BIOINFORMATICS ORIGINAL PAPER

Vol. 28 no. 3 2012, pages 403–410
doi:10.1093/bioinformatics/btr663

Databases and ontologies

Advance Access publication November 30, 2011

SNPxGE^2: a database for human SNP–coexpression associations

Yupeng Wang[1,2,*], Sandeep J. Joseph[3], Xinyu Liu[4], Michael Kelley[2] and Romdhane Rekaya[1,2,4*]

[1]Institute of Bioinformatics, [2]Department of Animal and Dairy Science, University of Georgia, Athens, GA 30602, [3]Department of Medicine, Division of Infectious Diseases, Emory University School of Medicine, Atlanta, GA 30322, USA and [4]Department of Statistics, University of Georgia, Athens, GA 30602, USA

Associate Editor: Jonathan Wren

ABSTRACT

Motivation: Recently, gene–coexpression relationships have been found to be often conditional and dynamic. Many studies have suggested that single nucleotide polymorphisms (SNPs) have impacts on gene expression variations in human populations.

Results: The SNPxGE^2 database contains the computationally predicted human SNP–coexpression associations, i.e. the differential coexpression between two genes is associated with the genotypes of an SNP. These data were generated from a large-scale association study that was based on the HapMap phase I data, which covered 269 individuals from 4 human populations, 556 873 SNPs and 15 000 gene expression profiles. In order to reduce the computational cost, the SNP–coexpression associations were assessed using gap/substitution models, proven to have a comparable power to logistic regression models. The results, at a false discovery rate (FDR) cutoff of 0.1, consisted of 44 769 and 50 792 SNP–coexpression associations based on single and pooled populations, respectively, and can be queried in the SNPxGE^2 database via either gene symbol or reference SNP ID. For each reported association, a detailed information page is provided.

Availability: http://lambchop.ads.uga.edu/snpxge2/index.php

Contact: wyp1125@uga.edu, rrekaya@uga.edu

Received on August 7, 2011; revised on October 30, 2011; accepted on November 27, 2011

1 INTRODUCTION

It is widely accepted that eukaryotic gene expression variation may be potentially modulated by many genetic and epigenetic factors such as transcription factors, *cis*-regulatory elements, enhancers/repressors, transposons, miRNAs, nucleosome positioning, DNA methylation, histone modifications and by environmental factors (Idaghdour *et al.*, 2010; Weirauch and Hughes, 2010; Wilson and Odom, 2009; Wray *et al.*, 2003). Moreover, much of gene expression variation is heritable (Cheung *et al.*, 2003).

Recently, many studies have suggested that single nucleotide polymorphisms (SNPs) also have impacts on gene expression variations in human populations. Stranger *et al.* (2005) found that many regions within 1 Mb of 374 expressed genes of interest had significant association of SNPs with expression variation. Subsequently, Stranger *et al.* (2007a) performed association analyses of expression levels of 14 925 transcripts with SNPs and copy number variants using multiple linear regression and suggested that SNPs captured 83.6% of the total detected genetic variation in gene expression. Later on, Stranger *et al.* (2007b) carried out another association analysis of >2.2 million common SNPs with gene expression using multiple linear regression and identified at least 1348 and 180 genes with association signals in *cis* and *trans* effects, respectively. Spielman *et al.* (2007) suggested that common genetic variants accounted for differences in gene expression as well as prevalence of complex diseases among ethnic groups. Duan *et al.* (2008) also reported several thousands of significant associations between expression quantitative nucleotides and transcript clusters in two of HapMap populations. In a separate study, Zhang *et al.* (2008) identified that some genes, which were differentially expressed among human populations, were also associated with particular biological processes. Using a Bayesian hierarchical model, Veyrieras *et al.* (2008) found strong enrichments of quantitative trait loci for gene expression (eQTLs) in 250 bp upstream regions of transcription end site (TES) and around the transcription start site (TSS). Similarly, Pickrell *et al.* (2010) identified genetic variations in more than a thousand genes, which influenced the overall gene expression or splicing.

In addition to individual gene expression profiles, the relationships among transcripts such as coexpression are also often studied for deciphering gene regulatory mechanisms. The identification of human gene–coexpression relationships using microarray data has often relied on overall correlations between gene expression profiles across many experiments or conditions (Lee *et al.*, 2004; Obayashi and Kinoshita, 2011). Recently, some studies have suggested that the coexpression relationships between genes are often dynamic and conditional, which may be dependent on cellular states (Li *et al.*, 2004), developmental stages (Adryan and Teichmann, 2010), disease status (case or control) (Choi *et al.*, 2005) or human populations (Nayak *et al.*, 2009). However, few studies have related such differential coexpression to genetic bases. The effects of SNPs on gene expression variation in humans have been well documented, and so it is reasonable to conjecture that the differential coexpression between two genes may be associated with the genotypes of an SNP, which is termed as an SNP–coexpression association in this study. Kayano *et al.* (2009) showed a biological switch mechanism in coexpression between correlation and inverse correlation of two genes, controlled by the genotypes of an SNP. However, the switch mechanism in two genes' coexpression controlled by an SNP could be only part of the whole picture of SNP–coexpression associations because it is also possible that two genes are well coexpressed under individual genotypes of an SNP

*To whom correspondence should be addressed.

while their coexpression cannot be detected if different genotypes of the SNP are pooled, or that two genes are well coexpressed under one or two genotypes of an SNP while are not coexpressed under others. Furthermore, Kayano *et al.*'s method was only able to assess 3×10^8 SNP–expression–expression combinations, of which only 142 gene expression profiles were included, more than two orders of magnitude fewer than what is typical in a genome-wide expression data.

