Evaluating basic approaches to post-hoc analysis for commonly used, gene-based rare variant tests of association

To date, gene-based rare variant testing approaches have focused on maximizing statistical power to identify genes showing significant association with disease. The test statistics can accommodate a combination of risk-increasing, risk-reducing, and non-causal variants and can weight each variant. Increasingly complex test statistics and weighting strategies may improve power, but may also hinder interpretation of a significant association. Identifying causal variant(s) in the gene and estimating their effect is crucial for planning replication studies and characterizing the genetic architecture of the locus. Recent work by our group has classified general characteristics of two classes of gene-based rare variant tests. Using this framework, we have explored the ramifications of choice of gene-based test on post-hoc analyses attempting to identify causal variants. Furthermore, we have evaluated the overall quality and consistency of different single-marker association statistics in identifying causal variants within a gene. To conclude, we offer suggestions regarding the future use of post-hoc analysis methods.

Introduction

In 2003, the Human Genome Project accomplished what had never been done before: the sequencing of the entire human genome. Since then, sequencing technology has advanced rapidly and brought with it a vast amount of complicated data, necessitating the development of new statistical methodology. The possibilities presented by this quantity of data and its potential for broader impacts are fascinating. These data can be used to attack complex problems such as heart disease, cancer, and mental illness that affect the life of nearly every human being.

An early approach to analyzing this genotypic data, genome-wide association studies (GWAS), uses single-marker tests to identify common genetic single nucleotide variants (SNVs) associated with diseases (otherwise known as causal variants). However, studies have shown that these identified common variants do not account for all heritability known to be associated with many of the complex diseases that have been studied [Schork et al., 2009; Manolio et al., 2009; Eichler et al., 2010]. This has given rise to the search for rare variants which are, unsurprisingly, difficult to find. In recent years, various gene-based association tests have been proposed that combine signals across rare variants within the same gene in order to improve power [Morgenthaler and Thilly, 2007; Li and Leal, 2008; Madsen and Browning, 2009; Han and Pan, 2010; Morris and Zeggini, 2010; Price et al., 2010; Zawistowski et al., 2010; Basu and Pan, 2011; Feng et al., 2011; Ionita-Laza et al., 2011; Lin and Tang, 2011; Neale et al., 2011; Pan and Shen, 2011; Sul et al., 2011; Wu et al., 2011; Zhang et al., 2011; Dai et al., 2012]. Previous work has used a geometric framework to classify these gene-based tests into two categories: length and joint tests [Liu et al., 2013].

The goal of gene-based, rare variant tests is to reduce the multiple testing penalties associated with performing single-marker tests across the entire genome and improve power in the identification of causal rare variants. In case-control studies, most of these gene-based tests consider the null hypothesis that there is no significant difference in the number of minor alleles observed among the cases versus the controls. In other words, they test the null hypothesis that no variant within the gene is significantly associated with the phenotype. Thus a significant gene-based rare variant test result informs you that at least one variant in the gene is significantly associated with the disease phenotype.

It is important to note, however, that a significant gene-based test result does not tell you precisely which variant(s) are associated. A very natural follow-up question, and the principal question of this research project, is: Which of these variants is most likely to be associated with the disease phenotype? This situation is in many ways analogous to a one-way ANOVA from which one can conclude that at least one group mean differs from the others and its post-hoc tests to determine precisely which of those group means is different.

Approaches to post-hoc analysis for gene-based rare variant tests vary greatly. In this paper we (1) review existing post-hoc analysis methods and use a comprehensive simulation study to (2) evaluate the overall quality and consistency of different single-marker association statistics in identifying the most likely causal variants within a gene and (3) explore the ramifications of the choice of initial gene-based test on results of post-hoc analysis.

Background

When there is evidence that at least one variant in a gene is significantly associated with the phenotype, it seems straightforward to conclude that the variant with the strongest single-marker association statistic (difference in minor allele counts (*d*) between cases and controls, relative risk (*r*), Permutation Test with test statistic *d*, or Fisher's Exact Test statistic, for example) is the most likely of these variants to be associated with the disease phenotype. However, using this approach researchers have often been able to identify only a single, reasonably common variant and struggle to identify causal variants with low minor allele frequency (MAF) [Tintle et al., 2011]. Alternatively, many individuals attempt to incorporate biological information into analysis by, for example, filtering non-synonymous variants or variants in a certain location [Khetarpal et al., 2011; Stitziel et al., 2011; Bick et al., 2012; Kathiresan and Srivastava, 2012]. This provides further information about which variants or set of variants are the most plausible candidates for disease phenotype association. Other researchers have proposed Bayesian methods which attempt to incorporate prior biological information via prior distributions [Maller et al., 2012; Zhang et al., 2012]. Continued research is underway to develop methods to best identify the one variant most likely to be causal or to better identify the entire subset of causal variants.

Unfortunately, success in post-hoc analysis has been limited, and based on a comprehensive literature review (details not provided here) it is clear that there is a general lack of consensus about best practices for identification of the most likely causal variant(s) in post-hoc analysis. Identification of rare causal variants has the potential to improve our understanding of complex diseases such as heart disease, cancer, and depression and can lead to improvements in personalized medicine. Thus it becomes very important to improve upon existing methods of post-hoc analysis. This, along with the limited success and lack of consensus regarding best practices for post-hoc analysis, motivates our work.

