Supplement to “U-PASS: unified power analysis and forensics for qualitative traits in gentic association studies”

Zheng Gao*[1], Jonathan Terhorst[1], Cristopher Van Hout[2], and Stilian Stoev[1]

[1]Department of Statistics, University of Michigan, Ann Arbor, MI 48109, USA
[2]Regeneron Genetics Center, Tarrytown, NY 10591 USA.

August 6, 2019

Abstract

We provide a brief overview of power analysis for genetic association studies in Section 1. The relationship between two approaches to specifying the alternative hypothesis in power calculations for association studies is clarified in Section 2. We develop a unified analysis of statistical power for common association tests in Section 4, and provide a simple rule for finite-sample correction of the asymptotic approximations in Section 5. The performance of this correction is examined through numerical experiments in Section 7. Proof of the main result is found in Section 6.

Contents

*Corresponding author: gaozheng@umich.edu

1 Overview of power analysis in genetic association studies

In power analysis of genetic association studies, disease models are used to specify the distribution of observations under the alternative hypothesis. In this document, we discuss the challenges in disease model specifications, and discuss its relationship with an alternative approach that accomplishes the same goal through a set of ”cannonical parameters”.

We shall first briefly recall the main steps of a typical power analysis for genetic association studies.

1.1 Disease model specifications

A typical power analysis for genetic association studies begin by specifiying an alternative hypothesis through a disease model (dominant, recessive, multiplicative, additive, etc.), which assumes:

- The genotype relative risks (GRR).
- Risk allele frequency in the general population (p).
- Disease prevalence in the general population (Prev).

The disease model and parameters determine the joint distribution of the genotypes and phenotyes *in the population*, shown in the following table.

	Risk allele copies		
Population Prob.	0 copies	1 copy	2 copies
Cases	π_{10}	π_{11}	π_{12}
Controls	π_{20}	π_{21}	π_{22}

In the disease models, the conditional probabilities of having the disease, given the risk allele copy numbers, satisfy the following relations,

$$\frac{\pi_{10}}{\pi_{10}+\pi_{20}} : \frac{\pi_{11}}{\pi_{11}+\pi_{21}} : \frac{\pi_{12}}{\pi_{12}+\pi_{22}} = \begin{cases} 1 : \text{GRR} : \text{GRR}^2, & \text{Multiplicative} \\ 1 : \text{GRR} : 2 \times \text{GRR} - 1, & \text{Additive} \\ 1 : \text{GRR} : \text{GRR}, & \text{Dominant} \\ 1 : 1 : \text{GRR}, & \text{Recessive} \end{cases} \tag{1}$$

where GRR is strictly greater than 1 under the alternative, and equal to 1 under the null hypothesis.

The disease prevalence determines the sum of the probabilities of cases in the population,

$$\pi_{10} + \pi_{11} + \pi_{12} = \text{Prev.} \tag{2}$$

The risk allele frequency in the general population, p, assuming Hardy-Weinberg equilibrium, satisfies

$$\pi_{10} + \pi_{20} = (1-p)^2, \quad \pi_{11} + \pi_{21} = 2p(1-p), \quad \pi_{12} + \pi_{22} = p^2. \tag{3}$$

The population probabilities are determined by the disease model and its parameters (GRR, Prev, and p). The six unknowns $(\pi_{10}, \ldots, \pi_{22})$ and are solved for using the six equations above: two from Relation (1), one from (2), and three from (3).

1.2 Sampling adjustments

Next, the probabilities of observing each genotype-phenotype combination are adjusted according the number of cases and controls recruited *in the studies*, where the sample sizes are specified with

1. The number of cases (n_1) and controls (n_2), or equivalently, the fraction of cases (ϕ) and total number of subjects (n).

	Risk allele copies		
Prob. in study	0 copies	1 copy	2 copies
Cases	$\pi_{10}\frac{\phi}{\text{Prev}}$	$\pi_{11}\frac{\phi}{\text{Prev}}$	$\pi_{12}\frac{\phi}{\text{Prev}}$
Controls	$\pi_{20}\frac{1-\phi}{1-\text{Prev}}$	$\pi_{21}\frac{1-\phi}{1-\text{Prev}}$	$\pi_{22}\frac{1-\phi}{1-\text{Prev}}$

As an example, if $\phi > \text{Prev}$, the probabilities are adjusted to account for over-sampling of cases.

The relative frequencies of allele type-phenotype combinations *in the study* are then calculated as follows.

	Allele variant	
Prob. in study	Risk allele	Non-risk allele
Cases	$\phi\left(\frac{\pi_{12}}{\text{Prev}}+\frac{\pi_{11}}{2\times\text{Prev}}\right)$	$\phi\left(\frac{\pi_{11}}{2\times\text{Prev}}+\frac{\pi_{10}}{\text{Prev}}\right)$
Controls	$(1-\phi)\left(\frac{\pi_{22}}{1-\text{Prev}}+\frac{\pi_{21}}{2(1-\text{Prev})}\right)$	$(1-\phi)\left(\frac{\pi_{21}}{2(1-\text{Prev})}+\frac{\pi_{20}}{1-\text{Prev}}\right)$

To simplify notation, we denote the relative frequencies of allele type-phenotype combinations with $\mu = (\mu_{11}, \mu_{12}, \mu_{21}, \mu_{22})$.

	Allele variant		
Prob. in study	Risk allele	Non-risk allele	Total by phenotype
Cases	μ_{11}	μ_{12}	$\phi = \mu_{11} + \mu_{12}$
Controls	μ_{21}	μ_{22}	$1 - \phi = \mu_{21} + \mu_{22}$

1.3 Power calculations of association tests

Finally, the power of an statistical test is calculated as the probability of (a correct) rejection, assuming that the data (i.e., tabulated counts of the allele type-phenotype combinations) follow a multinomial distribution with probability matrix μ and sample size $2n$, since each individual has a pair of alleles.

Some common association tests include the likelihood ratio test, Pearson's chi-square test, tests of zero slope coefficient in logistic regressions, as well as t-tests for equal proportions. In principal, power analysis has to be tailored to the association test used. Fortunately, many of these tests are asymptotically equivalent in terms of power, and results of the power analysis applies to all equivalent tests; see Section 4.

These steps form the basis of the calculations implemented in the most existing tools, including the GAS calculator (Johnson *et al.*, 2017). Although not explicitly stated, the GAS calculator assumes the test of association to be the Welch t-test.

2 Two approaches to specifying the alternative hypothesis

In the power calculations outlined above, the disease models are used to describe the distribution of the data under the alternative hypothesis. Specifically, they are used to specify the conditional distributions of the allele variants, given the phenotypes. The probability of observing a risk allele in the control group,

$$f := \mathbb{P}[\,\text{risk allele}\,|\,\text{control group}\,] = \frac{\mu_{21}}{1-\phi} = \frac{\pi_{22}+\pi_{21}/2}{\pi_{22}+\pi_{21}+\pi_{20}}, \tag{4}$$

is fully determined by the disease model. Similarly, the odds ratio between the two allele variants,

$$R := \frac{\mu_{11}\mu_{22}}{\mu_{12}\mu_{21}} = \frac{(\pi_{12}+\pi_{11}/2)(\pi_{20}+\pi_{21}/2)}{(\pi_{10}+\pi_{11}/2)(\pi_{22}+\pi_{21}/2)}, \tag{5}$$

is also determined by the disease model and its parameters.

In turn, the parameters (f, R), together with the sample sizes (ϕ, n), fully describe the distribution of our data under the alternative hypothesis (by determining the probability vector μ and the sample size n). Power of association tests, therefore, depends on (and only on) the set of "canonical parameters":

- Risk allele frequency among the controls (f).
- Odds ratio (R) of having the defined trait between the two allele variants.
- One of the two equivalent ways of parametrizing the sample sizes.

We illustrates the common process of power analysis implemented in existing power calculators in Figure 1.

Our discussion above shows that in power calculations, *we can either describe the alternative hypothesis with a disease model, or through the canonical parameters (f, R). Both approaches are sufficient for the purpose of power analysis.*

We remind readers that the risk allele frequency in the control group (f) is not equivalent to the risk allele frequency in the general population (p); odds ratio (R) is not equivalent to genotype relative risk (GRR).

2.1 Conversion between the two specifications

Notice that, as illustrated in Figure 1, the power calculations are mediated through the canonical parameters, which are invariant to different model specifications. That is, different disease model specifications may lead to the *same* set of canonical parameters (f, R),

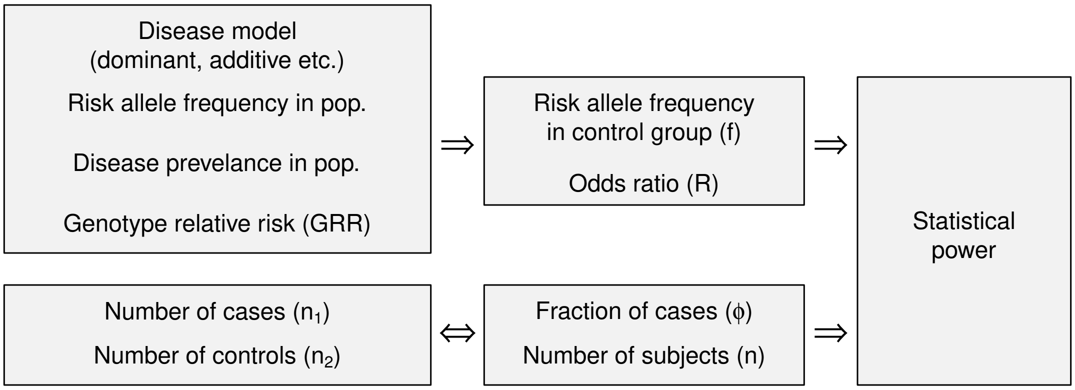

Figure 1: The process of a typical power analysis for genetic association tests. The quantities depend on, and can be calculated from the values of their parents in the directed graph. Power can be calculated as long as one set of parameters in each branch is known. While there is a one-to-one correspondence between the sample size specifications (n_1, n_2) and (ϕ, n), the mapping from disease model specifications to (f, R) is many-to-one.

and consequently the) *same* distributions of the allele variant counts. From a statistical perspective, the disease models that map to the same set of canonical parameters are equivalent in terms of power.

For example, the following set of disease models and parameters imply the same set of canonical parameters ($f = 0.290, R = 1.575$), and therefore enjoy the same power at the same sample sizes ($n_1 = n_2 = 1000$).

Disease Model	$(\text{Prev}, p, \text{GRR})$	(f, R)	Power
Multiplicative	$(0.1, 0.3, 1.500)$	$(0.290, 1.575)$	0.990
Additive	$(0.1, 0.3, 1.588)$	$(0.290, 1.575)$	0.990
Dominant	$(0.1, 0.3, 1.909)$	$(0.290, 1.575)$	0.990
Recessive	$(0.1, 0.3, 2.666)$	$(0.290, 1.575)$	0.990

Conversely, different disease models with the same parameters, map to drastically different canonical parameters. For example, the default disease model parameters in the GAS calculator,

$$\text{Disease prevalence in the population}: \quad \text{Prev} = 0.1 \tag{6}$$

$$\text{Risk allele frequency in the population}: \quad p = 0.5 \tag{7}$$

$$\text{Genotype relative risk}: \quad \text{GRR} = 1.5. \tag{8}$$

map to very different canonical parameters under different disease model assumptions (assuming the same sample sizes of $n_1 = n_2 = 1000$), which leads to drastically different statistical power.

Disease Model	$(\text{Prev}, p, \text{GRR})$	(f, R)	Power
Multiplicative	$(0.1, 0.5, 1.5)$	$(0.489, 1.568)$	0.995
Additive	$(0.1, 0.5, 1.5)$	$(0.491, 1.453)$	0.920
Dominant	$(0.1, 0.5, 1.5)$	$(0.495, 1.224)$	0.282
Recessive	$(0.1, 0.5, 1.5)$	$(0.494, 1.281)$	0.098

In the application, we provide users with a "Disease model converter" that implements this many-to-one conversion from the disease model specifications to the canonical parameters.

