



Published in final edited form as:

Genet Epidemiol. 2011 ; 35(Suppl 1): S115–S119. doi:10.1002/gepi.20660.

Effect of Linkage Disequilibrium on the Identification of Functional Variants

Alun Thomas¹, Haley J Abel², Yanming Di³, Laura L Faye⁴, Jing Jin⁵, Jin Liu⁶, Zheyang Wu⁷, and Andrew D Paterson⁸

¹Division of Genetic Epidemiology, University of Utah, Salt Lake City, UT

²Division of Statistical Genetics, Washington University, St Louis, MO

³Department of Statistics, Oregon State University, Corvallis, OR

⁴Dalla Lana School of Public Health, University of Toronto, Toronto, Canada

⁵Department of Applied Mathematics and Statistics, State University of New York at Stony Brook, Stony Brook, NY

⁶Department of Statistics and Actuarial Science, University of Iowa, Iowa City, Iowa

⁷Department of Mathematical Sciences, Worcester Polytechnic Institute, Worcester, MA

⁸Hospital for Sick Children, Toronto, Toronto, Canada

Abstract

We summarize the contributions of Group 9 of Genetic Analysis Workshop 17. This group addressed the problems of linkage disequilibrium and other longer range forms of allelic association when evaluating the effects of genotypes on phenotypes. Issues raised by long-range associations, whether a result of selection, stratification, possible technical errors, or chance, were less expected but proved to be important. Most contributors focused on regression methods of various types to illustrate problematic issues or to develop adaptations for dealing with high-density genotype assays. Study design was also considered, as was graphical modeling. Although no method emerged as uniformly successful, most succeeded in reducing false-positive results either by considering clusters of loci within genes or by applying smoothing metrics that required results from adjacent loci to be similar. Two unexpected results that questioned our assumptions of what is required to model linkage disequilibrium were observed. The first was that correlations between loci separated by large genetic distances can greatly inflate single-locus test statistics, and, whether the result of selection, stratification, possible technical errors, or chance, these correlations seem overabundant. The second unexpected result was that applying principal components analysis to genome-wide genotype data can apparently control not only for population structure but also for linkage disequilibrium.

Keywords

score tests; two-stage study designs; robust regression; higher criticism; principal components analysis; graphical modeling

Introduction

The themes that emerged from the contributions of Group 9 from Genetic Analysis Workshop 17 (GAW17) were those encountered in addressing the multiple testing issues inherent in high-density single-nucleotide polymorphism (SNP) data that are assayed through sequencing using gene-based rather than locus-based methods. Focusing on the aggregate effects of variants in a gene on a phenotype should, intuitively, be far more rewarding than considering the data as belonging to anonymous marker loci, because methods that do this can account for and exploit allelic heterogeneity. Although allelic heterogeneity and linkage disequilibrium (LD) give rise to within-gene correlations between observations and tests, long-range correlations are also seen as a result of selection for traits that exhibit locus heterogeneity and in data that are stratified by population, covariates, and, it seems, other unknown effects. In accounting for these correlations, we hope to better detect influential genes and, moreover, the specific effects of variants within genes, or at least an assessment of the causal effects of groups of variants. Some of the complex patterns of allelic association seen by Group 9 contributors were anticipated, but others were more unexpected yet had an impact on results.

All six of the papers from Group 9 in the published proceedings (*BMC Proceedings*, v. 5, suppl. 9) analyzed the mini-exome sequence data for 697 unrelated individuals with their fixed genotype data and simulated case-control status. All used multiple simulations to some degree to assess statistical power and size. Most of the contributions involved some form of regression analysis. Various phenotypes were used, with Q1 being particularly popular. The continuous phenotypes were analyzed both as presented and in derived categorical or dichotomous forms. All the covariates (Age, Sex, Smoking status, and Population of origin) were used by at least two work groups.

Broadly, two of the six contributions focused on highlighting problematic issues that resulted from correlations between noncausal SNPs and multiple causal variants in multiple genes. Di et al. [2011] showed that the p -value in a single-SNP score test of association depended on the sum of correlations with truly causal SNPs, whereas Faye and Bull [2011] considered the effects of LD on the power to detect causal variants and on bias in the estimation of the magnitude of effects when two-stage study designs were used. The other four contributions sought to model and account for allelic associations in a variety of ways. Liu et al. [2011] developed a penalized regression approach to include a term for LD to control for multiple testing with genetic data. He and Wu [2011] developed a form of the higher criticism criterion for multiple tests specifically for use with genetic data. Jin et al. [2011] showed that using principal components analysis to provide stratifying covariates in regression analysis can control the size of statistical tests. In contrast to these regression approaches, Abel and Thomas [2011] used graphical modeling to fit the joint distributions of phenotypes, covariates, and imputed phased genotype states.

