Published in final edited form as:
Genet Epidemiol. 2012 January ; 36(1): 3–16. doi:10.1002/gepi.20632.

Using the Gene Ontology to Scan Multi-Level Gene Sets for Associations in Genome Wide Association Studies

Daniel J. Schaid[1], **Jason P. Sinnwell**[1], **Gregory D. Jenkins**[1], **Shannon K. McDonnell**[1], **James N. Ingle**[2], **Michiaki Kubo**[4], **Paul E. Goss**[5], **Joseph P. Costantino**[6], **D. Lawrence Wickerham**[7], and **Richard M. Weinshilboum**[3]

[1]Division of Biomedical Statistics and Informatics, Department of Molecular Pharmacology and Experimental Therapeutics, Mayo Clinic, Rochester, MN

[2]Division of Medical Oncology, Department of Molecular Pharmacology and Experimental Therapeutics, Mayo Clinic, Rochester, MN

[3]Division of Clinical Pharmacology, Department of Molecular Pharmacology and Experimental Therapeutics, Mayo Clinic, Rochester, MN

[4]RIKEN Center for Genomic Medicine, Tokyo, Japan

[5]Massachusetts General Hospital Cancer Center, Harvard University, Boston, MA

[6]Department of Biostatistics, University of Pittsburgh, Pittsburgh, PA

[7]Section of Cancer Genetics and Prevention, Allegheny General Hospital, Pittsburgh, PA

Abstract

Gene-set analyses have been widely used in gene expression studies, and some of the developed methods have been extended to genome wide association studies (GWAS). Yet, complications due to linkage disequilibrium (LD) among single nucleotide polymorphisms (SNPs), and variable numbers of SNPs per gene and genes per gene-set, have plagued current approaches, often leading to ad hoc “fixes”. To overcome some of the current limitations, we developed a general approach to scan GWAS SNP data for both gene-level and gene-set analyses, building on score statistics for generalized linear models, and taking advantage of the directed acyclic graph structure of the gene ontology when creating gene-sets. However, other types of gene-set structures can be used, such as the popular Kyoto Encyclopedia of Genes and Genomes (KEGG). Our approach combines SNPs into genes, and genes into gene-sets, but assures that positive and negative effects of genes on a trait do not cancel. To control for multiple testing of many gene-sets, we use an efficient computational strategy that accounts for LD and provides accurate step-down adjusted p-values for each gene-set. Application of our methods to two different GWAS provide guidance on the potential strengths and weaknesses of our proposed gene-set analyses.

Correspondence: Dr. Daniel J. Schaid Division of Biostatistics Harwick 7, Mayo Clinic 200 First Street SW Rochester, MN 55905, USA Tel. +1 507 284 0639 Fax +1 507 284 9542 schaid@mayo.edu.

Electronic Information
MACH
http://www.sph.umich.edu/csg/abecasis/MACH/index.html
Gene Ontology
http://www.geneontology.org
KEGG
http://www.genome.jp/kegg/

Keywords

gene-sets; genome wide association; pathways; score statistics

Introduction

Genome wide association studies (GWAS) that survey ½ to 1 million single nucleotide polymorphisms (SNPs) for their associations with traits have had an enormous impact on the understanding of some complex traits. More critical, perhaps, is that they provide a glimpse at the large number of genes likely involved in the genetic basis of many common traits. As of July, 2011, there have been 957 published GWAS with individual single-nucleotide polymorphisms (SNPs) associated with over 165 traits, with single-SNP p-values < 5E-8 [1]. Yet, many of the associated SNPs explain only a small fraction of the expected heritability [2; 3]. Several recent studies suggest that a very large number of SNPs, most with very small effects on traits, would be required to explain a large fraction of the heritability of some traits [4; 5]. In this case, extremely large sample sizes would be needed to reliably detect the individual SNPs.

In contrast to thinking of effects of individual SNPs, there is growing experimental evidence that sets of genes, possibly in overlapping pathways, tend to function as sub-networks having a larger effect on a complex trait than individual genes. A good example was shown for tissue-specific gene expression data combined with genotypes measured on a segregating mouse population [6]. The different sub-networks were enriched for a number of biological processes and were interpreted to represent key functional units that underlie processes specific to the different cell types. This suggests that common diseases might result from many genes that are highly interconnected in networks. Following these ideas, and stimulated by gene expression studies, it is anticipated that grouping SNPs into genes, and then genes into sets of similar function, might facilitate biological interpretations of GWAS, and perhaps account for genetic heterogeneity, as well as improve statistical power to detect relevant genes [7]. The benefit of grouping genes into sets would be greatest if many genes in a given set are associated with a trait — by averaging over the scores for genes in a set, the signal for association with a trait can be larger for the average than that for individual gene-scores, because of the reduced variability of the average.

A variety of gene-set methods have been widely used for gene expression studies [8; 9]. But, we take a different strategy for GWAS. To understand our strategy, and appreciate why it differs from many methods used for gene expression studies, it is helpful to briefly review gene-set analyses developed for gene expression studies. A typical gene expression study compares the expression of tens of thousands of genes between two groups, searching for differential expression, resulting in a list of p-values for the different genes. A common approach to analyze a set of genes is to choose a cut-off of statistical significance, and then cross-classify the discrete indicator of significance with whether the genes are in a particular set or not, creating a 2×2 contingency table. The fraction of significant genes can then be compared between those in the set versus the remainder not in the set, using Fisher's exact test. An alterative approach that does not require a cut-off is to order the p-values from smallest to largest (or perhaps some other measure of differential expression), and then test whether the genes in a set tend to clump toward smaller p-values (Gene Set Enrichment Analysis – GSEA). Many other approaches have been developed for lists of p-values, such as adaptively truncating the p-values in a set in order to emphasize the smaller p-values [10], or ad hoc methods that attempt to account for the number of SNPs per gene and the correlation structure among SNPs by use of arbitrary correlation thresholds and heuristic adjustments for both linkage disequilibrium and the number of SNPs [11].

An advantage of many of the proposed methods is that they require only lists of p-values, making them appealing for pooling across multiple studies in meta analyses. Because of their simplicity, they have been adapted and used for GWAS [12-19]. However, there are important limitations, well discussed elsewhere [7-9; 20]. Some of the major issues relevant for GWAS follow. First, by comparing genes in the set versus those not in the set, for their associations with a trait, the null hypothesis focuses on "competition" between the set and its complement. Lack of enrichment does not mean that the genes in the set are not associated with the trait, but rather that they are associated, on average, as much as those not in the set. Second, when genes are correlated, the Type-I error rate can be inflated (related to whether the sample or the genes are the random units under the null hypothesis). Third, "size-biased" sampling haunts analyses in two ways; at the gene-level and at the set-level. At the gene-level, most statistical tests require a single p-value per gene, so when there are multiple p-values for a gene, a common strategy is to choose the smallest p-value to represent the gene [13]. This, however, results in a bias towards detecting associations with genes having more p-values (e.g., larger genes having more SNPs). The magnitude of this bias depends not only on the number of SNPs in a gene, but also on the correlations among the SNPs — weaker correlations allow more extreme statistics. This extreme p-value approach might also have weak power when a gene contains multiple SNPs that are jointly associated with a trait, yet individually have small effects. Size-bias at the set-level occurs when the number of genes varies over sets: sets that have a large number of genes can be biased if only the most extreme p-value among all genes in a set is used to represent the set. For further discussion about these issues, and references, see [7]. Fourth, most gene-set methods focus on one set at a time (comparing a set to its complement), and so the resulting p-values across multiple sets are correlated, making it difficult to accurately evaluate the Type-I error rates across all gene-set tests. Bonferroni correction based on the number of sets tested would be overly conservative.

Because of the above concerns, our strategy takes a "direct" approach to test the association of each gene-set with a trait; the null hypothesis is that the gene-set is not associated with a trait, not the competitive hypothesis discussed above. This seems more relevant for GWAS that are expected to have a much smaller number of statistical associations than gene expression studies that typically have many differentially expressed genes. Our approach allows SNPs to be mapped to multiple genes, and directly accounts for correlations among SNPs, both within and between genes. By doing so we are able to normalize genes, and gene-sets, to avoid size-biased sampling. This also avoids using arbitrary cut-offs to choose SNPs with low correlations, or to assign SNPs to haplotype blocks [21]. Furthermore, we consider all possible gene-sets, and provide a simulation-based method for the entire genome that is computationally efficient and accounts for the testing of many gene-sets. Our methods also allow for overlapping gene-sets, which can occur when genes map to multiple gene-sets, or when sets can be nested or overlap when grouping genes according to their function. These new methods resolve some of the unsolved analytic issues for GWAS gene-set analyses that have been highlighted elsewhere[22].

A variety of public resources are widely used to create gene-sets, such as Kyoto Encyclopedia of Genes and Genomes (KEGG [23]), BioCarta, The Reactome Project, NetPath, Pathway Interaction Databases, and the Gene Ontology (GO [24]). A complete list of over 300 resources for biological pathways can be found at the Pathguide website (http://www.pathguide.org). Because the GO has the largest amount of information and is well structured, we primarily use it to create gene-sets. However, our approach is sufficiently general that it could easily be adapted to gene-sets created from other sources of information, as we illustrate with KEGG, or even gene-sets created from multiple sources (e.g., GO + KEGG + others).