2.2 Comparisons of the two approaches

While the disease models may carry additional insights into the biological process, the canonical parameters also have their unique advantages. We offer an incomplete list of comparisons of the two approaches, and discuss their usage in practice.

2.2.1 Interpretability and communicability

In general, geneticists and biostatisticians seem to agree that disease models are more interpretable. The concept of genotype relative risks, in particular, seems easier to reason about than odds ratios in the canonical parameters definition. Disease models also seem to be the de facto mode of model specification when performing power analysis for study planning and grant applications.

The "nonparametric" approach to model specification through the canonical parameters is somewhat lesser known to the statistical genetics community. The canonical parameters are typically estimated and reported as outcomes of the research, but not used as inputs to the power analysis for planning purposes.

2.2.2 Availability of parameter estimates

The canonical parameters f and R can be estimated from data collected in the study. They are also reported and curated in GWAS catalogs such as the NHGRI-EBI Catalog (MacArthur *et al.*, 2016).

On the other hand, accurate information regarding the disease model parameters can be more difficult to obtain, partly because some parameters in the disease models cannot be estimated from the association studies alone.

In particular, disease prevalence in population (Prev), as well as RAF in population (p), must be obtained from other studies or surveys targeting the general population; the association studies, unless matching the proportion of cases in the population vs in the study ($\phi = \text{Prev}$), cannot produce estimates without using external information. Genetic association studies rarely explicitly estimate the disease model and its parameters. In fact, we are not aware of a GWAS catalog that reports and curates the disease models and their estimated parameters.

This paucity of information on disease model parameters is not an issue if we are planning to study a trait for which we have little prior knowledge. In this case, the purpose of power analysis is to determine the range of models and parameters that lead to discovery of associations, given the study designs.

In contrast, in confirmatory / follow-up studies and systematic reviews, our main interest is in the statistical validity of the reported findings. Power analysis then serves to find efficient designs, and to validate the claims made. Knowledge obtained in prior studies (in the form of parameter estimates) are indeed necessary.

2.2.3 Robustness against model misspecification

Disease models are useful in as much as they help us understand the biology behind the data we observe. Unfortunately, like all models, they can be misspecified. For example, the following genotype relative risks,

$$\frac{\pi_{10}}{\pi_{10} + \pi_{20}} : \frac{\pi_{11}}{\pi_{11} + \pi_{21}} : \frac{\pi_{12}}{\pi_{12} + \pi_{22}} = 1 : 3 : 4,$$

does not follow any of the common disease models. In this case, different studies may come up with different disease models (say, Dominant and Additive), and of course, different parameter estimates.

Suppose a researcher wishes to perform a meta analysis or confirmatory experiment of the existing results, where the literature reports inconsistent estimates of disease models and parameters, he would a have a difficult time pooling the information from these different sources. And even when they are pooled, the resulting model usually does not fall in one of the familiar categories — there is no existing tool with which to perform power analysis. The researcher will likely have to forgo the information from one model, and use estimates from only the other.

On the other hand, the canonical parameters are invariant to the disease model choices, and accommodate models falling outside the usual categories. They can also be easily combined to produce pooled estimates. This universality allows us to perform power analysis in a unified fashion, regardless of the disease models assumed. This also paves the way for the "OR-RAF diagram", as well as systematic reviews of statistical validity of existing studies.

2.2.4 Robustness against human errors

The disparity in availability of parameter estimates we mentioned earlier can lead to unintended consequences, one of which is potentially incorrect usage of power calculators.

This issue, although minor, affects correctness of the results from power analysis.

Recall that the specification of a disease model requires as input the risk allele frequency (RAF) in the general population (p). The RAF reported in the NHGRI-EBI Catalog (MacArthur *et al.*, 2016), in contrast, refers to RAF in the control group (f). With RAF in population often unavailable, it is tempting to substitute the RAF in control group into the calculations. While the two quantities may be close when diseases prevalence and penetrance are low, their difference becomes non-negligible if either of the two conditions are violated, leading to grossly distorted results.

Performing power analysis with the canonical parameters is not guaranteed to prevent this human error, as mistake in the other direction could also happen. But perhaps it is more robust to such mistakes, since what is readily available matches with what is required as input.

2.2.5 Compatibility of parameters

We make another minor comment regarding correct usage of disease models.

We caution users that not all values of the disease model parameter combinations are valid. For example, in a multiplicative model, the parameters

$$p = 0.1, \quad \text{Prev} = 0.5, \quad \text{GRR} = 1.5,$$

would result in the conditional probability of having two risk allele copies greater than 1. (In this case, the GAS calculator (Johnson *et al.*, 2017) would produce the error message: "I don't like the genetic model you requested!", without explicitly pointing to the compatibility issue.)

Although an experienced geneticist would immediately notice the impossibility of the disease model parameter combinations, these contradictions may not be obvious to the untrained eye. The end user of the software – experienced or not – is ultimately responsible for inputting valid values when specifying a disease model.

On the other hand, any combination of

$$f \in (0, 1), \quad \text{and} \quad R \in (0, +\infty)$$

is valid. Parameter compatibility is not an problem for the set of canonical parameters.

2.3 Recommendations on model specification in power analysis

Since both the disease models and the canonical parameters are sufficient for the purpose of power analysis, a natural question then arises: why (and when) should one take the canonical parameters approach, given that the more familiar disease models would also suffice?

We believe that either approach may be preferred, depending on the use cases. Recall that power analysis is useful in at least three scenarios:

1. *In planning for an exploratory study, where little is known about the associations.*

   In this case, the top branch in Figure 1 is unknown to the researcher. The goal is to find out the range of disease models and parameters that are discoverable given the study designs. Power analysis is also to some extent exploratory in nature.

2. *In planning for a confirmatory study, where something is known about the associations and one wishes to validate the findings with an efficient design.*

   In this case, the top branch in Figure 1 is known, and the variables in the bottom branch is what we are solving for. The goal of power analysis is to provide a set of efficient study designs with sufficient power.

3. *In reviewing the reported findings and verify the statistical validity of the claims made.*

   In the third case, one looks to find out whether the claims of statistical significance are congruent with the evidence from data. A claim supported by very weak or contradictory evidence should invite further investigations. In this case, both branches in Figure 1 have to be available.

In view of the discussions above and in Section 2.2, we propose the following general guideline for power analysis in genetic association studies.

- When designing an association study where little to no prior information is available:
  - Either approach is valid.
  - Disease models are easier to interpret and communicate.
- When designing a follow-up / confirmatory study, or conducting a systematic review:
  - Choose the approach for which the parameter estimates are available, or of better quality.
  - Typically the canonical parameters are better estimated, reported, and curated.

There are, of course, exceptions to these rules. The minor comments we made about the two approaches should not be taken as criticisms, but rather as reminders of the potential pitfalls in power analysis. In either approach, care needs to be exercised in order to produce valid results.

3 Use case illustrations

We detail each of the three functionalities of the application, and illustrate with examples.

3.1 Reviewing reported findings in the GWAS catalog

The "OR-RAF power diagram" tab of the application provides a tool for reviewing reported associations from existing studies. The application calculates statistical power based the core parameters common to models of qualitative traits:

- Sample sizes, i.e., the number of cases and controls i.e., (n_1, n_2) or (ϕ, n).
- The canonical parameters (f, R).

Users need only prescribe the sample sizes, by one of two ways provided in the first box, i.e., total sample size + fraction of cases, or number of cases + number of controls.

Statistical power of familiar association tests, including the likelihood ratio test, chi-square test, Welch's t-test, and LR test for logistic regressions, have the same asymptotic power curves (see the Section 4 for details). This common power limit is calculated as a function of RAF and OR, and visualized as a heatmap in the OR-RAF diagram.

3.1.1 Adaptive and interactive display for exploring GWAS catalogs

We provide options for users to load and overlay findings reported in the NHGRI-EBI GWAS Catalog (MacArthur *et al.*, 2016), or upload data from other sources compliant with the Catalog's data format.

The initial sample sizes are dynamically adjusted, and automatically determined from texts of the article reporting the user selected loci. Since the sampling structures are many and varied across different studies, and no uniform reporting format is enforced in the catalog, the initial sample sizes are best estimates from the extracted texts.

Information of the selected loci and the study is also dynamically displayed below the diagram.

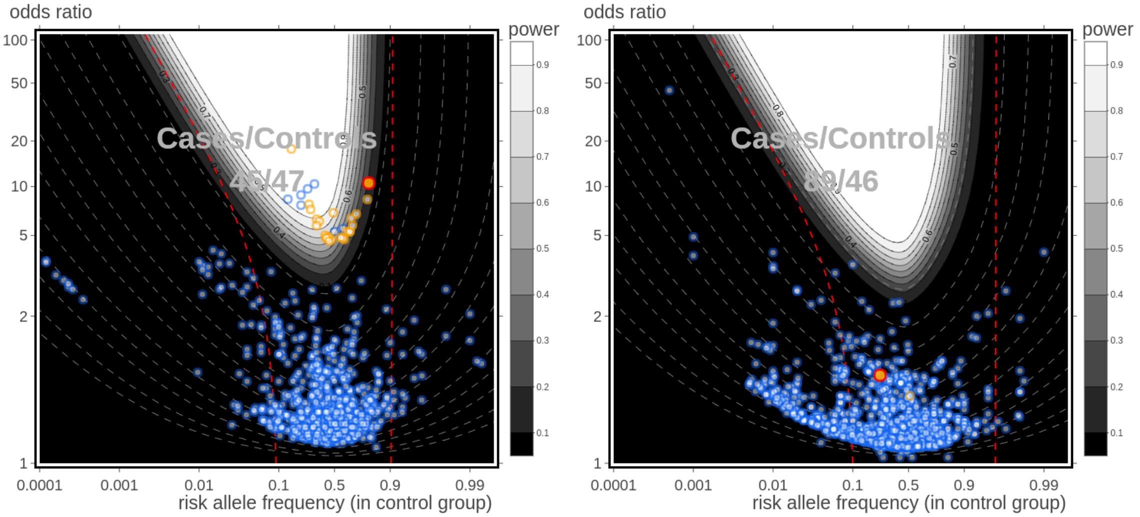

Figure 2: Two studies where gross misalignment was identified. Left: Dominguez-Cruz *et al.* (2018), right: Haryono *et al.* (2015). We reached out to the authors of the study in the left panel (Dominguez-Cruz *et al.*, 2018), who confirmed that this is the result of a problem in the data curation process of the GWAS Catalog (Dominguez-Cruz, personal communication). In particular, the RAF reported in the Catalog were based on all subjects in the study, as opposed to only the control group, while the Catalog requires that RAF be reported in the control group only. As a consequence, the RAFs are systematically overestimated, shifting the reported findings to the right in the diagram. The study on the right, though may very well be valid, calls for further scrutiny of its statistical methodologies given the apparent incongruity of its conclusions at the reported sample sizes.

3.1.2 An example: forensics of existing studies

The unified power analysis allows us to examine results from different studies employing different models and applicable tests, in the same diagram, with the same power limits. It allows for a systematic review of reported findings for their statistical validity.

In particular, a reported association predicted to have low power given the study's sample size – lying in the dark regions of the OR-RAF diagram – while not impossible, invites further scrutiny.

It should be noted that a reported association predicted to have high power is not automatically accurate, as survival bias induced by multiple testing may inflate the reported OR and RAF estimates.

Studies where reported associations show misalignment with the predicted powered curves may be further investigated for potential problems in the data curation process. The following figure shows one such study, where gross misalignment was identified.

In general, we expect this aspect of our software to be useful for discovering problems with data entry and catalog curation process, as well as for assessing the reproducibility and robustness of reported findings.

3.2 Designing association studies

The “Design my studies” tab of the application provides a tool for finding optimal designs of association studies. The tool requires inputs in a four-step process.

1. Model specification.
2. Sample size constraints specification.
3. False discovery Criteria specification.
4. Target power specification.