Methods

Tests used

We break our rare variant association testing procedure into two stages. For each gene, we begin at Stage 1 with a gene-based test for association between the gene and disease phenotype. If a significant result is achieved at Stage 1 (p-value < 0.05), we proceed to Stage 2: post-hoc analysis. If a significant result is not achieved at Stage 1, we do not conduct post-hoc analysis.

Stage 1: Gene-based tests

As noted earlier, most gene-based rare variant tests of association can be classified into one of two broad classes of tests: length or joint [Liu et al., 2013]. This means that most rare variant test statistics can be written as functions of the generally stated test statistics as defined immediately below:

General Length Test Statistic:

$$L_p = (\sum_{i=1}^{m} |\frac{c_i^+}{2N^+}|^p)^{1/p} - (\sum_{i=1}^{m} |\frac{c_i^-}{2N^-}|^p)^{1/p}$$

General Joint Test Statistic:

$$J_p = (\sum_{i=1}^{m} |\frac{c_i^+}{2N^+} - \frac{c_i^-}{2N^-}|^p)^{1/p}$$

We define *m* to be the number of single nucleotide variants within the gene; N^+ and N^- indicate the number of cases and controls, respectively; c_i^+ and c_i^- are the observed number of minor alleles at variant *i = 1,…, m*, within the case and control samples, respectively; and *p* reflects the choice of L^p norm. Thus, in light of the geometric framework introduced by Liu et al. [2013], length tests compare the lengths (or magnitudes) of the *m*-long vector of minor allele frequency estimates for the cases and controls, while joint tests compare both the lengths of these minor allele frequency vectors and the angle between them.

To date, most published length tests use *p=1*, while most joint tests use *p=2*. In our first set of simulations we conduct Proportion Regression (PR) [Morris and Zeggini, 2010] and Sequence Kernel Association Test [Wu et al., 2011] on each gene. These two gene-based rare variant tests are approximately equivalent to the generic length test with *p=1* and a joint test with *p=2*, respectively [Liu et al., 2013]. In our second set of simulations we consider generic versions of length and joint tests with *p=1, 2, 4,* and *Infinity*.

Additionally, for each simulation setting we keep track of Stage 1 test results so that we may calculate Stage 1 power, or the percent of the 10,000 simulated data sets under a particular simulation setting that yield a p-value smaller than 0.05 for each Stage 1 test.

Stage 2: Post-hoc analysis

Post-hoc analysis is carried out on all variants within significant Stage 1 genes (p-value < 0.05) using four straightforward methods. For each iteration of each simulation setting, we calculate the difference in minor allele counts between cases and controls ($d = c_i^+ - c_i^-$), relative risk ($r = c_i^+ / c_i^-$), and p-value for a permutation test (1,000 permutations) with test statistic *d* (*pp*) and Fisher's Exact Test (*fp*). Variants are then ranked from most likely causal variant to least likely, or from largest to smallest by *d* and *r* and smallest to largest by value of *pp* and *fp*.

It is important to note limitations with the calculation of *r*. In the simulation of rare variants, it is not uncommon to observe values of $c_i^- = 0$, which causes a major issue for the computation of $r = c_i^+ / c_i^-$. One way of avoiding this issue is to instead calculate $r_t = (c_i^+ + 0.001) / (c_i^- + 0.001)$. In the calculation of r_t we are able to avoid the issue of division by zero. However, it could be argued that this transformation is still not ideal (see discussion, below).

Simulations

We use two simulation studies to evaluate the overall quality and consistency of different single-marker association statistics in identifying the most likely causal variants within a gene and explore the ramifications of the choice of initial gene-based test on results of post-hoc analysis.

Simulation settings: Initial study

In the first simulation study, we simulate case-control data for 1,000 cases and 1,000 controls. Each simulated gene contains either eight or sixteen variants. There are four distributions of risk-increasing, risk-neutral, and risk-reducing variants with ratios (1) 25:75:0, (2) 50:50:0, (3) 75:25:0, and (4) 25:50:25. In half of the simulations, relative risk is minor allele frequency (MAF)-independent: risk-increasing variants have a constant relative risk of 1.5 and risk-reducing variants have a constant relative risk of 0.67. The other half of simulations use an MAF-dependent relative risk: (1) causal variants with MAF=0.01% have a relative risk of 5 (risk-increasing) or 0.2 (risk-reducing), (2) causal variants with MAF=0.1% have a relative risk of 3 (risk-increasing) or 0.33 (risk-reducing), and (3) causal variants with MAF=1% have a relative risk of 1.5 (risk-increasing) or 0.67 (risk-reducing). Finally, the distribution of MAF of the variants within each gene vary: either (1) all MAF are constant at (a) 0.01%, (b) 0.1%, or (c) 1% or (2) MAF varies at a 3:1 or 7:1 ratio of (a) 0.01% to 0.1%, (b) 0.01% to 1%, or (c) 0.1% to 1%. Note that in the case of a 7:1 ratio of low to high MAF, all high MAF variants are neutral. All combinations of number of variants (2), risk distributions (8), and MAF distributions (9) are considered for a total of 144 simulation settings. We simulate 10,000 sets of genotype-phenotype data for each setting. One length test (PR) and one joint test (SKAT) are applied to the simulated genotype-phenotype matrices.

