Genetic Analysis Workshop 16: Introduction to Workshop Summaries

Jean W MacCluer[1,*], **Christopher I Amos**[2], **Peter K Gregersen**[3], **Nancy Heard-Costa**[5,6], **Monica Lee**[5,6], **Aldi T Kraja**[7], **Ingrid B Borecki**[7], **L Adrienne Cupples**[4,5], and **Laura Almasy**[1]

[1]Department of Genetics, Southwest Foundation for Biomedical Research, San Antonio, TX

[2]Department of Epidemiology, University of Texas M.D. Anderson Cancer Center, Houston, TX

[3]Robert S. Boas Center for Genomics & Human Genetics, Feinstein Institute for Medical Research, Manhasset, NY

[4]Department of Biostatistics, Boston University School of Public Health, Boston, MA

[5]Framingham Heart Study, Framingham, MA

[6]Department of Neurology, Boston University School of Medicine, Boston, MA

[7]Division of Statistical Genomics, Washington University School of Medicine, St. Louis, MO

Abstract

Genetic Analysis Workshop 16 GAW16) was held September 17-20, 2008 in St. Louis, Missouri. The focus of GAW16 was on methods and challenges in analysis of single-nucleotide polymorphism (SNP) data from genome-wide scans. GAW16 attracted 221 participants from 12 countries. The 168 contributions were organized into 17 discussion groups of 6 to 17 papers each. Three data sets were available for analysis. Two of these were data from ongoing studies, generously provided by the investigators. The North American Rheumatoid Arthritis Consortium provided case-control data on rheumatoid arthritis, and the Framingham Heart Study made available information on cardiovascular risk factors for participants in three generations of pedigree data. The third data set included simulated phenotypes for participants in the Framingham Heart Study, using actual pedigree structures and genotypes. This volume includes a paper for each of the 17 discussion groups, summarizing their main findings.

Keywords

single-nucleotide polymorphism; SNP; genome-wide scan; association; linkage; haplotype

INTRODUCTION

The biennial Genetic Analysis Workshops (GAWs) are devoted to evaluation and comparison of statistical methods for mapping, identifying and characterizing the genetic contribution to complex diseases and their precursors and risk factors. For each GAW, topics are chosen that are relevant to current analytical problems in genetic epidemiology. Approximately 6 to 8 months before each GAW, organizers send an e-mail memo to the GAW mailing list, which includes nearly 2,600 investigators worldwide. The memo announces the availability of the data sets, together with a short description of the data and a

*Corresponding author: Department of Genetics Southwest Foundation for Biomedical Research 7620 NW Loop 410 San Antonio, TX 78227-5301 jean@sfbrgenetics.org .

data request form. The form contains a statement to be signed by each investigator acknowledging that the data are confidential and agreeing not to use them for any purpose other than the Genetic Analysis Workshop without written permission from the provider(s). For GAW16, two of the data sets were made available through the National Institutes of Health (NIH) database of Genotypes and Phenotypes (dbGaP). Investigators submit written summaries of their analyses approximately 6 weeks before the Workshop, and these are assembled and distributed to all participants approximately 2 weeks before the Workshop. Workshop organizers assign submissions to discussion groups based on keywords listed by the authors (Table I). Discussion groups interact by e-mail and phone before GAW, and meet on the first day of GAW to discuss their contributions and to prepare for an integrated presentation to the larger Workshop audience. In recent GAWs, a few contributions have been selected for presentation in a Novel Methods session.

Attendance at GAW is limited to investigators who 1) provide data, 2) submit results of their analyses for presentation at the Workshop, or 3) are involved in GAW organization. At the Workshops, each topic is introduced by an invited speaker or a member of the organizing committee, and individuals who have provided data sets are invited to give brief descriptions of the data. The organizers for each discussion group present a comparison of methods and summary of results, and individual contributions may be presented in poster sessions. Several hours are typically devoted to discussion of each topic. These discussions sometimes evolve into continuing collaborations among data providers and participants. In recent years, participant contributions have been peer-reviewed and published by BMC (most recently in *BMC Proceedings*), and summary papers from each discussion group have been published in *Genetic Epidemiology*.