Each of the steps can be specified in a number of alternative ways.

3.2.1 Model specification

We provide two three ways to describe the model for biological process of the disease or trait of interest.

As outlined in Section 2, the distribution of observations can be specified through the canonical parameters, risk allele frequency in the control group (f) and odds ratio (R). Estimates for these quantities in previous studies of the same trait can be found in GWAS catalogs such as the NHGRI-EBI Catalog. See Section 2 for their definitions.

Alternatively, users may opt to specify through the disease models, of which we implement the four most popular ones: additive, multiplicative, dominant, and recessive. See Section 1 for the definitions of the quantities involved in the disease models. We remind users the difference between the risk allele frequency in the control group (f) versus risk allele frequency in the general population (p); only the latter is used in the disease model specifications.

Advanced users may choose to use a more succinct “signal size per sample” option, which directly parametrizes the signal sizes (λ/n). Details can be found in Section 6.

3.2.2 Sample size specifications

The second step requires users input the sample size constraints of the study. The three available options are “Budget / total number of subjects”, “Number of cases”, and “Fraction of cases”. In the subsequent calculations, the selected and specified quantities are treated as fixed. With only one unknown parameter left in the flow of power calculations (recall the flowchart in Fig. 1), we calculate power as a function of the remaining specified parameter. In particular,

- If the constraint is total budget, power is shown as a function of the fraction of cases.
- If the constraint is number of cases, power is shown as a function of the number of controls.
- If the constraint is fraction of cases, power is shown as a function of the total number of subjects.

3.2.3 Type I and Type II errors specifications

The final two steps require as input the desired level of false discovery and false non-discovery control. Both specification can be done through the marginal levels, i.e., Type I error and Type II error, or alternatively, through the multiple testing-adjusted levels, i.e., family-wise error rate (FWER) and family-wise non-discovery rate (FWNR).

3.2.4 An example: designing prospective studies

A researcher wishes to find out how many controls are needed in order to detect an association between a risk variant described by a multiplicative model with parameters:

$$\mathrm{RGG} = 1.2, \quad p = 0.3, \quad \mathrm{Prev} = 0.1.$$

The study has recruited 1000 subjects in the case group, and is aiming for power of 80% with FWER controlled at 5% level adjusted for the multiplicity of 10^6 tests.

In the application, we input the disease model parameters in the first step. In the second step, we select the sample size constraint as "number of cases" and set to 1000. The third step, we selected FWER as the criteria, and set the appropriate levels and multiplicity; a p-value cut off ($0.05/10^6 = 5 \times 10^{-8}$) is automatically calculated and displayed. The final step, we choose "Type II error / (1-power)" as the target and select $1 - 80\% = 20\%$.

The result of the calculation shows that the targeted power cannot be achieved at the current number of cases, no matter how many controls are recruited. Therefore, the researcher should consider recruiting more subjects in the case group in order to in crease power. For example, if there are instead 4000 subjects in the case group, then we would need only roughly 4929 controls in order to achieve the desired level of power.

3.3 Converting disease models into canonical parametrization

The “Disease model converter” tab of the application provides a tool for converting disease models into their implied canonical parameters.

The converter implements the mapping from disease models to the canonical parameters as detailed in Section 2 and illustrated in Figure 1. The tool also allows users to copy the model parameters into the “Design my study” tab for power calculations. Several numerical examples, discussed in Section 2, are provided in the tool.

4 Unified asymptotic analysis of power

In genetic association studies, researchers have the freedom to choose their favorite statistical procedure to perform hypothesis tests with the data collected. In this section, we answer the natural question that arises: does the choice of statistical test influence the power of scientific discovery?

It turns out, perhaps unsurprisingly, that most common association tests have asymptotically equivalent power. We shall quantify this shared power limit, and provide practical formulas for power calculations.

4.1 A model-invariant parametrization

In association studies for qualitative traits, counts of subjects in each phenotype-genetic variant combination are tabulated in the form of a contingency table. For a 2-allele-variant-by-2-phenotype definition, we have the following table of counts.

	Genotype		
# Observations	Variant 1	Variant 2	Total by phenotype
Cases	O_{11}	O_{12}	n_1
Controls	O_{21}	O_{22}	n_2

Statistics are then calculated based on the counts, to test for associations between the genotypes and phenotypes, at levels adjusted for multiplicity. Performance of a test is measured in terms of power, i.e., probability of correct rejection under an alternative hypothesis.

As we have seen in Section 2, power analysis typically starts by assuming an alternative distribution, typically (though not necessarily) described by a disease model. Power of a test is approximated either based on large sample asymptotics, or by simulating the empirical distribution of the statistic under the alternative.

Recall the 2-by-2 multinomial distributions with probability matrix $\mu = (\mu_{ij})_{2\times 2}$,

	Allele type		
Prob. in study	Risk allele	Non-risk allele	Total by phenotype
Cases	μ_{11}	μ_{12}	$\phi = \mu_{11} + \mu_{12}$
Controls	μ_{21}	μ_{22}	$1 - \phi = \mu_{21} + \mu_{22}$

We may assume – by relabelling, and hence without loss of generality – that allele Variant 1 is positively associated with the Cases, and referred to as the risk allele/variant.

The multinomial probability matrix μ can be fully parametrized by the parameter triple:

- Fraction of Cases ϕ, i.e., marginal distribution of phenotypes.
- Conditional distribution of risk variant among Controls, i.e., risk allele frequency (RAF) in the Control group
$$f := \mu_{21}/(1-\phi). \tag{9}$$
- Odds ratio (OR) of the allele Variant 1 to Variant 2
$$\mathrm{R} := \frac{\mu_{11}}{\mu_{21}}\bigg/\frac{\mu_{12}}{\mu_{22}} = \frac{\mu_{11}\mu_{22}}{\mu_{12}\mu_{21}}. \tag{10}$$

An alternative hypothesis, e.g., a disease model, determines the canonical parameters (f, R) implicitly, and therefore fully determines statistical power for a specific test at given sample sizes. Alternatively, power can also be calculated by directly prescribing the canonical parameters and the sample sizes.

It is worth pointing out that while disease models play no role beyond specifying the alternative, they do sometimes inform our choice of a test statistic, hence influencing statistical power in higher order contingency tables. These tests include, e.g., the Cochran-Armitage test, and variations thereof; see González *et al.* (2008); Li *et al.* (2008) for further examples where tests are tailored to disease models.

We make the important distinction between RAF *in the Control group* (f), versus RAF *in the study* $(\mu_{11} + \mu_{21})$, and RAF *in the general population* (p). In the following sections, unless otherwise stated, RAF will refer to the risk allele frequency in the Control group, consistent with the reporting standards of the NHGRI-EBI Catalog (MacArthur *et al.*, 2016).

4.2 Conditional vs unconditional tests

Readers familiar with the underlying assumptions of association tests in contingency tables may have noticed that we have described a multinomial distribution of the cell counts. That is, we have only conditioned on the total number of observations in the study. This is indeed the assumption behind tests such as (the original, unconditional version of) the likelihood ratio test, and the Person chi-square test.

It is not, however, the assumption behind some other tests. For example, analysis of t-tests typically assumes the observed number in each arm of the study are given. That is, we would condition on the phenotype marginals when comparing the proportions of genetic variants among the Cases and Controls. It is perhaps an assumption most close to reality, where the number of samples collected in each arm of the study are pre-determined. In this case, we have two binomial observations, $\text{Binom}(n_1, p_1)$ and $\text{Binom}(n_2, p_2)$, instead of a multinomial observation.

	Allele type		
Cond. Prob.	Variant 1	Variant 2	Counts by phenotype
Cases	p_1	$1 - p_1$	n_1
Controls	p_2	$1 - p_2$	n_2

RAF and OR can be similarly defined,

- Marginal distribution of Cases, fixed at $\phi := n_1/(n_1 + n_2)$,
- Risk allele frequency (RAF) in the Control group is the synonymous with the conditional distribution of risk variant among Controls,

$$f := p_2,$$

- Odds ratio (OR)

$$\mathrm{R} := \frac{p_1\phi}{p_2(1-\phi)} \Big/ \frac{(1-p_1)\phi}{(1-p_2)(1-\phi)} = \frac{p_1(1-p_2)}{(1-p_1)p_2}.$$

Alternative hypotheses may be formed as in the multinomial case with parameters ϕ, f and R, along with the total samples size.

Finally, we mention the assumptions behind the Fisher's exact test. The Fisher exact test conditions on the number of observations of both the phenotype variants and genetic variants, leading to a hypergeometric distribution of the first cell count O_{11} given the marginals n_1, n_2, and $O_{11} + O_{21}$, under the null hypothesis. We found no easy parametrizations of alternative hypotheses under this framework. Indeed, existing power calculations for Fisher's exact test resort to simulations under the two-binomial assumptions (Smyth *et al.*, 2017).

We refer interested readers to the recent work by Ripamonti *et al.* (2017) and Choi *et al.* (2017) which elucidate the controversies regarding the choices of conditioning when performing statistical inferences on 2-by-2 contingency tables. We do not attempt to resolve the controversies in this work. Our goal is to state clearly the assumptions behind the tests, and show the asymptotic equivalence in terms of power, under their respective assumptions.

4.3 A test-independent power analysis

We now present the main result allowing a unified power analysis, applicable for a wide range of common association tests in 2-by-2 tables.

If we consider a fixed parameter values of (f, R) under the alternative, no matter how close to the null, the probability of rejection of the null hypothesis by any reasonable test should approach one as sample size increases ($n \to \infty$). On the other hand, the probability of rejection is less than one in finite samples, making this type of asymptotics useless for approximation.

Therefore, in order to find finer approximations of power, we study alternatives close to the null. In particular, we take a sequence of alternatives approaching a limit point in the null space, in the hope that limiting rejection probability is between 0 and 1. It turns out – see, e.g., Ferguson (2017) Chapter 10, and Lehmann (2004) Chapter 5 – that the appropriate rate at which the alternatives should shrink towards the limit point is $1/\sqrt{n}$.

Under the multinomial assumption, let $\mu^{(0)}$ be the probability matrix of the (independent) 2-by-2 multinomial distribution, with marginals $(\theta, 1-\theta)$ for the genetic variants, and marginals $(\phi, 1-\phi)$ for the phenotypes. We require that $\theta \in (0,1)$ and $\phi \in (0,1)$ be bounded away from 0 and 1. Let $\mu = \mu^{(n)}$ be the sequence of alternatives such that

$$\sqrt{n}(\mu^{(n)} - \mu^{(0)}) \to \delta \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}, \tag{11}$$

where δ is a positive constant.

Equivalently, under the two-binomial assumption, let $p_1^{(0)} = p_2^{(0)} = \theta$ be the null hypothesis, with fixed marginals $(\phi, 1-\phi)$ for the phenotypes. Let $(p_1, p_2) = p^{(n)}$ be the sequence of alternatives such that

$$\sqrt{n}\phi(p_1 - \theta) \to \delta \quad \text{and} \quad \sqrt{n}(1-\phi)(p_2 - \theta) \to -\delta, \tag{12}$$

where δ is a positive constant. It is easy to see that with the same δ the two sequences of alternatives have the same RAF and OR, and therefore have the same expected number of observations in each cell.

Theorem 1. *In 2-by-2 contingency tables, under the assumption that the counts in the contingency table follow the multinomial distributions.*

- *The likelihood ratio test for independence,*

- *the likelihood ratio test for zero slope in logistic regressions,*
- *Person's chi-squared test for Independence,*

and under the assumption that counts in the contingency table follow the two binomial distributions,

- *the two-sided Welch's t-test for equal proportions*

have the same asymptotic power curves. Specifically, for the sequence of alternatives defined in (11) *and* (12)*, all of the listed tests, at level α, have statistical powers converging to*

$$\mathbb{P}[\chi^2(\lambda) \geq q_\alpha], \tag{13}$$

where q_α is the upper α quantile of the central chi-square distribution, and $\chi^2(\lambda)$ is a non-central chi-square distribution with non-centrality parameter

$$\lambda = \delta^2/(\theta(1-\theta)\phi(1-\phi)). \tag{14}$$

The proof of Theorem 1 is detailed in Section 6 below.

