Published in final edited form as:
Genet Epidemiol. 2009 ; 33(Suppl 1): S68–S73. doi:10.1002/gepi.20475.

Detecting Gene-Environment Interactions in Genome-Wide Association Data

Corinne D. Engelman[1], **James W. Baurley**[2], **Yen-Feng Chiu**[3], **Bonnie R. Joubert**[4], **Juan P. Lewinger**[2], **Matthew J. Maenner**[1], **Cassandra E. Murcray**[2], **Gang Shi**[5], and **W. James Gauderman**[2]

[1]Department of Population Health Sciences, University of Wisconsin School of Medicine and Public Health, Madison, WI

[2]Department of Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles, CA

[3]Division of Biostatistics and Bioinformatics, Institute of Population Health Sciences, National Health Research Institutes, Taiwan, Republic of China

[4]Department of Epidemiology, Gillings School of Global Public Health, University of North Carolina, Chapel Hill, NC

[5]Division of Biostatistics, Washington University School of Medicine, Saint Louis, MO

Abstract

Despite the importance of gene-environment (G×E) interactions in the etiology of common diseases, little work has been done to develop methods for detecting these types of interactions in genome-wide association study data. This was the focus of Genetic Analysis Workshop 16 Group 10 contributions, which introduced a variety of new methods for the detection of G×E interactions in both case-control and family-based data using both cross-sectional and longitudinal study designs. Many of these contributions detected significant G×E interactions. Although these interactions have not yet been confirmed, the results suggest the importance of testing for interactions. Issues of sample size, quantifying the environmental exposure, longitudinal data analysis, family-based analysis, selection of the most powerful analysis method, population stratification, and computational expense with respect to testing G×E interactions are discussed.

Keywords

GAW; case-control; family-based; cross-sectional; longitudinal; rheumatoid arthritis; Framingham Heart Study

INTRODUCTION

Most common diseases are believed to be a result of the combined effect of genes, environmental factors, and their interactions. However, current genome-wide association studies (GWAS) are designed to detect the main effect, that is, the direct association of a single-nucleotide polymorphism (SNP) or cluster of SNPs with disease [Browning and Browning, 2007; Zhao et al., 2006]. Investigators may therefore miss important genetic

Corresponding author: Corinne D. Engelman, Department of Population Health Sciences, University of Wisconsin School of Medicine and Public Health, 610 Walnut Street, 1007A WARF, Madison, WI 53726-2397, Telephone: 608-265-5491 Fax : 608-263-2820, cengelman@wisc.edu.

variants that are specific to subgroups of the population defined by some environmental exposure. In fact, main effect tests will have no power to detect variants that have effects in opposite directions within subgroups (crossing interaction). Despite the potential importance of gene-environment (G×E) interactions in the etiology of common diseases, little work has been done to develop methods for detecting these types of interactions in GWAS data.

The Group 10 contributions to the Genetic Analysis Workshop 16 (GAW16) used both the Framingham Heart Study (FHS) and the North American Rheumatoid Arthritis Consortium (NARAC) data to develop and/or test methods for detecting G×E interactions in GWAS data. The methods used include regression models, latent components (factor) analysis, and machine learning approaches.

METHODS AND RESULTS

Table I shows the study population, research design, and statistical methods used by each contribution to Group 10. Following is a brief description of each contribution, organized according to the GAW16 data set they analyzed.

NARAC: Problem 1

Three groups analyzed the NARAC data. Rheumatoid arthritis (RA) affection status was their primary phenotype. Unless otherwise stated, the groups analyzed all 545,080 SNPs genotyped using the Illumina 550k platform. Following are the specific methods and results for each group.

Arya et al [2009] used a logistic regression approach implemented in PLINK to assess the effects of covariates and genotype×sex interactions on the GWAS analysis of RA after accounting for the effects of population stratification. In the single-SNP analysis, only two non-HLA SNPs, on chromosomes 4 and 20, were significant after Bonferroni correction; neither of these SNPs showed a significant genotype×sex interaction. In the genotype×sex interaction analysis, 30 SNP×sex interactions were significant at the uncorrected $p<1.0\times10^{-4}$ level, although none of these survived the stringent Bonferroni correction.