Genetic Analysis Workshop 16 (GAW16) took place September 17-20, 2008 in St. Louis, Missouri, immediately following the meeting of the International Genetic Epidemiology Society. The overall focus of GAW16 was on methods for analysis of large sets of genome-wide data. Several months before GAW16, the GAW mailing list was informed of the availability of three data sets that were to be the focus of GAW16. The first data set (Problem 1) was derived from studies of rheumatoid arthritis (RA), Problem 2 included genotypic and phenotypic data from the Framingham Heart Study, and Problem 3 consisted of simulated phenotypic data utilizing the pedigrees and genotypic data provided to GAW16 by the Framingham Heart Study (FHS).

Investigators used these data to investigate methods for genome-wide association studies (GWAS) of discrete and quantitative phenotypes; joint linkage and association analyses using high density single-nucleotide polymorphism (SNP) data; detecting gene-gene and gene-environment interactions; haplotype-based analyses; controlling false-positive rates in genome-wide screening; detecting and correcting for population stratification; analysis of longitudinal data; joint analysis of multiple phenotypes in GWAS; quality control and error checking; machine learning; multistage approaches to GWAS; family-based association analysis; defining new phenotypes related to rheumatoid arthritis and heart disease; incorporating *a priori* information in GWAS; and gene- or region-based association analysis.

The diverse themes of the 17 discussion groups listed in Table I are a reflection of the range of analytical problems confronting investigators who are attempting take advantage of evolving molecular and statistical methodologies to discover the genetic basis of complex traits. This *Genetic Epidemiology* Supplement presents summaries of the findings from each of the 17 GAW16 discussion groups. Some papers submitted for GAW16 were important components in the group discussions but were not published. In the summary papers these are referred to as "unpublished".

PROBLEM 1: ASSOCIATION ANALYSIS OF RHEUMATOID ARTHRITIS DATA

For Problem 1, data were provided for genome-wide association analysis of RA. SNP genotype data were provided for 868 cases and 1194 controls that had been assayed using an Illumina 550k platform. In addition, phenotypic data were provided from genotyping *DRB1* alleles, which were classified according to the rheumatoid arthritis shared epitope, and levels of anticitrullinated cyclic protein (CCP), and rheumatoid factor IgM. Several questions could be evaluated using the data, including analysis of genetic associations using single SNPs or haplotypes, gene-gene interaction, and genetic analysis of SNPs for qualitative and quantitative factors.

The cases that were made available for analysis by GAW16 participants comprised independent individuals who had met the American College of Rheumatology criteria for rheumatoid arthritis. These cases comprise 445 individual cases who were studied as a part of the North American Rheumatoid Arthritic Consortium (NARAC) because they had at least one additional sibling with rheumatoid arthritis, and an additional 423 cases who were not selected for family history. The cases were recruited from across the United States. Cases are predominantly of Northern European origin. The controls were derived from the New York Cancer Project and were enrolled in the New York metropolitan area [Mitchell et al., 2004]. These controls are somewhat enriched for individuals of Southern European or Ashkenazi Jewish ancestry compared with cases. The GAW16 RA data are part of ongoing studies to identify genetic associations of RA [Plenge et al., 2005]. The data that were provided to GAW16 included results from genotyping 868 cases and 1194 controls after the application of quality control procedures that included removing individuals who had a low overall call rate (<95%) of SNPs, removing first degree relatives, and removing duplicated and contaminated samples. The data that were provided as a part of GAW16 were included in a prior publication [Plenge et al., 2007] that identified the *TRAF1/C5* locus as contributing to susceptibility to rheumatoid arthritis. This earlier publication included additional data that were not provided to the Genetic Analysis Workshop from a study of early onset rheumatoid arthritis conducted in Sweden. Aside from the *TRAF1/C5* locus, there were significant effects from the HLA region and *PTPN22* that can be readily discerned from the data. Data that were provided to GAW16 participants included affection status with RA, sex, *DRB1* alleles detected by serology and further defined using DNA probes for *DRB1*04* and *DRB1*01* alleles, number of shared epitopes carried, the anti-CCP titer, rheumatoid factor IgM level, and 545,080 genotypes derived from Illumina genotyping arrays. All RA cases and 589 controls were genotyped on the HumanHap500 v1, 358 controls were genotyped using the HumanHap500 v3.0, and 247 controls were genotyped using HumanHap300 and HumanHap240 arrays.

A more detailed description of the Problem 1 data is given in Amos et al. [2009].