There were five additional work groups who contributed to the discussions of Group 9 but whose work is not represented in the proceedings. Armentrout and Carlson [2010] used cross-model test statistics to identify rare variants of large effects, and King and Nicolae [2010] modeled sequence association using lessons learned from population genetics. Lee et al. [2010] used gene set analysis for rare variants. Only two contributions to this group addressed family data. Paterson [2010] used family-based tests for joint detection of linkage and association, and Song et al. [2010] analyzed family data using mixed models with a sliding windows approach.

Methods and Results

Single-Locus Score Tests

Di et al. [2011] showed that, when single-SNP score tests are used to detect association, the probability of declaring association at a locus, whether relevant to the trait or not, is largely determined by the sum over truly causal loci of the effect at those loci and by the correlation with the tested locus. More specifically, Di and colleagues considered a regression of a continuous phenotype on covariates and counts of rare alleles at SNP loci. A covariate was also included to correct for population stratification. The probability of rejecting the null hypothesis was shown to be governed by the noncentrality parameter

$$\lambda = (N - 1) \left(\sum_{j=1}^J r_{\tau,j} h_j \right)^2, \quad (1)$$

where N is the sample size, h_j is a measure of the effect of locus j on the phenotype, and $r_{\tau,j}$ is the correlation between the tested locus τ and the causal locus j .

Although this noncentrality parameter shows that LD can increase the detection probability for loci in or near an influential gene, it also demonstrates why spurious correlations give high false-positive results. A particularly effective demonstration of this result is that a cluster of SNPs on chromosome 12 (C12S704 to C12S709) was declared associated in all simulated replicates despite the fact that no genes on this chromosome contributed to the simulated phenotypes. The summed effect by correlation for C12S706 was 0.21, which was higher than the summed effect for all causal SNPs except C13S522.

Di et al. [2011] demonstrate clearly that single-SNP tests are unreliable for association. The factors contributing to this unreliability include allelic and locus heterogeneity, which means that a large number of loci with which a test SNP can appear correlated are present. Also, rare alleles at the test or causal loci can give positively skewed estimates of correlation in even moderately sized samples.

Two-Stage Study Designs

By extracting SNPs with a high minor allele frequency from the exome sequence data, Faye and Bull [2011] were able to mimic a two-stage study design. In the first stage the extracted SNPs were treated as though they had been sampled using a typical SNP genotyping assay; full sequencing was done in the second stage. In such studies where the same sample is used for both discovery and effect estimation, biases in this estimation are introduced. Faye and Bull described and quantified two competing sources of bias. One was an upward bias resulting from selection of genes in the first stage: Genes whose effect was, by chance, larger in the sample than in the general population had higher probability of passing through the first stage selection. The second form of bias was downward and occurred when the correlations between the tested SNPs and the causal SNPs were weak. The effects of selection bias and attenuation resulting from LD were more complicated when a common tag was in LD with multiple rare causal SNPs.

Faye and Bull [2011] explored these phenomena using regression models for both continuous and discrete traits. At each gene an individual i was allocated an independent variable equal to r_i/n_i , where, for each individual, n_i is the number of SNPs in the gene with a small minor allele frequency and r_i is the number of times that the individual carries the

rarer allele. Faye and Bull considered three scenarios, which were based on the frequency of the minor allele at the tested and causal SNPs.

Faye and Bull found that selection at the tag SNP induced bias in the effect estimate at the causal SNP, even when the correlation between the tag SNP and the causal SNP was not strong. At the extremes of correlation between tag and causal SNPs the biases were larger. In earlier work, Faye et al. [2011] proposed methods to correct for selection bias in genetic effect estimates in single-stage designs. Extensions of the methods that account for selection bias and attenuation resulting from LD are needed for two-stage designs. These biases are important to understand and address, particularly when planning follow-up studies, because of the danger that a replication study will be underpowered.