Chiu et al. [2009] used an extension of a generalized estimating equations approach to conduct linkage disequilibrium (LD) mapping with a sliding window of 100 SNPs in a region containing 1,561 SNPs located between 28,000 and 40,000 cM on 6p21. They also tested for interactions between these SNPs and both shared epitope (SE) alleles and sex. In the single-SNP analysis, 251 out of 1,561 SNPs had a *p*-value $\leq3.2\times10^{-5}$ (the Bonferroni corrected significance level). There was a significant interaction between the SE allele number and the estimated disease locus at 32.6 cM. However, there was no interaction between sex and the estimated disease locus.

Zhuang and Morris [2009] compared a standard test of association between RA and genotype, using an additive genetic model and adjusting for covariates, to a sex-differentiated test [Kraft et al., 2007], and to a test of interaction with sex using a logistic regression framework. They found that signals of association in the major histocompatibility complex (MHC), which are now well established for RA, demonstrated strong effects in both sexes. However, they also identified eight novel SNPs that demonstrated genetic effects in only one sex, or reciprocal effects on risk in males and females. These SNPs were not significant in the standard main effect test of association that ignores possible heterogeneity of effect between sexes.

FHS: Problem 2

Three groups analyzed the FHS data, using several different phenotypes as their outcome. Unless otherwise stated, the groups analyzed all 550,000 SNPs genotyped using the Affymetrix GeneChip® Human Mapping 500k Array Set and the 50k Human Gene Focused Panel. Following are the specific methods and results for each group.

Gu et al. [2009] used a type of latent components (factor) analysis called supervised statistical learning approach for multivariate analysis (SLAM) for longitudinal data and generalized estimating equations to account for familial correlation. They used data from the Offspring Cohort at Exams 1, 3, 5, and 7. The primary phenotype of interest was coronary heart disease (CHD) status, and the data on ten variables including CHD endophenotypes (body mass index, total cholesterol, high-density lipoprotein cholesterol, triglycerides, systolic blood pressure (SBP), diastolic blood pressures, and fasting glucose) and environmental covariates (age at visit, cigarette smoking, and alcohol use) were used for the latent-component analysis. They identified several genes, including two well known CHD candidate genes (*SCNN1B* and *PKP2*) with potential time-dependent G×E interactions, and several others including a novel cardiac-specific kinase gene (*TNNI3K*), with potential G×E interactions independent of time and marginal effects.

Joubert et al. [2009] used a novel variance component method for the estimation of age-dependent genetic effects on longitudinal SBP using 57,837 SNPs on chromosomes 17–22 genotyped in the Offspring Cohort. Three SNPs reached genome-wide statistical significance for association with longitudinal SBP using a Bonferroni corrected *p*-value of 8.6×10^{-7}. Of these three SNPs, one corresponded to the main genetic effects, and the remaining two were for the 2 degrees of freedom (df) test, which simultaneously estimated the main genetic effect and genotype×age interactions for each SNP. There were no significant genotype×age interactions for SBP in these data.

Maenner et al. [2009] used a case-only study design and the random forests (RF) algorithm, a type of machine learning, to identify SNPs that may be involved in gene×smoking interactions related to the age at onset of CHD. They used data from the Original and Offspring Cohorts. After ranking the covariate importance score in each of four runs of RF using 500 trees each, one SNP (rs2011345) ranked as the most important SNP and was within the top ten of all ranked covariates in three of the four runs. Using generalized estimating equations to adjust for sex and account for familial correlation, there was significant evidence of a main effect for both the SNP and smoking status, as well as significant evidence of an interaction between the two.

FHS simulated data: Problem 3

Shi et al. [2009] applied a three-level hierarchical linear mixed-effects model for testing genetic main effects and gene×age interactions affecting coronary artery calcification, while accounting for correlation due to the family-based longitudinal data. Genome-wide association analyses using the 50k chip were conducted based on cross-sectional data (i.e., each of the three single visit data sets separately) and also on the longitudinal data (i.e., using data from all three visits simultaneously). They had prior knowledge of the simulation schema and answers. Results showed that the association tests using longitudinal data were more powerful than those using cross-sectional data. Out of the five simulated major gene SNPs of coronary artery calcification, association with rs17714718 (τ_2) was detected only when using the longitudinal data. SNP rs213952 (τ_5) was found to be significant with both longitudinal and cross-sectional data, but the former yielded a more significant result. None of the other major gene SNPs were found to be significant. No significant gene×age interactions were observed.