PROBLEM 2: THE FRAMINGHAM HEART STUDY DATA

GAW16 Problem 2 presented data from the FHS, an observational, prospective study of risk factors for cardiovascular disease begun in 1948. Data have been collected in three generations of family participants in the Study and the data presented for GAW16 included phenotype data from all three generations, with four examinations of data collected repeatedly for the first two generations. The trait data consisted of information on blood pressure, hypertension treatment, lipid levels, diabetes and blood glucose, smoking, alcohol consumed, weight, and coronary heart disease incidence. Additionally, genotype data obtained through a genome-wide scan (FHS SHARe) of 550,000 SNPs from Affymetrix chips were included with the GAW16 data. The genotype data were also used for GAW16

Problem 3, in which simulated phenotypes were generated using the actual FHS genotypes and pedigrees. These data served to provide investigators with a rich resource to study the behavior of genome-wide scans with longitudinally collected family data and to develop and apply new procedures.

The FHS data sets for GAW16 include pedigree, genotype, and phenotype data. The phenotypic data provide information on those participants who have consented to anyone's use, including those at for-profit and not-for-profit institutions. The pedigree file contains all biologically related participants in the FHS and is not limited to the 7230 participants with full consent. A total of 7130 participants have phenotype data: 373 Original Cohort, 2760 Offspring Cohort, and 3997 Generation 3 participants. No phenotypes were included from the 100 fully consented nonoffspring spouses. Of the 7230 consented participants, 6979 are members of pedigrees and 251 are unrelated. Overall, there are a total of 6848 participants who are genotyped, including 6621 in pedigrees and 227 unrelated participants. There are 766 pedigrees with 2 to 301 genotyped participants, including 47 pedigrees with more than 20 genotyped participants.

Data for GAW16 were selected from a subset of examinations. These exams were chosen so that data from FHS participants of approximately the same age from the three cohorts were considered. Data for only one exam were available for Generation 3 participants. Original Cohort participants with data included only the select few who survived ~40 years to have DNA collected and to provide consent for the SHARe project.

Genotype data sets contained approximately 550,000 genotypes for each participant. Genotype data were cleaned for familial relationships, evaluated for consistency with reported familial relationships, and checked for unknown (cryptic) first degree relationships between families. In some cases familial relationships were altered as a result. Cleaning at this stage could result in all genotypes of some individuals being deleted. The genotype data set included legacy DNA samples, which were of poorer quality with a higher rate of missing genotypes. Files with allele intensities and confidence scores for each marker and cel files were also available at dbGaP.

The family structure file, defining the pedigree structures, was provided. There were 8732 participants in this file who have been genotyped. However, only data for those participants who consented to general use (both for-profit and not-for-profit) were available to GAW16.

Three phenotype files were provided: 1) Original Cohort participants, 2) Offspring participants, 3) Generation 3 participants. These files provide information on demographics (sex and age), height, weight, traditional risk factors for coronary heart disease (blood pressure and hypertension, diabetes and blood glucose, smoking, alcohol, and lipid levels), and on incident coronary heart disease and age at onset. Also included are age at onset of diabetes, age at death, and age at last contact.

Further details for the GAW16 Problem 2 data are provided in Cupples et al. [2009b].

PROBLEM 3: SIMULATION OF HERITABLE LONGITUDINAL CARDIOVASCULAR PHENOTYPES BASED ON ACTUAL GENOME-WIDE SNPS IN THE FRAMINGHAM HEART STUDY

GAW16 Problem 3 is composed of simulated phenotypes emulating the lipid domain and its contribution to cardiovascular disease risk. For each replication there were 6476 subjects in families from the FHS, with their actual genotypes for Affymetrix 550k SNPs and simulated phenotypes. Phenotypes were simulated at three visits, 10 years apart, parallel to the

