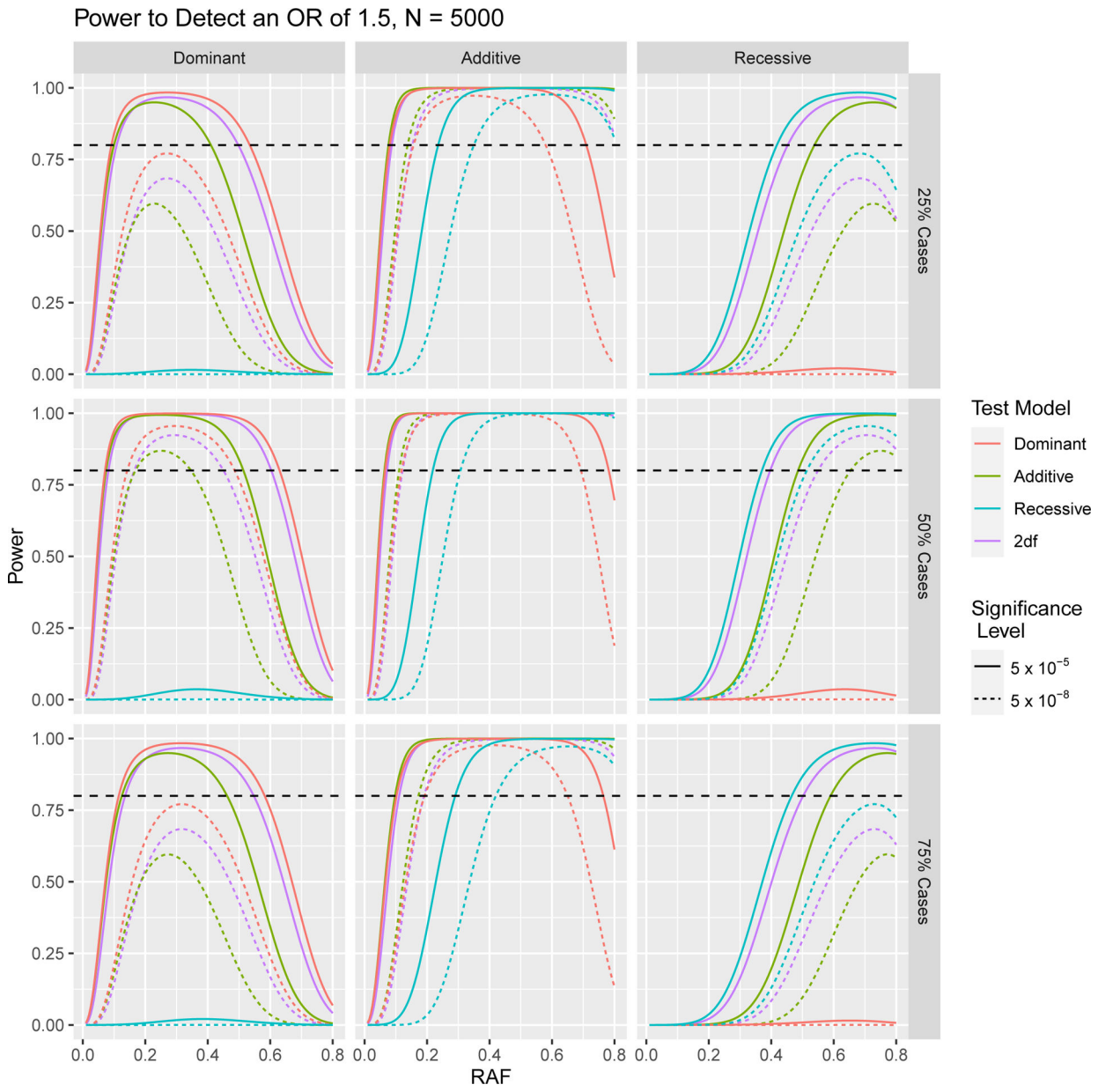

Fig. 2.

