FLAGS: A flexible and adaptive association test for gene sets using summary statistics

Jianfei Huang¹, Kai Wang², Peng Wei³, Xiangtao Liu¹, Xiaoming Liu⁴, Kai Tan⁵, Eric Boerwinkle⁴ *,
James B. Potash¹, Shizhong Han¹

1) Department of Psychiatry, University of Iowa, Iowa City, Iowa, USA, 52242.
2) Department of Biostatistics, University of Iowa, Iowa City, Iowa, USA, 52242
3) Department of Biostatistics, University of Texas School of Public Health, Houston, TX, USA, 77225
4) Human Genetics Center, University of Texas Health Science Center at Houston, Houston, Texas, USA, 77030
5) Department of Internal Medicine, University of Iowa, Iowa City, Iowa, USA, 52242

¹ Interdisciplinary Graduate Program in Genetics, University of Iowa, Iowa City, Iowa, USA, 52242
² Human Genome Sequencing Center, Baylor College of Medicine, Houston, Texas, USA, 77030

Copyright 2016.
Running title: Set-based association test using summary statistics

Key words: Complex disease; gene set; association; GWAS; summary statistics

Corresponding Author:

Shizhong Han, PhD,
Department of Psychiatry, University of Iowa Carver College of Medicine,
200 Hawkins Drive, Iowa City, IA, 52242, USA
Tel: +319 3538773, Fax: +319 353 8656
Email: shizhong-han@uiowa.edu
Abstract

Genome-wide association studies (GWAS) have been widely used for identifying common variants associated with complex diseases. Despite remarkable success in uncovering many risk variants and providing novel insights into disease biology, genetic variants identified to date fail to explain the vast majority of the heritability for most complex diseases. One explanation is that there are still a large number of common variants that remain to be discovered, but their effect sizes are generally too small to be detected individually. Accordingly, gene set analysis of GWAS, which examines a group of functionally related genes, has been proposed as a complementary approach to single marker analysis. Here, we propose a flexible and adaptive test for gene sets (FLAGS) using summary statistics. Extensive simulations showed that this method has an appropriate type I error rate and outperforms existing methods with increased power. As a proof of principle, through real data analyses of Crohn’s disease GWAS data and bipolar disorder GWAS meta-analysis results, we demonstrated the superior performance of FLAGS over several state-of-the-art association tests for gene sets. Our method allows for the more powerful application of gene set analysis to complex diseases, which will have broad use given that GWAS summary results are increasingly publicly available.
Introduction

Genome-wide association studies (GWAS), which examine millions of single nucleotide polymorphisms (SNPs) across the genome, have been widely used for identifying common variants associated with complex diseases (MCCARTHY et al. 2008). These studies have uncovered numerous risk variants and provided novel insights into disease biology. Despite these successes, genetic variants identified to date explain only a small fraction of the heritability for most complex diseases, which raises the question of “missing heritability” (MANOLIO et al. 2009). One explanation is that a large number of common variants remain to be discovered, but their effect sizes are generally too small to be detected individually. In the search for additional common variants in complex diseases, more sophisticated analyses of GWAS, rather than SNP-level analysis, can enhance the identification of true genetic signals and advance our understanding of disease biology.

As a complementary approach to SNP-level analysis, gene set analysis, also referred to as pathway analysis, of GWAS has been proposed (WANG et al. 2007; WANG et al. 2010). Gene set analysis examines groups of functionally related genes, each of which might contribute a small and individually undetectable effect on the phenotype. The hypothesis is that when examined jointly, the combined effect of all the genes would rise to the detectable level. Gene sets are usually predefined canonical pathways, gene ontology terms, or derived from protein-protein interaction networks or other gene networks. Gene set analysis is based on the premise that although many genes may be involved in complex disease, these genes are unlikely to be randomly distributed; rather they should converge on biological pathways and molecular networks underlying the disease. Compared to SNP-level analysis, gene set analysis has several advantages. For example, it can increase power by reducing multiple testing and aggregating weak signals distributed across a gene set; it is arguably more suited for complex diseases characterized by genetic heterogeneity; and it can provide more insights into disease biology. Indeed, gene set analysis has shown great success in identifying biological pathways underlying many diseases, including autoimmune
diseases (Wang et al. 2009), cancer (Chen and Gyllensten 2014), psychiatric disorders (Consortium 2015), and others.

A variety of methods exist for gene set analysis of GWAS, many of which were evolved from those for gene expression data. Gene set analysis approaches can be broadly categorized into competitive or self-contained methods based on the null hypothesis tested (Wang et al. 2010). The null hypothesis for competitive methods is that genes in a gene set are not more strongly associated with the disease than genes outside the gene set. Examples of competitive tests include gene set enrichment analysis (Wang et al. 2007), the hypergeometric test, and their extensions, such as ALIGATOR (Holmans et al. 2009). For self-contained gene set analysis, the null hypothesis assumes that none of the genes in a gene set are associated with the disease. Therefore, it directly tests association for a gene set and does not depend on genes outside the set. Some self-contained methods include the SNP-ratio test (SRT) (O’dushlaine et al. 2009), the sequence kernel association test (SKAT) (Wu et al. 2010), PLINK-set (Purcell et al. 2007), GRASS (Chen et al. 2010), aSPUpath (PAN et al. 2015), MAGMA (De LEEUW et al. 2015) and methods combining p-values. Comprehensive reviews and comparisons of existing methods have been offered by several authors (Atias et al. 2013; MOONEY et al. 2014; Wang et al. 2010; Wang et al. 2011). No consensus has been reached as to which method is optimal, but it is generally believed that competitive test approaches are more conservative than self-contained tests (GOEMAN and BüHLMANN 2007).

Existing self-contained methods for gene set analysis generally suffer from several important statistical and computational issues. First, most current methods test each gene set equally and do not consider their unique association patterns. It is well known that not all genes in a gene set are causal and the proportion of risk genes may also vary across gene sets. A test that is optimal for one gene set may be underpowered for another if the two sets have very different proportions of risk genes. For instance, if a gene set contains a large proportion of risk genes, a test that combines all genes may be the most powerful; in contrast, the same test will be less powerful for another set that contains only a few causal genes. As we never know the true association pattern of a gene set, it is statistically important and challenging to adaptively aggregate information over causal genes while minimizing the noise of
noncausal genes. Second, many current methods use a permutation-based approach to derive p-values. Not only is this computationally demanding, but, in addition, there are situations in which permutation is not easy or not possible, such as in family-based GWAS designs and gene set analyses for GWAS meta-analytic results from large consortia. Third, current methods often require raw genotype data as input, but it is not always possible to obtain such data.

