Genome-Wide Association Analyses of Expression Phenotypes

Gary K. Chen,[1] Tian Zheng,[2] John S. Witte,[1]* and Ellen L. Goode[3] on behalf of Group 1[1]†

[1] *Department of Epidemiology and Biostatistics, Institute for Human Genetics, University of California, San Francisco, California*
[2] *Department of Statistics, Columbia University, New York, New York*
[3] *Department of Health Sciences Research, Mayo Clinic College of Medicine, Rochester, Minnesota*

A number of issues arise when analyzing the large amount of data from high-throughput genotype and expression microarray experiments, including design and interpretation of genome-wide association studies of expression phenotypes. These issues were considered by contributions submitted to Group 1 of the Genetic Analysis Workshop 15 (GAW15), which focused on the association of quantitative expression data. These contributions evaluated diverse hypotheses, including those relevant to cancer and obesity research, and used various analytic techniques, many of which were derived from information theory. Several observations from these reports stand out. First, one needs to consider the genetic model of the trait of interest and carefully select which single nucleotide polymorphisms and individuals are included early in the design stage of a study. Second, by targeting specific pathways when analyzing genome-wide data, one can generate more interpretable results than agnostic approaches. Finally, for datasets with small sample sizes but a large number of features like the Genetic Analysis Workshop 15 dataset, machine learning approaches may be more practical than traditional parametric approaches. *Genet Epidemiol* 31 (Suppl. 1): S7–S11, 2007. © 2007 Wiley-Liss, Inc.

Key words: Genetic Analysis Workshop; linkage; association; machine learning approaches; expression data; single nucleotide polymorphisms

*Correspondence to: John S. Witte, UC San Francisco, 513 Parnassus Avenue S-965, San Francisco, CA 94143-0794. E-mail: jwitte@ucsf.edu

†A complete list of the Group 1 primary contributing authors is given in the Acknowledgment section.

Published online in Wiley InterScience (www.interscience.wiley.com).
DOI: 10.1002/gepi.20275

INTRODUCTION

Genotype and expression arrays provide researchers with a wealth of data for deciphering the genetic basis of common diseases. The enormity of data leads to a number of challenges regarding how to best analyze the ensuing data and properly interpret results. These issues were addressed at the Genetic Analysis Workshop 15 [Cordell et al., 2007], including by the 10 presentations focused on association testing of expression data [Group 1; Chen et al., 2007; Goode et al., 2007; Hu et al., 2007; Suh et al., 2007; Szymczak et al., 2007; Zheng et al., 2007].

The data analyzed in this group consisted of 3,554 transcript expression levels and 2,882 single nucleotide polymorphism (SNP) genotypes on 194 individuals in 14 Centre d'Etude du Polymorphisme Humain families from Utah [Cheung et al., 2005; Cheung and Spielman, 2007]. Some participants [Chen et al., 2007; Goode et al., 2007] included SNP genotype data from HapMap [International HapMap Consortium, 2005] as well. The high density of genotype data and large number of continuous phenotypes gave participants the flexibility to generate a diverse set of biological hypotheses as well as to design studies without a priori assumptions of affection status. In this summary, we discuss the potential use of such data in biological applications and the resulting design and analysis challenges. In particular, we first highlight some of the more salient biological inferences underlying the Genetic Analysis Workshop 15 papers. We then consider study design and statistical methods for making inferences and addressing issues unique to large data sets.

BIOLOGICAL INFERENCE

Evaluation of mRNA abundance has increased our understanding of gene regulatory activities. Genes are usually profiled by their expression levels across multiple hybridization experiments. This expression profile information is frequently considered an ''exposure'' and correlated with phenotypes (e.g., response to treatment, survival time). However, these data can also be considered as a phenotype and compared across germline genotypes. Genome-wide genotype information enables assessment of the association between germline sequence variation and expression levels. Such data can lead to many biological hypotheses and ensuing experiments, including regulatory relations between genes, extent of *cis*-regulation (i.e., regulation from a nearby SNP) versus *trans*-regulation (i.e., regulation from a SNP elsewhere), and common regulatory activities in a pathway or concerning genes related to a human disorder.