The GO provides structured information on the properties of the products of genes, using three general domains: cellular components (the parts of a cell or its extracellular environment), molecular function (the activities of gene products at the molecular level), and biological process (sets of molecular events with a defined beginning and end). For each of these domains, there are well-defined terms that describe gene product properties. For example, a single gene product might be described by a molecular function term that describes its enzymatic activity, a biological process term that describes a sequence of molecular activities, and a cellular component term that describes where the activities occur, such as a cell membrane. Furthermore, terms can be related to each other. The more specific terms are called child terms, and these are linked to more general terms, called parent terms. If every child term had only one parent term, the terms would form a hierarchy. But, because a child term can have more than one parent, the terms are related in a directed acyclic graph (DAG). Consistent with terminology for graphs, we refer to the GO terms in the DAG as vertices, and the connections between them as edges. The majority of edges are “is_a” relationships, meaning that a child is a subtype of its parents. Other terms are “part_of” (meaning a child is necessarily part of its parent, so the parent must exist when the child exists), “has_part” (meaning a parent necessarily has its child as a part, so the child must exist when the parent exists), and terms related to how a gene product regulates another gene product. All directions go from child to parent, except for “has_part”, which is directed from parent to child.

A number of methods have been developed that capitalize on the GO data. Holmans, like others, used a list of p-values from a GWAS [25]. By applying an arbitrary threshold to determine statistical significance, a gene was categorized as significant if it had at least one significant SNP: then, a count of the number of significant genes was used in gene-set analyses. The gene-sets were created by mapping genes to the GO vertices, so that each vertex is a gene-set. Additional sets were created by recursively building larger sets by uniting child vertices with parent vertices. The null distribution of the statistics was simulated by randomly sampling the SNP p-values from the list of all p-values. This process, however, assumes that the level of linkage disequilibrium among SNPs is approximately equal among all genes and among all GO categories, which seems unlikely. Analyses based on p-values are limited by the inability to appropriately account for the underlying correlation structure among SNPs, making it questionable whether Holman's approach adequately corrects for LD among SNPs, variable gene sizes, and overlapping genes.

Others have also capitalized on the GO DAG structure, but mainly to control the family-wise error rate (FWER) when testing gene-sets using a list of p-values as input. One approach is “bottom-up”, which starts with the most specific terms, the “child-vertices”, and applies a FWER control for these vertices. The procedure moves up to parent vertices only if the null hypothesis for the child vertex can be rejected. Another approach is “top-down”, which starts with the root vertices and moves down to child-vertices only if the root can be rejected. Meinhausen developed this procedure to control the FWER when hypotheses can be structured in a hierarchy, by starting at the highest level of a hierarchy, and continuing down to the next level of the hierarchy until no tests are significant [26]. To apply this to GO, the DAG would have to be transformed to a hierarchical tree structure (requiring duplication of terms so each child has only one parent). However, this approach would not be useful for GWAS, because the procedure would likely stop at the highest GO vertex that contains all genes. Another alternative is to start testing at a specific “focus level” of the GO graph, which determines the level of specificity of the GO terms used in the analyses [27]. This approach controls the FWER by a sequentially rejective multiple testing procedure that exploits the GO DAG structure. Starting at a specified level, the procedure sequentially moves away (up and down the DAG) until no gene-sets are statistically significant.

However, this approach does not account for the correlations among SNPs. Furthermore, for the typical agnostic GWAS, it is difficult to specify a specific focus level. Another approach developed for gene-expression studies is to model the joint distribution of p-values using a hidden Markov model after transforming the DAG into a hierarchical tree [28]. A limitation of these methods for GWAS data is that they are based on a list of p-values, and so they cannot account for the correlation structure for SNPs.

As we emphasize throughout, accounting for SNP correlations is impossible when only a list of p-values is available, and at best a crude approximation when external reference correlation structures are used to infer correlations in the GWAS data that generated the p-values. Because of this, we use the subject-level genotypes to directly account for SNP correlations. This allows us to avoid unrealistic assumptions or arbitrary thresholds, as well as develop more powerful methods for gene-level and gene-set analyses. While developing our methods, we recognize the balance between computer-intensive methods required to fit sophisticated statistical models and the need to be able to handle large amounts of data that can have complex data structures, as well as the need to compute simulation-based p-values. To achieve this balance, we focus on score statistics, which are rapid to compute and tend to be optimal for alternative hypotheses that are “local” to the null hypothesis, which in our context means small gene effects, which is expected for many GWAS.

The paper begins with a detailed description of our methods to score SNPs, genes, and gene-sets, along with how gene-sets are created from the GO structure. We then apply these new method to several GWAS, contrasting results from gene-set analyses with traditional single-SNP analyses. Simulations are used to explore and highlight some of the statistical properties of our proposed methods, although a thorough simulation that evaluates the power of our new methods across a broad spectrum of possible genetic architectures is beyond the scope of this initial presentation. We then discuss strengths and weaknesses of gene-set analyses, and areas for potential research directions.

Methods

Mapping SNPs, Genes, and GO Terms

To map SNPs to genes, and genes to the GO structures, we used publically available physical maps for both HapMap SNPs and known genes, as well as the GO data files. These files and the steps we used to process them are described in Appendix A. A few points are worth emphasizing. For a SNP to be included in the gene-set analyses, SNP genotypes for cases and controls must be available, the SNP must map to within a user-specified distance to known genes, and the gene(s) it maps to must map to the GO structure. To create gene-sets, we use a recursive method so that genes are included in a gene-set if the genes map to a specific GO term or any of its descendant GO terms. At the highest level, such as a single DAG root, the gene-set would be all genes that map to the entire GO structure. At the lowest level, such as a leaf, the gene-set would be only those genes that map to the leaf. See Appendix A for more details.

Scoring SNPs

We first describe how we score individual SNPs for their association with a trait, and then describe how we combine these scores into scores for genes, and then scores for gene-sets. To score the effects of SNPs on a trait, we use score statistics from regression models. To simplify our exposition, we focus on case-control data using logistic regression, but then discuss simple ways to use our methods for traits that can be analyzed with generalized linear models, such as quantitative traits using linear regression. Let y_i have values of 1 for cases and 0 for controls, and let x_{ij} represent the score for the j^{th} SNP of the i^{th} subject ($i =$

1,. . .*n*, ;*j*1=, . . . ,*m*). We let x_{ij} count the number of minor alleles for a SNP (x_{ij} = 0,1,2), although other genotype scoring could be used. With this set up, the log-likelihood for a logistic regression model with only the j^{th} SNP is

$$\ln \quad L = \sum_{i=1,\ldots,n} y_i \quad \log \quad p_i + (1 - y_i) \log (1 - p_i),$$

where

$$p_i = \frac{\exp\left(\alpha + \beta_j x_{ij}\right)}{1 + \exp\left(\alpha + \beta_j x_{ij}\right)}.$$

The score statistic for β_j can be found by taking the first partial derivative of ln *like* with respect to β_j, and then setting $\beta_j = 0$. The contribution of the i^{th} subject to this score is $U_{if} = (y_i - p_{oi})x_{if}$, where p_{oi} is the null probability $P(y_i = 1|\alpha, \beta_j = 0)$. As a side note, it is important to recognize that adjusting covariates could be included in the analyses by including them in the logistic regression model and using them, along with the maximum likelihood estimates of their regression coefficients, to compute p_{oi}. Then, the efficient score statistic to test whether $\beta_j = 0$ can be computed as

$$z_j = \frac{\sum U_{ij}}{\sqrt{\sum_i U_{ij}^2}}.$$

Note that the denominator of z_j is asymptotically equivalent to the square-root of the observed information for β_j, adjusted for estimating the intercept.

To apply our methods to other types of traits, such as quantitative, one simply needs to scale the U_{ij} scores by the appropriate dispersion parameter. To see this, note that the log-likelihood for a subject's trait y_i based on a generalized linear model (GLM) for exponential family data can be expressed as $\ln L = [y\eta - b(\eta)]/a(\phi) + c(y, \phi)$, where *a, b*, and *c* are known functions, ϕ is a dispersion parameter, and $\eta = Z'\beta$; Z denotes a vector of covariates to be adjusted out, and β the vector of corresponding regression coefficients. The expected value of the trait is $\tilde{y} = f^{-1}(Z'\beta)$, where f is the link function. Then, for a GLM, the corresponding score is $U_{if} = [(y_i - \tilde{y}i)/a(\phi)]x_{if}$. For binomial and Poisson data, $a(\phi) = 1$, giving the results we present above for logistic regression. For quantitative traits that have a normal distribution, $a(\phi) = \sigma^2_{mse}$, the mean-squared error for the residuals, $(y_i - \tilde{y}_i)$, after regressing out the covariates. This works because under the null hypothesis the expected value of $\sum_i U_{ij}^2$ is equal to Fisher's information, the variance of $\sum_i U_{ij}$.

Scoring Genes

We now need to consider how to combine the z_j scores for all SNPs in a single gene to create a gene-level score. Some guidance can be gleaned from the literature. Critically, Newton et al. [29] compared two different ways to score sets of genes for gene expression studies. One way was to consider the fraction of genes in a set that had p-values below a threshold. The other way was to average scores across genes in a set. Their simulations showed that power was greater for the averaging method when a set contains a large fraction of associated genes, each with small effects. In contrast, thresholding was more powerful

when there was a small fraction of genes with large effects. For GWAS, we expect many genes with small effects, so averaging scores seems appropriate. Furthermore, for gene-set analysis of expression data, others have found that squaring the statistics to allow for both positive and negative directions of the genetic effects and then averaging these squared terms gave greater power than other popular methods, including the enrichment scoring method of the GSEA [8]. They also found that squared statistics gave similar results as using absolute values of the statistics. Because of this, we too use averages of squared statistics to score genes. So, to create gene-level scores, we use the average of z_j^2 across all SNPs in a gene,

$$gene = \frac{1}{m}\sum_{j\in g} Z_j^2 = \frac{1}{m} Z_g' Z_g.$$

Here, there are m SNPs in gene g, and the vector Z_g contains the z_j scores for all SNPs in gene g. By using vector notation, we emphasize that the *gene* score is a quadratic form. It should be noted that this quadratic form is equivalent to a logistic kernel-machine test when using a linear kernel [30]. This opens opportunities to consider alternative ways to score genes, as discussed in Appendix B. Under the null hypothesis, it is well known that Z_g has an asymptotic multivariate normal distribution with mean vector 0 and correlation (and covariance) matrix R. This correlation matrix is based on the observed information matrix, $V = \sum_i p_{oi}(1 - p_{oi}) X_i X_i'$, where the first element of X_i is the intercept and the remaining elements are the doses of the minor alleles for the m SNP genotypes. After accounting for estimating α in the score statistic and assuming that there are no adjusting covariates (p_{oi} is the same for all subjects), it can be shown that R is equal to the correlation matrix for the x_{ij}'s.

