Published in final edited form as:

Genet Epidemiol. 2009 ; 33(Suppl 1): S13–S18. doi:10.1002/gepi.20466.

Genome-Wide Association Analyses of Quantitative Traits: The GAW16 Experience

Saurabh Ghosh^{*}

Human Genetics Unit, Indian Statistical Institute, Kolkata, India

Abstract

The group that formed on the theme of genome-wide association of quantitative traits (Group 2) in the Genetic Analysis Workshop 16 comprised eight sets of investigators. Three data sets were available: one on auto-antibodies related to rheumatoid arthritis provided by the North American Rheumatoid Arthritis Consortium; the second on anthropometric, lipid, and biochemical measures provided by the Framingham Heart Study (FHS); and the third a simulated data set modeled after FHS. The different investigators in the group addressed a large set of statistical challenges and applied a wide spectrum of association methods in analyzing quantitative traits at the genome-wide level. While some previously reported genes were validated, some novel chromosomal regions provided significant evidence of association in multiple contributions in the group. In this report, we discuss the different strategies explored by the different investigators with the common goal of improving the power to detect association.

Keywords

linkage disequilibrium mapping; model-free genetic analyses; rheumatoid arthritis; cardiovascular disorder

INTRODUCTION

Most complex genetic disorders are defined by binary clinical end-point (affected/unaffected) traits. Thus, the phenotypic variation between individuals is minimal. For example, in a genetic study on type 2 diabetes, all individuals diagnosed with the disease are indistinguishable in the analyses. However, these clinical end-point traits are typically determined on the basis of values of a set of heritable quantitative characters. For example, increased body mass index (BMI), fasting blood sugar levels, and hemoglobin A_{1C} levels, are precursors of type 2 diabetes. These precursor variables contain more information on inter-individual variability than binary clinical end-points. Thus, it has been argued that analyzing quantitative precursors may be statistically a more powerful strategy for deciphering the architecture of a complex trait [Majumder and Ghosh, 2005].

The current availability of high-throughput single-nucleotide polymorphism (SNP) genotype data and the inability of the candidate gene approach to detect novel genetic associations have resulted in increased interest in performing genome-wide association (GWA) scans. The advantage of GWA studies is that they are unbiased with minimum biological priors on the location of putative genes possibly involved in disease pathogenesis and hence, provide

insights into novel regions of the genome that would probably not have been considered in candidate-gene approaches. While GWA analyses for binary clinical traits are being routinely performed, the most well known being the Wellcome Trust Case-Control Consortium (WTCCC) analyses of seven complex disorders, each based on 2,000 cases and a common set of 3,000 controls [WTCCC, 2007]. Similar analyses for quantitative traits correlated with the clinical end-points are gaining popularity due to the increased information on trait variability as discussed in the preceding paragraph.

GWA analyses pose multiple analytical challenges that are difficult to address using classical statistical tools. These include the issues of handling and managing huge data sets, genetic interpretation of specific statistical tests, determination of sample size and power, and correction for multiple tests and population stratification. While the challenges mentioned above pertain to the analyses of both binary and quantitative traits, there exist additional statistical issues in analyzing quantitative traits, the most crucial being the development of robust distribution-free methodologies that are not overly sensitive to the violation of underlying distributional assumptions on the quantitative trait values while not compromising too much power to detect associations.

