SNPxGE²: a database for human 3-way SNP-expression associations

Yupeng Wang¹,², Sandeep J Joseph³, Xinyu Liu¹,², Michael Kelly¹ and Romdhane Rekaya¹,²,4,*

¹Department of Animal and Dairy Science, University of Georgia, Athens, GA30605, USA

²Institute of Bioinformatics, University of Georgia, Athens, GA30605, USA

³Department of Medicine, Division of Infectious Diseases, Emory University School of Medicine, Atlanta, GA30322, USA

⁴Department of Statistics, University of Georgia, Athens, GA30605, USA

*Corresponding author

E-mails:
YW: wyp1125@uga.edu
SJ: sandeepjosejoseph@gmail.com
XL: xinyu81@uga.edu
MK: mck@uga.edu
RR: rrekaya@uga.edu
Abstract

Recently, gene co-expression relationships have been found to be often conditional and dynamic. Besides, many studies have suggested that single nucleotide polymorphisms (SNPs) have impacts on gene expression variations in human populations. The SNPxGE² database contains the computationally predicted human 3-way SNP-expression associations, that is, the differential co-expression between 2 genes is associated with a genotype/SNP. This data was generated from a large scale association study that was based on the HapMap data, which covered 269 individuals from 4 human populations, 701,202 SNPs and 15,000 gene expression profiles. Two models of 3-way SNP-expression associations were considered: gap/substitution and on/off. The implementation was carried out using 64 Linux cluster nodes in ~30 days and assessed a total of $9.21 \times 10^{12}$ SNP-expression combinations. The results, including 4,713 on/off associations and 36,170 gap/substitution associations at a p-value cutoff of 0.001, can be queried in the SNPxGE² database via either gene name or reference SNP ID. For each reported association, a detailed information page is provided. The SNPxGE² database can be freely accessed at http://tunisia.ads.uga.edu/SNPxGE2/index.php.
Introduction

It is widely accepted that eukaryotic gene expression variation may be potentially modulated by many genetic and epigenetics factors such as cis regulatory elements, enhancers, transposons, nucleosome positioning, DNA methylation, and by environmental factors (1-4). Moreover, much of the expression variations are heritable (5).

Recently, many studies have suggested that SNPs also have impacts on gene expression variations in human populations (6-9). Stranger at al. (6) found that many regions within 1Mb of 374 expressed genes of interest had significant association of SNPs with expression variation. Stranger et al. (7), performing association analyses of expression levels of 14,925 transcripts with SNPs and copy number variants (CNVs) using multiple linear regression, suggested that SNPs captured 83.6% of the total detected genetic variation in gene expression. Stranger et al. (8) carried out an association analysis of over 2.2 million common SNPs with gene expression using multiple liner regression and identified at least 1,348 genes with association signals in cis and at least 180 in trans. Veyrieras et al. (9), using a Bayesian hierarchical model, found strong enrichments of quantitative trait loci for gene expression (eQTLs) in the 250 bp upstream of transcription end site (TES) and around the transcription start sites (TSS). Pickrell et al. (10) identified more than a thousand genes at which genetic variation influences overall expression levels or splicing.

Besides individual gene expression profiles, the relationships among them such as co-expression are also often studied for deciphering gene regulatory mechanisms. The identification of human gene co-expression relationships using microarray data has often relied on reliable correlations between gene expression profiles across many experiments or conditions (11,12). But recently, some studies have suggested that the co-expression relationships between genes are often dynamic and conditional, e.g. dependent on cellular states (13), developmental stages (14), disease status (health or cancer) (15), or human populations (16). However, few studies have linked such differential co-expression to genetic bases. SNP effects on gene expression variation in humans has been well
documented, and hence, it is reasonable to conjecture that the differential co-expression between two genes may be associated with the genotypes of an SNP, which is termed as a 3-way SNP-expression association in this study. Kayano et al. (17) proved a biological switching mechanism in expression between correlation and inverse-correlation of two genes, controlled by a genomic SNP. However, the switching mechanism in two genes’ co-expression controlled by a SNP could be only part of the whole picture of 3-way SNP-expression associations because it is also possible that two gene are well co-expressed under different genotypes of an SNP while their co-expression can not be detected if different genotypes of the SNP are pooled, or that two gene are well co-expressed under 1 or 2 genotypes of an SNP while are not co-expressed under others. Besides, Kayano et al.’s method was only able to assess $3 \times 10^8$ three-way 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. The biggest difficulty in conducting 3-way association analysis on a genome-wide scale is computational intractability. Thus, a resource providing the results of a comprehensive analysis of 3-way SNP-expression associations, which are based on more practical association models and large-scale computation, should be informative to the scientific community. To this end, we adopted a more efficient method presented by Dettling et al. (18), which enabled us to assess up to $9.21 \times 10^{12}$ three-way combinations. The results were deposited in the SNPxGE$^2$ database.