Power to Detect an Odds Ratio of 1.5 in Case Control Study with a Total of 5,000 Subjects for Different Ratios of Cases to Controls, at 5×10^{-5} and 5×10^{-8} Significance Levels. The horizontal dashed line indicates 80% power.

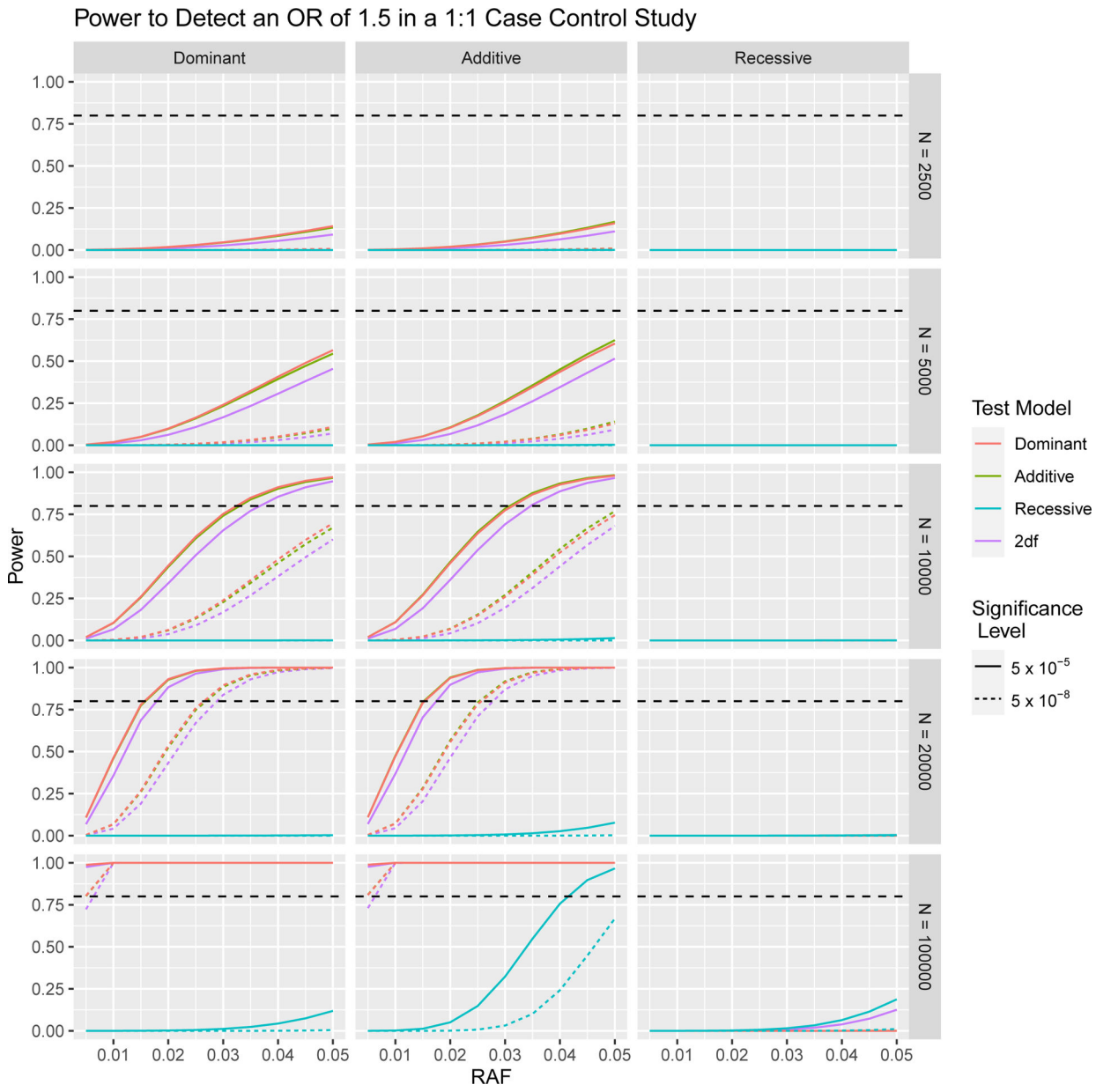

Fig. 3.