To address the limitations of existing methods, we propose a flexible and adaptive test for gene sets (FLAGS) using summary statistics. FLAGS takes into account the unique association pattern of each gene set by adaptively aggregating gene signals derived from SNP-level summary statistics. Extensive simulations show that FLAGS has an appropriate type I error rate and outperforms existing methods with increased power. We further demonstrate by comparative analysis of two real datasets that FLAGS is competitive with and complementary to existing methods. In summary, FLAGS has several advantages over existing methods, including: 1) It does not require raw genotype data, rather it only needs SNP-level p-values and genotype data of ancestry-matched reference samples; 2) Permutation is not needed to calculate the gene set p-values. Instead, it uses a computationally efficient simulation approach based on a multivariate normal distribution; 3) It is more flexible and can be applied to almost any GWAS design that has SNP-level p-values available; and 4) It can adaptively choose the most likely subset of causal genes (if any) and maintain high performance across a wide range of disease models.

Materials and Methods

An overview of FLAGS

An overview of FLAGS is illustrated in Figure 1. Briefly, given a gene set, SNP-level p-values, and genotype data of ancestry-matched reference samples, we first compute a p-value for each gene using VEGAS, a popular gene-based test using SNP-level summary statistics (Liu et al. 2010). The gene-based p-value is then converted to a z-score according to an inverse normal distribution function. The converted z-score is positive (negative) if the p-value is less (larger) than 0.5. Under the null
hypothesis of no genetic effect, the z-scores of all genes in a gene set should jointly follow a multivariate normal distribution with mean 0 and covariance matrix $\Sigma(G)$, which reflects the correlation of z-scores for all gene pairs. We note that the covariance matrix can be estimated based on the linkage disequilibrium (LD) information of ancestry-matched reference samples as shown in Figure S1. Once the covariance matrix is estimated, FLAGS will simulate a sufficient number of multivariate normally distributed vectors with mean 0 and covariance matrix $\Sigma(G)$, to serve as the null distribution of joint z-scores of the gene set. To adaptively combine gene signals, FLAGS will construct a class of test statistics that sum over the z-scores for various proportions of the most significant genes (for example, the top 10%, 20%, …, 100% of most significant genes). The empirical p-value of each test statistic is then the proportion of simulated null statistics that are equal to or larger than the observed statistic. To capture the most likely risk genes, FLAGS will scan the series of test statistics aforementioned and take their minimum p-value as an adaptive test statistic. The null distribution of the adaptive test statistic (minimum p-value) can be obtained from the same set of simulated z-score vectors. The p-value of the adaptive statistic is the proportion of null minimum p-values that are equal to or less than the observed minimum p-value.

**Estimation of gene correlation matrix**

To obtain the null distribution of joint z-scores for a gene set, it is necessary to obtain the covariance matrix of z-scores among genes. Two genes should show correlation of their z-scores when they are physically close on the same chromosome. The closer the two genes are located, the higher the expected correlation due to LD. Inspired by VEGAS, we estimated the correlation of z-scores for a given gene pair based on the LD matrix of SNPs within the two genes. The LD information can be obtained from ancestry-matched reference samples, such as those from HapMap, 1000 Genomes, or a custom set of individuals if genotype data is available. Figure S1 shows the schematic diagram of our simulation process to estimate the correlation of z-scores between two genes. In brief, for any given pair of genes, we simultaneously generate $K$ number of multivariate normal vectors $Z$ for both genes with mean 0 and covariance matrix equal to the pairwise LD matrix of SNPs within the two genes. $Z$ is then transformed
into a vector of Chi-squared variables Q with 1 degree of freedom. For each simulation sample of Q, we calculate a gene-based statistic by summing over the same proportion of top ranked Q values as done for the observed gene-based statistic in VEGAS. The null distribution of the gene-based statistic for each Q sample is approximated by the same gene-based statistics calculated from the remaining Q samples. We then derive a gene-based p-value for each simulated Q sample based on its gene-based statistic and null distribution. Therefore, we can get K number of gene-based p-values from each of the K simulated Q samples. After converting K number of gene-based p-values into z-scores, we obtain K number of z-scores for both genes simultaneously. The correlation of the z-score between two genes can then be calculated based on the K number of z-scores obtained for each gene.

To investigate the validity of this LD-based simulation approach, we compared gene correlations estimated by the simulation approach to those obtained from a standard permutation-based approach. Specifically, we constructed null data sets with individual-level genotype data for the first 500 genes on chromosome 22 using real GWAS data from the Atherosclerosis Risk in Communities (ARIC) study (INVESTIGATORS 1989). We randomly selected 500 samples as cases and 500 samples as controls, and permuted case-control status 10,000 times. For each permuted phenotype, we performed an association test for each SNP using PLINK logistic regression (PURCELL et al. 2007). For each gene, we obtained 10,000 p-value vectors with each recording p-values of SNPs within that gene from one permuted phenotype. P-value vectors were then transformed into vectors of Chi-squared variables Q with 1 degree of freedom. We then calculated a gene-based p-value for each gene in the same way as we did in the LD-based simulation approach. We further converted gene-based p-values to z-scores and obtained 10,000 z-scores for each gene. The correlations of all pairwise gene z-scores for these 500 genes were then calculated and compared with those obtained by the LD-based simulation approach.

**A class of tests and an adaptive test**

The proposed gene set method will employ VEGAS to compute a gene-based p-value for each gene using SNP-level p-values and ancestry-matched LD information. The gene-based p-value will then be converted to a z-score according to $z = \Phi^{-1}(1 - p)$, where $\Phi^{-1}$ denotes the inverse normal
distribution function. FLAGS will construct a number of proportion test statistics that are the sum of z-scores over various proportions of the most significant genes (for example, the top 10%, 20%, ..., 100% of the most significant genes). Formally, we define the proportion test statistic as $T(\alpha) = \sum_{i \in \alpha} Z_i$, where $\alpha$ represents the top $\alpha\%$ of the most significant genes, and $T(\alpha)$ refers to the sum of the top $\alpha\%$ of z-scores. With various values of $\alpha$, we obtain a class of proportion tests. Under the null hypothesis of no genetic effect, the z-scores of the gene set jointly follow a multivariate normal distribution with mean 0 and covariance matrix $\Sigma(G)$, which can be estimated based on the simulation approach described above.