Theorem 1 is the central result that paves the way for a unified power analysis. It allows us to chart findings from different studies employing the applicable tests in the same diagram, with the same power limits. In particular, for large samples, tests for zero slopes in logistic regressions should report approximately the same set of loci as Welch's t-tests for equal proportions on the same dataset, after the same family-wise error rate adjustments. The estimated odds ratios (in the case of logistic regression, estimate slopes exponentiated) and RAF's, when charted on the OR-RAF diagram, should also follow the same power limits.

To use this result for power calculations, we start with an alternative hypothesis, defined by the canonical parameters (f, R), and sample sizes (ϕ, n).

	Allele type	
Probabilities	Variant 1	Variant 2
Cases	$fR\phi/(fR+1-f)$	$(1-f)\phi/(fR+1-f)$
Controls	$f(1-\phi)$	$(1-f)(1-\phi)$

Elementary algebra yields $\theta = fR\phi/(fR+1-f) + f(1-\phi)$, and $\delta = \sqrt{n}(\theta - f)(1-\phi)$. If tests are based on allele type counts, accounting for the fact that each genetic location has a pair of alleles, the effective sample sizes should be doubled, and the appropriate non-centrality parameter becomes $\delta = \sqrt{2n}(\theta - f)(1-\phi)$. Power may then be approximated using the formula in (13).

5 Finite-sample corrections

The result of our power calculations above are only accurate to the extend that the asymptotic approximations are applicable. In practice, of course, we have only finite samples, and the asymptotic approximations no longer hold when cell counts are low. While existing tools have completely ignored this issue, we offer here a simple correction in finite samples by resorting to exact tests.

Specifically, we calculate the minimum number of observations of the genetic variants needed for Fisher's exact test to be correctly calibrated, referred to as the **minimum calibration numbers**. As we shall see in simulations in Section 7, they provide a useful lower bound on the variant counts necessary for any asymptotic approximations to apply.

For a contingency table with marginal phenotype counts (n_1, n_2), and marginal genetic variant counts (m_1, m_2), we calculate the p-values of the most extreme observations according to Fisher's exact test. For rare risk alleles, this corresponds to the following table.

	Allele type		
# Observations	Variant 1	Variant 2	Counts by phenotype
Cases	m_1	$n_1 - m_1$	n_1
Controls	0	n_2	n_2

If the p-values do not fall below the desired type I error threshold, then the rejection region (for O_{11}) must lie beyond m_1. Under the fixed marginal assumptions of Fisher's exact test, no contingency tables with the given marginals can be rejected at the specified level. In other words, we have given up all power to achieve proper type I error control. Therefore, the minimum counts needed for the risk allele count must exceed m_1, in order for association tests to have any power.

For rare non-risk alleles, the most extreme observation corresponds to the following table.

	Allele type		
# Observations	Variant 1	Variant 2	Counts by phenotype
Cases	n_1	0	n_1
Controls	$n_2 - m_2$	m_2	n_2

We can similarly determine the minimum number of non-risk allele counts needed to achieve non-zero power at the given type I error target, for a given phenotype marginals (n_1, n_2).

For correctly calibrated tests, an alternative hypothesis with expected variant counts less than the minimum calibration numbers should have power close to zero; asymptotic power approximations do not apply for these alternatives. We correct the asymptotic approximations laid out in Section 4.3 by setting the predicted statistical power for alternatives in this "rare-variant zone" to zero.

In the web-based application U-PASS, we mark the "rare-variant zone" with red dashed lines in the OR-RAF diagram. We also provide options for users to specify the rare-variant threshold by absolute number of counts, or as a fraction of the total number of subjects.

6 Proof of Theorem 1

6.1 Asymptotic equivalence of likelihood ratio tests and the chi-square test

The asymptotic equivalence of the likelihood ratio (LR) test and the chi-square test in 2-by-2 tables can be found in standard texts on asymptotic theory. See, e.g., Ferguson (2017) Chapter 10 and Chapter 24, and Lehmann (2004) Chapter 5; see also, Hunter (2002) for an accessible derivation of the formula (14).

Recall the likelihood ratio statistic in LR test

$$LR = \frac{\sup_{\mu \in H_1} L(\mu)}{\sup_{\mu \in H_0} L(\mu)}.$$

where H_0 is the set of independent probabilities, and H_1 all valid 2-by-2 multinomial probabilities. To see the asymptotic equivalence with the LR statistic under logistic regressions with binary predictors, we reparametrize the likelihood as

$$\begin{aligned} L(\mu) &= \mu_{11}^{O_{11}} \mu_{12}^{O_{12}} \mu_{21}^{O_{21}} (1 - \mu_{11} - \mu_{12} - \mu_{21})^{O_{22}} \\ &= \phi^{n_1} (1-\phi)^{n_2} p_1^{O_{11}} (1-p_1)^{O_{12}} p_2^{O_{21}} (1-p_2)^{O_{22}} \end{aligned}$$

where we have omitted the multinomial coefficient. In the latter parametrization, it is easy to show that the maximizers are $\widehat{\phi} = n_1/n$, $\widehat{p_1} = O_{11}/n_1$, and $\widehat{p_1} = O_{21}/n_2$ under the alternative, and $\widehat{\phi} = n_1/n$, $\widehat{p_1} = \widehat{p_1} = (O_{11} + O_{21})/n$ under the null. Therefore, the terms involving ϕ cancels in the LR statistic, and the LR statistic coincides with the logistic regressions likelihood ratio, where p_1 and p_2 are further reparametrized as

$$\begin{aligned} p_1 &= \exp(\beta_0 + \beta_1)/(1 + \exp(\beta_0 + \beta_1)), \\ p_2 &= \exp(\beta_0)/(1 + \exp \beta_0). \end{aligned}$$

Hence, the logistic regressions likelihood ratio follows the same distribution as in the original likelihood ratio test. Notice that this is not an immediate consequence of the invariance property of the likelihood ratio tests. Rather, it follows because n_1, n_2 are ancillary for inference of the odds ratios.

6.2 Asymptotic equivalence with Welch's t-test

We now work with the two-binomial assumption, conditioning on the phenotype marginals n_1, n_2, and show that Welch's t-test has asymptotically the same power. Recall the Welch t-statistic

$$t = \frac{\widehat{p_1} - \widehat{p_2}}{\sqrt{\frac{\widehat{p_1}(1-\widehat{p_1})}{n_1} + \frac{\widehat{p_2}(1-\widehat{p_2})}{n_2}}}, \tag{15}$$

where $\widehat{p_1} = O_{11}/n_1$ and $\widehat{p_2} = O_{21}/n_2$. By the (Lindeberg-Feller) central limit theorem, for the sequence of alternatives defined in (12) we have

$$\sqrt{n_i}(\widehat{p_i} - p_i)\Big/\sqrt{p_i(1-p_i)} \Rightarrow \mathrm{N}(0,1), \quad \text{for } i = 1, 2.$$

and therefore, by independence of the two binomial distributions, we have

$$t - (p_1 - p_2)\Big/\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}} \Rightarrow \mathrm{N}(0,1). \tag{16}$$

By the definition of the alternatives in (12), we know that

$$\sqrt{n}(p_1 - p_2) \to \delta/(\phi(1-\phi)). \tag{17}$$

On the other hand, for the denominator, we have

$$\begin{aligned}
&\sqrt{n}\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}} \sim \sqrt{\frac{p_1(1-p_1)}{\phi} + \frac{p_2(1-p_2)}{1-\phi}} \\
&= \left(\frac{(\theta + O(1/\sqrt{n}))(1-\theta+O(1/\sqrt{n}))}{\phi} + \frac{(\theta + O(1/\sqrt{n}))(1-\theta+O(1/\sqrt{n}))}{1-\phi}\right)^{1/2} \\
&\sim \sqrt{(\theta(1-\theta))/(\phi(1-\phi))}.
\end{aligned} \tag{18}$$

Dividing (17) by (18), in view of (16), we conclude that the centers of the distribution of t converges to

$$\delta/\sqrt{\theta(1-\theta)\phi(1-\phi)},$$

which is precisely the square root of the non-centrality parameter in (14). Finally, the conclusion in Theorem 1 follows from the fact that the square of a normal distribution with mean $\sqrt{\lambda}$ is equal in distribution to a chi-square distribution with non-centrality parameter λ.

7 Numerical illustrations

We examine the accuracy of the asymptotic approximations in Theorem 1 in finite samples, and of the correction by minimum calibration number introduced in Section 5, via numerical simulation.

7.1 Power as a function of the sample sizes

We compare the theoretical predictions with empirical power of the following tests

- Fisher's exact test
- Welch's t-test
- Pearson's chi-square test
- Likelihood ratio test, and
- LR test with logistic regression.

Theses tests are performed on data simulated from the two-binomial model at following ranges of parameter values:

- Sample size total, n: $100, 150, 200, 300, \ldots, 900, 1000, 1500, 2000, \ldots$, up to 10^5,
- Fraction of cases in the study, ϕ: 15%, 25%, 35%, 50%, 85%,
- Risk allele frequencies in the control group, f: 0.01%, 0.1%, 1%, 10%, 50%, 90%,
- Odds ratio, R: 1.05, 1.1, 1.2, 1.5, 3.0.
- p-value cutoffs: 5×10^{-5}, 5×10^{-8}.

Each of the $41 \times 5 \times 6 \times 5 \times 2$ parameter value combinations was simulated 200 times. In the interest of space, only the results for p-value cutoff at 5×10^{-8}, and odds ratios $1.05, 1.1, 1.5$ are visualized. The figures are organized by increasing fraction of cases in the study ϕ, in Figures 3 to 7.

Each design (and test) were examined under small to moderate odds ratios (left to right panels), and under rare to common risk allele frequencies (top to bottom panels). The Fisher's exact test (black circles), Welch's t-test (red triangles), Pearson's chi-square test (blue crosses), Likelihood ratio test, and equivalently, LR test with logistic regression (light blue diamonds) are compared against the theoretical predictions (purple nabla).

The accuracy of the theoretical predictions will be commented in Section 7.3.

