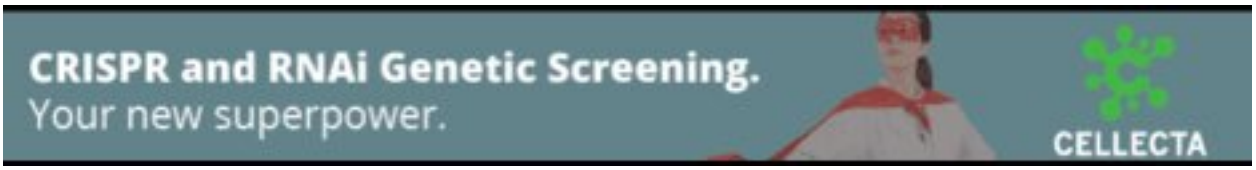
Integrating genetic and gene expression evidence into genome-wide association analysis of gene sets

Qing Xiong, Nicola Ancona, Elizabeth R. Hauser, et al.

Genome Res. published online September 22, 2011
Access the most recent version at doi:10.1101/gr.124370.111



Integrating genetic and gene expression evidence into genome-wide association analysis of gene sets

Qing Xiong,[1] Nicola Ancona,[2] Elizabeth R. Hauser,[3] Sayan Mukherjee,[4,5,6] and Terrence S. Furey[1,5,6]

[1]*Department of Genetics, Department of Biology, Lineberger Comprehensive Cancer Center, and Carolina Center for Genome Sciences, The University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599, USA;* [2]*Institute of Intelligent Systems for Automation National Research Council, Bari IT 70126, Italy;* [3]*Center for Human Genetics and Section of Medical Genetics, Department of Medicine, Duke University, Durham, North Carolina 27710, USA;* [4]*Departments of Statistical Science, Computer Science, and Mathematics, Institute for Genome Sciences & Policy, Duke University, Durham, North Carolina 27708, USA*

Single variant or single gene analyses generally account for only a small proportion of the phenotypic variation in complex traits. Alternatively, gene set or pathway association analyses are playing an increasingly important role in uncovering genetic architectures of complex traits through the identification of systematic genetic interactions. Two dominant paradigms for gene set analyses are association analyses based on SNP genotypes and those based on gene expression profiles. However, gene–disease association can manifest in many ways, such as alterations of gene expression, genotype, and copy number; thus, an integrative approach combining multiple forms of evidence can more accurately and comprehensively capture pathway associations. We have developed a single statistical framework, Gene Set Association Analysis (GSAA), that simultaneously measures genome-wide patterns of genetic variation and gene expression variation to identify sets of genes enriched for differential expression and/or trait-associated genetic markers. Simulation studies illustrate that joint analyses of genomic data increase the power to detect real associations when compared with gene set methods that use only one genomic data type. The analysis of two human diseases, glioblastoma and Crohn's disease, detected abnormalities in previously identified disease-associated pathways, such as pathways related to PI3K signaling, DNA damage response, and the activation of NFKB. In addition, GSAA predicted novel pathway associations, for example, differential genetic and expression characteristics in genes from the ABC transporter family in glioblastoma and from the HLA system in Crohn's disease. These demonstrate that GSAA can help uncover biological pathways underlying human diseases and complex traits.

[Supplemental material is available for this article.]

Dissecting the genetic and molecular mechanisms underlying complex traits, including many diseases, is one of the key scientific goals in the post-genomic era. In the past decade, genome-wide association studies (GWAS) have emerged as one of the main strategies in finding genetic variants associated with trait variation, and a large number of genetic associations have been identified for a wide variety of common complex diseases as listed in the GWAS catalog (Hindorff et al. 2009) (http://www.genome.gov/gwastudies). Despite the enormous success of these GWAS studies in uncovering important genetic effects, the identified single nucleotide polymorphisms (SNPs) explain only a small proportion of the phenotypic variation, and the predictive power of these SNPs remains low for many complex diseases (Manolio et al. 2009).

The majority of current GWAS analyze individual loci independently to identify causal variants. These analyses have two related limitations. The first is that an initial GWA scan can yield a large number of statistically significant loci that may include causal variants as well as spurious associations, such as those that result from cryptic population structure or other sources of error. It is also difficult to distinguish causal variants from markers in strong linkage disequilibrium (LD) with causal variants. A standard strategy to minimize the number of false associations is to focus on the most significant SNPs for biological validation. This gives rise to the second limitation. Even if these significant SNPs correspond to causal variants, they generally only account for a small proportion of the phenotypic variation. These analyses are not ideal for diseases where the common pattern of allelic architecture consists of potentially hundreds of susceptibility loci that increase the risk of disease. Under this model in which only a few of these variants have large effects and most have small effects, the latter category may not be identified by association analyses of individual variants (Wang et al. 2005).

To overcome these challenges of current GWAS approaches, gene set/pathway association analyses have been developed that identify variation in pathway activity or function associated with trait variation. Compared with single-gene or single-SNP analyses, set-based approaches can potentially (1) reduce the false positives or decrease the uncertainty around causal genes or variants by inferring associations over sets of biologically related genes; (2) facilitate interpretation of the results by providing insights into the functional links between implicated genes or variants; and (3) uncover a significant biological effect distributed over multiple loci even if changes in any individual locus have a small effect. Arguably, this strategy is better suited to capture the allelic architectures of complex diseases. Two dominant paradigms for gene set analyses are association analyses based on gene expression profiles and

[5]These authors contributed equally to this work.
[6]Corresponding authors.
E-mail sayan@stat.duke.edu.
E-mail tsfurey@email.unc.edu.
Article published online before print. Article, supplemental material, and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.124370.111. Freely available online through the *Genome Research* Open Access option.

those based on SNP genotypes. Numerous expression-based strategies have been developed (Goeman et al. 2004; Boorsma et al. 2005; Kim and Volsky 2005; Mansmann and Meister 2005; Subramanian et al. 2005; Dinu et al. 2007; Maglietta et al. 2007; Newton et al. 2007; Luo et al. 2009), and each method has its advantages, limitations, and assumptions (for reviews and comparisons, see Khatri and Draghici 2005; Allison et al. 2006; Goeman and Buhlmann 2007; Liu et al. 2007; Abatangelo et al. 2009). More recently, a few methods have been developed for pathway analyses using SNP data (Wang et al. 2007; Holden et al. 2008; O'Dushlaine et al. 2009; Chen et al. 2010; De la Cruz et al. 2010; Peng et al. 2010; Zhong et al. 2010).

Gene set analyses based on a single data type, for example, gene expression data or SNP data, have successfully revealed altered cellular processes associated with complex diseases (Baranzini et al. 2009; Ooi et al. 2009; Liu et al. 2010; Zhang et al. 2010). However, an integrative statistical framework and computational platform for set-based analyses that can simultaneously leverage information across both expression data and SNP data is still lacking. Valuable associations may be discarded in single data type analyses. For instance, genes with only genetic alterations are not considered in gene set analyses based solely on expression data. Similarly, genes with only expression changes cannot be captured by a purely SNP-based approach. This may miss inferences of gene–disease association that result, in part, from the complex interplay of genetic alterations and gene expression changes leading to the development and progression of diseases. These issues create a need to integrate both genetic and gene expression evidence into the association analysis of gene sets. The observation that expression quantitative trait loci (eQTLs) or eSNPs were more likely detected as disease variants in association studies than SNPs not associated with expression differences (Gorlov et al. 2009; Nicolae et al. 2010) is additional evidence supporting the need for methods integrating gene expression analysis and SNP analysis.

In this study, we propose a novel integrative method called Gene Set Association Analysis (GSAA) for the joint analysis of gene expression and SNP data using a pathway-based strategy for more accurate and comprehensive inference of associations. GSAA is based on the idea of integrating evidence across multiple levels of analyses into a single statistical framework to analyze genetic variation across all SNPs mapped to genes and expression variation over all genes simultaneously. The model integrates these two types of genomic information as overall evidence for gene set association analysis.

Using extensive simulation studies, we illustrated that GSAA outperforms each of three gene set methods that use only one genomic data source and that joint analyses reduced the false discovery rate (FDR) in all simulated scenarios and increased the power in nearly all simulated scenarios. We further validated the ability of GSAA to identify association signals in human disease using data sets from glioblastoma and Crohn's disease. We found significant associations of well-known disease-associated pathways, such as pathways related to PI3K signaling and DNA damage response in glioblastoma and pathways involved in the activation of NFKB in Crohn's disease. These integrative analyses also revealed novel pathway associations we did not find in single data analyses, for example, aberrations in ABC transporter family in glioblastoma and in the human leukocyte antigen (HLA) system in Crohn's disease.

Java-based software implementing GSAA is freely available at http://gsaa.unc.edu. The software includes a user-friendly and straightforward graphical user interface and provides full support for the visualization of results. In addition, we also provide a separate module called gene set association analysis-SNP (GSAA-SNP) that was used in this study and performs pathway-based analysis based solely on SNP genotype data. The GSAA platform now supports the gene set association analysis of all species with complete genome sequences available in the Ensembl database (http://ensembl.org/).

Results

A summary of the multilevel model that was implemented via GSAA is outlined in Figure 1 (for details, see Methods). The key steps consist of (1) calculating gene level scores for differential expression and genetic association, respectively (Fig. 1, Differential gene expression score, Single-SNP association score/SNP-set association score); (2) combining the two gene level scores using a single metric (Fig. 1, Gene association score); and (3) evaluating these combined scores in terms of gene sets or pathways (Fig. 1, gene set/pathway association test).

Simulation studies

We conducted a comprehensive simulation study to illustrate the power of GSAA under various conditions of genetic association and differential expression and to justify the use of certain summary statistics for data integration. These simulations were designed to evaluate two primary questions: (1) What is the relative performance and what are the advantages of the various statistical summaries explored to integrate evidence in our pathway-based approach? (2) Can we reduce the FDR and increase the power of association tests by integrating expression and genotypic data into pathway-based analyses?