Simulation settings: Follow-up study

In the follow-up simulation study we consider additional gene-based rare variant tests in order to explore the impact of the choice of norm *p* on post-hoc analysis. In this study, we again simulate 1,000 cases and 1,000 controls. Genes contain either eight or thirty-two variants. Relative risks are constant at 1.5 for risk-increasing variants, 1 for neutral variants, and 0.67 for risk-reducing variants. Ratios of risk-increasing, neutral and risk-reducing variants are (1) 25:75:0, (2) 50:50:0, (3) 75:25:0, (4) 25:50:25, and (5) 50:25:25. The distribution of MAF of the variants is either (1) constant at 0.5% or 0.05%, (2) distributed at a 3:1 ratio of low (0.05% or 0.5%) to high (1% or 20%) MAF among risk-increasing, neutral, and risk-reducing variants, or (3) low for all variants except one neutral variant. All combinations of number of variants (2), risk distributions (5), and MAF distributions (10) are considered for a total of 100 simulation settings. Again, we simulate 10,000 data sets at each setting. Four length tests (*p=1, 2, 4,* and *Infinity*) and four joint tests (*p=1, 2, 4,* and *Infinity*) are applied to each simulated genotype-phenotype matrix.

Results

Evaluation of ability to rank causal variants

Overall, performance of all considered post-hoc methods was unsatisfying. Before detailing the results of this analysis, it is important to mention the way in which we have evaluated the “performance” of post-hoc analysis methods. There are many methods of comparison that might have been used. However, our group has decided to focus first on the ability of each method to correctly identify a causal variant as the top-ranked (largest *d* or *r*, smallest *fp* or *pp*) variant. We feel this method of evaluation is justified in that a significant Stage 1 test result tells us that at least one variant within that gene is associated with the disease phenotype, and thus any successful post-hoc method should at least be able to identify one variant most likely to be causal. Later we also consider the ability of each method to correctly identify a causal variant as the second, third, etc. most likely to be causal, depending on the true number of causal variants in the gene.

The poor performance of simple single-marker association statistics *d* and *r* is not entirely surprising. Alone, these statistics do not capture all necessary information. For example, an observed number of minor allele counts for cases-controls of 0-2 and 8-10 have the same value of $d = -2$, but represent two situations that we might in reality want to distinguish between. Similarly, an observation of 1-2 or 5-10 represents the same value of $r = 0.5$, but again represents two distinct situations that do not provide the same amount of evidence for association with disease phenotype. We might, for example, consider an observation of minor allele counts in cases-controls of 5-10 to be more convincing evidence of a true relative risk of 0.5, while an observation of 1-2 is more likely to have happened by chance under the null hypothesis of no association with disease phenotype. With this in mind, it is not surprising to note the relatively poor performance of these single-marker post-hoc methods in ranking causal variants. We find that under some simulation settings *d* proves effective in identifying the top-ranked variant, and transformed relative risk (r_t) in others (Table 1, Fig. 1). However, both *d* and r_t struggle when it comes to the identification of any causal variants beyond the most common (Fig. 2).

Furthermore, even other, slightly more sophisticated methods—such as single-marker p-values from Fisher's Exact Test or a permutation test that rely on *d* as their test statistic but incorporate additional information, such as variability, in hopes of better post-hoc analysis results—still struggle in the identification of causal variants. We see that these methods perform best when all variants within the gene have the same low MAF (Table 1, Fig. 3). However, the methods struggle in other settings and do not always out-perform simpler methods such as *d* and *r,* as we might have hoped (Table 1, Fig. 1).

It is clear from Table 1 that none of the post-hoc methods evaluated here is ideal across all genetic architectures in the correct identification of the most likely causal variant. We have broken down our 144 simulation settings into three main categories that seem to have the greatest impact on which test is most effective at identifying a risk-increasing variant. The first category includes all simulation settings for which all variants have the same minor allele frequency. Under this condition, *d* is the single method that most often correctly identifies the most likely causal variant better than the other methods, though *pp* performs relatively well under this condition also. The second condition is a 3:1 ratio of low MAF to high MAF variants, spread across both causal and non-causal variants. Again under this condition *d* performs best most often, and *pp* performs best less often than in the category. Finally, the third category represents simulation settings with a 7:1 ratio of low to high MAF variants in which only risk-neutral variants have the high MAF. In this category of simulation settings, r_t performs best most often and *d* never performs best.

Type of Gene	Stage 1 Test	Difference in minor allele counts (*d*)	Transformed relative risk (*r_t*)	Permutation Test p-value (*pp*)	Fisher's Exact Test p-value
Constant MAF	SKAT	54.2%	0.0%	37.5%	8.3%
	PR	62.5%	0.0%	33.3%	4.2%
3:1 Low:High MAF	SKAT	87.5%	0.0%	12.5%	0.0%
	PR	54.2%	4.2%	25.0%	16.7%
7:1 Low:High MAF	SKAT	0.0%	83.3%	16.7%	0.0%
	PR	0.0%	83.3%	16.7%	0.0%

Table 1. Percent of simulation settings under which each of the post-hoc methods (difference in minor allele counts, relative risk, and permutation test p-value) identified a causal variant as the top-ranked variant more frequently than the other post-hoc methods.

