Genes **2014**, *5*, 460-476; doi:10.3390/genes5020460

OPEN ACCESS
genes
ISSN 2073-4425
www.mdpi.com/journal/genes

Article

GWAS to Sequencing: Divergence in Study Design and Analysis

Christopher Ryan King [1] and Dan L. Nicolae [2,*]

[1] Department of Health Studies, University of Chicago, Chicago, IL 60637, USA;
E-Mail: c.ryan.king@gmail.com

[2] Departments of Medicine, Statistics, and Human Genetics, University of Chicago, Chicago, IL 60637, USA

* Author to whom correspondence should be addressed; E-Mail: nicolae@galton.uchicago.edu;
Tel.: +1-773-702-4837.

Received: 27 December 2013; in revised form: 13 May 2014 / Accepted: 15 May 2014 / Published: 28 May 2014

Abstract: The success of genome-wide association studies (GWAS) in uncovering genetic risk factors for complex traits has generated great promise for the complete data generated by sequencing. The bumpy transition from GWAS to whole-exome or whole-genome association studies (WGAS) based on sequencing investigations has highlighted important differences in analysis and interpretation. We show how the loss in power due to the allele frequency spectrum targeted by sequencing is difficult to compensate for with realistic effect sizes and point to study designs that may help. We discuss several issues in interpreting the results, including a special case of the winner's curse. Extrapolation and prediction using rare SNPs is complex, because of the selective ascertainment of SNPs in case-control studies and the low amount of information at each SNP, and naive procedures are biased under the alternative. We also discuss the challenges in tuning gene-based tests and accounting for multiple testing when genes have very different sets of SNPs. The examples we emphasize in this paper highlight the difficult road we must travel for a two-letter switch.

Keywords: sequencing; GWAS; prediction; power of genetic association

1. Introduction

The Human Genome Project has paved the way to the data revolution in complex disease genetics, by permitting the development of databases of genetic variation, such as HapMap [1], and machinery for producing genome-wide data, such as genotyping arrays and high-throughput sequencing technologies. Our understanding of the genetic risk factors for complex traits has evolved from a few loci discovered with positional cloning approaches in the 1990s to thousands of replicated associations from genome-wide association studies (GWAS), available to the public, as well as scientists in the NHGRI catalog [2]. Interest has shifted recently to discovering disease association with data from whole-genome or whole-exome sequencing studies, and so far, these have had limited success. GWAS has delivered on their early promise to speed up the search for disease genes, and there are bold predictions about what sequencing can achieve [3] on the way to the era of personalized medicine. Sequencing could offer a complete picture of genetic variation—from SNPs to Copy Number Variants (CNVs) and insertions-deletions—for the subjects in the study and for future patients and has led to successful discoveries in Mendelian diseases. So far, sequencing has had limited success for complex diseases, mostly in candidate gene studies. Whole-genome and whole-exome sequencing investigations have only demonstrated the complicated architecture of common traits, sometimes indirectly through a lack of findings in single-SNP low-frequency analyses.

The success of GWAS and of the corresponding analytical tools leads naturally to an investigation of what is different between the two strategies. The goal of this paper is to compare some of the divergent aspects of GWAS and sequencing studies with the hope of guiding future sequencing investigations. We focus on two key distinctions. First, we look at consequences that follow from investigating SNPs with low minor allele frequency (MAF), including the ability to detect novel SNPs. It is important to reiterate that GWAS analyses cover, directly (through genotyping or imputation) or indirectly (through linkage disequilibrium), most of the common variants in the studied populations. This implies that the goal of sequence-based studies is to detect association with low frequency and rare variants. Even though sequencing studies can be used to investigate high MAF SNPs, we ignore their role, since traditional genotyping is dramatically more cost effective. Furthermore, we do not discuss the fact that sequencing studies permit the investigation of structural variation, an important characteristic for diseases, such as autism, where these variants play an important role. In Section 2, we develop a simple analytical formula for the power of a burden test and use it to illustrate the factors affecting power with a contrast to GWAS and scenarios for improving them. In Section 3, we illustrate two novel problems with estimation and prediction using sequencing data. First, we show that case-control studies, which add rare SNPs into a super-SNP or test the distribution of case- and control-private SNPs, can be misleading if analyzed naively. Second, we show that the optimal prediction for previously observed and novel rare SNPs can be strikingly different.

Second, we turn to issues surrounding the use of gene-based tests. In Section 4, we discuss the difficulty of selecting and tuning gene-based test statistics and contrast this to the case in GWAS. We show the alternative hypothesis, which would recommend that a particular procedure can be quite unstable even with seemingly irrelevant details of a gene. We do not recommend a particular testing procedure, but highlight concerns guiding the tuning parameter selection. Finally, in Section 5, we highlight the sharp distinctions between multiple-testing-adjustment strategies for GWAS and gene-based tests.

2. Power of Sequencing *versus* GWAS

The relatively minor number of associations with rare variants seems surprising to many, but was predicted by prior knowledge on the genetics of complex phenotypes. For example, the lack of major linkage loci for diseases, like type 2 diabetes [4], suggests that there are no genes with many rare variants with very large effects. Given this lack of observed associations, it is useful to investigate the relative contributions of factors driving power. We will illustrate with a burden-style test for which an analytical power calculation is straightforward.

The goal here is not to calculate power nor to find realistic sample sizes for genetic association studies with rare variants. Existing software (e.g., [5]) can perform such calculations. Our aim is to use simple analytical calculations to gain insight into what drives power and what are possible strategies for designing optimal investigations. A comparison to GWAS will illustrate the challenges ahead of us. One important set of shared assumptions for GWAS and WGAS is that of the unconfoundedness of associations. Recent work has suggested that approaches to adjusting for population structure, which work well in GWAS, may not in WGAS [6–8]. However, the literature on this topic is rapidly evolving, and we will set this problem aside for purposes of discussion.