longitudinal data available in the FHS. There were up to six "major" genes influencing variation in high- and low-density lipoprotein cholesterol (HDL, LDL) and triglycerides (TG), and 1000 "polygenes" simulated for each trait. Some polygenes had pleiotropic effects. The locus-specific heritabilities of the major genes ranged from 0.1-1.0%, under additive, dominant, or overdominant modes of inheritance. The locus-specific effects of the polygenes ranged from 0.002-0.15%, with effect sizes selected from negative exponential distributions. All polygenes acted independently and had additive effects. A group of 39 polygenes influencing HDL were clustered within 0.5 Mb on chromosome 11; otherwise, the polygenes for each trait were randomly distributed throughout the genome. At each simulated visit, the value for LDL of each subject was checked, and individuals in the upper tail were designated as medicated. The proportion of subjects that were medicated increased across visits at 2%, 5% and 15%. Carotid arterial calcification (CAC) was simulated as a phenotype that takes many years to develop. Whether a subject smoked during the period before a visit influenced the risk of a myocardial infarction (MI). At first visit, men had a 27% chance to be smokers and women had a 23% chance. Each smoker had an 8% chance to permanently quit smoking before each of the subsequent visits. The resulting smoking rates are commensurate with rates reported by the Centers for Disease Control. The risk of having an MI event before each visit was determined primarily by CAC, but also by smoking and two independent genetic loci interacting with CAC.

The FHS pedigrees, distributed as GAW16 Problem 2 [Cupples et al., 2009b], formed the basis of the simulation distributed as Problem 3 [Kraja et al., 2009]. In total, there were 6476 subjects who had genotypes and simulated phenotypes. After the simulations began, additional FHS subjects provided broad consent for data sharing; because of time limitations, these additional subjects were not included in the simulations. To ensure comparable data to that which was simulated, a file that defined precisely which subjects were included and their relationships within families was provided. The ~550k measured SNP genotypes, distributed for GAW16 Problem 2 from both the genome-wide scan and the additional candidate gene platform (GeneChip® Human Mapping 500k Array Set (Nsp and Sty) and the 50k Human Gene Focused Panel), comprised the genotypes for GAW16 Problem 3. As mentioned, novel fictitious phenotypes were simulated for subjects. The overall effect of each trait-specific polygenic component was scaled to achieve the target total trait heritabilities of 60%, 55% and 40% for HDL, LDL, and TG, respectively. The remaining variance was uncorrelated among family members, with the exception of a simulated dietary effect on TG levels that accounted for a correlation of 0.05 among family members, regardless of their coefficient of relationship. The phenotypes generated from this genetic model were scaled to the empirically derived means and variances for the actual HDL, LDL, and TG traits within 13 age strata (in 5-year intervals) and sex.

Although family members of the FHS attended various exams at different times, depending on the generation, we modeled our study as if all subjects were recruited at one time. The simulation model included up to six "major" genes for the lipid phenotypes HDL, LDL, and TG, and 1000 polygenes for each trait. Several polygenes had pleiotropic effects. The locus-specific heritabilities of the major genes ranged from 0.1-1.0%, under additive, dominant, or overdominant modes of inheritance, with minor allele frequencies at least 5%, with one exception, where the minor allele frequency was 1%. The locus-specific effects of the polygenes were on average an order of magnitude smaller, ranging from 0.002 to 0.15%, with effect sizes extracted from negative exponential distributions. All polygenes acted independently and had additive effects. HDL, TG and LDL shared 40% of their polygenes in common, and HDL and TG shared an additional 20%. A group of 39 polygenes influencing HDL were clustered within 0.5 Mb on chromosome 11; otherwise, the polygenes for each trait were randomly distributed throughout the genome. The overall effect of each trait-specific polygenic component was scaled to achieve the target total trait heritabilities of

60%, 55%, and 40% for HDL, LDL, and TG, respectively. The remaining variance was uncorrelated among family members, with the exception of a simulated dietary effect on TG levels that accounted for a correlation of 0.05 among family members, regardless of their coefficient of relationship. The phenotypes generated from this genetic model were scaled to the empirically derived means and variances for the actual HDL, LDL, and TG traits within 13 age strata (in 5-year intervals) and sex.

Two hundred realizations of the generating model were simulated. The simulated data are archived in the dbGAP of the National Center for Biotechnology Information under the name "GAW16 Framingham and Simulated Data." A more detailed description of the Problem 3 data is given in Kraja et al. [2009].