The results summarized in Table 1 are also supported by Figure 1. This figure shows that none of the methods consistently has the best power in terms of correct identification of a causal variant as the top-ranked variant. It is interesting to note that when SKAT was the Stage 1 test, the permutation test p-values are consistently within five percent or greater power in comparison to difference in minor allele counts. This pattern, however, does not hold when we conducted PR at Stage 1.

What's more, Figure 2 shows that identification of causal variants beyond the most likely becomes even more complicated. In Figure 2 we see that a vast majority of the time the neutral variant with MAF=1% has a larger value of *d* than the three causal variants with MAF=0.05%. Thus a ranking strategy based on *d* often incorrectly ranks the common, neutral variant as the second-most likely causal variant. This pattern holds across many other simulation settings and post-hoc methods, although details are not provided here.

Understanding why performance is poor

Winner's curse is a well-known phenomenon in the analysis of single-marker data (for example, in GWAS) [Lohmueller et al., 2003; Zollner and Pritchard, 2007; Xiao and Boehnke, 2009], whereby the estimated effects of causal markers are substantially upwardly biased. In other words, the estimated effects of significant markers tend to over-estimate the true effects of these markers. The extent of the impact of winner's curse has been shown to be directly related to the power of the study, with more powerful study designs exhibiting fewer problems with winner's curse.

We have found that winner's curse has a substantial impact on post-hoc analysis methods. The distribution of causal markers *r* and d are substantially higher than in the population, especially for variants with relatively large MAF. But this problem is not only observed among causal variants. Common, neutral variants also have distributions of *r* that are centered at numbers further from 1, and distributions of *d* centered at numbers further from 0 than would be expected under the null hypothesis of no association with disease phenotype. This could explain the poor performance of post-hoc analysis methods. Consider the example of a Fisher's Exact Test or permutation test that test the null hypothesis that $d = 0$, when in fact we have discovered that in many cases $d \neq 0$ under the null hypothesis of no association with disease phenotype.