SUMMARY AND CONCLUSIONS

The contributions from Group 10 introduced a variety of new methods for the detection of G×E interactions in both case-control and family-based data using both cross-sectional and longitudinal study designs. These contributions also illustrated a number of challenges that arise when considering G×E interactions in a GWAS.

One of the most difficult challenges of GWAS is how to deal with the large *p*, small *n* problem that arises when the number of variables considered (*p*) is much larger than the number of subjects (*n*). The problem becomes even more pronounced when one seeks to test interactions between SNPs and one or more environmental exposures in addition to determining the main effect of each SNP. One approach commonly used to reduce the number of tests performed is to select a subset of SNPs to be tested for interactions with known or hypothesized environmental predictors of the phenotype. This can be done by first conducting a single-SNP analysis for each SNP in the genome-wide data, in which one SNP at a time, along with relevant covariates, is tested for association with the phenotype, and then only the most significant SNPs are followed up in G×E interaction testing. Several Group 10 contributions showed that the use of this strategy may miss potentially important SNPs that have a very small main effect, but a significant G×E effect. For example, Arya et al. [2009] tested genotype×sex interactions for association with RA and found 30 SNP×sex interactions that were nominally significant ($p<1.0\times10^{-4}$), but none of these were significant in the single-SNP analysis. Similarly, Zhuang and Morris [2009] also tested genotype×sex interactions for association with RA and identified eight novel SNPs that demonstrated genetic effects in only one sex, or reciprocal effects on risk in males and females, but these SNPs were not significant in the single-SNP analysis. In the FHS, Maenner et al. [2009] found significant evidence for an interaction between smoking and a SNP selected by a RF algorithm as the most important, but this SNP ranked as only the 2,111th smallest *p*-value in the single-SNP analysis (out of 355,649 SNPs).

The inability of many of the groups to detect a G×E interaction that reached a genome-wide level of significance is likely to be due to inadequate sample sizes. To explore the power to detect an interaction in a GWAS, we adopt a standard logistic model framework for a disease outcome (*D*), with form $\text{logit}[\Pr(D=1|G,E)] = \beta_0 + \beta_g G + \beta_e E + \beta_{ge} G\times E$. This model parameterizes the baseline disease prevalence (β_0), the main effects of *G* (β_g) and *E* (β_e), and the G×E interaction (β_{ge}). The quantity of interest is the interaction $OR_{ge} = \exp(\beta_{ge}) = OR_{G|E=1} / OR_{G|E=0}$, or, in other words, the odds ratio for a given SNP (*G*) in exposed (*E*=1) individuals divided by the odds ratio for *G* in unexposed (*E*=0) individuals. The epidemiologist may want to adopt the alternative exposure-based interpretation for OR_{ge}, specifically $OR_{E|G=1} / OR_{E|G=0}$. Table II shows the required number of case-control pairs required to achieve 80% power for detecting an interaction, for various underlying values of OR_{ge}, minor allele frequencies, and exposure prevalences. The range of exposure prevalences represents that of many common environmental exposures, including physical inactivity, obesity, and smoking. An exposure prevalence of 0.1 is representative of physical inactivity in non-Hispanic whites (10.9% according to Centers for Disease Control (CDC) statistics for 2007) and that of 0.5 is representative of obesity in non-Hispanic black women (53% according to National Health and Nutrition Examination Survey (NHANES) statistics from 2003–2006). The prevalence of physical inactivity in other racial/ethnic groups, obesity in other racial/ethnic/sex groups, and smoking (19.8% according to CDC statistics for 2007) falls between 0.1 and 0.5. The required sample size to detect a significant G×E interaction of reasonable magnitude in a GWAS at $p<10^{-7}$ is approximately two to three times larger than that needed to test a single variant at $p<0.05$ due to the correction for multiple testing.

