Identifying the Genetic Variation of Gene Expression Using Gene Sets: Application of Novel Gene Set eQTL Approach to PharmGKB and KEGG

Ryan Abo[1], **Gregory D. Jenkins**[2], **Liewei Wang**[1], **Brooke L. Fridley**[2]*

1 Division of Clinical Pharmacology, Department of Molecular Pharmacology and Experimental Therapeutics, Mayo Clinic, Rochester, Minnesota, United States of America, 2 Department of Health Sciences Research, Mayo Clinic, Rochester, Minnesota, United States of America

Abstract

Genetic variation underlying the regulation of mRNA gene expression in humans may provide key insights into the molecular mechanisms of human traits and complex diseases. Current statistical methods to map genetic variation associated with mRNA gene expression have typically applied standard linkage and/or association methods; however, when genome-wide SNP and mRNA expression data are available performing all pair wise comparisons is computationally burdensome and may not provide optimal power to detect associations. Consideration of different approaches to account for the high dimensionality and multiple testing issues may provide increased efficiency and statistical power. Here we present a novel approach to model and test the association between genetic variation and mRNA gene expression levels in the context of gene sets (GSs) and pathways, referred to as gene set – expression quantitative trait loci analysis (GS-eQTL). The method uses GSs to initially group SNPs and mRNA expression, followed by the application of principal components analysis (PCA) to collapse the variation and reduce the dimensionality within the GSs. We applied GS-eQTL to assess the association between SNP and mRNA expression level data collected from a cell-based model system using PharmGKB and KEGG defined GSs. We observed a large number of significant GS-eQTL associations, in which the most significant associations arose between genetic variation and mRNA expression from the same GS. However, a number of associations involving genetic variation and mRNA expression from different GSs were also identified. Our proposed GS-eQTL method effectively addresses the multiple testing limitations in eQTL studies and provides biological context for SNP-expression associations.

Citation: Abo R, Jenkins GD, Wang L, Fridley BL (2012) Identifying the Genetic Variation of Gene Expression Using Gene Sets: Application of Novel Gene Set eQTL Approach to PharmGKB and KEGG. PLoS ONE 7(8): e43301. doi:10.1371/journal.pone.0043301

Editor: Zhaoxia Yu, University of California Irvine, United States of America

Received April 5, 2012; **Accepted** July 19, 2012; **Published** August 14, 2012

Funding: This research was supported by the National Institutes of Health (NIH grants GM61388, CA140879, GM86689, CA130828, CA138461, CA102701), the Minnesota Partnership, and the Mayo Foundation. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: fridley.brooke@mayo.edu

Introduction

Establishing genetic variation (e.g., single nucleotide polymorphisms (SNPs)) associated with variation in mRNA gene expression is a key component to further understand the molecular basis of human traits, including complex disease and response to drug therapies. The genetics of human mRNA expression level has been extensively studied and many mRNA expression regulatory loci or expression quantitative trait loci (eQTL) have been identified using a variety approaches, often based on the use of cell line model systems [1]. Yet, additional research in this area is needed to fully characterize and understand the mechanisms by which eQTLs regulate mRNA gene expression. A basic understanding regarding the locations of eQTLs relative to the genes in which they regulate has been established. A *cis*-acting eQTL, or *cis*-eQTL, describes a DNA sequence variant located within or outside the gene transcription unit up to a couple mega-bases away [2,3], while *trans*-acting eQTLs, or *trans*-eQTLs, are considered to be located much further from the associated transcription unit. *Trans*-eQTLs that are associated with many mRNA gene expressions are termed "hotspots" or "master regulators", and are presumed to influence many biological functions [4]. Mapping eQTLs in humans could help to identify the functional loci contributing to variation in human traits and has been applied to the study of many complex traits, such as asthma [5], type 2 diabetes [6], adult height [7], Crohn's disease [8], and celiac disease [9].

Identification of eQTLs in humans has been performed using analytical methods previously developed for disease-risk genetic studies by treating each mRNA gene expression level as a quantitative trait with linkage analysis methods for family-based data [10–13] and association analysis methods for unrelated individuals [14–17]. More recently, the rapid development and cost reduction of genomic arrays to capture genome-wide single nucleotide polymorphism (SNP) and mRNA expression data have resulted in the use of genome-wide association (GWA) analyses using independent samples [18–22]. The eQTL mapping approach with genome-wide data involves assessing the association between all possible SNP-expression pairs. These eQTL GWA studies have resulted in a large number of expression associated SNPS (eSNPs) [21,23,24]. The success of eQTL association mapping methods as compared to disease-risk studies may be due to the strength of eQTL signals and lack of phenotype heterogeneity; however, there are much greater multiple testing issues to consider with eQTL association mapping and therefore, a possible substantial loss in statistical power to detect the weaker associations.

A recent approach to reduce multiple testing and improve inference in genomic association analysis involves the consolidation of SNPs or expression probes into sets of related genes [i.e., gene sets (GS)], followed by a determination if the gene set is associated with a trait [25,26]. Gene set analysis (GSA) was initially proposed for microarray expression data as a Gene Set Enrichment Analysis (GSEA) [6,27]. The GSEA method made use of a priori biological knowledge of genes to determine the GSs, such as biochemical pathways. While many GSA methods for expression have been developed [28], recent GSA methods for expression studies have been extended for use with genome-wide SNP data [25,29]. GSA methods designed for expression and SNP data fall into two separate categories, competitive or self-contained, based on the null hypothesis tested and within each category, methods differ widely in the statistics used for the GSs and how to assess the significance of these statistics [26,29]. A common feature among most of the methods developed for GSA is the use of databases to define the GSs. These databases usually group genes that fall into a biological pathway or have similarly defined characteristics. A number of databases exist with different approaches and definitions for grouping genes, such as Gene Ontology [30], KEGG [31], and PharmGKB [32].

A number of recent efforts have been applied the GS enrichment methodology towards identifying eQTLs. While this strategy provides a reasonable follow up analysis to the SNP-expression pair-wise analyses, it still requires the exhaustive pair-wise tests to be performed and the necessary permutations for unbiased association testing [33,34]. In particular, Li et al. proposed a method in which the eQTL p-values within a GS are combined using Fisher's method, followed by approximation of the distribution of the test statistic under the null hypothesis using Satterwhite's approximation [35]. An alternative approach to methods based on summary statistics (i.e., p-values) is one in which the association of SNP genotypes with mRNA gene expression levels within a given pathway is assessed with using multivariate model [36]. Examining eQTLs in the context of the Protein Interaction Network has also been done [37].

In addition to the use of GS analysis method for reducing the dimensionality of genomic data, the use of principal components analysis (PCA) has also been used in the analysis of high dimensional genomic data as a means to extract the features (e.g., components) with the most variation. The selected subset of principal components (PCs) accounting for a majority of the overall variation observed in the genomic data can then be analyzed in a manner similar to the original data. Gauderman et al. introduced the use of PCA for assessing the association of multiple SNPs within a candidate gene [38].

In this paper, we present a new approach to identify genetic variation associated with the mRNA expression by modeling SNP and mRNA expression variables within the context of pre-defined GSs. This method, referred to as gene set eQTL (GS-eQTL), is illustrated using data from a cell line model system [39–41] and the GSs (or pathways) defined in PharmGKB (http://www.pharmgkb.org/) [32] and KEGG (http://www.genome.jp/kegg/) [31]. Application of GS-eQTL to these two sets of GSs enable us to detect 28,597 GS-eQTL associations with an empirical false discovery rate (FDR) less than 0.05 (436 GS-eQTLs in PharmGKB and 28161 GS-eQTLs in KEGG). Replication of two of these top GS-eQTL associations using data in HapMap was also completed resulting in GS-eQTL p-values <0.05 (e.g., replication of the GS-eQTL).