Studies of associations between larger number of transcripts and genotypes data are computationally demanding. Thus, databases that deposit pre-computed associations are informative and helpful to biological researchers. For example, the SCAN database provides queries of the results of association of HapMap variants with gene expression at user-specified thresholds (Gamazon *et al.*, 2010). Genevar allows researchers to investigate expression quantitative trait loci (eQTL) associations within a gene locus of interest in real time (Yang *et al.*, 2010). However, detecting genome-wide SNP–coexpression associations, a type of three-way associations, is more computationally intractable. Thus, a resource providing the results of a comprehensive analysis of SNP–coexpression associations, which are based on practical three-way association models and large-scale computation, should be very informative to the scientific community. In this study, we adopted an efficient method presented by Dettling *et al.* (2005), which enabled us to assess up to $\sim 10^{13}$ SNP-expression-expression combinations. The results are deposited in a freely available database named SNPxGE[2].

2 METHODS

2.1 The raw data

There are different expression datasets available on the HapMap samples [e.g. exon array (Duan *et al.*, 2008), Illumina (Stranger *et al.*, 2007a), Affy focus (Spielman *et al.*, 2007)]. In this study, normalized gene expression values (generated by Illumina Human WG-6 array) of 269 HapMap individuals from four populations including 90 CEPH trios of European descent (CEU), 90 Yoruba trios from Nigeria (YRI), 45 unrelated Han Chinese from Beijing (CHB) and 44 unrelated Japanese from Tokyo (JPT) were downloaded from GENEVAR (ftp:// ftp.sanger.ac.uk/pub/genevar/ CEU_children_ norm_march2007.zip, CEU_parents_norm_march2007.zip, CHB_norm_march2007.zip, JPT_ norm_march2007.zip, YRI_children_norm_march2007.zip & YRI_parents_ norm_march2007.zip) (Stranger *et al.*, 2007a; Yang *et al.*, 2010). Annotation of Illumina Human WG-6 array was downloaded from ftp://ftp.sanger.ac.uk/pub/genevar/illumina_Human_WG-6_array_content.csv. For each probe set (transcript), only the lowest level gene ontology (GO) terms within each of the three categories including biological process, molecular function and cellular component were considered. Out of the 47 294 transcripts, 15 000 genes with the highest expression level variations, which are more likely to have gene expression variations rather than random variations, were selected for further analysis. The SNP genotypes from phase I HapMap (ftp://ftp.ncbi.nlm.nih.gov/hapmap/genotypes/2005-06_16c_phaseI/full/redundant-filtered/) were used (Gibbs *et al.*, 2003). For detailed experimental procedures, please refer to Gibbs *et al.* (2003); Stranger *et al.* (2007a).

2.2 Modeling SNP–coexpression associations

The SNP–coexpression associations of interest, i.e. the differential coexpression between two genes is associated with the genotypes of an SNP, have been proven to exist in humans by Kayano *et al.* (2009) based on 142 gene expression profiles belonging to five disease pathways and 366 140 SNPs. However, Kayano *et al.*'s method, which incorporated several different statistical tests and logistic regression, was not efficient for a comprehensive study on a genome scale. Also, they addressed only one case of SNP–coexpression association. Therefore, we considered the approach introduced by Dettling *et al.* (2005), which consisted of single statistical tests and was proven to be as powerful as logistic regression, but at the same time the computational cost in terms of searching for differentially expressed gene combinations were considerably low. Although Dettling *et al.*'s work dealt with binary association, the three-class association problem in our SNP–coexpression associations can be modeled by finding the best one from the three possible binary associations.

For this study, we adopted the following convention: the correlation between two expression profiles was measured by Pearson's correlation coefficient. Given a dataset in which expression levels and genotypes are measured at once for each individual, for a pair of expression profiles E_1, E_2 and an SNP locus with genotypes AA, AB or BB, the genotype-dependent correlations between E_1 and E_2 (the correlations based on a subset of individuals with the same genotype) are denoted by R_{AA}, R_{AB} and R_{BB}, respectively, and the overall correlations between E_1 and E_2 in any two of the genotypes are denoted by $R_{\mathrm{AA+AB}}$, $R_{\mathrm{AA+BB}}$ and $R_{\mathrm{AB+BB}}$. The on/off model finds a maximum difference among genotype-dependent correlations, which is calculated as

$$\begin{aligned}\text{score\_}o = \max(&|R_{\mathrm{AA}} - R_{\mathrm{AB}}|,\\ &|R_{\mathrm{AA}} - R_{\mathrm{BB}}|, |R_{\mathrm{AB}} - R_{\mathrm{BB}}|)\end{aligned}$$