GSAA uses summary statistics to summarize evidence at two distinct steps. The first occurs when evidence of differential expression is combined with evidence of genetic association at the

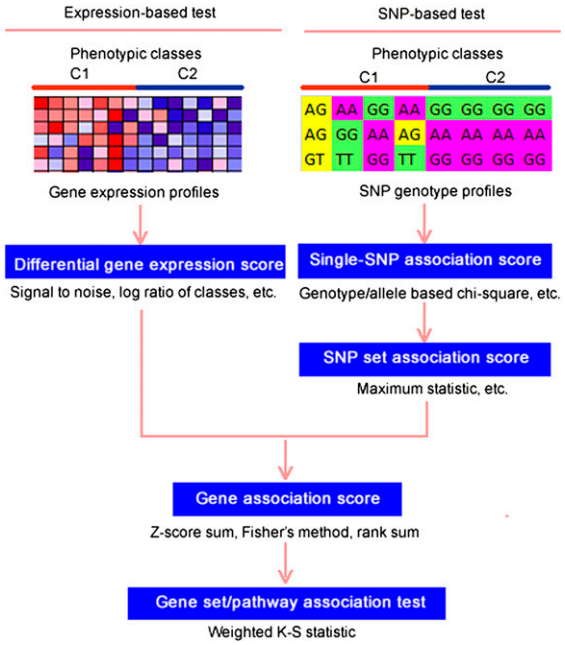


Figure 1. Overview of GSAA. Differential gene expression scores are computed for each gene from gene expression profiles (*left*). Independently, SNP set association scores for each gene are likewise computed based on the SNPs assigned to the respective gene (*right*). The gene association scores integrate evidence from both the expression signature and the genotype signature for each gene. Finally, the pathway association test identifies gene sets associated with samples of a single phenotype by integrating evidence across genes in the gene sets.

level of single genes (Fig. 1, gene association score). The second is when this single gene evidence is aggregated across genes within an a priori defined gene set (Fig. 1, gene set/pathway association test). To combine expression and association evidence at the gene level, we evaluated the following statistics (for details, see Methods): a *Z*-score based sum statistic (zs), a statistic based on Fisher's method (fm), and a rank sum statistic (rs). To integrate evidence across genes, we used a weighted Kolmogorov-Smirnov (K-S) statistic (ks). This results in variations of GSAA indexed by these two statistics, for example, GSAAzs-ks uses the *Z*-score-based sum statistic at the gene level and the weighted K-S statistic at the gene set level. The variations of GSAA we compared were zs-ks, fm-ks, and rs-ks. We added to these comparisons an SNP-based gene set association analysis called GSAA-SNP (see Methods), the previously developed Gene Set Enrichment Analysis (GSEA) (Subramanian et al. 2005) that performs expression based enrichment analysis, and a variant of GSEA we implemented that ranks genes based on the absolute value of their expression differences that we refer to as GSEAndes (see Methods).

We designed five scenarios in our simulation study to test how each of the methods evaluated performed under varying magnitudes and presence of signals in the SNP and expression data with respect to case versus control samples. In each scenario, we simulated genotype and expression data for 1000 genes and 100 gene sets. Only one of 100 defined genes sets contained causal genes. This "causal gene set" contained 16 genes of which a subset were genetically associated and differentially expressed. The remaining 99 gene sets corresponded to a null model and were composed of a random subset of the remaining 984 non-causal genes. Note that a non-causal gene may be assigned to multiple gene sets by this design (for details, see Methods). The five scenarios are distinguished by the degree and type of signal embedded in the genes constituting the causal gene set:

S1: Eight of the 16 genes are both genetically associated with the phenotype and differentially expressed with these eight genes up-regulated in cases.

S2: Four genes are both genetically associated and differentially expressed, all up-regulated in the cases. Four other genes are only differentially expressed, all up-regulated in cases.

S3: Four genes are both genetically associated and differentially expressed, all up-regulated in cases. Four other genes are only genetically associated.

S4: Eight genes are both genetically associated and differentially expressed, six of them up-regulated in cases and the other two down-regulated in cases.

S5: Four genes are only up-regulated in cases. Four other genes are only genetically associated.

In each scenario, the strength of association at the causal loci in the SNP data is determined by the odds ratios at the loci where the odds ratio is drawn independently from a uniform distribution (see Methods). The extent of differential expression is determined by the effect size at each gene in a regression model with the effect size drawn independently from a uniform distribution (see Methods). This process allows for the modeling of a spectrum of effect sizes at the causal loci as well as of the differentially expressed genes.

As stated above, the two objectives of the simulation study were, first, to select the summary statistic to use for integration across genomic data sources at the single-gene level and, second, to quantitate the benefit of an integrative model. Results using the simulated data described above for GSAA and for the three single data type analyses—GSAA-SNP, GSEA, and GSEAndes—are reported in Table 1. For these results, the odds ratio for simulating the genetic association was drawn from a uniform distribution (U [1.1, 1.3]). Results using U [1.2, 1.4] are shown in Supplemental Table S1. The average *P*-value, FDR, and FWER for the causal gene set over 200 replicates and the power of each method in each scenario are reported. The *P*-value, FDR, and FWER were calculated based on 2000 permutations of phenotype labels. Power was calculated as the proportion of replicates for which the *P*-value for the causal gene set was <0.05.

With respect to our first objective, these results show that when combining information at the single gene level, the *Z*-score-based sum statistic performs considerably better than the rank sum statistic and slightly better than the Fisher's method statistic with respect to the FDR and FWER (Table 1; Supplemental Table S1). Interestingly, the rank sum statistic tends to have the lowest power. This may suggest that the loss of information when using only

Table 1. Simulation results based on scenarios S1–S5

	GSAAzs-ks				GSAAfm-ks				GSAArs-ks			
Scenario	*P*-value	FDR	FWER	Power	*P*-value	FDR	FWER	Power	*P*-value	FDR	FWER	Power
S1	0.00136	0.00861	0.00678	0.995	0.00226	0.01069	0.00989	0.99	0.02232	0.20798	0.20691	0.87
S2	0.00929	0.05686	0.05937	0.95	0.00988	0.05678	0.06230	0.97	0.07403	0.43896	0.47427	0.61
S3	0.00232	0.01158	0.01049	0.995	0.00521	0.02769	0.02620	0.99	0.06899	0.40380	0.42313	0.655
S4	0.00100	0.00378	0.00359	1	0.00151	0.00587	0.00567	1	0.02679	0.23876	0.25247	0.845
S5	0.01059	0.07656	0.07863	0.955	0.01431	0.10589	0.10882	0.94	0.13899	0.59330	0.65140	0.38

	GSAA-SNP				GSEA				GSEAndes			
Scenario	*P*-value	FDR	FWER	Power	*P*-value	FDR	FWER	Power	*P*-value	FDR	FWER	Power
S1	0.00300	0.08506	0.08966	0.985	0.00354	0.13129	0.12922	0.99	0.00526	0.32252	0.32704	0.975
S2	0.02956	0.36058	0.39286	0.88	0.00288	0.15532	0.15265	0.995	0.00566	0.35774	0.35181	0.97
S3	0.00300	0.08506	0.08966	0.985	0.05040	0.49751	0.51888	0.74	0.03596	0.48409	0.50091	0.8
S4	0.00300	0.08506	0.08966	0.985	0.08978	0.72820	0.80797	0.465	0.00378	0.31637	0.31944	0.985
S5	0.03866	0.45631	0.49498	0.795	0.04868	0.50246	0.53658	0.705	0.03313	0.44409	0.48971	0.82

Three versions of GSAA were evaluated where each varied the summary statistic used for combining genetic and gene expression evidence at the single gene level: GSAAzs-ks, GSAAfm-ks, and GSAArs-ks. Also shown are results for GSAA-SNP, GSEA, and GSEAndes. For these simulations, the odds ratios for causal loci were drawn from U [1.1, 1.3], and 200 simulated replicates were used.

rank information causes a decrease in power. In addition, we note that Fisher's method for combining *P*-values has limitations. It treats large and small *P*-values asymmetrically and is asymmetrically sensitive to small *P*-values compared with large *P*-values (Whitlock 2005). This asymmetry could result in a bias in certain conditions that may lead to worse overall performance.

As to our second objective, we evaluated the advantage of using an integrative approach by comparing the FDR and power for GSAAzs-ks and three single-source methods: GSAA-SNP, GSEA, and GSEAndes. Our results indicate that GSAAzs-ks substantially outperforms each of these three single data type analyses (Table 1; Supplemental Table S1). Overall, FDRs from GSAAzs-ks are consistently smaller than GSAA-SNP, GSEA, and GSEAndes in all simulated scenarios. The power of GSAAzs-ks is also better than or equal to those of single data type analyses in all simulated situations except for scenario S2 under the odds ratio setting U [1.1, 1.3], in which GSEA and GSEAndes is slightly better. Gene association scores calculated using GSAAzs-ks are higher for genes with alterations in both gene expression and genotype compared with genes with a single type of alteration. As expected, GSAAzs-ks performs better than single data type analyses when all or part of causal genes contain both gene expression and genetic alterations (S1–S4). GSAAzs-ks also shows increased ability to detect effects when the gene set includes both types of alterations, but in which no single gene simultaneously contains both types (S5). GSAA-SNP performance decreases when fewer genes are associated with phenotype (S2 and S5), and both variants of GSEA suffer when fewer genes are differentially regulated (S3 and S5). Interestingly, the power of traditional GSEA decreases markedly when the causal gene set contains both up-regulated and down-regulated genes (S4), but GSEAndes still retains power similar to scenario S1 in this case. Since both GSAA and GSEAndes use a non-directional differential expression score, there was no loss of power when up-regulation and down-regulation of genes coexist in the gene set.

Analyses of glioblastoma data

Human glioblastoma, the most common type of primary adult brain cancer, was the first cancer targeted for analysis within The Cancer Genome Atlas (TCGA) project. Several types of molecular data were generated including gene expression and SNP genotype data as well as DNA copy number and sequence data (Cancer Genome Atlas Research Network 2008). We applied GSAAzs-ks, GSAA-SNP, GSEA, and GSEAndes to these data using gene sets from the Molecular Signatures Database (MSigDB) that included 357 canonical pathways (see Methods). We used GSAAzs-ks based on our simulations that indicated that the *Z*-score-based data integration tends to have the highest ability to detect effects of association. Full results are shown in Supplemental Tables S2–S8. Consistent with previous analyses using GSEA, we use a FDR cutoff of 0.25 (Subramanian et al. 2005). For these results, we have designated gene sets as being associated with the phenotypic class in which genes are up-regulated overall, although in many gene sets there will be some genes that are down-regulated in the assigned class. GSAAzs-ks identified 39 canonical pathways significantly associated with tumor samples with FDR $\leq$ 0.25 (Table 2; Supplemental Table S2).