Regularized Regression

When faced with a regression analysis on a large number of independent variables, regularized regression methods such as the least absolute shrinkage and selection operator (LASSO) [Tibshirani, 1996] have an intuitive appeal as a way to control the number of nonzero effects estimated. Liu et al. [2011] developed such a method by combining a minimax concave penalty [Zhang, 2010] to control the number of significant parameters with a smoothing term to model the similar genetic effects expected at adjacent loci. This approach is based on minimizing with respect to β the penalized least-squares criterion

$$\frac{1}{2} \sum_j \frac{1}{n_j} \sum_i (y_{i,j} - x_{i,j} \beta_j)^2 + \sum_j \rho_1(|\beta_j|; \lambda_1, \gamma) + \frac{\lambda_2}{2} \sum_j \zeta_j (|\beta_j| - |\beta_{j+1}|)^2, \quad (2)$$

where $y_{i,j}$ is the phenotype of the i th subject from a total of n_j subjects with data at locus j with nonmissing genotypes at the j th SNP and $x_{i,j}$ is the genotype for subject i at SNP j . The β are the regression parameters, λ_1 and γ are tuning parameters, and ζ_j is chosen to be the absolute value of the correlation between genotypes at SNPs j and $j + 1$. The function $\rho_1(|\beta_j|; \lambda_1, \gamma)$ is the minimax concave penalty given in detail by Liu et al. [2011]. The first term in this function is the standard sum of squared residuals, and the second term controls the number of parameters. From our point of view, it is the third term that is interesting and novel because it models LD by imposing smoothness on the parameter estimates as we progress along a chromosome.

The optimization of this quantity is complex because it requires dealing with missing genotypes, setting values of the tuning parameters using an extended Bayesian information criterion, and evaluating significance using a leave-one-out method. Nonetheless, this optimization is possible and results in a small number of loci with nonzero effect estimates. Of these, the strongest positive finding seen across the genome is a true-positive cluster on chromosome 13. It is interesting to note that in the initial analysis, the cluster on chromosome 12 found as a false positive by Di et al. [2011] was once again indicated as a hit. However, in the final analysis an adjustment was made for population stratification, which removed this false signal.

Higher Criticism

The higher criticism statistic was introduced by Tukey [1976] as a fair method to simultaneously test multiple hypotheses. If $\rho_{(1)}$ and $\rho_{(L)}$ are ordered p -values from L tests, then the higher criticism statistic is defined as

$$HC = \max_{\{j: \frac{1}{L} \leq p_{(j)} \leq \frac{1}{2}\}} \left\{ \frac{L^{1/2} [(j/L) - p_{(j)}]}{[p_{(j)}(1 - p_{(j)})]^{1/2}} \right\}. \quad (3)$$

The idea is to find the largest standardized difference between the observed and the expected significance under the null distribution.

Among several variants of the higher criticism statistic is innovated higher criticism [Hall and Jin, 2010], which has been shown to provide higher power when tests are related by a distance metric and the correlation between tests decays polynomially with distance. Rather than using the standard joint density, He and Wu [2011] further developed the innovated higher criticism statistic, a clear candidate for LD modeling, to estimate gene effects from marginal densities. This extension addresses the problem of high-dimensional joint densities being sparsely populated with small to moderate data sets.

In an association analysis of Q1, Q2, and disease status over all 200 replicates of the GAW17 data set, He and Wu [2011] compared their new statistic with other forms of higher criticism as well as with ridge regression and a naive minimal p -value approach. Their novel statistics gave, on average, higher ranks to truly causal genes than the other methods did. However, none of the methods consistently ranked the causal genes above other hits; the highest mean rank was 54 for the true-positive gene *FLT1*. He and Wu concluded that, as methods to detect sparse, weak signals, higher criticism approaches deal well with the problem of rare causal alleles, although they are still susceptible to unstable estimates obtained from regression analyses with insufficiently large samples.

Principal Components Analysis

Hidden stratification in samples from structured populations can lead to distortions in the error probabilities in regression analysis. Principal components analysis is a familiar tool that has already been used to find population structure from genotypes assayed at a large number of polymorphisms [Li et al., 2008]. Jin et al. [2011] applied principal components analysis to the GAW17 mini-exome data in order to control type I error rates in regression analyses for the continuous traits Q1 and Q2 and also applied logistic regression on dichotomized variables derived from them.

Using all 200 replicates, Jin and colleagues evaluated a baseline model that regressed the dependent variable on SNP status and the covariates Age, Sex, and Smoking status on three types of SNPs: noncausal SNPs in noncausal genes, noncausal SNPs located in genes where other SNPs were causal, and causal SNPs. They found that type I error rates were inflated for Q1 (11–14%) and slightly inflated for Q2 (about 6%). The error rates were assessed using randomization.

Jin et al. [2011] then fitted two further models, one adding the reported population of origin as a covariate and the other adding the first 10 principal components as assessed from analysis of each replicate. The principal components approach gave a better correction than using reported ethnicity, controlling the type I error to below 6%. There was some cost to this: The type II error rate was increased more by the principal components approach, although it is not clear whether there would be a real loss of power if the critical values for the baseline model and the reported population models were raised to control the type I error to 5%.