GENE NETWORKS

There are a number of fundamental biological questions about the interaction of genes within pathways. Deciphering the topology of gene networks can provide hints as to gene function. Woo et al. [personal communication] studied 393 known biological pathways with less than 30 genes, searching for genome regulators of correlated transcriptional activities within each of these pathways. Among 61 pathways with significant evidence of association (*Bonferroni* corrected $p<0.00001$), they found an interesting association between SNP rs129408 within the *SPSB1* gene (SPRY domain-containing SOCS box protein 1) at 1p36.22 and expression values of 23 proteins involved in the inflammatory response pathway. Their finding confirms results from previous work showing that the *SPSB1* gene encodes a protein that is implicated in immune response regulation [Alexander and Hilton, 2004].

Zhang et al. [personal communication] inferred networks of gene expression phenotypes and separately of SNPs using a measure of information content between all possible pairs of genes and SNPs. The authors identified SNP and expression ''nodes'', which were defined as any SNP or gene transcript, and through a measure derived from information theory (described in more detail later) inferred connections between the nodes. In addition, SNP and expression ''hub nodes'' were defined as nodes with a large number of connections to other SNP and expression nodes. respectively. Based on an assumption that the number of genes connected to any node follows a Poisson distribution, they were able to identify significant hub nodes for a given α level. They identified 115 expression phenotypes and 49 SNPs that served as expression and SNP hub nodes ($p<0.001$), respectively. They further found that 27 SNP hub nodes mapped within five megabases of 25 expression hub nodes, suggesting that their method may be useful for identifying SNP clusters that have *cis*-regulatory functions.

Using an alternative approach, Gao et al. [personal communication] identified SNP and gene ''modules'', which were considered to be sets of SNPs and genes with similar genotype and transcript expression profiles, respectively. The authors applied a hierarchical clustering approach of gene-SNP correlation statistics (*t*-values) to infer 541 gene modules and 255 SNP modules. They found a correlation between genotypes at an intronic SNP rs1355776 in the gene *TMEM108* on chromosome 3 and a gene module consisting of six expression phenotypes (five mapped to 6p21.3) with Gene Ontology [Ashburner et al., 2000] annotation keywords such as ''chromosome organization'' and ''nucleosome assembly'', suggesting that a cluster of genes responsible for chromatin organization may be regulated by rs1355776 or a nearby SNP (or neighboring genes *BFSB2*, *NPHP3*). In addition, they found a SNP module containing eight SNPs, where three SNPs mapped to chromosome 6 and two SNPs mapped to chromosome 4, suggesting that this method can also be applied to find unlinked genetic polymorphisms that non-random cluster, possibly due to selective forces. To infer whether clusters of genes are regulated by clusters of SNPs, the investigators examined regions on a two-dimensional matrix of *t*-values where gene and SNP modules intersected. They found 364 regions which were more correlated than would be expected by chance at a 1% significance level. Because the approaches of Gao et al. and Zhang et al. both seek to cluster SNPs and related phenotypes, future work measuring the degree of overlap of results discovered from both methods will inform on the sensitivity and specificity of these two approaches.

CIS-REGULATION VERSUS TRANS-REGULATION

Hu et al. [2007] aimed to discriminate *cis*-acting SNPs from *trans*-acting SNPs. *Cis*-acting SNPs were identified through linear regression of the expression phenotypes against SNPs that mapped within or near (<1 Mb) genes included in the expression array, with gender used as a covariate; the residuals from these regressions were viewed as variation adjusted for gender and *cis*-acting SNPs. Linkage analysis of these residuals identified candidate *trans*-acting SNPs as those under the peaks and among these *trans*-acting SNPs, those showing association were identified through linear regression. Among 3,462 expression phenotypes, Hu et al. found that 1,514 phenotypes were more strongly influenced by *trans*-regulators, 309 phenotypes were more strongly influenced by *cis*-regulators, and the 1,639 phenotypes did not show any significant association to any SNPs. Heritability of the residuals from the regression models was shown to be lower than heritability of the original expression traits, indicating that part of the heritability of the expression traits is accounted for by the regulatory loci identified in this study. To summarize, by decomposing expression variation into discrete components, Hu et al. were able to quantitatively compare the distribution of *cis*- versus *trans*-regulators in this dataset.