Now, because $Z_g \sim MVN(0, R)$, the moments of the quadratic form are known to be $E\left[Z_g' Z_g\right] = tr(R) = m$ and $Var\left(Z_g' Z_g\right) = 2tr(RR) = 2\sum_i \sum_j r_{ij}^2$, where $tr(A)$ means trace of matrix A[31]. This implies that the expected value of a gene score is $E[gene] = 1$, and its variance is $Var(gene) = 2tr(RR)/m^2$. It is important to recognize that this allows us to compute the variance of the gene-scores using all the SNPs within a gene while accounting for the correlations among them. To illustrate, when there is no LD among SNPs within a gene, $Var(gene) = 2/m$. At the other extreme, if all SNPs within a gene were in complete LD, so that $r_{ij}^2 = 1$, then $Var(gene) = 2$. As an alternative to averaging z_j^2, we considered averaging the absolute value, $|z_j|$, but avoided this because the resulting truncated multivariate normal distribution would require numerical evaluation of multivariate integrals [32].

Scoring Gene-Sets

To score a gene-set, we use the weighted average of the gene-scores within a set,

$$set = \sum_{j=1,\ldots G} w_j gene_j,$$

where $w_j = 1/Var(gene_j)$. Under the null, $E[gene_j] = 1$, so $E[set] = 1$. The variance of the set-score is

$$Var(set) = \frac{1}{\left(\sum_i w_i\right)^2} \sum_i \sum_j w_i w_j Cov\left(gene_i, gene_j\right).$$

Like the methods used to derive $Var(gene)$, the covariance between two *gene* scores can be shown to be

$$Cov\left(gene_i, gene_j\right) = \frac{2tr\left(R_{ij} R_{ij}'\right)}{m_i m_j}.$$

This covariance depends on the matrix of cross-correlations (R_{if}) of SNPs for genes i and j, and the number of SNPs for gene i (m_i). and gene j(m_j). These derivations allow us to measure the correlation structure among genes in a set so we can normalize the scores for gene-sets by subtracting the null expectation and dividing by the standard error of the set-score,

$$z_{set} = \frac{set - E[set]}{\sqrt{Var(set)}}.$$

This approach directly accounts for the correlation among SNPs, eliminating the need to specify arbitrary correlation thresholds for including or excluding SNPs, as well accounts for the number of SNPs per gene and the number of genes per set. This approach also accounts for situations when a SNP maps to multiple genes within the same set, a likely occurrence when a gene has alternative splicing.

After applying our methods to real data and evaluating its properties by simulations, we found a few transformations that aided the distributional properties of z_{set}. First, the scores from logistic regression, $U_{if} = (y_i - p_{oi})x_{if}$, are based on binomial y_i and trinomial x_{ij}, and so asymptotic normality of the SNP scores might not hold if n is not very large. Centering the x_{ij}'s about their sample means helped to prevent the mean of the set scores to depend on the set size. Second, because we are using quadratic statistics (i.e.., $Z_g' Z_g$) to score genes, sets with few genes tended to have positively skewed values of z_{set}. To reduce this skewness, we used the square-root transformation for the gene scores, $gene^* = \sqrt{gene} = \frac{1}{\sqrt{m}} \sqrt{Z_g' Z_g}$. Using a 2nd order Taylor-series approximation, the null expected value of this square-root transformation is $E[gene^*] \approx 1 - Var(gene)/8$ and its variance is $Var(gene^*) \approx Var(gene)/4$. Likewise, the covariance between two transformed gene-scores is $Cov\left(gene_i^*, gene_j^*\right) \approx Cov\left(gene_i, gene_j\right)/4$. Hence, the final set-scores are based on the weighted average of the transformed $gene^*$ scores, along with $E[gene^*]$ and $Cov\left(gene_i^*, gene_j^*\right)$ to create z_{set} based on the square-root transformation.

When applying our methods, a critical issue is to decide on the maximum number of genes in a set. Without any restrictions, the largest set will contain all genes which might not be sensible. One approach would be to limit the depth of the GO DAG structure. However, this could still result in sets containing many genes, potentially washing out the association signal by many random non-associated genes. For this reason, we allow the user to specify the maximum number of genes in a set so that the scan for gene-set associations would only

cover sets that meet this constraint. For the applications we illustrate, we set the maximum set-size to be 30. Avoiding large sets not only avoids diluting out signals of association from many non-associated genes, but it also speeds the computations by reducing the number of covariance terms to compute for genes in the same set.

Controlling for Population Structure or Other Covariates

As many have emphasized, it is important to correct for systematic genomic inflation of test statistics [22], such as by using the eigenvectors of the independent SNPs from a GWA study panel as adjusting covariates in logistic regression models [33; 34]. A way to perform this adjustment for our gene-set scanning approach is to use adjusted U_{ij} scores

$$U_{ij} = \left[y_i - p_i\left(\widehat{\beta}, e_i\right)\right] r_{ij},$$

where $p_i\left(\widehat{\beta}, e_i\right)$ is the logistic regression predicted status based on the eigenvectors for the i^{th} subject and the maximum likelihood estimates, $\widehat{\beta}$, of the eigenvector regression coefficients, and r_{ij} is the residual from linear regression of the j^{th} SNP on the eigenvectors. As discussed elsewhere [35], this method works because the residuals are uncorrelated with the eigenvectors and testing the residuals is equivalent to testing the allele dosage score adjusted for the eigenvectors.

Missing Data

To account for missing genotypes, it is most informative to impute them using powerful methods and then use the allele dosage scores for the missing genotypes [36]. Absent this imputation, one could set the U_{ij} score to 0 for missing genotypes, which is equivalent to replacing missing x_{ij} values with their sample means.

Computing p-values

To determine the null distribution of the set-score statistics, z_{set}, we could permute the trait over all subjects, preserving the correlation structure among the SNPs, and then recompute all statistics. This, however, is computationally intensive, particularly due to the need to compute $Cov(gene_i, gene_j)$. Lin[37] proposed a more efficient method, capitalizing on the multivariate normal distribution of the score-statistics from logistic regression. Suppose that we place the U_{ij} scores for all SNPs into a vector for the i^{th} subject, U_i. If n were very large, $\sum_i U_i \sim MVN(\mu, \Sigma)$. We could approximate the null distribution by simply multiplying the observed vector U_i by a standard normal random variable (scalar), $s_i \sim N(0,1)$, to create $U_i^* = s_i U_i$. Conditional on the observed data, this new vector also leads to a multivariate normal statistic, $\sum_i U_i^* \sim MVN(0, \Sigma)$, with mean 0 and the same covariance matrix as that for the observed statistic, $\sum_i U_i$. Because we standardized U scores by dividing by the square root of their variances, we work with the correlation matrix, R, for the x allele doses. This means that we only need to compute the R matrix for the observed data once. Furthermore, the entire R matrix is not used — only sub-matrices that correspond to genes in the same set. Hence, to further speed calculations, we only compute $Cov(gene_i, gene_j)$ at the first time it is needed, and then store it for future use. This allows us to avoid computing $Cov(gene_i, gene_j)$ for genes that are never in the same set, dramatically reducing the computational burden. For the simulations, we only need to create $U_i^* = s_i U_i$ for all n subjects in each simulation loop, and then use this U_i^* for all subsequent calculations to create scores for genes, and scores for gene-sets, but use the once-calculated terms for $Cov(gene_i, gene_j)$.

Another benefit of Lin's approach is a simulation-based step-down procedure to compute adjusted p-values that control the FWER. This step-down procedure first orders the observed statistics (e.g., gene-sets) from largest to smallest, denoted $z_{(1)},z_{(2)},...,z_{(m)}$. The corresponding hypotheses are labeled $H_{(1)}, H_{(2)},...,H_{(m)}$. Starting with $H_{(1)}$, it is rejected if $\Pr\left(\max_{1\le k\le m} s_k \ge z_{(1)}\right) \le \alpha$, where s_k is the simulated statistic and α is the Type-I error rate. If $H_{(1)}$ rejects, $H_{(2)}$ is tested, and it is rejected if $\Pr\left(\max_{2\le k\le m} s_k \ge z_{(2)}\right) \le \alpha$, and so forth. This means that $H_{(j)}$ is tested only if all prior hypotheses $H_{(1)}, H_{(2)},...,H_{(j-1)}$ are rejected. By noting that the p-value is the α - level test at which the null hypothesis would be rejected, we can use this step-down approach so the p-values for the ordered statistics, $z_{(1)},z_{(2)},...,z_{(m)}$, are monotonically increasing, and the FWER is controlled. Two important points are: 1) the p-value for $z(1)$ is simply how frequent the maximum simulated statistic is at least as large at the maximum observed statistic, as one would desire for the most extreme statistic, and 2) this approach accounts for the correlations among the statistics, avoiding the conservative approximation when using the Bonferroni correction.