Perhaps the most striking feature of Jin and colleagues' results, however, is that the results for noncausal SNPs in LD with causal SNPs matched those for noncausal SNPs in noncausal genes. That is, the principal components correction seems to account not only for population stratification but also for local LD structure.

Graphical Modeling

Graphical models aim to represent the joint distribution of a large set of discrete variables as products of terms from marginal distributions on subsets of variables of low dimension [Lauritzen and Spiegelhalter, 1988]. For discrete variables, these marginal distributions are contingency tables whose dimensions can vary adaptively to represent varying complexities in the conditional independence structure of the variables. Fitting methods have been developed to estimate graphical models for LD from unphased data [Thomas 2005]. Abel and Thomas [2011] made further developments that allowed variables that represent phenotypes and covariates to be included in the model estimation procedure.

Because model estimation requires computationally intensive Markov chain Monte Carlo methods, Abel and Thomas [2011] analyzed only 50 replicates. The results of such an analysis can be shown as a conditional independence graph that connects variables that have direct influence on each other, and Abel and Thomas give three examples of such graphs.

Over all the replicates, Abel and Thomas's method consistently reconstructed the relationships between the traits (disease status, Q1, Q2, Q4) and the covariates (Age, Sex, Smoking status, and Population of origin). Finding associations with SNPs, however, was more problematic, and although one or two true associations were found in each replicate, there were several false positives. It is worth noting, though, that in more than 90% of cases for which SNPs from a gene were found to be correlated with a trait, a true causal SNP was indicated as directly associated. Other SNPs in LD with the causal SNP linked to it were connected to the trait only indirectly through the causal SNP. This indicates that graphical modeling has the power to extract interesting correlations from a background of confounding ones.

Discussion

The conclusion from GAW17 Group 9 is that allelic association must be recognized and accounted for, whether it is a local phenomenon caused by LD or a longer range effect resulting from selection and population stratification. Although the importance of LD was expected in advance of the GAW17 meetings, long-range association, even across chromosomes (as well illustrated by Di et al. [2011]), was less anticipated. Some of these cross-chromosome correlations may be due to technical problems: GAW17 used about 22,000 SNPs, whereas the 1000 Genomes Project analysis of the same sequence data using low sequence coverage called only 12,000 SNPs [1000 Genomes Project Consortium, 2010]. In addition, some sequences might have been mismapped. Sporadic correlations that appear by chance must also be dealt with, and it is perhaps well to remember that, although the problem of multiple testing grows linearly with the increasing number of genetic loci assayed, the number of pairs of loci and hence the opportunities to observe false correlations grow quadratically.

To a substantial extent, many of these LD and scaling issues can be resolved by clustering SNPs in some way. For example, King and Nicolae [2010] pooled observations of rare variants in genes, and Abel and Thomas [2011] modeled correlations between SNPs in the same gene and between SNPs, covariates, and phenotypes simultaneously in a multigene graphical model. Over the multiple replicates, graphical modeling consistently found one or two true positives, although there were also many false ones. Perhaps the most successful

aspect of the graphical modeling approach is its ability to both reconstruct and display in an intuitively appealing way the relationships between phenotypes, covariates, and, when correctly estimated, SNP haplotypes. Faye and Bull [2011] showed that potential biases were still possible for gene-based effect estimation in two-stage studies. Previous work by these investigators has used bootstrapping to address biases in one-stage studies, and this may prove to be a productive line in the two-stage case also.

Rather than using gene-based clustering, Liu et al. [2011] added a smoothing term to a penalized least-squares regression to achieve both parameter reduction and modeling of interlocus correlation. Conceptually similar though quite different technically was the use by He and Wu [2011] of a higher criticism statistic, which models not correlations between alleles but correlations between observed multiple p -values using a function that decays as the physical distance between loci increases. Both of these approaches succeed to some extent in controlling for the number of and correlations between parameters or tests. The top hit using regularized regression was a true positive, but many true effects were missed. Over the 200 replications, higher criticism consistently gave higher average rank to true positives than to false positives, but the highest ranked loci were not consistently the truly causal ones. A common feature of these methods is that correlations between loci are assumed to decay strictly with increasing distance. In contrast, Abel and Thomas [2011] showed that, often, the strongest local correlations are not those between adjacent loci, and Di et al. [2011] highlighted the effect of correlations from a distance. Perhaps further developments to these methods that relax the requirement of monotonic decay would be productive.