The Genetic Analysis Workshop 16 (GAW16) provided an excellent opportunity to explore the various statistical issues pertaining to GWA analyses of quantitative precursors of rheumatoid arthritis (RA) and cardiovascular disorder (CVD). The RA data (Problem 1 of GAW16) is the initial whole-genome association screen performed by the North American Rheumatoid Arthritis Consortium (NARAC) comprising 868 cases and 1,194 controls. Data were available on two quantitative autoantibodies: anti-cyclic citrullinated peptide (anti-CCP) and rheumatoid factor (RF)-IgM. Both the phenotypes are known to have high specificities for RA [Rantapaa-Dahlqvist et al., 2003; Criswell et al., 2007] and the levels are correlated with disease severity [Kastbom et al., 2004; del Val Del Amo et al., 2006]. Genotype data were available for 531,689 SNPs generated on the Illumina Infinum HumanHap550 chip. The CVD data (Problem 2 of GAW16) provided by the Framingham Heart Study (FHS) comprised data on 373 individuals from the Original Cohort, 2,760 from the Offspring Cohort, and 3,997 from the Generation 3 Cohort. There are several quantitative anthropometric, lipid, and biochemical traits that are highly correlated with CVD. These include BMI, systolic blood pressure (SBP), diastolic blood pressure (DBP), fasting total and high-density lipoprotein (HDL) cholesterol levels, fasting triglyceride levels, fasting blood glucose levels, and measures of cigarette and alcohol use. Information was also available on whether individuals were under treatment for hypertension and high cholesterol. Genotype data were provided for 550,000 SNPs generated on the Affymetrix GeneChip Human Mapping 500k Array Set and the 50k Human Gene Focused Panel. The simulated data set (Problem 3 of GAW16) was modeled after the FHS data set provided in GAW16. It included a total of 6,476 individuals with both phenotype and genotype data in 942 pedigrees distributed among three generations and 188 unrelated individuals. The pedigree structures and SNP genotypes were identical to the FHS. The clinical end-point considered in the simulated data was the fictitious condition fatty brain disease (FBD) that develops as a function of circulating lipid levels in the blood, measured in three fractions by triglyceride, HDL, and low-density lipoprotein levels. The degree of subclinical FBD is measured by cranial adipose cumulation. The details of the simulation process are provided in the GAW16 website http://www.gaworkshop.org/readme_problem3.doc and in Kraja et al. [2009].

There were eight contributions in our group, of which four analyzed the NARAC data, three the FHS data, and one the simulated data set mimicking the FHS data. While the quantitative phenotypes analyzed in the NARAC data were common for most contributions, there was heterogeneity in the quantitative phenotypes analyzed in the FHS data. We were able to

integrate the wide spectrum of statistical issues addressed by the different investigators in the group into a few unifying themes as discussed in the subsequent sections.

DATA PROCESSING PRIOR TO ASSOCIATION ANALYSES

Genome-wide SNP data require substantial quality control before they can be used for association analyses. It has been suggested to exclude SNPs not satisfying certain criteria from the association analyses because they may lead to misleading inferences. These criteria include low genotype call rates, low minor allele frequencies, deviation from Hardy-Weinberg equilibrium (HWE), and presence of population stratification. While the different contributors in our group used standard quality control measures to exclude SNPs, there were minor differences in their criteria. For example, the minimum threshold of genotype call rate to include an SNP varied between 90 and 95%. Based on the “Common Variant Common Disease” hypothesis investigators used 0.05 or 0.01 as a threshold of minor allele frequency to exclude SNPs from the analyses. Deviation from HWE, particularly in the control population of a complex disorder, is often an indicator of genotyping errors and can result in false positives. Since testing for HWE at the genome-wide level involves multiple tests, the nominal P -value of 0.05 is not the appropriate threshold to reject the null hypothesis of HWE. Different thresholds ranging between 10^{-7} and 10^{-3} were used to remove SNPs not in HWE. To control for possible population stratification, Lin et al. [2009], Sun et al. [2009], and Xing and Xing [2009] used principal components to capture the variation in ancestry of the different SNPs. Similar to the rationale behind excluding SNPs with low call rates, individuals were removed if genotype data on a certain proportion (5 or 10%) of SNPs were not available. A majority of the contributions used the software PLINK [Purcell et al., 2007] for the data processing and the software EIGENSTRAT [Price et al., 2006] for the principal-components analyses. However, it has been observed that PLINK misses some errors, such as Mendelian incompatibility, and hence may not be optimal as a quality control package for family data.