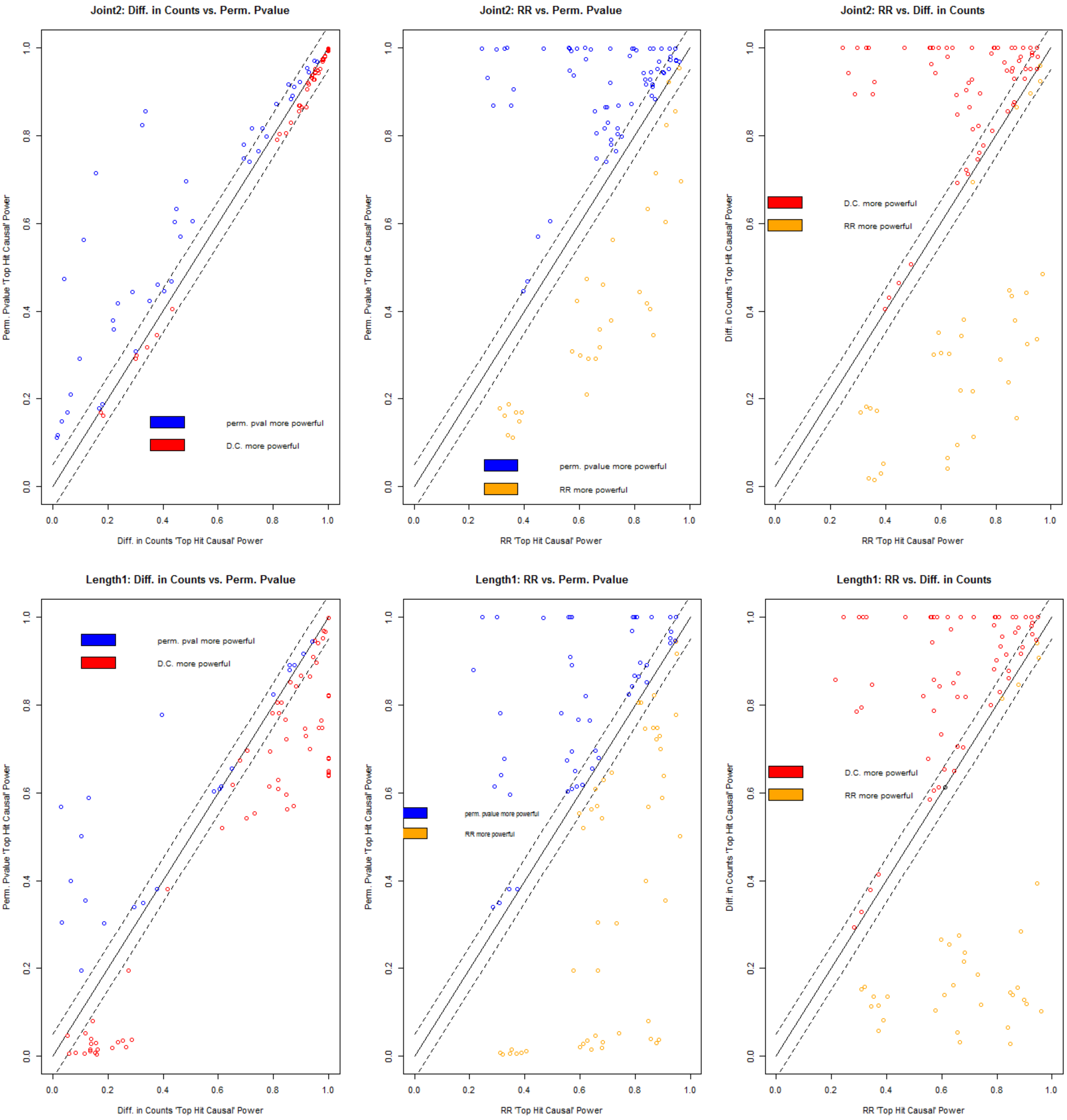

Figure 1. Pairwise comparisons of single-marker association statistics *d, r,* and *pp,* comparing the percent of times that the methods correctly identified a causal variant as the top-ranked variant for each simulation setting.

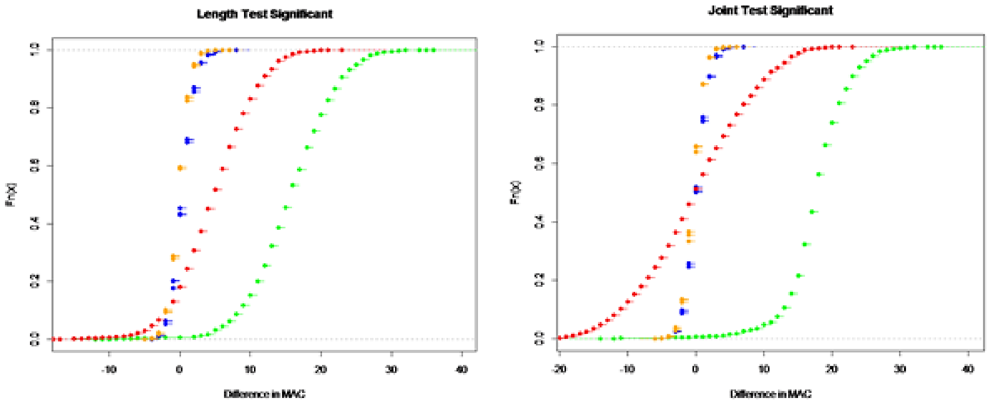

Figure 2. Empirical cumulative distribution function for difference in minor allele counts (*d*) for each variant for the simulation setting with eight variants, a 50:50:0 ratio of risk-increasing to neutral to risk-reducing variants, and a 3:1 ratio of low (0.05%) to high (1%) MAF. Results are shown after PR (Length Test) and SKAT (Joint Test) were conducted at Stage 1. The green ECDF represents the risk-increasing variant with MAF=1%, red represents the neutral variant with MAF=1%, blue represents the risk-increasing variants with MAF=0.05%, and orange represents the neutral variants with MAF=0.05%.

Bias is introduced when we condition post-hoc analysis on Stage 1 significance. We refer to this bias as “Stage 1 bias.” Figure 2 shows the impact that winner’s curse has on post-hoc methods. We see that, especially when a length test with norm 1 (PR) is conducted at Stage 1, a common neutral variant exhibits a larger value of *d* than a rare causal variant almost eighty-five percent of the time. We would expect, based on the distributional properties of *d*, that in an un-biased situation this should only occur fifty percent of the time (and in the ideal post-hoc analysis situation it would never occur). This bias severely affects our ability to correctly identify causal variants for all post-hoc methods based on *d* or *r* as their test statistic.

Under simulation settings where Stage 1 bias is not as present, we see that post-hoc performance is better. From Table 1 we see that *pp* and *fp* are best at identifying the top-ranked variant as causal when all variants have a constant, low MAF. Under this condition, these methods also do a better job of identifying causal variants beyond just the one most likely (Fig. 3). We see in Figure that 3 that after both a length test (p=1) and a joint test (p=2), we are able to see a separation of the causal variants (blue and green) and neutral variants (red and orange). Although *pp* is not often smaller than 0.05 even for truly causal variants, it is at least the case that *pp* tends to be smaller for causal variants relative to non-causal variants.

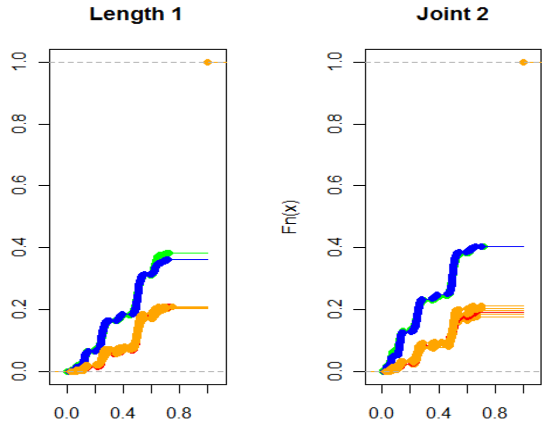

Figure 3. Empirical cumulative distribution function for permutation test p-values at each variant for the simulation setting with eight variants, constant MAF and a 50:50:0 ratio of risk-increasing to neutral to risk-reducing variants. As in Fig. 2, the green and blue ECDs represent the risk-increasing variants and the orange and red ECDFs represent the neutral variants.