Assume a balanced design with n cases and n controls. It can be shown (see Appendix A for the assumptions used in the derivation of this) that the non-centrality parameter (NCP) for burden tests [9,10] can be approximated by:

$$\sqrt{n}\,\frac{k_1}{\sqrt{k}}\,\frac{E_M}{\sqrt{V_M + E_M - E_M^2}}\,(\gamma - 1) \tag{1}$$

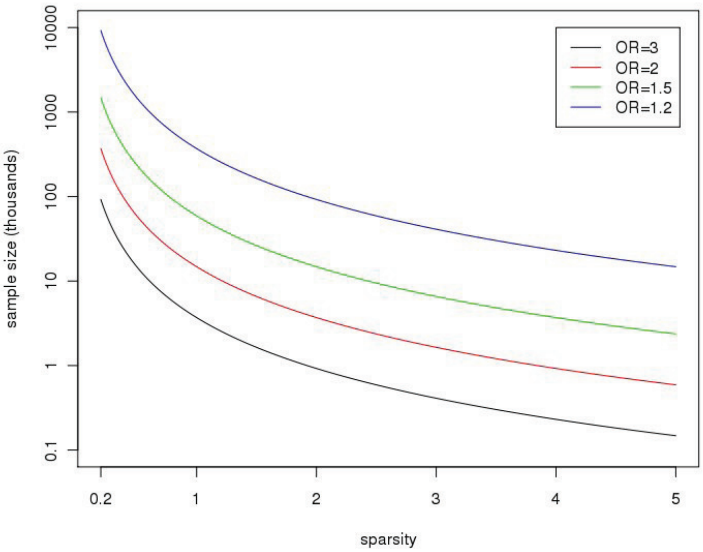
where the test is done on a set of k SNPs, out of which, k_1 are associated with a common odds-ratio (OR) of γ, and $E_M,\ V_M$ are the mean and variance of the minor allele frequency (MAF) for the SNPs in the set. This formula works for single SNP analyses, as well, with $k_1/\sqrt{k} = 1$ and the term about frequency replaced by the corresponding function of MAF. Note that the power of the test is approximately linear in the NCP in the interesting range of moderate values.

All the terms, except the one containing elements of the MAF distribution, are easy to calculate and interpret. The MAF term can be approximated using 1000 Genomes Project data and calculations conditional on an SNP being polymorphic in a study. For 5000 cases and 5000 controls of European descent, and filtering to SNPs with MAF $< 1\%$, that term is close to 0.046, and the non-centrality parameter when $k = 100$, $k_1 = 10$ and γ = 3 is approximately 6.47. Those settings yield a power of 87% at the genome-wide 5×10^{-8} significance level. We will discuss the four terms in Formula (1) and contrast the results between GWAS and sequencing.

Sample size: The simplest way to double the NCP is to increase the sample size by a factor of four. This requires the least amount of innovation, but takes a huge effort and expense, especially when using existing cohorts, since ascertaining and phenotyping additional samples comparable with existing data is very difficult. As is common with many GWAS meta-analyses, a cost-effective increase in the sample size requires the use of ancestry-diverse populations. Additional diversity increases heterogeneity and will affect power to a larger degree than in GWAS, both because the effective MAF decreases (many rare alleles are population-specific) and because a similarly defined set of SNPs (e.g., all exonic SNPs in a given gene) will have different elements in different populations, with powerful tests requiring the presence of functional/causal variants in each (sub)population. We also anticipate that cryptic gene-environment interactions (GxE) provides a substantial amount of heterogeneity in effect sizes. GxE has been long known to exist for some complex traits (e.g., for a review in psychiatric phenotypes, see [11]); given how difficult is it to anticipate relevant modifiers, measure them accurately and statistically detect them [12], it seems likely that unknown environmental modifiers are not uniformly distributed across populations. We will not expand on the difficulties inherent to adjusting for structure in diverse samples, but note that this is much more challenging in sequencing, since rare SNPs can be specific for relatively recent and small-scale demographic events [9,10].

Sparsity of signals and variant annotation: The next term in Formula (1) has to do with the number of associated SNPs relative to investigated SNPs, which we call the sparsity of the signal. Figure 1 shows the impact of sparsity on sample sizes needed to design powerful association studies. For GWAS ($k_1 = k = 1$), the sparsity term is equal to one for associated SNPs. In sequencing studies, it is possible to increase the power by reducing the number of non-associated SNPs (for sets where k is large compared to k_1). The annotation of SNPs through functional status, eQTL (expression quantitative trait loci) studies, ENCODE, prior data, *etc.*, will allow more useful definitions for the analyzed sets by excluding SNPs with a low *a priori* likelihood of being associated. This is a fruitful area of current research and one that is implicit in some study designs, such as exome sequencing.

Figure 1. The plot shows the sample sizes (on the y-axis, in thousands) needed to achieve 80% power at the 10^{-6} significance level as a function of "sparsity", $k_1/\sqrt{k}$ (on the x-axis), with k and k_1 as defined in the text. It is assumed for these calculations that the k SNPs are independent (no linkage disequilibrium), with the minor allele frequency (MAF) sampled from a beta distribution with parameters selected to match allele frequencies from the CEU of the 1000 Genomes Project, $B(0.14,0.73)$; the distribution is truncated at 0.01 (so SNPs have MAF $< 1\%$) and only polymorphic SNPs when sequencing 10,000 subjects are selected. Calculations are based on the NCP in Equation (1). OR, odds ratio.



The MAF distribution: The dominant term in the denominator of Formula (1) is given by the mean MAF, so a simple approximation to the third component of NCP is $\sqrt{E_M}$. This is the term in NCP that explains most of the difference between sequence association and GWAS. The corresponding term for a single-SNP test with a risk allele frequency of 0.2 is approximately 10 times larger than our 1000 Genomes-based estimate. In order to have comparable power between sequencing and GWAS, this loss would have to be balanced by the other terms (sample size, sparsity and effect size). There is no easy strategy to increase this term in unrelated individuals, but one available route is to shift the study design to families or isolated populations, where alleles which are rare in the larger population are locally common.

Phenotyping/environment: We can also increase power by analyzing datasets with a larger effect size; this corresponds to the last term in Formula (1), $(\gamma - 1)$. This can be done using stratified analysis: by analyzing sub-phenotypes and/or by accounting for environment (when GxE is present). This is a common issue for GWAS and sequencing, and we illustrate the impact of stratification on the effect size using a single SNP as the unit of analysis. Let us assume that γ_T is the mean effect corresponding to the cases in the most at risk strata (with the rest of cases being "controls" with respect to the variants in the set under investigation). Let α be the proportion of relevant cases, and let p be the control MAF. It follows that in the full set of cases (relevant and irrelevant), the MAF is approximated by $\alpha p \gamma_T + (1-\alpha)p$, and the corresponding effect size is $\gamma - 1 \approx \alpha(\gamma_T - 1)$. Therefore, if phenotyping or sub-setting by environment allows one to find the relevant cases, analyzing a smaller sample size (of αn) leads to an increase of $1/\alpha$ in the fourth term of NCP and to a $1/\sqrt{\alpha}$ overall increase in NCP.