Application to GWAS Data

1) GWAS for Musculoskeletal Adverse Events from Breast Cancer Treatment—Aromatase inhibitors are well established adjuvant therapies for postmenopausal women with early-stage breast cancer, yet about half of the patients receiving these types of treatments have joint-related musculoskeletal complaints, which likely contributes to decreased compliance. A GWAS was performed to identify SNPs associated with musculoskeletal adverse events (MSAEs) in women treated with aromatase inhibitors for early breast cancer [38]. A nested case-control design was used to select patients enrolled onto the MA.27 phase III trial that compared the aromatase inhibitors anastrozole and exemestane. Cases (n=293) were defined as patients with MSAEs. Controls (n=585) did not experience any of the MSAEs, were followed for at least two years, and had at least six months longer follow-up than their matched case. Genotyping was performed with the Illumina Human610- Quad BeadChip. After quality control checks, a total of 551,358 SNPs were analyzed. The smallest P-values (< 1E-4) occurred for 14 SNPs that were on chromosomes 1-4, 8, 9, 14, 15, 22, and X. The smallest P-values were for 4 SNPs on chromosome 14 (P-values 2.23E-06 to 6.67E-07). The gene closest to these 4 SNPs is the T-cell leukemia 1A (TCL1A), 926-7000 bp distance from the SNPs. Further details of the study design and main results can be found in the primary publication [38]. The analyses presented here include only the SNPs measured on the Illumina platform. Imputed SNPs were not included.

2) GWAS for Breast Cancer Risk among High-Risk Women in Breast Cancer Prevention Trials—The selective estrogen receptor modulators (SERM) tamoxifen and raloxifene have been evaluated for their ability to prevent breast cancer in high-risk women in large randomized clinical trials [39; 40]. To further reduce the risk of breast cancer, indentifying SNPs and genes associated with breast cancer occurrence in these women would provide important leads to better understand why some women develop breast cancer despite SERM therapy, and might even lead to better selection of patients for preventive therapy. To achieve this aim, cases (n=594) and controls (n=1,172) were selected from the tamoxifen arm in the NSABP prevention trial P-1 and from the tamoxifen and raloxifene arms in the NSABP prevention trial P-2. Cases were women who experienced an invasive breast cancer or ductal carcinoma in-situ (DCIS); controls had neither. Genotyping was performed with the Illumina Human610- Quad BeadChip. After quality control checks, a total of 547,356 SNPs were analyzed. The smallest P-values (< 1E-4) occurred for 11 SNPs that were on chromosomes 3, 4, 8, 9, 13, and 16. The smallest P-value was for a SNP on

chromosome 16 (P-value 2.12E-06), near the ZNF423 zinc fingerprint gene. Further details of the study design and main results can be found in the primary publication [41].

Results

Results from MA.27 Analyses

Results from gene-set analyses for the MA.27 study are presented in Table 1. Although no gene-sets achieved statistical significance at the global level, it is intriguing that the most extreme gene-set statistic occurred for the GO cellular component term relating to actinomysin, the actin part of any complex of actin, myosin, and accessory proteins. The single gene that maps to this GO term is ACTC1 actin, a highly conserved protein found in muscle tissues, making it a suspect for musculoskeletal adverse events in the MA.27 study. The second most extreme gene-set for cellular component relates to the acrosome, an organelle in sperm cells. Although this is non-sensible for musculoskeletal adverse events in women, one of the three genes in this gene-set codes the L protein of the mitochondrial glycine cleavage system. Other gene-sets related to mitochondrial cellular components and biological processes had extreme gene-set statistics, suggesting that mitochondria could be related to musculoskeletal adverse events.

Gene-level analyses for MA.27 are presented in Table 2. The most extreme z-score was for a microRNA, which are short non-coding RNAs involved in post-translational regulation of gene expression. The second most extreme statistic was for a hypothetical protein (predicted gene with unknown function). It is important to see that the TCL1B gene identified by single-SNP analyses had the 3rd largest gene-level statistic. This gene is near the TCL1A gene, which was near the most significant single-SNP test. In fact, the gene-level statistics for these two gene were quite similar (z-scores of 4.313 for TCL1B and 3.897 for TCL1A), and each had approximately 40 SNPs mapping to the genes. But, based on p-values, TCL1A ranked 25 from the top. This is partly because of the 40 SNPs mapping to TCL1A, some have weak associations and others very strong, resulting in a reduced average gene-level score. It is possible that the SNPs in this region, found by the strongest single-SNP tests, have functional roles for both genes, or are in LD with functional SNPs.

To contrast the single-SNP analyses with the gene-set analyses, we examined how the most striking single SNP result contributed to the gene-set results. The SNP rs7158782 near the TCL1A gene had a p-value of 6.67E-07. This gene maps to the GO term “stem cell maintenance” (ID 0019827), within the biological process domain. There were 23 other genes in the gene-set for this GO term, with a gene-set z-statistic = -0.3, suggesting no evidence of the genes in this set to be associated with MSAE's. This illustrates a potential limitation of gene-set analyses when only a single gene in a large set is associated with a trait.

We also considered why the actin gene was not revealed in the single-SNP analyses. There were 43 SNPs in the actin gene region with single-SNP p-values ranging 0.00034 to 0.9728. Only 7 SNPs had p-values < 0.02; their distance spanned 24 kb, with 4 other intervening SNPs having p-values ranging 0.26 to 0.86. Based on single-SNP analyses, these SNPs individually were not near the top of the list of interesting SNPs. This illustrates that single-SNP analyses might miss important statistical associations when their individual signals are small, and by combining over a gene, or a gene-set, there is potential to strengthen the statistical association signal.

Results from NSABP Analyses

Results from gene-set analyses for the NSABP study are presented in Table 3. None of the gene-set analyses achieved global statistical significance, and the most extreme gene-set z-

score was for biological process related to activation of Janus kinase activity, a process which activates the JAK protein by introducing a phosphate group to a tyrosine residue of a JAK protein. The two genes in this gene-set are IL6R interleukin 6 receptor, a cytokine that regulates cell growth and differentiation and plays an important role in immune response, and GH1 growth hormone, which plays an important role in growth control. The second most extreme z-statistic for biological process was for the GO term virus-host interaction, which included 21 genes. Many of these genes are involved in immune response (e.g. THO complex) or regulation of gene expression or growth hormones. This set also includes SMAD3, a signal transducer and transcriptional modulator that mediates signaling pathways. In particular, this protein is a transcriptional modulator activated by transforming growth factor-beta and is thought to play a role in the regulation of carcinogenesis. This suggests that immune-related genes and those related to growth factor regulation would be worth while to pursue as candidates for associations with breast cancer occurrence among women treated with tamoxifen or raloxifene.

Gene-level analyses for NSABP are presented in Table 4. The gene with the most extreme statistic relates to a mitochondrial ribosomal protein, but currently interpreted as a pseudo-gene. The single-SNP analyses led to the ZNF423 gene, which has the second largest z-score for the gene-level analyses.

KEGG Set Analyses

In addition to the recursively defined GO gene-sets, we used the KEGG pathways to define gene-sets. Here, there were 221 gene-sets, with the number of genes per set ranging 1-1088, and an average of 70 genes per set. Unlike the GO-set analyses, we did not restrict the number of genes per set for the KEGG sets, primarily because there is no DAG structure among the KEGG sets. Furthermore, with many fewer gene-sets, the time to compute the analyses was reasonable, despite the larger number of genes per set. Result from both the MA.27 and NSABP studies are presented in Table 5. No KEGG sets were near statistical significance, and the top-ranking sets do not suggest reasonable biological hypotheses for these sets.

The results presented above for both MA.27 and NSABP were not adjusted for population stratification by regression-based eigenvector adjustment. These adjusted analyses were run, but they did not differ substantially from those presented, so conclusions would not change if eigenvectors were used for adjustment.

Evaluation of Type-I Error Rate

To evaluate the Type-I error rate for the GO scanning method, we used the NSABP data for bootstrap sampling. For each of 1,000 simulations, we sampled with replacement 1,766 random subjects. We then randomly assigned case-control status (594 cases and 1,172 controls, as for the observed data) and computed the scores for all observed SNPs. The SNP scores were analyzed by gene-sets for the GO biological process domain. The Type-I error rate for the smallest p-value was 0.042, which illustrates that the global FWER is well controlled at the nominal level of 0.05. The type-I error rates for the next 4 smallest p-values were 0.016, 0.010, 0.009, 0.009. These monotonically decreasing error rates result from the way the step-down p-values were created.

Software

Our software, called *gene_set_scan*, can perform gene-level analyses, gene-set analyses determined by the GO structure, or gene-set analyses determined by KEGG pathways. Although a number of input files are required, only one file must be created by a user. This file contains score statistics for each SNP and each subject. The other supplied files describe

how SNPs map to genes, and how genes map to gene-sets (determined by GO data or KEGG data). Types of analyses are controlled by input options. The output from *gene_set_scan* is structured as HTML files, which provide statistical results with links to web-based annotation (e.g., NCBI GENE information, Amigo for GO terms, or Genome in Japan for KEGG terms). The output is also provided as comma-delimited, for input to other software (e.g., R statistical software for plotting or additional analyses).