High as the sample sizes are in Table II, they are underestimates because in reality both G and E are likely measured with error. As discussed below, accurate measurement of environmental exposures is the exception rather than the rule. Genotypes are also subject to measurement errors but these are generally small compared with errors in environmental exposures. However, in GWAS, the vast majority of the common SNPs are not measured directly but rather are captured through tag SNPs. When testing for SNP main effects, it is well known that imperfect tagging inflates the required sample size by a factor of approximately $1/r^2$ [Pritchard and Przeworski, 2001]. Under certain assumptions, this applies more generally to covariates measured with error, specifically the required sample is inflated by the reciprocal of the square of the correlation coefficient between the true value of the covariate and its measurement [Devine and Smith, 1998]. Thus, if r_G^2 is the LD between a tag SNP and the causal SNP, and r_E^2 is the squared correlation coefficient between the true exposure and the measured exposure, we can expect an approximate sample size inflation of $1/r_G^2 r_E^2$. For example, if r_G^2=0.9 and r_E^2=0.8, then the required sample size to detect a G×E interaction with a given power will be 39% higher than the required sample size if both G and E were measured without error.

Unfortunately, given the realities of epidemiological research and the desire to continue to use valuable existing studies (e.g., FHS), the required sample sizes are often not practical. Moreover, recent discussions have pointed out that corrections for multiple testing, such as the Bonferroni correction, are too conservative because they do not take into account correlations between the tests due to LD [Rice et al., 2008]. Rice et al. point out that the effect sizes of susceptibility alleles (and G×E interactions) will rarely reach the required level of significance in GWAS if a Bonferroni correction is used. Although the Bonferroni correction is easy and straightforward to calculate, less conservative methods, such as permutation testing, false-discovery rate, and sequential methods (splitting the data into a test set and replication set), may need to be applied to balance the type I and type II errors (false positives and false negatives, respectively). Alternately, Maenner et al. [2009] initially used a machine learning approach, which is not based on *p*-values so a Bonferroni correction is not applicable. Machine learning approaches can screen large amounts of data and take into account interaction effects as well as main effects without requiring model specification. They then selected a very small number of variables with the highest variable importance scores and tested these for interactions using traditional regression approaches.

Accurately quantifying environmental exposures at the individual level is a challenge that is widely recognized and the topic of the NIH Genes, Environment and Health Initiative. The difficulty of this task, however, may be most evident when contrasted with the detail and volume of information obtained from genetic samples. In all three data sets, measures of environmental variables were comparatively crude, if available at all. Although the FHS data includes a variable for cigarettes smoked per day at each visit, missing data and the lack of information about smoking before and between visits make it difficult to quantify smoking behavior. Given the structure of smoking data, a choice has to be made between creating crude categories of exposure (e.g., ever or never smoked) and basing the ascertainment of exposure on a limited time period (e.g., average number of cigarettes smoked per day across visits, or last known smoking status). More complete measurements of environmental exposures across time would not only better represent the exposure of interest, but would also allow greater flexibility to replicate findings and to compare with other studies in which exposures are computed differently. Several groups selected sex as an “environmental” variable of interest, which could represent a proxy for different environmental exposures [Arya et al., 2009]. While the measurement of sex is certainly more straightforward than smoking or alcohol consumption, it does not necessarily provide more insight into the causal pathways of specific environmental exposures. Because many

health outcomes are thought to be a complex combination of environmental and genetic factors, future studies should strive to create new methods for the collection of meaningful environmental information that is as reliable and comprehensive as the genotype data.