3. Prediction Using Rare and Novel SNPs: A Different Winner's Curse

Aside from association discovery, one of the major goals of GWAS is to estimate the effect sizes of SNPs on traits, which can be used for the prediction of unrealized phenotypes on newly sequenced individuals. For example, SNP genotyping platforms have recently been used for risk and pharmacogenomic prediction by several companies, such as 23andMe, Life Technologies, and Pathway Genomics. Prediction using SNPs discovered in a sequencing study can be performed analogously to GWAS, as long as adequate data has been gathered. One major difference between GWAS and sequencing is that newly-sequenced individuals will regularly carry novel SNPs in disease-associated genes, and most discovered SNPs will have too little information for accurate per-SNP estimates [13,14]. Quantitative estimates of personal risk based on sequencing association studies will therefore require an evidence-based estimate of the effect of previously unobserved and seldomly observed rare SNPs. Given that mutations in the gene in question have already been associated with disease and that harmful SNPs are thought to be more likely to be rare [15–20], ignoring these SNPs (setting the effect to zero) is unlikely to be accurate. Naively, we could estimate a "rare SNP effect" based on the rare SNPs observed in previous sequencing studies and apply that estimate to new SNPs and known rare SNPs alike. We illustrate two problems with no analog in GWAS that occur when rare SNPs are lumped into a super-SNP for estimation or prediction. The major results are that: (1) rare alleles in a sequencing study can cumulatively have a substantial per-allele OR, which depends on disease prevalence, even if log odds-ratios (lORs) are centered at zero; (2) the prediction of new samples based on that OR is substantially inaccurate.

First, unlike GWAS, prediction with new SNPs depends non-trivially on the variability of rare SNP effects. With GWAS, previous data will give the investigator an estimate of the effect of each SNP; a plug-in prediction can be formed using these estimates: $logit(\hat{Y}_i) = G_i\hat{\beta}+\alpha$, where $logit$ is the logistic function, $\hat{Y}_i$ is the predicted probability for person i, G_i is the vector of genotypes for that person, α is an intercept, which depends on disease frequency, and β are SNP lORs. The naive plug-in prediction is not quite correct due to the uncertainty in SNP effects; however, the inaccuracy with GWAS-based estimates tends to be negligible for reasons discussed below. In contrast, the effect of a rare SNP in an associated gene is not precisely known, and the impact of that uncertainty on prediction is substantial. For example, even if SNPs in an associated gene are as likely to be risk-decreasing as risk-increasing, the correct prediction in the context of a rare disease for a newly sequenced individual is that carrying a novel SNP increases their odds of being affected. Qualitatively, the uncertainty in SNP effects makes one less confident in the plug-in estimate and pushes the best estimate from the raw prevalence closer to 50:50. To give a numerical example, if the population of lORs for novel SNPs is Gaussian, with a mean of zero and standard deviation of one, and the disease frequency is 1%, then the marginal OR for carrying an allele (*versus* no minor alleles) is 1.9.

This is a well-known phenomenon from the literature comparing marginal and conditional random effects [21–23]; derivation of the effect size and additional explanation is offered in the Appendix. A useful formula for the risk associated with carrying new SNPs can be derived under the assumption that their lORs are Gaussian distributed with mean μ and standard deviation σ along with standard logistic regression assumptions. Define c as a constant related to the disease prevalence (the threshold in the Appendix $c = -log(\frac{p}{1-p}) \approx -log(p)$) and g_i as the number of alleles in that individual, then:

$$logit\left(Pr\{Y_i = 1|g_i, \mu, \sigma\}\right) \approx \frac{-c+\mu g_i}{\sqrt{1+\nu^2\sigma^2 g_i}} \tag{2}$$

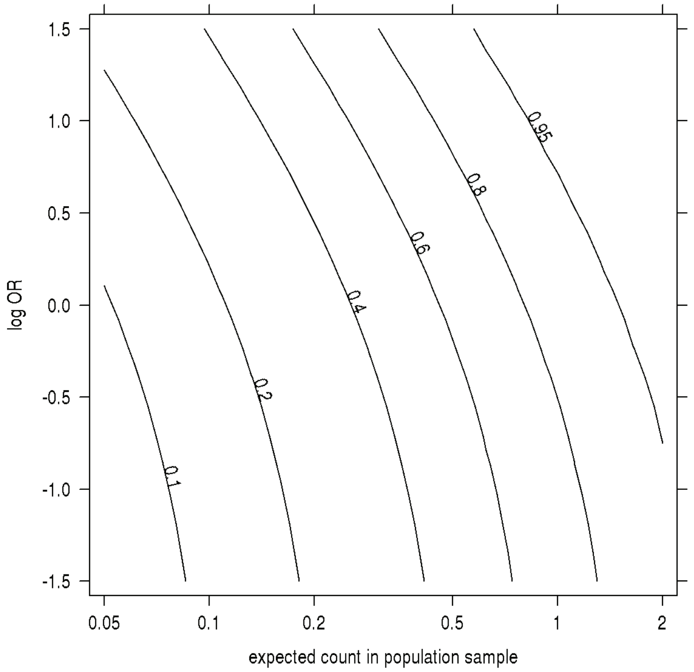
where $\nu \approx 0.625$. When the mean lOR is zero and the standard deviation is not large (it is almost assuredly less than one in areas without overwhelming evidence for linkage), the lOR for having an SNP (*versus* no SNPs) is approximately $\frac{c\nu^2\sigma^2}{2\sqrt{1+\nu^2\sigma^2}}$, which increases sharply with the standard deviation of lORs and scales with the negative log of disease prevalence. This is the lOR that we estimate when regressing the outcome on the number of SNPs carried and that we would use for prediction, absent other information about the effects of particular SNPs.

A related result occurs for GWAS-based plug-in estimates, the details of which depend on the choice of statistical estimators used for effect estimates and the pattern of linkage disequilibrium. For any estimate of an SNP's lOR for which a central limit theorem applies, σ^2 in Equation (2) can be replaced with the square of the standard error of the estimate, which will usually scale as the inverse of the sample size and the MAF. Given the relatively low cost of GWAS data acquisition and the need to overcome the burden of genome-wide multiple testing, we are accustomed to gathering enough data for precise estimates. For example, the expected standard error of a lOR of zero with a MAF of 0.3 and 3000 cases and 3000 controls is 0.06. If the standard deviation of the population of novel SNP lORs is 0.5, then the marginal lOR for a new SNP is 70 times bigger than the previously observed SNP.