Power to Detect an Odds Ratio of 1.5 in 1:1 Case Control Study with a Total of N Subjects, at 5×10^{-5} and 5×10^{-8} Significance Levels at Low RAFs. The horizontal dashed line indicates 80% power.

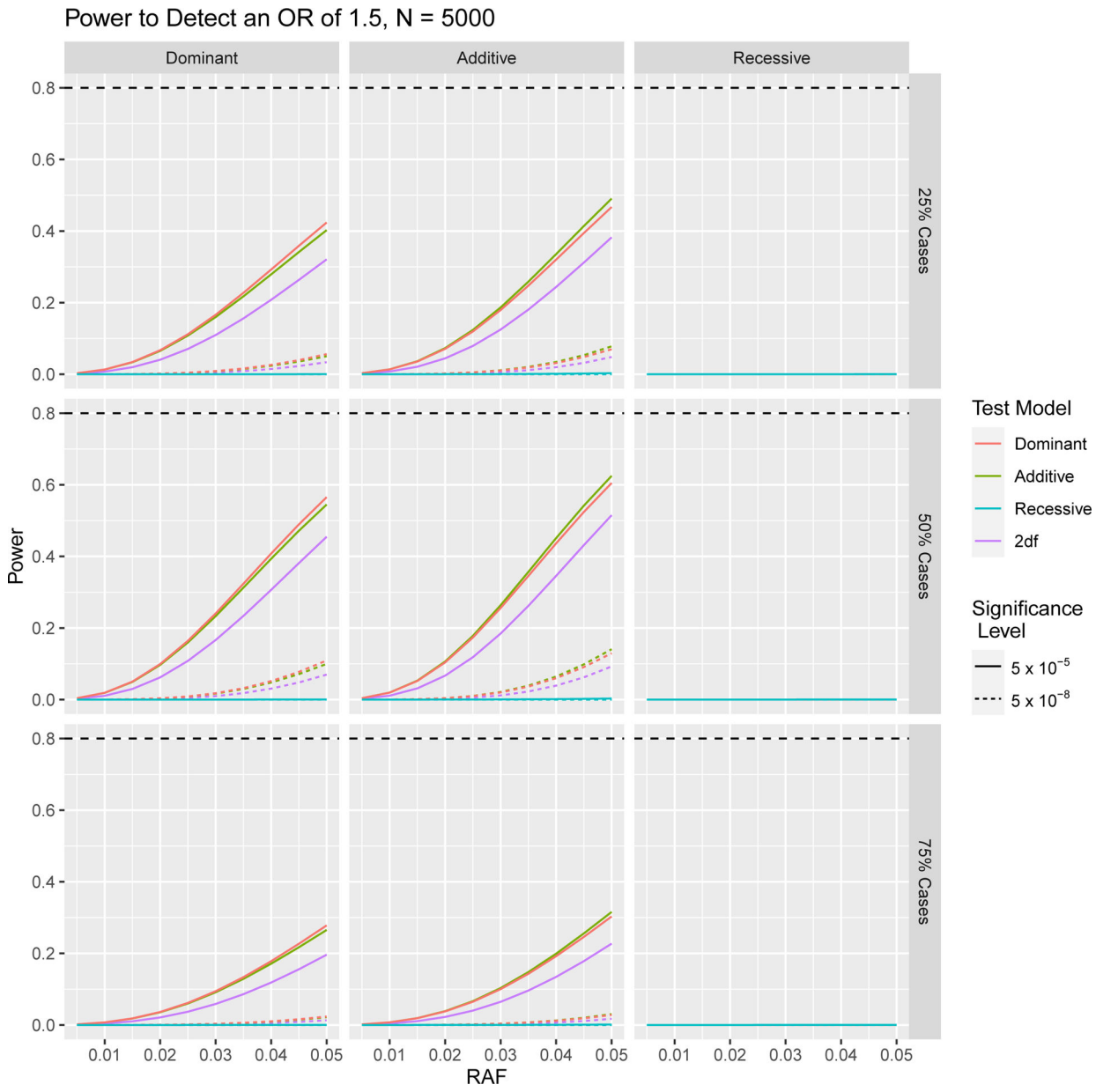

Fig. 4.

Power to Detect an Odds Ratio of 1.5 in Case Control Study with a Total of 5,000 Subjects for Different Ratios of Cases to Controls, at 5×10^{-5} and 5×10^{-8} Significance Levels at Low RAFs. The horizontal dashed line indicates 80% power.