Implications of Stage 1 significance

The Figures and Table above show that post-hoc analysis performance varies depending on which Stage 1 test was conducted (PR or SKAT). Particularly, we notice that length tests appear to be more affected by Stage 1 bias (winner's curse) than joint tests. In Figure 2 we see that the common, neutral variant is ranked as second most likely to be causal based on difference in minor allele counts (*d*) almost eighty-five percent of the time when a length test with norm 1 (PR) was conducted at Stage 1, versus only fifty percent of the time when a joint test with norm 2 (SKAT) was conducted at Stage 1.

Observing this difference in post-hoc performance after distinct Stage 1 tests motivates our follow-up simulation study, which was developed to explore the impact of choice of norm on post-hoc results. In general, we observe that higher norms tend to be less affected by Stage 1 bias. Across various simulation settings considered below in Table 2, the length test with *p=Infinity* is, of the length tests, least affected by Stage 1 bias, but it is still more affected by bias than joint tests. For example, we see from Table 2 that for a gene with eight variants, a 50:50:0 ratio of risk-increasing to neutral to risk-reducing variants, and low MAF (0.5%) for all variants except one neutral variant with higher MAF (1%), relative risk of the neutral variant with higher MAF is overestimated as often as sixty-four percent of the time after a length test with norm 1, versus closer to fifty percent of the time for the other Stage 1 tests considered. In an unbiased situation we would expect this to happen, on average, fifty percent of the time.

Simulation Setting		**Joint, p=2**	**Joint, p=Inf**	**Length, p=1**	**Length, p=Inf**
Relative risk distribution	**MAF distribution**				
50:50:0	3:1, .05% to 1%	0.4889	0.4792	0.515	0.4388
50:50:0	3:1, .5% to 1%	0.4774	0.467	0.6082	0.4857
25:75:0	3:1, .05% to 1%	0.4945	0.4841	0.5323	0.4578
50:50:0	7:1, .5% to 1%	0.4896	0.4774	0.6433	0.5043
50:50:0	7:1, .05% to 1%	0.5348	0.5282	0.5741	0.4993

Table 2. The proportion of the data sets simulated under each genetic architecture (and significant at each specified Stage 1 test) in which the empirical relative risk for the high MAF, neutral variant is larger than 1. All simulated genes considered here have eight variants.

Discussion

The identification of rare variants could lead to improvements in our understanding of complex diseases and developments in personalized medicine. Unfortunately, the task of identifying these variants, which by definition are so difficult to find, is a challenging one. Numerous methods have been proposed to combine signals across multiple single nucleotide variants in order to create one gene-based test statistic, reduce multiple testing penalties, and improve power. However, there is no consensus as to which methods should then be used to identify precisely which variants within these genes are associated with the disease phenotype. We have evaluated four straightforward post-hoc analysis methods and have identified various issues that are presented by the task of post-hoc analysis.

Across a variety of genetic architectures, there is no one post-hoc method that consistently outperforms the others. The four single-marker association statistics we evaluated in this study (*d, r, fp,* and *pp*) often exhibit low power even in the detection of the most likely causal variant (Fig. 1). Identification of any causal variant beyond the most likely quickly becomes even more problematic.

We have briefly discussed above that there are limitations to simply using the difference in minor allele counts (*d*) or relative risk (*r*) as the single-marker association statistic, given that each of these statistics alones fails to capture all the necessary information (such as variability) about the difference between the case and control groups. Single-marker association tests that do account for some of this missing information such as Fisher's Exact Test or a permutation test, still struggle to identify the correct subset of causal variants across many genetic architectures. We hypothesize that this poor performance is in many ways a result of a phenomenon similar to winner's curse which we have called Stage 1 bias.

In retrospect, it is not surprising that Stage 1 bias plays a role in post-hoc analysis. Since we are conducting post-hoc analysis on genes that have significant Stage 1 test results, we actually already know some information about the gene and its variants that we are testing. In particular, these Stage 1 tests use a test statistic that incorporates c_i^+ and c_i^- for each $i=1,\ldots,m$. So when considering post-hoc single-marker association statistics that also incorporate c_i^+ and c_i^-, the distributions of these post-hoc statistics will be conditional upon Stage 1 significance. Correctly conditioning upon Stage 1 significance when conducting post-hoc analysis could improve performance. We have already seen that in scenarios when Stage 1 bias is not as present, post-hoc performance improves: we are able to correctly identify the top-ranked variant as causal and better separate causal variants from non-causal.