**Construction**

**The raw data**

Normalized gene expression values of 269 HapMap individuals from 4 populations (CEU, CHB, JPT, YRI) were downloaded from GENEVAR (http://www.sanger.ac.uk/genevar/) (19). Out of the 47,294 transcripts, 15,000 genes with the highest variations were selected for further analysis. The SNP genotypes from phase I HapMap (http://hapmap.ncbi.nlm.nih.gov/) were used. Data from the four populations were pooled together to make the analysis more reliable (correlations with small sample sizes are unstable).
Modeling 3-way SNP-expression associations

The 3-way SNP-expression associations of interest, that is, the differential co-expression between 2 genes was associated with a genomic genotype / SNP, have been proven to exist in humans by Kayano et al. (17) based on 142 gene expression profiles out of 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. Besides, they addressed only one type of 3-way SNP-expression association. Thus, we considered the approaches introduced by Dettling et al. (18), which were single statistical testes and were proven to be as powerful as logistic regression, but have a markedly lower computational cost in terms of searching for differentially expressed gene combinations. Although Dettling et al.’s work dealt with binary association, the 3-class association problem in our 3-way SNP-expression associations can be solved by finding the best one from 3 possible binary associations.

For this study we adopted the following convention. The correlation between gene two expression profiles was measured by Pearson’s correlation coefficient. Given a dataset in which expressions 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 genotypes are denoted by $R_{AA+AB}$, $R_{AA+BB}$ and $R_{AB+BB}$. The on/off model finds a maximum difference among genotype dependent correlations, which is calculated as

$$score_o = \max(|R_{AA} - R_{AB}|, |R_{AA} - R_{BB}|, |R_{AB} - R_{BB}|)$$

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

$$score_g = \max(|R_{AA} + R_{AB} - \alpha R_{AA+AB}|, |R_{AA} + R_{BB} - \alpha R_{AA+BB}|, |R_{AB} + R_{BB} - \alpha R_{AB+BB}|)$$
where \( \alpha = 1.5 \) as suggested by Dettling et al. A good gap/substitution association may be interpreted as: the expression patterns of two genes are well correlated under different genotypes at a genomic locus, while their overall correlation can not be observed. Note that in the above equations, genotype dependent correlations with less than 27 samples were not counted because small sample sizes increase the instability of correlations.

**The computational strategy for detecting 3-way SNP-expression associations**

To detect 3-way SNP-expression associations, any two expression profiles that have the same GO term were considered for the subsequent downstream analysis. This restriction resulted in 3,284,179 combinations of 2 expression profiles (out of 15,000 expression profiles) that share common GO terms. For each such combination of 2 expression profiles, all of the 701,202 SNPs were assessed using the two association models: on/off and gap/substitution models and the best of each were recorded. To find the significant 3-way associations, we permuted the genotype data and the above procedure was repeated. This strategy resulted in the assessment of \( 9.21 \times 10^{12} \) SNP-expression combinations, that is:

\[
3,284,179 \text{(expression profile combinations)} \times 701,202 \text{(SNPs)} \times 2 \text{(association models)} \times 2 \text{(real and permuted genotype data)} = 9.211492 \times 10^{12}
\]

The computation time on a 64 dual-processor quad-core nodes (Intel Xeon, 2GB RAM/core) was about 30 days, compared to more than 1000 days if Kayano et al.’s method were used.

The distributions of on/off scores and gap/substitution scores based on real genotype data showed a clear right shift compared with those based on permuted genotype data (Figure 1), which suggests the existence of 3-way associations between differential gene co-expression and SNPs in human genome. Let \( \text{score}_\text{or} \) and \( \text{score}_\text{op} \) denote the on/off scores on real genotype data and permuted genotype data respectively. For an on/off score \( x \), the p-value was assessed by:

\[
\frac{\# \{ \text{score}_\text{op}_i > x, \ i = 1, 2, ..., 3284179 \}}{3284179}
\]

and the false discovery rate (FDR) was estimated by (20)
Figure 1. Comparison of score distributions between real genotype data and permuted genotype data.

\[
\frac{\#\{\text{score}_{op_i} > x, i = 1,2,\ldots, 3284179\}}{\#\{\text{score}_{or_i} > x, i = 1,2,\ldots, 3284179\}}
\]

The p-value and FDR for a gap/substitution score were computed in a similar manner.