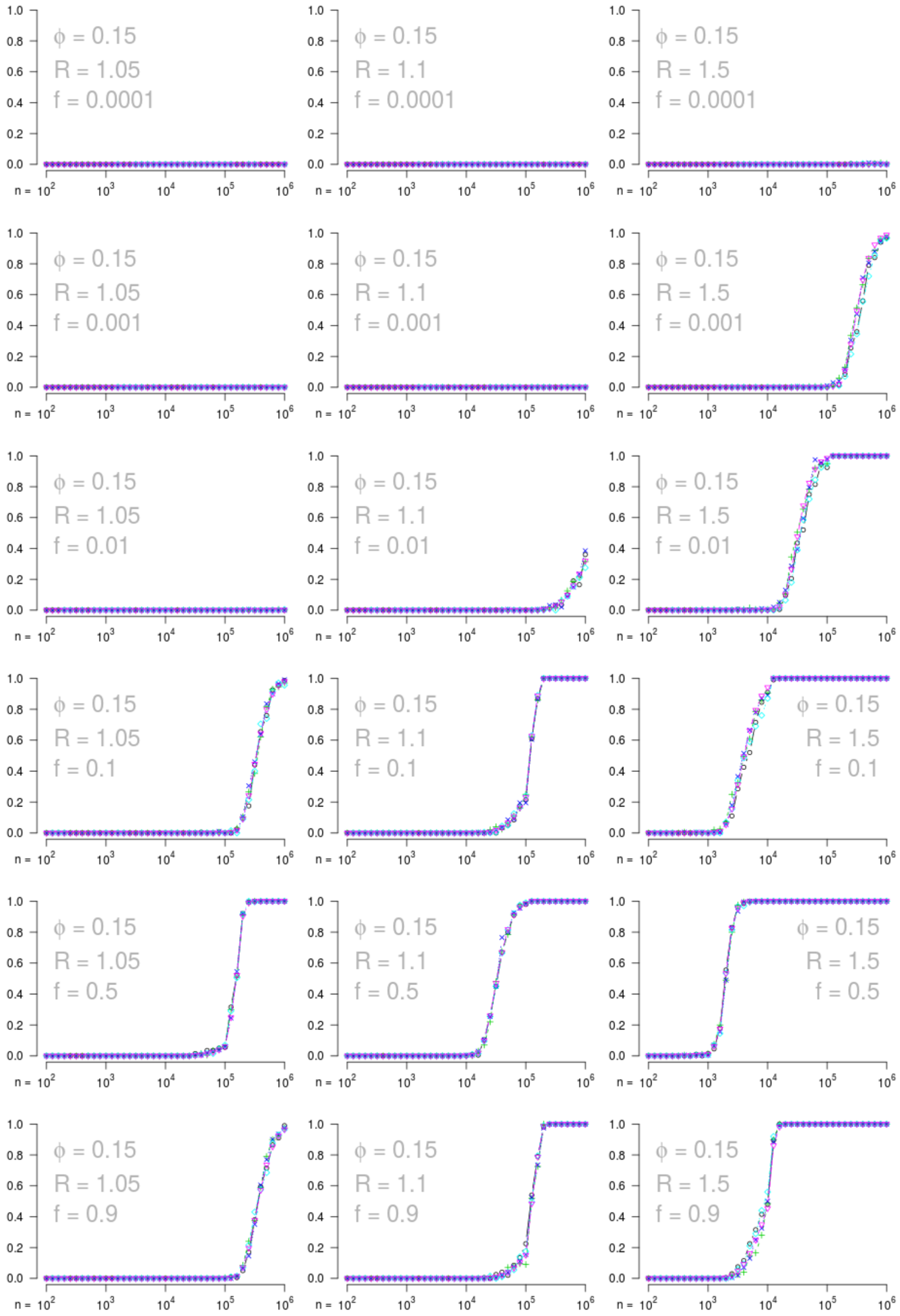

Figure 3: Empirical power as a function of sample sizes, under small to moderate odds ratios (left to right panels), and under rare to common risk allele frequencies (top to bottom panels). The Fisher's exact test (black circles), Welch's t-test (red triangles), Pearson's chi-square test (blue crosses), Likelihood ratio test, and equivalently, LR test with logistic regression (light blue diamonds) are almost identical to the theoretical predictions (purple nabla), at sample sizes ranging from 100 to 100,000. Fraction of cases $\phi = 15\%$.

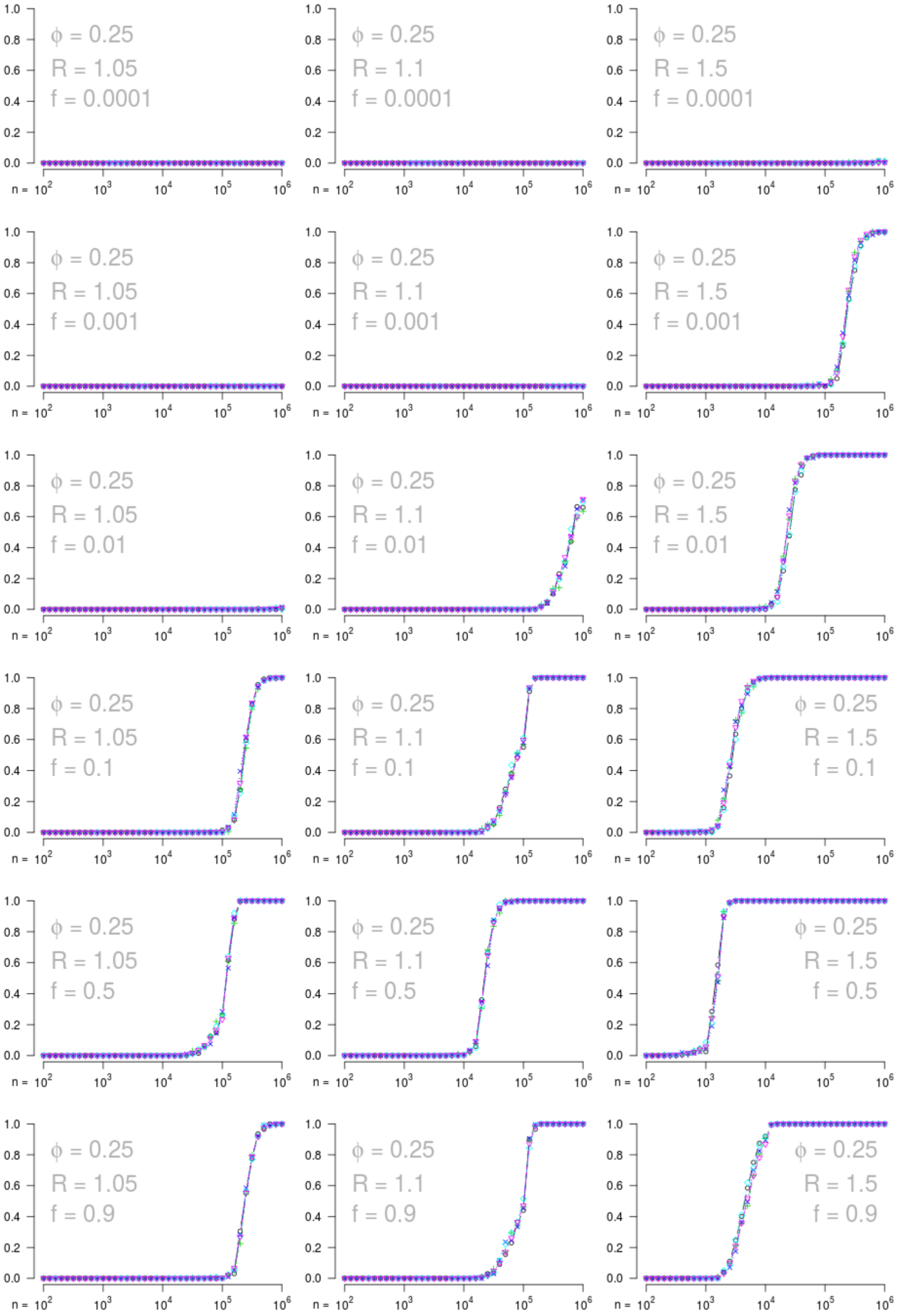

Figure 4: Empirical power as a function of sample sizes, under small to moderate odds ratios (left to right panels), and under rare to common risk allele frequencies (top to bottom panels). The Fisher's exact test (black circles), Welch's t-test (red triangles), Pearson's chi-square test (blue crosses), Likelihood ratio test, and equivalently, LR test with logistic regression (light blue diamonds) are almost identical to the theoretical predictions (purple nabla), at sample sizes ranging from 100 to 100,000. Fraction of cases $\phi = 25\%$.

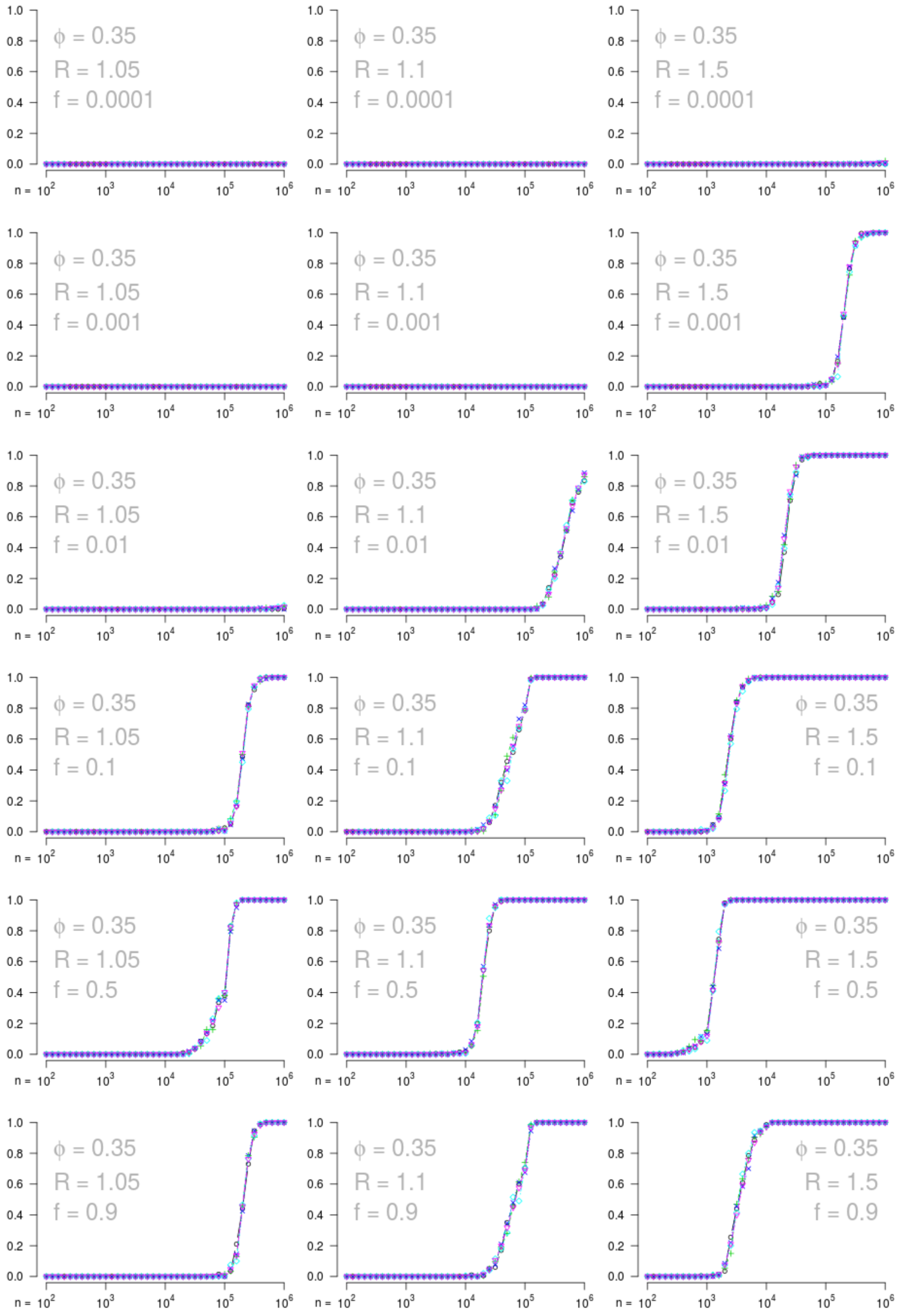

Figure 5: Empirical power as a function of sample sizes, under small to moderate odds ratios (left to right panels), and under rare to common risk allele frequencies (top to bottom panels). The Fisher's exact test (black circles), Welch's t-test (red triangles), Pearson's chi-square test (blue crosses), Likelihood ratio test, and equivalently, LR test with logistic regression (light blue diamonds) are almost identical to the theoretical predictions (purple nabla), at sample sizes ranging from 100 to 100,000. Fraction of cases $\phi = 35\%$.