The above effect is observable regardless of the sample size and MAF of SNPs used in the calculation. One might expect that since case control-based estimates of ORs are consistent for prospective associations, that this effect would be corrected by empirically estimating the per-allele OR and using that for future data; however, there is a unique twist for the group of SNPs with MAFs, such that they are reasonably likely to be monomorphic in the original study. The observed lOR for all rare SNPs together does estimate the marginal effect of future rare SNPs, but that prediction breaks down when stratified by whether or not the SNP was observed as polymorphic in the case-control study. Case-control designs are somewhat more efficient for discovering rare risk-increasing SNPs compared to risk-decreasing SNPs [20,24–27], so as a group, previously observed SNPs are more harmful in future samples than newly observed SNPs.

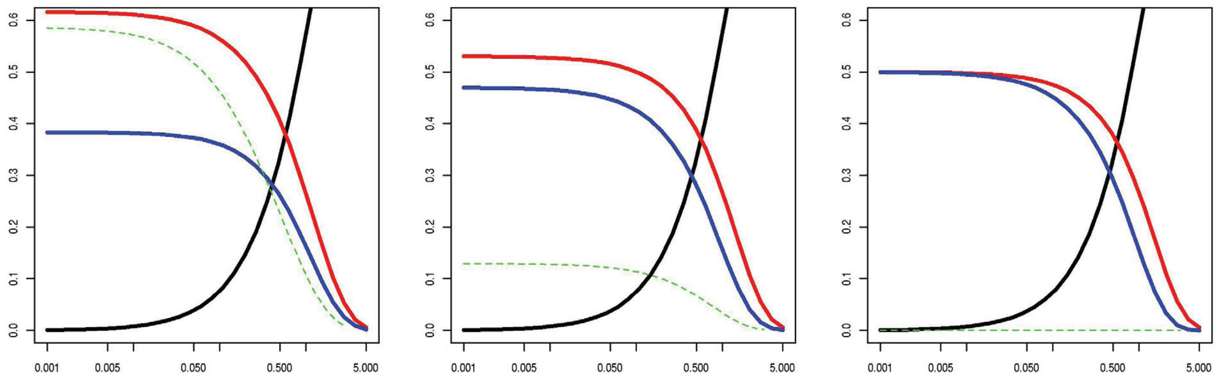
In Figure 2, we plot how the probability of an SNP being discovered (the minor allele is observed in at least one participant) in a case-control study depends on both the OR and the MAF when the MAF is low compared to the sample size. The absolute probability of a SNP appearing at least once in the study increases markedly with OR in this range of MAF. Intuitively, compared to a population sample, a risk-increasing SNP's greater frequency among cases more than makes up for the decline in its frequency among controls. As a result, the observed odds of a rare SNP appearing in a case are inflated, and the finite pool of remaining SNPs at that MAF contains a preponderance of protective and small effects.

Figure 2. Sampling probability by MAF, log odds-ratio. The contour plot has on the x-axis the allelic expected count in a population sample the same size as the control group (sample sizes times MAF) and, on the y-axis, the log-odds ratio. Contours are the absolute probability of being sampled in a case-control study of 100 cases and 100 controls when prevalence equals 1%.



This is similar to the first problem discussed above, except that we have selectively observed SNPs based on their true odds ratio. Figure 3 shows that when the lORs of SNPs in a gene are assumed to come from a population with high variance, the odds of an SNP appearing only in cases and the expected lOR of observed SNPs varies with MAF and substantially favors an increase in risk. This is not a Bayesian argument; it relies only on the rate at which SNPs appear in the sample, even for fixed SNP effects. To give a numerical example, with a sample size of 100, a prevalence of 5%, 125 SNPs with a MAF of 0.002 and lORs drawn from a standard normal, an average of 36% of the SNPs are discovered in the original sequencing study with an average per-allele estimated lOR of 0.33. In a new replication or prediction sample, the average per-allele lOR based on previously discovered SNPs is 0.66, but the per-allele lOR of new SNPs is only 0.05. The numerical result depends heavily on a number of parameters; we have deferred a detailed exploration of the phenomenon to another work [28]. The longer report is available for download, and re-demonstration of the importance of each factor is beyond the scope of this paper. However, there are three notable features to which we wish to briefly draw attention: (1) the tail behavior of SNP lORs is very influential; large ORs enrich even rare SNPs into the population of cases; (2) the effect occurs at a MAF around the minimum observable in a study; regardless of the size of the original dataset observed, rare SNPs are unrepresentative of future rare SNPs; (3) the effect vanishes under the null hypothesis that no SNPs in a gene are associated.

Figure 3. Observed data probabilities by MAF. X-axis N·MAF. The y-axis shows the probability of each special data type conditional on the SNP being polymorphic: occurring only in cases (red), once in controls and zero times in cases (blue) and all other (black). Green = expected log-odds-ratio (OR) of sampled SNPs (same numeric scale). The log-ORs are assumed to be distributed left: $N(0,1)$; center: $N(0,0.5^2)$; right: $N(0,0.25^2)$; other settings are as in Figure 2.



4. Implicit and Explicit Models in Association Studies

In contrast to GWAS, genetics practitioners with sequencing data are currently faced with a dizzying selection of methods to test for an association between genotype and phenotype, each of which has tuning parameters. In GWAS, a simple allelic test is the overwhelmingly most commonly used test. The additive allelic model performs well regardless of the true risk model when linkage disequilibrium between a tested marker and causal allele is imperfect [29]. While some authors have suggested GWAS schemes that incorporate prior knowledge and more complex risk models through explicit Bayesian calculations or alpha-spending procedures (see also Section 5), the dominant technique in the literature is to report SNP-level evidence and the allelic effect from a particular dataset. While the commonly used methods make usual regression-type assumptions about the distribution of the trait, the effects of confounders and covariates and the measurement error of SNPs, they make minimal assumptions about the effects of other SNPs or how effect size varies with SNP- or gene-level features.