The on/off model could be regarded as an alternative approach to address the switch mechanism introduced by Kayano *et al.* (2009). The gap/substitution model tests whether the sum of genotype-dependent correlations is significantly higher than the overall correlation, which is calculated as

$$\begin{aligned}\text{score\_}g = \max(&|R_{\mathrm{AA}} + R_{\mathrm{AB}} - \alpha R_{\mathrm{AA+AB}}|,\\ &|R_{\mathrm{AA}} + R_{\mathrm{BB}} - \alpha R_{\mathrm{AA+BB}}|,\\ &|R_{\mathrm{AB}} + R_{\mathrm{BB}} - \alpha R_{\mathrm{AB+BB}}|)\end{aligned}$$

where α is a pre-defined constant. A good gap–substitution association may be interpreted as the expression patterns of two genes are well correlated under individual genotypes at a genomic locus, while their overall correlation cannot be observed. In the above equations, if the frequency of a genotype is $<10\%$, we only consider the interactions in the other two genotypes. In such case, the above equations become the original forms proposed by Dettling *et al.* (2005), i.e. assuming that BB is not considered,

$$\text{score\_}o = \left|R_{\mathrm{AA}} - R_{\mathrm{AB}}\right|$$

$$\text{score\_}g = \left|R_{\mathrm{AA}} + R_{\mathrm{AB}} - \alpha R_{\mathrm{AA+AB}}\right|.$$

Note that rare alleles/genotypes may also be involved in SNP–coexpression associations. We do not consider them in the current study because of frequent high correlations by random chance (i.e. the P-values of correlations are relatively high due to small degrees of freedom).

2.3 Simulation study for comparing different models for SNP–coexpression associations

Logistic regression was assumed to be a reasonable approach for modeling differential gene expression combinations and SNP–coexpression associations despite of heavy computational burden (Dettling *et al.*, 2005; Kayano *et al.*, 2009). To test the effectiveness of on/off and gap/substitution scores for modeling SNP–coexpression association, a simulation study was carried out. Every effort was made such that the simulation scenario will mimic the real data used in this study. The simulated data consisted of 269 individuals. SNP genotypes $g_j\,(1 \leqslant j \leqslant 269)$ were generated according to a pre-defined minor allele frequency (MAF) and assuming Hardy–Weinberg equilibrium. One thousand SNP–coexpression associations and a similar number of random SNP-expression-expression combinations were

simulated. In a random three-way combination, because the two genes were assumed to be not associated, their expression data were generated according to a normal distribution with mean and variance equal to 8 [denoted by $N(8, 8)$]. In an SNP–coexpression association, the expression data of the first gene was generated from $N(8, 8)$, whereas the expression data of second gene were generated by multiplying the expression levels of the first gene by a coefficient sampled from uniform distribution, $U[1.5,1.5]$, for each one of the three genotypes and then by adding a noise term generated from $N(0,8)$. Mathematically, the simulation process could be summarized as follows:

$$x_{1j} \sim N(8,8), \quad 1 \leqslant j \leqslant 269$$
$$a_j \sim U(-1.5,1.5), \quad 1 \leqslant j \leqslant 3$$
$$x_{2j} = \begin{cases} a_1 x_{1j} + e_j \text{ if } g_j = \text{AA} \\ a_2 x_{1j} + e_j \text{ if } g_j = \text{AB} \\ a_3 x_{1j} + e_j \text{ if } g_j = \text{BB} \end{cases} \quad 1 \leqslant j \leqslant 269,\ e_j \sim N(0,8)$$

To determine a proper α value for gap/substitution model, three choices including 1, 1.5 and 2 were tested. The fitness of logistic regression was evaluated using Akaike information criterion (AIC). A lower AIC indicates better fitness. For an SNP-expression-expression combination, the lowest AIC among three possible genotype combinations was used to represent its fitness.

Ten datasets using different MAF were simulated. The receiver operating characteristic (ROC) curves for comparing logistic regression, on/off model and gap/substitution model are shown in Figure 1. In general, smaller MAFs tend to result in lower performance. To show potential transition (between bad and good performance) of ROC curves, more figures were plotted between MAF = 0.1 and MAF = 0.3. The comparison shows that on/off and gap/substitution models are, at least, as powerful as logistic regression for detecting SNP–coexpression associations based on the data in this study. When MAF < 0.15, all models perform badly (i.e. close to the diagonal line). Thus, SNP loci with an MAF < 0.15 were removed from the HapMap genotype data. Then the loci in complete linkage disequilibrium were merged. The total numbers of available SNPs in the HapMap genotype data were 407 733, 360 109, 347 415, 481 321 and 556 873 for CEU, CHB, JPT, YRI and the pooled populations, respectively.

The performance of gap/substitution model is slightly affected by the choice of α. When MAF ranges from 0.2 to 0.275, the performance of gap/substitution model is comparable among different α values. When MAF $\leqslant$0.175 and MAF $\geqslant$3, gap/substitution models with α=2 and α=1 perform best, respectively. Dettling *et al.* (2005) suggested $\alpha = 1.5$ for detecting differential gene combinations. In this simulation study, gap/substitution model with $\alpha = 1.5$ is more stable across different MAFs, though not the best at each MAF. It seems that the results of this simulation study are not adequate for determining the best α value, although they generally show that the on/off and gap/substitution models can be effective in detecting SNP–coexpression associations. The determination of the best α value could be solved in real data applications.