Initial integrative pathway analyses carried out by TCGA using sequence and DNA copy number data, but not SNP genotype data, indicated that three major pathways are frequently altered in glioblastoma: RTK/RAS/PI3K signaling, p53 signaling, and RB signaling (2008). Consistent with this, we identified multiple pathways involved in different aspects of these three signaling pathways. Three pathways (Table 2: G20, G21, G28) contain core components of RTK/RAS/PI3K signaling (see Supplemental Table S2 for genes in these pathways) with pathways G20 and G28 directly describing PI3K/AKT signaling. PI3K acts as an upstream activator of AKT that has pivotal roles in apoptosis, proliferation, and cell survival, and alterations of genes in the PI3K/AKT pathway have been considered as causal forces underlying cancer (Osaki et al. 2004; Carnero et al. 2008). Pathway G21 is involved in the activation of the EGFR pathway. *EGFR* is reportedly overexpressed and mutated in a significant proportion of glioblastoma (Heimberger et al. 2005).

Alterations in p53 and RB signaling are reflected by another group of pathways: G8, G17, G22, G24, G25, and G29. These pathways are commensurate with a general oncogene-induced DNA damage model for cancer development and progression uncovered by several experimental studies (Kastan and Bartek 2004; Bartkova et al. 2005; Gorgoulis et al. 2005; Venkitaraman 2005; Halazonetis et al. 2008). The key contributors to the association of these pathways contain the core genes involved in checkpoint response to DNA damage such as *TP53*, *CHEK2* (also called *CHK2*), and *ATM* (Supplemental Table S2). In addition, the DNA damage model indicates the involvement of apoptosis and DNA replication in the oncogenesis. A considerable number of significant pathways in GSAA analysis are also related to apoptosis (G1, G2, G5, G11, G12, G14, G20, G27, G28, G35, G37, G39) and DNA replication (G26).

In addition to aberrations in these well-known signaling pathways, GSAA analysis also suggested a role for the family of ATP-binding cassette (ABC) transporters, the third-ranked pathway (G3). Gene expression profiles of many ABC transporter genes were significantly altered in tumor samples. Some genes, such as *ABCB7* and *ABCC4*, were up-regulated, while others, for example, *ABCC8* and *ABCG4*, were down-regulated. Significantly associated genomic variants were found in several of these genes including *ABCB7* rs6647618: $p = 1.0 \times 10^{-4}$) and *ABCC4* (rs7324277: $p = 1.0 \times 10^{-4}$). Normal and cancer stem cells have been shown to express high levels of ABC transporters that are normally inactive in more mature cells, and the overexpression of specific ATP transporters has been found to significantly affect the chemoresistance phenotype of cancer cells (Bronger et al. 2005; Bleau et al. 2009).

GSAA analysis also suggests a connection between the coagulation system and glioblastoma, represented by the signaling pathway involved in platelet activation (G4) and the intrinsic prothrombin activation pathway (G13). Patients with cancer, especially glioblastoma, have been shown to have increased risk for thrombosis, characterized by alterations in normal blood flow, injury to the vascular endothelium, and alterations in the constitution of blood (Lyman and Khorana 2009; Reynes et al. 2010). Glioblastoma cells have also been found to become both procoagulant and hypersensitive to TF/PAR-mediated signaling. The expression of *EGFR* and the most common *EGFR* mutant, *EGFRvIII*, is thought to drive this transformation (Magnus et al. 2010). In addition, our analysis indicated that the coagulation factor receptor *F2R* (also called *PAR1*) is the most significantly up-regulated gene in glioblastoma samples. *F2R* is also significantly associated in the SNP-based test (rs2227753: $p = 1.0 \times 10^{-4}$). F2R is functional in glioblastoma cells and can mediate anti-apoptotic signaling in the nervous system (Guo et al. 2004; Junge et al. 2004). Another pathway, the VEGF pathway (G19), is also related to blood flow being involved in tumor angiogenesis (Plate et al. 1994; Saharinen et al. 2011).

Two identified gene sets are related to other cancers, namely, bladder cancer (G32) and myeloid leukemia (G38), and have common components with other cancer types. Several pathways are involved in sugar metabolism (G16, G18, G33, G36).

Table 2. Significant pathways associated with glioblastoma tumor samples (FDR ≤ 0.25)

Index	Gene set name	*P*-value	FDR	Function
G1	RELAPATHWAY	0.0010	0.0463	Proliferation, migration and apoptosis
G2	CERAMIDEPATHWAY	0.0006	0.0676	Apoptosis, cellular differentiation, proliferation
G3	HSA02010 ABC TRANSPORTERS_GENERAL	0.0036	0.1247	ATP binding
G4	SPPAPATHWAY	0.0002	0.1305	Thrombin signaling, blood coagulation
G5	NFKBPATHWAY	0.0109	0.1361	Proliferation, migration, apoptosis
G6	HCMVPATHWAY	0.0118	0.1505	Proliferation, viral replication
G7	TOLLPATHWAY	0.0083	0.1555	Immune response, proliferation
G8	G2PATHWAY	0.0054	0.1658	Cell cycle checkpoints, DNA damage response
G9	INTEGRINPATHWAY	0.0043	0.1732	Cellular shape, mobility, cell cycle
G10	CARM ERPATHWAY	0.0060	0.1734	Activation of transcriptional factors, regulation of estrogen receptor
G11	TNFR1PATHWAY	0.0102	0.1830	Apoptosis
G12	FASPATHWAY	0.0140	0.1834	Apoptosis
G13	INTRINSICPATHWAY	0.0287	0.1858	Thrombin signaling, blood coagulation
G14	STRESSPATHWAY	0.0052	0.1931	Apoptosis
G15	HSA05040 HUNTINGTONS DISEASE	0.0078	0.1932	Huntington disease
G16	HSA00531 GLYCOSAMINOGLYCAN DEGRADATION	0.0499	0.1932	Glycosaminoglycan degradation
G17	ATMPATHWAY	0.0289	0.1952	Cell cycle checkpoints, DNA damage response
G18	FRUCTOSE AND MANNOSE METABOLISM	0.0288	0.1998	Fructose and mannose metabolism
G19	VEGFPATHWAY	0.0084	0.2031	Angiogenesis, blood vessel formation
G20	SIG PIP3 SIGNALING IN B LYMPHOCYTES	0.0223	0.2065	PI3K/AKT signaling, apoptosis, proliferation
G21	CARDIACEGFPATHWAY	0.0125	0.2106	EGFR signaling
G22	CELLCYCLEPATHWAY	0.0332	0.2106	Cell cycle checkpoints, DNA damage response
G23	ST ERK1 ERK2 MAPK PATHWAY	0.0172	0.2150	ERK1/ERK2 MAPK signaling
G24	RACCYCDPATHWAY	0.0281	0.2152	Cell cycle
G25	CELL CYCLE KEGG	0.0722	0.2164	Cell cycle checkpoints, DNA damage response
G26	TELPATHWAY	0.0590	0.2252	DNA replication, cell division
G27	ST FAS SIGNALING PATHWAY	0.0177	0.2310	Apoptosis
G28	AKTPATHWAY	0.0510	0.2312	PI3K/AKT signaling, apoptosis, proliferation
G29	ARFPATHWAY	0.0746	0.2334	Cell cycle checkpoints, DNA damage response
G30	HSA04120 UBIQUITIN MEDIATED PROTEOLYSIS	0.0412	0.2351	Proteolysis
G31	IL7PATHWAY	0.1139	0.2388	B- and T-cell development
G32	HSA05219 BLADDER CANCER	0.0217	0.2403	Bladder cancer
G33	HSA00051 FRUCTOSE AND MANNOSE METABOLISM	0.0416	0.2404	Fructose and mannose metabolism
G34	INTEGRIN MEDIATED CELL ADHESION KEGG	0.0108	0.2411	Cellular shape, mobility, cell cycle
G35	TNFR2PATHWAY	0.0722	0.2444	Apoptosis, proliferation
G36	HSA01032 GLYCAN STRUCTURES DEGRADATION	0.0886	0.2453	Glycan degradation
G37	ST P38 MAPK PATHWAY	0.0275	0.2489	p38 MAPK signaling, proliferation, differentiation, apoptosis
G38	HSA05220 CHRONIC MYELOID LEUKEMIA	0.0119	0.2495	Myeloid leukemia
G39	SPRYPATHWAY	0.0947	0.2497	Proliferation, differentiation, apoptosis

For full results, see Supplemental Table S2.

We also investigated the 83 pathways associated with normal samples at FDR ≤ 0.25 (Supplemental Table S3). Perturbations in calcium-mediated signal transduction pathways may be involved in the pathogenesis of glioblastoma since the top three pathways were calcineurin-mediated pathways and 26 of the 83 significant pathways were calcium/calmodulin-dependent. The expression level of catalytic subunit A of calcineurin (*PPP3CA*) and some genes in the families of calmodulins (CALM) and calcium/calmodulin-dependent protein kinases (CAMK), for example, *CALM1*, *CALM2*, *CALM3*, *CAMK2A*, and *CAMK2B*, were down-regulated in tumor tissues. Some genes also include statistically significant genomic variants, such as *PPP3CA* (rs12647627: $p = 1.0 \times 10^{-4}$) and *CALM1* (rs2300500: $p = 1.0 \times 10^{-4}$). It has been shown that a high level of activity of calcineurin predisposes neuronal cells to apoptosis (Asai et al. 1999; Wang et al. 1999). Calcium-mediated signaling also has a role in the regulation of the cell cycle (Baksh et al. 2000; Kahl and Means 2003). Interestingly, a gene set defined based on genes involved in glioma (G69) is associated with normal samples. We noticed that this pathway, unlike other cancer pathways, contains many CALMs and CAMKs that were down-regulated in tumor tissues, although other genes, such as *EGFR* and *TP53*, were up-regulated (see Supplemental Table S3 for genes in this pathway). This may explain why this pathway was more associated with normal samples and supports the connection between calcium processing and glioblastoma. Other top pathways were related to the release of neurotransmitters such as glutamate (G1, G6) and ATPase signaling (G16, G19, G22, G31). These results suggest that these particular processes were down-regulated in diseased tissues.

Neither GSEA nor GSAA-SNP identified significant pathways associated with tumor samples at FDR ≤ 0.25 (Supplemental Tables S4, S6). However, in both GSEA and GSAA-SNP, the 10 most significant pathways included ones related to cell cycle checkpoints and the coagulation system. GSEA also reported pathways involved in apoptosis and DNA replication, while PI3K/AKT signaling pathways were highly ranked by GSAA-SNP. Similar pathways as those found significantly associated with normal samples by GSAA were likewise identified by GSEA (Supplemental Tables S3, S5). GSEAndes

did not identify any pathways as significant in either direction (Supplemental Tables S7, S8).