NOVEL APPROACHES

Association analyses of quantitative traits have traditionally used the analysis of variance (ANOVA) approach to test whether the conditional means of the quantitative trait values given the marker genotypes are equal. While this formulation of genetic association is in the standard statistical framework, ANOVA suffers from some inherent limitations. The classical ANOVA model assumes that the underlying variable being studied is distributed normally. Many human quantitative traits are known to have skewed distributions. Moreover, the method is valid in a strict statistical sense only under the assumption of homoskedasticity of the variable in the different groups. Such an assumption is unrealistic for quantitative precursors (e.g., fasting glucose levels) of complex disorders (e.g., type 2 diabetes) because it is likely that the variance of trait values within the high-risk genotype group will be higher than the other genotype groups [Vimalaewaran et al., 2005]. Simulation studies have shown that ANOVA may result in misleading inferences of association both in terms of an inflated rate of false positives as well as reduced power if these assumptions are not valid [Ghosh and De, 2007]. While analytic transformations on the underlying variables may circumvent these problems to a certain degree, it is still of interest to explore for robust alternatives to ANOVA.

If there is no association between the alleles of a quantitative trait locus (QTL) and an SNP, the allele frequencies at the SNP should be uniform across the different quantile intervals of the quantitative trait.

In the presence of association, the frequency of the marker allele in positive linkage disequilibrium with the QTL allele predisposing to high values of the quantitative trait will increase with the quantile intervals. Ghosh et al. [2009] proposed a novel method based on the correlation between the frequencies of the marker alleles within different quantile intervals and

the quantile values. They analyzed anti-CCP and RF-IgM levels in the NARAC data set and found that while many of the association findings obtained using an asymptotic test based on ANOVA could not be validated by permutation tests, a majority of those obtained using the asymptotic test based on the proposed correlation statistic were validated by permutation tests, thereby indicating that the proposed method is more robust to possible violations in distribution assumptions.

It has been argued that a selective genotyping strategy wherein only individuals with extreme phenotypes are genotyped and used in the analyses may lead to increase in power to detect association [Wallace et al., 2006; Huang and Lin, 2007]. Xing and Xing [2009] explored the efficiency of this strategy in the NARAC data set with respect to the proportion of individuals to be selectively genotyped, the stringency of the level of the association test, and the frequency of the minor allele at an SNP. They observed that the statistical significance of the associations using a selective genotyping strategy is strongly dependent on the level of the test, with the significance decreasing linearly with sample size for very stringent levels of significance. On the other hand, the significance increased with increase in the minor allele frequency irrespective of the level of the test, indicating that the strategy may be optimal in identifying common variants.

Because complex disorders are controlled by multiple loci, a majority of which have minor genetic effects and possibly interact epistatically, association analyses with multiple SNPs are likely to be more powerful than single-marker analyses [Waldron et al., 2006; Ma and Huang, 2007]. While haplotype-based methods have been proposed to address this issue, additional statistical challenges such as inferring of haplotypes, assessing the impact of rare haplotypes in the analyses, and defining the boundaries of haplotype blocks are introduced in these analyses. Thus, there has been considerable interest in developing novel methods that can simultaneously analyze multiple SNPs without considering haplotypes. These statistical methods are primarily model-selection algorithms such as LASSO (least absolute shrinkage and selection operator), SSVS (stochastic search variable selection) and POCRE (penalized orthogonal components regression).

Srivastava and Chen [2009] compared the performances of LASSO and SSVS using a simulated data set of 60 individuals from the Hapmap CEU population and the NARAC data set. The LASSO [Tibshirani, 1996] minimizes the residual sum of squares in the least squares regression framework subject to a constraint on the sum of the absolute values of regression coefficients. On the other hand, SSVS [George and McCulloch, 1993] uses a hierarchical Bayes model based on latent variables indicating whether an explanatory variable (e.g., an SNP in GWA scan) is included in the model. The posteriors on the latent variables are obtained using Gibbs sampling. The simulations showed that the SSVS outperformed both the LASSO and the single-marker analyses based on an area under curve criterion. However, the LASSO performs no better than single-marker analyses when a marker effect is small. For each of the two RA-related quantitative phenotypes, there were a large number of SNPs in the same region that were identified by both the LASSO and SSVS. Lin et al. [2009] used the POCRE approach to simultaneously identify all SNPs associated with anti-CCP based on a single multiple regression. POCRE [Zhang et al., 2008] is a novel regression approach to deal with the “small n (observations), large p (regressors)” problem as encountered in GWA studies and involves the sequential construction of sparsely loaded orthogonal explanatory variables subject to a penalty function and normalization of eigenvectors. The sparse loadings force most of the regression coefficients to be zero, thereby identifying the markers with non-zero loadings as the ones possibly associated with the quantitative phenotype. They obtained significant evidence of association close to some genes, which though not reported in any previous study on RA are known to encode proteins that may have potential roles in the pathogenesis of RA.