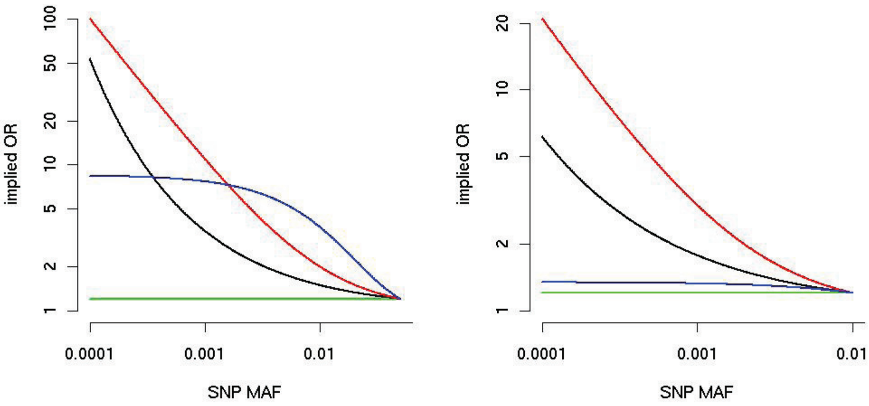
Because of the small amount of information at each rare SNP, all sequencing association tests of which we are aware pool information in some way across SNPs, which are regarded as belonging to a unit (gene) or being "similar," and some pool information across genes that are "similar." These techniques have tuning parameters appropriate under a particular alternative hypothesis and that may suffer a substantial loss of power under other alternatives. A full comparison of proposed tests for sequencing data is beyond the scope of this article; however, we will discuss a few of the most common tests. In this section, we will discuss the role and meaning of some of these tuning parameters. Ignoring these tuning parameters as if the investigator were still using relatively assumption-free GWAS techniques is unlikely to work well, and the importance of these analytic decisions represents a substantial divergence from GWAS.

One extreme of this approach is to try to swap tuning parameters for explicit models and assumptions. We have advocated multi-level modeling of effect sizes or lORs using SNP- and gene-level features as predictors, with stated assumptions, such as the functional form of associations, the linearity and additivity of associations, distributional requirements and exchangeability between SNPs, where required [30].

However, most proposed tests are not model-based summaries. In some cases, we can gain insight into these tests by constructing a map from the tuning parameters to a genetic model, which would imply those as optimal in some way. For example, $C(\alpha)$ and diagonal kernel SKAT [5] can be derived from an explicit model with weights on the $j-th$ SNP, $w_j \propto E[lOR_j^2]$ [31], and therefore, any scheme of weights as a function of MAF can be understood in terms of the implied variance of SNP effects. The addition of correlation structures to SNP effects also follows a simple model-based logic; for example, if effects are expected to substantially go the same direction (such as a group of loss-of-function SNPs), one can balance between the burden-type and variance-component-type test [5]. Similarly, several authors [32,33] have pointed out that optimal weights for burden-type tests are proportional to per-SNP lORs. Figure 4 shows curves for three MAF-depended weights proposed in the literature [34]. Notably, the implied lOR curve depends heavily on the upper limit of MAF included in the pooling procedure.

However, the tuning parameters of some proposed tests are more challenging. For example, SKAT with the Gaussian kernel does not map to a meaningful model of SNP effects, but is suspected to work reasonably well under several alternatives and detects some non-linear effects and epistatic interactions [35–37]. The kernel itself is a tuning parameter for SKAT and related methods; kernel-based techniques are sensitive to the choice of kernel and, aside from a few special cases [35,38], are difficult to choose between *a priori*. Additionally, while most kernels can up- or down-weight SNPs, transforming prior information into a calibrated SNP "similarity" or "distance" measure is a task for which we have little guidance. While SKAT and related tests can be motivated by a simple variance components model, they are not automatically robust to non-Gaussian SNP effects, such as a mixture of causal and non-causal alleles [39,40]. This is not to suggest that the many variants of kernel-based tests are poorly applied tools or that they perform poorly compared to other tests, just to highlight the fundamental difficulty of interpreting and guiding the key analytic decision.

Figure 4. Implied alternative OR (on the y-axis, logarithmic scale) as a function of MAF (x-axis) for three burden weighting schemes. The black line corresponds to the Madsen–Browning weight [10]; the red line corresponds to the attributable risk weight [41], and the blue line corresponds to the default in SKAT, Beta(25,1) [42]; the green line is for equal weights. For the **left** panel, the MAF is truncated at 5%, and for the **right** panel at 1%. We assume that the OR of the SNP with the largest MAF is 1.2.



There are numerous specific deviations from linear Gaussian SNP effects, which hypothetically should influence the tuning parameter selection. In the implicit model tests described above, these issues are difficult to address in planning and power analysis. When considering sequencing data as potential negative evidence in replication studies, each has to be explored on a case-by-case basis. Explicit-model methods have the advantage of facilitating graphical model checks (for an example, see [30]), posterior-predictive diagnostics [43] and prior-data conflict summaries [44,45]. The price of these checks is a relatively high computational burden, stricter distributional assumptions, additional investigator effort eliciting the model (and suitable priors for Bayesian methods [46]) and an unclear definition of a "good enough" model.

5. On Multiplicity

The common strategy used in GWAS for ranking and follow-up of new discoveries is to focus on the SNPs with the most significant *p*-values. As we discussed in Section 2, association tests of common single SNPs yield *p*-values that reflect a combination of sample size, effect size and MAF; however, for the range of MAFs in GWAS, rankings based on *p*-values correlate well with those based on effect size. Adjustment for multiple testing is usually done independent of any information on SNPs, and in a Bayesian framework, this corresponds to an equal prior probability of each SNP being associated. There have been approaches developed to incorporate prior knowledge, such as stratified false discovery control [47] and weighted Bonferroni criteria [48]. Although there are several sources of information for these procedures in GWAS (such as effect on expression, effect on related phenotypes, position relative to gene elements, MAF), formal methods have not been used extensively for several reasons. In general, the available information is difficult to translate to the right scale, and there is low prior confidence that information on tag-SNPs is useful, since the causal SNP is unobserved.

Sequencing-based association studies are even more challenging, because there is more variability in the units of analysis than in GWAS. Gene units vary enormously in the number of SNPs, linkage disequilibrium (LD) pattern, the plausible ratio of causal SNPs, MAF spectrum and annotations. For example, what is more likely to be associated, a gene with two non-synonymous SNPs or a gene with ten non-synonymous SNPs? A gene with ten singletons (variants with only one observed copy of the non-reference allele) *versus* a gene with 10 total minor alleles with varying MAF? A set of ten non-synonymous SNPs *versus* a set of ten intronic SNPs? Furthermore the calculation of optimal weights will include an interplay of subject matter knowledge (e.g., assumptions on the effect sizes for different annotations) and the choice of statistical methods (e.g., some methods will accommodate signal sparsity well).