Analyses of Crohn's disease data

To explore the performance of GSAA within the context of a non-cancer complex disease, we applied GSAAzs-ks, GSAA-SNP, GSEA, and GSEAndes to previously published gene expression and SNP genotype data from patients with and without Crohn's disease (CD). As before, MSigDB canonical pathways (352) were used for this analysis. Full results are shown in Supplemental Tables S9–S15. GSAA analysis identified 12 canonical pathways significantly associated with case samples at FDR ≤ 0.25 (Table 3; Supplemental Table S9).

Proteasome activity was found to be highly associated with disease because the fifth and eleventh most significant pathways were directly related to the proteasome complex. Recent studies have demonstrated that the transcription factor NFKB is a key regulator of epithelial integrity and intestinal immune homeostasis (Ben-Neriah and Schmidt-Supprian 2007; Nenci et al. 2007; Zaph et al. 2007; Atreya et al. 2008; Spehlmann and Eckmann 2009). Deficiency in or hyperactivation of NFKB is one of the core mechanisms leading to chronic inflammatory bowel disease (IBD). NFKB signaling is primarily regulated by inhibitory IκB proteins and the IκB kinase complex. Proteasomes play a crucial role in the degradation of inhibitory IκB proteins and the activation of NFKB (Mattson and Meffert 2006; Visekruna et al. 2006; Atreya et al. 2008). We found that most of the proteasome-related genes were up-regulated in disease samples including *PSMB8* (also called *LMP7*) and *PSMB9* (also called *LMP2*). *PSMB9* also contains a significant genomic variant (rs20547: $p = 0.0115$). PSMB8 and PSMB9 are two subunits of the immunoproteasome encoded by the HLA region and are required for the degradation of phosphorylated IκB proteins and for processing of the NFKB precursor (Hayashi and Faustman 2000; Visekruna et al. 2006). Overexpression of immunoproteasomes in the inflamed intestine of CD patients has been observed in multiple studies and has been found correlated to the excessive NFKB activation (Visekruna et al. 2006, 2009a,b). In addition, there is increasing evidence that a bacterial or viral infection and the host reaction to that infection play an important role in the onset of Crohn's disease (Irving and Gibson 2008). Immunoproteasomes can be induced and replace standard proteasomes quickly in response to the viral infection (Yewdell 2005). This process involves the rapid expression of immunoproteasomes, possibly explaining the relevance of two high-ranked pathways related to aminoacyl-tRNA biosynthesis (Table 3).

A gene set encapsulating key genes underlying type I diabetes (T1D) ranked third among canonical pathways. This pathway includes a large number of genes belonging to the HLA system (see Supplemental Table S9 for genes in this pathway). The HLA system encodes cell surface molecules specialized to present antigenic peptides to the T-cell receptor (TCR) on T cells and plays a critical role in the immune system and autoimmunity (Benacerraf 1981; Fernando et al. 2008). It has been shown that T1D and inflammatory bowel disease share common susceptibility pathways (Wang et al. 2010). Multiple loci within HLA genomic region have been reported to be associated with CD (Forcione et al. 1996; Reinshagen et al. 1996; Lombardi et al. 2001; Newman et al. 2004), and additional susceptibility loci may remain undiscovered.

GSAA found five pathways, in addition to the T1D pathway (G3), that are relevant to immune response (G1, G4, G6, G8, G9). Thrombopoietin signaling (G10) was also identified and has been previously reported to be disturbed in CD (Kapsoritakis et al. 2000).

Sixteen pathways were significantly associated with control samples at FDR ≤ 0.25 (Supplemental Table S10). Seven (G1, G2, G3, G4, G7, G8, G10) are related to G-protein-coupled receptor (GPCR) signaling or PI3K/AKT signaling. GPCRs are upstream regulators of PI3K/AKT signaling. PI3K has important roles in lymphocyte development, differentiation, and activation (Okkenhaug and Vanhaesebroeck 2003; Okkenhaug and Fruman 2010). Multiple studies have shown the correlation between the PI3K pathway and IBD (Fukao and Koyasu 2003; Zhao et al. 2008).

GSEA analysis identified 17 pathways significantly associated with case samples at FDR ≤ 0.25 (Supplemental Table S11). Except for two proteasome-related pathways (G4, G8), GSEA identified multiple additional pathways involving the activation of NFKB. The RelA (G3) and NFKB (G14) pathways describe NFKB signaling. The NTHi pathway (G17) describes the induction of an inflammatory response through activation of NFKB triggered by bacterial infection. Six pathways (G1, G5, G6, G7, G13, G17) contain genes that participate in the immune system or inflammatory response. In addition, another identified pathway (G11) is responsible for the activation of matrix metalloproteinases and the degradation of the extracellular matrix. It has been shown that TNF mediated

Table 3. Significant pathways associated with case samples (FDR ≤ 0.25)

Index	Gene set name	*P*-value	FDR	Function
G1	SIG BCR SIGNALING PATHWAY	0.0004	0.0231	Immune response
G2	AMINOACYL TRNA BIOSYNTHESIS	0.0014	0.0239	tRNA biosynthesis
G3	HSA04940 TYPE I DIABETES MELLITUS	0.0008	0.0294	Autoimmunity, immune response, inflammation
G4	ST G ALPHA I PATHWAY	0.0000	0.0306	Chemotaxis, G-protein signaling, immune response
G5	PROTEASOME	0.0586	0.0357	Activation of NFKB
G6	ST DICTYOSTELIUM DISCOIDEUM CAMP CHEMOTAXIS PATHWAY	0.0000	0.0389	Chemotaxis, G-protein signaling, immune response
G7	HSA00970 AMINOACYL TRNA BIOSYNTHESIS	0.0085	0.0476	tRNA biosynthesis
G8	SA B CELL RECEPTOR COMPLEXES	0.0008	0.0675	Immune response
G9	HSA04514 CELL ADHESION MOLECULES	0.0000	0.0744	Immune response, inflammation
G10	TPOPATHWAY	0.0041	0.1203	Thrombopoietin Signaling
G11	PROTEASOMEPATHWAY	0.0715	0.1374	Activation of NFKB
G12	HSA00565 ETHER LIPID METABOLISM	0.0136	0.1824	Ether lipid metabolism

For full results, see Supplemental Table S9.

up-regulation of matrix metalloproteinases results in severe damage of the extracellular matrix and mucosal degradation (Pallone and Monteleone 2001; Atreya et al. 2008). An altered apoptosis pathway (G10) may contribute to inappropriate T-cell accumulation and subsequently chronic inflammation (Ina et al. 1999). Only one pathway associated with control samples reached significance in the gene expression analysis of GSEA (Supplemental Table S12). GSAA-SNP identified seven significant pathways, of which four (G1, G2, G5, G7) are related to PI3K signaling (Supplemental Table S13). GSEAndes identified four significant pathways of which the top three were proteasome pathways (Supplemental Table S14). The fourth gene set is related to the activation of matrix metalloproteinases also suggested by GSEA. No pathways associated with normal samples reached significance in the GSEAndes analysis (Supplemental Table S15).

Discussion

Genome-wide gene expression profiling and genotyping offer unparalleled opportunities to elucidate the underlying mechanisms of complex traits or diseases. In this study, we developed a novel statistical framework that simultaneously integrates gene expression data and genotype data into genome-wide association analysis of biological pathways or gene sets. Combining evidence from these two genomic data sources facilitates identification of genes with differential gene expression, genetic alterations, or both characteristics that are associated with phenotypic traits. Results from our simulation study and the analyses of glioblastoma and Crohn's disease data showed that GSAA captured association signals that occur in either type of genomic data as well as across both genomic data sources.

Many functionally relevant variants that are deleterious have minor allele frequencies of <5% and therefore are not well represented on SNP chips used in GWAS. However, new sequencing technologies are enabling the better identification of both common and rare variants. An area of future development is to adapt GSAA to detect associations of both common variants and rare variants in gene sets by integrating sequence analysis and gene expression analysis. There is no conceptual difference between using sequence data and SNP data in our method. GSAA is well suited to capture concordant association signals over a gene set or multiple loci even if the association information carried by each gene or locus is weak. This is a key strength of GSAA since current research has shown that most complex human diseases are associated with the presence of multiple common or rare variants, each with a low marginal effect, and not simply a few common variants (Kryukov et al. 2007; Gorlov et al. 2008; Robinson 2010). Similarly, other genomic data such as copy number variation, methylation, and microRNA expression will be explored as inputs to GSAA.

GSAA requires a mapping of SNPs to genes. Currently, it is not known exactly what genomic regions affect the function of each particular gene. In our analyses, to each gene we assigned SNPs that were within the region spanning 1 kb upstream of the TSS to the end of the transcribed bases. We know that for many cases, this may not include variants in distal regulatory regions hundreds of kilobases away that influence gene expression levels, but it should include those in the core and proximal promoter regions and part of those in the distal promoter (Bortoluzzi et al. 2005). This mapping can include information about distal regulatory variants if they are in LD with those included in our mapping intervals or if they are within the mapping intervals of other genes in the gene set. In two previous studies (Wang et al. 2007; Peng et al. 2010), one mapped SNPs to the closest gene, while another used all SNPs within a gene to represent that gene. GSAA software provides users the ability to define their own SNP mapping criteria by specifying how many base pairs upstream of and/or downstream from a gene a SNP must be included. Hopefully, the current influx of functional genomic data, especially chromatin data, will eventually allow more accurate mappings.

The optimal way to assess the joint contribution of multiple SNPs mapped to the same gene in association analyses is unknown. The region of association for a gene may harbor only one risk variant or may harbor multiple risk variants that independently contribute to the overall association signal. Compared to test statistics that combine correlation scores or *P*-values across all SNPs, we believe that the maximum statistic we used can more effectively eliminate the negative effects of correlation structure between SNPs and differences in SNP set size on association inference. The maximum statistic should be the best way to measure association signals when the gene region only contains a single risk variant, but contain multiple markers in strong LD with the risk variant that may artificially inflate the association. However, this statistic cannot accurately capture the overall association information when multiple independent risk variants coexist. GSAA would benefit by the development of new algorithms that more effectively assess joint contributions of SNPs to the trait variation. Given its modular framework, new algorithms like these could be easily incorporated.