OPTIMAL UTILIZATION OF FAMILY-BASED AND POPULATION-BASED DATA

The success of a GWA scan depends strongly on multiple issues pertaining to the study design such as the choice between a population-based case-control study and a family-based transmission-disequilibrium study, determination of an optimal sample size, optimal resource allocation in an initial genome-wide analysis and/or a follow-up confirmation study, an optimal choice of SNPs, and optimal genomic coverage. While population-based case-control analyses are, in general, more powerful than family-based association studies, case-control analyses are susceptible to population stratification and hence have higher rates of false-positive rates. Similarly, while one would ideally like to achieve greater genomic coverage using a higher density of SNPs, such a strategy would require more stringent corrections for multiple testing. In order to assess these issues, Cui et al. [unpublished] used data on HDL in the simulated data set of GAW16 that was modeled after the FHS data. Three types of samples were generated from the provided data set: unrelated individuals, independent sib-pairs, and three generational cohorts. Their results were consistent with intuitive expectations: sample size plays the most vital role in determining the success of the genome-wide scans and the sample size required to detect genes with very minor effects may be unrealistic. In particular, one would require more than 1,000 individuals to have sufficient power to detect association with SNPs that explain 1% of the variance in the quantitative trait. The cohort design yielded more power than the other two sample designs in detecting SNPs associated with HDL. However, ignoring the structure of pedigrees and assuming that related individuals were unrelated led to inflated false-positive rates, but did not compromise the power. While reduced genomic coverage resulted in loss of power, the false-positive rate was almost the same as that based on the maximum coverage available in the FHS simulated data set.

Sun et al. [2009] addressed the issue of possible population stratification in the FHS via a modified version of FamCC, a unified association method based on principal components that incorporates both unrelated and family samples [Zhu et al., 2008]. The FamCC approach computes principal components based on the marker data to capture the population's genetic diversity and then tests for association between the quantitative phenotype and a marker by evaluating the correlation between the residual phenotypes and the residual genotypes after regressing out the effect of population stratification. The FamCC approach was compared with the transmission-disequilibrium test procedure implemented in the software FBAT [Rabinowitz and Laird, 2000] to analyze hypertension, SBP, and DBP in the real FHS data set. While many of the SNPs identified using FamCC were consistent with those obtained using FBAT, the levels of significance of the associated SNPs were different for the two methods. However, this can be explained by the fact that the alternative hypothesis of FBAT is the presence of both linkage and association, while that of FamCC is the presence of only association. Moreover, the power of FBAT depends on the number of informative allele transmissions, while that of FamCC depends on the total number of individuals; hence, FBAT is likely to be less powerful than FamCC. They also found that the modified FamCC method was more robust than the original method proposed by Zhu et al. [2008].

IDENTIFYING MATERNAL EFFECTS

There is now increasing evidence that offspring phenotypes such as insulin resistance [Gonzalez-Ortiz and Martinez-Abundis, 1999] and cognitive function [Julvez et al., 2007] may be influenced by maternal prenatal environment. While these studies are restricted to the influence of measured phenotypes on offspring phenotypes, Kent et al. [2009] addressed a more general issue of the influence of measured genotypes on an offspring's quantitative phenotypes in the FHS data. They developed a mixed variance-components model in the software SOLAR [Almasy and Blangero, 1998] to estimate the maternal random effects on

height, weight, BMI, SBP, DBP, fasting total cholesterol, HDL cholesterol, and triglycerides. Using the estimated maternal effects as a null model, genome-wide measured genotype association analyses were performed on height, which provided the maximum evidence of maternal effects. In the variance-components model, the maternal genotype was included in addition to the offspring genotype and three separate tests of association were carried out for the offspring's genotype, mother's genotype, and the mother's genotype conditional on the offspring genotype. They found that modeling random maternal effects modestly increased the statistical significance of the associations at various loci, particularly in the *TRAPPC9* gene on chromosome 8.