DISEASE PATHWAYS

Two contributions to this GAW15 group investigated the expression of candidate genes involved in common diseases, one focused on a metabolic pathway and another on breast cancer risk. Suktitipat et al. used a principal-components approach [PCA; Ott and Rabinowitz, 1999] to combine expression levels of 10 gene transcripts involved in cholesterol synthesis pathway into a new single composite trait which contained higher heritability information [Suktitipat, personal communication]. Although no SNP associations were significant based on a false discovery rate of 5%, the most significant association was between rs575030 and the composite trait; this SNP maps within 100 kb of lathosterol oxidase (*SC5DL*), a gene encoding an enzyme of the cholesterol biosynthesis pathway. *SC5DL* was previously shown to elevate lathosterol and alter cholesterol synthesis in a sample of human subjects [Brunetti-Pierri et al., 2002]. Zheng et al. [2007] studied 18 transcripts from genes thought to be involved in the etiology of breast cancer. By aggregating linkage signals and association signals, they identified regulatory hotspots for expression of genes involved in breast cancer. Notably, two of the regulatory hotspots were the regions containing *BARD1* and *BRCA1*, loci known to harbor breast cancer genes [Rauch et al., 2005].

STUDY DESIGN

SELECTION OF INFORMATIVE INDIVIDUALS

One design question in genetic association studies is whether to study unrelated individuals only or to include family data. Suh et al. [2007] compared association results between a founders-only analysis (56 individuals) and an analysis that included all family members (194 individuals). A linear mixed model was fit to account for intra-familial correlation. The authors found that using a larger sample such as family data improved one association but decreased evidence for several others. They concluded that pedigree data can be used to filter out some false positives and thus can act as a validation tool for any putative associations detected when analyzing only the founders.

SELECTION OF INFORMATIVE SNPS

Use of tagging SNPs, which capture much of the variation across the genome, can substantially reduce the number of SNPs required for genotyping in association studies with small losses in power [de Bakker et al., 2005]. Goode et al. [2007] compared tagging SNP selection methods implemented in Tagger, ldSelect, and TagSNPs and found that pairwise SNP selection methods often performed better than multi-marker haplotype-based methods. Nonetheless, this report concludes that the most appropriate tagging tool depends on the true genetic model of association. If the discovery set includes the causal SNP, pairwise or multi-marker methods are optimal; if the discovery set defines a haplotype carrying the causal allele, then a haplotype-based tagging approach is optimal. It is critical that one considers the disadvantages, as well as the advantages, of using a particular tagging SNP method.

Alleles associated with common diseases are thought to be more easily detected in association studies, as opposed to linkage studies. Chen et al. [2007] postulated that allele frequencies and effect sizes of associated SNPs inside linkage regions are distributed differently from those of associated SNPs outside linkage regions, because negatively selected traits of large effect, which are rare by definition, are best detected through linkage analysis. The implication is that investigators can select SNPs based on the disease model of interest and SNP properties such as allele frequency. Results were inconclusive due to insufficient power given the available sample size, yet trends in their results were consistent with their hypothesis. Therefore, genome-wide association studies (GWAS) may best detect common, low-penetrant alleles.

NOVEL STATISTICAL APPROACHES

ASSOCIATION TESTING OF MULTI-SNP INTERACTIONS

Two papers [Szymczak et al., 2007; Zheng et al., 2007] applied machine learning approaches to answer the question: which groups of SNPs can best explain phenotypic variation? Although these approaches are potentially computationally intensive, analyses can especially benefit from these nonparametric approaches in the context of ″small n, large p″ situations (i.e., when data is of particularly high dimension relative to the number of observations).

Novel association testing methods were described that used information metrics to select sets of SNPs that best explained phenotypic variance. Szymczak et al. grouped SNPs in a classification setting. In an unsupervised learning phase, classes were defined as clusters of individuals with similar phenotypic profiles. In the following step, SNPs were selected using a novel permutation-based technique. This approach improves classification performance in comparison to classical parametric procedures by identifying a smaller list of markers with higher accuracy.