Gene set association analysis takes advantage of prior knowledge of biological pathways. Operating at the pathway level aids in interpreting results, especially across different experimental platforms or strategies. However, this creates a dependence on a priori knowledge. Inaccurate or incomplete information about these pathways may lead to inaccurate association inferences. With the accumulation of our knowledge on biological processes, pathway annotations are becoming increasingly more accurate, which will continue to increase the power of gene set association tests.

In summary, we report here a novel statistical framework that is capable of effectively identifying the biological pathways or gene sets associated with complex traits or diseases by integrating genetic and gene expression evidence into genome-wide association analysis of gene sets. Compared with gene set methods that use only one genomic data type, our proposed method reduces the FDR in all simulated scenarios and increases the power in nearly all simulated scenarios. In real settings, it not only confirmed the associations of well-known pathways but also provided new insights into the etiology of disease.

Methods

Gene set association analysis

GSAA is based on multiple layers of association tests. The advantage of a multi-layer approach is that evidence for an association signal is aggregated from individual SNPs to individual genes to gene sets. See Figure 1 for a graphical overview of the method. The methodology formulated here is for the case in which samples belong to one of two phenotypic classes. This multi-level procedure consists of five individual calculations: (1) computation of a differential gene expression score; (2) computation of a single-SNP association score; (3) computation of a SNP set association score; (4) computation of a gene association score; and (5) a gene set association test. The following describes each of these in detail.

Differential gene expression score

The differential expression score reflects the degree to which a gene is differentially expressed between two phenotypic classes. It can be computed by a variety of suitable test statistics. In this study, the

test statistic used is the difference of the class means scaled by the standard deviation. The absolute magnitude of the statistic indicates the strength of the correlation between the gene expression profile and the phenotype, and the sign indicates the direction of this correlation. In our software, we provide five different statistics that can be used to calculate this differential expression score, similar to GSEA (for more details, see Supplemental Document S1).

Single–SNP association score

Five different methods to calculate single-SNP association scores are provided in our software: a genotype-based χ^2 statistic; an allele-based χ^2 statistic; a statistic based on frequency differences in major/minor alleles between the two classes; and two statistics extended from genotype-based and allele-based χ^2 statistics, respectively (for more details, see Supplemental Document S1). Other suitable test statistics for the categorical phenotypes can be used. The results described in this paper used the allele-based χ^2 statistic because it had greater power than genotype-based χ^2 statistic for our simulated SNP data that were based on an additive model (data not shown).

SNP set association score

Genotype data are complicated to analyze in gene-focused analyses because there is not a standard mapping of SNPs to genes and multiple SNPs can cover each gene and its regulatory region. To assign SNPs to genes, we define a genomic interval encompassing each gene and some specified number of bases upstream of and downstream from the transcribed region. All SNPs within this interval are used to represent the gene. Given these SNPs, we calculate an SNP set association score for a gene using a maximum statistic. The maximum statistic is the maximum single-SNP score over all the SNPs assigned to the gene. This idea of selecting the maximum score or minimum *P*-value is an example of a minP procedure used in resampling-based multiple testing (Westfall and Young 1993).

Gene association score

The differential gene expression score and SNP set association score for each gene are combined to generate a single gene association score. This composite correlation integrates evidence for association across the gene expression and SNP data. The differential gene expression scores used by GSEA (Subramanian et al. 2005) have directionality: positive values indicate greater expression in class 1, while negative values indicate greater expression in class 2; therefore, the weighted K-S test used in GSEA analysis may only capture differential expression signals from one direction. However, genes in the same pathway are not always differentially expressed in the same direction. Some pathways may contain both up-regulated and down-regulated genes associated with the disease condition because there exist feedback loops in some pathways, such as p53 pathway (Harris and Levine 2005), where the increase in expression of one gene leads to the increased expression of some genes but the decreased expression of others. The directionality derived from the SNP-based test is not biologically meaningful because for each locus it is not known which allele is actually associated with disease. For the SNP set association scores, we do not know the directionality.

Therefore, we take the absolute values of the differential gene expression scores before data integration in order to capture both up-regulation and down-regulation in pathways and to be consistent with the form of SNP set association scores. Directionality is then resolved at the gene set association test step below. In GSAA, three methods are used to integrate the evidence from gene expression analysis and SNP analysis to produce gene association scores.

Z–score sum

For each differential expression score or SNP set association score, we first generate its null distribution by a phenotype-based permutation procedure. Then we standardize these scores by the mean and standard deviation of its null distribution. More specifically, suppose $\{e_1, \ldots, e_N\}$ are the absolute values of differential expression scores for N genes and $\{s_1, \ldots, s_N\}$ are the SNP set association scores for the same genes. The standard expression scores $\{z_{e1}, \ldots, z_{eN}\}$ are computed as

$$z_{ei} = \frac{e_i - \mu_e}{\sigma_e},$$

and the standard SNP set association scores $\{z_{s1}, \ldots, z_{sN}\}$ for the same genes are similarly computed as

$$z_{si} = \frac{s_i - \mu_s}{\sigma_s},$$

where (μ_e, μ_s) and (σ_e, σ_s) are the means and standard deviations of the null distributions corresponding to e_i and s_i, respectively. This transformation brings the scores from different statistical tests or on different scales onto a common scale so that these scores are directly comparable with each other. The *Z*-score transformation results in both positive values and negative values. Negative scores indicate a lack of association. For convenience, we shift all these scores to be positive by adding a constant c that is the absolute value of the most negative score across all standard gene expression scores and standard SNP set association scores. The addition of this constant transforms all negative standard gene expression scores and standard SNP set association scores to positive scores without changing the shapes of their distributions, particularly the right tail, and has no meaningful effect on the subsequent analysis.

The gene association scores are the sum of these standard scores $g_i = (z_{ei} + c) + (z_{si} + c)$.

Fisher's method

For each differential gene expression score and SNP set association score, we first generate its null distribution by a phenotype-based permutation procedure, and then we estimate its *P*-value by comparing the score with its null distribution. Fisher's method (Fisher 1932; Peng et al. 2010), also known as Fisher's combined probability test, is used to combine *P*-values from the expression-based test and the SNP-based test to produce the integrative gene association score:

$$g_i = -2\sum_{j=1}^{K} \log_e(p_{ij}),$$

where K is the number of independent tests, in this case $K = 2$, namely, expression-based test and SNP-based test, and p_{ij} is the *P*-value for gene i in test j.

Rank sum

For each differential expression score or SNP set association score, we first generate its null distribution by a phenotype-based permutation procedure, and then we transform every score and its corresponding null scores into ranks. Tied values are assigned the average of the applicable ranks. For example, (2, 5, 6, 5) is ranked as (1, 2.5, 4, 2.5). Gene association scores are then computed as

$$g_i = r_{ei} + r_{si},$$

where r_{ei} and r_{si} are the ranks of gene i in the expression-based test and SNP-based test, respectively.

Gene set association test

Given the gene association scores, we use a weighted Kolmogorov-Smirnov (K-S) test to determine which gene sets have the greatest combined evidence for association with the given phenotype. Essentially, the weighted K-S test determines for each gene set whether the genes belonging to that gene set are preferentially near the top of the ranked ordered list based on gene association scores. More formally, given a particular gene set S including H genes and the rank ordered gene association scores $\{g_1, \ldots, g_N\}$ for

all genes in the expression data set, a running association score $RAS_S(i)$ for the rank ordered genes in positions $i = 1, \ldots, N$ is computed as

$$RAS_S(i) = \frac{1}{N_S} \sum_{j=1}^{i} |g_j| I(j \in S) - \frac{1}{N - H} \sum_{j=1}^{i} I(j \notin S),$$

$$N_S = \sum_{j=1}^{N} |g_j| I(j \in S),$$

where $I(j \in S)$ is an indicator variable that is one if the jth gene in the rank ordered list is in gene set S and is otherwise zero. Similarly, $I(j \notin S)$ takes the value of zero if the jth gene is in the gene set and is otherwise one. The gene set association score, $AS(S)$, is the maximum deviation from zero of the running association score over the positions $i = 1, \ldots, N$

$$AS(S)_+ = \max_{i=1,\ldots,N} [RAS_S(i)],\ AS(S)_- = \min_{i=1,\ldots,N} [RAS_S(i)].$$

Finally, if $|AS(S)_+| > |AS(S)_-|$ then the final gene set association score $AS(S) = AS(S)_+$, otherwise, $AS(S) = AS(S)_-$.

The gene association scores we used lack directionality, so a negative $AS(S)$ means that there is no association between the gene set and the phenotype. We here set $AS(S) = 0.0001$ if $AS(S) < 0$ so that negative AS scores will not confuse the following assignment of the direction. One advantage for the standard GSEA analysis is that its association score suggests the direction of an association. In the K-S test used by GSAA to calculate the integrative gene set association score, we aim to capture both up-regulated and down-regulated genes in a gene set, so we do not assign directionality at this point. Instead, we perform an additional K-S test based solely on the directed differential gene expression scores to get a corresponding expression-based association score (EAS) for each gene set. We impose directionality on the integrative AS based on the sign of the EAS for the same gene set, $AS(S) = AS(S) \times \text{sign}(EAS(S))$.

In GSAA, we integrate gene expression information and genotype information at the gene level. However, some genes may not have associated SNPs. For these genes, the gene association score is just derived from the expression-based test. To account for possible heterogeneity of information at each gene locus, we standardize the original gene association score by the mean and standard deviations of its null distribution using the same method as we used to calculate the standard expression score before performing the K-S test.

The absolute magnitude of the AS score indicates the strength of the association between the gene set and the phenotype, and the sign indicates which phenotypic class the gene set is associated with. Finally, a normalized association score (NAS) for each gene set is calculated to adjust for difference in gene set size. Similar to GSEA, we use a mean-based method and normalize the positive and negative scores separately.

Assessment of statistical significance and adjustment for multiple hypothesis testing

We assess the statistical significance of the gene set association score and adjust for multiple hypothesis testing based on a phenotype-based permutation procedure. This procedure preserves LD structure in SNP data and gene–gene correlation structure in gene expression data. A nominal P-value is calculated relative to a null distribution generated by shuffling the phenotypic class labels and recalculating the gene set association score many times. If the gene expression and SNP data come from the same samples, i.e., from matched data, GSAA will perform better. Since it may be difficult to obtain matched genomic data and to be able to use GSAA on existing GWA and gene expression data that may not be matched, we designed GSAA to allow for both matched and unmatched data. When the data are matched, permutations for the expression-based test and SNP-based test are not independent, and GSAA uses the same permutation template for both. This can result in greater power to identify real associations.