VALIDATING PREVIOUS FINDINGS

One of the major concerns with regard to GWA scans has been the failure to replicate association findings in independent populations. It is essential to validate significant results in multiple populations to increase the probability that association findings are true positives so that they can be followed up through functional studies. Zhao et al. (unpublished) focused on selected genes (*FTO*, *TMEM18*, *MC4R*, *NYD-SP18*, *GNPDA2*, *SH2B1*, *MTCH2*, *KCTD15*, *NEGR1*, and *INSIG2*) that have been recently reported to modulate BMI levels and explored whether association of SNPs near these genes with BMI could be validated using the FHS data set. However, because data on SNPs in the genes *GNPDA2*, *SH2B1*, *MTCH2*, *KCTD15*, and *NEGR1* were not available in the Affymetrix 500k GeneChip, genotypes at the candidate SNPs were imputed for all individuals based on the HapMap CEU sample via the software IMPUTE v0.5 [Marchini et al., 2007]. Standard association analyses were complemented by the genome-wide rapid association analyses using mixed model and regression as implemented in the software GRAMMAR [Aulchenko et al., 2007]. Previously reported SNPs in or near the *FTO*, *TMEM18*, *MC4R*, and *INSIG2* genes were validated in their analyses, but the non-Affymetrix SNPs that were imputed did not exhibit any evidence of association. However, an inflation factor, measured by the ratio of the observed chi-squares and the expected order statistics values [Clayton and Leung, 2007], was relatively high, indicating the possibility of population substructure and other sources of bias or confounding in the FHS data.

DISCUSSION

The common aim of our GAW16 group (Group 2) was to explore various techniques of improving power to identify SNPs that exhibit significant evidence of association with quantitative precursors of RA and heart-related disorders. Multiple studies suggested that one needs not only to look beyond ANOVA for association analyses of population-based quantitative trait data [Ghosh et al., 2009], but also to develop methods incorporating both population and family samples to have increased power in the association studies [Cui et al., personal communication; Sun et al., 2009]. Moreover, it may be a more prudent strategy to analyze multiple SNPs simultaneously [Lin et al., 2009; Srivastava and Chen, 2009] instead of the standard practice of analyzing single SNPs. Kent et al. [2009] showed through a novel concept of modeling maternal genetic effects that there can be an increase in the power to detect associations. Xing and Xing [2009] showed that one could adopt a selective genotyping strategy for detecting common variants without compromising too much power. Zhao et al. (unpublished) emphasized the need to replicate previously reported association findings to enhance the likelihood that these findings are true positives.

It is well known that the major histocompatibility complex-class II-DR beta 1 (*HLA-DRB1*) in the human leukocyte antigen region on 6p21.3 is a major genetic factor for RA [Kerlan-Candon et al., 2001]. The protein tyrosine phosphatase non-receptor type 22 (*PTPN22*) on 1p13.2 is the only other gene with established association with RA [Begovich et al., 2004]. As mentioned in the *Introduction*, levels of anti-CCP and RF-IgM are highly correlated with the RA affection

status. Thus, it was expected that SNPs within both the *HLA-DRB1* and *PTPN22* genes would show significant evidence of association in the genome-wide scans of the two quantitative phenotypes. However, none of the analyses on either anti-CCP or RF-IgM in our group obtained any significant finding in these two genomic regions [Ghosh et al., 2009; Lin et al., 2009; Srivastava and Chen, 2009; Xing and Xing, 2009]. While there is a possibility that the pathway involved in modulating the quantitative traits may not be the same as that of the clinical end-point of RA, a more plausible explanation of this inconsistency is that the genetic effect sizes are small for individual quantitative traits and hence require much larger sample sizes compared with the end-point RA phenotype for detecting associations. Moreover, data on the quantitative traits were available only on RA-affected individuals, leading to reduced trait variation and possible loss of power.