2.4 Computational strategy for detecting SNP–coexpression associations

The SNP–coexpression associations were computed based on both single populations and pooled populations. To detect SNP–coexpression associations, any two transcripts that have the same GO terms (at the lowest levels) were considered for the subsequent downstream analysis. Note that extremely big GO terms (i.e. involving >900 of the 15 000 selected transcripts) were excluded. These restrictions resulted in 3 284 179 combinations of two expression profiles (out of 15 000 expression profiles) that share common GO terms. For each combination of two expression profiles, all available SNPs were assessed using on/off model and gap/substitution models with $\alpha = 1, 1.5$ and 2, and the best of each were recorded. This procedure was conducted to all single and the pooled populations, resulting in the assessment of $\sim 10^{13}$ SNP-expression-expression combinations. The computation time on 128

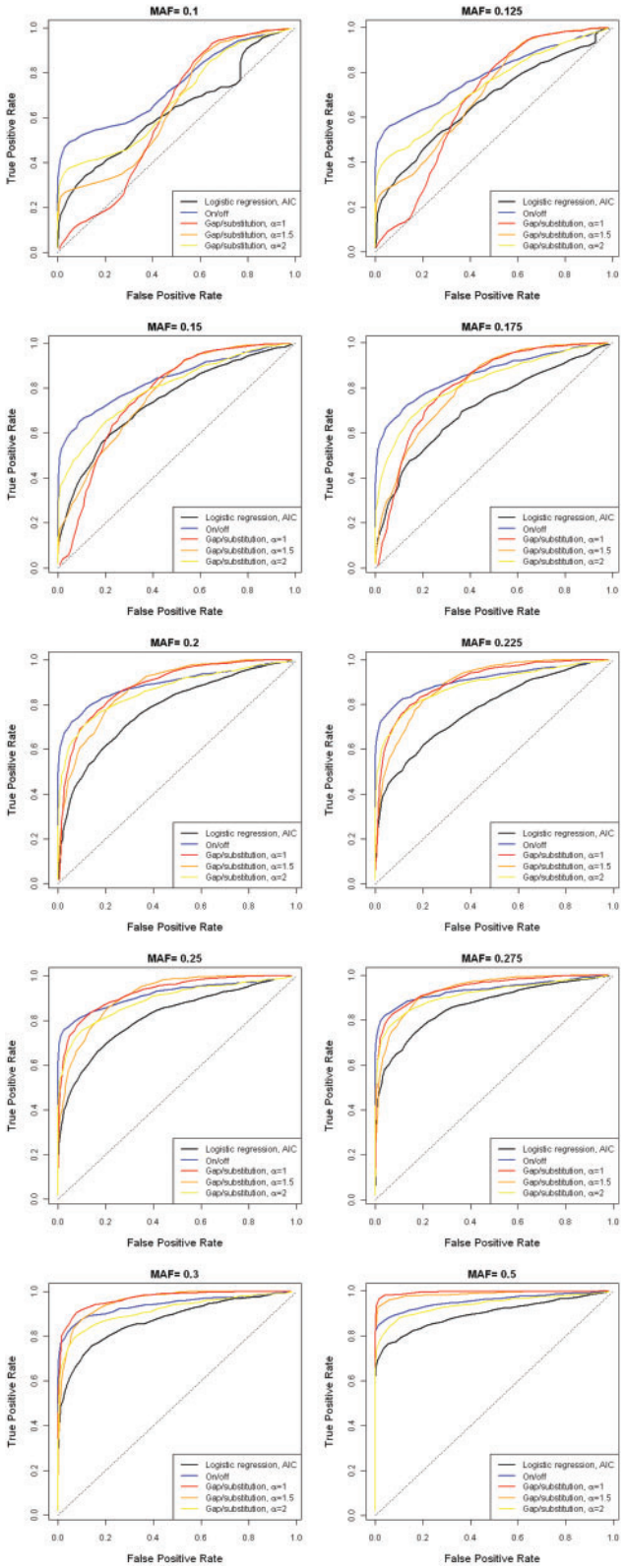

Fig. 1. ROC curves for comparing logistic regression, on/off model and gap/substitution model.

dual-processor quad-core nodes (Intel Xeon, 2 GB RAM/core) was ~15 days, compared to >1000 days if Kayano *et al.*'s method were used.

As a primary method for assessing the significance of SNP–coexpression associations, we permuted the genotype data and repeated the above procedure, to calculate an FDR for each potential SNP–coexpression association. Let score_*or* and score_*gr* and score_*op* and score_*gp* denote the on/off and gap/substitution scores on real genotype data and permuted genotype data, respectively. For an on/off score x and a gap/substitution score y, the FDRs can be computed according to the basic idea (Storey and Tibshirani, 2003):

$$\mathrm{FDR} = E[\frac{F}{S}]$$

where F is the number of false positives, and S is the total number of associations called significant. This formula is adapted to the following two equations for this study:

$$\frac{\#\{\text{score\_}op_i > x,\ i = 1,2,...,3284179\}}{\#\{\text{score\_}or_i > x,\ i = 1,2,...,3284179\}}$$

and

$$\frac{\#\{\text{score\_}gp_i > y,\ i = 1,2,...,3284179\}}{\#\{score\_gr_i > y,\ i = 1,2,...,3284179\}}$$

Since the procedure for FDR estimation has the same computational needs as the computation procedure on real genotype data, it is impractical to do multiple permutations. However, the FDR estimation in this study should be reliable because of the large number of permuted/uncorrelated SNP loci, equivalent to on average 10 000 permutations at an MAF interval of 0.01.