THE GENETIC ANALYSIS WORKSHOPS: HISTORY

GAW16 marks the transition of leadership of the GAWs from Jean MacCluer, who has retired, to Laura Almasy. The GAWs began in 1982, when GAW1 was held at the American Society of Human Genetics meeting in Detroit. The idea for the Genetic Analysis Workshops began with a suggestion by Newton Morton at the 1981 American Society of Human Genetics meeting in Dallas. There had been controversy concerning the best approach to detecting the contribution of major genes to quantitative traits. The method of choice at the time was complex segregation analysis, and there were at least three computer packages that often obtained disparate results in analyses of data for complex traits. A contest was proposed, in which computer-simulated data for determination of a complex trait would be generated, for which the mode of inheritance of the trait in question was known. The impetus was initially to determine the numerical accuracy of the algorithms, to examine the robustness of the methodologies to violations of assumptions, and finally, to compare the range of conclusions that could be drawn from a single set of data. The data would be distributed to all interested investigators, and the task would be to determine the genetic contribution to the trait. Four simulated data sets were distributed, and seven groups of investigators participated in GAW1. There was lively discussion, and it became apparent that the skill of the analyst was at least as important an element in success as was the choice of computer program [MacCluer et al., 1983]. The Workshops have evolved to include consideration of problems related to analyses of specific diseases, but the focus has always been on analytical methods.

Previous Genetic Analysis Workshops have been devoted to evaluation of methods of segregation analysis of quantitative traits; methods for detecting and interpreting disease-marker associations and linkage; methods for multipoint mapping; resolution of differences in genetic maps obtained by different investigators; new approaches to analysis of data sets for specific diseases; issues of genetic heterogeneity and genotype-environment interaction; linkage analysis using information from affected relative pairs; resolution of physical and genetic maps; recent progress in statistical genetic methods; evaluation of methods for detecting genetic effects on quantitative risk factors for complex diseases; methods for detecting the genes that contribute to complex oligogenic diseases using genomic scan data; methods for genetic analysis of longitudinal data; and methods for analysis of genome-wide data.

The Workshops have utilized both computer-simulated and real data. The real data sets have included family data for insulin-dependent diabetes mellitus, celiac disease, multiple sclerosis, Huntington's disease, breast cancer, affective disorders, melanoma, Alzheimer's disease, coronary heart disease, alcoholism, asthma, and rheumatoid arthritis. For GAW15, microarray RNA expression data from CEPH families were distributed. Workshops focusing on mapping have utilized chromosome 11 genetic markers, chromosome 21 genetic and

physical data, chromosome 17q data from the Breast Cancer Linkage Consortium, chromosomes 5 and 18 data from several bipolar data sets, and genomic scan data from the Collaborative Study on the Genetics of Alcoholism.

Simulated data sets have been used to address problems of power and false positives, as well as other issues in which investigators wish to know the true mode of inheritance of the trait in order to evaluate the new analytical methods that they are developing. GAW1 [MacCluer et al., 1983], GAW2 [MacCluer et al., 1984], and GAW3 [MacCluer et al., 1985] used only simulated data because the primary concern in those years was whether some genetic analysis programs were better than others at determining the true genetic contribution to complex traits. For GAW4 (1985) through GAW8 (1992), the emphasis was mostly (but not exclusively) on real data because participants were eager to apply their methods to complex disorders for which they might not otherwise have ready access to high-quality data sets. Since then, participants have consistently indicated their strong interest in having access to both real and simulated data because the two types of data each have distinct advantages and serve different purposes.

The Workshops have provided an opportunity, rare outside a workshop setting, for participants to interact in addressing methodological issues, to test novel methods on the same well characterized data sets, to compare results and interpretations, and to discuss current problems in genetic analysis. The Workshop discussions are valuable for investigators who are evolving new methods of analysis as well as for those who wish to gain experience with existing methods. Over the years, the people who have contributed data have been the most crucial element in the success of the Genetic Analysis Workshops, and they report that they have thoroughly enjoyed the experience.

The GAWs are felt to be an excellent learning experience, both for young researchers and for those established in the field. Participants appreciate the opportunity to explore the behavior of new methods in a setting in which the advantages and disadvantages of each method can be discussed in depth. They benefit by observing the many ways in which different investigators tackle the same problem. Many young investigators (graduate students and postdoctoral fellows) have participated and are thereby being drawn into genetic epidemiology, a field in desperate need of new talent. For recent GAWs, scholarships have been awarded to pre- and post-doctoral students to help defray their travel expenses. Funding for these scholarships has been obtained from a variety of sources, both governmental and private.