In contrast, many genes previously implicated in obesity such as *FTO* and *MC4R* were validated in the association analyses of BMI levels in the real FHS data set [Zhao et al., personal communication]. It is likely that since obesity is defined by a simple function of BMI levels, the analyses of the binary end-point trait and BMI levels yielded consistent results. However, the significant association findings in the *PRG3* gene for the binary hypertension phenotype (defined in terms of a union of SBP and DBP thresholds) did not validate in the analyses on either SBP or DBP [Sun et al., 2009]. Thus, there continues to be a debate regarding the utility of using quantitative traits as a surrogate for binary clinical end-points in association analyses of complex traits. It is likely that the variation in the clinical end-point trait cannot be captured by a single quantitative trait and hence it may be more prudent to use a vector of quantitative phenotypes, with each component phenotype correlated with the end-point trait. However, since none of the contributors in our group adopted a multivariate phenotype approach in their analyses, the relative efficiency of using multivariate phenotypes over binary end-points could not be assessed. We would like to highlight that the analyses of quantitative traits in both the NARAC and the FHS data sets are not without certain caveats. As mentioned earlier, information on anti-CCP and RF-IgM levels in the RA data were available only on RA cases. This raises the question whether an association analysis based on such data is identifying variants modulating the quantitative traits, per se, or is it mapping some sort of severity measure of RA captured through differential elevations of anti-CCP and RF-IgM levels in RA cases from the mean normal levels. The other major issue is that the effect of medication on association analyses of quantitative traits still remains unclear. While this effect could be estimated and regressed out if data on medication history are available, studies have shown that such an adjustment may lead to misleading inferences [Cui et al., 2003; Tobin et al., 2005]. Kent et al. [2009] and Sun et al. [2009] used an ad hoc correction factor by elevating SBP and DBP values marginally as recommended by Tobin et al. [2005]. It is also important that multiple methods yield similar association findings to increase the probability that the findings are true positives. Ghosh et al. [2009] and Lin et al. [2009] obtained significant association in the 6p22 region with anti-CCP and Zhao et al. (unpublished) validated some previously established genes for obesity. However, because the various contributors analyzing the FHS data studied different quantitative phenotypes, it was not possible to evaluate whether the same genomic regions could be identified by different approaches for the same phenotype. The other issues that were not addressed by any of the contributors in our group included the effects of incorporating environmental covariates and the effects of adjusting for population stratification in the association analyses. However, these issues have been discussed in the summary article of the group formed on the theme of population structure [Hinrichs et al., 2009].

Finally, we note that most current analytical methods for association analyses of quantitative traits use empirical thresholds rather than asymptotic thresholds to determine the significance of the test statistic value. While this increases the robustness of the test with respect to possible violations in distributional assumptions, it clearly requires extensive computational resources

in a GWA study. One way of reducing the computational load is to perform asymptotic tests at the genome-wide level and then confirm the positive association findings empirically using permutations and search algorithms.

Acknowledgments

The authors are grateful to all the members of the group for their contributions to the group discussions during the GAW16 meeting and during the preparation of this article. This work was supported by Fogarty International Center, NIH through R01 grant TW006604 to Saurabh Ghosh. The Genetic Analysis Workshops are supported by NIH grant R01 GM031575 from the National Institute of General Medical Sciences.

Contract grant sponsor: NIH Fogarty International Center; Contract grant number: R01 TW006604; Contract grant sponsor: NIH National Institute of General Medical Sciences; Contract grant number: R01 GM031575.