The significance of SNP–coexpression associations was also evaluated based on another permutation strategy. The top 100 000 SNP–coexpression associations based on each model were retrieved. For each of these associations, the genotypes of the SNP were permuted 1000 times, resulting in 1000 scores on null data. The null distribution of on/off or gap/substitution scores was simulated based on the 1000 permutation of top 100 000 associations, i.e. 10^8 null scores. The P-values for real scores were initially estimated by the number of null scores greater than a query real score divided by 10^8, and subsequently adjusted for multiple comparisons using Benjamini and Hochberg's method (Benjamini and Hochberg, 1995).

3 RESULTS

3.1 Identified SNP–coexpression associations

We first tested whether the identified SNP–coexpression associations are meaningful, i.e. distinguishable from those based on permuted SNP data. Comparisons of score distributions for top 20 000 SNP–coexpression associations between real and permuted genotype data are shown in Figure 2. In general, the distributions of on/off scores are indistinguishable between real and permuted genotype data, suggesting that human genome is likely to generate random on/off SNP–coexpression associations. For all populations, the distributions of gap/substitution scores based on real genotype data have obvious right shifts when compared with those based on permuted genotype data, indicating that the gap/substitution model is more powerful in detecting SNP–coexpression associations than the on/off model. The difference between two models may also suggest that in human genome, the scenario where genes show coexpression under individual genotypes but not under pooled genotypes is more widespread than the scenario where genes switch coexpression relations between positive and negative values under different genotypes.

In Figure 2, it appears that the difference of gap/substitution scores between real and permuted genotype data are generally larger at $\alpha = 1.5$ than at $\alpha = 1$ or 2. Further, with an FDR threshold of 0.1, gap/substitution model with $\alpha = 1.5$ generated much more SNP–coexpression associations than $\alpha = 1$ or 2 (Table 1). These analyses suggest that the gap/substitution model is more powerful at $\alpha = 1.5$ than at $\alpha = 1$ or 2. To provide more useful information, we considered the SNP–coexpression associations with FDR <0.1 under the gap/substitution model with $\alpha = 1.5$ as a reasonable set of results, including 44 769 and 50 792 SNP–coexpression associations based on single and pooled populations, respectively.

The HapMap populations are indeed different ethnic groups. The results suggest that SNP–coexpression associations exist both within an ethnic group and among ethnic groups. Further, we found that there are striking differences in the number of SNP–coexpression associations among ethnic groups. CHB and JPT have <2000 SNP–coexpression associations, whereas CEU and YRI have 9522 and 33 346 SNP–coexpression associations, respectively. Between ethnic groups, there may be up to ~87% common gene coexpression relations in the SNP–coexpression associations (Table 2), indicating that some SNP-dependent coexpression relations may be widespread in humans. However, we found that common coexpression relations are often associated with different loci in different ethnic groups, reflecting substantial genetic variations among ethnic groups. In addition, only ~40% coexpression relations are shared between single and pooled populations (Table 2), suggesting that a large number of SNP–coexpression associations can be detected only if genetic variations are increased when different ethnic groups are studied together.

Previous studies on associations between SNPs and expression levels show that associated SNPs tend to be proximal (*cis*-) to their respective genes (Duan *et al.*, 2008; Spielman *et al.*, 2007; Stranger *et al.*, 2005; Stranger *et al.*, 2007a; Stranger *et al.*, 2007b; Veyrieras *et al.*, 2008). Out of the 95 561 SNP–coexpression associations identified in this study, 48 143 have information of chromosomal locations according to the Ensembl database (http://www.ensembl.org/). We found that in 47 407 (98.5%) SNP–coexpression associations, the SNPs are located remotely from the coexpressed genes (i.e. not between 1 kb upstream and downstream of the coexpressed genes). In addition, in 41 406 (85.3%) SNP–coexpression associations, the SNPs are located on a different chromosome from the coexpressed genes. These analyses suggest that most of SNP–coexpression associations are *trans*-associations, distinct from the case of associations between SNPs and expression levels.

Further, we defined that an SNP is related to a gene if it is located between 1 kb upstream and downstream of the gene. Out of the 95 561 SNP–coexpression associations, there were 43 924 unique associated SNPs, which were found to be related to a total of 4644 unique genes. A GO term enrichment analysis was applied to these SNP-related genes using Fisher's exact test with Bonferroni correction. The top 10 associated GO terms (not including cellular component) are summarized in Table 3. It appears that the SNP-related genes are more likely to be regulatory genes (i.e. genes that regulate the activity of other genes), which are widely involved in binding, signal transduction and transcriptional regulation.