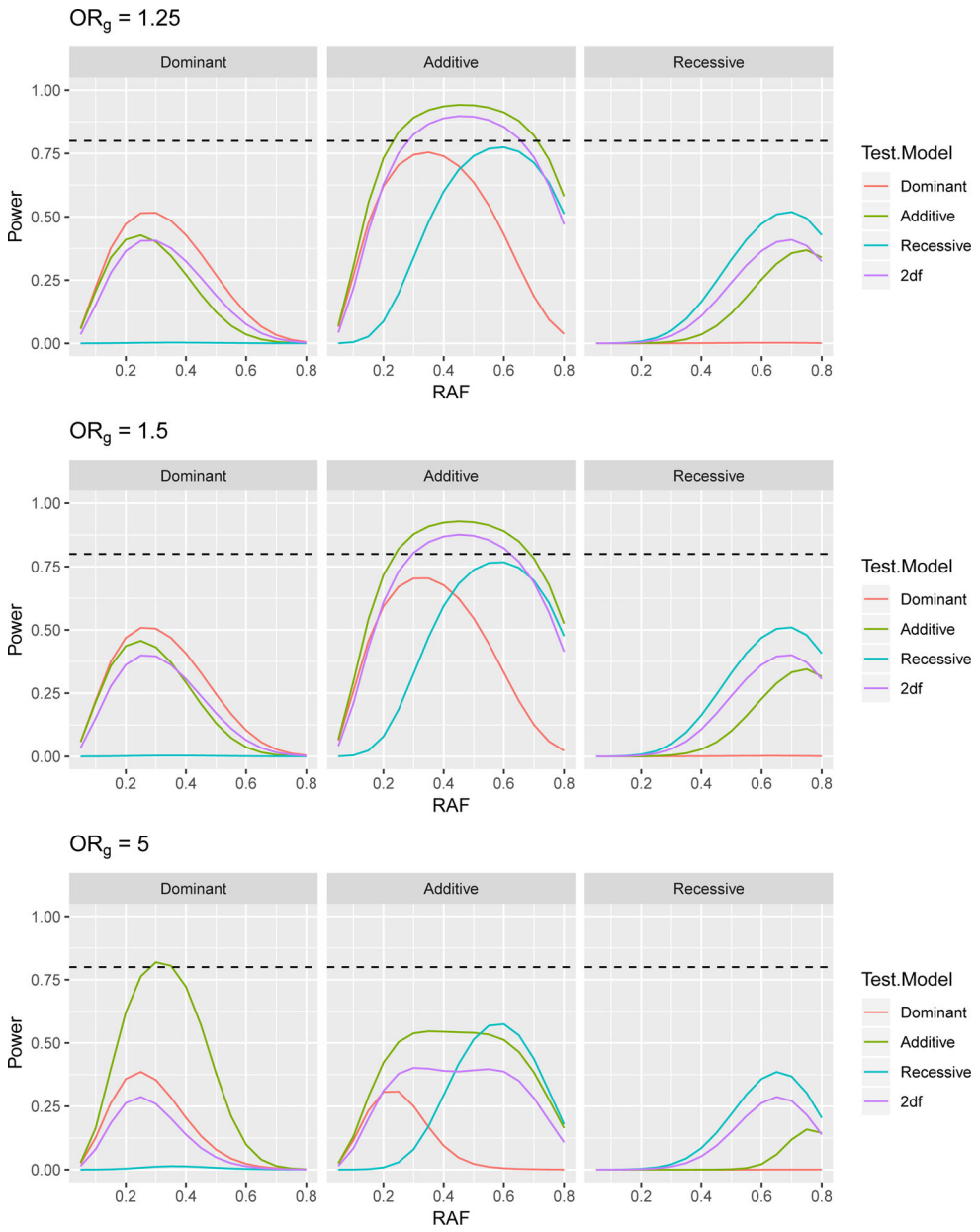

Fig. 5.

Power to Detect a Gene × Environment (G × E) Odds Ratio of 1.5 in a Case Control Study with 3,624 Cases and 4,442 Controls, at 5×10^{-5} Significance Level, Assuming a Genetics Odds Ratio of OR_g when E=0. The dashed line indicates 80% power.