In summary, our proposed approach has demonstrated its applicability and potential for analyzing the associations between SNP and mRNA expression data beyond the traditional single marker eSNP analyses. The use of GSs reduces the multiple testing and focuses on biologically relevant hypotheses. The current study, involving cell line data and PharmGKB and KEGG GSs, illustrates these two attractive features. Such methods and subsequent findings will become increasingly important in aiding the functional translation of disease risk or pharmacogenomic association findings.

Materials and Methods

Cell Line Model System

EBV-transformed lymphoblastoid cell lines (LCLs) from 96 African-American (AA), 96 Caucasian-American (CA), and 96 Han Chinese–American (HCA) unrelated subjects (sample sets HD100AA, HD100CAU, HD100CHI) were purchased from the Coriell Cell Repository. NIGMS collected and anonymized the samples, and all subjects provided written consent for their experimental use.

DNA from the LCLs was genotyped using Illumina HumanHap 550 K and 510 S BeadChips, which assayed 561,298 and 493,750 SNPs, respectively. Genotyping was performed in the Genotype Shared Resource at the Mayo Clinic. The genotyping data had been described previously [39–41]. SNP quality control procedures consisted of removal of SNPs with low call rate ($<95\%$), low minor allele frequency (MAF) (<0.05), and departures from Hardy Weinberg Equilibrium ($p<0.001$). Subjects with call rates $<95\%$ were also removed from the analysis. SNP genotypes were coded in terms of the number of minor alleles (e.g., 0, 1 or 2) (i.e., additive genetic model). Missing genotypes were imputed with the mean dosage value for the SNP. Population stratification was assessed due to the use of cell lines representing multiple races/ ethnic groups, as discussed in Li et al. [40] and Niu et al. [41], in which an eigen analysis was used to detect and adjust for population stratification [42].

Total RNA was extracted the cell lines using Qiagen RNeasy Mini kits (QIAGEN, Inc.). RNA quality was tested using an Agilent 2100 Bioanalyzer, followed by hybridization to Affymetrix U133 Plus 2.0 Gene-Chips, which contains a total of 54,613 probe sets, in two batches. mRNA expression array data were obtained for all of the cell lines with no missing data and normalized on a log2 scale using GCRMA [43]. The data had been used in previous reported studies (NCBI Gene Expression Omnibus, http://www.ncbi.nlm.nih.gov/geo, SuperSeries accession number GSE24277) [39–41]. The mean and standard deviations (SD) were calculated for each mRNA expression probe set with the GCRMA normalized values. Outliers with mRNA expression values more than 4 SD from the mean expression value were replaced with the maximum outlier value (mean expression value $+/-4$ SD). Similar to the genotype data, prior to GSA the expression values were adjusted using the same model including the race effect, population stratification eigenvectors, gender, and batch effect.

Gene set eQTL association analysis (GS-eQTL)

SNPs within 20 kb of the flanking sequence of a gene were mapped to the gene, with multiple SNP-to-gene mappings allowed. mRNA expression probe sets were also mapped to their respective genes. All PharmGKB [32] and KEGG [31] GSs were downloaded and genes mapped to GSs with multiple gene-to-GS mappings allowed (http://www.pharmgkb.org/, http://www.

genome.jp/kegg/). For GSs containing SNP genotypes (GS$_{SNP}$) or mRNA expression values (GS$_{expression}$), we define a "*cis*-GS" association to reference an association between SNPs and expression probe sets that mapped to the same GS (GS$_{SNP}$ = GS$_{expression}$). A "*trans*-GS" association is defined to represent the association between SNPs and expression probe sets that mapped to different GSs (GS$_{SNP}$ ≠ GS$_{expression}$). Hierarchical clustering using *hclust*, an R function [44], was used to visualize the overlap existing between the PharmGKB and KEGG GSs. We defined a distance measure between GSs to be 1–τ, where τ represents the average proportion of genes shared between the GSs.

With SNPs and expression probes mapped to GSs, we sought to model the association between all GS$_{SNP}$ and GS$_{expression}$ within PharmGKB and KEGG GSs using a multivariate linear model. Let GS$_{SNP}$ and GS$_{expression}$ represent all the adjusted SNP genotypes and expression probe set values, respectively, mapped to genes contained in a GS. For each set of SNPs within the given GS, GS$_{SNP,}$ we performed a principle component analysis (PCA) to reduce the dimensionality of GS$_{SNP}$ [45]. This approach has been applied with success in other GSA methods to produce a lower-dimensional GS [46,47]. In addition, PCA is a commonly used approach for modeling the association of multiple SNPs within a gene, as opposed to GS [38,48]. The design matrix was then constructed using the components that explain 80% of the variance of the adjusted SNP genotypes within the GS of interest (i.e., design matrix of predictors variables is defined as $\mathbf{X}$ = PCA$_{80\%}$(GS$_{SNP}$)). Similarly, PCA is also applied to GS$_{expression}$, where we also keep the components that explain 80% of the variance of the adjusted mRNA expression values (i.e., response variable is defined as $\mathbf{Y}$ = PCA$_{80\%}$(GS$_{expression}$)).

Next, we define the GS-eQTL model as $\mathbf{Y} = \mathbf{B}_0 + \mathbf{B}_1 * \mathbf{X} + \boldsymbol{\varepsilon}$, where $\mathbf{B}_1$ represents the vector of SNP effects (represented by the principal components needed to explain 80% of the variation), $\mathbf{B}_0$ represents the intercept and $\boldsymbol{\varepsilon}$ is the error assumed to follow a normal distribution with mean zero and common variance, $N(0,\sigma^2)$. The test of association between the expression and SNP GSs is then completed by assessing $\mathbf{B}_1$ using a multivariate analysis of variance with a Wilk's lambda test statistic where under the null hypothesis this vector of effects equals zero (H_0: $\mathbf{B}_1 = \mathbf{0}$). To account for the multiple testing and correlation between GS tests we computed false discovery rates (FDR) using 10,000 permutations [49]. Permutations were completed by shuffling the samples' expression values while holding the SNP data fixed and re-performing all GS-eQTL analyses for each permutation.

Replication of top two GS-eQTL results

Replication of two KEGG GS-eQTL (one *cis* and one *trans*) associations detected in the analysis of SNP and mRNA expression data measured on the Coriell cell lines was completed using publically available data on the HapMap cell lines. SNP data was downloaded for the Phase 2 HapMap CEU (unrelated) cell lines, while gene expression data was downloaded from the Gene Expression Omnibus (GEO) for GEO7792 (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc = GSE7792). GS-eQTL analyses were completed in a similar fashion as outlined for the GS-eQTL analysis of the Coriell cell line model system mRNA expression and SNP data.

Results

Gene set mapping

A total of 60 and 201 GSs were downloaded from PharmGKB and KEGG databases, respectively, and were used to map SNPs and expression probe sets measured on the LCLs. **Table 1** summarizes the total number of genes, SNPs and expression probe sets mapped to PharmGKB and KEGG GSs, as well as, GS sizes and amount of gene overlap between different GSs for both resources. For PharmGKB the GS sizes ranged from 2–64 genes with an average size of 14 genes compared to a range of 1–1100 genes with an average of 70 genes for KEGG GSs. The number of genes overlapping between different GSs was also larger for KEGG GSs as compared to PharmGKB GSs. When only considering GSs with overlapping genes, KEGG averaged a 10.5 gene overlap while PharmGKB averaged a 2.7 gene overlap.

General GS categories and sub-categories (KEGG only) were also identified for PharmGKB and KEGG GSs (**Table S1**). The PharmGKB categories designate different therapeutic groups, while the KEGG categories delineate biological functions or areas to classify the GSs. For the PharmGKB and KEGG GSs used in our analysis, there were 10 and 7 GS categories ranging between 1 and 25 and 1 and 85 GSs, respectively (**Figure 1**).