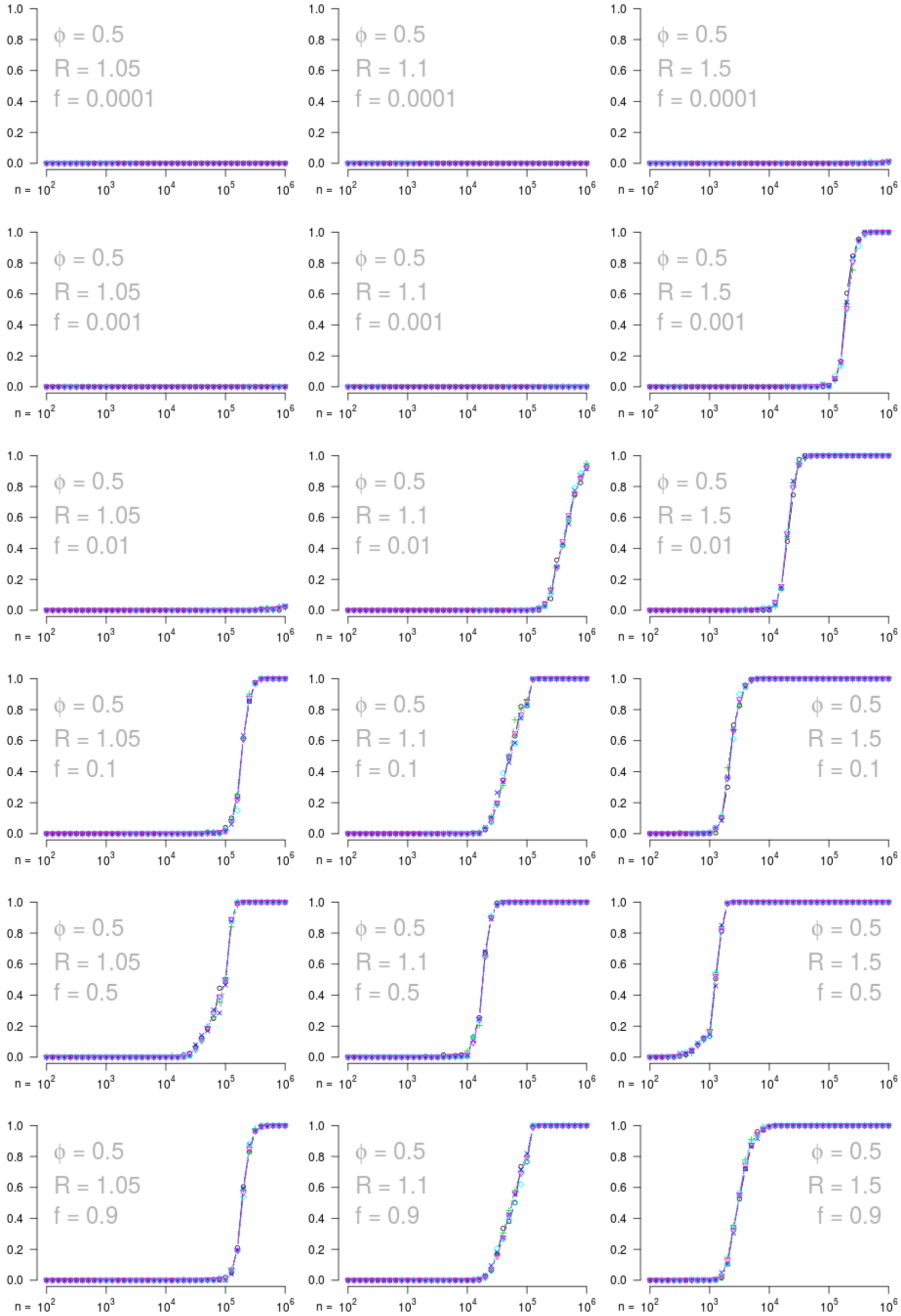

Figure 6: Empirical power as a function of sample sizes, under small to moderate odds ratios (left to right panels), and under rare to common risk allele frequencies (top to bottom panels). The Fisher's exact test (black circles), Welch's t-test (red triangles), Pearson's chi-square test (blue crosses), Likelihood ratio test, and equivalently, LR test with logistic regression (light blue diamonds) are almost identical to the theoretical predictions (purple nabla), at sample sizes ranging from 100 to 100,000. Fraction of cases $\phi = 50\%$.

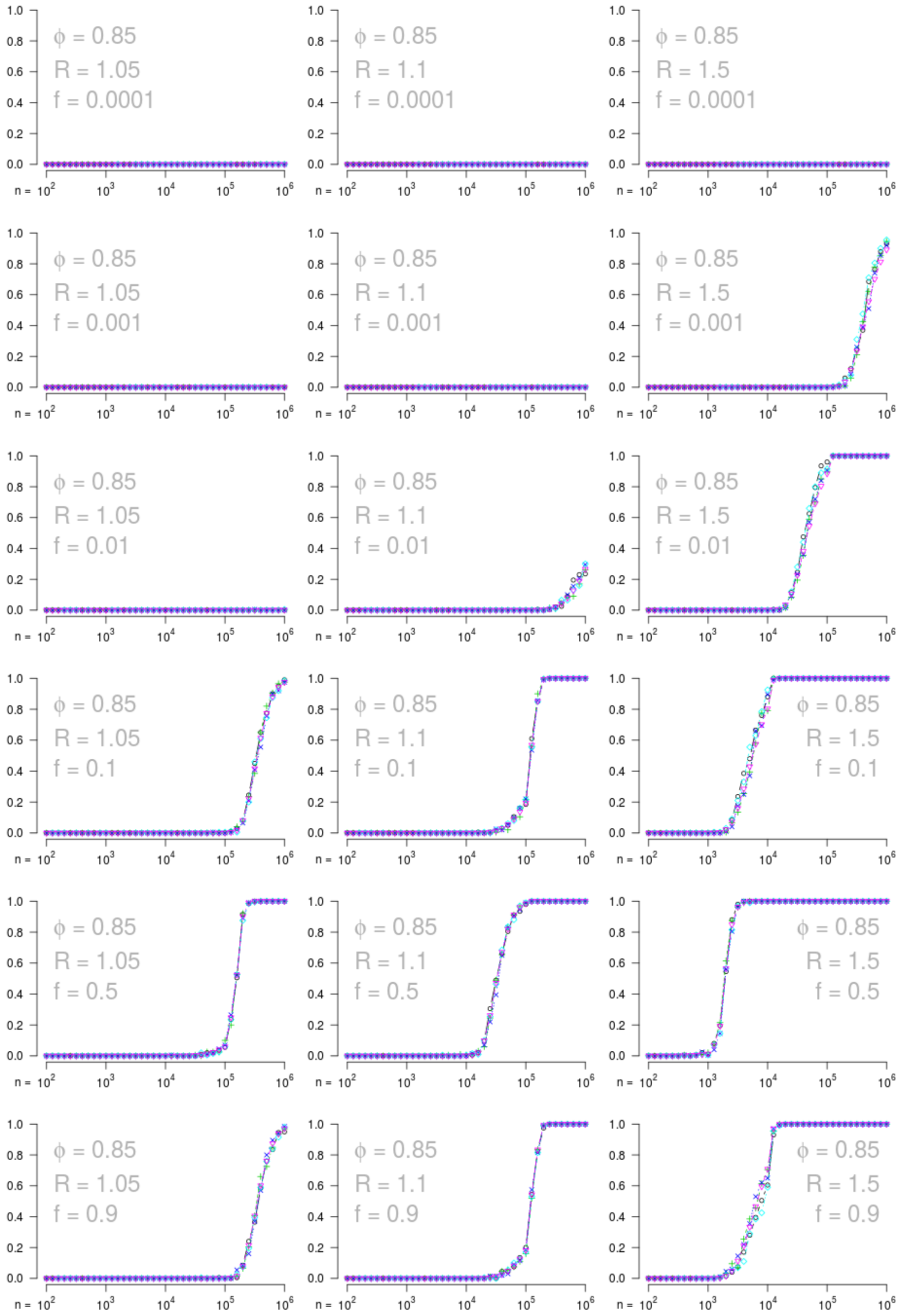

Figure 7: Empirical power as a function of sample sizes, under small to moderate odds ratios (left to right panels), and under rare to common risk allele frequencies (top to bottom panels). The Fisher's exact test (black circles), Welch's t-test (red triangles), Pearson's chi-square test (blue crosses), Likelihood ratio test, and equivalently, LR test with logistic regression (light blue diamonds) are almost identical to the theoretical predictions (purple nabla), at sample sizes ranging from 100 to 100,000. Fraction of cases $\phi = 85\%$.

7.2 Power as a function of the canonical parameters

To fully explore the accuracy of the theoretical predictions, we extend the simulation to a wider range of model parameters.

- Sample size total, n: 10^2, 10^3, 10^4, 10^5,
- Fraction of cases in the study, ϕ: 5%, 15%, 25% 50%, 85%,
- Risk allele frequencies in control group, f: a grid of 100 values ranging from 0.01% to 99.5%. The values are non-regularly spaced; more values are selected towards the end points of the interval $(0, 1)$.
- Odds ratio, R: a grid of 100 values ranging from 1 to 100. The values are regularly spaced on the log-scale.
- p-value cutoffs: 5×10^{-5}, 5×10^{-8}.

Each of the $4 \times 4 \times 100 \times 100 \times 2$ parameter value combinations were simulated 1000 times.

We compare the theoretical results with powers of Fisher's exact test obtained by simulations. We choose to compare against exact tests for their superior performance in finite samples — while approximate tests like the chi-square test and likelihood ratio tests may fail to protect against type I error inflation when sample sizes are small, exact tests maintain the correct levels.

The results for p-value cutoff at 5×10^{-8} are visualized in the form "OR-RAF diagrams" to better highlight settings under which the theoretical predictions and the empirical powers differ.

The results are shown in Figures 8 to 12, organized by increasing fraction of cases in the study ϕ. Left columns of the figures show the theoretical predictions obtained by Theorem 1, the middle columns shows the estimated powers from simulation, and the right columns visualize their differences. Sample sizes ranging from 100 to $100,000$ are arranged from top to bottom in each figure.

7.3 Accuracy of theoretical predictions

We find that the asymptotic approximations in Theorem 1 to be robust for small to moderate odds ratios ($R < 2$), as seen in all simulation studies (see Figures 3 to 12).

The theoretical approximations are extremely accurate (within 1 percentage point in absolute deviations) for balanced designs, even at very small sample sizes (see Figure 11), as well as in settings with moderately more controls than controls at large sample sizes (see Figures 9 and 10, $n = 100,000$). Some discrepancies can be seen under designs with too