Knowing that most Stage 1 tests can be classified either as joint or length tests [Liu et al., 2013] that incorporate minor allele counts for cases and controls into their test statistic, it is not surprising that we found evidence of Stage 1 bias across all simulation settings and Stage 1 tests. However, it does seem that post-hoc analysis is often less problematic for joint tests versus length tests, and for tests with higher norms. This raises the question of whether Stage 1 bias is directly related to Stage 1 test power, as is the case for winner's curse.

It is also important to discuss the over-arching goal of post-hoc analysis. The question of post-hoc analysis has been considered extensively in the context of one-way ANOVA. Saville [1990] argues that there is an inherent problem with conducting post-hoc analysis in that it attempts to formulate and test hypotheses in the same study simultaneously. In other words, he might argue in the context of gene-based rare variant tests that we should not use post-hoc analysis to both generate hypotheses regarding which variants are most likely to be causal *and* to test those variants to discover whether they are causal. Rather, we should use post-hoc analysis to generate hypotheses about which variants are causal and future replication studies to see if we can confirm those hypotheses. O'Neill and Wetherill discuss further issues with multiple comparison methods post-one-way ANOVA that discourage the idea of trying to identify the exact subset of causal variants in the gene. Thus we may need to shift our view of post-hoc analysis not as a means to an end, but as an exploratory process in search of evidence of causality that can only be confirmed through further experimentation.

Our study has some limitations. First, we only considered single-marker testing approaches in our evaluation of post-hoc analysis and there are many other single-marker tests that we might have explored. The post-hoc methods that we considered are quite basic and straightforward. We feel, however, that our focus on these straightforward methods has allow us to better understand the basic issue of Stage 1 bias that we believe would affect any post-hoc method, although future work is needed to confirm this.

We have already briefly mentioned an additional limitation in our calculation of relative risk (*r*), given the many instances in which we are dividing by zero. We attempted to correct for this limitation via transformation of relative risk ($r_t = (c_i^+ + 0.001) / (c_i^- + 0.001)$), but this transformation is not ideal. Suppose for a particular variant there is one minor allele observed among the cases and zero among the controls. Then $r_t = 1.0001/0.001 = 1001.$ This is a very high relative risk for a situation that could easily have been a false positive. Future work needs to consider more carefully how best to transform relative risk to avoid the issue of dividing by zero.

Another limitation of our study is our choice of simulation settings. We did our best to consider as many genetic architectures as possible, but it is possible that we have failed to consider a various genetic architectures, under some of which the patterns we observed in our analysis may not exactly hold. The difficulty with generating simulated genotype data is that we still no so little about which genetic architectures are actually observed in nature that it is difficult to simulate “life-like,” realistic data.

Future work might consider more complicated methods that incorporate prior biological information or involve more complicated test statistics. We hypothesize that any method that conditions on the correct biological information, such as dropping out variants that are not expected to be associated with the disease based on biological information, could improve post-hoc performance (especially if the variants that are dropped out are common, neutral variants that are often affected by Stage 1 bias). However, we also believe these more sophisticated methods will still be affected by Stage 1 bias provided their test statistic is based on the difference in minor allele counts between cases and controls (or relative risk, though this is not as often chosen as a test statistic). Future work could be conducted to confirm these hypotheses.

Future work might also explore whether Stage 1 bias is really just a question of Stage 1 power. For example, we might explore the impact of increased sample size on post-hoc analysis performance. We could also explore methods that have been developed for correction of winner’s curse, or methods used in one-way ANOVA post-hoc analysis to see how similar problems have been addressed in other fields.

This summer I plan to continue my work with this project. I hope to begin by quantifying the effect of different factors such as minor allele frequency and relative risk on Stage 1 bias. From there, I will develop an empirical correction for this bias and eventually quantify the conditional distribution of minor allele counts given Stage 1 significance. Once I have completed this task, I will incorporate this incorporate this information into methodology. I will apply any correction developed during the first stage of my research into existing post-hoc methods which depend on minor allele counts. Eventually I will work toward developing new methods specifically designed to account for Stage 1 bias.

Once I develop these methods, I will test them against existing methods on simulated genotype data in order to compare power and false positive rates. I will then apply any tests that produce promising results when applied to simulated data to real genotypic data.

Future work in this area has the potential for great impact. The wide dissemination of improved methods for post-hoc analysis will lead to the identification of more rare variants associated with complex diseases and further understanding of these diseases. Educational and training initiatives will inform people of the new understanding of these diseases, which will lead to more effective genetic counseling, personalized health care, and even decreased stigma associated with diseases such as obesity, heart disease, diabetes, and mental disorders which affect every one of our lives.

References

Basu, Saonli, and Wei Pan. "Comparison of statistical tests for disease association with rare variants." *Genetic epidemiology* 35.7 (2011): 606-619.

Bick, Alexander G., et al. "Burden of rare sarcomere gene variants in the Framingham and Jackson Heart Study cohorts." *The American Journal of Human Genetics* 91.3 (2012): 513-519.

Dai, Yilin, Renfang Jiang, and Jianping Dong. "Weighted selective collapsing strategy for detecting rare and common variants in genetic association study. "*BMC genetics* 13.1 (2012): 7.

Eichler, Evan E., et al. "Missing heritability and strategies for finding the underlying causes of complex disease." *Nature Reviews Genetics* 11.6 (2010): 446-450.