Top GS-eQTL associations

There were 436 PharmGKB GS associations between overall genetic and gene expression variation with FDR values <0.05. The top 20 PharmGKB GS-eQTL associations are presented in **Table 2.** For KEGG GSs, there was a large number of highly significant GS associations (minimum nominal $p = 3.84 \text{x} 10^{-129}$); however, the majority of the top results were driven by the 13 GSs which include the human leukocyte antigen (HLA) genes. Among the top 100 results (nominal $p < 10^{-40}$), only 7 involved GSs without HLA genes. After removing these HLA genes from the GS-eQTL analysis for the KEGG GSs, there were 28161 GS associations with FDR values <0.05. The top 20 KEGG GS-eQTL associations are presented in **Table 3.**

For both the PharmGKB and KEGG GSs, *cis*-GS associations were the most significant: PharmGKB "VEGF pathway" ($p = 7.46 \times 10^{-18}$) and KEGG "Metabolic pathways" (nominal $p = 7.86 \times 0^{-85}$). All PharmGKB and KEGG GS associations with FDR<0.05 are displayed in the heatmaps (**Figure 2**), with SNP and expression GSs indexed on the x- and y-axis, respectively. The SNP and expression GSs are ordered on the axes by the order established using hierarchical clustering with distances between GSs, where distance is 1 – τ (τ = average proportion of genes shared between GSs). The clusters are shown on the left and upper axes in **Figure 2**, with the colors indicating the GS categories. While the average "distance" between different GSs for PharmGKB and KEGG are 0.97 and 0.91, respectively, there are clusters of GSs due to overlaps of genes. However, **Figure 2** does not indicate a strong clustering among GSs within the same category, with a lower average distance between GSs within the same category as compared to the average distance between GSs in different categories (PharmGKB = 0.47 verses 0.97; KEGG = 0.49 verses 0.98).

Figure 2 also provides a visual for the GSs which are involved in a large number of significant associations, either as a SNP or expression GS. The highly associated SNP GSs appear as vertical lines in the heatmaps, indicating their association with a large number of expression GSs, while the highly associated expression GSs similarly appear as horizontal lines. **Table 4** lists the five SNP and expression GSs involved in the most GS associations for PharmGKB and KEGG. For PharmGKB, "EGFR Inhibitors Pathway PD" had the most associations (31 associations) as an expression GS, while "Antiarrhythmic Drug Pathways" had the most associations (32 associations) as a SNP GS. For the analysis of the KEGG GSs, the expression GS involved in the most associations was "Pathways in cancer" (142 associations), while the "Calcium signaling pathway" was involved in the most

Table 1. Summary of gene expression and SNP GS mappings for PharmGKB and KEGG.

Source	Total genes mapped	Genes per GS			Gene overlap			SNPs mapped		Expression probe sets mapped	
		Avg.	Max	Min	Avg.	Max	Min	Max	Min	Max	Min
PharmGKB	511	13.93	64	2	0.76	18	0	4384	172	192	2
KEGG	5333	70.02	1100	1	1.96	126	0	50871	35	2149	1

doi:10.1371/journal.pone.0043301.t001

associations (142 associations) as a SNP GS. **Figure 3** is a plot of the number of associations for each GS against each GS's average distance (based on the proportion of overlap of genes) to the GSs associated with it. The five highly associated GSs for PharmGKB and KEGG all have average distances >0.87, indicating that the large number of associations is not simply due to an overlap between GSs.

Next, **Figure 4** shows boxplots of the log transformed p-values for all SNP and expression GS associations by GS category. The largest PharmGKB GS category, Antineoplastic and Immunomodulating Agents, also contained the most significant association, a *cis*-GS association for the "VEGF pathway" (nominal $p = 7.46 \times 10^{-18}$). The KEGG category, Global Map, contained the most significant *cis*-GS association for the GS "Metabolic pathways" (nominal $p = 7.86 \times 10^{-85}$). Comparing the SNP and expression GS associations by PharmGKB and KEGG categories, little differences were observed. The level of association results also appeared to be evenly distributed amongst the categories, other than the KEGG Global Map category having many more highly significant associations than the other categories for KEGG.

Cis- verses trans- gene set associations

To assess if we observed more *cis*- or *trans*- GS associations for both the PharmGKB and KEGG GSs, we tested whether a disproportionate amount of *cis*- or *trans*-GS associations among the findings with FDR <5% were observed. For the PharmGKB associations with FDR <5%, there were 19 out of 60 (32%) *cis*- and 417 out of 3,540 (11.8%) *trans*-GS associations (empirical $p < 4.0 \times 10^{-4}$). In contrast, for the KEGG GS-eQTL analysis, 188 out of 201 (94%) *cis*-GS associations had FDR <5% while and 27,973 out of 40,200 (70%) *trans*-GS associations had FDR <5% (empirical $p < 1.0 \times 10^{-4}$). **Figure 5** illustrates this larger number of significant GS associations for *cis*-relationships as compared to *trans*-relationships for both PharmGKB and KEGG GS-eQTL analyses.

Replication of top two GS-eQTLs

Replication of the top KEGG GS-eQTL associations involving the Pathways in Cancer and Neuroactive ligand-receptor interaction was completed using publically available SNP and mRNA expression data measured on the CEU HapMap LCLs. The analysis *cis* GS-eQTL association between the Pathways in Cancer GS in the HapMap data resulted in a p-value of 0.00076. Similarly, the *trans*-association between the variation in mRNA expression levels for genes within the Pathways in Cancer GS and the genetic variation for genes within the Neuroactive ligand-receptor interaction GS was replicated with a p-value of 0.036.

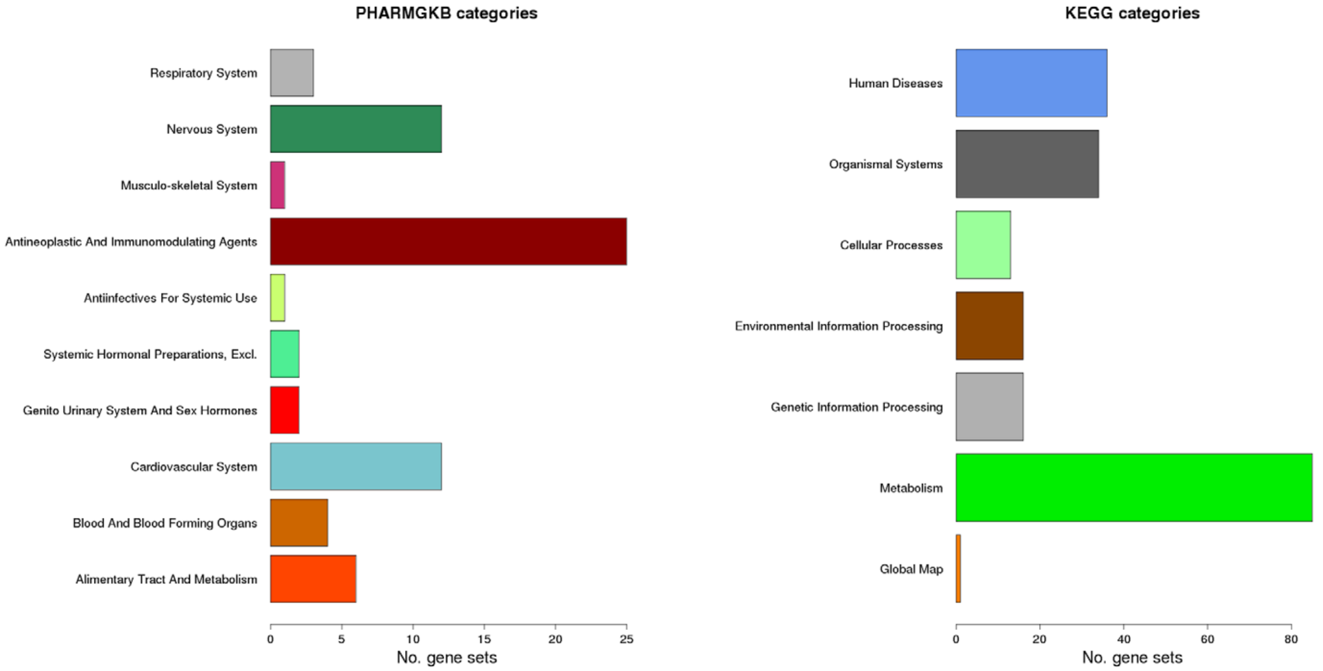

Figure 1. Barplots for number of GSs in PharmGKB and KEGG categories.
doi:10.1371/journal.pone.0043301.g001

Table 2. Top 20 PharmGKB GS-eQTL associations.