To estimate the significance of each proportion test statistic, FLAGS will simulate $B$ number of multivariate normal vectors with mean 0 and covariance matrix $\Sigma(G)$. We calculated the null statistic $T^{(b)}(\alpha) = \sum_{i \in \alpha} Z_{i}^{(b)}$, $b = 1, 2, ..., B$. The p-value for the $T(\alpha)$ statistic is:

$$p(\alpha) = \left( \sum_{b=1}^{B} \left[ I \left( T^{(b)}(\alpha) \geq T(\alpha) \right) + 1 \right] \right) / (B + 1),$$

where $I(x)$ is an indicator function.

Our empirical observation is that, depending on the proportion of causal genes in a gene set, some proportion tests may be more powerful than others. For example, if 10% of the genes in a set are causal, the test statistic summing over the top 10% of most significant genes tends to be the most powerful. Since we do not know the true proportion of causal genes in a gene set, it cannot be known in advance which proportion test statistic will be the most powerful. Nonetheless, we do know that one of these proportion tests should be more powerful than others. To adaptively aggregate information over genes and maintain high power across different association patterns, FLAGS will scan a class of proportion test statistics and take their minimum p-value as a data-adapted statistic. Formally, suppose that we have a class of values of $\alpha$ in $\Gamma$, e.g., $\Gamma = \{10, 20, 30, ..., 100\}$, and suppose that the p-value of $T(\alpha)$ is $p(\alpha)$, then our adaptive statistic is $T = \min_{\alpha \in \Gamma} p(\alpha)$. The null distribution of the adaptive statistic (minimum p-value) can be obtained from the same set of simulated z-score vectors. For each simulated z-score vector, we calculate the corresponding test statistic $T^{(b)}(\alpha)$, and their p-values as
\[ p^{(b)}(\alpha) = \left( \sum_{b \neq b} I \left( T^{(b)}(\alpha) \geq T^{(b)}(\alpha) \right) \right) + 1) / B. \]

Thus, we have \( T^{(b)} = min_{\alpha \in \Gamma} p^{(b)}(\alpha) \), and the p-value of the adaptive test statistic is

\[ p = \left( \sum_{b=1}^{B} I( T^{(b)} \leq T) \right) + 1) \left( B + 1 \right). \]

**A brief review of some existing tests**

We compared the power of FLAGS with several existing self-contained tests. Below, we briefly review a number of tests that have been widely applied to the analysis of GWAS data (SRT, PLINK-set, SKAT and GRASS), or that have been recently reported (aSPUpath and MAGMA).

**SNP ratio test (SRT):** SRT evaluates the proportion of significant SNPs to all SNPs within genes of a gene set. It computes an empirical p-value using a permutation-based approach. Specifically, the test statistic of SRT represents the ratio of the number of significant SNPs (p < 0.05) to the total number of SNPs in a gene set. The null distribution of the test statistic is obtained by permutation. An empirical p-value of the test statistic is the proportion of null statistics that are equal to or greater than the observed statistic.

**PLINK-set:** PLINK-set is a SNP set analysis implemented in PLINK, and can be naturally extended for gene set analysis. For a given gene set, it first conducts a single SNP-level analysis. It then performs a p-value informed LD pruning to select the top independent SNPs with p-value below a predefined threshold (e.g., p < 0.05). The test statistic is the mean of these single SNP statistics from the top independent SNPs. The p-value of the test statistic is estimated by permutation, which is the proportion of permuted statistics that are equal to or exceed the observed one.

**SKAT:** SKAT is a SNP set analysis approach that is conducted in a kernel-machine regression framework. Similar to PLINK-set, SKAT can also be applied to gene set analysis. It aggregates signals of variants through a kernel matrix. It is computationally efficient because it does not need permutation to
derive p-values. It first fits the null model in which the phenotype is regressed on the covariates alone. A variance component score test is then applied to calculate the p-value analytically.

**GRASS:** GRASS investigates the association of a gene set with the phenotype using principle component analysis (PCA) and regularized regression. It first derives principle components or “eigenSNPs” of each gene by PCA. It then uses a regularized regression technique, termed group ridge regression, to select representative eigenSNPs for each gene and assess their joint association with disease risk. The test statistic for a gene set is defined as $T = (\beta_1^2 + \cdots + \beta_G^2)^{1/2}$, where $\beta_i, i = 1, \cdots G$ are the standardized regression estimates for each gene $i$ in the gene set, and $G$ is the number of genes in the gene set. The p-value of a test statistic for the gene set can be derived through the same permutations, and it is defined as the proportion of permuted statistics that are equal to or larger than the observed statistic.

**aSPUpath:** The aSPUpath test is a test for gene sets extended from an adaptive gene-based test— the adaptive sum of powered score test (SPU) (Pan et al. 2014; Pan et al. 2015). The essence of aSPUpath test is that it adaptively aggregates signals at both the gene and pathway level through a power-weighting scheme. At the power of one, aSPUpath treats each variant and gene equally and takes the average signals for both genes and pathway. If the power is raised to a very large number, it tends to use the maximum signal as the test statistic. By varying the value of the power, aSPUpath adaptively identifies a subset of variants and genes that are most likely to be causal. However, it needs individual level data and permutations to derive p-values for a gene set.

**MAGMA:** MAGMA is a novel tool for gene and gene set analysis of GWAS data (de Leeuw et al. 2015). To perform gene set analysis, it converts the gene-based p-value to a z-score based on the inverse normal distribution function. It then tests whether the mean of z-scores of all genes in a gene set deviates from zero in a regression model while accounting for the correlations among genes. It does not require permutation to derive p-values.