The FHS presented additional challenges beyond that of multiple comparisons and measurement of environmental exposures when testing for G×E interactions. The longitudinal, family-based design resulted in data that were correlated in two ways: repeated measurements taken in the same individual at multiple time points and measurements taken in members of the same family. Shi et al. [2009] addressed both types of correlation by applying a three-level hierarchical linear mixed-effects model to account for correlation due to the family-based longitudinal data. Using the simulated data, they found this model to be generally more powerful than using a cross-sectional model that accounted for familial correlation. Joubert et al. [2009] used a novel variance-component method to account for both repeated measures and familial correlation. Maenner et al. [2009] utilized the longitudinal data by analyzing age at onset of CHD as the outcome. Gu et al. [2009] used a two-level factor analysis for longitudinal data. **Maenner et al.** and **Gu et al.** used a generalized estimating equations model to confirm the results from their primary analysis while accounting for familial correlation. The use of longitudinal data in studies of G×E interactions is particularly appealing because it may help overcome some of the pitfalls discussed above. Specifically, some of the power lost by conducting a G×E analysis using GWA data may be recaptured by the use of longitudinal data and having multiple measurements of the environmental exposure may lessen the problem of measurement error.

As discussed previously, because tests for G×E interaction are generally less powerful than those to detect main effects and current GWAS are typically only powered to detect main effects, it is important for investigators to choose an analysis method that has the most power to detect G×E interactions. Group 10 has investigated the use of many different methods, including traditional logistic regression [Arya et al., 2009; Zhuang and Morris, 2009], latent components analysis [Gu et al., 2009], machine learning algorithms [Maenner et al., 2009], an extended generalized estimating equations approach [Chiu et al., 2009], and hierarchical modeling [Shi et al. 2009]. As discussed previously, many of these analyses identified markers involved in G×E interactions that would have been missed if tested for main effects alone. Zhuang and Morris [2009] and Joubert et al. [2009] applied a two degree of freedom test that has been shown to be a more robust choice to detect markers involved in disease risk because it jointly tests for main effect and interaction [Kraft et al., 2007]. The case-only analysis has been shown to be a powerful alternative to test for G×E [Khoury and Flanders, 1996; Piegorsch et al., 1994]. However, in a genome-wide setting, the assumption of G×E independence in the population is untenable across the large number of markers. Recently, Murcray et al. [2009] developed a two-step method that uses a case-only style screening step on the combined case-control sample to reduce the number of markers tested formally for interaction. They show that their two-step test is more powerful than a traditional logistic regression model for interaction under a wide range of scenarios, even in the presence of an association between gene and environment in the population. Mukherjee and Chatterjee [2008] developed an empirical Bayes-type shrinkage estimator to model G×E interactions with the efficiency of the case-only design and unbiasedness of a case-control design. By combining case-control and case-only analysis, Li and Conti [2009] developed a Bayes model averaging approach to obtain a single estimate of the interaction effect. Through simulation, their Bayes model averaging approach was shown to be more powerful and robust to violations of independence than traditional approaches. Although complex disease is likely to be more complicated than can be defined by simple two-way interaction models, the development of powerful tools that incorporate the joint effects of genes and environment is an important step toward understanding disease outcomes.

Although only one of the Group 10 contributions tested for population stratification [Arya et al., 2009], it should be considered in all studies of G×E interaction. Population stratification can occur when systematic differences in allele frequencies exist between subgroups of a population, often corresponding to distinct genetic ancestry. This is an issue for GWAS because it can result in erroneous associations with the outcome. For analyses including G×E interactions, population stratification is an issue if population membership is associated with the outcome, the genetic effect, and the environmental exposure. One way to determine this is first to test for population stratification, i.e., by employing principal-components analyses using the software EIGENSTRAT [Patterson et al., 2006; Price et al., 2006]. If distinct populations are found, then population membership can be tested for association with the environmental variable as well as with the outcome. A priori criteria for a measure of association should be determined by the investigator. In this scenario, it is assumed that the association with the genetic effect is established through the principal-components analysis. If population membership is also found to be associated with the environmental exposure and the outcome, then the final analyses should adjust for population stratification.