The *gene_set_scan* software is written in C++, and uses the BOOST graphical library for computations on the GO DAG structure. All computations were accomplished on a Beowulf-style Linux cluster. This cluster has a master node that is a dual Athlon 2800+ MP with 2 GB memory. The 20 compute nodes are a mixture of Athlon 2400+ XP CPU with 1 GB memory (PC133 or DDR). The approximate times to compute the analyses are given in Table 6. A key factor that determines the time for computations is the number of genes in a set, due to the calculation of covariance terms among gene-scores within the same set. The GO analyses were restricted to no more than 30 genes in a set. The longer compute times for KEGG analyses were because of the large sizes of some gene-sets, with an average set size of 70 genes. The GO analyses would take longer if the maximum set size were not restricted to 30 genes per set. For examples, when the maximum set size was allowed to be as large as 1,000, the compute times were 26h:36m for NSABP and 14h:50m for MA.27, still quite achievable.

Discussion

We developed a general approach to scan GWAS SNP data for gene-set analyses, primarily based on score statistics for individual SNPs. Although our presentation was for a binary trait modeled with logistic regression, our approach can be used for any type of trait that can be modeled by a generalized linear model. This strategy allows adjustment for covariates, and is computationally efficient. It also capitalizes on the well known statistical advantages of score statistics, such as most power for "local" alternative hypotheses — relatively small SNP effects on a trait. When creating gene-scores, we averaged the square of SNP-level *z*-scores, so that positive and negative effects of SNPs on a trait would not cancel. This allowed us to then combine gene-level scores into a gene-set score, using known properties of quadratic forms, allowing us to directly account for the covariances among gene-scores within a gene-set. A further computational advantage of our method is use of Lin's approach to compute p-values, based on asymptotic multivariate normal distribution of score statistics, and step-down adjusted p-values.

Some advantages of our proposed quadratic statistics to score genes can be gleaned from both theoretical and simulation studies. By moving from the assumption that the vector β of regression coefficients for p SNPs in a gene is a fixed effect to the assumption of a random effect, as in a mixed model, the null hypothesis $H_o: \beta = 0$ can be formulated as testing a variance component, $H_o: \nu = 0$, where the covariance matrix of β is νK, with K a known matrix. Goeman et al. [42] show that the quadratic statistic is the score statistic for testing $H_o: \nu = 0$, when K is the identity matrix. This quadratic statistic has optimal power for alternative hypotheses that are "local" to the null hypothesis. They further showed that the quadratic statistic has greater power than an F-test for linear regression of a trait on all covariates (e.g., all SNPs in a gene) whenever there are a limited number of latent variables, such that the latent variables are captured by a few of the large variance principal components of the SNP correlation matrix. This would occur, for example, when there are multiple tag-SNPs for a gene, but much fewer unmeasured causal variants. One could apply principal components to the SNPs in a gene to choose a few principal components to be used in a regression model F-test, as many have suggested [43-45]. But, as Goeman et al. emphasize, this is not necessary when using the locally most powerful quadratic statistic that already has

greater power than the F-test. Another surprising result by Goeman et al. is that the quadratic statistic has greater power than using the most extreme statistic (e.g., smallest p-value) even when there is only a single non-zero regression coefficient. Once again, these conclusions hinge on the correlation of the tag-SNPs with the latent casual variants and the causal variants having small effects on the trait. As we emphasized in the development of our methods, and Appendix B, this mixed model formulation provides a powerful framework to develop other types of gene-scores. This is based on the connection between generalized linear mixed models and kernel-machine learning [30], primarily driven by choice of the kernel matrix.

A number of simulation-based studies compared different multi-marker statistics to detect associations between a gene and disease [46-49]. These include Fisher's method for combining P-values, the minimum P-value of all SNPs in a gene, a Fourier-transform-based approach [50], regression on principal components, and the quadratic score statistic of Goeman et al. [42]. The power advantage of the competing methods depends on the underlying genetic mechanisms, but the general conclusions were that the variance components score statistic and principal components methods provided similar power, often greater than competing methods. The greatest power occurred when there were multiple tag-SNPs associated with disease, and when the minor allele frequencies of the tag-SNPs were not too small (eg., MAF > 10%).

There is some similarity in our gene-scoring approach with that proposed by Luo et al, [51], who also used linear combinations of SNP scores to score a gene, and used the SNP correlation matrix to account for their correlations. A critical distinction, however, is that we used a linear combination of the squared statistics, so that positive and negative associations would not cancel. Luo et al. were vague about whether their methods of converting p-values to normal distribution z-scores assumed 1-sided or 2-sided p-values. It is clear that their approach works for signed statistics (1-sided), but it is not valid for when using absolute values of the statistics (2-sided), because in this case, the moments of a truncated multivariate normal would be required [32]. Our gene-scoring approach is also similar in spirit to that implemented in the VEGAS software, except that we use the GWAS data to estimate the correlation of SNPs, in contrast to VEGAS that uses a reference population to estimate the correlation structure, and then uses this correlation matrix to simulate from a multivariate normal distribution to determine the null distribution of gene-scores [52].

Although our approach requires genotype data used for GWAS analyses, it overcomes the size-bias problem when using the smallest p-value among all SNPs in a gene. Many reports on gene-set analyses recognize the potential for gene size and SNP correlation structure to bias the scoring of genes by extreme p-values, yet few have found adequate solutions. A clever approach proposed by Segre et al. [53] is to convert the extreme p-value to a z-score, and then attempt to adjust for gene characteristics correlated with the z-score. Using linear regression, they found the following gene characteristics to be associated with extreme p-values: gene size (in kilobases, kb), number of SNPs per kb, number of independent SNPs per kb, number of recombinant hot-spots per kb, and linkage disequilibrium units per kb. These characteristics, however, were based on reference data, not the GWAS data that generated the p-values. All of these factors are likely correlated among themselves, and are merely indirect reflections of the underlying correlation structure of the SNPs in a gene. In contrast, our approach uses the observed correlation structure of the GWAS data to directly account for the SNP correlation structure, both within and between genes.

We defined recursive gene-sets that take advantage of the DAG structure of GO. This can be advantageous if in fact the associations of genes with a trait are well represented by the DAG structure. A strength of GO is the large amount of information, with over 16 million

annotations. Yet, a limitation is that over 95% of the annotations are computationally derived, not experimentally validated [54]. This is achieved by taking experimentally derived biological knowledge from a limited number of model organisms to infer knowledge about similar gene products in other organisms using a phylogenetic framework [55]. As of November 1, 2010, the GO web pages show that 50% of the 198,207 annotations for humans have been electronically inferred. Hence, errors are likely to exist about how genes map to the GO, and the DAG structure of the GO. For the GWAS of musculoskeletal adverse events, the strongest single-SNP test was near the TCL1A gene. Further functional studies in cell lines revealed a totally new and unexpected function for TCL1A— estrogen dependent regulation of cytokine expression. This illustrates the limitations of databases if our prior knowledge is incomplete.

One way to begin modeling incomplete prior knowledge is by using ideas from fuzzy-set theory [56], whereby errors and ambiguity could be addressed in the analyses by weighting GO terms according to the accuracy of their information or according to the probability of function (either at the gene level or the GO term level). Alternatively, the DAG structure could be enhanced by using edge weights, to differentially weight relationships when creating gene-sets. However, at this point in time, it is difficult to develop a biologically sensible weighting scheme. Perhaps clever use of gene expression and proteomic experiments to model the DAG would help, but this is beyond the scope of this current work.

As discussed by Holmans [22], there are multiple sources for pathway information to create gene-sets, and unresolved issues relate to which pathways to use (quality vs. quantity of information). One approach is to try different pathway resources, but it can be difficult to decide which gives the most reasonable results in light of the large number of genes and gene-sets. Another might be to use multiple sources of information to create larger numbers of gene-sets, recognizing that genes are likely to overlap for the more well-established sets. Our approach would allow for these types of complex sets, accounting for the within-set correlations of genes by our analytic derivations, while allowing for between-set correlations in the way that we compute p-values (i.e., simulation-based p-values allow for between-set correlations that are not conservative like Bonferroni-adjusted p-values).

Application of our methods to two different GWAS provided guidance on the potential strengths and weaknesses of our gene-set analyses. Although no gene-set results were close to statistical significance, hints of genes and gene-sets worthy to evaluate in functional studies were suggested. Whether there are true effects of gene-sets in either of these two data sets is unknown, and so simulations would be useful to evaluate the power of our approach. However, empirical power studies would be best when comparing the relative power of a number of competing analysis strategies across a variety of assumed genetic effects. This is beyond the scope of this initial work, but we outline our future directions. First, we described ways to score genes and score gene-sets, based on averaging, centering about the null mean, and then scaling by the null standard error. If we replace averaging with a general function, but still center about the null mean of the general function, and scale by the standard error of the general function, we can extend our methods yet still account for size biases. We no longer have multivariate normal theory to guide us on how to compute the null mean and standard error, so we would need to use simulations to compute them. This approach could allow a general function to incorporate thresholding of SNPs based on their SNP-level z-scores, effectively filtering out a large number of random noise variables. Whether this approach is more powerful than the averaging approach we described in this paper will require simulations. In fact, a wide range of general scoring functions could be considered, requiring a wide range of generic scenarios to evaluate which scoring functions are best for different genetic mechanisms.

Besides considering different functions to score genes and gene-sets, an alternative strategy is to reduce the number of statistical tests to improve power. One strategy, similar to Goeman's focus-level, is to scan all sets defined by the GO DAG to find the most extreme statistic. If that set is statistically significant, then look at the statistics of its neighbors (parents and children). If any of those are significant, then look at their neighbors, and so forth. This approach reduces the multiple testing burden, but at the price of reduced power if gene sets unrelated to each other contain causal genes. A second strategy would be to begin scanning only the gene-sets defined by the leaves of the GO DAG; if any are significant, then scan their parents, and so forth. This bottom-up approach might be more powerful when the causal genes map to the most specific GO terms, otherwise it could miss gene-sets that are at higher levels in the DAG.

