Distribution Agreement

In presenting this thesis as a partial fulfillment of the requirements for a degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis in whole or in part in all forms of media, now or hereafter now, including display on the World Wide Web. I understand that I may select some access restrictions as part of the online submission of this thesis. I retain all ownership rights to the copyright of the thesis. I also retain the right to use in future works (such as articles or books) all or part of this thesis.

Joshua Weinstock April 15, 2015

An Investigation of the Properties of Gene-Based Tests of GWAS Data: Simulations and Analyses of ADHD GWAS Data

by

Joshua Weinstock

Advisor:

Irwin Waldman, Ph.D.

Department of Psychology

Irwin Waldman, Ph.D.

Adviser

David Cutler, Ph.D.

Committee Member

Nancy Bliwise, Ph.D.

Committee Member

Lee Cooper, Ph.D.

Committee Member

April 15th, 2015

An Investigation of the Properties of Gene-Based Tests of GWAS Data:

Simulations and Analyses of ADHD GWAS Data

by

Joshua Weinstock

Advisor:

Irwin Waldman, Ph.D.

An abstract of
a thesis submitted to the Faculty of Emory College of Arts and Sciences
of Emory University in partial fulfillment
of the requirements of the degree of
Bachelor of Arts with Honors

Department of Psychology

2015

Abstract

An Investigation of the Properties of Gene-Based Tests of GWAS Data:
Simulations and Analyses of ADHD GWAS Data
By Joshua Weinstock

Genome-wide association studies (GWAS) have become a useful tool in recent years for elucidating the etiology of a variety of complex traits. Although these studies have shown promising findings in the field of psychiatric genetics, conventional single-SNP testing methods pose a number of challenges of both a statistical and biological nature. These challenges make alternative association testing methods, such as gene-based tests, attractive. In the current study, I used simulations and analyses of real data to evaluate the properties of several of the aforementioned gene-based tests, as well as to attempt to answer some questions raised by the application of several gene-based tests to real ADHD GWAS data from four samples. The issues addressed by the simulations included an examination of their Type I Error rates and statistical power across several plausible scenarios, as well as the impact of differing LD among the SNPs on the gene-based tests that use summary data and differences in test performance with genes of differing length, number of LD blocks, and number of causal variants per block. Analyses of the real ADHD GWAS data were used to evaluate differences among the gene-based tests in finding associated genes.

An Investigation of the Properties of Gene-Based Tests of GWAS Data: Simulations and Analyses of ADHD GWAS Data

By

Joshua Weinstock

Adviser:

Irwin Waldman, Ph.D.

A thesis submitted to the Faculty of Emory College of Arts and Sciences
of Emory University in partial fulfillment
of the requirements of the degree of
Bachelor of Arts with Honors

Department of Psychology

2015

Acknowledgements

I would like to thank Dr. Irwin Waldman for his mentorship and guidance. It goes without saying that he was an incredible resource and that I learned a great deal under his supervision. Next I would like to thank Dr. Hasse Walum, who was immensely helpful with the construction of the simulations. I would also like to express my gratitude towards Dr. David Cutler, Dr. Nancy Bliwise, and Dr. Lee Cooper for serving on my honors committee.

Table of Contents

List of Tables

An Investigation of the Properties of Gene-Based Tests of GWAS Data:

Simulations and Analyses of ADHD GWAS Data

Joshua S. Weinstock, advised by Dr. Irwin Waldman

Emory University

Abstract

Genome-wide association studies (GWAS) have become a useful tool in recent years for elucidating the etiology of a variety of complex traits. Although these studies have shown promising findings in the field of psychiatric genetics, conventional single-SNP testing methods pose a number of challenges of both a statistical and biological nature. These challenges make alternative association testing methods, such as gene-based tests, attractive. In the current study, I used simulations and analyses of real data to evaluate the properties of several of the aforementioned gene-based tests, as well as to attempt to answer some questions raised by the application of several gene-based tests to real ADHD GWAS data from four samples. The issues addressed by the simulations included an examination of their Type I Error rates and statistical power across several plausible scenarios, as well as the impact of differing LD among the SNPs on the gene-based tests that use summary data and differences in test performance with genes of differing length, number of LD blocks, and number of causal variants per block. Analyses of the real ADHD GWAS data were used to evaluate differences among the gene-based tests in finding associated genes.

Introduction

Genome-wide association studies (GWAS) have become a useful tool in recent years for elucidating the etiology of a variety of non-Mendelian diseases. These studies have shown promising findings in the field of psychiatric genetics, including locating relevant genetic loci for schizophrenia (Ripke, 2013) and ADHD (Neale et al., 2010). The canonical GWAS methodology typically consists of estimating the association between a given single nucleotide polymorphism (SNP) and the phenotype of interest for each of the millions of SNPs that are genotyped in a given study. This results in a separate association test statistic for each of the millions of genotyped and imputed genetic markers.

Single-SNP methods pose a number of challenges of both a statistical and biological nature. In a statistical sense, the number of association tests performed results in the need for very stringent p-values (Corvin et al., 2009), typically on the order of p = 5 X 10-8 (Dudbridge & Gusnanto, 2008). Statistical power in this context if often poor due to small effect sizes, relatively modest sample sizes given the number of statistical tests performed, as well as the number of markers vastly exceeding the number of individuals genotyped. These challenges make alternative association testing methods, such as gene-based tests, attractive (Neale & Sham, 2004). Gene-based tests perform a synthesis of the effect-sizes and/ or p-values provided for all the SNP's mapped to a given gene and its flanking region, taking into account the correlation among the SNPs, referred to as their linkage disequilibrium (LD). This is an attractive approach, as the gene is the actual functional biological unit of inheritance. Indeed, the functions of genes and pathways have been more thoroughly investigated than that of SNP's (Peng et al., 2009). Gene-based tests also create opportunities for different patterns of association among SNP's in a gene to result in the same gene-based test statistic; this provides the potential

for genes in which each SNP contributes a small amount of risk to still produce a significant test-statistic and a more appreciable effect size than any of the constituent SNPs.

A number of competing gene-based tests have emerged which provide different analytic routes for calculating their gene-based test statistics. Naïve formulations include Fisher's procedure for combining p-values and simply choosing the most significant SNP within a gene, known as Sidak's combination test (Peng et al., 2009). Fisher's combination procedure is not well-suited to this context given its assumption of independent p-values. Due to linkage disequilibrium (LD), groups of nearby SNP's, such as within a gene, often appear in correlated blocks. This makes test statistics that assume independent p-values or effect sizes are suspect in this context, given that LD is often of a non-trivial magnitude. Sidak's combination test only uses a single SNP statistic as the overall gene-best statistic, which leads to the exclusion of other SNPs from affecting the results, thus also is not ideal for this context.

The GATES (Gene-Based Association Test Using Extended Simes Procedure) method (Li, 2011) provides an alternative analytic procedure for combining p-values. This test estimates the gene-based p-value as the minimum of the effective number of SNP p-values, where the effective number of p-values is estimated using the LD of the SNPs. This is found by performing an eigen decomposition of the LD correlation matrix and diminishing the effective number of independent SNP's according to eigenvalues that are greater than one, as these indicate the presence of a dependency structure among the SNP's.

Li et al. (2011) evaluated their algorithm through a series of simulation studies. The goal of such studies is to assess the type 1 error rate and statistical power across a range of realistic scenarios. The authors generated 30 SNP's under Hardy-Weinberg equilibrium (HWE) and each SNP was grouped within a block of correlated SNPs. HWE refers to the expected prevalence of

observed alleles at a bi-allelic marker, and violations may suggest genotyping errors or population stratification, among other explanations. Blocks were simulated using varying levels of within-block correlation, or LD. Blocks were assumed to be independent of one another. While the authors used three disease models (I.e., additive, multiplicative, and a null model where the SNPs have no effect on disease risk) for constructing a phenotype, only the additive model is relevant to this discussion. They also varied the effect-sizes of individual SNPs. Their simulation studies indicate a distinct advantage for the GATES algorithm over other evaluated gene-based testing methods in the presence of LD in terms of reduced type 1 error. In terms of statistical power GATES did not perform meaningfully better than the other algorithms tested, which included logistic regression, Fisher's combination procedure, Simes test, and VEGAS (Liu et al., 2010).

As an extension of the GATES algorithm, Li et al. (2012) proposed the hybrid set-based test (HYST) as a gene-based test. This algorithm uses GATES on each of *n* LD blocks within a gene to synthesize a single p-value for each block. A scaled chi-squared test is then used to combine the *n* block p-values into an overall test statistic for each gene, which is a modification of Fisher's combination procedure that takes the correlation among constituent SNPs (I.e., their LD) into account. Li et al. then simulated a variety of scenarios for protein-protein interaction based association analysis, finding that HYST generally outperformed alternative gene-based tests in terms of statistical power. Among the competing algorithms evaluated were GATES and the scaled chi-squared test, in addition to the Sequence Kernel Association Test (SKAT), which will be discussed below.

Canonical correlation analysis (CCA) was proposed by Ferreira et al. (2012) and provides an alternative analytic route to gene-based testing. CCA is a multivariate test of association that

maximizes the correlation among unique, orthogonal linear combinations of the two sets of variables on the "predictor" and the "outcome" sides of the equation. In the case of a single continuous phenotype and multiple SNP's, Ferreira et al. note that CCA is equivalent to linear multiple regression. SNP's in high multi-collinearity are removed through a pruning stage relying on variance inflation factor analysis. Wilk's lambda is then used to test the significance of all canonical correlations. It should be noted that a crucial difference between CCA and the above algorithms is that CCA does not work with summary data (I.e., SNP-based p-values or effect sizes) from a SNP-based GWAS, but rather uses the original genotype / dosage and phenotype data.

Ferreira et al. assessed the viability of CCA as a gene-based test through simulations. They compared the results of CCA to two permutation based approaches implemented in PLINK (Purcell et al., 2007), as well as the gene-based test GWiS, which will be described below. They generated data for 2000 individuals with a normally distributed quantitative trait. This differs from the above simulations which generated a categorical diagnostic phenotype. Ferreira et al. manipulated the proportion of phenotypic variance explained by each gene (heritability), the number of quantitative trait loci (QTL), gene length (ranging from 220 to 500 kb), QTL allele frequency, and allele frequency of surrounding SNP's. Naturally, heritability was also manipulated to assess type-1 error rate. The results suggested that CCA performs better in the presence of small genes (better meaning more statistical power and appropriate type-1 error rates) and when the number of QTL's is small. In contrast, CCA was less powerful in the context of larger genes. The results suggested that in most scenarios PLINK's all-SNP method provides similar or better performance.