3.2 Availability of the database

The identified SNP–coexpression associations were deposited in the SNPxGE^2 database. The SNPxGE^2 database is freely available at http://lambchop.ads.uga.edu/snpxge2/index.php. The SNP–coexpression associations can be searched on website via

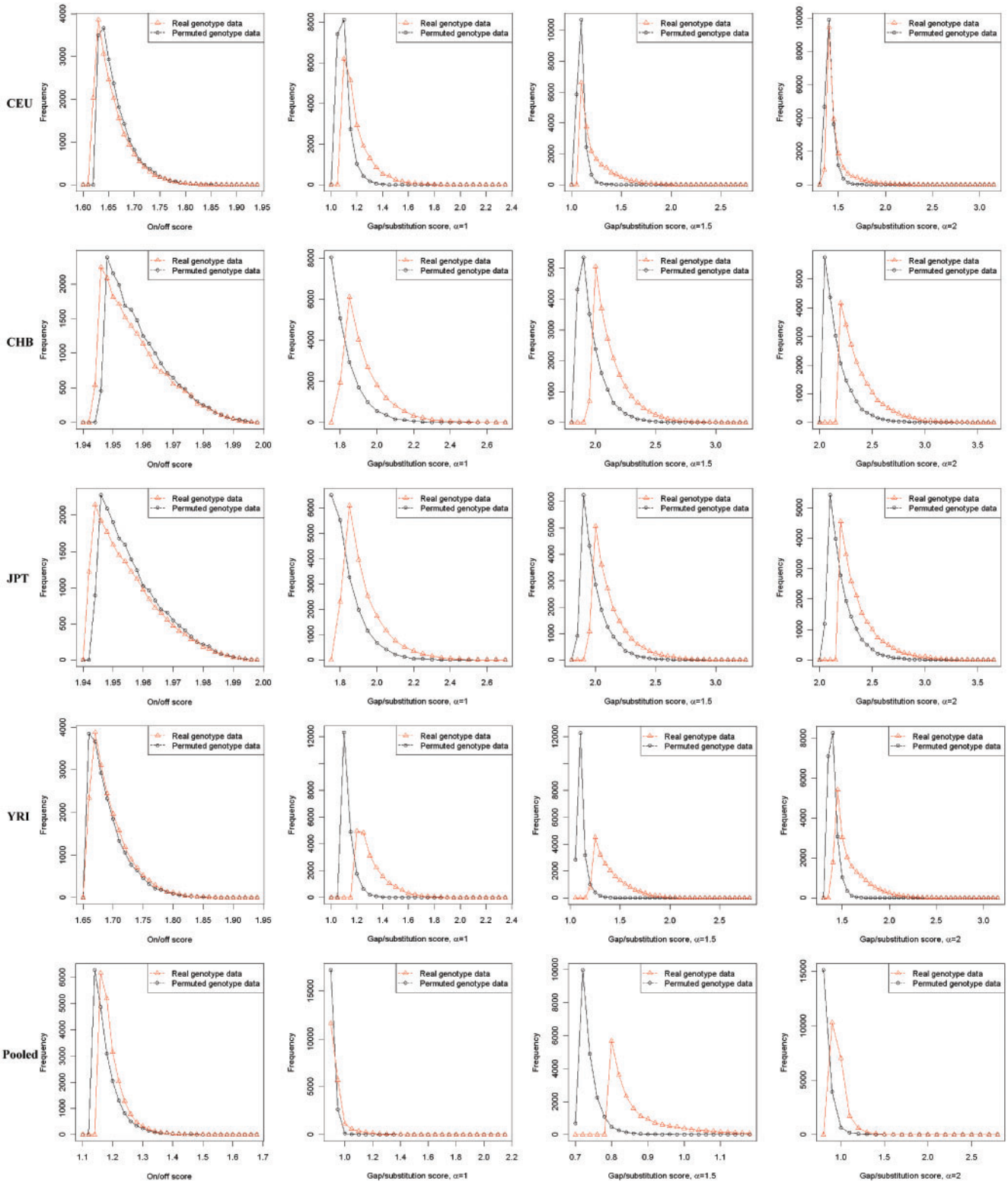

Fig. 2. Comparison of the score distributions of top 20 000 SNP–coexpression associations between real and permuted genotype data based on different statistical models.

gene symbol or reference SNP ID, which represent units of SNP–coexpression associations. Alternatively, all the associations can be downloaded as a plain text file. On the home page, a quick search engine which supports keyword search and batch search is provided. An advanced search page is also provided, which allows the user to choose a cutoff FDR, to search by GO terms or to use exact search. When the user submits a query (e.g. XPO4) on the home page, a brief information page is returned (Fig. 3), on which the icon for a particular association can be clicked on for detailed information. On the detailed information page (Fig. 4), three pieces of information are provided: (i) parameters for the SNP–coexpression association, including population, gap/substitution score, *P*-value, FDR and associated GO terms; (ii) genomic information for the three units of the SNP–coexpression association, including positions, RefSeq IDs and Ensembl IDs for the two transcripts, and position, function and related gene (if the SNP is located between 1 kb upstream and downstream of a gene) for the SNP; (iii) a plot showing the SNP–coexpression association, where points represent HapMap individuals, which are marked in red, blue or green corresponding to different genotypes. Note that the gap/substitution score is computed based on the best two genotype combinations out of three possible combinations. The expression correlations and regression lines under the two genotypes that comprise the gap/substitution model are shown.