The complicated assessment of prior probabilities for a set of SNPs is one of the issues in using *p*-values for ranking genes and for deciding on efficient follow-up studies. *p*-values might be a poor proxy for the probability of replication, especially when the signals come from the very rare alleles that might not appear in the subjects used for replication. *p*-values contain little information on strategies for functional validation, because they do not inform on the best variants to be investigated. Relying only on *p*-values for decision-making has a bigger impact in sequencing studies than in GWAS, and we hope that developing better and more diverse measures of significance will become a more active area of research.

6. Discussion

Many people have been surprised by the lack of substantial findings from the recent studies on rare variants performed with whole-genome or whole-exome sequencing and from platforms, such as the exome chip. The reality is that for complex traits, there was little prior evidence in favor of genetic models that would give such studies high power (with multiple rare variants with a large effect per unit of study). The whole literature of the recent past, which is too extensive to be cited here, on investigating low frequency variants using imputation from population-based sequencing shows that large effect SNPs are uncommon for the diseases where they exist. This advocates for the development of more efficient strategies than the brute force sequencing of large, poorly phenotyped cohorts. The detailed annotation of variants should improve the sparsity of signals in the units of analysis, and careful phenotyping and incorporation of environmental factors should lead to the discovery of larger effects.

Much of the analytical effort on the association with sequencing data has been put into the development of novel testing tools. We argue in this paper that it is equally important to focus on other aspects of the process, from the design of the study to the interpretation of results. Furthermore, hypothesis tests and multiplicity adjustments should fit into the paradigm of a careful design that we set out above; model-based tests should incorporate the complexity that we expect without resorting to black boxes or poorly characterized weights. We should also be on guard for excessive parsimony; lumping together rare SNPs into a super-SNP creates a variable with properties that depend on the sampling scheme, minor allele frequency distribution and effects on phenotype distribution in complex ways.

Acknowledgments

The research was supported in part by the NIH grants, U01DK085501, P50MH094267 and R01MH101820. Dr. King was supported by National Institute of General Medical Sciences (NIGMS) T32GM007281 and F30HL103105. We are grateful to Nancy J. Cox and Hae Kyung Im for helpful discussions.

Author Contributions

Both authors conceived and designed the study, performed analytical calculations and simulations, and wrote the paper.

Appendix

A. The Derivation of the Power Formula

The following assumptions are used for the calculation of power: (1) there are n cases and m controls; (2) the association test is performed on a set of k SNPs, out of which, k_1 are associated, and the calculations are done conditional on k and k_1, ignoring the variability in those numbers that is associated with sequencing; (3) MAFs are sampled from a distribution with mean E_M and variance V_M; (4) for simplicity, we assume that the effect sizes of the associated SNPs are independent of MAF; (5) the SNPs are in linkage equilibrium; (6) all associations are with the rare allele; and (7) the effect sizes are sampled from a distribution with a mean odds ration equal to γ.

The association method used for illustration is the "burden" test, where, for each individual, we calculate a score based on the genotypes for the k SNPs. Let G_j denote the number of rare alleles at the j-th SNP, and let w_j be a fixed prespecified weight (it does not depend on the observed data; this is needed to simplify the analytical calculations). For each subject, we calculate:

$$S = \sum_{j=1}^{k} w_j G_j$$

then correlate this with the trait for the detection of association. For case-control studies, this can be done using a two-sample *t*-test. Note that power for a two sample normal test is governed by the non-centrality parameter,

$$\sqrt{\frac{nm}{n+m}}\,\frac{\mu_1 - \mu_2}{\sigma}$$

with classical notation (and assuming equal variance). Note that for large k and n, the burden test will be close to a two sample normal test, and the mean and variance of the scores are the key determinants of power. The approximation is fairly accurate, even for small k, as long as n remains large.

If we denote with P_j the MAF for the j-th SNP, we have that:

$$\mathrm{E}(S) = \mathrm{E}\left[\mathrm{E}(S|P)\right] = \mathrm{E}\left(\sum_{j=1}^{k} w_j 2P_j\right) = \sum_{j=1}^{k} 2w_j E_M = 2\bar{w}kE_M$$

where $\bar{w}$ is the average weight. Similarly,

$$\mathrm{Var}(S) = \mathrm{Var}\left[\mathrm{E}(S|P)\right] + \mathrm{E}\left[\mathrm{Var}(S|P)\right] = \mathrm{Var}\left(2\sum_{j=1}^{k} w_j P_j\right) + \mathrm{E}\left(2\sum_{j=1}^{k} w_j^2 P_j(1-P_j)\right) =$$

$$4V_M \sum_{j=1}^{k} w_j^2 + 2\left(\sum_{j=1}^{k} w_j^2\right)(E_M - V_M - E_M^2) = 2k\overline{w^2}\left[V_m + E_M - E_M^2\right]$$

In the case of equal weights ($w_j = 1$),

$$\mathrm{E}(S) = 2kE_M, \quad \mathrm{Var}(S) = 2k(V_M + E_M - E_M^2)$$

For a rare associated SNP, its MAF is approximated by the product of the MAF in controls and the odds ratio. Because MAF and the odds ratios are independent (Assumption 4), we obtain that the mean MAF is approximated by $E_M\gamma$. This leads to the following mean score in cases (assuming equal weights),

$$\mathrm{E}(S) \approx 2(k-k_1)E_M + 2k_1E_M\gamma = 2kE_M + 2k_1E_M(\gamma - 1)$$

One can similarly derive a formula for $\mathrm{Var}(S)$ in the cases.