The PLINK all-SNP method provides the canonical method for permutation based set-based statistics. The PLINK set-based statistic non-parametrically generates a null distribution of p-values by permuting the phenotype labels and keeping the genotypic information constant. For each of these iterations, the same test-statistic as before is calculated. Afterwards, the original test-statistic is compared to this null distribution to generate an empirical p-value. The SNPs used are selected to be in low LD with one another. This method seems to perform well according to the above mentioned simulation analyses, but is limited in practice due to the computational expense of permutation.

Another regression based approach is the SKAT (Wu et al., 2011) method for rare-variants. SKAT conducts linear regression and performs hypothesis testing on a group of markers using a variance-component score test. The null-hypothesis is generated by only using specified covariates (I.e., no SNPs) to predict the phenotype. This is then compared to a model that also includes genotypic information on the constituent SNPs. The variance-component score test also includes a matrix whose elements correspond to the genetic similarity between two subjects. The kernel in this case is linear, which means that the two sets of genotypic information are compared in a manner similar to covariance. It is not explicitly stated why other types of kernels (polynomial, Gaussian) were not explored, but is likely that the linear kernel simply provides much better computational performance. In this context a larger variance-component score test corresponds to a larger effect of the gene due to the contribution of genotypic information over merely including covariates. This statistic follows a chi-squared distribution, lending itself well to analytical calculation of p-values.

The Gene-Wide Significance test (GWiS) provides a Bayesian alternative (Huang et al., 2011) to calculation of gene-based p-values. GWiS uses Bayesian model selection to determine

the best subset of SNPs that maximizes the total probability of the model given the phenotype and the genotypic data. Bayesian model selection is a procedure for discriminating between competing models. In accordance with Bayesian methods, a prior is specified over each possible model which in this case is maximized when the number of SNPs in the model is equal to the effective number of independent tests, which is calculated using the LD of the gene. The likelihood term is simply the likelihood of a multiple linear regression model for the gene, with the SNPs as predictors and the phenotype as a dependent variable. The prior and likelihood term are multiplied to yield a value proportional to the likelihood of the model given the phenotype and genotypic information. GWiS is designed to choose a single model for each gene that maximizes the above model likelihood term. GWiS then uses a permutation procedure to convert the GWiS test-statistic to a p-value, and applies a p-value adjustment procedure to the results.

The Current Study

In the current study, I used simulations and analyses of real data to evaluate the properties of several of the aforementioned gene-based tests, as well as to attempt to answer some questions raised by the application of several gene-based tests to real ADHD GWAS data from four samples. The issues addressed by the simulations included an examination of their Type I Error rates and statistical power across several plausible scenarios, as well as the impact of differing LD among the SNPs on the gene-based tests that use summary data and differences in test performance with genes of differing length, number of haplotype blocks, and number of causal variants per block. Analyses of the real ADHD GWAS data were used to evaluate differences among the gene-based tests in finding associated genes.

Methods

Simulation

The simulated data were generated using a two-step process. First, genotype data are simulated. The foundational unit for the genotype simulations are LD blocks, which occur in real data when certain sets of SNPs in close genetic proximity to one another are transmitted together during meiosis, as their close genetic distance prevents them from being split up due to recombination. To simulate LD blocks, we first specified the theoretical correlation matrix Σ among SNPs in a block, and then simulate covariance matrices based on randomized draws from a Wishart distribution with Σ as a parameter. Markers are produced in a dosage format, where each column in the data matrix corresponds to a single marker. For *N* individuals and *P* markers per block, an *N x P* matrix is simulated with each column being a marker. The *K* LD blocks are then concatenated together to yield an *N x P*K* matrix, representing a gene. The similarity between adjacent blocks is measured through canonical correlation, which effectively seeks to maximize the correlation among linear combinations between the two LD blocks. Due to the inherent randomness in the generation of the data, the markers do not depart from HWE. The parameters currently specified are minor allele frequency, Σ, number of individuals, number of blocks, number of markers per block, and between block correlation.

The phenotype data are simulated by randomly choosing a set of causal LD blocks, and then subsequently choosing a specified number of causal markers within those LD blocks. For a given effect size, the explanatory power is distributed equally over the causal markers. This per-marker effect size is then converted to an expected slope based on the effect size, marker variance, and phenotype variance. The slope is then multiplied by a randomly sampled sign (-1 or 1) to produce both positive and negative marker effects. Each of these causal markers is then combined in an additive manner based on their slopes to produce the "true" phenotype. A degree of random noise is then added to the phenotype based on the total explanatory power specified

before (specified as an R-squared value). The distribution of this noise is Gaussian, which can later be converted into a binomial variable through the sigmoid function if the user desires a categorical phenotype (e.g., a diagnosis).

Currently a few gene-based tests are implemented in the simulation package: Sidak's, Fisher's, VEGAS, GATES, HYST, and CCA gene-based test. Naturally others are in the process of being implemented (GWiS, SKAT) as discussed before. It is important to distinguish among two general types of gene-based tests. Only CCA uses the raw genotypic data to associate the gene with the phenotype, whereas the other tests use summary statistics of and the LD among the constituent SNPs. This is a fundamental difference which has performance consequences that will be assessed through simulation. Once implementation was completed, a series of tests were conducted. The gene-based tests were first assessed for proper type-1 error by simulating a gene with no causal effect on the phenotype and measuring the proportion of simulations in which a given test produces a significant p-value. Type-1 error was checked under different conditions of LD. Next, power analyses were conducted by simulating genes with a non-zero causal effect on the simulated phenotype, and measuring the proportion of outcomes where the gene-based tests produces a significant p-value.

To distinguish between the two types of gene-based tests, another parameter was added to the simulation. When using a summary-statistic based gene-based test that relies on LD information, in practice it is most common to use a corresponding reference panel (I.e., the 1000 genomes samples; provide reference) for the measurement of LD. This has an important consequence in that this LD information may lead to spurious results if the 1000 genomes reference panel has very different LD from the population of interest. Gene-based tests which do not rely on summary statistics do not rely on an external 1000 genomes sample for LD

information. To incorporate this effect into the simulations, a multivariate Gaussian noise matrix was produced with a user specified amount of variance for each variable. The noise matrix is of the same dimensions as the LD matrix. In addition, each of the simulated random variables has a mean of zero with no covariance, leading to a diagonal covariance matrix. In an effort to mimic the form of a correlation matrix, which is a Hermitian matrix, the noise matrix was coerced into a symmetric matrix by equating the lower and upper triangular regions. This guarantees the absence of complex eigenvalues in the sum of the noise matrix and original LD matrix. After the symmetric noise matrix is produced it is summed with the original LD matrix. This procedure is only applied to those gene-based tests which rely on external LD information. The motivation behind this addition of noise is to replicate the lack of complete concordance between the 1000 genomes reference panel LD and the LD of the population of interest. Varying amounts of noise can be specified to create more or less noisy LD matrices. This feature will provide insights into the degree of robustness of the summary-statistic based gene-based tests to noisy inferences of population LD.

The simulation tools were written in R 3.1.0 (R core team, 2014) and compiled into an R package, gwassim. A small portion of the simulation is written in C++ for speed gains. The package was developed on a personal laptop with limited computing power. The simulations were run on the SURFsara Lisa computing cluster housed in the Netherlands. Each simulation test constitutes the combined results of 500 independent simulations. Naturally more simulations would lead to a more precise result, but there is a tension between computational tenability and precision that is sufficiently well struck by the number of simulations used herein.

Psychiatric Genomics Consortium (PGC) ADHD GWAS data

Samples and participants

ADHD GWAS data were used from four samples. The first sample was collected in Germany, where families were recruited at outpatient clinics in Wuerzburg, Homburg, and Trier. The families had a German and Caucasian ancestry. The children were at least six years of age and all met DSM-IV criteria for ADHD. This sample will be referred to in this paper as GERM. This sample had 494 cases and 1,297 controls.

The next sample was collected from behavioral clinics affiliated with the Children's Hospital of Philadelphia. This sample consisted of 3588 family trios (i.e., parents and an offspring). Families of European ancestry were included. The children were ADHD probands of at least age six. This will be referred to as the CHOP sample.

The next sample was collected from three locations: the Massachusetts General Hospital, UCLA, and Washington University. The sample consisted 712 family trios. Families were invited if at least one child displayed at least three inattentive systems during an initial interview. This will be referred to as the PUWM sample.

The final sample was collected from a variety of European countries. Family members were Caucasians and of European descent. Germany, Ireland, the Netherlands, Spain, Switzerland, the United Kingdom, and Israel were among the countries included. The probands had been clinically diagnosed as having ADHD. The sample consisted of 866 family trios. This will be referred to as the IMAGE1 sample

Genotyping and imputation

Cases were genotyped using an Affymetrix array at the State University of New York Upstate Medical University, while controls were genotyped using an Affymetrix array at the

Broad Institute National Center for Genotyping and analysis. The data were imputed using the BEAGLE software (Browning & Browning, 2009).

Quality control

In the original GWASs, Covariates controlling for sex and population stratification were included. Population stratification was addressed by including the first six components from principal component analysis (PCA) on the genotypic data. Markers with allele frequency less than .05 or genotyping score of less than .80 were excluded from analysis. This data filtering was performed on each sample separately.

Statistical analyses

For each single SNP, logistic regression was conducted to estimate a regression coefficient for the given SNP as well as its standard error. The coefficients were then exponentiated to put in odds-ratio form. P-values were generating using the Wald statistic, which provides an asymptotic estimate of the given marker P-value using a chi-squared distribution with a single degree of freedom. These analyses were conducted using PLINK.

Application of gene-based tests

After the above analysis was conducted gene-based tests were performed. Specifically, using the package KGG (Li, 2010), I applied the gene-based tests GATES and HYST to the existing PLINK outputs of SNP-based results combined with LD information from the 1000 genomes European reference panel (Genomes Project Consortium, 2010). For the CCAs, I selected the most significant genes from GATES and HYST and ran CCAs on them using Ferreira's test which I implemented in R. I then compared the various gene p-values using various statistical tests (e.g. paired sample t-tests) and statistical graphic methods (e.g. scatter

plots with fitted regression lines). Only the above three gene-based tests were run on the data due to considerations of computational tenability and available software.