Table 1. Number of SNP–coexpression associations identified by gap/substitution models with different α values

Population	Number of SNP–coexpression associations		
	$\alpha=1$	$\alpha=1.5$	$\alpha=2$
CEU	5066	9522	2977
CHB	816	1427	2337
JPT	333	474	809
YRI	20 216	33 346	11 876
Pooled	3495	50 792	51
Total	29 926	95 561	18 050

3.3 A stand-alone tool for analyzing any two transcripts

In the current study, in order to make the genome-wide SNP–coexpression association studies feasible based on our current computational capabilities, the two transcripts of an SNP–coexpression association were restricted to the same GO terms. However, we are aware that the two transcripts that are associated with the genotypes of an SNP may belong to different GO terms. Thus, we provide a stand-alone program on the 'download' page, for the search of associated SNPs for any two transcripts of interest. The usage of the program is described in its incorporated manual. The user may use the 'download' web page to plot an SNP–coexpression association of interest.

4 DISCUSSION

Unlike traditional eQTL analyses, in this study, we have associated differential coexpression relations between two genes with the genotypes of an SNP on a genome scale. We found that most of the associated SNPs in SNP–coexpression associations are located remotely from the coexpressed genes, and that the SNP-related genes are enriched with regulatory functions. A potential mechanism underlying SNP–coexpression associations could be that

Table 2. Common coexpression relations in the SNP–coexpression associations between ethnic groups

Population	No. of SNP–coexpression associations	No. of common coexpression relations			
		CEU	CHB	JPT	YRI
CEU	9522	–	–	–	–
CHB	1427	700	–	–	–
JPT	474	362	96	–	–
YRI	33 346	7653	858	414	–
Pooled	50 792	7086	801	392	10 145

Table 3. Top 10-enriched GO terms associated with the SNP-related genes in SNP–coexpression associations

GO term	Description	*P*-value
GO:0 005 515	Protein binding (F)	0
GO:0 046 872	Metal ion binding (F)	7.49×10^{-84}
GO:0 000 166	Nucleotide binding (F)	5.11×10^{-77}
GO:0 007 165	Signal transduction (P)	7.62×10^{-74}
GO:0 005 524	ATP binding (F)	7.25×10^{-66}
GO:0 007 155	Cell adhesion (P)	3.29×10^{-44}
GO:0 008 270	Zinc ion binding (F)	4.92×10^{-41}
GO:0 006 468	Protein amino acid phosphorylation (P)	3.33×10^{-38}
GO:0 006 355	Regulation of transcription, DNA-dependent (P)	8.74×10^{-38}
GO:0 006 811	Ion transport (P)	1.05×10^{-37}

F, molecular function; P, biological process.

Index	Population	Expression #1	Expression #2	SNP	SNP-related gene	Score	FDR	Details
1	CEU	SEC13L	XPO4	rs861085		1.23973	0.053757	
2	YRI	XPO4	STX10	rs11795672		1.44112	0.004162	
3	YRI	STX16	XPO4	rs11795672		1.20386	0.061733	
4	pooled	XPO4	TIMM22	rs2586306	ABR	0.8869	0.021371	
5	pooled	SEC5	XPO4	rs6090449	SRMS	0.850816	0.029481	
6	pooled	XPO4	BRDG1	rs10764383	MSRB2	0.79381	0.060610	
7	pooled	KDELR1	XPO4	rs2002157	RP11-291B21.2	0.793012	0.061313	
8	pooled	APBA3	XPO4	rs616153	KB-1562D12.1	0.774794	0.077517	
9	pooled	SEC8	XPO4	rs10000062	STK32B	0.767968	0.085403	

Fig. 3. Brief information page.

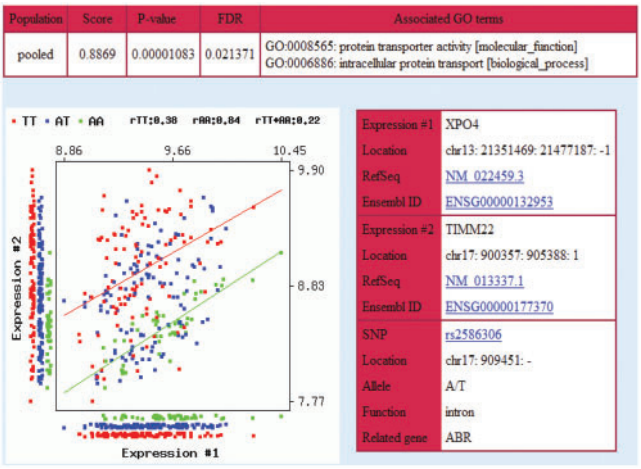

Fig. 4. Detailed information page.

the associated SNPs affect the functions or expression of their related genes, which are regulators of downstream coexpressed genes, and thereby these associated SNPs could fine tune coexpression relationships between genes.

Gene expression is important for understanding the genetics of complex diseases. This study suggests that natural variations in human gene expression may exist in coexpression relationships. The analysis of genetically determined variation in coexpression relationships may enhance understanding of both the underlying genetics and the population differences in complex diseases.

The SNPxGE^2 database provides information on computationally predicted human SNP–coexpression associations. The interfaces of SNPxGE^2 are friendly and easy to use. Biologists may use this database to find whether their genes of interest have coexpressed/interacting partners associated with a particular distribution of genotypes, and/or their potential upstream regulatory genes, which are difficult to detect using routine correlation analyses. Biologists may also use this database to study potentially dynamic and complex relationships among the members of a biological process using GO term search.