Finally, our applications illustrated that when a single gene with strong associations with a trait is grouped in a set with many other null genes, the signal for the strong gene can be lost. This is a consequence of using score statistics, all computed under the null hypothesis of no associations. An alternative approach would be use of different kernels to score genes (see Appendix B). Another viable strategy is to use penalized models to select the strongest genes and gene-sets. We are actively developing such models for the DAG structure of the GO, and other gene-set structures, using lasso-type penalized generalized linear models. These penalized models will be compared with our score-statistic methods presented in this paper, in terms of their practical utility in large GWAS data, and their relative power evaluated with simulations.

Acknowledgments

This research was supported by the U.S. Public Health Service, National Institutes of Health, contract grant number GM065450.

Acknowledgments

This research was supported by the U.S. Public Health Service, National Institutes of Health (NIH), contract grant number GM065450 (DJS, JPS). The NSABP trials were supported by NIH contract grant numbers U10-CA-37377 and U10-CA-69974. The MA.27 trail was supported by NIH contract grant numbers U01GM61388, U01GM63173, P50CA116201, U10CA77202, and the National Cancer Institute of Canada contract grant number CCS 015469.The administrative coordination and data analyses were supported by NIH contract number U19 GM6138 (Mayo PGRN). The genotyping was supported the Biobank Japan Project funded by the Ministry of Education, Culture, Sports, Science and Technology, Japan. Genotyping was performed at the RIKEN Center for Genomic Medicine, Japan.

Appendix A. Mapping SNPs, Genes, and Gene-Sets

To map SNPs to genes, and genes to the Gene Ontology (GO), we used publically available maps for both SNPs and known genes, and the GO data files. These files and steps are described below.

Converting SNP maps from NCBI human genome build 36.3 to build 37.1

Because the HapMap SNP maps are based on NCBI human genome build 36.3, but the gene maps are based on build 37.1, we needed to first convert the HapMap SNP map to build 37.1. For build 36.3, there are 4,098,136 SNPs in HapMap, which includes most, if not all, SNPs on panels that are commercially available. To convert HapMap from build 36.3 to build 37.1 positions, we wrote a perl script, *updateMapPos.pl*, that updates the positions in the PLINK formatted SNP map file using SNP positions for build 37.1 from the NCBI site. The files needed for this step are publicly available at these sites:

PLINK file with build 36.3 positions

http://pngu.mgh.harvard.edu/~purcell/plink/res.shtml#hapmap

hapmap_r23a.bim is within hapmap_r23a.zip

NCBI file with build 37.1 positions

ftp://ftp.ncbi.nih.gov/snp/organisms/human_9606/database/organism_data/b131_SNPChrPosOnRef_37_1.bcp.gz

Of the HapMap SNPs, a total of 11,949 SNPs were dropped when converting from build 36.3 to build 37.1, leaving 4,086,187 SNPs. The counts and reasons for the excluded SNPs are given in Table A1.

Mapping SNPs to Genes

To map the HapMap SNPs to genes, we used the updated HapMap genome build 37.1 map file described above, and the "Seq_Gene" file from NCBI for gene start/stop positions, available at the site ftp://ftp.ncbi.nih.gov/genomes/MapView/Homo_sapiens/sequence/BUILD.37.1/updates/seq_gene.md.gz. Our Perl script, *snp2gene.pl*, creates a text file with columns Gene_ID, rs_ID, and position, where position is the position of the SNP relative to the gene. Position has the value zero if the SNP is within the gene's start/stop position. Otherwise, position is an integer for the number of base pairs outside of the gene, with a negative value if the SNP occurs before the gene start position, and a positive value if the SNP occurs after the gene stop position. Our default base pair maximum absolute distance for SNPs outside a gene is 50,000 bp. Our software, however, allows the user to choose shorter distances for analyses. If distances larger than 50,000 bp are required, one would need to set the maximum distance in the script *snp2gene.pl*, and rerun it to create a new file.

To process the "Seq_Gene" file, we used the following steps.

1. Only match SNPs to entries in Seq_Gene labeled as "GENE" in the "Primary Assembly" group.
2. Some genes (e.g. GeneID:100126533) are on multiple chromosomes. Our Perl script includes all of them by creating multiple lines in the output file, one for each chromosome. This means that a gene can be paired with different SNPs that occur on different chromosomes.
3. The "X" chromosome in Seq_Gene file is labeled "23" in the SNP map file, and chromosome "Y" in seq_gene is labeled "24" in the SNP map file; our script accounts for this.
4. Some genes are not localized to a particular position on a chromosome, indicated with a different format in Seq_Gene — our script excludes these genes.

Mapping Genes to GO Terms

To map genes to terms in the GO, we use the file gene2go, available from the NCBI ftp site: ftp://ftp.ncbi.nih.gov/gene/DATA/gene2go.gz. The gene2go file contains human and other organisms, so we subset to humans, based on taxonomy identifier for humans, "9606".

Creating and Processing the GO DAG Structure File

To create the directed acyclic graph (DAG) structure to represent the GO terms (vertices) and their connecting edges, we use the GO OBO format file available from http://www.geneontology.org/ontology/obo_format_1_2/gene_ontology_ext.obo.

We wrote a Perl script, *go2edges.pl*, that parses the OBO file to create a comma-separated file that contains the three GO namespaces (Biological Process, Molecular Function, and Cellular Component), a child-term, an edge relationship, and a parent-term, where the parent-term is more general than the child-term. The output file format is namespace, child-term, relationship, parent-term.

The edge relationship is one of the following: “is_a”, “part_of”, “has_part”, “regulates”, “negatively_regulates” or “positively_regulates”. This file is read into our software and processed as follows. The edge relationships point from child to parent, except for the “has_part” relationship, which points from parent to child, so the child and parent are swapped for this type of relationship. We used the Boost Graph Library, BGL, [57] to create a DAG from an input file. For this, direction goes from left vertex to right vertex. So, when reading the above file format, the child-term is read as left vertex and parent-term as right vertex. This would make the direction to go from child to parent. However, we prefer to have directions to go from parent to child, so our software puts the parent-term as the left vertex and the child-term as the right vertex. By reversing direction, we consider the most general GO terms as the “roots” and the most specific GO terms as the “leaves”. This allows us to use the depth-first-search algorithm of the BGL to traverse the DAG structure, flagging leaf vertices when first visited, and then recursively flagging parent vertices after their child vertices are flagged. This flagging operation allows us to determine the order to visit vertices of the DAG so that we can then create gene-sets in a recursive manner. That is, a recursive gene-set for a specific GO term is the set of genes that map to the specific GO term, and all the genes that map to GO terms that descend from the specific GO term. At the highest level, say a single DAG root, the gene-set would be all genes that map to the entire GO structure. At the lowest level, say a leaf, the gene-set would be only those genes that map to the leaf.

At this point, we have not included weights for the edge relationships. Including weights would be simple, by adding weights to the input records, and making use of the edge weights in the BGL algorithms. However, at this point it is not clear how best to specify weights, so this feature will be a topic of future research.

Merging Observed SNP Scores with Above Mappings

For a SNP to be used in analyses, it must exist in the SNP-score file, and also exist in the gene-SNP map file described above. For a gene to be used in analyses, it must exist in the gene-SNP file, and also exist in the GO-gene mapping file. Hence, for a SNP to be included in a gene-set analysis, the SNP must exist in the SNP-score file, it must map to a gene, and the gene it maps to must map to the GO structure.

Appendix B. Scoring Genes by Kernel Matrices

Based on work by Lin [58], Wu et al. [30] derived powerful variance component score tests for binary traits, called logistic kernel-machine tests, that have the quadratic form $Q = (y - \hat{p}_o)' K(y - \hat{p}_o)$, where y is a vector of length n for binary indicators of disease status, $\hat{p}_o$, also a vector of length n, contains the logistic regression fitted values based on an intercept, and possibly adjusting covariates, and the n x n symmetric matrix K is a positive semidefinite kernel matrix. If the n x m matrix X contains the SNP allele dosage scores, then the linear

kernel is $K = XX'$. Substituting this linear kernel into the expression for Q, reduces to $Q = U'U$, where U is a vector containing the U-scores. If the U-scores are standardized by their standard errors, $Q = Z'Z$, which is the quadratic form we use to score genes in the main text. Wu et al. evaluated the power of this linear kernel, along with a radial-basis Gaussian kernel and an identity-by-state (IBS) kernel. Although the linear kernel had good power, they found that the IBS kernel had greater power when there was interaction among SNPs, on the logit scale. This suggests that a variety of kernels might be considered when scoring genes. Here, we illustrate the form of a general kernel gene-score, following results in the Appendix A of Wu et al. [30].

Suppose that kernel matrix K_i is used for gene i to compute the gene score $Q_i = (y - \hat{p}_o)' K_i (y - \hat{p}_o)$. For example, if IBS were used, then K_i would be an n x n matrix of IBS scores based on SNPs in the i^{th} gene. Then, the null expected value of Q_i is $E[Q_i] = tr(K_i V_o)$, where V_o is the null covariance matrix for the residual vector $(y - \hat{p}_o)$, which can be estimated by $V_o = D_o - D_o A(A' D_o A)^{-1} A' D_o$. The matrix A is the design matrix for the intercept and any adjusting covariates, and D_o is a diagonal matrix with binomial variance terms $\hat{p}_{oi}(1 - \hat{p}_{oi})$. Furthermore, the covariance of Q_i and Q_j is $Cov(Q_i, Q_j) = 2tr(K_i V_o K_j V_o)$. Knowing these moments, we can follow our procedure of computing weighted averages of gene-scores to compute a set-score, and then create a normalized set-score statistic by subtracting the null value of the expected set-score and dividing by the standard error of the set-score. This illustrates that different kernels could be used for different genes, although it might be most sensible to use consistent kernel definitions, such as linear kernels or IBS kernels across all genes. Note that if the design matrix A has only an intercept, then $\hat{p}_{oi}$ is constant over all subjects (the fraction of cases in the sample), and $V_o = p_o(1 - p_o)[I - \frac{1}{n}J]$, where I is the identity matrix and J is a matrix of all 1's.