We use the false discovery rate (FDR) and the family-wise error rate (FWER) based on the normalized gene set association scores to correct for multiple hypothesis testing and to control the proportion of false positives below a certain threshold. Given m gene sets $\{S_1, \ldots, S_m\}$ and label permutations $\pi = 1, \ldots, \Pi$, the FDR for each gene set S_i with $NAS(S_i) \geq 0$ is computed as

$$FDR(S_i) = \frac{\%\ \text{of}\ NAS(S_j, \pi)_+ \geq NAS(S_i)\ \text{for}\ j = 1, \ldots, m\ \text{and}\ \pi = 1, \ldots, \Pi}{\%\ \text{of}\ NAS(S_j)_+ \geq NAS(S_i)\ \text{for}\ j = 1, \ldots, m},$$

If $NAS(S_i) < 0$, the FDR is computed as

$$FDR(S_i) = \frac{\%\ \text{of}\ NAS(S_j, \pi)_- \leq NAS(S_i)\ \text{for}\ j = 1, \ldots, m\ \text{and}\ \pi = 1, \ldots, \Pi}{\%\ \text{of}\ NAS(S_j)_- \leq NAS(S_i)\ \text{for}\ j = 1, \ldots, m},$$

where $NAS(S_j, \pi)$ is the normalized association score for gene set j with label permutation π. $NAS(S_j, \pi)_+$ and $NAS(S_j, \pi)_-$ denote positive and negative $NAS(S_j, \pi)$, respectively. $NAS(S_j)$ is the normalized association score for gene set j. $NAS(S_j)_+$, $NAS(S_j)_-$ denote positive and negative $NAS(S_j)$, respectively.

The FWER for a gene set S_i with $NAS(S_i) \geq 0$ is computed as

$$FWER(S_i) = \%\ \text{of}\ [\max_{j=1,\ldots,m} [\text{NAS}(S_j, \pi)_+]] \geq \text{NAS}(S_i)\ \text{for}\ \pi = 1, \ldots, \Pi.$$

If $NAS(S_i) < 0$, the FDR is computed as

$$FWER(S_i) = \%\ \text{of}\ [\min_{j=1,\ldots,m} [\text{NAS}(S_j, \pi)_-]] \leq \text{NAS}(S_i)\ \text{for}\ \pi = 1, \ldots, \Pi.$$

Computational efficiency of GSAA

With respect to computational efficiency, GSAA took ~0.9 h and 4 h for the analyses of glioblastoma data and Crohn's disease data, respectively, using one computational node with eight processors (Intel Xeon CPU E5520 @ 2.27 GHz). It only took ~3.5 min for the simulated data sets.

Gene set association analysis–SNP (GSAA–SNP)

GSAA-SNP was created to perform gene set association analysis based solely on SNP data. In GSAA-SNP, we remove the module of the differential expression test in GSAAzs-ks and use the original SNP set association score as the gene association score. Otherwise, it is the same as GSAAzs-ks.

Gene set enrichment analysis based on non–directional differential expression scores (GSEAndes)

Both GSAA and GSAA-SNP are based on the non-directional association analysis at the gene level. To compare them with the GSEA more fairly, we created GSEAndes. GSEAndes is an extension of the original GSEA software and also an expression-based version of GSAAzs-ks. In GSEAndes, we remove the two modules of single-SNP association test and SNP set association test in GSAAzs-ks and use the original differential expression score as the gene association score. Otherwise, it is the same as GSAAzs-ks.

Generation of simulated data

We generated simulated gene expression data and SNP genotype data to study the power of various integrative methods and single-source methods. Modeling a case-control setting, we simulated 200 cases and 200 controls for each data set.

Gene set data

For each simulation we generated 100 gene sets. Only the first gene set (causal gene set) included risk genes. We randomly chose a pathway, P53PATHWAY, that contains 16 genes from the Molecular Signatures Database (MSigDB, http://www.broadinstitute.org/gsea/msigdb/index.jsp) as a prototype to simulate the causal gene set. The gene expression and genotype information of P53PATHWAY were obtained from the glioblastoma data generated through The Cancer Genome Atlas (TCGA, http://cancergenome.nih.gov) project. The remaining 99 gene sets were simulated from null models, namely, none of the genes in these gene sets were associated with the phenotype of interest with respect to gene expression profiles or genotypes. The sizes of null gene sets were randomly drawn from U[15, 30]. Genes within null gene sets were randomly drawn from a pool of 984 non-causal genes.

SNP data

Each simulated SNP data set included 1000 genes, each gene with one genotyped SNP for a total of 1000 SNPs. Some of these SNPs were considered causally related to the phenotype of interest. We simulated the causal SNPs in the causal gene set based on the genotype information of P53PATHWAY. We first assigned SNPs that were within the region 1 kb upstream of the transcription start site (TSS) to the end of the transcribed bases to each gene in the P53PATHWAY, then we removed SNPs with minor allele frequency (MAF) <0.05 and chose the SNP with the highest score in the χ^2 test as the tag SNP of the gene. We set the allele frequencies of causal SNPs in the simulated causal gene set the same as the allele frequencies of corresponding tag SNPs in the P53PATHWAY. The heterozygote odds ratio for each causal SNP was generated from U[1.1, 1.3] and U[1.2, 1.4], respectively. We used an additive disease model for the causal loci, and the disease prevalence was set to 0.02. We drew allele frequencies from a Beta distribution, Beta(0.1, 0.1), for null SNPs with no association with the phenotype based on the approximation for the unconditional distribution of allele frequencies in the HapMap populations stated in Coram and Tang (2007). Based on these parameter settings, the genotype data were generated by PLINK (Purcell et al. 2007) (http://pngu.mgh.harvard.edu/purcell/plink/). We then assigned the case-control status based on the model

$$\text{logit}\{\Pr(Y_j = 1)\} = \sum_{i=1}^{N} x_{G_{ji}}\beta_i + e_j,$$

where N is the number of causal SNPs in the causal gene set. $x_{G_{ji}}$ denotes the coding of the genotype at causal SNP i for sample j with effect size β_i that is the log-odds ratio at SNP i. e_j denotes a random sample-specific error term for sample j; e_j is sampled from a standard normal distribution.

Gene expression data

Each simulated gene expression data set consisted of 1000 genes corresponding to the 1000 genes in the SNP data set. Some of these genes were considered risk genes that were differentially expressed in cases and controls. We first generated baseline expression levels for genes in the causal gene set from a multivariate normal distribution $X \sim N(\mu, \Sigma)$. The mean vector μ and the covariance matrix $\sum$ were estimated from the P53PATHWAY based on the glioblastoma data. Next, we added disease effect to the causal genes in the causal gene set based on the model $x_{ji} = x_0(1 + x_{G_{ji}}\beta)$, where x_{ji} is the expression level of gene i in sample j, x_0 is the baseline expression level of gene i in sample j, and $x_{G_{ji}}$ denotes the coding of the genotype at SNP i for sample j in the SNP data. β is the effect size of the genotype on gene expression and reflects the degree to which the gene expression is correlated with the genotype of tag SNP of the same gene. β was drawn from either U[0.5, 0.8] or U[−0.5, −0.8]. The sign of β indicates up- or down-regulation of the gene. In our simulation, gene expression variations of a causal gene in the expression data were determined by the genotypes of the same gene in the SNP data. However, it is not realistic that all causal genes contain both gene expression variation and genotypic variation associated with the phenotype. To address this issue, we also simulated scenarios in which some causal genes contain only gene expression variation and others include just genotypic variation. In the former case, we first simulated a causal SNP in the SNP data set and then simulated a causal gene in the expression data set based on this causal SNP. Finally, we replaced this SNP with a null SNP. In the latter case, we only added disease effect to part of the causal genes in the expression data set. Gene expression values for null genes were also drawn from a multivariate normal distribution, $X_0 \sim N(\mu_0, \Sigma_0)$. We estimated the average values of means and variances of all genes in glioblastoma data and use these average values to set μ_0 and $\sum_0$.

Glioblastoma and Crohn's disease data

Data generated through The Cancer Genome Atlas (TCGA, http://cancergenome.nih.gov) project for glioblastoma samples were obtained through their data portal. The expression data set includes 258 tumor samples and 10 normal samples. The SNP data set includes 205 tumor samples and 89 normal samples. For Crohn's disease (CD), expression data were generated by Wu et al. (2007) and are available in the NCBI Gene Expression Omnibus (GEO, http://www.ncbi.nlm.nih.gov/geo, GSE6731). The expression data set we used contains 23 samples with seven cases versus 16 controls. Cases were obtained from biopsies from affected regions of colons of CD patients. Controls were derived from biopsies from unaffected regions of colons of CD patients and from the colons of healthy adults. The SNP data set includes 1748 cases and 2938 controls obtained from a published large-scale GWA study (Wellcome Trust Case Control Consortium 2007), available from the Wellcome Trust Case Control Consortium (WTCCC, https://www.wtccc.org.uk/info/access_to_data_samples.shtml).

We imputed missing SNP data using fastPHASE 1.4.0 (Scheet and Stephens 2006) (http://depts.washington.edu/uwc4c/express-licenses/assets/fastphase). We assigned SNPs that were within the region 1 kb upstream of the TSS to the end of the transcribed bases to be associated with a gene. Although 1 kb upstream of the TSS may be insufficient to cover the entire regulatory region for all genes, it should include both the core and proximal promoter regions and at least some of the distal regulatory elements (Bortoluzzi et al. 2005). The canonical pathways from the Molecular Signatures Database (MSigDB, http://www.broadinstitute.org/gsea/msigdb/index.jsp) were used in this analysis. Pathways with less than 15 genes or more than 100 genes in the expression data set were filtered to avoid overly narrow or broad functional categories. This resulted in 357 canonical pathways for glioblastoma data and 352 canonical pathways for the Crohn's disease data. Data for genes not contained in any of the gene sets were filtered prior to performing GSAA analysis since these data would not affect the gene set association analysis. We assessed the statistical significance of association scores of gene sets and adjusted for multiple hypothesis testing using 10,000 permutations of phenotypic class labels.

Data access

GSAA software is freely available at http://gsaa.unc.edu.

Acknowledgments

We thank Xuejun Qin (Duke University) for discussions on simulations. We thank Nianjun Liu (University of Alabama at Birmingham) for discussions on identifying the causal variants. We thank Jenny Tung (The University of Chicago) for discussions on trait mapping.