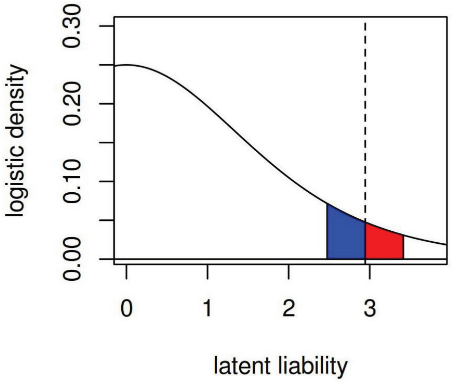
Assuming that the variances of the scores are not greatly different in cases and controls (valid with mild assumptions), the non-centrality parameter is approximated by:

$$\sqrt{\frac{2nm}{n+m}}\,\frac{k_1}{\sqrt{k}}\,\frac{E_M}{\sqrt{V_M + E_M - E_M^2}}\,(\gamma - 1)$$

B. Marginal Effects with Uncertainty

The marginal effect of a SNP whose true log odds ratio comes from a known distribution is easy to quantitatively analyze when using a model of the binary disease outcome as a dichotomized latent liability plus an SNP effect. That is, one can recast a traditional logistic regression model as a model where each individual has an unobserved quantitative trait, and individuals whose quantitative trait is greater than some threshold demonstrate a positive binary trait. Effects of covariates (such as SNPs) add or subtract to the unobserved quantitative trait; when the liability has a logistic distribution (slightly heavier tailed than a Gaussian distribution), the effects on the latent scale are the same as lORs. Consider a logistic model shown in Figure A1; the plotted curve is the density of the latent liability and the dotted line the threshold above which individuals are affected by disease and below which they are unaffected if the base rate of the disease is 5%. The area under the curve to the right of the threshold are affected individuals and to the left of the threshold unaffected. The blue area under the curve is the fraction of individuals in the population who, with a moderately risk-increasing SNP, would cross the liability threshold and become affected by the disease. The smaller red area is the individuals who, with a risk-decreasing SNP of the same magnitude, lOR would cross the liability threshold in the other direction and cease to be affected.

Figure A1. Density of latent trait before SNP effects. The dotted line indicates the case threshold. The blue area corresponds to controls that become cases if possessing an SNP with OR = 1.6. The red area indicates cases that become controls if possessing an SNP with lOR = 1/1.6.



When G_i is a vector of genotypes for person i, β are SNP lORs Gaussian distributed with mean μ and standard deviation σ, c is the threshold above, $I()$ the indicator function and X_i the latent liability, then we can write:

$$Y_i|g_i, \mu, \sigma = I(X_i + G_i\beta > c) \tag{A1}$$

One can approximate a logistic variable by a Gaussian scaled by 1.6, yielding:

$$Y_i|g_i, \mu, \sigma = I(Z_i\sqrt{1.6^2 + \sigma^2 g_i} + \mu g_i > c) \tag{A2}$$

for Z_i, a standard normal. One can then re-apply the normal-logistic approximation:

$$Y_i|g_i, \mu, \sigma = I(X_i^* + \frac{\mu g_i - c}{\sqrt{1 + \sigma^2 g_i/1.6^2}} > 0) \tag{A3}$$

where X^* is again logistic distributed, returning to a usual form for logistic regression.

Conflicts of Interest

The authors declare no conflict of interest.

References

1. The International HapMap Consortium. A haplotype map of the human genome. *Nature* **2005**, *437*, 1299–1320.
2. Hindorff, L.A.; Sethupathy, P.; Junkins, H.A.; Ramos, E.M.; Mehta, J.P.; Collins, F.S.; Manolio, T.A. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl. Acad. Sci. USA* **2009**, *106*, 9362–9367.
3. Burn, J. Should we sequence everyone's genome? Yes. *BMJ* **2013**, *346*, doi:10.1136/bmj.f3133.
4. Guan, W.; Pluzhnikov, A.; Cox, N.J.; Boehnke, M.; International Type 2 Diabetes Linkage Analysis Consortium. Meta-analysis of 23 type 2 diabetes linkage studies from the International Type 2 Diabetes Linkage Analysis Consortium. *Hum. Hered.* **2008**, *66*, 35–49.

5. Lee, S.; Wu, M.C.; Lin, X. Optimal tests for rare variant effects in sequencing association studies. *Biostatistics* **2012**, *13*, 762–775.
6. Mathieson, I.; McVean, G. Differential confounding of rare and common variants in spatially structured populations. *Nat. Genet.* **2012**, *44*, 243–246.
7. Liu, Q.; Nicolae, D.L.; Chen, L.S. Marbled inflation from population structure in gene-based association studies with rare variants. *Genet. Epidemiol.* **2013**, *37*, 286–292.
8. Babron, M.C.; de Tayrac, M.; Rutledge, D.N.; Zeggini, E.; Gnin, E. Rare and Low Frequency Variant Stratification in the UK Population: Description and Impact on Association Tests. *PLoS One* **2012**, *7*, e46519.
9. Li, B.; Leal, S. Methods for Detecting Associations with Rare Variants for Common Diseases: Application to Analysis of Sequence Data. *Am. J. Hum. Genet.* **2008**, *83*, 311–321.
10. Madsen, B.E.; Browning, S.R. A Groupwise Association Test for Rare Mutations Using a Weighted Sum Statistic. *PLoS Genet.* **2009**, *5*, e1000384.
11. Caspi, A.; Moffitt, T.E. Gene-environment interactions in psychiatry: Joining forces with neuroscience. *Nat. Rev. Neurosci.* **2006**, *7*, 583–590.
12. Hunter, D.J. Gene-environment interactions in human diseases. *Nat. Rev. Genet.* **2005**, *6*, 287–298.
13. Coventry, A.; Bull-Otterson, L.M.; Liu, X.; Clark, A.G.; Maxwell, T.J.; Crosby, J.; Hixson, J.E.; Rea, T.J.; Muzny, D.M.; Lewis, L.R.; *et al*. Deep resequencing reveals excess rare recent variants consistent with explosive population growth. *Nat. Commun.* **2010**, *1*, 131.
14. Keinan, A.; Clark, A.G. Recent explosive human population growth has resulted in an excess of rare genetic variants. *Science* **2012**, *336*, 740–743.
15. Pritchard, J.K.; Cox, N.J. The allelic architecture of human disease genes: Common disease-common variant... or not? *Hum. Mol. Genet.* **2002**, *11*, 2417–2423.
16. Pritchard, J.K. Are Rare Variants Responsible for Susceptibility to Complex Diseases? *Am. J. Hum. Genet.* **2001**, *69*, 124–137.
17. Manolio, T.A.; Collins, F.S.; Cox, N.J.; Goldstein, D.B.; Hindorff, L.A.; Hunter, D.J.; McCarthy, M.I.; Ramos, E.M.; Cardon, L.R.; Chakravarti, A.; *et al*. Finding the missing heritability of complex diseases. *Nature* **2009**, *461*, 747–753.
18. Eyre-Walker, A. Evolution in Health and Medicine Sackler Colloquium: Genetic architecture of a complex trait and its implications for fitness and genome-wide association studies. *Proc. Natl. Acad. Sci. USA* **2010**, *107*, 1752–1756.
19. Gorlov, I.P.; Gorlova, O.Y.; Sunyaev, S.R.; Spitz, M.R.; Amos, C.I. Shifting Paradigm of Association Studies: Value of Rare single-nucleotide polymorphisms. *Am. J. Hum. Genet.* **2008**, *82*, 100–112.
20. Li, B.; Leal, S.M. Discovery of Rare Variants via Sequencing: Implications for the Design of Complex Trait Association Studies. *PLoS Genet.* **2009**, *5*, e1000481.
21. Zeger, S.L.; Liang, K.; Albert, P.S. Models for Longitudinal Data: A Generalized Estimating Equation Approach. *Biometrics* **1988**, *44*, 1049–1060.
22. Neuhaus, J.M.; Kalbfleisch, J.D.; Hauck, W.W. A Comparison of cluster-specific and population-averaged Approaches for Analyzing Correlated Binary Data. *Int. Stat. Rev. Rev. Int. Stat.* **1991**, *59*, 25–35.