References

1. Hindorff, LA.; Junkins, HA.; Hall, PN.; Mehta, JP.; Manolio, TA. [July 20, 2011] A Catalog of Published Genome-Wide Association Studies. Available at: www.genome.gov/gwastudies.
2. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, et al. Finding the missing heritability of complex diseases. Nature. 2009; 461:747–753. [PubMed: 19812666]
3. Maher B. Personal genomes: The case of the missing heritability. Nature. 2008; 456:18–21. [PubMed: 18987709]
4. Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, Madden PA, Heath AC, Martin NG, Montgomery GW, et al. Common SNPs explain a large proportion of the heritability for human height. Nat Genet. 2010; 42:565–569. [PubMed: 20562875]
5. Park JH, Wacholder S, Gail MH, Peters U, Jacobs KB, Chanock SJ, Chatterjee N. Estimation of effect size distribution from genome-wide association studies and implications for future discoveries. Nat Genet. 2010; 42:570–575. [PubMed: 20562874]
6. Chen Y, Zhu J, Lum PY, Yang X, Pinto S, MacNeil DJ, Zhang C, Lamb J, Edwards S, Sieberts SK, et al. Variations in DNA elucidate molecular networks that cause disease. Nature. 2008; 452:429–435. [PubMed: 18344982]
7. Cantor RM, Lange K, Sinsheimer JS. Prioritizing GWAS results: A review of statistical methods and recommendations for their application. Am J Hum Genet. 2010; 86:6–22. [PubMed: 20074509]
8. Ackermann M, Strimmer K. A general modular framework for gene set enrichment analysis. BMC Bioinformatics. 2009; 10:47. [PubMed: 19192285]
9. Song S, Black MA. Microarray-based gene set analysis: a comparison of current methods. BMC Bioinformatics. 2008; 9:502. [PubMed: 19038052]
10. Yu K, Li Q, Bergen AW, Pfeiffer RM, Rosenberg PS, Caporaso N, Kraft P, Chatterjee N. Pathway analysis by adaptive combination of P-values. Genet Epidemiol. 2009; 33:700–709. [PubMed: 19333968]

11. Hong MG, Pawitan Y, Magnusson PK, Prince JA. Strategies and issues in the detection of pathway enrichment in genome-wide association studies. Hum Genet. 2009; 126:289–301. [PubMed: 19408013]
12. Chasman DI. On the utility of gene set methods in genomewide association studies of quantitative traits. Genet Epidemiol. 2008; 32:658–668. [PubMed: 18481796]
13. Wang K, Li M, Bucan M. Pathway-based approaches for analysis of genomewide association studies. Am J Hum Genet. 2007; 81:1278–1283. [PubMed: 17966091]
14. Menashe I, Maeder D, Garcia-Closas M, Figueroa JD, Bhattacharjee S, Rotunno M, Kraft P, Hunter DJ, Chanock SJ, Rosenberg PS, et al. Pathway analysis of breast cancer genome-wide association study highlights three pathways and one canonical signaling cascade. Cancer Res. 2010; 70:4453–4459. [PubMed: 20460509]
15. Hosgood HD 3rd, Menashe I, Shen M, Yeager M, Yuenger J, Rajaraman P, He X, Chatterjee N, Caporaso NE, Zhu Y, et al. Pathway-based evaluation of 380 candidate genes and lung cancer susceptibility suggests the importance of the cell cycle pathway. Carcinogenesis. 2008; 29:1938–1943. [PubMed: 18676680]
16. Perry JRB, McCarthy MI, Hattersley AT, Zeggini E, the Wellcome Trust Case Control, C. Weedon MN, Frayling TM. Interrogating Type 2 Diabetes Genome-Wide Association Data Using a Biological Pathway-Based Approach. Diabetes. 2009; 58:1463–1467. [PubMed: 19252133]
17. Zhang L, Guo YF, Liu YZ, Liu YJ, Xiong DH, Liu XG, Wang L, Yang TL, Lei SF, Guo Y, et al. Pathway-based genome-wide association analysis identified the importance of regulation- of-autophagy pathway for ultradistal radius BMD. J Bone Miner Res. 2010; 25:1572–1580. [PubMed: 20200951]
18. Peng G, Luo L, Siu H, Zhu Y, Hu P, Hong S, Zhao J, Zhou X, Reveille JD, Jin L, et al. Gene and pathway-based second-wave analysis of genome-wide association studies. Eur J Hum Genet. 2010; 18:111–117. [PubMed: 19584899]
19. Torkamani A, Topol EJ, Schork NJ. Pathway analysis of seven common diseases assessed by genome-wide association. Genomics. 2008; 92:265–272. [PubMed: 18722519]
20. Goeman JJ, Buhlmann P. Analyzing gene expression data in terms of gene sets: methodological issues. Bioinformatics. 2007; 23:980–987. [PubMed: 17303618]
21. Elbers CC, van Eijk KR, Franke L, Mulder F, van der Schouw YT, Wijmenga C, Onland-Moret NC. Using genome-wide pathway analysis to unravel the etiology of complex diseases. Genet Epidemiol. 2009; 33:419–431. [PubMed: 19235186]
22. Holmans P. Statistical methods for pathway analysis of genome-wide data for association with complex genetic traits. Adv Genet. 2010; 72:141–179. [PubMed: 21029852]
23. Kanehisa M, Goto S. KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 2000; 28:27–30. [PubMed: 10592173]
24. Ashburner M, Ball C, Blake J, Botstein D, Butler H, Cherry J, Davis A, Dolinski K, Dwight S, Eppig J, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet. 2000; 25:25–29. [PubMed: 10802651]
25. Holmans P, Green EK, Pahwa JS, Ferreira MA, Purcell SM, Sklar P, Owen MJ, O'Donovan MC, Craddock N. Gene ontology analysis of GWA study data sets provides insights into the biology of bipolar disorder. Am J Hum Genet. 2009; 85:13–24. [PubMed: 19539887]
26. Meinshausen N. Hierarchical testing of variable importance. Biometrika. 2008; 95:265–278.
27. Goeman JJ, Mansmann U. Multiple testing on the directed acyclic graph of gene ontology. Bioinformatics. 2008; 24:537–544. [PubMed: 18203773]
28. Liang K, Nettleton D. A hidden Markov model approach to testing multiple hypotheses on a tree-transformed gene ontology graph. Amer Statist Assoc To Appear. 2010
29. Newton M, Quintana F, denBoon J, Sengupta S, Ahlquist P. Random-set methods identify distinct aspects of the enrichment signal in gene-set analysis. Annals of Applied Statistics. 2007; 1:85–106.
30. Wu MC, Kraft P, Epstein MP, Taylor DM, Chanock SJ, Hunter DJ, Lin X. Powerful SNP-set analysis for case-control genome-wide association studies. Am J Hum Genet. 2010; 86:929–942. [PubMed: 20560208]

31. Lancaster H. Traces and cumulants of quadratic forms in normal variables. J Royal Stat Soc, Ser B. 1954; 16:247–254.
32. Tallis G. The moment generating function of the truncated multi-normal distribution. J Royal Stat Soc, Ser B. 1961; 23:223–229.
33. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet. 2006; 38:904–909. [PubMed: 16862161]
34. Patterson N, Price AL, Reich D. Population structure and eigenanalysis. PLoS Genet. 2006; 2:e190. [PubMed: 17194218]
35. Potter DM. A permutation test for inference in logistic regression with small- and moderate-sized data sets. Stat Med. 2005; 24:693–708. [PubMed: 15515134]
36. Li Y, Willer C, Sanna S, Abecasis G. Genotype imputation. Annu Rev Genomics Hum Genet. 2009; 10:387–406. [PubMed: 19715440]
37. Lin DY. An efficient Monte Carlo approach to assessing statistical significance in genomic studies. Bioinformatics. 2005; 21:781–787. [PubMed: 15454414]
38. Ingle JN, Schaid DJ, Goss PE, Liu M, Mushiroda T, Chapman JA, Kubo M, Jenkins GD, Batzler A, Shepherd L, et al. Genome-Wide Associations and Functional Genomic Studies of Musculoskeletal Adverse Events in Women Receiving Aromatase Inhibitors. J Clin Oncol. 2010; 28:4674–4682. [PubMed: 20876420]
39. Fisher B, Costantino JP, Wickerham DL, Redmond CK, Kavanah M, Cronin WM, Vogel V, Robidoux A, Dimitrov N, Atkins J, et al. Tamoxifen for prevention of breast cancer: report of the National Surgical Adjuvant Breast and Bowel Project P-1 Study. J Natl Cancer Inst. 1998; 90:1371–1388. [PubMed: 9747868]
40. Vogel VG, Costantino JP, Wickerham DL, Cronin WM, Cecchini RS, Atkins JN, Bevers TB, Fehrenbacher L, Pajon ER Jr. Wade JL 3rd, et al. Effects of tamoxifen vs raloxifene on the risk of developing invasive breast cancer and other disease outcomes: the NSABP Study of Tamoxifen and Raloxifene (STAR) P-2 trial. JAMA. 2006; 295:2727–2741. [PubMed: 16754727]
41. Ingle J, Liu M, Wickerham D, Schaid D, Wang L, Mushiroda T, Kubo M, Costantino J, Goetz M, Ames M, et al. Genome-wide assocations of breast events and functional genomic studies in high-risk women receiving tamoxifen or raloxifene on NSABP P1 and P2 prevention trials. A Pharmacogenomics Research Network-RIKEN-NSABP Collaboration. Cancer Res. 2010; 70(24 Suppl):110. (Abstract).
42. Goeman J, van de Geer S, van Houwelingen H. Testing against a high dimensional alternative. J Royal Stat Soc Ser B. 2006; 68:477–493.
43. Gauderman WJ, Murcray C, Gilliland F, Conti DV. Testing association between disease and multiple SNPs in a candidate gene. Genetic Epidemiology. 2007
44. Wang K, Abbott D. A principal components regression approach to multilocus genetic association studies. Genet Epidemiol. 2008; 32:108–118. [PubMed: 17849491]
45. Chen X, Wang L, Hu B, Guo M, Barnard J, Zhu X. Pathway-based analysis for genome- wide association studies using supervised principal components. Genet Epidemiol. 2010; 34:716–724. [PubMed: 20842628]
46. Ballard DH, Cho J, Zhao H. Comparisons of multi-marker association methods to detect association between a candidate region and disease. Genet Epidemiol. 2010; 34:201–212. [PubMed: 19810024]
47. Han F, Pan W. Powerful multi-marker association tests: unifying genomic distance-based regression and logistic regression. Genet Epidemiol. 2010; 34:680–688. [PubMed: 20976795]
48. Pan W. Asymptotic tests of association with multiple SNPs in linkage disequilibrium. Genet Epidemiol. 2009; 33:497–507. [PubMed: 19170135]
49. Chapman J, Whittaker J. Analysis of multiple SNPs in a candidate gene or region. Genet Epidemiol. 2008; 32:560–566. [PubMed: 18428428]
50. Wang T, Elston RC. Improved power by use of a weighted score test for linkage disequilibrium mapping. Am J Hum Genet. 2007; 80:353–360. [PubMed: 17236140]
51. Luo L, Peng G, Zhu Y, Dong H, Amos CI, Xiong M. Genome-wide gene and pathway analysis. Eur J Hum Genet. 2010; 18:1045–1053. [PubMed: 20442747]