Feng, Tao, Robert C. Elston, and Xiaofeng Zhu. "Detecting rare and common variants for complex traits: sibpair and odds ratio weighted sum statistics (SPWSS, ORWSS)." *Genetic epidemiology* 35.5 (2011): 398-409.

Han, Fang, and Wei Pan. "A data-adaptive sum test for disease association with multiple common or rare variants." *Human heredity* 70.1 (2010): 42-54.

Ionita-Laza, Iuliana, et al. "A new testing strategy to identify rare variants with either risk or protective effect on disease." *PLoS genetics* 7.2 (2011): e1001289.

Kathiresan, Sekar, and Deepak Srivastava. "Genetics of human cardiovascular disease." *Cell* 148.6 (2012): 1242-1257.

Khetarpal, Sumeet A., et al. "Mining the LIPG allelic spectrum reveals the contribution of rare and common regulatory variants to HDL cholesterol." *PLoS genetics* 7.12 (2011): e1002393.

Li, Bingshan, and Suzanne M. Leal. "Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data." *The American Journal of Human Genetics* 83.3 (2008): 311-321.

Lin, Dan-Yu, and Zheng-Zheng Tang. "A general framework for detecting disease associations with rare variants in sequencing studies." *The American Journal of Human Genetics* 89.3 (2011): 354-367.

Liu, Keli, Shannon Fast, Matthew Zawistowski, and Nathan Tintle. ”A Geometric Framework for Evaluating Rare Variant Tests of Association.” *Genetic Epidemiology* 37.4 (2013): 345-57.

Lohmueller, Kirk E., et al. "Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease." *Nature genetics* 33.2 (2003): 177-182.

Madsen, Bo Eskerod, and Sharon R. Browning. "A groupwise association test for rare mutations using a weighted sum statistic." *PLoS genetics* 5.2 (2009): e1000384.

Maller, Julian B., et al. "Bayesian refinement of association signals for 14 loci in 3 common diseases." *Nature genetics* 44.12 (2012): 1294-1301.

Manolio, Teri A., et al. "Finding the missing heritability of complex diseases. "*Nature* 461.7265 (2009): 747-753.

Morgenthaler, Stephan, and William G. Thilly. "A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: a cohort allelic sums test

(CAST)." *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis* 615.1 (2007): 28-56.

Morris, Andrew P., and Eleftheria Zeggini. "An evaluation of statistical approaches to rare variant analysis in genetic association studies." *Genetic epidemiology* 34.2 (2010): 188-193.

Neale, Benjamin M., et al. "Testing for an unusual distribution of rare variants."*PLoS genetics* 7.3 (2011): e1001322.

O'Neill, R., and G.B. Wetheril. "The Present State of Multiple Comparison Methods." *Journal of the Royal Statistical Society* 33.2 (1971): 218-50. *JSTOR*. Web. 15 Oct. 2013.

Pan, Wei, and Xiaotong Shen. "Adaptive tests for association analysis of rare variants." *Genetic epidemiology* 35.5 (2011): 381-388.

Price, Alkes L., et al. "Pooled association tests for rare variants in exon-resequencing studies." *The American Journal of Human Genetics* 86.6 (2010): 832-838.

Saville, D.J. "Multiple Comparison Procedures: The Practical Solution." *The American Statistician* 44.2 (1990): 174-80.

Schork, N.J., Sarah Murray, Kelly Frazer, and Eric Topol. "Common vs. Rare Allele Hypotheses for Complex Diseases." *Current Opinion in Genetics & Development* 19.3 (2009): 212-219.

Stitziel, Nathan O., Adam Kiezun, and Shamil Sunyaev. "Computational and statistical approaches to analyzing variants identified by exome sequencing." *Genome Biol* 12.9 (2011): 227.

Sul, Jae Hoon, et al. "An optimal weighted aggregated association test for identification of rare variants involved in common diseases." *Genetics* 188.1 (2011): 181-188.

Tintle, Nathan, et al. "Inflated type I error rates when using aggregation methods to analyze rare variants in the 1000 Genomes Project exon sequencing data in unrelated individuals: summary results from Group 7 at Genetic Analysis Workshop 17." *Genetic epidemiology* 35.S1 (2011): S56-S60.

Wu, Michael C., et al. "Rare-variant association testing for sequencing data with the sequence kernel association test." *The American Journal of Human Genetics* 89.1 (2011): 82-93.

Xiao, Rui, and Michael Boehnke. "Quantifying and correcting for the winner's curse in genetic association studies." *Genetic epidemiology* 33.5 (2009): 453-462.

Zawistowski, Matthew, et al. "Extending rare-variant testing strategies: analysis of noncoding sequence and imputed genotypes." *The American Journal of Human Genetics* 87.5 (2010): 604-617.

Zhang, Qunyuan, et al. "A data-driven method for identifying rare variants with heterogeneous trait effects." *Genetic epidemiology* 35.7 (2011): 679-685.

Zöllner, Sebastian, and Jonathan K. Pritchard. "Overcoming the winner's curse: estimating penetrance parameters from case-control data." *The American Journal of Human Genetics* 80.4 (2007): 605-615.