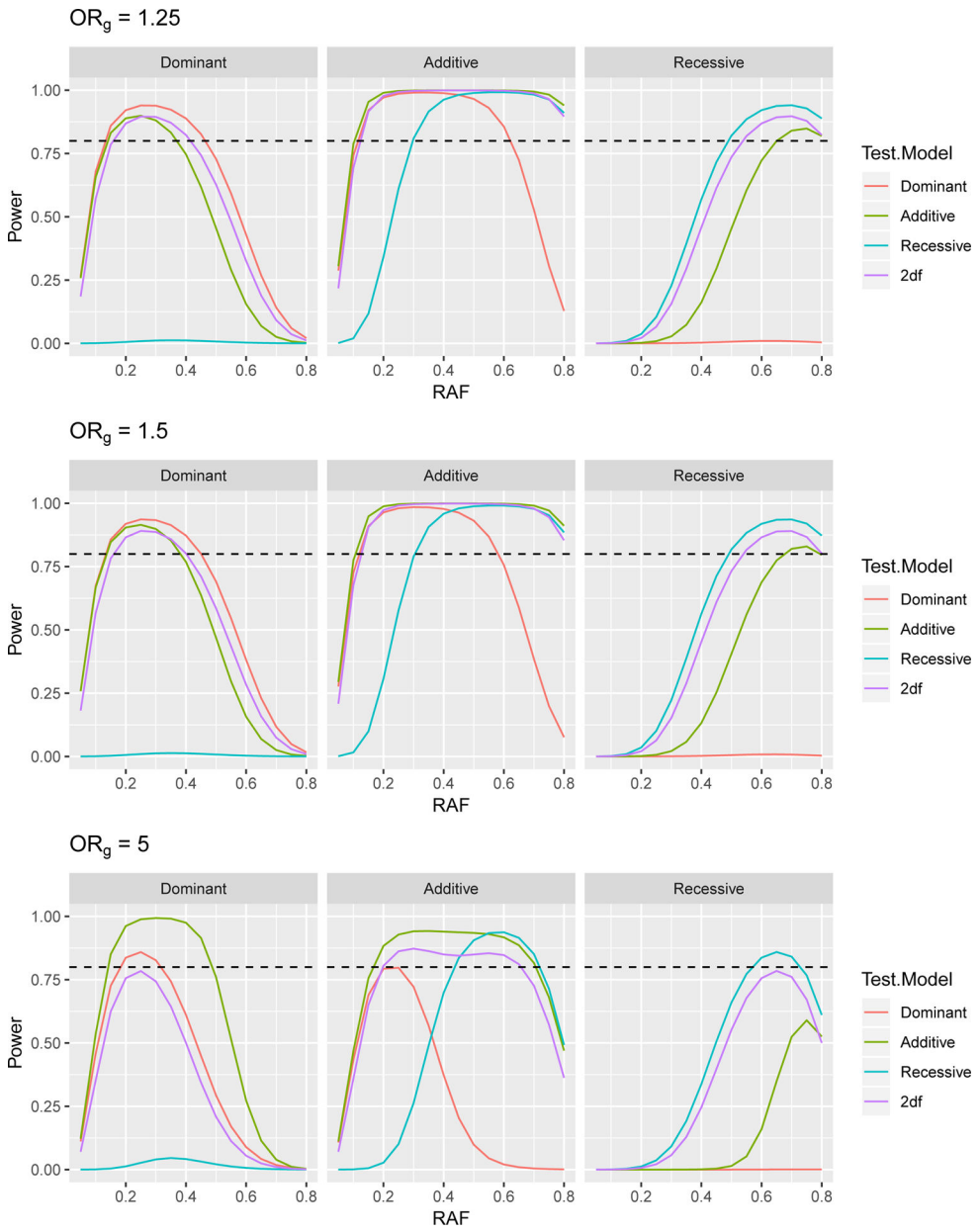

Fig. 6.

Power to Detect a Gene × Environment (G × E) Odds Ratio of 1.75 in a Case Control Study with 3,624 Cases and 4,442 Controls, at 5×10^{-5} Significance Level, Assuming a Genetics Odds Ratio of OR_g when E=0. The dashed line indicates 80% power.