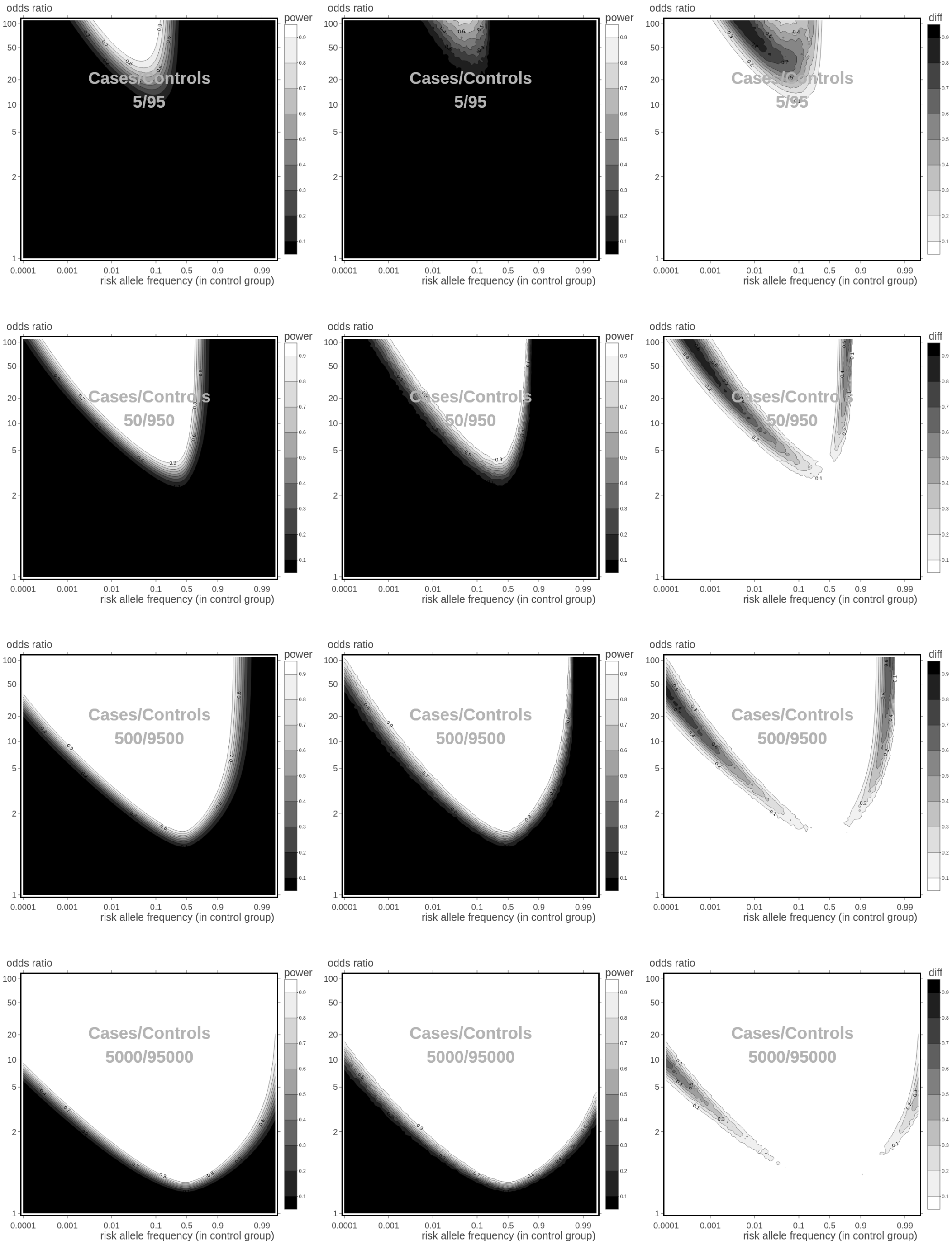

Figure 8: Statistical powers for the OR-RAF combinations, obtained from theoretical predictions (left column) and by simulations (middle column). Their absolute differences (right column) are shown at sample sizes ranging from 100 to $100,000$ (top to bottom). p-value threshold is at 5×10^{-8}. Fractions of cases $\phi = 5\%$.

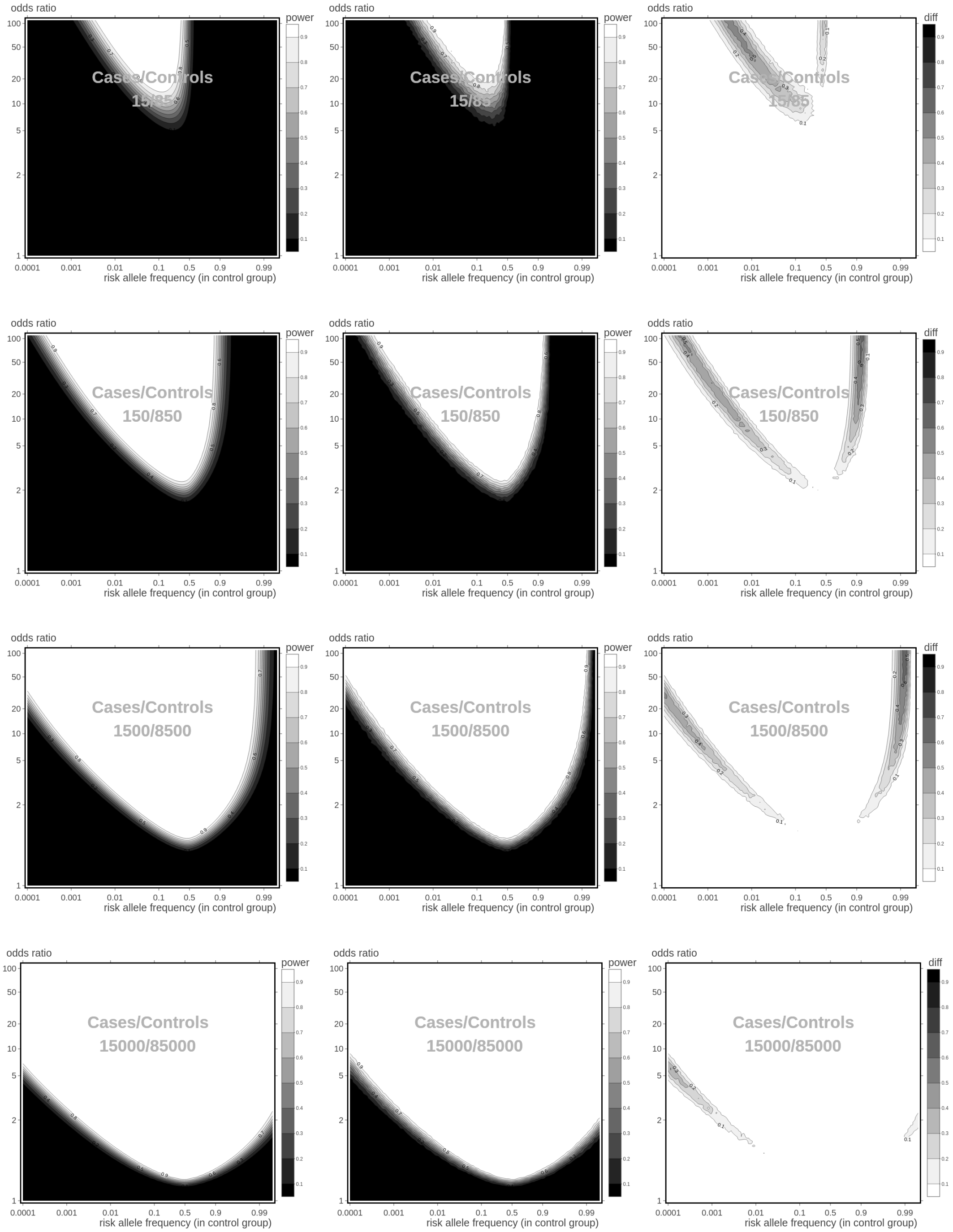

Figure 9: Statistical powers for the OR-RAF combinations, obtained from theoretical predictions (left column) and by simulations (middle column). Their absolute differences (right column) are shown at sample sizes ranging from 100 to $100,000$ (top to bottom). p-value threshold is at 5×10^{-8}. Fractions of cases $\phi = 15\%$.

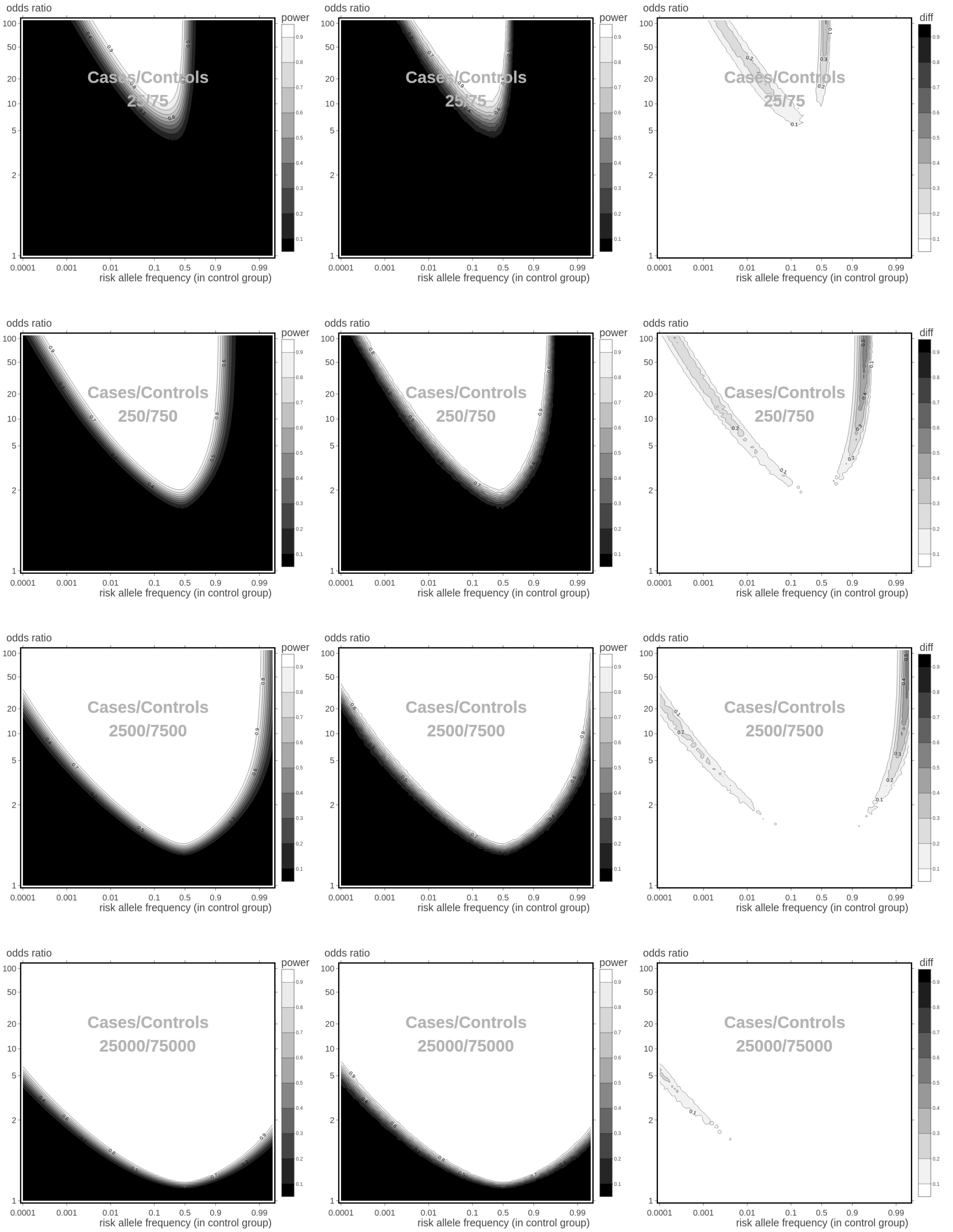

Figure 10: Statistical powers for the OR-RAF combinations, obtained from theoretical predictions (left column) and by simulations (middle column). Their absolute differences (right column) are shown at sample sizes ranging from 100 to $100,000$ (top to bottom). p-value threshold is at 5×10^{-8}. Fractions of cases $\phi = 25\%$.