**Simulations**
We evaluated the performance of FLAGS using genotype data with realistic LD patterns. Specifically, we conducted simulations using ARIC GWAS data from ~9,000 samples (INVESTIGATORS 1989). To reflect the complex configurations of gene sets, we simulated data using gene sets in MSigDB database (SUBRAMANIAN et al. 2005). We performed standard quality control for the GWAS data before it was used to generate simulated data. Briefly, we excluded samples and SNPs based on predetermined quality control metrics, including sample call rate ≤ 95%, SNP call rate ≤ 95%, MAF ≤ 0.01, and p-values of Hardy-Weinberg Equilibrium (HWE) tests ≤ 10^{-5}. We computed principal components (PCs) for GWAS samples using EIGENSOFT (PRICE et al. 2006). We removed outlier samples defined as subjects whose ancestry was at least three standard deviations from the mean of one of the two largest PCs. SNPs were mapped to genes based on NCBI 37.3 gene definitions.

**Type I error**

We aimed to examine the type I error rates of FLAGS, and to evaluate the extent to which the type I error would be influenced if gene correlations within a gene set were not considered. For this purpose, we generated null datasets for two pathways selected from MSigDB database, representing gene sets with a low or high level of gene correlations. Specifically, the gene set of low level gene correlations has 19 genes, and contains only one gene pair with a z-score correlation ≥ 0.2. The gene set of high level gene correlations includes 13 genes, among which eight gene pairs have a z-score correlation ≥ 0.2. To generate null data sets, 1,000 random samples were drawn without replacement from the ARIC GWAS data. We randomly assigned half the samples as cases and the other half as controls. We then permuted case-control labels 10,000 times to generate 10,000 null phenotypes, and hence 10,000 null datasets for each gene set. For each null dataset, we ran PLINK logistic regression to get SNP-based p-values, followed by applying FLAGS to get the gene set-based p-values with and without considering gene correlations. The empirical type I error rate was calculated as the proportion of 10,000 p-values from the null data that were equal to or smaller than a given \( \alpha \) level (0.05, 0.01, and 0.001). To evaluate the effect of sample size on type I error, we repeated the simulation for a sample size of 2,000 (1,000 cases and 1,000 controls).
Power

To compare the power of FLAGS and other existing methods, we simulated data under alternative hypotheses using curated gene sets from MSigDB database, including KEGG, BioCarta, and Reactome. We reasoned that a simulation based on a variety of gene sets across the genome would ensure that the power estimation was not dependent on the choice of a specific set, and hence it would more realistically reflect the performance of various methods. We selected 1,000 gene sets with 10-200 genes as “causal gene sets”. The phenotype was generated based on the ARIC GWAS data for each causal set. Specifically, for each gene set, we randomly selected K number of genes as causal genes, and one causal SNP (0.05<MAF<0.25) was randomly taken from each causal gene, yielding K causal genotypes \((g_1, g_2, \cdots, g_K)\). Then, we generated a quantitative trait by a simple linear regression model:

\[
y = \beta_1 g_1 + \beta_2 g_2 + \cdots + \beta_K g_K + \epsilon,
\]

where \(\beta_i (i = 1, 2, \cdots, K)\) is the additive effect for the causal variants, and \(\epsilon\) is a random term that follows a standard normal distribution \(N(0,1)\). A dichotomous trait was created by taking samples with the highest (\(\geq 75^{th}\) percentile) and lowest quartiles (\(\leq 25^{th}\) percentile) of the simulated quantitative trait as cases and controls, respectively. The genetic effect of all causal variants was set as the same as \(\log(1.1)\).

We considered four different settings for the proportion of causal genes contained in each causal gene set (10\%, 20\%, 30\% or 40\%). We therefore created 1,000 causal datasets under each disease model. For each causal dataset, FLAGS and other competing methods were run to get a gene set-based p-value. The power was defined as the proportion of p-values from 1,000 causal datasets that were smaller than a given \(\alpha = 0.01\) level. We considered two sample sizes of 1,000 (500 cases and 500 controls) and 2,000 (1,000 cases and 1,000 controls).

Real data analysis

To evaluate the performance of FLAGS on real data sets and compare it with existing methods, we applied FLAGS and other appropriate methods to two datasets: 1) Wellcome Trust Case Control Consortium (WTCCC) GWAS data for Crohn’s disease (CD, [MIM: 266600]) (CONSORTIUM 2007), and
2) GWAS meta-analysis results for bipolar disorder (BD, [MIM: 125480]) from the Psychiatric Genomic Consortium (PGC) (GROUP 2011). The CD data include the raw data, and therefore enabled us to compare FLAGS with four other methods that need raw data as input (aSPUpath, GRASS, SKAT, MAGMA). We did not include STR and PLINK-set for CD, because these two methods are too computational demanding. In the comparative analysis of BD, we only applied MAGMA because it is the sole method that can take summary statistics as input. We focused our analysis on autosomal SNPs. We assigned a SNP to a gene if it was located within the gene, based on NCBI 37.3 gene annotation, or within 20 kb upstream and downstream of the gene, to capture regulatory variants. To reduce multiple testing and choose gene sets that are relatively well defined, we only used gene sets from the KEGG database (KANEHISA et al. 2010). Following previous authors, to facilitate interpretation of the results, we excluded gene sets that were either unusually small (<10 genes) or unusually large (>500 genes). In total, there were 64,557 and 178,873 SNPs mapped to 4,572 and 4,904 genes for CD and BD, respectively, which resulted in 197 gene sets for CD and 186 gene sets for BD. To reduce systematic bias and minimize the chance of false positive findings, we used genomic control-corrected SNP p-values for both disorders. For gene set analyses that require gene-based p-values, we computed these with VEGAS using a statistic summarizing the top 5% of most significant SNPs. For CD, we used genotype information from healthy controls in the CD GWAS as reference samples to compute the gene-based p-values and estimate gene correlations, whereas for BP, we did so by using the unrelated CEU samples from the 1000 Genome Project. Because the gene correlation in z-scores is close to zero when two genes were more than 2Mb away (Figure S2), we only calculated correlations for gene pairs within 2Mb.

R code and data examples for implementing FLAGS were posted at http://www.medicine.uiowa.edu/psychiatry/han/software.