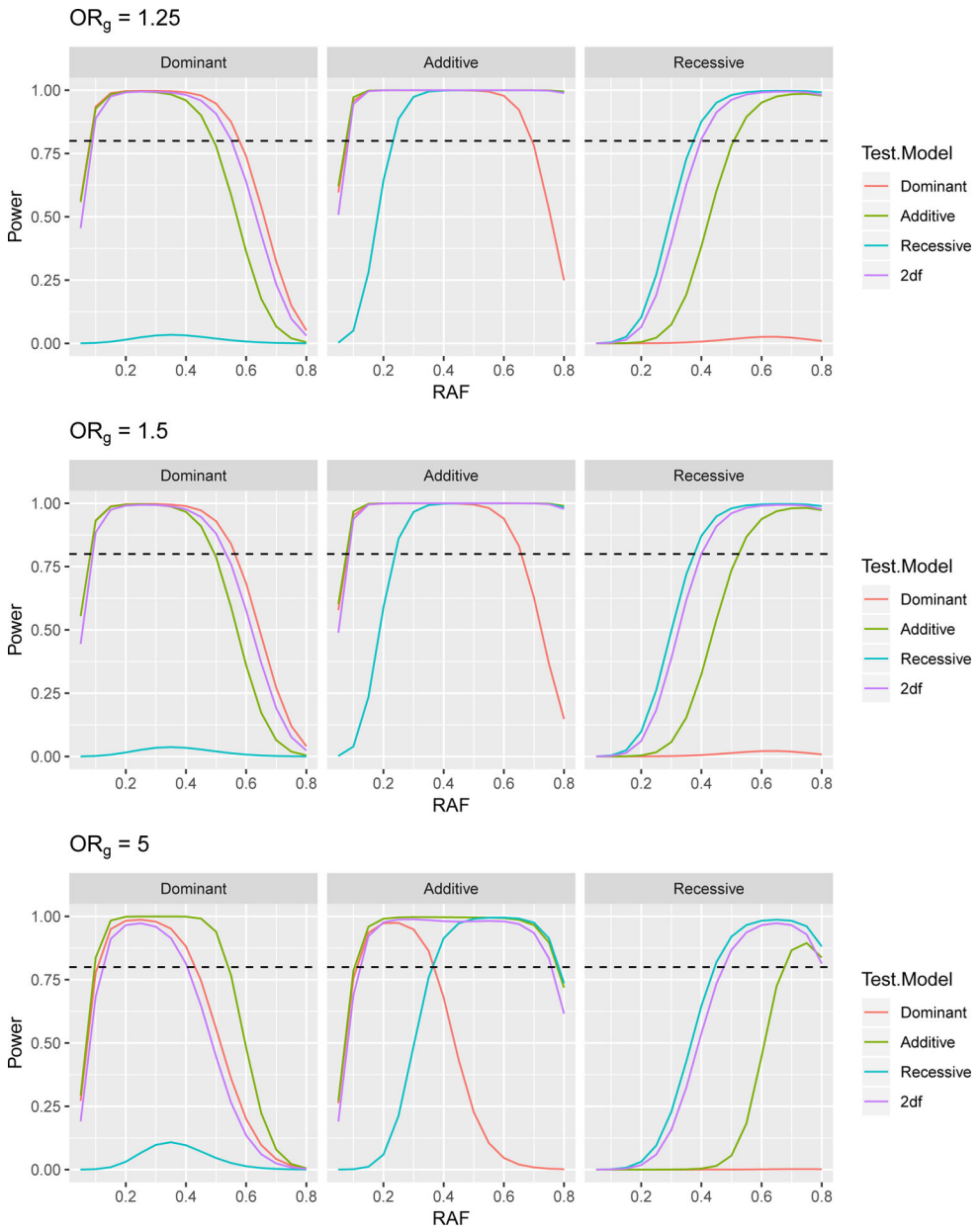

Fig. 7.

Power to Detect a Gene × Environment (G × E) Odds Ratio of 2 in a Case Control Study with 3,624 Cases and 4,442 Controls, at 5×10^{-5} Significance Level, Assuming a Genetics Odds Ratio of OR_g when E=0. The dashed line indicates 80% power.

Table 1:

Coding of Genetic Covariates for Inclusion in Regression Models

Model	Genotype	X_1	X_2
Dominant	aa	0	-
	aA	1	-
	AA	1	-
Additive	aa	0	-
	aA	1	-
	AA	2	-
Recessive	aa	0	-
	aA	0	-
	AA	1	-
2df	aa	0	0
	aA	1	0
	AA	0	1

Table 2:

Smallest Risk Allele Frequency (RAF) with at Least 80% Power to Detect an Odds Ratio of 1.5 for a 1:1 Case Control Study with a Total of N Subjects, at 5×10^{-8} Significance Level

		True Mode of Genetic Susceptibility		
N	**Model Coding**	**Dominant**	**Additive**	**Recessive**
5,000	Dominant	0.145	0.120	-
	Additive	0.170	0.115	0.655