Meta-analysis

For each gene there were 12 total p-values – four each (one for each of the four samples) for CCA, GATES, and HYST. For each gene-based test, each of the four sample p-values was combined using Fisher's combination procedure. Given that genes are assumed to be relatively independent it seems unlikely that Fisher's combination procedure suffers the same pitfalls as when it is used on SNP level p-values. The Benjamini-Hochberg (Benjamini & Hochberg, 1995) procedure was subsequently applied to the resulting p-values to correct for the number of tests.

Results

Type-1 simulation results

As described above, the first simulation conducted was a test of type-1 error, the results of which are displayed in figure 1.

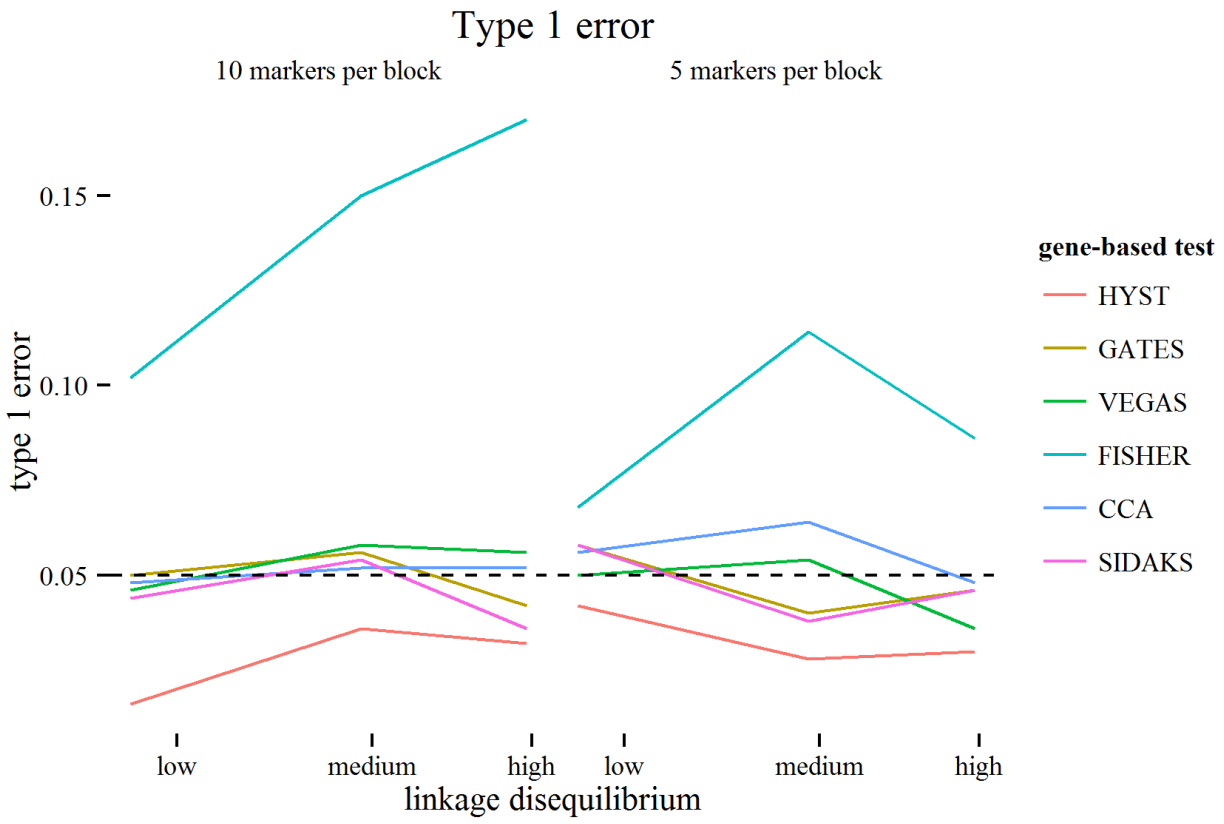

Figure 1: type 1 error rates

Tests for type-1 error were varied over two parameters: level of LD and number of markers per block. The nominal type-1 error rate was .05, which is displayed in figure 1 with a dashed line. The test with the most divergent performance from our expectation of .05 was Fisher's procedure, which showed elevated levels of type-1 error throughout. In particular, as expected the level of type-1 inflation of Fisher's appears to increase as the level of LD increases. The other tests display type-1 error rates very close to the expected .05 rate, with HYST appearing as slightly conservative.

PGC ADHD GWAS sample results

GERM

It is important to consider the inflation of p-values when working with high-dimensional genomic data. P-value inflation occurs when we observe systematically lower p-values across the genome than expected due to a spurious association. To evaluate this potential inflation it is useful to examine quantile-quantile plots. Were we to observe p-value inflation we would notice that the plotted values would systematically fall above the diagonal in Figure 2. Here we observe limited departure from the diagonal line. The gene-based p-values are from GATES.

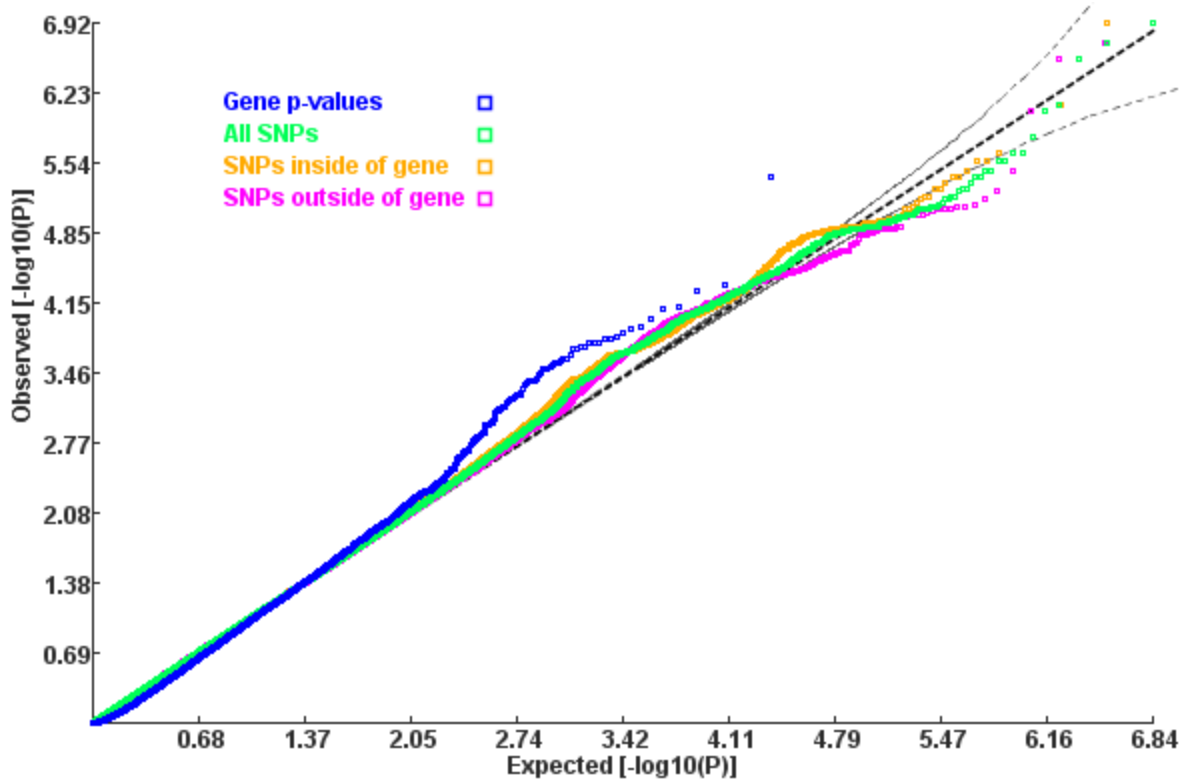

Figure 2: QQ-plot of Germany gene-based p-values

Table 1 indicates a list of the top 10 hits for the gene-based p-values according to GATES. As will be shown later, there is a very high degree of agreement between GATES and HYST, with CCA departing from the other two tests in its estimates. It is important to note that the p-values listed in table 1 are nominal – that is, they have not been corrected for multiple

testing. The p-values will be corrected for multiple testing after the multiple samples are combined through meta-analysis. The results indicate that the strongest signal occurs in chromosome four, closely followed by signals in chromosomes two and eleven, but that none reach genome-wide significance.

Gene	gatesp	chrom	startPosition	Length	hystp	ccap
COX7B2	3.93E-06	4	46736846	174407	3.32E-05	NA
BCL11A	4.63E-05	2	60678301	102333	0.000371	0.061662
SPI1	5.22E-05	11	47376408	23720	4.27E-05	0.004787
MIR1286	7.84E-05	22	20236656	79	7.84E-05	NA
MAP2K3	8.18E-05	17	21191347	27205	8.18E-05	NA
GPX6	0.000103	6	28471072	12499	0.000137	NA
MYBPC3	0.000122	11	47352956	21298	0.000502	0.04235
TRIM67	0.000127	1	231298673	58642	0.000219	NA
PLEKHM1	0.000137	17	43513265	54882	0.000124	NA
C1orf131	0.000149	1	231359508	17426	0.000144	NA

Table 1: Top 10 gene hits according to GATES

It is worthwhile to more explicitly compare the results of the three-gene based tests. A beneficial transformation of p-values is to apply $-log_{10}$ to the values. This aids the visual

interpretation of the p-values by transforming small p-values into large values, and by distorting the distance between small p-values.