Type	$GS_{expression}$				GS_{SNP}				GS-eQTL p-value	FDR
	Gene Set	No. Genes	No. Probe sets	No. PCs	Gene Set	No. Genes	No. SNPs	No. PCs		
cis	VEGF Pathway	15	41	9	VEGF Pathway	15	700	57	1.4×10^{-23}	0
trans	EGFR Inhibitors Pathway PD	64	192	26	VEGF Pathway	15	700	57	1.4×10^{-12}	0
trans	Thiopurine Pathway	32	82	14	Antiarrhythmic Drug Pathways	55	3667	143	5.1×10^{-11}	0
cis	Glucocorticoid and Inflammatory genes Pathway PD	9	31	6	Glucocorticoid and Inflammatory genes Pathway PD	9	571	31	1.5×10^{-10}	0
cis	Methotrexate Pathway	29	74	16	Methotrexate Pathway	29	2415	87	3.9×10^{-10}	0
cis	EGFR Inhibitors Pathway PD	64	192	26	EGFR Inhibitors Pathway PD	64	4384	130	3.9×10^{-10}	0
trans	Fluoropyrimidine PK	24	62	12	Antiarrhythmic Drug Pathways	55	3667	143	7.7×10^{-10}	0
trans	EGFR Inhibitors Pathway PD	64	192	26	Antiarrhythmic Drug Pathways	55	3667	143	7.9×10^{-9}	1.1×10^{-5}
cis	Etoposide Pathway	12	33	4	Etoposide Pathway	12	1104	47	4.3×10^{-8}	3.8×10^{-5}
cis	Bisphosphonate Pathway	19	55	13	Bisphosphonate Pathway	19	775	51	4.9×10^{-8}	4.0×10^{-5}
trans	Bisphosphonate Pathway	19	55	13	Antiarrhythmic Drug Pathways	55	3667	143	5.1×10^{-8}	4.0×10^{-5}
trans	Fluoropyrimidine PK	24	62	12	Methotrexate Pathway	29	2415	87	7.7×10^{-8}	4.9×10^{-5}
cis	Statin Pathway Cholesterol and Lipoprotein Transport PD	26	55	11	Statin Pathway Cholesterol and Lipoprotein Transport PD	26	1286	83	8.0×10^{-8}	4.9×10^{-5}
trans	Selective Serotonin Reuptake Inhibitors SSRI Pathway	28	72	12	Antiarrhythmic Drug Pathways	55	3667	143	9.9×10^{-8}	5.2×10^{-5}
cis	Fluoropyrimidine PK	24	62	12	Fluoropyrimidine PK	24	2483	83	1.1×10^{-7}	5.2×10^{-5}
trans	EGFR Inhibitors Pathway PD	64	192	26	Selective Serotonin Reuptake Inhibitors SSRI Pathway	28	1975	98	1.2×10^{-7}	5.2×10^{-5}
cis	Nicotine PD Pathway Dopaminergic Neuron	20	42	9	Nicotine PD Pathway Dopaminergic Neuron	20	1049	74	1.9×10^{-7}	6.3×10^{-5}
trans	Methotrexate Pathway	29	74	16	Antiarrhythmic Drug Pathways	55	3667	143	2.0×10^{-7}	6.3×10^{-5}
cis	Antiarrhythmic Drug Pathways	55	127	14	Antiarrhythmic Drug Pathways	55	3667	143	2.9×10^{-7}	9.3×10^{-5}
trans	VEGF Pathway	15	41	9	EGFR Inhibitors Pathway PD	64	4384	130	4.1×10^{-7}	0.0001

doi:10.1371/journal.pone.0043301.t002

Table 3. Top 20 KEGG GS-eQTL associations after removing the genes within the HLA region.

Type	$GS_{expression}$				GS_{SNP}				GS-eQTL p-value	FDR
	Gene Set	**No. Genes**	**No. Probe sets**	**No. PCs**	**Gene Set**	**No. Genes**	**No. SNPs**	**No. PCs**		
cis	Metabolic pathways	1100	2149	67	Metabolic pathways	1100	50871	189	7.9×10^{-85}	$<5\times10^{-8}$
trans	Metabolic pathways	1100	2149	67	Neuroactive ligand receptor interaction	302	14725	173	2.6×10^{-58}	$<5\times10^{-8}$
trans	Metabolic pathways	1100	2149	67	Calcium signaling pathway	178	12520	172	5.3×10^{-50}	$<5\times10^{-8}$
trans	Metabolic pathways	1100	2149	67	Pathways in cancer	330	18667	175	1.1×10^{-45}	<5x10–^{8}
trans	Metabolic pathways	1100	2149	67	Vascular smooth muscle contraction	125	8588	162	5.9×10^{-45}	<5x10^{-8}
trans	MAPK signaling pathway	273	683	48	Metabolic pathways	1100	50871	189	7.5×10^{-43}	$<5\times10^{-8}$
trans	Metabolic pathways	1100	2149	67	Cytokine cytokine receptor interaction	278	9377	162	1.1×10^{-42}	$<5\times10^{-8}$
trans	Metabolic pathways	1100	2149	67	Focal adhesion	201	12846	169	7.3×10^{-42}	$<5\times10^{-8}$
trans	Pathways in cancer	330	893	50	Metabolic pathways	1100	50871	189	1.3×10^{-39}	$<5\times10^{-8}$
trans	Metabolic pathways	1100	2149	67	Axon guidance	129	9757	163	2.8×10^{-39}	$<5\times10^{-8}$
trans	Metabolic pathways	1100	2149	67	Dilated cardiomyopathy	92	7964	155	4.5×10^{-39}	$<5\times10^{-8}$
trans	Metabolic pathways	1100	2149	67	Chemokine signaling pathway	190	9422	157	7.5×10^{-39}	$<5\times10^{-8}$
cis*	Pathways in cancer	330	893	50	Pathways in cancer	330	18667	175	4.9×10^{-38}	$<5\times10^{-8}$
Trans	Metabolic pathways	1100	2149	67	Neurotrophin signaling pathway	126	6436	146	2.1×10^{-37}	$<5\times10^{-8}$
Trans	Metabolic pathways	1100	2149	67	Arrhythmogenic right ventricular cardiomyopathy ARVC	76	8274	158	2.7×10^{-37}	$<5\times10^{-8}$
Trans	Metabolic pathways	1100	2149	67	Hypertrophic cardiomyopathy HCM	89	7492	154	3.2×10^{-37}	$<5\times10^{-8}$
Trans	Metabolic pathways	1100	2149	67	Purine metabolism	159	9165	159	7.2×10^{-37}	$<5\times10^{-8}$
trans*	Pathways in cancer	330	893	50	Neuroactive ligand receptor interaction	302	14725	173	1.9×10^{-36}	$<5\times10^{-8}$
trans	MAPK signaling pathway	273	683	48	Pathways in cancer	330	18667	175	2.1×10^{-36}	$<5\times10^{-8}$
trans	Metabolic pathways	1100	2149	67	GnRH signaling pathway	101	6256	152	4.1×10^{-36}	$<5\times10^{-8}$

*GS-eQTL associations assessed in HapMap for replication.