Results

Estimation of gene correlations

We estimated gene correlation in z-scores using an LD-based simulation approach. There was a high correlation ($r^2=0.96$) between results obtained from our simulation-based approach (10,000
simulations) and those from the permutation-based approach (Figure 2). To investigate how many simulations are needed to obtain a stable estimate of gene correlations, we varied the number of simulations and compared the corresponding results with those from the permutation-based approach. We found that 3,000 simulations will generally be enough for a stable estimation of gene correlations ($r^2=0.9$, Figure S2). We also observed that the strength of gene correlation was mainly determined by physical distance. The correlation of z-scores is close to zero when two genes were more than 2Mb apart (Figure S3). Based on this empirical observation, we recommend that the estimation of gene correlation is only necessary for closely located gene pairs, for example, within 2Mb.

**Type I error**

We evaluated the type I error of FLAGS, and to what extent the type I error would be influenced when gene correlations are not considered within a gene set. Table 1 shows the empirical type I error rates at three nominal error rates ($\alpha=0.05$, $\alpha=0.01$, and $\alpha=0.001$) for two gene sets, which represent a low and high level of gene correlation, respectively. Overall, all FLAGS test statistics maintained an appropriate type I error rate when gene correlation was accounted for. However, the type I error rates tended to be inflated if gene correlations were ignored. The extent of inflation of type I error was dependent on the level of gene correlation within the gene set. For example, the type I error rates were only slightly higher than the nominal levels for the gene set in the presence of a low level of gene correlation. However, the type I error rates were considerably inflated when gene correlation was ignored for the gene set with a high level of gene correlation. We found a similar pattern of type I error rates based on simulations with a larger sample size (1,000 cases and 1,000 controls; Table S1).

**Power**

Using a number of simulated disease models, we investigated the power of FLAGS and compared its performance with several other self-contained tests (Figure 3). We considered four disease models with various proportions of causal genes (10%, 20%, 30% and 40%). As expected, among all FLAGS proportion tests, we observed the highest power for tests that combined the top significant genes with a proportion that was equal or close to the proportion of causal genes. For example, for disease models with
10% and 20% causal genes, the highest power was observed for the test that combined the top 10% and 20% of significant genes, respectively. The power of the adaptive test was close to that of the most powerful proportion test. It was noted that, for disease models with a low proportion of causal genes (10% or 20%), the performance of FLAGS proportion tests deteriorated dramatically when an increasing proportion of genes was included in the simulations. On the other hand, for the disease model with a high proportion of causal genes (40%), FLAGS proportion tests did not suffer from substantial power loss when more genes were added into the simulations. In comparison with other self-contained tests, the FLAGS adaptive test was clearly the winner in all simulated models. The power of the FLAGS adaptive test is around 10%-30% higher than other methods across various disease models. In the scenario of a low proportion of causal genes (10% and 20%), SRT and MAGMA have the lowest power, possibly due to their low performance in the situation of low signal to noise ratio. However, aSPUpath performs slightly better than other methods, perhaps because of its adaptive nature. In the case of a high proportion of causal genes (40%), MAGMA performs the best among all other compared methods. Power simulations using a larger sample size (1,000 cases and 1,000 controls) yielded the same conclusions (Figure S4).

Real data

Crohn’s disease (CD): We applied the FLAGS adaptive test, as well as four other methods, to the GWAS data for CD. Table 2 shows 17 KEGG gene sets that were identified by FLAGS and remained significant after multiple testing correction ($p < 2.5\times10^{-4}$). Interestingly, five positive control gene sets that have been confirmed to be associated with CD are all among the 17 gene sets. These gene sets are all related to the immune system and represent the best understood pathways underlying CD. It is noteworthy that only FLAGS and aSPUpath identified all positive controls at $p$-values $< 2.5\times10^{-4}$. All methods detected the JAK-STAT signaling pathway (hsa04630) and the cytokine-cytokine receptor interaction pathway (hsa04060). However, GRASS did not identify the T cell receptor signaling pathway (hsa04660, $p = 0.078$) and the chemokine signaling pathway (hsa04062, $p = 0.0019$). The NOD-like receptor signaling pathway (hsa04621) was missed by MAGMA ($p = 0.39$) and SKAT ($p = 0.002$). FLAGS also identified three additional gene sets — Graft-versus-host disease (hsa05332, p=5×10^{-5}), cell adhesion
molecules (hsa04514, p=1×10^{-5}), and vasopressin-regulated water reabsorption (hsa04962, p=1.4×10^{-4}) — which were not significant by aSPUpath, but were significant by GRASS or SKAT.

**Bipolar disorder (BD):** FLAGS and MAGMA were applied to the BD GWAS meta-analysis results. We only applied MAGMA for comparison with FLAGS because it is the sole method that can take summary statistics as input. The significance threshold is p = 2.7×10^{-4} after Bonferroni correction for the number of gene sets tested for the BD data set. To evaluate both approaches, we included one positive control gene set from GO, histone H3-K4 methylation (term 51568). This gene set was the most strongly associated one for BD in a recent study (CONSORTIUM 2015). Interestingly, FLAGS detected the positive control gene set with a p-value that remained significant after multiple testing correction (p = 2.3×10^{-5}), but it showed only nominal significance by MAGMA (p = 0.003). Besides the positive control gene set, there were eight other gene sets that showed p-values < 0.001 by either method (Table 3). Of note, none of these survived multiple testing correction in the MAGMA analysis. However, FLAGS detected five additional gene sets that were significant after such correction, including Type II diabetes mellitus (p = 2.2×10^{-5}), MAPK signaling pathway (p = 1.1×10^{-4}), glycosphingolipid biosynthesis-globo series (p = 1.6×10^{-4}), GnRH signaling pathway (p = 2.2×10^{-4}), and gap junction (p = 2.4×10^{-4}).

**Discussion**

Motivation for the proposed FLAGS method is based on the consideration that only a fraction of genes are causal within a gene set that underlies disease susceptibility. Moreover, the proportion of risk genes may also vary across different gene sets. A robust test statistic should take into account the unique association patterns of gene sets. Accordingly, FLAGS was designed to adaptively aggregate information over the most likely causal genes while minimizing the noise of noncausal genes. FLAGS showed consistent high performance over a wide range of disease models. The robustness of FLAGS is consistent with a recent study that found the adaptive rank truncated product (ARTP) to perform best in a comparison of a number of set-based association tests (Su et al. 2015). The ARTP adaptively seeks a subset of SNPs that yields the best evidence for genetic association. The ARTP method is analogous to
FLAGS in that both methods form a model selection that aims to maximize the statistical evidence. However, the ARTP method needs raw data and permutations to derive p-values, which is not only computationally intensive, but, additionally, not possible or very difficult in many situations. In contrast, FLAGS needs only SNP-level p-values and ancestry-matched LD information. Moreover, FLAGS does not use permutation to estimate p-values; instead it uses a simulation-based approach to derive p-values, and it is therefore much faster and more computationally efficient.