**Utility**

**Availability of the database**

The results, including 4,731 on/off associations and 36,170 gap/substitution associations at a p-value cutoff of 0.001, were deposited in the SNPxGE^2^ database. Note that occasionally a reported 3-way association in SNPxGE^2^ may have a p-value>0.001 because it is also suspected to have such association based on the fact that its score is not high, but is much higher than its own expected maximum (the best SNP association for the 2 expression profiles based on permuted genotype data).

The SNPxGE^2^ database is freely available at [http://tunisia.ads.uga.edu/SNPxGE2/index.php](http://tunisia.ads.uga.edu/SNPxGE2/index.php). The 3-way SNP-expression associations can be searched via gene name or reference SNP ID, which represent units of 3-way
associations. 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 specific 3-way association model and cutoff p-value, to search by gene ontology (GO) terms, or to use exact search. When the user submits a query (e.g. KPNB1) on the home page, a brief information page is returned (Figure 2), on which the icon for a particular association can be clicked on for detailed information. On the detailed information page (Figure 3), three pieces of information are provided:

1) Parameters for the 3-way association, including the 3-way association model (on/off or gap/substitution), score, expected maximum score based on permuted genotype data, p-value, FDR and associated GO terms.

| Index | Expression #1 | Expression #2 | Control SNP | Related gene | Score | P-value | Details |
|-------|---------------|---------------|-------------|--------------|-------|---------|---------|
| 1     | KPNB1         | SCFD2         | rs4861533   |              | 1.30799 | 0.000211 |         |
| 2     | KPNB1         | NAPB          | rs2019502   |              | 1.24493 | 0.000737 |         |

| Index | Expression #1 | Expression #2 | Control SNP | Related gene | Score | P-value | Details |
|-------|---------------|---------------|-------------|--------------|-------|---------|---------|
| 1     | HDAC6         | KPNB1         | rs10928056  |              | 0.830513 | 0.000164 |         |
| 2     | APOBEC3G      | KPNB1         | rs7183178   |              | 0.773439 | 0.000955 |         |
| 3     | PRDM2         | KPNB1         | rs2075052   |              | 0.801297 | 0.000386 |         |
| 4     | GGA1          | KPNB1         | rs4131470   |              | 0.772495 | 0.000988 |         |
| 5     | APBA3         | KPNB1         | rs572589    |              | 0.773978 | 0.000938 |         |
| 6     | KPNB1         | FGD3          | rs10928056  |              | 0.859775 | 0.000075 |         |
| 7     | KPNB1         | RNF19         | rs705308    |              | 0.831035 | 0.000161 |         |
| 8     | KPNB1         | SCFD2         | rs4861533   |              | 0.773834 | 0.000941 |         |

Figure 2. The brief information page.
2) Genomic information for the 3 units of the 3-way association, including positions, RefSeq IDs and Ensembl IDs for the 2 transcripts, and position, function and related gene for the SNP.

3) A plot showing the 3-way SNP-expression association.

**A stand-alone tool for analyzing any two transcripts**

In order to make the genome-wide 3-way association studies feasible based on our current computational capabilities, the 2 transcripts of a 3-way association were restricted to the same GO terms, in this study. However, we are aware that the 2 transcripts that are associated with a genomic SNP may be also in different GO terms. Thus, we provide a stand-alone program on the “download” page, for the search of associated SNPs for any 2 transcripts of interest. The usage of the program is described in its incorporated manual. The user may use the “download” web page to plot a 3-way association of interest.
Conclusion and future plans

The SNPxGE\(^2\) database provides information on computationally predicted human 3-way SNP-expression associations. The interfaces of SNPxGE\(^2\) are friendly and easy to use. SNPxGE\(^2\) is a continuing project and is expected to grow substantially over the coming years as next-generation sequencing technologies like Illumina and SOLID have made the data generation cheaper and newer mapping technologies will lead to enormous amounts of RNASeq data for studying gene expression (21), which is quantitative enough to be included in human 3-way SNP-expression association analyses. Furthermore, mapping differential gene co-expression to more genetic features like DNP copy numbers (array CGH data) is also under investigation and is planned to be incorporated into SNPxGE\(^2\) in future.

Availability and requirements

Project name: SNPxGE\(^2\)

Project home page: [http://tunisia.ads.uga.edu/SNPxGE2/index.php](http://tunisia.ads.uga.edu/SNPxGE2/index.php)

Operating system(s): Platform independent

Programming language: C++

Other requirements: None

License: None for usage

Any restrictions to use by non-academics: None

Acknowledgements

We thank Dr. Wensheng Zhang, Dr. Xiyin Wang for critical reading of the manuscript. 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.
References

1. Wray, G.A., Hahn, M.W., Abouheif, E., Balhoff, J.P., Pizer, M., Rockman, M.V. and Romano, L.A. (2003) The evolution of transcriptional regulation in eukaryotes. *Mol Biol Evol*, 20, 1377-1419.