doi:10.1371/journal.pone.0043301.t003

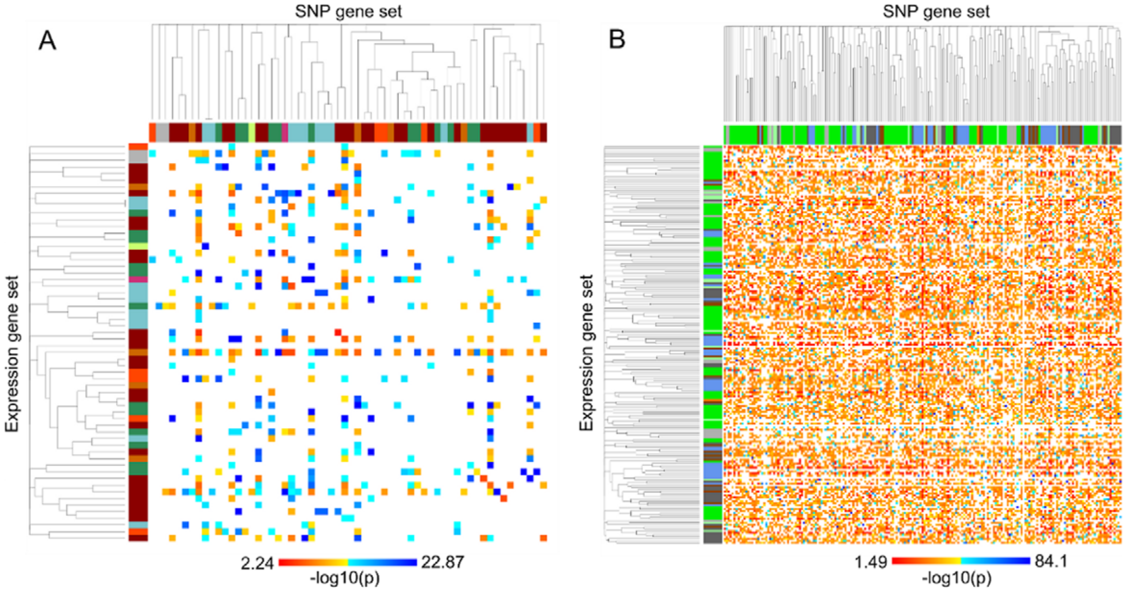

Figure 2. Heatmaps with points indicating associations (FDR <5%) between SNP (x-axis) and expression (y-axis) GSs. SNP and expression GSs are indexed based on hierarchical clustering using distances between GSs (distance determined by average proportion of genes shared between GSs). The color of the points indicate the level of association significance (blue = less significant, red = more significant)

doi:10.1371/journal.pone.0043301.g002

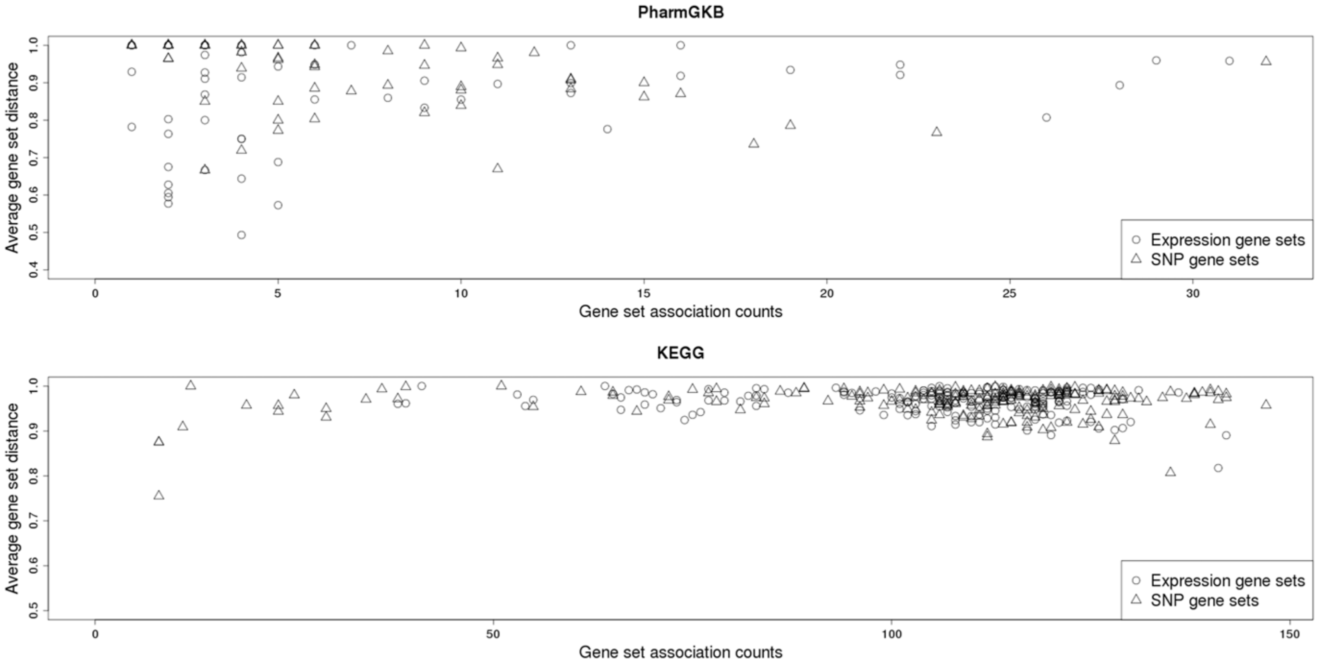

Figure 3. Scatter plot of the number of associations (FDR <5%) each GS was involved in against the average distance between each GS and GSs associated with it.
doi:10.1371/journal.pone.0043301.g003

Discussion

Using SNP and expression genome-wide data collected on a cell based model system we applied a new approach, GS-eQTL analysis, to identify genetic variation associated with mRNA gene expression in the context of GSs or pathways. By modeling the genetic variation and expression using GSs we were able to increase statistical power by reducing the multiple testing inherit

Table 4. Top five SNP and expression GSs involved in the most associations (FDR <5%).

Source	Data type	Gene Set	No. significant associations
PharmGKB	Expression	EGFR Inhibitors Pathway PD	31
		Selective Serotonin Reuptake Inhibitors SSRI Pathway	29
		Methotrexate Pathway	28
		Doxorubicin Pathway	26
		Antiarrhythmic Drug Pathways	22
	SNP	Antiarrhythmic Drug Pathways	32
		Taxane Pathway	23
		Imatinib	19
		Doxorubicin Pathway	18
		Fluoropyrimidine PK	16
KEGG	Expression	Pathways in cancer	142
		Metabolic pathways	141
		Cysteine and methionine metabolism	136
		ABC transporters	131
		Insulin signaling pathway	130
	SNP	Calcium signaling pathway	147
		Tyrosine metabolism	142
		Ether lipid metabolism	142
		Nucleotide excision repair	141
		Antigen processing and presentation	141