Although the FLAGS method has been developed for GWAS data, the same adaptive framework can be applied to gene set analysis of rare variants, such as those generated from whole genome or exome sequencing studies. However, in order to calculate gene correlations based on rare variants, we recommend using a standard permutation-based approach, which requires that the raw data be available. In the current application of FLAGS we used VEGAS to compute gene-based p-values, although other gene-based tests may also be applied that can take SNP-level summary statistics as input. However, we point out that the current approach to estimating gene correlations was based on gene-level test statistics from VEGAS that use LD information from ancestry-matched reference samples. To estimate gene correlations based on other gene test statistics, we recommend that a permutation-based approach be employed using genotype data from reference samples. In addition, our method introduces a natural way to incorporate weights on genes that are more likely to be causal. Specifically, the weights can be incorporated in the null distributions of joint z-scores of genes within a gene set. The weights may come from prior information, such as gene expression differences, linkage signals, and other biological knowledge.

In the analysis of two real datasets for CD and BD, FLAGS not only detected all positive control gene sets for both diseases, but also identified novel gene sets that may underlie the pathophysiology. For example, one top hit gene set for CD is the “calcium signaling pathway”. Calcium signals are crucial for the proper activation of lymphocytes, regulation of cell differentiation, and effector functions (V1G and KINET 2009). Dysregulated calcium responses have been implicated in several autoimmune and inflammatory diseases, such as systemic lupus erythematosus, rheumatoid arthritis, and multiple sclerosis.
Detection of the calcium signaling pathway may reveal new insights into CD pathogenesis and provide potential targets for treatment. In the test of BD meta-analysis results, the most significant pathway was identified for “type II diabetes mellitus” which is in line with prior evidence supporting a shared genetic component between diabetes and BD (TORKAMANI et al. 2008). The second top gene set was the “MAPK signaling pathway”. Interestingly, a number of studies have shown that the two most commonly used anti-manic agents, lithium and valproate, activate the MAPK signaling cascade, which may in part be responsible for their therapeutic effects (BOECKELER et al. 2006). The third ranked gene set “glycosphingolipid biosynthesis-globo series” produces glycosphingolipids (GSLs) that are especially abundant in the nervous system. Mounting evidence supports regulatory roles of GSLs in neurogenesis and signal transduction in the developing human brain (FURUKAWA et al. 2014). FLAGS also identified another significant gene set, “gap junction,” which is consistent with its important roles in the nervous system.

In large-scale data analysis, not only statistical power but also computational demand is a major consideration. We compared the running time of five different methods applied to CD data based on a 16 core Xeon Phi (2.6GHz) with 256 GB of RAM (Table S2). In the analysis of 197 KEGG gene sets using only a single core, it took five and 57 minutes for MAGMA and SKAT, respectively. The running time for FLAGS and aSPUpath was 42 and 68 hours, respectively. GRASS is the most computationally demanding, as it can take more than 10 days. However, in our real practice where we have submitted jobs in parallel, the computational time was greatly reduced for FLAGS, aSPUpath, and GRASS. For example, it took ~11 hours for FLAGS, 13 hours for aSPUpath, and five days for GRASS, through running 80 parallel jobs on the Linux clusters.

In summary, we have developed a flexible and powerful adaptive test for gene sets using summary statistics. Although our method was simulated and applied based on binary traits, it can be utilized for other types of traits, such as quantitative and ordinal ones. Given the fact that GWAS summary results are increasingly publicly available, and given the difficulties that can arise in accessing
raw data, our method will allow for the broader application of gene set analysis to GWAS of complex
diseases.

Acknowledgments

This study was supported by National Institutes of Health grants R01AA022994. This study was
also supported by the NARSAD Young Investigator Award (S.H.). We are grateful for the insightful
comments and suggestions by the reviewers. This study made use of data generated by the Wellcome
Trust Case Control Consortium. A full list of the investigators who contributed to the generation of the
WTCCC data is available from http://www.wtccc.org.uk. Funding for the WTCCC project was provided
by the Wellcome Trust under award 076113.
References


CONSORTIUM, W. T. C. C., 2007 Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447: 661-678.


Figure 1. Overview of the FLAGS method. First, the input information includes a gene set, p-values of SNPs mapped to the gene set, and ancestry-matched LD information. Second, it computes a gene-based p-value using VEGAS and converts p-values to z-scores for each gene according to an inverse normal distribution function. To derive the null distributions of test statistics, FLAGS will simulate a sufficient number of z-score vectors based on a multivariate variate normal (MVN) distribution function. Third, FLAGS will construct a class of test statistics that sum over the z-scores for various proportions of the genes.
most significant genes, followed by an adaptive test statistic with a p-value estimated using the same simulated z-score vectors.

Figure 2. Comparison of gene correlations in z-scores estimated by an LD-based simulation approach to those obtained from a permutation-based approach. Comparisons were made for all gene pairs from the first 500 genes on chromosome 22 using real GWAS data from the Atherosclerosis Risk in Communities (ARIC) study. We found that both methods produced similar results and correlation of the two sets of results was very high based on 10,000 simulations ($r^2=0.96$).
Figure 3. Empirical power estimation for various methods under different disease models with 500 cases and 500 controls. A) gene sets contain 10% causal genes, B) gene sets contain 20% causal genes, C) gene sets contain 30% causal genes, and D) gene sets contain 40% causal genes.
Table 1. Empirical type I error rates for two gene sets at three nominal error rates (\(\alpha=0.05\), \(\alpha=0.01\), and \(\alpha=0.001\)) based on 500 cases and 500 controls