Finally, analyzing G×E interactions can be computationally intensive. For dichotomous traits, testing G×E interaction under a logistic regression framework requires maximizing the likelihood function numerically; for quantitative traits, testing the interaction with family-based longitudinal data using a mixed-effects model also relies on numerical optimization, both of which are computationally much more extensive than contingency table or regular regression approaches. Bayesian methods are, inherently, computationally demanding as well, but may allow consideration of more complex models. When scaling up to genome-wide data with hundreds of thousands of SNPs, care should be taken to choose an appropriate statistical model, analysis software, and necessary computing hardware. As demonstrated by Shi et al. [2009], mixed-effects models with Kronecker and hierarchical structures yielded comparable model fitness. However, the Kronecker analysis required about 5 minutes for a single model fitting, while the hierarchical model required only 3 seconds, both using SAS PROC MIXED. Due to the parallel nature of the genome-wide scan, cluster computing with tens or hundreds of computing units working simultaneously can significantly reduce the overall computation time. SAS Grid computing enables SAS applications to automatically utilize grid resources. PLINK version 1.05 [Purcell et al., 2007] provides an R interface, which allows the use of abundant analytical resources developed under R and also offers an option for cluster computing at no cost.

The analysis of G×E interaction is likely to be of increasing importance as researchers attempt to unravel the etiology of complex diseases using high-volume genetic data. A researcher primarily interested in environmental risk factors may be interested in identifying genes that modify the effect of a target environmental risk factor for a disease, while a researcher primarily interested in genetic risk factors may want to know how an environmental factor affects the penetrance of a gene on a disease. Either situation can be viewed as G×E interaction, and for both, the researcher will be charged with conducting the most efficient analysis possible. At a minimum, this will include consideration of sample size and power, potential population stratification, and the best way to measure the environmental exposure. The Group 10 contributions have provided examples of several approaches one might take in testing G×E interactions, for example, jointly testing for a main and interaction effect, testing for population stratification, and the use of longitudinal data with multiple measurements of the environmental exposure to lessen the problem of measurement error. Although several new issues arise when one analyzes G×E interactions in a GWAS, Group 10 demonstrated that such analysis may hold the promise of uncovering new genes that might not otherwise be detected.

Acknowledgments

The Genetic Analysis Workshops are supported by NIH grant R01 GM031575 from the National Institute of General Medical Sciences. We thank the Group 10 participants for their contributions.

REFERENCES

Arya R, Hare E, del Rincon I, Jenkinson CP, Duggirala R, Almasy L, Escalante A. Effects of covariates and interactions on a genome-wide association analysis of rheumatoid arthritis. BMC Proc. 2009; 3 Suppl 7:S84. [PubMed: 20018080]

Browning BL, Browning SR. Efficient multilocus association testing for whole genome association studies using localized haplotype clustering. Genet Epidemiol. 2007; 31:365–375. [PubMed: 17326099]

Chiu Y-F, Kao H-Y, Chen Y-S, Hsu F-C, Yang H-C. Assessment of gene-covariate interactions by incorporating covariates into association mapping. BMC Proc. 2009; 3 Suppl 7:S85. [PubMed: 20018081]

Devine OJ, Smith JM. Estimating sample size for epidemiologic studies: The impact of ignoring exposure measurement uncertainty. Stat Med. 1998; 17:1375–1389. [PubMed: 9682326]

Gu CC, Yang W, Kraja AT, de las Fuentes L, Dávila-Román VG. Genetic association analysis of coronary heart disease by profiling gene×environment interaction based on latent components in longitudinal endophenotypes. BMC Proc. 2009; 3 Suppl 7:S86. [PubMed: 20018082]

Joubert BR, Diao G, Lin D, North KE, Franceschini N. Longitudinal age-dependent effect on systolic blood pressure. BMC Proc. 2009; 3 Suppl 7:S87. [PubMed: 20018083]

Khoury MJ, Flanders WD. Nontraditional epidemiologic approaches in the analysis of gene-environment interaction: Case-control studies with no controls! Am J Epidemiol. 1996; 144:207–213. [PubMed: 8686689]

Kraft P, Yen YC, Stram DO, Morrison J, Gauderman WJ. Exploiting gene-environment interaction to detect genetic associations. Hum Hered. 2007; 63:111–119. [PubMed: 17283440]

Li D, Conti DV. Detecting gene-environment interactions using a combined case-only and case-control approach. Am J Epidemiol. 2009; 169:497–504. [PubMed: 19074774]

Maenner MJ, Denlinger LC, Langton A, Meyers KJ, Engelman CD, Skinner HG. Detecting gene-by-smoking interactions in a genome-wide association study of early-onset coronary heart disease using random forests. BMC Proc. 2009; 3 Suppl 7:S88. [PubMed: 20018084]