SNPxGE^2 is a continuing project and is expected to grow substantially over the coming years as next-generation sequencing technologies like Illumina, 454 Plus PacBio and Ion Torrent have made the data generation cheaper and newer mapping technologies will lead to enormous amounts of RNASeq data for studying gene expression (Joseph and Read, 2010; Langmead *et al.*, 2010), which is quantitative enough to be included in human SNP–coexpression association analyses. Furthermore, mapping differential gene coexpression to more genetic features like DNA copy numbers (generated from array CGH data) is also under investigation and will be incorporated into SNPxGE^2 in the next release. In addition, this study establishes preliminary results for constructing human genotype-dependent coexpression networks, which will be carried out more extensively in the future.

ACKNOWLEDGEMENTS

We thank Drs Wensheng Zhang and Xiyin Wang for critical reading of the manuscript.

Funding: This study was supported in part by resources and technical expertise from the University of Georgia Research Computing Center, a partnership between the Office of the Vice President for Research and the Office of the Chief Information Officer.

Conflict of Interest: none declared.

REFERENCES

Adryan,B. and Teichmann,S.A. (2010) The developmental expression dynamics of Drosophila melanogaster transcription factors. *Genome Biol.*, **11**, R40.

Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate - a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B*, **57**, 289–300.

Cheung,V.G. *et al.* (2003) Natural variation in human gene expression assessed in lymphoblastoid cells. *Nat. Genet.*, **33**, 422–425.

Choi,J.K. *et al.* (2005) Differential coexpression analysis using microarray data and its application to human cancer. *Bioinformatics*, **21**, 4348–4355.

Dettling,M. *et al.* (2005) Searching for differentially expressed gene combinations. *Genome Biol.*, **6**, R88.

Duan,S. *et al.* (2008) Genetic architecture of transcript-level variation in humans. *Am. J. Hum. Genet.*, **82**, 1101–1113.

Gamazon,E.R. *et al.* (2010) SCAN: SNP and copy number annotation. *Bioinformatics*, **26**, 259–262.

Gibbs,R.A. *et al.* (2003) The International HapMap Project. *Nature*, **426**, 789–796.

Idaghdour,Y. *et al.* (2010) Geographical genomics of human leukocyte gene expression variation in southern Morocco. *Nat. Genet.*, **42**, 62–67.

Joseph,S.J. and Read,T.D. (2010) Bacterial population genomics and infectious disease diagnostics. *Trends Biotechnol.*, **28**, 611–618.

Kayano,M. *et al.* (2009) Efficiently finding genome-wide three-way gene interactions from transcript- and genotype-data. *Bioinformatics*, **25**, 2735–2743.

Langmead,B. *et al.* (2010) Cloud-scale RNA-sequencing differential expression analysis with Myrna. *Genome Biol.*, **11**, R83.

Lee,H.K. *et al.* (2004) Coexpression analysis of human genes across many microarray data sets. *Genome Res.*, **14**, 1085–1094.

Li,K.C. *et al.* (2004) A system for enhancing genome-wide coexpression dynamics study. *Proc. Natl Acad. Sci. USA*, **101**, 15561–15566.

Nayak,R.R. *et al.* (2009) Coexpression network based on natural variation in human gene expression reveals gene interactions and functions. *Genome Res.*, **19**, 1953–1962.

Obayashi,T. and Kinoshita,K. (2011) COXPRESdb: a database to compare gene coexpression in seven model animals. *Nucleic Acids Res.*, **39**, D1016–D1022.

Pickrell,J.K. *et al.* (2010) Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature*, **464**, 768–772.

Spielman,R.S. *et al.* (2007) Common genetic variants account for differences in gene expression among ethnic groups. *Nat. Genet.*, **39**, 226–231.

Storey,J.D. and Tibshirani,R. (2003) Statistical significance for genomewide studies. *Proc. Natl Acad. Sci. USA*, **100**, 9440–9445.

Stranger,B.E. *et al.* (2005) Genome-wide associations of gene expression variation in humans. *PLoS Genet.*, **1**, e78.

Stranger,B.E. *et al.* (2007a) Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science*, **315**, 848–853.

Stranger,B.E. *et al.* (2007b) Population genomics of human gene expression. *Nat. Genet.*, **39**, 1217–1224.

Veyrieras,J.B. *et al.* (2008) High-resolution mapping of expression-QTLs yields insight into human gene regulation. *PLoS Genet.*, **4**, e1000214.

Weirauch,M.T. and Hughes,T.R. (2010) Conserved expression without conserved regulatory sequence: the more things change, the more they stay the same. *Trends Genet.*, **26**, 66–74.

Wilson,M.D. and Odom,D.T. (2009) Evolution of transcriptional control in mammals. *Curr. Opin. Genet. Dev.*, **19**, 579–585.

Wray,G.A. *et al.* (2003) The evolution of transcriptional regulation in eukaryotes. *Mol. Biol. Evol.*, **20**, 1377–1419.

Yang,T.P. *et al.* (2010) Genevar: a database and Java application for the analysis and visualization of SNP-gene associations in eQTL studies. *Bioinformatics*, **26**, 2474–2476.

Zhang,W. *et al.* (2008) Evaluation of genetic variation contributing to differences in gene expression between populations. *Am. J. Hum. Genet.*, **82**, 631–640.