| Test statistics                           | Type-I error rate (gene correlation controlled for?) |  |
|-------------------------------------------|-----------------------------------------------------|--|---|---|---|---|---|---|---|---|---|---|---|
|                                           | (Yes)                                               | (No) | (Yes) | (No) | (Yes) | (No) | (Yes) | (No) | (Yes) | (No) | (Yes) | (No) | (Yes) | (No) |
|                                           | \(\alpha=0.05\)                                   | \(\alpha=0.01\) | \(\alpha=0.001\) | \(\alpha=0.05\) | \(\alpha=0.01\) | \(\alpha=0.001\) | \(\alpha=0.05\) | \(\alpha=0.01\) | \(\alpha=0.001\) | \(\alpha=0.05\) | \(\alpha=0.01\) | \(\alpha=0.001\) |
| Gene set with a low level of gene correlations |                                                     |     |     |     |     |     |     |     |     |     |     |     |     |     |
| 10\%                                      | 0.051                                               | 0.01 | 0.0012 | 0.057 | 0.012 | 0.0013 | | | |
| 20\%                                      | 0.052                                               | 0.01 | 0.001 | 0.057 | 0.013 | 0.0012 | | | |
| 30\%                                      | 0.051                                               | 0.01 | 0.0009 | 0.056 | 0.013 | 0.0011 | | | |
| 40\%                                      | 0.049                                               | 0.01 | 0.0008 | 0.055 | 0.012 | 0.001 | | | |
| 50\%                                      | 0.049                                               | 0.0099 | 0.0006 | 0.055 | 0.013 | 0.0008 | | | |
| 60\%                                      | 0.049                                               | 0.0099 | 0.0007 | 0.055 | 0.013 | 0.0009 | | | |
| 70\%                                      | 0.048                                               | 0.0097 | 0.0009 | 0.053 | 0.012 | 0.0011 | | | |
| 80\%                                      | 0.046                                               | 0.0092 | 0.001 | 0.053 | 0.012 | 0.0011 | | | |
| 90\%                                      | 0.048                                               | 0.0089 | 0.0009 | 0.054 | 0.012 | 0.0012 | | | |
| 100\%                                     | 0.046                                               | 0.0097 | 0.0007 | 0.053 | 0.012 | 0.001 | | | |
| Adaptive                                   | 0.050                                               | 0.01 | 0.001 | 0.056 | 0.012 | 0.0014 | | | |
| Gene set with a high level of gene correlations |                                                     |     |     |     |     |     |     |     |     |     |     |     |     |     |
| 10\%                                      | 0.049                                               | 0.0088 | 0.001 | 0.094 | 0.027 | 0.0049 | | | |
| 20\%                                      | 0.051                                               | 0.0092 | 0.0008 | 0.1 | 0.032 | 0.0061 | | | |
| 30\%                                      | 0.051                                               | 0.0095 | 0.0008 | 0.1 | 0.034 | 0.0064 | | | |
| 40\%                                      | 0.051                                               | 0.0096 | 0.0008 | 0.1 | 0.035 | 0.0067 | | | |
| 50\%                                      | 0.050                                               | 0.0099 | 0.0008 | 0.1 | 0.036 | 0.0072 | | | |
| 60\%                                      | 0.050                                               | 0.0099 | 0.001 | 0.1 | 0.036 | 0.0079 | | | |
| 70\%                                      | 0.050                                               | 0.01 | 0.0011 | 0.1 | 0.036 | 0.0087 | | | |
| 80\%                                      | 0.050                                               | 0.01 | 0.001 | 0.1 | 0.035 | 0.0079 | | | |
| 90\%                                      | 0.049                                               | 0.0094 | 0.0014 | 0.1 | 0.036 | 0.0074 | | | |
| 100\%                                     | 0.048                                               | 0.0087 | 0.0011 | 0.094 | 0.034 | 0.0073 | | | |
| Adaptive                                   | 0.051                                               | 0.0096 | 0.0008 | 0.11 | 0.037 | 0.0078 | | | |
Table 2. Results of the WTCCC CD GWAS data application: KEGG gene sets with p-values < $2.5 \times 10^{-4}$ by FLAGS and corresponding results by aSPUpath, GRASS, MAGMA and SKAT