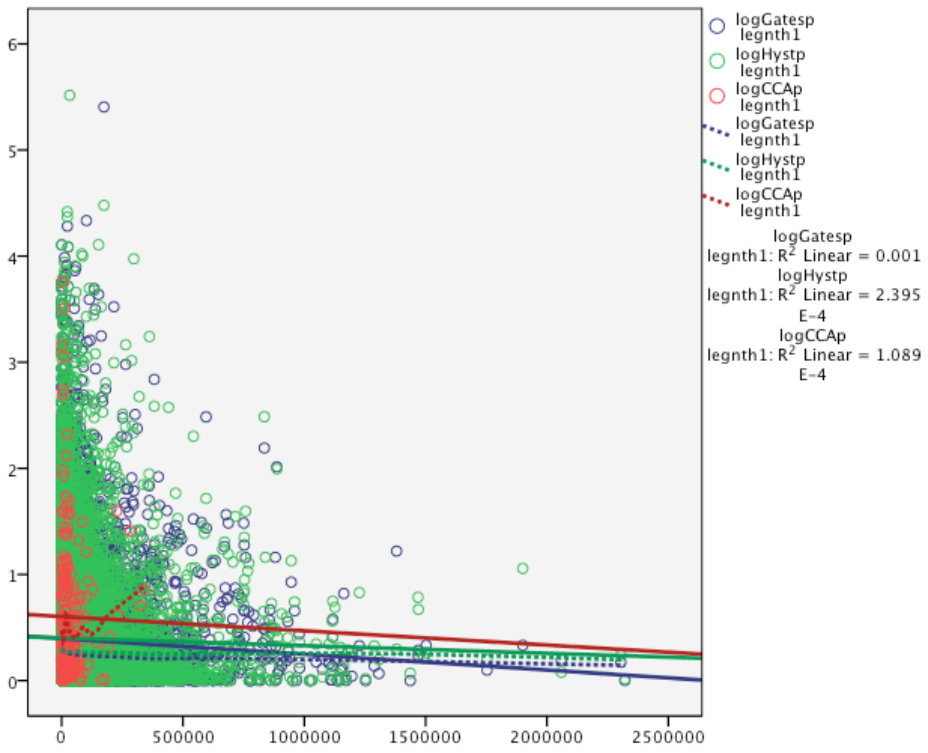

Figure 3: gene-based pvalues by gene length

As indicated in Figure 3, we observe a slight downward slope as the mean p-value decreases, suggesting a lack of power on large genes for these gene-based tests. We also observe that CCA results in more significant p-values on average.

In an effort to tease out the effect of gene-length, it can be useful to dichotomize the gene-lengths into small genes and large genes. While dichotomizing a continuous variable can reduce statistical power this choice is purely to aid exploratory analysis. After performing this split, it is informative to look at differences in mean –log(p-value) among the results.

Pair	Difference	t	Df	Sig. (2-tailed)
logGATESp – logHYSTp	-.00052	-.920	12005	.358
logGATESp - logCCAp	-.07705	-1.317	110	.190
logHYSTp - logCCAp	-.07437	-1.271	110	.206

Table 2: Differences in mean -log(p-values) for GERM sample for small genes

As indicated in Table 2, we did not observe any significant differences in mean difference of –log(p-values) when examining shorter genes in the GERM sample.

Pair	Difference	t	df	Sig. (2-tailed)
logGATESp – logHYSTp	**-.01683**	**-.10.600**	**12005**	**< .0001**
logGATESp - logCCAp	-.06245	.864	110	.389
logHYSTp – logCCAp	-.05219	.759	110	.449

Table 3: Differences in mean -log(p-values) for GERM sample for large genes

As indicated in Table 3, we observed a significant difference between the mean –log(p-values) of GATES and HYST in large genes, with HYST have a slightly higher mean. This may suggest an advantage for HYST when working with large genes. This suggestion is furthered by Figure 1 which indicated that HYST is actually slightly conservative. We explored possible explanations for this difference using simulations which will be discussed below.

CHOP

We first examine the QQplot of gene-based p-values for the sample.

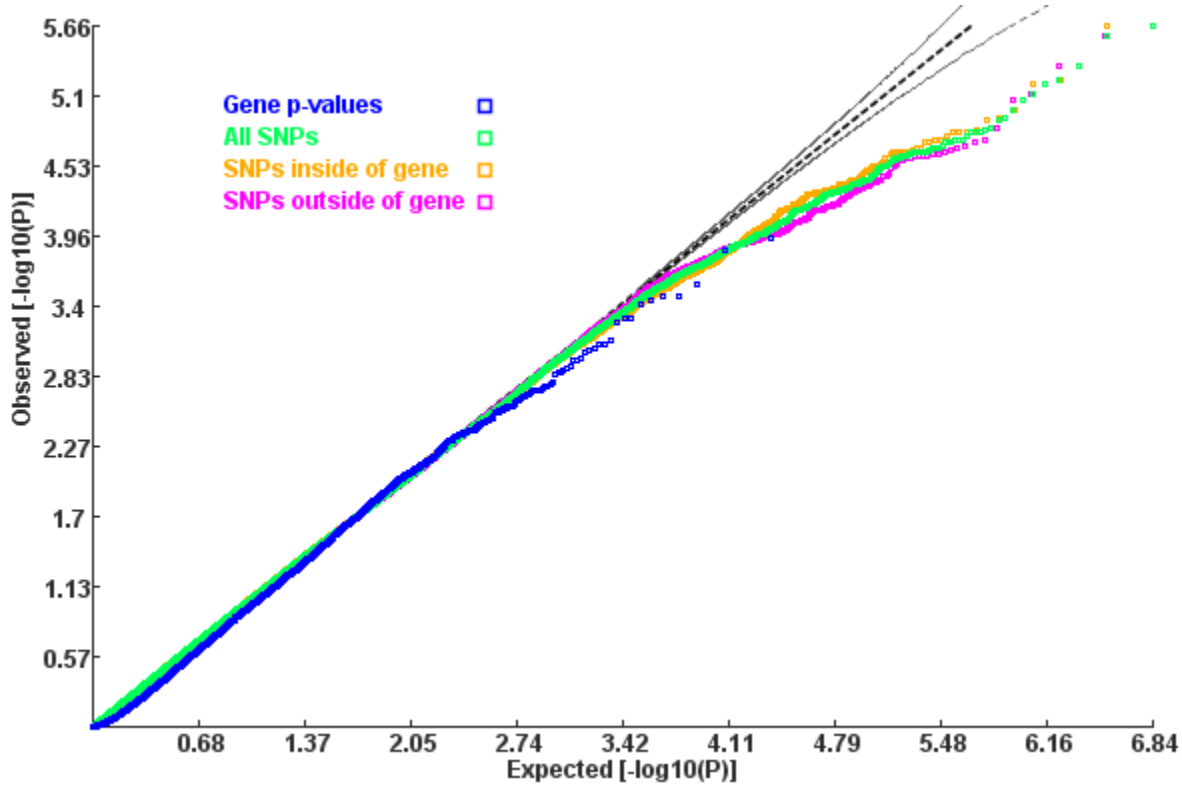

Figure 4: QQplot of GATES p-values for CHOP sample

As we observed with the GERM sample, visual examination of the QQplot does not suggest inflation in p-values.

gene	gatesp	chrom	startPosition	length	hystp	ccap
MIR3180-4	0.000112	16	15248706	154	0.000112	NA
SH3BP2	0.000139	4	2820540	22284	0.000107	NA

GSTM5	0.00027	1	1.1E+08	6028	0.00148	NA
MIR3666	0.000327	7	1.14E+08	112	0.000327	NA
ZNF781	0.000331	19	38158649	24568	0.000386	NA
AICDA	0.000361	12	8754761	10682	0.000361	NA
TES	0.000387	7	1.16E+08	35834	0.00281	NA
PHLPP1	0.000494	18	60382671	265006	0.000186	NA
SEC16B	5.00E-04	1	1.78E+08	40810	0.000809	NA
UBXN10-AS1	0.000532	1	20510735	2245	0.000532	NA

Table 4: Top hits by GATES for CHOP

As indicated in table four the top hits by Gates indicate the strongest signal arises at gene MIR3180-4 in chromosome 16. The next strongest signal is gene SH3BP2 in chromosome 4. Again, none of the genes reach genome-wide significance. As we observed before there is a strong congruence between HYST and GATES.

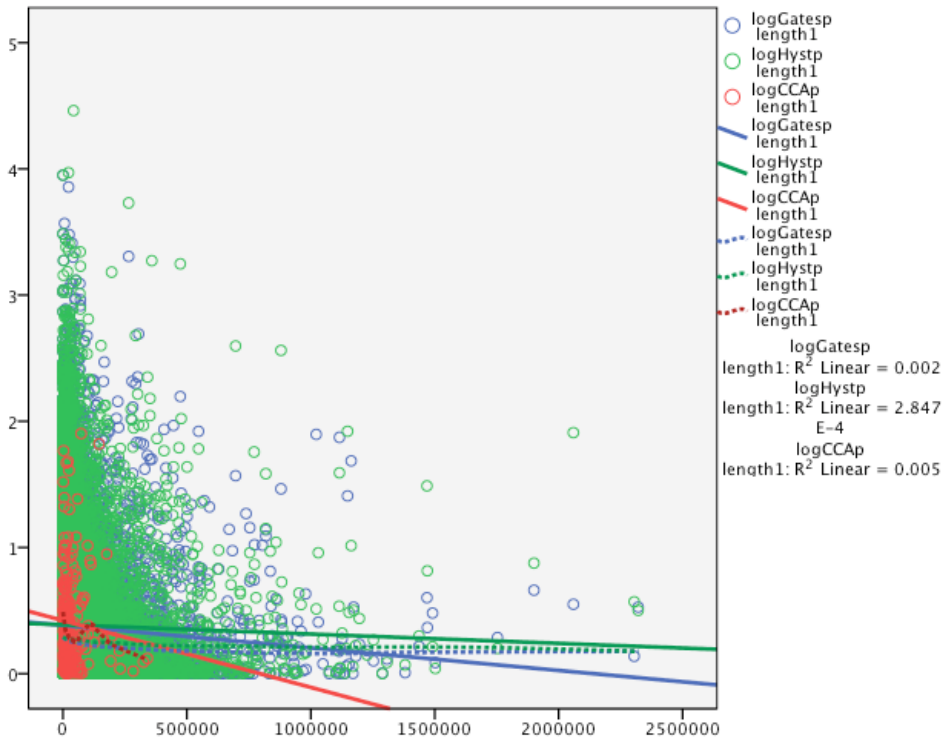

Figure 5: scatterplot of -log(p-values) by gene length

Figure 5 depicts a similar pattern to what was observed in the GERM sample – HYST appears to have smaller p-values on larger genes. In this sample we also observe CCA to produce larger p-values as gene length increases.

Pair	Difference	t	df	Sig. (2-tailed)
logGATESp – logHYSTp	-.00080	-1.303	12025	.193
logGATESp - logCCAp	-.05696	-1.621	113	.108
logHYSTp - logCCAp	-.05703	-1.609	113	.110

Table 5: Differences in mean -log(p-values) for GERM sample for small genes

As indicated in Table 6, we did not observe any significant differences in mean difference of – log(p-values) when examining shorter genes in the CHOP sample.