doi:10.1371/journal.pone.0043301.t004

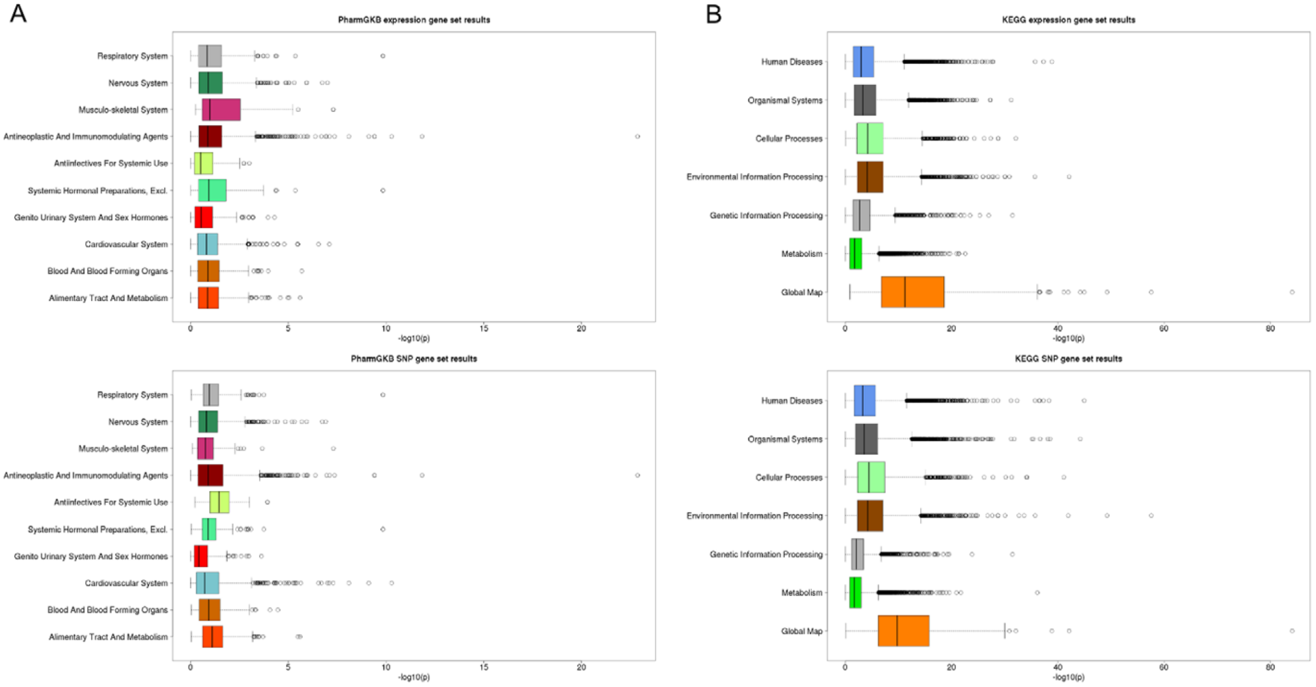

Figure 4. Boxplots of log transformed p-values for GS-eQTL association results by category. (A) PharmGKB and (B) KEGG.
doi:10.1371/journal.pone.0043301.g004

with high dimensional genomic data and combining the genetic variation and mRNA expression of functionally related genes, as defined by PharmGKB and KEGG. GSs have recently been used in a variety of settings for increased power [50–53]; however, limited research has been completed to apply the ideas of GSA to the study the genetics of gene expression and eQTL analysis.

After adjusting for multiple testing, we determined a large number of significant GS-eQTL associations (FDR <5%) for both GSs in PharmGKB and KEGG. Replication was attempted for two of the top association for KEGG GSs using the publically available data on the HapMap samples. The "Pathways in Cancer" *cis* GS-eQTL and the *trans* GS-eQTL association between the variation in mRNA expression levels for genes within the "Pathways in Cancer" GS and the genetic variation for genes within the "Neuroactive ligand-receptor interaction" GS were both replicated using the publically available data from HapMap

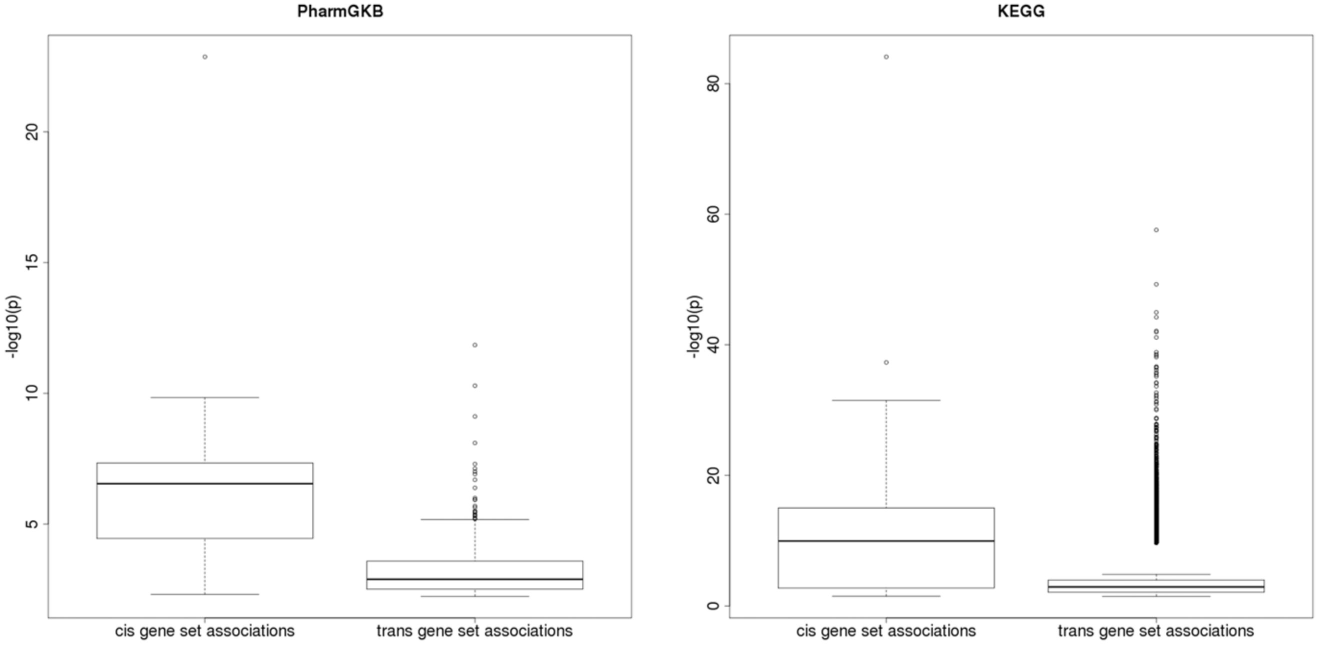

Figure 5. Boxplots of log transformed p-values for *cis*- and *trans*-GS association results. (A) PharmGKB and (B) KEGG.
doi:10.1371/journal.pone.0043301.g005

with p-values of 0.00076 and 0.036, respectively. The first canonical correlation between the mRNA gene expression and SNP genotypes for "Pathways in Cancer" *cis* GS-eQTL and the *trans* GS-eQTL association between the "Pathways in Cancer" and the "Neuroactive ligand-receptor interaction" GSs was 0.98 for both GS-eQTLs.

Examining all pairwise SNP-expression associations within the "Pathways in Cancer" found an association between rs2235529 within *WNT4* and the mRNA expression level of *CDC42* (208727_s) with a p-value of $1.61\times10-42$ (Bonferroni correction for testing all pairwise associations within this GS results in a p-value $=3.9\times10^{-35}$). These two genes are 25 kb apart, suggesting a typical *cis* regulatory relationship. There were an additional 31 eQTL associations within "Pathways in Cancer" GS with Bonferroni adjusted p-values $<1\times10^{-9}$, with *cis* associations observed for 5 genes within the GS. For the second replicated GS-eQTL between the "Pathways in Cancer" GS and the "Neuroactive ligand-receptor interaction" GS, the most significant eQTL association involved SNP rs1160198 from gene *GLRA2* and expression of *IGF1R* (Bonferroni adjusted p-value of 7.95×10^{-8}).

The large number of associations identified for KEGG may be due to the correlation structure that exists among the KEGG GSs or "master" regulating genes or GSs. Grouping the significant GS associations by category did not show a large difference between categories in terms of strength of association. There were also GSs involved in many GS associations either as a SNP or expression GS, which are analogous to eQTL "hotspots" in previous literature [4,54,55]. The SNP GSs with many associations can be considered "master regulator" GSs in terms of regulating the expression of other GSs, while expression GSs with many associations appear to be regulated by many different GSs. The concept of "master regulator" GSs may not be as straightforward as a single gene "master regulator" in a biological sense, but the GS associations may be indicating the interaction or regulation between many components involved in a complex system of biological processes or functions. Among the top findings for both GS resources, we also observed a greater proportion of *cis* GS-eQTL associations as compared to *trans* GS-eQTL associations, as one would expect from previous eQTL research [56].