	Recessive	-	0.310	0.515
	2df	0.165	0.125	0.550
20,000	Dominant	0.030	0.030	-
	Additive	0.030	0.030	0.335
	Recessive	-	0.130	0.230
	2df	0.030	0.030	0.240

Table 3:

Detectable Odds Ratios for a 1:1 Case Control Study with a Total of N Subjects, at 80% Power and 5×10^{-8} Significance Level

			True Mode of Genetic Susceptibility		
N	q_A	**Model Coding**	**Dominant**	**Additive**	**Recessive**
			Detectable OR		
5,000	0.1	Dominant	1.58	1.55	>100
		Additive	1.60	1.53	>100
		Recessive	15.10	3.59	10.39
		2DF	1.61	1.56	12.53
	0.3	Dominant	1.43	1.36	27.85
		Additive	1.48	1.32	2.63
		Recessive	3.26	1.51	1.88
		2DF	1.45	1.33	1.94
20,000	0.1	Dominant	1.26	1.24	>100
		Additive	1.26	1.23	>100
		Recessive	3.12	1.68	2.56
		2DF	1.27	1.25	2.68
	0.3	Dominant	1.19	1.16	2.95
		Additive	1.22	1.15	1.59
		Recessive	1.79	1.23	1.37
		2DF	1.20	1.15	1.39