Pair	Difference	t	df	Sig. (2-tailed)
logGATESp – logHYSTp	**-.01744**	**-11.646**	**12025**	**< .0001**
logGATESp - logCCAp	-.01577	-.324	125	.746
logHYSTp - logCCAp	-.00974	-.199	125	.842

Table 6: Differences in mean -log(p-values) for GERM sample for large genes

Here we observe the same difference between HYST and GATES that we did in the GERM sample – that is, HYST appears to have more significant p-values when examining larger genes.

IMAGE1

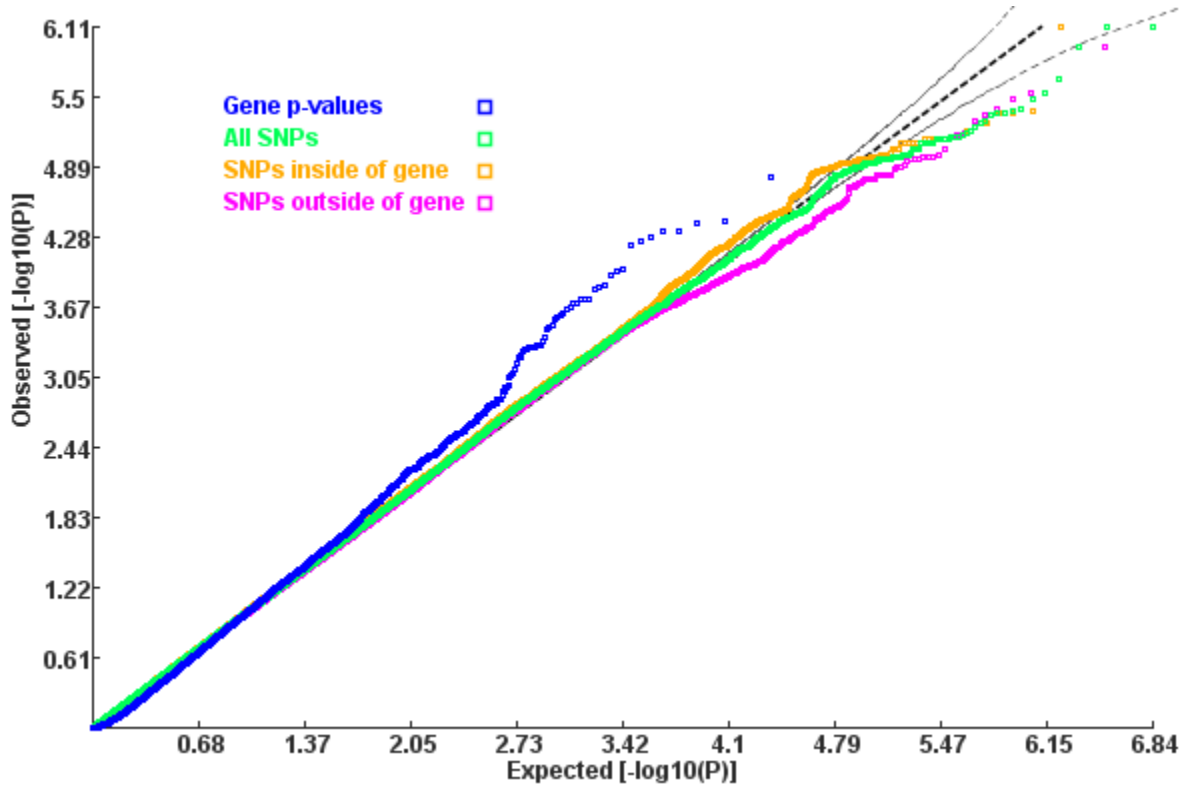

Figure 6: QQplot of GATES p-values for IMAGE1

Figure 6 suggests through visual inspection that systematic inflation of p-values is not occurring in IMAGE1.

gene	gatesp	chrom	startPosition	length	hystp	ccap
CKMT1A	1.62E-05	15	43985083	6338	1.62E-05	NA
MIR4257	3.88E-05	1	1.51E+08	87	3.88E-05	NA
HYPK	3.92E-05	15	44092618	2152	3.92E-05	NA
MFAP1	4.66E-05	15	44096732	20220	4.66E-05	0.003393

CATSPER2P1	4.70E-05	15	44028145	10352	4.70E-05	0.003247
PDIA3	5.26E-05	15	44038589	26216	5.26E-05	0.01136
SERF2-C15ORF63	5.78E-05	15	44084173	10597	5.78E-05	NA
ADAMTSL4	6.22E-05	1	1.51E+08	9381	6.22E-05	0.012174
DPH3P1	9.89E-05	20	61475917	1627	9.89E-05	NA
ELL3	0.000106	15	44064797	4706	0.000106	NA

Table 7: top hits according to GATES

We observe that gene CKMT1A has the strongest signal in the IMAGE1 sample, located in chromosome 15. We also observe genes HYPK, MFAP1, CATSPER2P1, PDIA3-SERF2, and C15ORF63 to also display strong signals in close physical proximity to CKMT1A, each being less than 1.2 mega base-pairs away. We also observe nominal significance according to CCA for each of the genes. Nonetheless, as with the previous samples, none of the genes reach genome-wide significance.

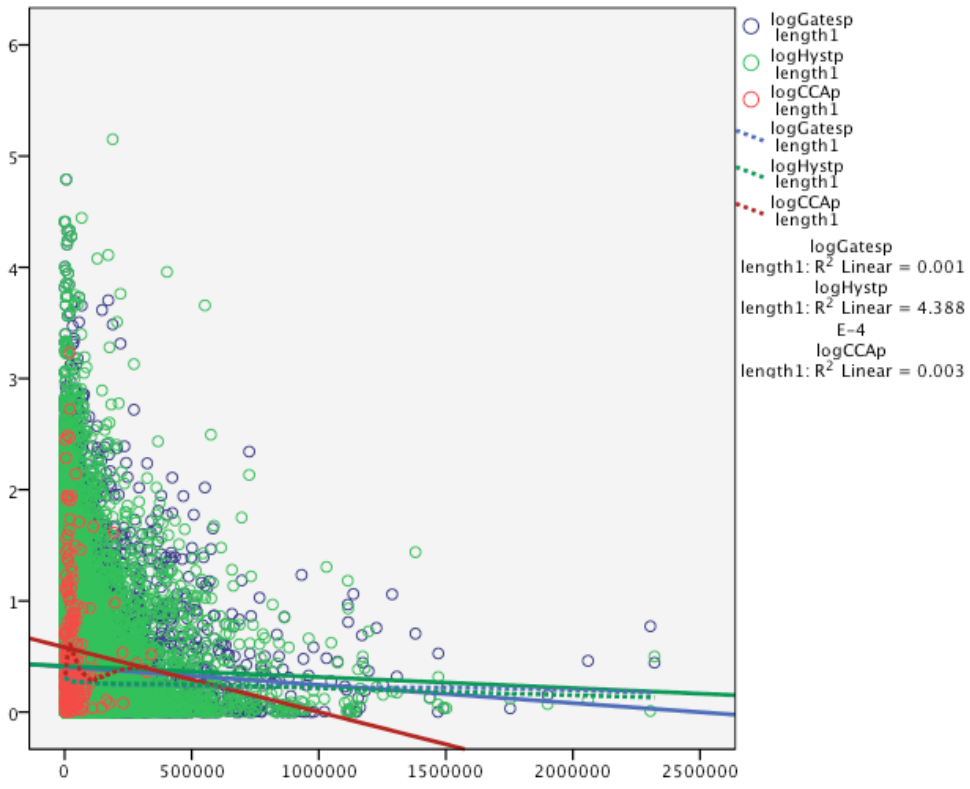

Figure 7: scatterplot of GATES -log(p-values) and gene length

Here we observe the same pattern as before with regards to HYST and GATES – that is, it appears HYST yields lower p-values than GATES on larger genes, whereas associations found with CCA decrease in significance with increasing gene length.

Pair	Difference	t	df	Sig. (2-tailed)
logGATESp – logHYSTp	**-.00144**	**-2.737**	**11918**	**.006**
logGATESp - logCCAp	-.06542	1.109	111	.270
logHYSTp - logCCAp	-.05746	.985	111	.327

Table 8: Differences in mean -log(p-values) for IMAGE1 sample for small genes

As indicated in Table 8, we observe that even in small genes HYST has smaller p-values than GATES.

Pair	Difference	t	df	Sig. (2-tailed)
logGATESp – logHYSTp	**-.01846**	**-11.707**	**11921**	**< .0001**
logGATESp - logCCAp	.06240	.909	123	.365
logHYSTp - logCCAp	.07539	1.050	123	.296

Table 9: Differences in mean -log(p-values) for IMAGE1 sample for large genes

Table 9 depicts the same result that we observed before – that is, on larger genes HYST yields smaller p-values than does GATES.