Zheng et al. applied a greedy screening algorithm to retain SNPs that contribute most to an information score. Using a random set of SNPs, a quantitative genotype trait distortion score was computed from a multi-locus genotype and the sum of the individuals' ranked phenotypes. The greedy phase of the algorithm then searched for a subset of these SNPs that maximizes the quantitative genotype trait distortion. After repeating the entire algorithm for many iterations, SNPs were ranked based on the number of times that they are retained at the greedy phase.

CHOICE OF TEST STATISTIC

At GAW15, a metric that was derived from information theory was commonly used to identify important genes and SNPs. Numerous authors proposed measuring association through entropy reduction as an alternative to traditional statistics such as the χ^2 deviate. Here, Shannon's entropy measure is shown in its basic form as

$$H(X) = -\sum_{x \in X} p(x) \log_2 p(x),$$

where X is a discrete random variable, x is a possible outcome, and $p(x)$ is $\Pr(X = x)$. Entropy reduction served as a basis for an information score, which was proposed in several papers [Szymczak et al. 2007; Zhang, personal communication]. The scoring functions introduced are nonparametric, thus being more robust to violations of normality. Szymczak et al. compared their score, known as the quantitative mutual information score to the analysis of variance and the Kruskal Wallis test (a nonparametric version of the analysis of variance). Substantial discrepancies in significance were observed, which can be explained by the underlying empirical distributions of the data. In particular, non-normality, skewness, and outliers have a different effect on the statistical methods of interest. Zhang et al. derived information content between all possible pairs of genes and SNPs as a function of expression differences and identity-by-descent probabilities provided from linkage analysis.

HANDLING NON-INDEPENDENCE

One issue to be wary of is potential non-independence of data points, either SNPs or phenotypes. Asymptotic p-values corrected for multiple comparisons that assume independent exchangeable observations are sure to be overly conservative. In situations where multiple traits are tested, one may want to account for any correlations. Such designs can improve power by (i.e., reduce the number of statistical tests) or support biological pathways. Suktitipat et al. [personal communication] analyzed a set of cholesterol-related phenotypes as a single phenotype through a generalized estimating equation and PCA and found discordant results between the abilities of both methods to detect linkage and association signals. Woo et al. [personal communication] compared PCA to a generalized estimating equation method called DACE (differential allelic co-expression) and found DACE to have more sensitivity in detecting associations between SNPs and pathways as a whole than PCA, which was biased toward clusters driven by a major gene effect.

CONCLUSIONS

These GAW15 contributions collectively show that reducing the dimensionality of data and incorporating biological information are crucial for an appropriate analysis. However, there is no single analytical approach that will work best in all situations. The efficacy of a diverse set of methods to detect multi-SNP associations or gene-SNP module interactions indicates that in practice one should generally consider more than one analytic technique. Consistency among results can provide an extra degree of confidence to an investigator; nevertheless, discordant results may be equally valuable in that further investigation can reveal violations of key assumptions in the data (i.e., some methods are more sensitive to outliers) or weaknesses in the model. If the latter, sensitivity analyses are warranted to refine the model.

This work leads to several specific lessons for the design and analysis of GWAS. Clearly, genome-wide association testing is more appropriate than linkage to detect common variants leading to low penetrance phenotypes. How one selects SNPs, individuals, and phenotypes for study is critical. Genome-wide SNP panels have initially differed in their approach to coverage of the genome and methods for tagging common variation; therefore, critical assessment of the methods used and applicability to a given dataset is needed. The use of family-based data may not necessarily improve power to detect some loci, even with an increased sample size. By focusing on a particular set of genes relevant to a candidate pathway, one can substantially alleviate the statistical burden of multiple testing. Based on these pathway studies searching for particular expression regulators, interesting results can still be weaned from modest sample sizes.

Finally, as technological developments rapidly drive down the costs of measuring genotypes and phenotypes, the dimensionality of data in many cases will grow at a faster rate than the available sample sizes. In this GAW15 dataset where sample sizes were modest, we learned that analysis of GWAS may benefit from the use of machine learning approaches. It would be of interest to the community to assess the applicability of these lessons in other study settings with relatively abundant genetic data such as the Framingham Heart Study [Dawber et al., 1963] or the Cancer Genetics Markers of Susceptibility initiative (http://cgems.cancer.gov).