52. Liu JZ, McRae AF, Nyholt DR, Medland SE, Wray NR, Brown KM, Hayward NK, Montgomery GW, Visscher PM, Martin NG, et al. A versatile gene-based test for genome-wide association studies. Am J Hum Genet. 2010; 87:139–145. [PubMed: 20598278]

53. Segre AV, Groop L, Mootha VK, Daly MJ, Altshuler D. Common inherited variation in mitochondrial genes is not enriched for associations with type 2 diabetes or related glycemic traits. PLoS Genet. 2010:6.

54. Rhee SY, Wood V, Dolinski K, Draghici S. Use and misuse of the gene ontology annotations. Nat Rev Genet. 2008; 9:509–515. [PubMed: 18475267]

55. Consoritium. The Gene Ontology in 2010: extensions and refinements. Nucleic Acids Res. 2009; 38:D331–335. [PubMed: 19920128]

56. Dudois, D.; Prade, H. Fuzzy sets and systems. Academic Press; New York: 1980.

57. Siek, J.; Lee, S-Q.; Lumsdaine, A. The Boost Graph Library. Addison-Wesley; New York: 2002.

58. Lin X. Variance component testing in generalised linear models with random effects. Biometrika. 1997; 84:309–326.

Table A1

SNPs excluded when converting from build 36.3 to build 37.1

Code	Reason	SNP Count
"Not On"	No longer mapped	716
"Multi"	Mapped to multiple chromosomes	3,113
No Position	Chromosome given, but no position	3,801
Merged	Merged to another SNP	4,319
Total		11,949

Table 1

MA27 GO gene-set scan results.

GO Namespace	GO Term ID	GO Term Definition	No. genes in set	z-score	p-value
Cellular Component	GO:0042643	actomyosin, actin part	1	3.382	0.47
	GO:0043159	acrosomal matrix	3	2.893	0.867
	GO:0031307	integral to mitochondrial outer membrane	6	2.868	0.884
	GO:0031306	intrinsic to mitochondrial outer membrane	7	2.654	0.971
	GO:0045254	pyruvate dehydrogenase complex	3	2.368	0.994
Biological Process	GO:0000266	mitochondrial fission	8	3.489	0.934
	GO:0045475	locomotor rhythm	3	3.472	0.936
	GO:0010534 GO:0010535	regulation of activation of JAK2 kinase activity	1	3.444	0.945
	GO:0032980	keratinocyte activation	1	3.242	0.992
	GO:0002312 GO:0002313	B cell activation involved in immune response	5	3.176	0.997
Molecular Function	GO:0008097	5S rRNA binding	1	3.542	0.882
	GO:0004756 GO:0016781	selenide, water dikinase activity phosphotransferase activity, paired acceptors	2	3.457	0.931
	GO:0048187	inhibin beta-B binding	2	3.282	0.978
	GO:0001595 GO:0004945	angiotensin receptor activity angiotensin type II receptor activity	4	3.104	0.993
	GO:0003868	4-hydroxyphenylpyruvate dioxygenase activity	2	3.088	0.995

Table 2

MA27 gene-level scan results.

Gene ID	Gene Definition	No. SNPs	z-score	p-value
100302269	MIR663B microRNA 663B	1	4.493	0.943
100132958	LOC100132958 hypothetical	3	4.483	0.943
9623	TCL1B T-cell leukemia/lymphoma 1B	40	4.313	0.984
100287215	LOC100287215 ubiquitin-like protein 5 pseudogene	15	4.296	0.989
100287080	LOC100287080 hypothetical protein	37	4.179	0.995
646915	ZNF806 zinc finger protein 806	4	4.134	0.996
399949	C11orf88 chromosome 11 open reading frame 88	13	4.130	0.997
100287733	LOC100287733 hypothetical protein	38	4.065	0.998
221476	PI16 peptidase inhibitor 16	38	4.065	0.998
222236	NAPEPLD N-acyl phosphatidylethanolamine phospholipase D	6	4.057	0.998

Table 3

NSABP GO gene-set scan results.

GO Namespace	GO Term ID	GO Term Definition	No. genes in set	z-score	p-value
Cellular Component	GO:0005869	dynactin complex	6	2.796	0.927
	GO:0000127	transcription factor TFIIIC complex	6	2.729	0.954
	GO:0033193	Lsd1/2 complex	1	2.671	0.972
	GO:0005896	interleukin-6 receptor complex	3	2.655	0.972
	GO:0005595	collagen type XII	1	2.627	0.938
Biological Process	GO:0010533 GO:0010536	regulation of activation of Janus kinase activity positive regulation of activation of Janus kinase activity	2	4.083	0.507
	GO:0019048	virus-host interaction	21	3.689	0.835
	GO:0046968	peptide antigen transport	3	3.527	0.936
	GO:0002690	positive regulation of leukocyte chemotaxis	23	3.474	0.954
	GO:0030237		1	3.453	0.959
	GO:0019101 GO:0042703	female sex determination female somatic sex determination menstruation			
Molecular Function	GO:0003709	RNA polymerase III transcription factor activity	10	3.677	0.790
	GO:0070119	ciliary neurotrophic factor binding	1	3.338	0.968
	GO:0004915 GO:0019981	interleukin-6 receptor activity interleukin-6 binding	2	3.137	0.997
	GO:0004890	GABA-A receptor activity	18	3.038	0.999
	GO:0004897	ciliary neurotrophic factor receptor activity	4	3.005	1.000

Table 4

NSABP gene-level scan results.

Gene ID	Gene Definition	No. SNPs	z-score	p-value
350297	MRPS21P8 mitochondrial ribosomal protein S21 pseudogene 8	31	6.031	0.058
23090	ZNF423 zinc finger protein 423	121	4.790	0.768
92335	STRADA STE20-related kinase adaptor alpha	7	4.698	0.826
729683	LOC729683 hypothetical protein	5	4.647	0.849
100289432	LOC100289432 hypothetical protein	5	4.647	0.849
80774	LIMD2 LIM domain containing 2	5	4.647	0.849
54491	FAM105A family with sequence similarity 105, member A	32	4.616	0.877
4215	MAP3K3 mitogen-activated protein kinase kinase kinase 3	8	4.584	0.901
90246	LOC90246 hypothetical	28	4.549	0.920
57003	CCDC47 coiled-coil domain containing 47	5	4.271	0.992

Table 5

Results from KEGG gene-set analyses.

Study	KEGG Pathway ID	KEGG Pathway Definition	No. genes in set	z-score	p-value
MA27	04614	Renin-angiotensin system	17	1.83	0.963
	00920	Sulfur metabolism	12	1.736	0.986
	00130	Ubiquinone and other terpenoid-quinone biosynthesis	7	1.618	0.992
	00785	Lipoic acid metabolism	3	1.576	0.997
	05020	Prion diseases	36	1.493	0.998
NSABP	00785	Lipoic acid metabolism	3	2.782	0.342
	04330	Notch signaling pathway	47	2.186	0.838
	04950	Maturity onset diabetes of the young	25	2.09	0.895
	04145	Phagosome	144	1.837	0.983
	00770	Pantothenate and CoA biosynthesis	16	1.475	1

Table 6

Time to Compute Gene and Gene-set Analyses*

Study	Number Subjects	Gene-level	GO			KEGG
			biological process	cellular component	molecular function	
MA.27	878	0:51	0:58	0:52	0:56	3:59
NSABP	1766	1:38	1:54	1:37	1:55	7:40

* Time in hours:minutes, based on 1,000 simulations for p-values.