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

3. Idaghdour, Y., Czika, W., Shianna, K.V., Lee, S.H., Visscher, P.M., Martin, H.C., Miclaus, K., Jadallah, S.J., Goldstein, D.B., Wolfinger, R.D. *et al.* (2010) Geographical genomics of human leukocyte gene expression variation in southern Morocco. *Nat Genet*, 42, 62-67.

4. 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.

5. Cheung, V.G., Conlin, L.K., Weber, T.M., Arcaro, M., Jen, K.Y., Morley, M. and Spielman, R.S. (2003) Natural variation in human gene expression assessed in lymphoblastoid cells. *Nat Genet*, 33, 422-425.

6. Stranger, B.E., Forrest, M.S., Clark, A.G., Minichiello, M.J., Deutsch, S., Lyle, R., Hunt, S., Kahl, B., Antonarakis, S.E., Tavare, S. *et al.* (2005) Genome-wide associations of gene expression variation in humans. *PLoS Genet*, 1, e78.

7. Stranger, B.E., Forrest, M.S., Dunning, M., Ingle, C.E., Beazley, C., Thorne, N., Redon, R., Bird, C.P., de Grassi, A., Lee, C. *et al.* (2007) Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science*, 315, 848-853.

8. Stranger, B.E., Nica, A.C., Forrest, M.S., Dimas, A., Bird, C.P., Beazley, C., Ingle, C.E., Dunning, M., Flicek, P., Koller, D. *et al.* (2007) Population genomics of human gene expression. *Nat Genet*, 39, 1217-1224.

9. Veyrieras, J.B., Kudaravalli, S., Kim, S.Y., Dermitzakis, E.T., Gilad, Y., Stephens, M. and Pritchard, J.K. (2008) High-resolution mapping of expression-QTLs yields insight into human gene regulation. *PLoS Genet*, 4, e1000214.

10. Pickrell, J.K., Marioni, J.C., Pai, A.A., Degner, J.F., Engelhardt, B.E., Nkadori, E., Veyrieras, J.B., Stephens, M., Gilad, Y. and Pritchard, J.K. (2010) Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature*, 464, 768-772.

11. Lee, H.K., Hsu, A.K., Sajdak, J., Qin, J. and Pavlidis, P. (2004) Coexpression analysis of human genes across many microarray data sets. *Genome Res*, 14, 1085-1094.

12. Obayashi, T., Hayashi, S., Shibaoka, M., Saeki, M., Ohta, H. and Kinoshita, K. (2008) COXPRESdb: a database of coexpressed gene networks in mammals. *Nucleic Acids Res*, 36, D77-82.

13. Li, K.C., Liu, C.T., Sun, W., Yuan, S. and Yu, T. (2004) A system for enhancing genome-wide coexpression dynamics study. *Proc Natl Acad Sci U S A*, 101, 15561-15566.
14. Adryan, B. and Teichmann, S.A. (2010) The developmental expression dynamics of Drosophila melanogaster transcription factors. *Genome Biol*, 11, R40.

15. Choi, J.K., Yu, U.S., Yoo, O.J. and Kim, S. (2005) Differential coexpression analysis using microarray data and its application to human cancer. *Bioinformatics*, 21, 4348-4355.

16. Nayak, R.R., Kearns, M., Spielman, R.S. and Cheung, V.G. (2009) Coexpression network based on natural variation in human gene expression reveals gene interactions and functions. *Genome Research*, 19, 1953-1962.

17. Kayano, M., Takigawa, I., Shiga, M., Tsuda, K. and Mamitsuka, H. (2009) Efficiently finding genome-wide three-way gene interactions from transcript- and genotype-data. *Bioinformatics*, 25, 2735-2743.

18. Dettling, M., Gabrielson, E. and Parmigiani, G. (2005) Searching for differentially expressed gene combinations. *Genome Biol*, 6, R88.

19. Thorisson, G.A., Smith, A.V., Krishnan, L. and Stein, L.D. (2005) The International HapMap Project Web site. *Genome Research*, 15, 1592-1593.

20. Manly, K.F., Nettleton, D. and Hwang, J.T. (2004) Genomics, prior probability, and statistical tests of multiple hypotheses. *Genome Res*, 14, 997-1001.

21. Langmead, B., Hansen, K.D. and Leek, J.T. (2010) Cloud-scale RNA-sequencing differential expression analysis with Myrna. *Genome Biol*, 11, R83.