The success of the Workshops is attributed at least in part to the informality of sessions, and the requirement that everyone who attends must have made a contribution. We constantly look for ways to improve Workshop format and to maintain the informality and the "roll up your sleeves" atmosphere that has been so important. The Genetic Analysis Workshops have provided investigators with the impetus to learn programs and methods that they otherwise would not have used, and data assembled for the Workshops often have stimulated further analyses.

The main purpose of GAW has always been to evaluate existing analytical methods and to provide an impetus for developing new methods. As indicated by the responses to our surveys, most GAW data sets have continued to be heavily used for this purpose for years after each Workshop. Investigators use GAW data to test new methods, and in grant applications, to estimate power and false-positive rates or to demonstrate the feasibility of statistical techniques for finding disease genes. GAW data also have been utilized extensively for dissertation research and in teaching. The data have been used for short courses and workshops, for independent study, and in formal classes.

CONCLUSIONS

This volume includes a summary paper for each of the 17 GAW16 discussion groups. These summaries provide an overview of the wide variety of topics addressed and conclusions drawn in the 168 papers presented at GAW16, 131 of which are published in *BMC Proceedings* [Cupples et al., 2009a]. Building on their long history, the GAWs continue to be a venue for development and testing of new statistical genetic methods, a forum for interaction among investigators and development of new collaborations, and a source of data sets for teaching and for methods development. Planning is currently underway for GAW17, which will be held October 12-15, 2010, in Boston, Massachusetts. For more information about the Genetic Analysis Workshops, visit http://www.gaworkshop.org/.

Acknowledgments

Tom Dyer, who worked on preparing the data for distribution, with the assistance of Richard Polich, Gene Hopstetter, and Juan Peralta, and who helped with GAW wiki with the assistance of Gerry Vest and Kent Polk. As for past GAWs, April Hopstetter, Director of Technical Publications and Printing at the Southwest Foundation for Biomedical Research, assisted with editing of the GAW16 proceedings, with the help of Maria Messenger and Malinda Mann. Rene Sandoval and Rudy Sandoval were responsible for putting together the final pre-GAW book.

Local arrangements for GAW16 required many hours of planning and organization. We are grateful to local organizers Michael Province, Ingrid Borecki, and Jeanne Cashman as well as volunteers Linus An, Mark Yong-Moon Park, Jevon Plunkett, Amy Sleeter, Kristy Smith, Jim Valentine, and Lorna Walters for welcoming us to St. Louis and for their efforts to assure a successful GAW.

Long-term planning and organization of the Genetic Analysis Workshops is the responsibility of the GAW Advisory Committee, which has a rotating membership. Its membership at the time of GAW16 included Laura Almasy (chairman), Joan Bailey-Wilson, Heike Bickeböller, Ingrid Borecki, Heather Cordell, Elizabeth Hauser, Jean MacCluer, Maria Martinez, John Witte, Xiaofeng Zhu, and Andreas Ziegler. We are grateful that in addition to serving on this committee, many of these individuals took on other tasks. You will see many of their names above in the lists of data providers and group leaders. Joan Bailey-Wilson and Heather Cordell also served as moderators for the general discussion session at the close of GAW16, and Xiaofeng Zhu and John Witte took on the responsibility of selecting papers for the Novel Methods session presented at the Workshop.

As always, we wish to express our appreciation to the GAW participants, without whose ongoing, enthusiastic support the GAWs could not have enjoyed their continuing success.

REFERENCES

Amos CI, Chen WV, Seldin MF, Remmers E, Taylor KE, Criswell LA, Lee AT, Plenge RM, Kastner DL, Gregersen PK. Data for Genetic Analysis Workshop 16 Problem 1, association analysis of rheumatoid arthritis data. BMC Proc. 2009; 3(Suppl 7):S2. [PubMed: 20018009]

Cupples LA, Beyene J, Bickeböller H, Daw EW, Fallin D, Gauderman J, Ghosh S, Goode E, Hauser E, Hinrichs A, Kent J Jr. Martin L, Martinez M, Neuman R, Province M, Szymczak S, Wilcox M, Ziegler A, MacCluer JW, Almasy L. Genetic Analysis Workshop 16. BMC Proc. 2009a; 3(Suppl 7):S1–135. [PubMed: 20017962]

Cupples LA, Heard-Costa N, Lee M, Atwood LD, for the Framingham Heart Study Investigators. Genetic Analysis Workshop 16 Problem 2: The Framingham Heart Study data. BMC Proc. 2009b; 3(Suppl 7):S3. [PubMed: 20018020]

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(Suppl 7):S4. [PubMed: 20018031]

MacCluer JW, Wagener DK, Spielman RS. Genetic Analysis Workshop I: Segregation analysis of simulated data. Am J Hum Genet. 1983; 35:784–92.