PUWM

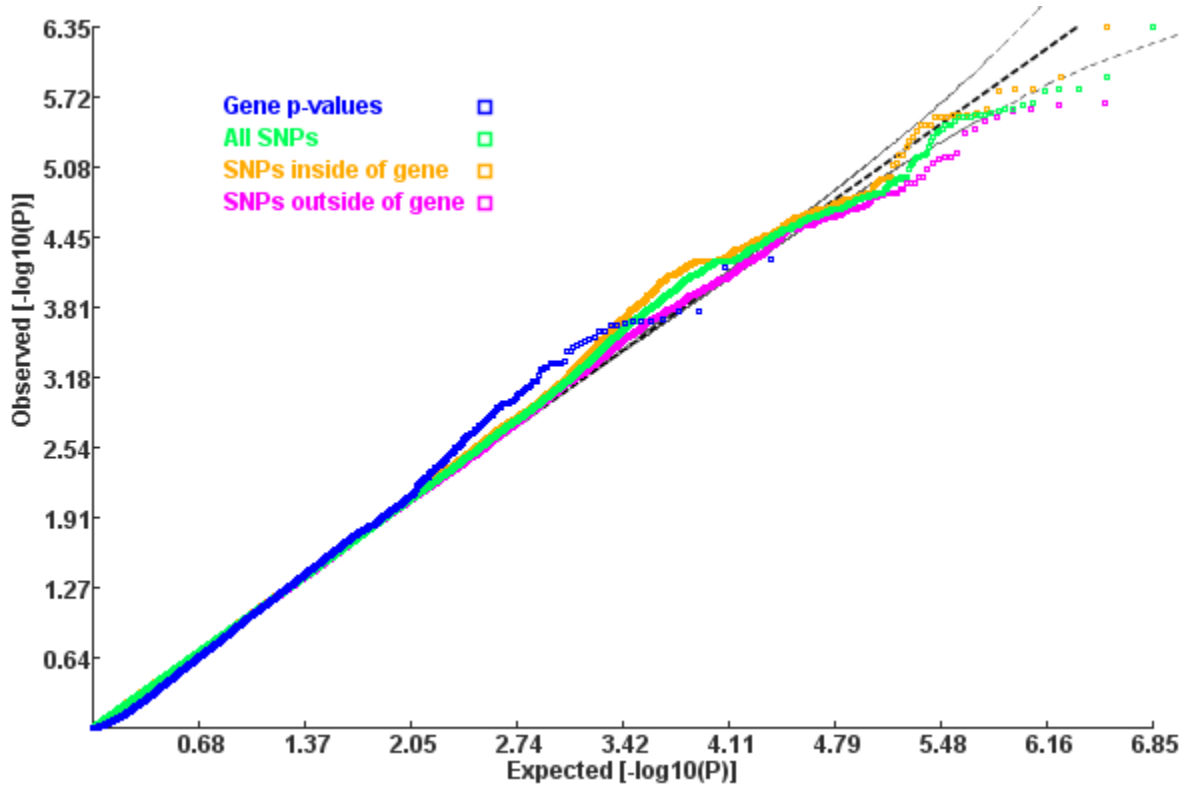

Figure 8: QQplot of GATES p-values

gene	gatesp	chrom	startPosition	length	hystp	ccap
ATP7B	5.74E-05	13	52506805	78826	0.000284	NA
LINC00865	6.73E-05	10	91589249	11370	0.000247	NA
HNRNPA1L2	0.000169	13	53191604	26316	0.000224	NA
SUGT1	0.000169	13	53226830	35604	0.000202	NA
UGT1A8	0.000198	2	2.35E+08	155656	0.000356	NA
BMPR1B	0.000203	4	95917382	162220	6.37E-05	NA
UGT1A9	0.000205	2	2.35E+08	101409	0.000914	NA
THSD1	0.000209	13	52951302	29328	0.000233	NA
VPS36	0.000212	13	52986736	37644	0.000218	NA
UGT1A10	0.000224	2	2.35E+08	136830	0.00178	NA

Table 10: Top hits by GATES in PUWM sample

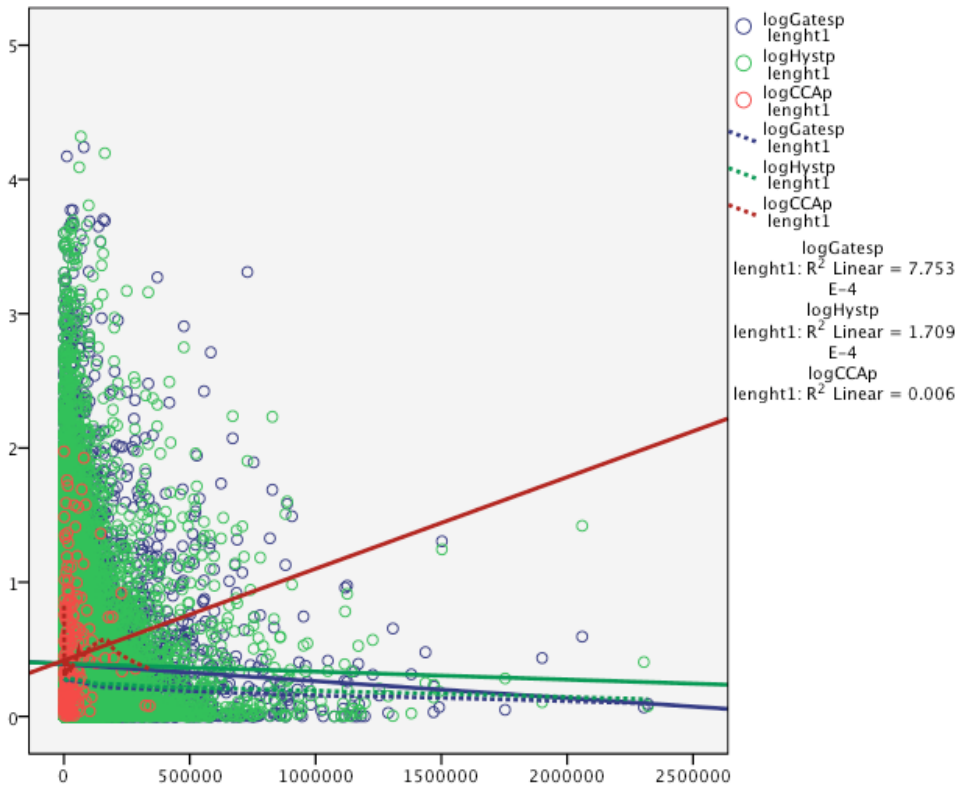

Figure 9: scatterplot of GATES -log(p-values) and gene length

Here and in the two tables below, we observe a similar pattern as before with HYST and GATES. We also observe a strong upward slope on linear CCA fit, but this is likely driven by outliers – the linear trend is for visual purposes and not for rigorous analysis.

Pair	Difference	t	df	Sig. (2-tailed)
logGATESp – logHYSTp	-.00070	-1.324	12087	.186
logGATESp - logCCAp	.02471	.463	114	.645
logHYSTp - logCCAp	.01729	.330	114	.742

Table 11: Differences in mean -log(p-values) for PUWM sample for small genes

Pair	Difference	t	df	Sig. (2-tailed)
logGATESp – logHYSTp	**-.01558**	**-9.886**	**12091**	**< .0001**
logGATESp - logCCAp	-.03277	-.591	125	.365
logHYSTp - logCCAp	-.02541	-.461	125	.296

Table 12: Differences in mean -log(p-values) for PUWM sample for large genes

Meta-analysis

After applying the meta-analysis procedure, a single gene, NOM1 on chromosome 7, resulted in nominal significance according to HSYT. GATES and CCA did not result in any significant genes. The top 10 hits according to HYST are displayed in Table 13.

gene	chrom	starPosition	length	ccap_pval	gatesp_pval	hystp_pval
NOM1	7	156742416	23461	NA	0.214309547	0.04956707
ABCA4	1	94458393	128313	NA	0.99999999	0.821187159
ADORA2A	22	24828086	10243	NA	0.875552356	0.821187159
ADORA2A-AS1	22	24834244	56540	NA	0.875552356	0.821187159
AGMO	7	15239942	361699	NA	0.99999999	0.821187159
BMPR1B	4	95917382	162220	NA	0.99999999	0.821187159

C1orf131	1	231359508	17426	NA	0.875552356	0.821187159
C9orf9	9	135754289	11130	NA	0.992133329	0.821187159
CAMTA1	1	6845383	86725	NA	0.99999999	0.821187159
CATIP	2	219221578	11240	NA	0.875552356	0.821187159

Table 13: top hits on meta-analyzed results according to HSYT

Power simulations

The above results on PGC data were used to inform hypotheses concerning these gene-based tests that could be examined in a controlled setting through simulation. Naturally this poses the risk of distancing ourselves from understanding the risks of the gene-based estimators in practice, but given the construction of the simulations it seems likely that they will contribute to our knowledge of which gene-based test to use in a given scenario. The use of real data results to generate hypothesis also has the desirable effect of creating dialogue between the simulations and real results, ensuring that relevant real world hypotheses are being tested in the simulations.

It is informative to begin with a general power analysis. Given the above focus on gene length I first performed a simulation on large genes. The genes had 10 markers per block, and 20 blocks in total. The total effect size was fixed at 1% of variance, and there were 4 total causal markers in separate blocks. The estimation of LD matrix for GATES and HYST had no added noise. Under these conditions we observe the results in Figure 10.

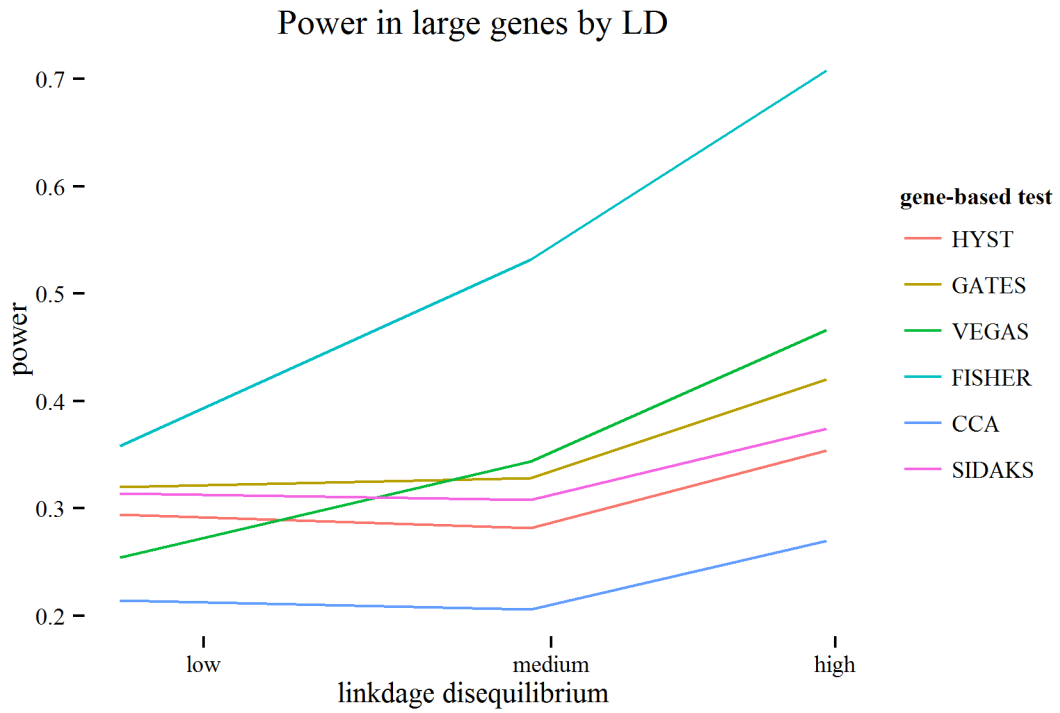

Figure 10: power in large genes by LD

The results indicate that the gene-based estimator with the most power in large genes is the Fisher combination procedure. However, as demonstrated before, this test also has incorrect type-1 error by a large degree, suggesting systematically low p-values. GATES, HYST, VEGAS, and Sidak's all perform similarly, with CCA appearing to have the lowest power in these conditions. It is also clear that the HYST performance in this simulation does not corroborate the PGC sample ADHD GWAS results. It is also apparent that an increase in LD within the blocks is beneficial for the gene-based tests. It is also clear the statistical power is low for all of the gene-based tests.