REFERENCES

- Almasy L, Blangero J. Multipoint quantitative-trait linkage analysis in general pedigrees. *Am J Hum Genet* 1998;62:1198–1211. [PubMed: 9545414]
- Aulchenko YS, de Koning DJ, Haley C. Genomewide rapid association using mixed model and regression: a fast and simple method for genomewide pedigree-based quantitative trait loci association analysis. *Genetics* 2007;177:577–585. [PubMed: 17660554]
- Begovich AB, Carlton VE, Honinberg LA, Schordi SJ, Chokkalingam AP, Alexander HC, Ardlie KG, Huang Q, Smith AM, Spoerke JM, Conn MT, Chang M, Chang SY, Saiki RK, Catanese JJ, Leong DU, Garcia VE, McAllister LB, Jeffery DA, Lee AT, Batliwalla F, Remmers E, Criswell LA, Seldin MF, Kastner DL, Amos CI, Sninsky JJ, Gregersen PK. A missense single-nucleotide polymorphism in a gene encoding a protein tyrosine phosphatase (*PTPN22*) is associated with rheumatoid arthritis. *Am J Hum Genet* 2004;75:330–337. [PubMed: 15208781]
- Clayton D, Leung HT. An R package for analysis of whole-genome association studies. *Hum Hered* 2007;64:45–51. [PubMed: 17483596]
- Criswell LA, Chen WV, Jawaheer D, Lum RF, Wener MH, Gu X, Gregersen PK, Amos CI. Dissecting the heterogeneity of rheumatoid arthritis through linkage analysis of quantitative traits. *Arthritis Rheum* 2007;56:58–68. [PubMed: 17195208]
- Cui JS, Hopper JL, Harrap SB. Antihypertensive treatments obscure familial contributions to blood pressure variation. *Hypertension* 2003;41:207–210. [PubMed: 12574083]
- del Val Del Amo N, Ibanez Bosch R, Fito Manteca C, Gutierrez Polo R, Loza Cortina E. Anti-cyclic citrullinated peptide antibody in rheumatoid arthritis: relation to disease aggressiveness. *Clin Exp Rheumatol* 2006;24:281–286. [PubMed: 16870095]
- George EL, McCulloch RE. Variable selection via Gibbs sampling. *J Am Stat Assoc* 1993;88:881–889.
- Ghosh S, De G. Association analysis of population-based quantitative trait data: an assessment of ANOVA. *Hum Hered* 2007;64:82–88. [PubMed: 17483600]
- Ghosh S, Sanapala KR, Ghosh A, Chakladar S. A quantile-based method for association mapping of quantitative phenotypes: an application to rheumatoid arthritis phenotypes. *BMC Proc* 2009;3:S18. [PubMed: 20018007]
- Gonzalez-Ortiz M, Martinez-Abundis E. Maternal effect of type 2 diabetes mellitus on insulin sensitivity and metabolic profile in healthy young Mexicans. *Diabetes Nutr Metab* 1999;12:32–36. [PubMed: 10517304]
- Hinrichs AL, Larkin EK, Suarez BK. Population stratification and patterns of linkage disequilibrium. *Genet Epidemiol* 2009;33:S88–S92. [PubMed: 19924707]
- Huang BE, Lin DY. Efficient association mapping of quantitative trait loci with selective genotyping. *Am J Hum Genet* 2007;80:567–576. [PubMed: 17273979]
- Julvez J, Ribas-Fitó N, Torrent M, Forns M, Garcia-Esteban R, Sunyer J. Maternal smoking habits and cognitive development of children at age 4 years in a population-based birth cohort. *Int J Epidemiol* 2007;36:825–832. [PubMed: 17550944]