Given the use of functionally defined GSs to perform GS-eQTL analysis, there are broader implications from these findings to consider beyond the standard SNP verses expression analyses, particularly for the regulatory function and/or regulation of drug pharmacokinetic (PK) and pharmacodynamic (PD) pathways. The PK and PD pathways are well characterized and studied pathways, and are composed of the elements involved in either the metabolism (PK) or targeted action (PD) of drugs. Thus, further understanding of these pathways has significant clinical impact. The *trans* GS-eQTL associations provide hypotheses to further pursue regarding the genetics of gene expression. For example, one of the top *trans*-GS associations for PharmGKB involved the expression of the "Thiopurine" pathway and the genetic variation of the "Anti-arrhythmic Drug" pathway. While these two pathways are curated for two completely different drugs, their genetic components appear to be associated.

Similarly with the KEGG results, there were many *trans*-GS associations which suggest novel hypotheses to be further explored. From the top 30 KEGG results, 25 involved the expression or genetic variation of the KEGG GS "Metabolic pathways," indicating a significant role for these genes in the genetics of human mRNA expression. Due to the non-specific nature of many GSs, other methods, such as gene level tests, will be needed to follow up on these initial findings to determine the potential "drivers" of these associations. Thus, this method is highlighted as an effective first step to help focus follow up association and/or functional studies to establish novel associations between genome-wide genetic sequence variation and mRNA gene expression.

The use of human cell lines from unrelated subjects (i.e. lymphoblastoid cell lines from HapMap samples) for eQTL studies have recently been successful in identifying many significant findings [24,57]; however, tissue-dependent patterns of gene expression may limit the generalization of our findings. A recent study suggests little eQTL overlap between tissues [58], while other work has found a more substantial eQTL overlap exists across tissues when considering sample size differences between eQTL studies [59]. Nonetheless, tissue dependent gene expression could play a considerable role in the context of our approach, especially when examining certain PK pathways that involve many genes that encode metabolic enzymes which are highly expressed in the liver. Future work is needed to consider GS-eQTLs studies where mRNA is measured in diverse tissue types, such as liver and adipose tissues.

In this manuscript, we focused on GS-eQTL analysis between GSs and pathways contained within PharmGKB and KEGG with SNPs mapped to within 20 kb of the 3′ and 5′ ends of each gene. Considering variation beyond 20 kb may include more functional variants, but studies have shown that much of the key variation lies within 20 kb of the gene transcription start and end sites [60]. Additionally, the current definitions of PharmGKB and KEGG pathways are incomplete and have a clear bias towards studies involving certain genes and therapeutic agents, and thus limit the scope of our conclusions. However, the novel GS-eQTL analysis proposed has the ability to easily be extended to other pathway or GS sources such as Gene Ontology (GO) [30].

Application of PCA in our GS-eQTL analysis method effectively reduced the dimensionality of the genomic data. However, in applying PCA one must deal with missing data. In our analysis, we removed SNPs with a call rate <95%. Due to the small amount of missing genotypic data, we chose to impute the mean SNP genotype (in terms of the number of minor alleles) for missing genotypes. Another approach to deal with missing genotypic data would be to use one of the various genotype imputation methods [61]. A second limitation is that PCA only assesses linear relationships as a means of dimension reduction between the data which may not be optimal for all GSs. Future work is on-going to determine an approach to reduce the dimensionally of the genetic and mRNA expression data using both linear and non-linear relationship, such as kernels [62,63], along with the application of this approach to other forms of genomic data, such as microRNA or methylation data.

In conclusion, we have demonstrated an efficient approach to analyze the high dimensional data for studying the genetics of gene expression with application of the GS-eQTL approach to determine novel relationships between GSs and pathways within PharmGKB and KEGG. A systems biology approach with GSs is a natural application towards studying the genetics of gene expression to reduce the high-dimensionality of the data and to make use of GSs grouped based on a biological process or function in which there already may be an expected relationship between the annotated GS processes or functions. Developing and applying new approaches, such as ours, to analyze the high-dimensional genomic data to identify associations is a necessary step towards establishing the regulatory relationships at the molecular level, which will help translate findings from disease risk or pharmacogenomic studies towards meaningful biology.

Supporting Information

Table S1 Categories of Gene Sets.
(XLSX)

References

1. Cheung VG, Spielman RS (2009) Genetics of human gene expression: mapping DNA variants that influence gene expression. Nat Rev Genet 10: 595–604.
2. Lettice LA, Horikoshi T, Heaney SJ, van Baren MJ, van der Linde HC, et al. (2002) Disruption of a long-range cis-acting regulator for Shh causes preaxial polydactyly. Proceedings of the National Academy of Sciences of the United States of America 99: 7548–7553.
3. Nobrega MA, Ovcharenko I, Afzal V, Rubin EM (2003) Scanning human gene deserts for long-range enhancers. Science 302: 413.
4. Breitling R, Li Y, Tesson BM, Fu J, Wu C, et al. (2008) Genetical genomics: spotlight on QTL hotspots. PLoS Genet 4: e1000232.
5. Moffatt MF, Kabesch M, Liang L, Dixon AL, Strachan D, et al. (2007) Genetic variants regulating ORMDL3 expression contribute to the risk of childhood asthma. Nature 448: 470–473.
6. Mootha VK, Lindgren CM, Eriksson KF, Subramanian A, Sihag S, et al. (2003) PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. Nat Genet 34: 267–273.
7. Lango Allen H, Estrada K, Lettre G, Berndt SI, Weedon MN, et al. (2010) Hundreds of variants clustered in genomic loci and biological pathways affect human height. Nature 467: 832–838.
8. Fransen K, Visschedijk MC, van Sommeren S, Fu JY, Franke L, et al. (2010) Analysis of SNPs with an effect on gene expression identifies UBE2L3 and BCL3 as potential new risk genes for Crohn's disease. Hum Mol Genet 19: 3482–3488.
9. Plenge RM (2010) Unlocking the pathogenesis of celiac disease. Nat Genet 42: 281–282.
10. Goring HH, Curran JE, Johnson MP, Dyer TD, Charlesworth J, et al. (2007) Discovery of expression QTLs using large-scale transcriptional profiling in human lymphocytes. Nat Genet 39: 1208–1216.
11. Emilsson V, Thorleifsson G, Zhang B, Leonardson AS, Zink F, et al. (2008) Genetics of gene expression and its effect on disease. Nature 452: 423–428.
12. Monks SA, Leonardson A, Zhu H, Cundiff P, Pietrusiak P, et al. (2004) Genetic inheritance of gene expression in human cell lines. Am J Hum Genet 75: 1094–1105.
13. Morley M, Molony CM, Weber TM, Devlin JL, Ewens KG, et al. (2004) Genetic analysis of genome-wide variation in human gene expression. Nature 430: 743–747.
14. Cheung VG, Spielman RS, Ewens KG, Weber TM, Morley M, et al. (2005) Mapping determinants of human gene expression by regional and genome-wide association. Nature 437: 1365–1369.
15. Stranger BE, Forrest MS, Clark AG, Minichiello MJ, Deutsch S, et al. (2005) Genome-wide associations of gene expression variation in humans. PLoS genetics 1: e78.
16. Veyrieras J-B, Kudaravalli S, Kim SY, Dermitzakis ET, Gilad Y, et al. (2008) High-resolution mapping of expression-QTLs yields insight into human gene regulation. PLoS genetics 4: e1000214.
17. Franke L, Jansen RC (2009) eQTL analysis in humans. Methods in molecular biology 573: 311–328.
18. Cheung VG, Spielman RS, Ewens KG, Weber TM, Morley M, et al. (2005) Mapping determinants of human gene expression by regional and genome-wide association. Nature 437: 1365–1369.
19. Dixon AL, Liang L, Moffatt MF, Chen W, Heath S, et al. (2007) A genome-wide association study of global gene expression. Nat Genet 39: 1202–1207.
20. Myers AJ, Gibbs JR, Webster JA, Rohrer K, Zhao A, et al. (2007) A survey of genetic human cortical gene expression. Nat Genet 39: 1494–1499.
21. Stranger BE, Forrest MS, Clark AG, Minichiello MJ, Deutsch S, et al. (2005) Genome-wide associations of gene expression variation in humans. PLoS Genet 1: e78.
22. Sieberts SK, Schadt EE (2007) Moving toward a system genetics view of disease. Mamm Genome 18: 389–401.
23. Chen L, Page GP, Mehta T, Feng R, Cui X (2009) Single nucleotide polymorphisms affect both cis- and trans-eQTLs. Genomics 93: 501–508.
24. Nicolae DL, Gamazon E, Zhang W, Duan S, Dolan ME, et al. (2010) Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. PLoS Genet 6: e1000888.
25. Wang K, Li M, Hakonarson H (2010) Analysing biological pathways in genome-wide association studies. Nat Rev Genet 11: 843–854.
26. Goeman JJ, Buhlmann P (2007) Analyzing gene expression data in terms of gene sets: methodological issues. Bioinformatics 23: 980–987.
27. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, et al. (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A 102: 15545–15550.
28. Song S, Black MA (2008) Microarray-based gene set analysis: a comparison of current methods. BMC Bioinformatics 9: 502.
29. Fridley BL, Biernacka JM (2011) Gene set analysis of SNP data: benefits, challenges, and future directions. Eur J Hum Genet 19: 837–843.