In effort to further examine the differences between GATES and HYST we produced another simulation that varies only the dispersal of causal effects within a gene. The simulated genes had five blocks, each consisting of five markers with a medium level of internal LD. The

effect size was held constant at 1% of total variance. In one scenario, there were four total causal SNPs, each within separate blocks. In another scenario, there was only a single causal SNP.

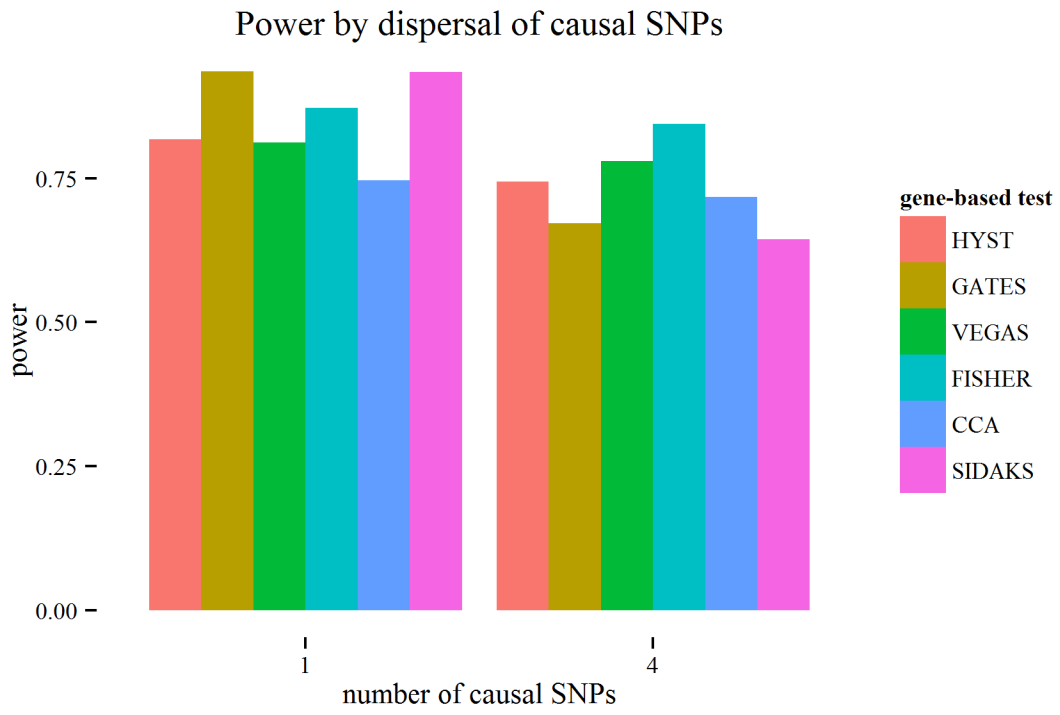

Figure 11: power by dispersal of causal SNPs

The results suggest that in general, gene-based tests will perform worse when the explanatory variance is dispersed evenly across disparate parts of a gene. We also observe that with a single causal SNP, GATES performs better than HYST, but when the causal etiology is dispersed across four SNPs, HYST performs better. Explanations for this switch in performance and its significance will be provided below.

Let us repeat the above experiment with a slight variation on size of gene. The same simulation was run except for an increase of five markers per block to 10 and an increase of total blocks to 10. That is, all other conditions were held constant and a gene four-times larger was used.

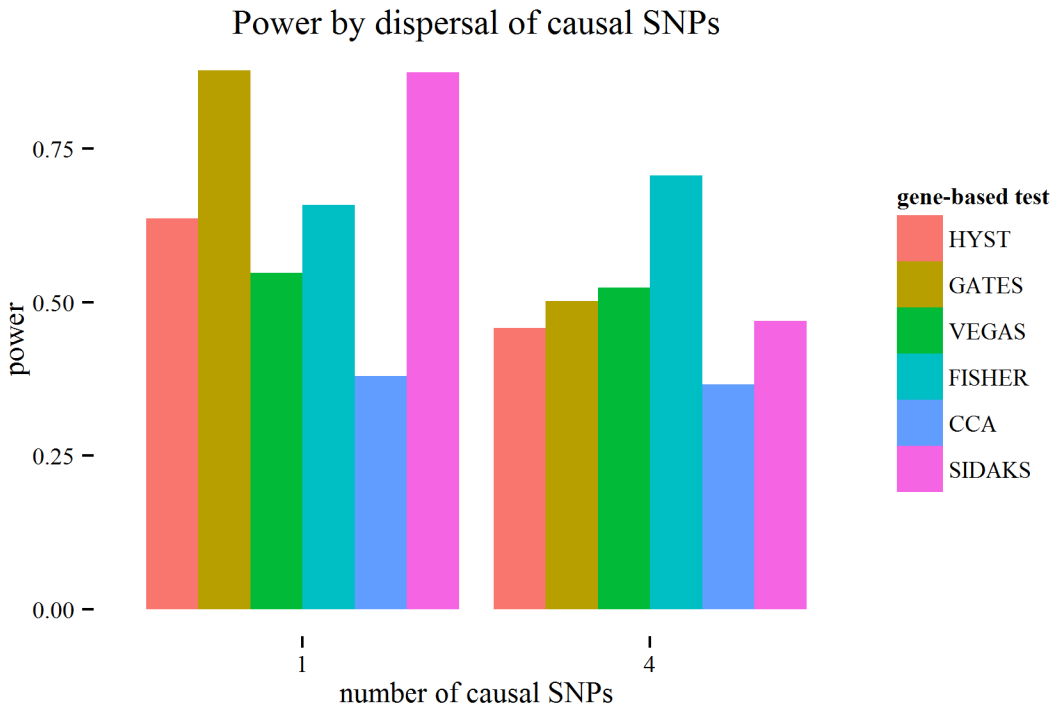

Figure 12: power by dispersal of causal SNPs in large genes

As depicted in Figure 12 we again observe the challenge posed to gene-based tests by genes with a wide dispersal of causal etiology. We also note that there is a large advantage of GATES and Sidak's over the other tests in the scenario of a single causal SNP. GATES also outperforms HYST by a large margin of over .20. In the case of four causal SNPs, the performance of HYST approaches GATES, but does not surpass it as we observed in the previous experiment.

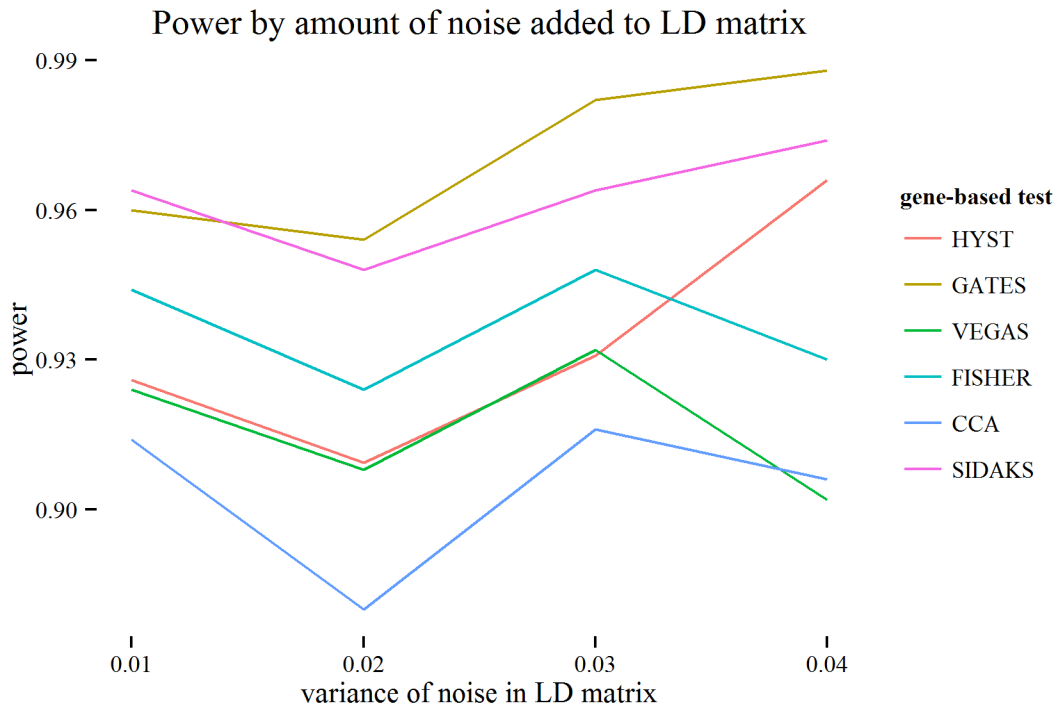

Figure 13: power by addition of noise

The simulation indicates the degree to which each gene-based test is robust to the addition of noise. This addition was only applied to GATES and HYST since these are the only gene-based tests included that both rely on LD information and a reference sample to calculate the LD. Due to this restriction we may conclude that the variation of the four other gene-based tests is completely due to noise in the simulations. It is apparent from Figure 13 that HYST and GATES do not display vulnerability to the addition of noise to the LD matrix. The variation in the power of HYST and GATES does not appear to exceed that of the other tests whose variation is completely due to randomness.

Discussion

Our first result in Figure 1 reaffirms that Fisher's combination procedure is inappropriate, even in cases of minimal LD, as it produces a much higher proportion of nominally significant p-values than the other tests. This is an expected result and is due to the assumption of Fisher's

combination procedure that each of the p-values be independent. This vulnerability appears to be exacerbated in particularly large genes. The other gene-based tests appear to exhibit correct type-1 error rates, with the possible exception of HYST. HYST appears to be slightly conservative, meaning that it produces slightly too few significant results. The inflation of Fisher's is in congruence with previous literature – Li et al. observed in 2012 a similar result. However, Li et al. do not observe the deflation of p-values from HYST. Given the small magnitude of our result, it seems likely that if HYST is producing a lower than expected type-1 error, then it is by an amount too insignificant to discourage practical use.

In all four of the above samples we observed that HYST has a higher –log(p-value) than GATES ($p < .0001$) in larger genes. This result inspired a particular simulation test – that is, a test of how HYST and GATES perform when the dispersion of causal etiology is manipulated. This test was born out of a hypothesis that larger genes in these samples may harbor a particularly wide dispersal of causal SNPs when compared to smaller genes. As indicated in Figure 9, HYST outperforms GATES by a small degree when examining genes with 16% of the SNPs being causal and their influence is very dispersed across separate blocks. We also notice that when we use a large gene but do not increase the number of causal markers that HYST no longer outperforms GATES, as indicated in Figure 10. These simulation results suggest a number of conclusions about when it is appropriate to use HYST instead of GATES.