<table>
<thead>
<tr>
<th>KEGG ID</th>
<th>Gene set names</th>
<th>Genes #</th>
<th>SNPs #</th>
<th>p-values</th>
<th>FLAGS</th>
<th>aSPUpath</th>
<th>GRASS</th>
<th>MAGMA</th>
<th>SKAT</th>
</tr>
</thead>
<tbody>
<tr>
<td>hsa04020</td>
<td>Calcium signaling pathway</td>
<td>155</td>
<td>4,343</td>
<td>&lt;1.0×10^{-5}</td>
<td>1.0×10^{-4}</td>
<td>0.04</td>
<td>2.5×10^{-4}</td>
<td>6.9×10^{-8}</td>
<td></td>
</tr>
<tr>
<td>hsa04060</td>
<td>Cytokine-cytokine receptor interaction</td>
<td>224</td>
<td>2,284</td>
<td>&lt;1.0×10^{-5}</td>
<td>&lt;1.0×10^{-5}</td>
<td>&lt;1.0×10^{-5}</td>
<td>4.8×10^{-11}</td>
<td>2.5×10^{-11}</td>
<td></td>
</tr>
<tr>
<td>hsa04310</td>
<td>Wnt signaling pathway</td>
<td>134</td>
<td>2,079</td>
<td>&lt;1.0×10^{-5}</td>
<td>&lt;1.0×10^{-5}</td>
<td>0.0015</td>
<td>0.029</td>
<td>0.0053</td>
<td></td>
</tr>
<tr>
<td>hsa04514</td>
<td>Cell adhesion molecules (CAMs)</td>
<td>119</td>
<td>3,277</td>
<td>&lt;1.0×10^{-5}</td>
<td>0.0011</td>
<td>&lt;1.0×10^{-5}</td>
<td>4.5×10^{-4}</td>
<td>6.5×10^{-6}</td>
<td></td>
</tr>
<tr>
<td>hsa04630</td>
<td>Jak-STAT signaling pathway *</td>
<td>139</td>
<td>1,328</td>
<td>&lt;1.0×10^{-5}</td>
<td>&lt;1.0×10^{-5}</td>
<td>&lt;1.0×10^{-5}</td>
<td>1.6×10^{-7}</td>
<td>6.3×10^{-11}</td>
<td></td>
</tr>
<tr>
<td>hsa04660</td>
<td>T cell receptor signaling pathway</td>
<td>98</td>
<td>1,371</td>
<td>&lt;1.0×10^{-5}</td>
<td>&lt;1.0×10^{-5}</td>
<td>0.078</td>
<td>2.5×10^{-4}</td>
<td>9.7×10^{-3}</td>
<td></td>
</tr>
<tr>
<td>hsa04940</td>
<td>Type I diabetes mellitus</td>
<td>39</td>
<td>684</td>
<td>&lt;1.0×10^{-5}</td>
<td>1.4×10^{-4}</td>
<td>&lt;1.0×10^{-5}</td>
<td>1.7×10^{-4}</td>
<td>5.9×10^{-6}</td>
<td></td>
</tr>
<tr>
<td>hsa05330</td>
<td>Allograft rejection</td>
<td>34</td>
<td>438</td>
<td>&lt;1.0×10^{-5}</td>
<td>&lt;1.0×10^{-5}</td>
<td>&lt;1.0×10^{-5}</td>
<td>7.3×10^{-7}</td>
<td>2.4×10^{-9}</td>
<td></td>
</tr>
<tr>
<td>hsa05416</td>
<td>Viral myocarditis</td>
<td>65</td>
<td>1,232</td>
<td>&lt;1.0×10^{-5}</td>
<td>5.0×10^{-3}</td>
<td>&lt;1.0×10^{-5}</td>
<td>9.1×10^{-5}</td>
<td>2.0×10^{-5}</td>
<td></td>
</tr>
<tr>
<td>hsa05332</td>
<td>Graft-versus-host disease</td>
<td>33</td>
<td>412</td>
<td>5.0×10^{-3}</td>
<td>7.3×10^{-4}</td>
<td>&lt;1.0×10^{-5}</td>
<td>0.0034</td>
<td>1.6×10^{-5}</td>
<td></td>
</tr>
<tr>
<td>hsa04640</td>
<td>Hematopoietic cell lineage</td>
<td>74</td>
<td>856</td>
<td>6.0×10^{-5}</td>
<td>9.0×10^{-3}</td>
<td>2.0×10^{-5}</td>
<td>9.1×10^{-5}</td>
<td>0.0015</td>
<td></td>
</tr>
<tr>
<td>hsa05310</td>
<td>Asthma</td>
<td>27</td>
<td>240</td>
<td>1.2×10^{-4}</td>
<td>2.6×10^{-4}</td>
<td>&lt;1.0×10^{-5}</td>
<td>9.3×10^{-3}</td>
<td>1.5×10^{-3}</td>
<td></td>
</tr>
<tr>
<td>hsa04962</td>
<td>Vasopressin-regulated water reabsorption</td>
<td>38</td>
<td>571</td>
<td>1.4×10^{-4}</td>
<td>0.002</td>
<td>8.0×10^{-3}</td>
<td>0.075</td>
<td>0.072</td>
<td></td>
</tr>
<tr>
<td>hsa05131</td>
<td>Shigellosis</td>
<td>55</td>
<td>782</td>
<td>2.0×10^{-4}</td>
<td>&lt;1.0×10^{-5}</td>
<td>6.0×10^{-3}</td>
<td>0.07</td>
<td>9.6×10^{-4}</td>
<td></td>
</tr>
<tr>
<td>hsa04622</td>
<td>RIG-I-like receptor signaling pathway</td>
<td>56</td>
<td>441</td>
<td>2.1×10^{-4}</td>
<td>&lt;1.0×10^{-5}</td>
<td>0.0018</td>
<td>0.35</td>
<td>0.055</td>
<td></td>
</tr>
<tr>
<td>hsa04621</td>
<td>NOD-like receptor signaling pathway *</td>
<td>51</td>
<td>480</td>
<td>2.3×10^{-4}</td>
<td>&lt;1.0×10^{-5}</td>
<td>&lt;1.0×10^{-5}</td>
<td>0.39</td>
<td>0.0022</td>
<td></td>
</tr>
<tr>
<td>hsa04062</td>
<td>Chemokine signaling pathway *</td>
<td>158</td>
<td>2,589</td>
<td>2.4×10^{-4}</td>
<td>&lt;1.0×10^{-5}</td>
<td>0.0019</td>
<td>3.9×10^{-5}</td>
<td>6.5×10^{-3}</td>
<td></td>
</tr>
</tbody>
</table>

Asterisks (*) indicate positive control gene sets.
Table 3. Results of the PGC BD GWAS meta-analysis application: KEGG gene sets with p-values < 0.001 by either FLAGS or MAGMA

<table>
<thead>
<tr>
<th>KEGG ID</th>
<th>Gene set names</th>
<th>Genes #</th>
<th>SNPs #</th>
<th>p-values</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>FLAGS</td>
</tr>
<tr>
<td>GO51568</td>
<td>histone H3-K4 methylation</td>
<td>13</td>
<td>303</td>
<td>2.3×10⁻⁵</td>
</tr>
<tr>
<td>hsa04930</td>
<td>Type II diabetes mellitus</td>
<td>45</td>
<td>2,576</td>
<td>2.2×10⁻⁵</td>
</tr>
<tr>
<td>hsa04010</td>
<td>MAPK signaling pathway</td>
<td>254</td>
<td>12,118</td>
<td>1.1×10⁻²</td>
</tr>
<tr>
<td>hsa00603</td>
<td>Glycosphingolipid biosynthesis - globo series</td>
<td>13</td>
<td>528</td>
<td>1.6×10⁻⁴</td>
</tr>
<tr>
<td>hsa04912</td>
<td>GnRH signaling pathway</td>
<td>97</td>
<td>5,699</td>
<td>2.2×10⁻⁴</td>
</tr>
<tr>
<td>hsa04540</td>
<td>Gap junction</td>
<td>85</td>
<td>5,969</td>
<td>2.4×10⁻⁴</td>
</tr>
<tr>
<td>hsa04720</td>
<td>Long-term potentiation</td>
<td>65</td>
<td>5,486</td>
<td>5.5×10⁻⁴</td>
</tr>
<tr>
<td>hsa04722</td>
<td>Neurotrophin signaling pathway</td>
<td>120</td>
<td>5,084</td>
<td>9.6×10⁻⁴</td>
</tr>
<tr>
<td>hsa05010</td>
<td>Alzheimer’s disease</td>
<td>151</td>
<td>6,574</td>
<td>0.0016</td>
</tr>
</tbody>
</table>

Asterisks (*) indicate positive control gene set