23. Subramanian, S.V.; O'Malley, A.J. Modeling neighborhood effects: The futility of comparing mixed and marginal approaches. *Epidemiology* **2010**, *21*, 475–478; discussion 479–481.
24. Longmate, J.A.; Larson, G.P.; Krontiris, T.G.; Sommer, S.S. Three Ways of Combining Genotyping and Resequencing in Case-Control Association Studies. *PLoS One* **2010**, *5*, e14318.
25. Curtin, K.; Iles, M.M.; Camp, N.J. Identifying rarer genetic variants for common complex diseases: Diseased *versus* neutral discovery panels. *Ann. Hum. Genet.* **2009**, *73*, 54–60.
26. Edwards, T.L.; Song, Z.; Li, C. Enriching Targeted Sequencing Experiments for Rare Disease Alleles. *Bioinformatics* **2011**, *27*, 2112–2118.
27. Yang, F.; Thomas, D.C. Two-Stage Design of Sequencing Studies for Testing Association with Rare Variants. *Hum. Hered.* **2011**, *71*, 209–220.
28. King, C.R.; Rathouz, P.J.; Nicolae, D.L. Generalizing from sequencing studies. *arXiv* **2013**, arXiv:1312.7714.
29. Clayton, D.; Chapman, J.; Cooper, J. Use of unphased multilocus genotype data in indirect association studies. *Genet. Epidemiol.* **2004**, *27*, 415–428.
30. King, C.R.; Rathouz, P.J.; Nicolae, D.L. An Evolutionary Framework for Association Testing in Resequencing Studies. *PLoS Genet.* **2010**, *6*, e1001202.
31. Zelterman, D.; Chen, C. Homogeneity Tests Against Central-Mixture Alternatives. *J. Am. Stat. Assoc.* **1988**, *83*, 179–182.
32. Morris, A.P.; Zeggini, E. An evaluation of statistical approaches to rare variant analysis in genetic association studies. *Genet. Epidemiol.* **2010**, *34*, 188–193.
33. Price, A.L.; Kryukov, G.V.; de Bakker, P.I.; Purcell, S.M.; Staples, J.; Wei, L.; Sunyaev, S.R. Pooled Association Tests for Rare Variants in Exon-Resequencing Studies. *Am. J. Hum. Genet.* **2010**, *86*, 832–838.
34. Bansal, V.; Libiger, O.; Torkamani, A.; Schork, N.J. An application and empirical comparison of statistical analysis methods for associating rare variants to a complex phenotype. In Proceedings of the Pacific Symposium on Biocomputing, Kohala Coast, HI, USA, 3–7 January 2011; pp. 76–87.
35. Schaid, D.J. Genomic Similarity and Kernel Methods I: Advancements by Building on Mathematical and Statistical Foundations. *Hum. Hered.* **2010**, *70*, 109–131.
36. Schaid, D.J. Genomic Similarity and Kernel Methods II: Methods for Genomic Information. *Hum. Hered.* **2010**, *70*, 132–140.
37. Pan, W. Relationship between genomic distance-based regression and kernel machine regression for multi-marker association testing. *Genet. Epidemiol.* **2011**, *35*, 211–216.
38. Hofmann, T. Kernel methods in machine learning. *Ann. Stat.* **2008**, *36*, 1171–1220.
39. Ladouceur, M.; Dastani, Z.; Aulchenko, Y.S.; Greenwood, C.M.T.; Richards, J.B. The Empirical Power of Rare Variant Association Methods: Results from Sanger Sequencing in 1,998 Individuals. *PLoS Genet.* **2012**, *8*, e1002496.
40. Xu, C.; Ladouceur, M.; Dastani, Z.; Richards, J.B.; Ciampi, A.; Greenwood, C.M.T. Multiple Regression Methods Show Great Potential for Rare Variant Association Tests. *PLoS One* **2012**, *7*, e41694.
41. Sul, J.H.; Han, B.; He, D.; Eskin, E. An Optimal Weighted Aggregated Association Test for Identification of Rare Variants Involved in Common Diseases. *Genetics* **2011**, *188*, 181–188.

42. Wu, M.; Lee, S.; Cai, T.; Li, Y.; Boehnke, M.; Lin, X. Rare-variant Association Testing for Sequencing Data with the Sequence Kernel Association Test. *Am. J. Hum. Genet.* **2011**, *89*, 82–93.
43. Meng, X.L. Posterior Predictive *p*-Values. *Ann. Stat.* **1994**, *22*, 1142–1160.
44. Bayarri, M.J.; Castellanos, M.E. Bayesian Checking of the Second Levels of Hierarchical Models. *Stat. Sci.* **2007**, *22*, 322–343.
45. Gelman, A. Comment: Bayesian Checking of the Second Levels of Hierarchical Models. *Stat. Sci.* **2007**, *22*, 349–352.
46. Yi, N.; Zhi, D. Bayesian analysis of rare variants in genetic association studies. *Genet. Epidemiol.* **2011**, *35*, 57–69.
47. Sun, L.; Craiu, R.V.; Paterson, A.D.; Bull, S.B. Stratified false discovery control for large-scale hypothesis testing with application to genome-wide association studies. *Genet. Epidemiol.* **2006**, *30*, 519–530.
48. Roeder, K.; Wasserman, L. Genome-Wide Significance Levels and Weighted Hypothesis Testing. *Stat. Sci.* **2009**, *24*, 398–413.