Firstly, there needs to be certain degree of dispersion of the causal influence for HYST to outperform GATES. In any gene with a single or a few causal block it seems likely that GATES will perform much better. Secondly, if a low proportion of blocks contain signal, then it seems likely that GATES will also outperform HYST. These results suggest that HYST is best used on genes that have a wide dispersal of causal influence within a sufficiently high proportion of

blocks. These results are supported to a degree by previous results (Li et al, 2012), who found HYST to generally outperform GATES in simulations, but to provide similar results in practice. It is also worthwhile to note that in most simulated scenarios with one or a few causal SNPs that Sidak's approach of simply selecting the best test-statistic performed comparably to GATES, and in most scenarios provided similar results. This combination of strong relative performance and simplicity in calculation make Sidak's a surprisingly attractive gene-based test.

As indicated in Figure 8, we also observe noticeably poor performance by the CCA test. This result is notable in that CCA is the only one of the above studied gene-based tests to directly take into account genotypic and phenotypic data without relying on summary statistics. The poor performance of CCA on large genes is in congruence with Tang and Ferreira results in 2012, who observe CCA to perform much better on smaller genes than large genes. A possible statistical explanation is grounded in the limitations of linear regression. As stated before, in the case of a single phenotype, CCA is identical to multiple linear regression. To avoid the instability of coefficients in the presence of high multi-collinearity, the CCA procedure relies on a SNP pruning procedure. However, it may be that this screening procedure is eliminating causal SNPs given that it does not take into account the test-statistics of the given marker. Use of a more lenient SNP pruning procedure might aid the performance of the CCA method.

There are computational considerations as well that limit the usability of CCA in practice. The test is implemented in R which is not suitable for speed intensive computations, which made it challenging to apply the test to more than a few genomic regions. In the simulations a small section of the test had to be rewritten in C++ to gain acceptably performant behavior. These computational barriers make it difficult to recommend the use of CCA in practice on large genomic datasets.

A theoretical advantage of CCA over gene-based tests that rely on summary data is that CCA is not vulnerable to imprecise estimation of population LD. As discussed above, GATES and HYST are typically applied with a corresponding 1000 Genomes reference panel sample used to estimate LD. This denigrates the performance of GATES and HYST relative to CCA if the 1000 Genomes reference sample is a poor surrogate for the population of interest. As indicated in Figure 11, it appears that even significant random perturbations of the true LD matrix in simulations does not meaningfully reduce the power of GATES and HYST relative to CCA. This suggests that GATES and HYST are robust to imprecise inferences of the population LD matrix. What follows are observations and explanations for this unexpected result.

Throughout the simulations, GATES often produced very similar results to Sidak's. This is expected if we consider the formulation of GATES, given that GATES effectively chooses the minimum marker p-value, only using the LD information to calculate an effective number of tests. In the case of no LD among SNPs it is trivial to show that GATES and Sidak's will produce the same result. Given that our noise matrix was a multivariate Gaussian with a mean of zero, we were not greatly disturbing this calculation of the total number of effective tests since local LD information has minimal effect on GATES. Given that HYST uses GATES as an essential step we can extend this explanation to HYST as well. This result is promising for GATES and HYST and bodes well for their practical use. It should be noted that this conclusion only holds in the case of unbiased addition of random noise.

The individual samples did not produce any genes with p-values at or below genome-wide significance for any of the gene-based tests. This result is consistent with our knowledge of ADHD as a complex phenotype, given that such traits are often comprised as the constellation of several small genetic effects, which can make individual signals hard to find. This is in

accordance with the existing literature on the PGC ADHD GWAS datasets, as a previous a case control analysis of these datasets did not find any SNPs at or below genome wide significance (Neale et al., 2010).

After meta-analysis, a single gene – NOM1 -- remained genome-wide significant after p-value correction at a nominally significant level ($p < .05$). Another gene in chromosome 7 – AGMO – within 5 Mbp suggests the possibility of this as a true signal. Given that the other gene-based test, GATES, applied to this gene was not significant, it is difficult to consider this as a robust finding. Rather, it seems most appropriate to highlight NOM1 as a genomic region of interest for further inquiry for ADHD. We will continue to evaluate this in the remaining 5 available PGC ADHD GWAS samples.

Limitations

It is important to consider the limitations of this study with regards to our ability to infer relevant genomic loci for ADHD. As suggested by Neale et al. in 2010, genomic loci may be implicated in epistasis, epigenetic effects, gene-environment correlation, none of which are accounted for in this study. Additionally, GWAS is principally focused on risk imposed by common variants. This may leave us blind to the true causal markers of ADHD if they are due to rare variants which we currently exclude from our single-SNP analysis. We have also imposed an additive model on each marker, rather than exploring alternative structures. Each of these factors may contribute variance to ADHD that we currently will be unable to powerfully detect.

Our study also has many limitations on our ability to simulate realistic genes. We simulated genes with minimal between block correlation, constant block sizes, and can only simulate one gene at a time. We also did not consider phenotypes that do not have an underlying Gaussian distribution. While these may be robust assumptions to a certain degree, they do limit

our ability to simulate genes that one would naturally observe in GWAS data. Each simulation was only ran 500 times. Although this amount is likely sufficient for our purposes, we still did observe some noise in repetitions of the same simulation.

Areas of future work

As mentioned previously, not every gene-based test of interest was implemented in the simulations or applied to the PGC data. This includes SKAT, GWiS, and the PLINK set-based tests. Given the prevalence of these methods in the literature it would be of great interest to examine their results. It would also be of interest to more formally compare the results of our power analyses to that of previously published simulation studies by Ferreira, Li, and others. This formal comparison would allow for a deeper understanding of how the design choices involved in the simulations affect the results.

Another potential area of inquiry is to compare our results from the gene-based tests with a purely SNP based analysis in each gene. It would be of particular interest to apply the q-value method (Storey & Tibshirani, 2003) to the SNP level results within each gene. The q-values provide a very attractive interpretation and may provide insights that we cannot gain from examining p-values alone. It would also be of great interest to expand our results from the four samples above to other five PGC datasets. The addition of these samples into our meta-analysis will bolster our power in finding significant genes.

Conclusion

This study complements existing literature on gene-based tests for ADHD in a few ways. To our knowledge, this is the first study to run simulation analyses of existing gene-based tests while not simultaneously proposing a new gene-based test. The lack of investment in a specific gene-based test's performance lends itself favorably to objective interpretation of the simulation

results. This study also made use of a dialogue between real PGC ADHD GWAS data and simulations to explore hypotheses in a rich field of inquiry about gene-based tests on ADHD GWAS data. The study also highlights NOM1 as a genomic location of potential relevance to our understanding of the etiology of ADHD. This line of inquiry is not complete by any means and future work into the PGC ADHD GWAS datasets and simulations are planned.

Works Cited

Corvin, A., Craddock, N., & Sullivan, P. F. (2010). Genome-wide association studies: a primer. *Psychological medicine*, *40*(07), 1063-1077.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 289-300.

Browning, B. L., & Browning, S. R. (2009). A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *The American Journal of Human Genetics*, *84*(2), 210-223.

Dudbridge, F., & Gusnanto, A. (2008). Estimation of significance thresholds for genomewide association scans. *Genetic epidemiology*, *32*(3), 227-234.

Huang, H., Chanda, P., Alonso, A., Bader, J. S., & Arking, D. E. (2011). Gene-based tests of association. *PLoS genetics*, *7*(7), e1002177.

Ferreira, M. A., & Purcell, S. M. (2009). A multivariate test of association.*Bioinformatics*, *25*(1), 132-133.

Li, M. X., Gui, H. S., Kwan, J. S., & Sham, P. C. (2011). GATES: a rapid and powerful gene-based association test using extended Simes procedure. *The American Journal of Human Genetics*, *88*(3), 283-293.

Li, M. X., Sham, P. C., Cherny, S. S., & Song, Y. Q. (2010). A knowledge-based weighting framework to boost the power of genome-wide association studies. *PloS one*, *5*(12), e14480.

Liu, J. Z., Mcrae, A. F., Nyholt, D. R., Medland, S. E., Wray, N. R., Brown, K. M., ... & Macgregor, S. (2010). A versatile gene-based test for genome-wide association studies. *The American Journal of Human Genetics*, *87*(1), 139-145.

Li, M. X., Kwan, J. S., & Sham, P. C. (2012). HYST: a hybrid set-based test for genome-wide association studies, with application to protein-protein interaction-based association analysis. *The American Journal of Human Genetics*, *91*(3), 478-488.

Neale, B. M., Medland, S. E., Ripke, S., Asherson, P., Franke, B., Lesch, K. P., ... & McGough, J. (2010). Meta-analysis of genome-wide association studies of attention-deficit/hyperactivity disorder. *Journal of the American Academy of Child & Adolescent Psychiatry*, *49*(9), 884-897.

Neale, B. M., & Sham, P. C. (2004). The future of association studies: gene-based analysis and replication. *The American Journal of Human Genetics*,*75*(3), 353-362.

Peng, G., Luo, L., Siu, H., Zhu, Y., Hu, P., Hong, S., ... & Xiong, M. (2009). Gene and pathway-based second-wave analysis of genome-wide association studies. *European Journal of Human Genetics*, *18*(1), 111-117.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., ... & Sham, P. C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics*,*81*(3), 559-575.

R Core Team (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing,Vienna, Austria. URL http://www.R-project.org/.

Ripke, S., O'Dushlaine, C., Chambert, K., Moran, J. L., Kähler, A. K., Akterin, S., ... & Steinberg, S. (2013). Genome-wide association analysis identifies 13 new risk loci for schizophrenia. *Nature genetics*, *45*(10), 1150-1159.

Storey, J. D., & Tibshirani, R. (2003). Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences*, *100*(16), 9440-9445.

Wu, M. C., Lee, S., Cai, T., Li, Y., Boehnke, M., & Lin, X. (2011). Rare-variant association testing for sequencing data with the sequence kernel association test. *The American Journal of Human Genetics*, *89*(1), 82-93.

1000 Genomes Project Consortium. (2010). A map of human genome variation from population-scale sequencing. *Nature*, *467*(7319), 1061-1073.