- Kastbom A, Strandberg G, Lindroos A, Skogh T. Anti-CCP antibody test predicts the disease course during 3 years in early rheumatoid arthritis (the Swedish TIRA project). *Ann Rheum Dis* 2004;63:1085–1089. [PubMed: 15308517]
- Kent JW Jr, Peterson CP, Dyer TD, Almasy L, Blangero J. Genome-wide discovery of maternal effect variants. *BMC Proc* 2009;3:S19. [PubMed: 20018008]
- Kerlan-Candon S, Combe B, Vincent R, Clot J, Pinet V, Eliaou JF. HLA-DRB1 gene transcripts in rheumatoid arthritis. *Clin Exp Immunol* 2001;124:142–149. [PubMed: 11359453]
- Kraja AT, Culverhouse R, Daw EW, Wu J, Van Brunt A, Province MA, Borecki IB. The Genetic Analysis Workshop 16 Problem 3: simulation of heritable longitudinal cardiovascular phenotypes based on actual genome-wide single-nucleotide polymorphisms in the Framingham Heart Study. *BMC Proc* 2009;3:S4. [PubMed: 20018031]
- Lin Y, Zhang M, Wang L, Pungpapong V, Fleet JC, Zhang D. Simultaneous genome-wide association studies of anti-cyclic citrullinated peptide in rheumatoid arthritis using penalized orthogonal-components regression. *BMC Proc* 2009;3:S20. [PubMed: 20018010]
- Ma S, Huang J. Combining multiple markers for classification using ROC. *Biometrics* 2007;63:751–757. [PubMed: 17825007]
- Majumder PP, Ghosh S. Mapping quantitative trait loci in humans. Achievements and limitations. *J Clin Invest* 2005;115:1419–1424. [PubMed: 15931376]
- Marchini J, Howie B, Myers S, McVean G, Donnelly P. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat Genet* 2007;39:906–913. [PubMed: 17572673]
- 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]
- Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ, Sham PC. PLINK: a tool set for whole-genome association and population-based linkage analysis. *Am J Hum Genet* 2007;81:559–575. [PubMed: 17701901]
- Rabinowitz D, Laird NM. A unified approach to adjusting association tests for population admixture with arbitrary pedigree structure and arbitrary missing marker information. *Human Hered* 2000;50:227–233.
- Rantapää-Dahlqvist S, de Jong BAW, Berglin W, Hallmans G, Wadell G, Stenlund H, Sundin U, van Venrooij WJ. Antibodies against cyclic citrullinated peptide and IgA rheumatoid factor predict the development of rheumatoid arthritis. *Arthritis Rheum* 2003;48:2741–2749.
- Srivastava S, Chen L. Comparison between stochastic search variable selection and the least absolute shrinkage and selection operator for genome-wide association studies of rheumatoid arthritis. *BMC Proc* 2009;3:S21. [PubMed: 20018011]
- Sun X, Feng T, Song Y, Elston RC, Zhu X. Comparison of a unified analysis approach for family and unrelated samples with the transmission-disequilibrium test to study associations of hypertension in the Framingham Heart Study. *BMC Proc* 2009;3:S22. [PubMed: 20018012]
- The Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* 2007;447:661–678. [PubMed: 17554300]
- Tibshirani R. Regression shrinkage and selection via the lasso. *J R Stat Soc* 1996;58:267–288.
- Tobin MD, Sheehan NA, Scurrah KJ, Burton PR. Adjusting for treatment effects in studies of quantitative traits: antihypertensive therapy and systolic blood pressure. *Stat Med* 2005;24:2911–2935. [PubMed: 16152135]
- Vimalaswaran KS, Radha V, Ghosh S, Majumder PP, Deepa R, Babu HN, Rao MR, Mohan V. Peroxisome proliferator-activated receptor-gamma co-activator-1alpha (PGC-1alpha) gene polymorphisms and their relationship to type 2 diabetes in Asian Indians. *Diabet Med* 2005;22:1516–1521. [PubMed: 16241916]
- Waldron ERB, Whittaker JC, Balding DJ. Fine mapping of disease genes via haplotype clustering. *Genet Epidemiol* 2006;30:170–179. [PubMed: 16385468]
- Wallace C, Chapman JM, Clayton DG. Improved power offered by a score test for linkage disequilibrium mapping of quantitative-trait loci by selective genotyping. *Am J Hum Genet* 2006;78:498–504. [PubMed: 16465623]

- Xing C, Xing G. Power of selective genotyping in genome-wide association studies of quantitative traits. BMC Proc 2009;3:S23. [PubMed: 20018013]
- Zhang D, Lin Y, Zhang M. Penalized orthogonal-components regression for large p small n data. Electron J Stat 2008;3:781–796.
- Zhu X, Li SQ, Cooper RS, Elston RC. A unified association analysis approach for family and unrelated samples correcting for stratification. Am J Hum Genet 2008;82:352–365. [PubMed: 18252216]