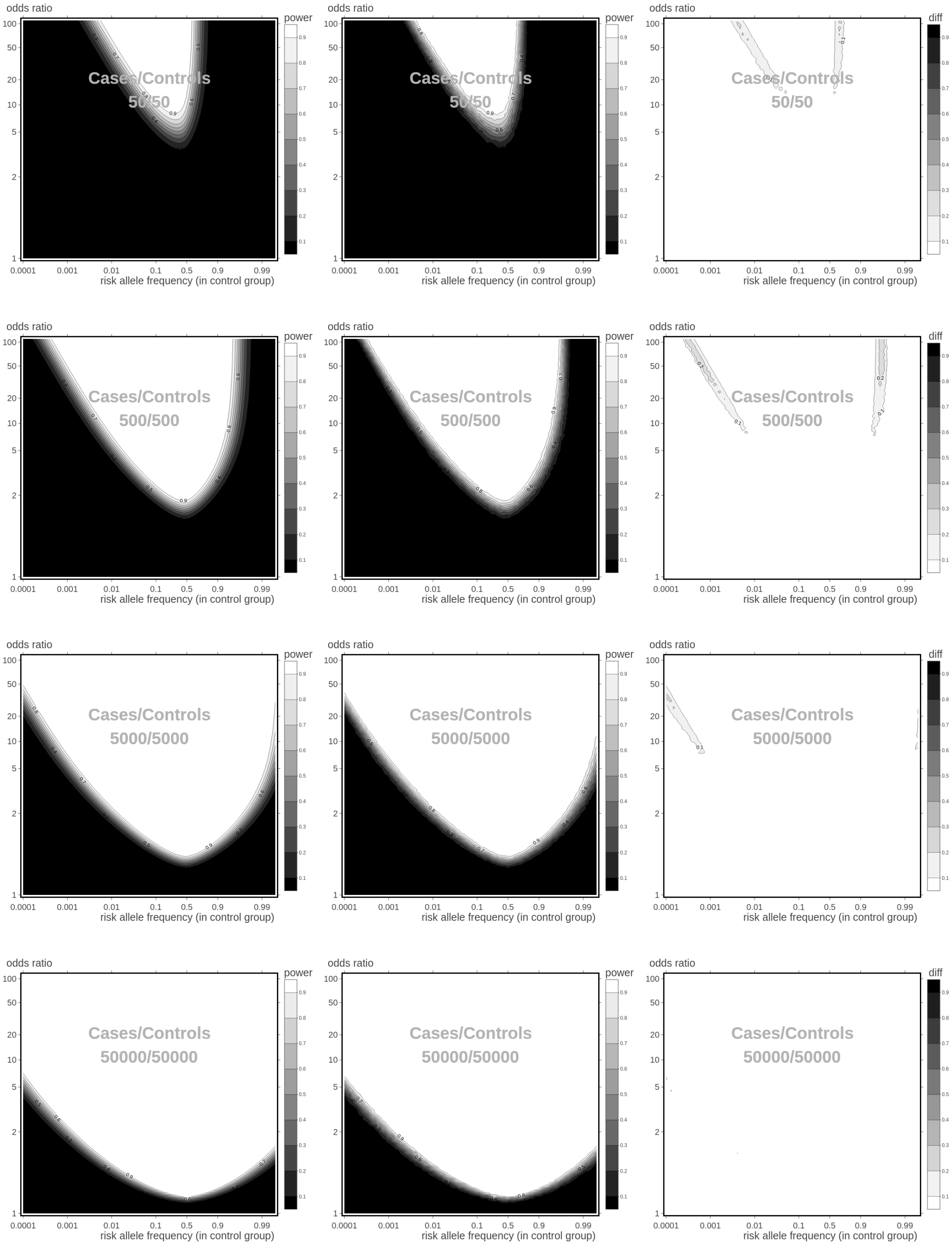

Figure 11: Statistical powers for the OR-RAF combinations, obtained from theoretical predictions (left column) and by simulations (middle column). Their absolute differences (right column) are shown at sample sizes ranging from 100 to $100,000$ (top to bottom). p-value threshold is at 5×10^{-8}. Fractions of cases $\phi = 50\%$.

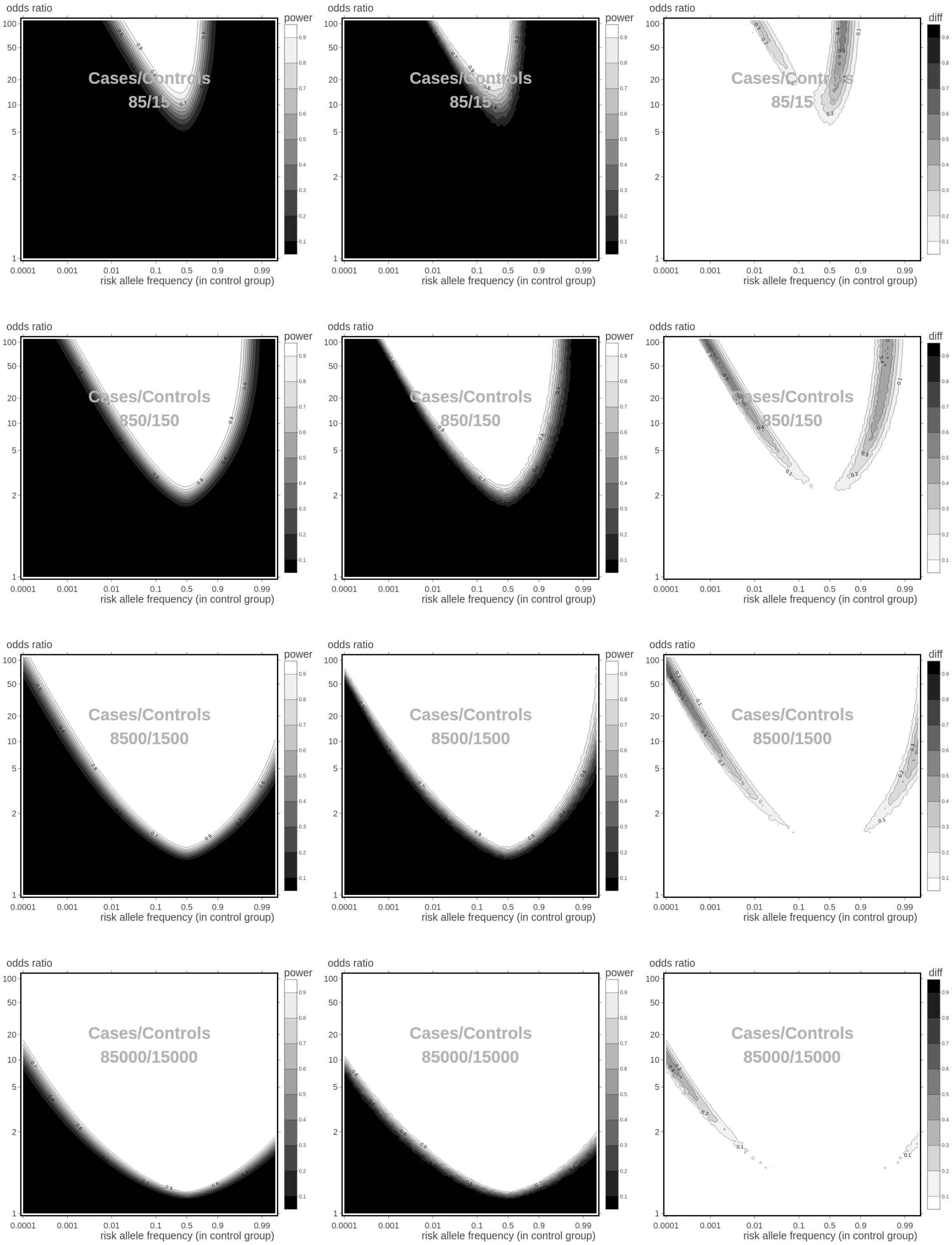

Figure 12: Statistical powers for the OR-RAF combinations, obtained from theoretical predictions (left column) and by simulations (middle column). Their absolute differences (right column) are shown at sample sizes ranging from 100 to $100,000$ (top to bottom). p-value threshold is at 5×10^{-8}. Fractions of cases $\phi = 85\%$.

few cases (see Figure 8) as well as too few controls (see Figure 12), although the differences vanish as the sample sizes increase.

	Fraction of cases ϕ				
sample size n	5%	15%	25%	50%	85%
100	4.61 (0.15)	2.39 (0.08)	1.44 (0.05)	0.78 (0.02)	2.00 (0.07)
$1,000$	5.00 (0.15)	3.37 (0.10)	2.35 (0.7)	1.03 (0.03)	3.44 (0.10)
$10,000$	4.73 (0.14)	3.06 (0.09)	1.98 (0.07)	0.50 (0.02)	2.43 (0.08)
$100,000$	1.74 (0.07)	0.73 (0.03)	0.43 (0.02)	0.18 (0.01)	0.88 (0.04)

Table 1: Mean absolute deviations in percentage points between the theoretical predictions and the simulated results, with standard errors in brackets. The p-values threshold are set at 5×10^{-8}.

In general, the mismatch between the theoretical predictions and empirical power only takes place at small sample sizes, and only for large odds ratios ($R > 3$). For small odds ratios, numerical experiments suggest that our power calculations are robust across a wide range of case-control ratios and sample sizes.

Theoretical predictions for p-value cutoff at 5×10^{-5} are up to 20% more accurate than for p-value cutoff at 5×10^{-8}, but qualitatively similar. For example, the average absolute deviation for unbalanced designs ($\phi = 15\%$) at sample size $n = 100,000$ is only 0.57 percentage points for p-value cutoff at 5×10^{-5}, down from 0.73 percentage points when p-value cutoff is at 5×10^{-8}. The same metric improves to 0.16, down from 0.18 percentage points, for balanced designs ($\phi = 50\%$) at the same sample sizes. In the interest of space, we only show the more challenging cases where the p-value cutoffs are set at 5×10^{-8}.

For sample sizes beyond $100,000$ (not reported), we find the empirical power to be almost indistinguishable from theoretical predictions.

Funding

This work is partially supported by NSF Grant DMS-1830293, Algorithms for Threat Detection.

References

Choi, L., Blume, J. D., & Dupont, W. D. (2015). Elucidating the foundations of statistical inference with 2 x 2 tables. *PloS one*, **10(4)**, e0121263.

Derkach, A., Zhang, H., & Chatterjee, N. (2017). Power Analysis for Genetic Association

Test (PAGEANT) provides insights to challenges for rare variant association studies. *Bioinformatics*, **34(9)**, 1506-1513.

Domínguez-Cruz, M. G., de Lourdes Muñoz, M., Totomoch-Serra, A., García-Escalante, M. G., Burgueño, J., Valadez-González, N., ... & Díaz-Badillo, Á. (2018). Pilot genome-wide association study identifying novel risk loci for type 2 diabetes in a Maya population. *Gene*, **677**, 324-331.

Ferguson, T. S. (2017). *A course in large sample theory.* Routledge.

González, J. R., Carrasco, J. L., Dudbridge, F., Armengol, L., Estivill, X., & Moreno, V. (2008). Maximizing association statistics over genetic models. *Genetic Epidemiology*, **32(3)**, 246-254.

Haryono, S. J., Datasena, I. G., Santosa, W. B., Mulyarahardja, R., & Sari, K. (2015). A pilot genome-wide association study of breast cancer susceptibility loci in Indonesia. *Asian Pacific journal of cancer prevention: APJCP*, **16(6)**, 2231–2235.

Hunter, D. (2002) Topic 25: Asymptotic Power of Pearsons Chi-Square Test in *Statistics 597A: Asymptotic Tools (Fall 2002), PSU Lecture notes,* http://personal.psu.edu/drh20/asymp/fall2002/lectures/ln11.pdf

Johnson, J. L., & Abecasis, G. R. (2017). GAS Power Calculator: web-based power calculator for genetic association studies. *bioRxiv*, 164343.

Lehmann, E. L. (2004). *Elements of large-sample theory.* Springer Science & Business Media.

Li, Q., Zheng, G., Li, Z., & Yu, K. (2008). Efficient approximation of Pvalue of the maximum of correlated tests, with applications to genomewide association studies. *Annals of human genetics*, **72(3)**, 397-406.

MacArthur, J., Bowler, E., Cerezo, M., Gil, L., Hall, P., Hastings, E., ... & Pendlington, Z. M. (2016). The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic acids research*, **45(D1)**, D896-D901.

Menashe, I., Rosenberg, P. S., & Chen, B. E. (2008). PGA: power calculator for case-control genetic association analyses. *BMC genetics*, **9(1)**, 36.

Purcell, S., Cherny, S. S., & Sham, P. C. (2003). Genetic Power Calculator: design of linkage and association genetic mapping studies of complex traits. *Bioinformatics*, **19(1)**, 149-150.

Ripamonti, E., Lloyd, C., & Quatto, P. (2017). Contemporary Frequentist Views of the 2×2 Binomial Trial. *Statistical Science*, **32(4)**, 600-615.

Sham, P. C. (1998). *Statistics in human genetics*. Wiley.

Skol, A. D., Scott, L. J., Abecasis, G. R., & Boehnke, M. (2006). Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies. *Nature genetics*, **38(2)**, 209.

Smyth, G., Hu, Y., Dunn, P., Phipson, B., & Chen, Y. (2017) statmod: Statistical Modeling. R package version 1.4.30.

Wang, G. T., Li, B., Lyn Santos-Cortez, R. P., Peng, B., & Leal, S. M. (2014). Power analysis and sample size estimation for sequence-based association studies. *Bioinformatics*, **30(16)**, 2377-2378.