ACKNOWLEDGMENTS

We thank the GAW15 participants who contributed to the studies described in this manuscript for their helpful comments. We also thank Vivian Cheung and the GAW15 organizers for providing data for these studies. Group 1 primary contributing authors include Gary K. Chen, Lei Gao, Ellen L. Goode, Pingzhao Hu, Young Ju Suh, Bhoom Suktitipat, Silke Szymczak, Jung Hoon Woo, Wei Zhang, Tian Zheng, and John S. Witte (Group Organizer).

REFERENCES

Alexander WS, Hilton DJ. 2004. The role of suppressors of cytokine signaling (SOCS) proteins in regulation of the immune response. Annu Rev Immunol 22:503–529.

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringswold M, Rubin GM, Sherlock G. 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25:25–29.

Brunetti-Pierri N, Corso G, Rossi M, Ferrari P, Balli F, Rivasi F, Annunziata I, Ballabio A, Russo AD, Andria G, Parenti G. 2002. Lathosterolosis, a novel multiple-malformation/mental retardation syndrome due to deficiency of 3beta-hydroxysteroid-delta5-desaturase. Am J Hum Genet 71:952–958.

Chen GK, Jorgenson E, Witte JS. 2007. A comparison between parameters of association inside and outside linkage regions. BMC Proceedings 1(Suppl 1):S5.

Cheung VG, Spielman RS. 2007. Data for Genetic Analysis Workshop (GAW) 15, problem 1: genetics of gene expression variation in humans. BMC Proceedings 1(Suppl 1):S2.

Cheung VG, Spielman RS, Ewens KG, Weber TM, Morley M, Burdick JT. 2005. Mapping determinants of human gene expression by regional and genome-wide association. Nature 437:1365–1369.

Cordell HJ, de Andrade M, Babron M-C, Bartlett CW, Beyene J, Bickeböller H, Culverhouse R, Cupples A, Daw EW, Dupuis J, Falk CT, Ghosh S, Goddard KA, Goode EL, Hauser ER et al. 2007. Genetic Analysis Workshop 15: gene expression analysis and approaches to detecting multiple functional loci. BMC Proceedings 1(Suppl 1):S1.

Dawber TR, Kannel WB, Lyell LP. 1963. An approach to longitudinal studies in a community: the Framingham Study. Ann N Y Acad Sci 107:539–556.

de Bakker PI, Yelensky R, Pe'er I, Gabriel SB, Daly MJ, Altshuler D. 2005. Efficiency and power in genetic association studies. Nat Genet 37:1217–1223.

Goode EL, Fridley BL, Sun Z, Atkinson EJ, Nord AS, McDonnell SK, Jarvik GP, de Andrade M, Slager SL. 2007. Use of tagging SNPs in association analyses. BMC Proceedings 1 (Suppl 1):S6.

Hu P, Lan H, Xu W, Beyene J, Greenwood C. 2007. Identifying cis- and trans-acting SNPs controlling lymphocyte gene expression in humans. BMC Proceedings 1(Suppl 1):S7.

International HapMap Consortium. 2005. A haplotype map of the human genome. Nature 437:1299–1320.

Ott J, Rabinowitz D. 1999. A principal-components approach based on heritability for combining phenotype information. Hum Hered 49:106–111.

Rauch T, Zhong X, Pfeifer GP, Xu X. 2005. 53BP1 is a positive regulator of the BRCA1 promoter. Cell Cycle 4: 1078–1083.

Suh YJ, Lee HS, Batliwalla F, Li W. 2007. A comparison of founder-only and all-pedigree-members genotype-expression associaiton by regression analysis. BMC Proceedings 1(Suppl 1):S8.

Szymczak S, Nuzzo A, Fuchsberger C, Schwarz DF, Ziegler A, Bellazzi R, Igl BW. 2007. Genetic association studies for gene expressions: Permutation-based mutual information in a comparison with standard ANOVA and as a novel approach for feature selection. BMC Proceedings 1(Suppl 1):S9.

Zheng T, Wang S, Cong L, Ding Y, Ionita-Laza I, Lo SH. 2007. Joint study of genetic regulators for expression traits related to breast cancer. BMC Proceedings 1(Suppl 1):S10.