We thank the GSEA team (Broad Institute) for providing GSEA software, code, and documentation. We thank The Cancer Genome Atlas (TCGA) and Wellcome Trust Case Control Consortium (WTCCC) for granting access to the raw genotype and phenotype data. We gratefully acknowledge funding from the Duke Comprehensive Cancer Center (Q.X.), Duke Institute for Genome Sciences & Policy (Q.X., S.M., T.S.F.), NIH grant 1RC1CA146849 (Q.X., T.S.F.), The University Cancer Research Fund at The University of North Carolina at Chapel Hill (Q.X., T.S.F.), NIH MH059528 (E.R.H., T.S.F.), CA123175-01A1 (S.M.), NIH Systems Biology Center Grant (S.M.), NSF grant DMS-0732260 (S.M.), NSF grant CCF-1049290 (S.M.), and R01 CA125618-01 (S.M).

References

Abatangelo L, Maglietta R, Distaso A, D'Addabbo A, Creanza TM, Mukherjee S, Ancona N. 2009. Comparative study of gene set enrichment methods. *BMC Bioinformatics* **10:** 275. doi: 10.1186/1471-2105-10-275.

Allison DB, Cui X, Page GP, Sabripour M. 2006. Microarray data analysis: From disarray to consolidation and consensus. *Nat Rev Genet* **7:** 55–65.

Asai A, Qiu J, Narita Y, Chi S, Saito N, Shinoura N, Hamada H, Kuchino Y, Kirino T. 1999. High level calcineurin activity predisposes neuronal cells to apoptosis. *J Biol Chem* **274:** 34450–34458.

Atreya I, Atreya R, Neurath MF. 2008. NF-κB in inflammatory bowel disease. *J Intern Med* **263:** 591–596.

Baksh S, DeCaprio JA, Burakoff SJ. 2000. Calcineurin regulation of the mammalian G_0/G_1 checkpoint element, cyclin dependent kinase 4. *Oncogene* **19:** 2820–2827.

Baranzini SE, Galwey NW, Wang J, Khankhanian P, Lindberg R, Pelletier D, Wu W, Uitdehaag BM, Kappos L, Polman CH, et al. 2009. Pathway and network-based analysis of genome-wide association studies in multiple sclerosis. *Hum Mol Genet* **18:** 2078–2090.

Bartkova J, Horejsi Z, Koed K, Kramer A, Tort F, Zieger K, Guldberg P, Sehested M, Nesland JM, Lukas C, et al. 2005. DNA damage response as a candidate anti-cancer barrier in early human tumorigenesis. *Nature* **434:** 864–870.

Benacerraf B. 1981. Role of MHC gene products in immune regulation. *Science* **212:** 1229–1238.

Ben-Neriah Y, Schmidt-Supprian M. 2007. Epithelial NF-κB maintains host gut microflora homeostasis. *Nat Immunol* **8:** 479–481.

Bleau AM, Huse JT, Holland EC. 2009. The ABCG2 resistance network of glioblastoma. *Cell Cycle* **8:** 2936–2944.

Boorsma A, Foat BC, Vis D, Klis F, Bussemaker HJ. 2005. T-profiler: Scoring the activity of predefined groups of genes using gene expression data. *Nucleic Acids Res* **33:** W592–W595.

Bortoluzzi S, Coppe A, Bisognin A, Pizzi C, Danieli GA. 2005. A multistep bioinformatic approach detects putative regulatory elements in gene promoters. *BMC Bioinformatics* **6:** 121. doi: 10.1186/1471-2105-6-121.

Bronger H, Konig J, Kopplow K, Steiner HH, Ahmadi R, Herold-Mende C, Keppler D, Nies AT. 2005. ABCC drug efflux pumps and organic anion uptake transporters in human gliomas and the blood–tumor barrier. *Cancer Res* **65:** 11419–11428.

Cancer Genome Atlas Research Network. 2008. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* **455:** 1061–1068.

Carnero A, Blanco-Aparicio C, Renner O, Link W, Leal JF. 2008. The PTEN/PI3K/AKT signalling pathway in cancer, therapeutic implications. *Curr Cancer Drug Targets* **8:** 187–198.

Chen LS, Hutter CM, Potter JD, Liu Y, Prentice RL, Peters U, Hsu L. 2010. Insights into colon cancer etiology via a regularized approach to gene set analysis of GWAS data. *Am J Hum Genet* **86:** 860–871.

Coram M, Tang H. 2007. Improving population-specific allele frequency estimates by adapting supplemental data: an empirical Bayes approach. *Ann Appl Stat* **1:** 459–479.

De la Cruz O, Wen X, Ke B, Song M, Nicolae DL. 2010. Gene, region and pathway level analyses in whole-genome studies. *Genet Epidemiol* **34:** 222–231.

Dinu I, Potter JD, Mueller T, Liu Q, Adewale AJ, Jhangri GS, Einecke G, Famulski KS, Halloran P, Yasui Y. 2007. Improving gene set analysis of microarray data by SAM-GS. *BMC Bioinformatics* **8:** 242. doi: 10.1186/1471-2105-8-242.

Fernando MM, Stevens CR, Walsh EC, De Jager PL, Goyette P, Plenge RM, Vyse TJ, Rioux JD. 2008. Defining the role of the MHC in autoimmunity: A review and pooled analysis. *PLoS Genet* **4:** e1000024. doi: 10.1371/journal.pgen.1000024.

Fisher RA. 1932. *Statistical methods for research workers*. Oliver and Boyd, Edinburgh.

Forcione DG, Sands B, Isselbacher KJ, Rustgi A, Podolsky DK, Pillai S. 1996. An increased risk of Crohn's disease in individuals who inherit the HLA class II *DRB3*0301* allele. *Proc Natl Acad Sci* **93:** 5094–5098.

Fukao T, Koyasu S. 2003. PI3K and negative regulation of TLR signaling. *Trends Immunol* **24:** 358–363.

Goeman JJ, Buhlmann P. 2007. Analyzing gene expression data in terms of gene sets: Methodological issues. *Bioinformatics* **23:** 980–987.

Goeman JJ, van de Geer SA, de Kort F, van Houwelingen HC. 2004. A global test for groups of genes: Testing association with a clinical outcome. *Bioinformatics* **20:** 93–99.

Gorgoulis VG, Vassiliou LV, Karakaidos P, Zacharatos P, Kotsinas A, Liloglou T, Venere M, Ditullio RA Jr, Kastrinakis NG, Levy B, et al. 2005. Activation of the DNA damage checkpoint and genomic instability in human precancerous lesions. *Nature* **434:** 907–913.

Gorlov IP, Gorlova OY, Sunyaev SR, Spitz MR, Amos CI. 2008. Shifting paradigm of association studies: Value of rare single-nucleotide polymorphisms. *Am J Hum Genet* **82:** 100–112.

Gorlov IP, Gallick GE, Gorlova OY, Amos C, Logothetis CJ. 2009. GWAS meets microarray: Are the results of genome-wide association studies and gene-expression profiling consistent? Prostate cancer as an example. *PLoS ONE* **4:** e6511. doi: 10.1371/journal.pone.0006511.

Guo H, Liu D, Gelbard H, Cheng T, Insalaco R, Fernandez JA, Griffin JH, Zlokovic BV. 2004. Activated protein C prevents neuronal apoptosis via protease activated receptors 1 and 3. *Neuron* **41:** 563–572.

Halazonetis TD, Gorgoulis VG, Bartek J. 2008. An oncogene-induced DNA damage model for cancer development. *Science* **319:** 1352–1355.

Harris SL, Levine AJ. 2005. The p53 pathway: Positive and negative feedback loops. *Oncogene* **24:** 2899–2908.

Hayashi T, Faustman D. 2000. Essential role of human leukocyte antigen-encoded proteasome subunits in NF-κB activation and prevention of tumor necrosis factor-α-induced apoptosis. *J Biol Chem* **275:** 5238–5247.

Heimberger AB, Suki D, Yang D, Shi W, Aldape K. 2005. The natural history of EGFR and EGFRvIII in glioblastoma patients. *J Transl Med* **3:** 38. doi: 10.1186/1479-5876-3-38.

Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA. 2009. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci* **106:** 9362–9367.

Holden M, Deng S, Wojnowski L, Kulle B. 2008. GSEA-SNP: Applying gene set enrichment analysis to SNP data from genome-wide association studies. *Bioinformatics* **24:** 2784–2785.

Ina K, Itoh J, Fukushima K, Kusugami K, Yamaguchi T, Kyokane K, Imada A, Binion DG, Musso A, West GA, et al. 1999. Resistance of Crohn's disease T cells to multiple apoptotic signals is associated with a Bcl-2/Bax mucosal imbalance. *J Immunol* **163:** 1081–1090.

Irving PM, Gibson PR. 2008. Infections and IBD. *Nat Clin Pract Gastroenterol Hepatol* **5:** 18–27.

Junge CE, Lee CJ, Hubbard KB, Zhang Z, Olson JJ, Hepler JR, Brat DJ, Traynelis SF. 2004. Protease-activated receptor-1 in human brain: Localization and functional expression in astrocytes. *Exp Neurol* **188:** 94–103.

Kahl CR, Means AR. 2003. Regulation of cell cycle progression by calcium/calmodulin-dependent pathways. *Endocr Rev* **24:** 719–736.

Kapsoritakis AN, Potamianos SP, Sfiridaki AI, Koukourakis MI, Koutroubakis IE, Roussomoustakaki MI, Manousos ON, Kouroumalis EA. 2000. Elevated thrombopoietin serum levels in patients with inflammatory bowel disease. *Am J Gastroenterol* **95:** 3478–3481.

Kastan MB, Bartek J. 2004. Cell-cycle checkpoints and cancer. *Nature* **432:** 316–323.

Khatri P, Draghici S. 2005. Ontological analysis of gene expression data: Current tools, limitations, and open problems. *Bioinformatics* **21:** 3587–3595.

Kim SY, Volsky DJ. 2005. PAGE: Parametric analysis of gene set enrichment. *BMC Bioinformatics* **6:** 144. doi: 10.1186/1471-2105-6-144.

Kryukov GV, Pennacchio LA, Sunyaev SR. 2007. Most rare missense alleles are deleterious in humans: Implications for complex disease and association studies. *Am J Hum Genet* **80:** 727–739.