Mukherjee B, Chatterjee N. Exploiting gene-environment independence for analysis of case-control studies: An empirical Bayes-type shrinkage estimator to trade-off between bias and efficiency. Biometrics. 2008; 64:685–694. [PubMed: 18162111]

Murcray CE, Lewinger JP, Gauderman WJ. Gene-environment interaction in genome-wide association studies. Am J Epidemiol. 2009; 169:219–226. [PubMed: 19022827]

Patterson N, Price AL, Reich D. Population structure and eigenanalysis. PLoS Genet. 2006; 2:e190. [PubMed: 17194218]

Piegorsch WW, Weinberg CR, Taylor JA. Non-hierarchical logistic models and case-only designs for assessing susceptibility in population-based case-control studies. Stat Med. 1994; 13:153–162. [PubMed: 8122051]

Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet. 2006; 38:904–909. [PubMed: 16862161]

Pritchard JK, Przeworski M. Linkage disequilibrium in humans: Models and data. Am J Hum Genet. 2001; 69:1–14. [PubMed: 11410837]

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, Sham PC. PLINK: A tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007; 81:559–575. [PubMed: 17701901]

Rice TK, Schork NJ, Rao DC. Methods for handling multiple testing. Adv Genet. 2008; 60:293–308. [PubMed: 18358325]

Shi G, Rice TK, Gu CC, Rao DC. Application of three-level linear mixed-effects model incorporating gene-age interactions for association analysis of longitudinal family data. BMC Proc. 2009; 3 Suppl 7:S89. [PubMed: 20018085]

Zhao J, Jin L, Xiong M. Nonlinear tests for genomewide association studies. Genetics. 2006; 174:1529–1538. [PubMed: 16816420]

Zhuang JJ, Morris AP. Assessment of sex-specific effects in a genome-wide association study of rheumatoid arthritis. BMC Proc. 2009; 3 Suppl 7:S90. [PubMed: 20018087]

TABLE I

Summary of contributions from Group 10

First author	Subset of data used	Genome-wide or candidate SNP	Phenotype	Environmental factor	Statistical methods
NARAC study: Problem 1					
Arya	N/A	Genome-wide	RA affection status	Sex	Logistic regression approach implemented in PLINK
Chiu	N/A	Candidate SNP (1,561 SNPs located between 28,000 and 40,000 cM on 6p21)	RA affection status	Shared epitope alleles and sex	LD mapping through an extension of a generalized estimating equations approach using a sliding window of 100 SNPs
Zhuang	N/A	Genome-wide	RA affection status	Sex	Standard test of association, sex-differentiated test, and test of interaction with sex using logistic regression framework
FHS: Problem 2					
Gu	Offspring Cohort	Genome-wide	CHD status	7 CHD endophenotypes, age, smoking, and alcohol use	Type of latent components (factor) analysis for longitudinal data and generalized estimating equations
Joubert	Offspring Cohort	57,837 SNPs on chromosomes 17–22	SBP	Age	Variance component method for longitudinal data
Maenner	Cases from Original and Offspring Cohorts	Genome-wide	Age at onset of CHD	Smoking	Random forests and generalized estimating equations
FHS simulated data: Problem 3					
Shi	N/A	Genome-wide (50k chip)	Coronary artery calcification	Age	Three-level hierarchical mixed-effects model for family-based longitudinal data

Genet Epidemiol. Author manuscript; available in PMC 2010 August 20.

TABLE II

Required number of case-control pairs to detect a G×E interaction

Interaction effect $OR_{G\times E}$	Exposure prevalence	Minor allele frequency*	
		0.05	0.40
2.0	0.1	6,238 / 12,110	1,364 / 2,748
	0.5	2,515 / 4,946	547 / 1,325
5.0	0.1	1,001 / 1,293	245 / 386
	0.5	459 / 657	113 / 320

*Required *N* assuming a SNP-specific significance level of $\alpha=0.05 / \alpha = 10^{-7}$, the latter corresponding to a GWA scan with Bonferroni adjustment for 500,000 tests.