MacCluer JW, Falk CT, Spielman RS, Wagener DK. Genetic Analysis Workshop II: Summary. Genet Epidemiol. 1984; 1:147–59. [PubMed: 14964293]

MacCluer JW, Falk CT, Wagener DK. Genetic Analysis Workshop III: Summary. Genet Epidemiol. 1985; 2:185–98.

Mitchell MK, Gregersen PK, Johnson S, Parsons R, Vlahov D. The New York Cancer Project: Rationale, organization, design, and baseline characteristics. J Urban Health. 2004; 81:301–10. [PubMed: 15136663]

Plenge RM, Padyukov L, Remmers EF, Purcell S, Lee AT, Karlson EW, Wolfe F, Kastner DL, Alfredsson L, Altshuler D, Gregersen PK, Klareskog L, Rioux JD. Replication of putative candidate-gene associations with rheumatoid arthritis in >4,000 samples from North America and

Sweden: Association of susceptibility with PTPN22, CTLA4, and PADI4. Am J Hum Genet. 2005; 77:1044–60. [PubMed: 16380915]

Plenge RM, Seielstad M, Padyukov L, Lee AT, Remmers EF, Ding B, Liew A, Khalili H, Chandrasekaran A, Davies LR, Li W, Tan AK, Bonnard C, Ong RT, Thalamuthu A, Pettersson S, Liu C, Tian C, Chen WV, Carulli JP, Beckman EM, Altshuler D, Alfredsson L, Criswell LA, Amos CI, Seldin MF, Kastner DL, Klareskog L, Gregersen PK. TRAF1-C5 as a risk locus for rheumatoid arthritis--a genomewide study. N Engl J Med. 2007; 357:1199–209. [PubMed: 17804836]

Table I

Genetic Analysis Workshop 16: Summary of Group Topics

Group	Topics	Problems[a]			Presented at GAW16	Published	Organizers[b]
		1	**2**	**3**			
1	Genome-wide association studies for discrete traits	X			16	13	Duncan Thomas/Ellen Goode
2	Genome-wide association analysis of quantitative traits	X	X	1	9	6	Saurabh Ghosh
3	Multistage analysis strategies for GWAS	X		1	10	6	Rosalind Neuman
4	Haplotype-based analysis	X			11	7	Elizabeth Hauser
5	Improving the signal-to-noise ratio in genome-wide association studies. GAW16		X	1	7	5	Lisa Martin
6	Analysis of multiple phenotypes	X	X	1	8	6	Jack Kent
7	Phenotype definition and development	X	X	X	9	6	Marsha Wilcox
8a	Quality control in genome-wide association studies and measures of clinical validity	X	X		6	6	Andreas Ziegler
8b	Machine learning in genome-wide association studies	X	1	X	13	11	Silke Szymczak
9	The challenge of detecting epistasis (G×G interactions)	X	X	X	17	13	Michael Province/Ping An
10	Detecting gene-environment interactions in genome-wide association data	X	X	1	9	7	Corinne Engelman/James Gauderman
11	Inclusion of a priori information in genome-wide association analysis	X	X	1	8	6	Nathan Tintle/Heike Bickeböller
12	Combining information from linkage and association methods	1	X	X	8	7	Ellen Wijsman/Warwick Daw/Elizabeth Marchani
13	Population stratification and patterns of linkage disequilibrium	X	X		10	8	Anthony Hinrichs/Brian Suarez
14	Use of longitudinal data in genetic studies in the genome-wide association studies era	X	X	1	8	8	Berit Kerner/Daniele Fallin
15	Family-based samples are useful in identifying common polymorphisms associated with complex traits	X	X	X	8	6	Stacey Knight/Maria Martinez
16	Gene- or region-based analysis of genome-wide association studies	X	X		11	10	Joseph Beyene

[a] X indicates problems worked on by each group; 1 indicates that only one member of the group worked on that problem.

[b] Organizers and first authors of summary papers, if different.