Liu Q, Dinu I, Adewale AJ, Potter JD, Yasui Y. 2007. Comparative evaluation of gene-set analysis methods. *BMC Bioinformatics* **8:** 431. doi: 10.1186/1471-2105-8-431.

Liu YJ, Guo YF, Zhang LS, Pei YF, Yu N, Yu P, Papasian CJ, Deng HW. 2010. Biological pathway-based genome-wide association analysis identified the vasoactive intestinal peptide (VIP) pathway important for obesity. *Obesity (Silver Spring)* **18:** 2339–2346.

Lombardi ML, Pirozzi G, Luongo V, Mercuro O, Pace E, Blanco Del Vecchio G, Cozzolino A, Errico S, Fusco C, Castiglione F. 2001. Crohn disease: Susceptibility and disease heterogeneity revealed by HLA genotyping. *Hum Immunol* **62:** 701–704.

Luo W, Friedman MS, Shedden K, Hankenson KD, Woolf PJ. 2009. GAGE: Generally applicable gene set enrichment for pathway analysis. *BMC Bioinformatics* **10:** 161. doi: 10.1186/1471-2105-10-161.

Lyman GH, Khorana AA. 2009. Cancer, clots and consensus: New understanding of an old problem. *J Clin Oncol* **27:** 4821–4826.
Maglietta R, Piepoli A, Catalano D, Licciulli F, Carella M, Liuni S, Pesole G, Perri F, Ancona N. 2007. Statistical assessment of functional categories of genes deregulated in pathological conditions by using microarray data. *Bioinformatics* **23:** 2063–2072.
Magnus N, Garnier D, Rak J. 2010. Oncogenic epidermal growth factor receptor up-regulates multiple elements of the tissue factor signaling pathway in human glioma cells. *Blood* **116:** 815–818.
Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, et al. 2009. Finding the missing heritability of complex diseases. *Nature* **461:** 747–753.
Mansmann U, Meister R. 2005. Testing differential gene expression in functional groups. Goeman's global test versus an ANCOVA approach. *Methods Inf Med* **44:** 449–453.
Mattson MP, Meffert MK. 2006. Roles for NF-κB in nerve cell survival, plasticity, and disease. *Cell Death Differ* **13:** 852–860.
Nenci A, Becker C, Wullaert A, Gareus R, van Loo G, Danese S, Huth M, Nikolaev A, Neufert C, Madison B, et al. 2007. Epithelial NEMO links innate immunity to chronic intestinal inflammation. *Nature* **446:** 557–561.
Newman B, Silverberg MS, Gu X, Zhang Q, Lazaro A, Steinhart AH, Greenberg GR, Griffiths AM, McLeod RS, Cohen Z, et al. 2004. CARD15 and HLA DRB1 alleles influence susceptibility and disease localization in Crohn's disease. *Am J Gastroenterol* **99:** 306–315.
Newton MA, Quintana FA, Den Boon JA, Sengupta S, Ahlquist P. 2007. Random-set methods identify distinct aspects of the enrichment signal in gene-set analysis. *Ann Appl Stat* **1:** 85–106.
Nicolae DL, Gamazon E, Zhang W, Duan S, Dolan ME, Cox NJ. 2010. Trait-associated SNPs are more likely to be eQTLs: Annotation to enhance discovery from GWAS. *PLoS Genet* **6:** e1000888. doi: 10.1371/journal.pgen.1000888.
O'Dushlaine C, Kenny E, Heron EA, Segurado R, Gill M, Morris DW, Corvin A. 2009. The SNP ratio test: Pathway analysis of genome-wide association datasets. *Bioinformatics* **25:** 2762–2763.
Okkenhaug K, Fruman DA. 2010. PI3Ks in lymphocyte signaling and development. *Curr Top Microbiol Immunol* **346:** 57–85.
Okkenhaug K, Vanhaesebroeck B. 2003. PI3K in lymphocyte development, differentiation and activation. *Nat Rev Immunol* **3:** 317–330.
Ooi CH, Ivanova T, Wu J, Lee M, Tan IB, Tao J, Ward L, Koo JH, Gopalakrishnan V, Zhu Y, et al. 2009. Oncogenic pathway combinations predict clinical prognosis in gastric cancer. *PLoS Genet* **5:** e1000676. doi: 10.1371/journal.pgen.1000676.
Osaki M, Oshimura M, Ito H. 2004. PI3K-Akt pathway: Its functions and alterations in human cancer. *Apoptosis* **9:** 667–676.
Pallone F, Monteleone G. 2001. Mechanisms of tissue damage in inflammatory bowel disease. *Curr Opin Gastroenterol* **17:** 307–312.
Peng G, Luo L, Siu H, Zhu Y, Hu P, Hong S, Zhao J, Zhou X, Reveille JD, Jin L, et al. 2010. Gene and pathway-based second-wave analysis of genome-wide association studies. *Eur J Hum Genet* **18:** 111–117.
Plate KH, Breier G, Weich HA, Mennel HD, Risau W. 1994. Vascular endothelial growth factor and glioma angiogenesis: Coordinate induction of VEGF receptors, distribution of VEGF protein and possible in vivo regulatory mechanisms. *Int J Cancer* **59:** 520–529.
Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, et al. 2007. PLINK: A tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* **81:** 559–575.
Reinshagen M, Loeliger C, Kuehnl P, Weiss U, Manfras BJ, Adler G, Boehm BO. 1996. HLA class II gene frequencies in Crohn's disease: A population based analysis in Germany. *Gut* **38:** 538–542.
Reynes G, Vila V, Martin M, Parada A, Fleitas T, Reganon E, Martinez-Sales V. 2010. Circulating markers of angiogenesis, inflammation, and coagulation in patients with glioblastoma. *J Neurooncol* **102:** 35–41.
Robinson R. 2010. Common disease, multiple rare (and distant) variants. *PLoS Biol* **8:** e1000293. doi: 10.1371/journal.pbio.1000293.
Saharinen P, Eklund L, Pulkki K, Bono P, Alitalo K. 2011. VEGF and angiopoietin signaling in tumor angiogenesis and metastasis. *Trends Mol Med* **17:** 347–362.
Scheet P, Stephens M. 2006. A fast and flexible statistical model for large-scale population genotype data: Applications to inferring missing genotypes and haplotypic phase. *Am J Hum Genet* **78:** 629–644.
Spehlmann ME, Eckmann L. 2009. Nuclear factor-κB in intestinal protection and destruction. *Curr Opin Gastroenterol* **25:** 92–99.
Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, et al. 2005. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci* **102:** 15545–15550.
Venkitaraman AR. 2005. Medicine: Aborting the birth of cancer. *Nature* **434:** 829–830.
Visekruna A, Joeris T, Seidel D, Kroesen A, Loddenkemper C, Zeitz M, Kaufmann SH, Schmidt-Ullrich R, Steinhoff U. 2006. Proteasome-mediated degradation of IκBα and processing of p105 in Crohn disease and ulcerative colitis. *J Clin Invest* **116:** 3195–3203.
Visekruna A, Joeris T, Schmidt N, Lawrenz M, Ritz JP, Buhr HJ, Steinhoff U. 2009a. Comparative expression analysis and characterization of 20S proteasomes in human intestinal tissues: The proteasome pattern as diagnostic tool for IBD patients. *Inflamm Bowel Dis* **15:** 526–533.
Visekruna A, Slavova N, Dullat S, Grone J, Kroesen AJ, Ritz JP, Buhr HJ, Steinhoff U. 2009b. Expression of catalytic proteasome subunits in the gut of patients with Crohn's disease. *Int J Colorectal Dis* **24:** 1133–1139.
Wang HG, Pathan N, Ethell IM, Krajewski S, Yamaguchi Y, Shibasaki F, McKeon F, Bobo T, Franke TF, Reed JC. 1999. Ca^{2+}-induced apoptosis through calcineurin dephosphorylation of BAD. *Science* **284:** 339–343.
Wang WY, Barratt BJ, Clayton DG, Todd JA. 2005. Genome-wide association studies: Theoretical and practical concerns. *Nat Rev Genet* **6:** 109–118.
Wang K, Li M, Bucan M. 2007. Pathway-based approaches for analysis of genomewide association studies. *Am J Hum Genet* **81:** 1278–1283.
Wang K, Baldassano R, Zhang H, Qu HQ , Imielinski M, Kugathasan S, Annese V, Dubinsky M, Rotter JI, Russell RK, et al. 2010. Comparative genetic analysis of inflammatory bowel disease and type 1 diabetes implicates multiple loci with opposite effects. *Hum Mol Genet* **19:** 2059–2067.
Wellcome Trust Case Control Consortium. 2007. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447:** 661–678.
Westfall PH, Young SS. 1993. *Resampling-based multiple testing: Examples and methods for p-value adjustment*. Wiley, New York.
Whitlock MC. 2005. Combining probability from independent tests: The weighted Z-method is superior to Fisher's approach. *J Evol Biol* **18:** 1368–1373.
Wu F, Dassopoulos T, Cope L, Maitra A, Brant SR, Harris ML, Bayless TM, Parmigiani G, Chakravarti S. 2007. Genome-wide gene expression differences in Crohn's disease and ulcerative colitis from endoscopic pinch biopsies: Insights into distinctive pathogenesis. *Inflamm Bowel Dis* **13:** 807–821.
Yewdell JW. 2005. Immunoproteasomes: Regulating the regulator. *Proc Natl Acad Sci* **102:** 9089–9090.
Zaph C, Troy AE, Taylor BC, Berman-Booty LD, Guild KJ, Du Y, Yost EA, Gruber AD, May MJ, Greten FR, et al. 2007. Epithelial-cell-intrinsic IKK-β expression regulates intestinal immune homeostasis. *Nature* **446:** 552–556.
Zhang L, Guo YF, Liu YZ, Liu YJ, Xiong DH, Liu XG, Wang L, Yang TL, Lei SF, Guo Y, et al. 2010. Pathway-based genome-wide association analysis identified the importance of regulation-of-autophagy pathway for ultradistal radius BMD. *J Bone Miner Res* **25:** 1572–1580.
Zhao L, Lee JY, Hwang DH. 2008. The phosphatidylinositol 3-kinase/Akt pathway negatively regulates Nod2-mediated NF-κB pathway. *Biochem Pharmacol* **75:** 1515–1525.
Zhong H, Yang X, Kaplan LM, Molony C, Schadt EE. 2010. Integrating pathway analysis and genetics of gene expression for genome-wide association studies. *Am J Hum Genet* **86:** 581–591.

Received April 13, 2011; accepted in revised form September 19, 2011.