Graphical modeling, regularized regression, and higher criticism methods all needed specific developments to meet the requirements of LD modeling. In contrast, Jin et al. [2011] applied principal components analysis in a relatively straightforward manner to provide covariates to account for population structure in regression analyses. That these covariates derived from principal components controlled type I error better than reported ethnicity is striking, and perhaps more surprising is that error rates for noncausal SNPs in genes with causal SNPs were similar to those for SNPs in completely unassociated genes. Although it is clear that principal components analysis should be able to account for population stratification, the mechanism by which it apparently controls for local correlations is an enigma that merits further research.

Acknowledgments

We would like to thank the reviewers who contributed to these proceedings with their constructive comments and evaluations. This work was supported in part by National Institutes of Health (NIH) grant R01 GM081417 awarded to AT. The work of JL was partially supported by NIH grant R01 CA120988 and National Science Foundation grant DMS 0805670. The work of LLF was funded by Canadian Institutes of Health Research (CIHR) grant MOP-84287, CIHR Doctoral Research Award MDR-88001, and a STAGE Training Grant in Genetic Epidemiology and Statistical Genetics.

References

- Abel HJ, Thomas A. Case-control association testing by graphical modeling for Genetic Analysis Workshop 17 mini-exome sequence data. *BMC Proc.* 2011; 5(suppl 9):S62.
- Armentrout SL, Carlson SE. Identifying rare variants of large effect using cross-model test statistics. 2010 Unpublished.
- Di Y, Mi G, Sun L, Dong R, Zhu H, Peng L. Power of association tests in the presence of multiple causal variants. *BMC Proc.* 2011; 5(suppl 9):S63.
- Faye LL, Bull SB. Two-stage study designs combining genome-wide association studies, tag single-nucleotide polymorphisms, and exome sequencing: accuracy of genetic effect estimates. *BMC Proc.* 2011; 5(suppl 9):S64.

- Faye LL, Sun L, Dimitromanolakis A, Bull SB. A flexible genome-wide bootstrap method that accounts for ranking and threshold-selection bias in GWAS interpretation and replication study design. *Stat Med.* 2011; 30:1898–912. [PubMed: 21538984]
- Hall P, Jin J. Innovated higher criticism for detecting sparse heterogeneous mixtures. *Ann Stat.* 2010; 38:1686–732.
- He S, Wu Z. Gene-based higher criticism methods for large-scale exonic single-nucleotide polymorphism data. *BMC Proc.* 2011; 5(suppl 9):S65.
- Jin J, Cerise JE, Kang SJ, Yoon EJ, Yoon S, Mendell NR, Finch SJ. Principal components ancestry adjustment for Genetic Analysis Workshop 17 data. *BMC Proc.* 2011; 5(suppl 9):S66.
- King CR, Nicolae DL. Population-genetics guided modeling of sequence association. 2010 Unpublished.
- Lauritzen SL, Spiegelhalter DJ. Local computations with probabilities on graphical structures and their applications to expert systems. *J R Stat Soc Ser B.* 1988; 50:157–224.
- Lee J, Kim K, Ahn S, Yeon B, Huang J. Gene set analysis for rare variants. 2010 Unpublished.
- Li JZ, Absher DM, Tang H, Southwick AM, Casto AM, Ramachandran S, Cann HM, Barsh GS, Feldman M, Cavalli-Sforza LL, Myers RM. Worldwide human relationships inferred from genome-wide patterns of variation. *Science.* 2008; 319:1100–4. [PubMed: 18292342]
- Liu J, Wang K, Ma S, Huang J. Regularized regression method for genome-wide association studies. *BMC Proc.* 2011; 5(suppl 9):S67.
- 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nat Genet.* 2010; 467:1061–73.
- Paterson A. Family-based tests of linkage/association for the mini-exome data of GAW17. 2010 Unpublished.
- Song J, Galbraith S, Motyer A, Pardy C, Wilson SR. Analysis of the GAW17 family data using mixed models with a sliding window approach. Unpublished. 2010
- Thomas A. Characterizing allelic associations from unphased diploid data by graphical modeling. *Genet Epidemiol.* 2005; 29:23–35. [PubMed: 15838847]
- Tibshirani R. Regression shrinkage and selection via the LASSO. *J R Stat Soc Ser B.* 1996; 58:267–88.
- Tukey, JW. Course notes. Princeton University; 1976. T13 n : The higher criticism.
- Zhang C-H. Nearly unbiased variable selection under minimax concave penalty. *Ann Stat.* 2010; 38:894–942.