Author Contributions

Conceived and designed the experiments: BLF RA. Performed the experiments: RA GDJ. Analyzed the data: RA GDJ. Contributed reagents/materials/analysis tools: LW. Wrote the paper: RA BLF.

30. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, et al. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25: 25–29.
31. Kanehisa M, Goto S (2000) KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res 28: 27–30.
32. Klein TE, Chang JT, Cho MK, Easton KL, Fergerson R, et al. (2001) Integrating genotype and phenotype information: an overview of the PharmGKB project. Pharmacogenetics Research Network and Knowledge Base. Pharmacogenomics J 1: 167–170.
33. Grieve IC, Dickens NJ, Pravenec M, Kren V, Hubner N, et al. (2008) Genome-wide co-expression analysis in multiple tissues. PloS one 3: e4033.
34. Wu C, Delano DL, Mitro N, Su SV, Janes J, et al. (2008) Gene set enrichment in eQTL data identifies novel annotations and pathway regulators. PLoS genetics 4: e1000070.
35. Li S, Williams BL, Cui Y (2011) A combined p-value approach to infer pathway regulations in eQTL mapping. Statistics and Its Interface 4: 389–402.
36. Li S, Lu Q, Cui Y (2010) A systems biology approach for identifying novel pathway regulators in eQTL mapping. Journal of biopharmaceutical statistics 20: 373–400.
37. Suthram S, Beyer A, Karp RM, Eldar Y, Ideker T (2008) eQED: an efficient method for interpreting eQTL associations using protein networks. Molecular systems biology 4: 162.
38. Gauderman WJ, Murcray C, Gilliland F, Conti DV (2007) Testing association between disease and multiple SNPs in a candidate gene. Genet Epidemiol 31: 383–395.
39. Li L, Fridley B, Kalari K, Jenkins G, Batzler A, et al. (2008) Gemcitabine and Cytosine Arabinoside Cytotoxicity: Association with Lymphoblastoid Cell Expression. Cancer Res 68: 7050–7058.
40. Li L, Fridley BL, Kalari K, Jenkins G, Batzler A, et al. (2009) Gemcitabine and arabinosylcytosin pharmacogenomics: genome-wide association and drug response biomarkers. PLoS One 4: e7765.
41. Niu N, Qin Y, Fridley BL, Hou J, Kalari KR, et al. (2010) Radiation pharmacogenomics: a genome-wide association approach to identify radiation response biomarkers using human lymphoblastoid cell lines. Genome Res 20: 1482–1492.
42. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, et al. (2006) Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 38: 904–909.
43. Wu Z, Irizarry RA, Gentleman R, Martinez-Murillo F, Spencer F (2004) A Model-Based Background Adjustment for Oligonucleotide Expression Arrays. Journal of the American Statistical Association 99: 909–917.
44. Team RDC (2010) R: A Language and environment for statistical computing. Vienna, Austria.
45. Mardia K, Kent J, Bibby J (1979) Multivariate Analysis. London: Academic Press.
46. Chai HS, Sicotte H, Bailey KR, Turner ST, Asmann YW, et al. (2009) GLOSSI: a method to assess the association of genetic loci-sets with complex diseases. BMC Bioinformatics 10: 102.
47. Tomfohr J, Lu J, Kepler TB (2005) Pathway level analysis of gene expression using singular value decomposition. BMC Bioinformatics 6: 225.
48. Ballard DH, Cho J, Zhao H (2010) Comparisons of multi-marker association methods to detect association between a candidate region and disease. Genet Epidemiol 34: 201–212.
49. Yekutieli D, Benjamini Y (1999) Resampling-based false discovery rate controlling multiple test procedures for correlated test statistics. Journal of Statistical Planning and Inference 82: 171–196.
50. O'Dushlaine C, Kenny E, Heron E, Donohoe G, Gill M, et al. (2011) Molecular pathways involved in neuronal cell adhesion and membrane scaffolding contribute to schizophrenia and bipolar disorder susceptibility. Mol Psychiatry 16: 286–292.
51. Chen LS, Hutter CM, Potter JD, Liu Y, Prentice RL, et al. (2010) Insights into colon cancer etiology via a regularized approach to gene set analysis of GWAS data. Am J Hum Genet 86: 860–871.
52. Jia P, Wang L, Meltzer HY, Zhao Z (2010) Common variants conferring risk of schizophrenia: a pathway analysis of GWAS data. Schizophr Res 122: 38–42.
53. Menashe I, Maeder D, Garcia-Closas M, Figueroa JD, Bhattacharjee S, et al. (2010) Pathway analysis of breast cancer genome-wide association study highlights three pathways and one canonical signaling cascade. Cancer Res 70: 4453–4459.
54. Loguercio S, Overall RW, Michaelson JJ, Wiltshire T, Pletcher MT, et al. (2010) Integrative analysis of low- and high-resolution eQTL. PLoS One 5: e13920.
55. Duarte CW, Zeng ZB (2011) High-confidence discovery of genetic network regulators in expression quantitative trait loci data. Genetics 187: 955–964.
56. Li J, Burmeister M (2005) Genetical genomics: combining genetics with gene expression analysis. Hum Mol Genet 14 Spec No. 2: R163–169.

57. Murphy A, Chu JH, Xu M, Carey VJ, Lazarus R, et al. (2010) Mapping of numerous disease-associated expression polymorphisms in primary peripheral blood CD4+ lymphocytes. Hum Mol Genet 19: 4745–4757.
58. Dimas AS, Deutsch S, Stranger BE, Montgomery SB, Borel C, et al. (2009) Common regulatory variation impacts gene expression in a cell type-dependent manner. Science 325: 1246–1250.
59. Ding J, Gudjonsson JE, Liang L, Stuart PE, Li Y, et al. (2010) Gene expression in skin and lymphoblastoid cells: Refined statistical method reveals extensive overlap in cis-eQTL signals. Am J Hum Genet 87: 779–789.
60. Veyrieras JB, Kudaravalli S, Kim SY, Dermitzakis ET, Gilad Y, et al. (2008) High-resolution mapping of expression-QTLs yields insight into human gene regulation. PLoS Genet 4: e1000214.
61. Marchini J, Howie B (2010) Genotype imputation for genome-wide association studies. Nat Rev Genet 11: 499–511.
62. Vapnik VN (1998) Statistical Learning Theory. NY: Wiley.
63. Schlkopf B, Smola AJ (2002) Learning with Kernels. Cambridge: MIT Press.