BMC Proceedings

BioMed Central

Proceedings

Open Access

Genetic Analysis Workshop 16: Strategies for genome-wide association study analyses

L Adrienne Cupples*[1], Joseph Beyene[2], Heike Bickeböller[3], E Warwick Daw[4], M Daniele Fallin[5], W James Gauderman[6], Saurabh Ghosh[7], Ellen L Goode[8], Elizabeth R Hauser[9], Anthony Hinrichs[4], Jack W Kent Jr[10], Lisa J Martin[11], Maria Martinez[12], Rosalind J Neuman[4], Michael Province[4], Silke Szymczak[13], Marsha A Wilcox[14], Andreas Ziegler[13], Jean W MacCluer[10] and Laura Almasy[10]

Addresses: [1]Department of Biostatistics, Boston University School of Public Health, 801 Massachusetts Avenue, Boston, MA 02130 and Framingham Heart Study, Framingham, Massachusetts, USA, [2]Research Institute of the Hospital for Sick Children and University of Toronto, 555 University Avenue, Toronto, Ontario M5G 1X8, Canada, [3]Department of Genetic Epidemiology, University Medical Center Göttingen, Humboldtallee 32, 37073 Göttingen, Germany, [4]Division of Statistical Genomics, Washington University School of Medicine, 4444 Forest Park Boulevard, Campus Box 8506, St. Louis, Missouri 63108, USA, [5]Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, 615 North Wolfe Street, Baltimore, Maryland 21205, USA, [6]University of Southern California, Department of Preventive Medicine, Division of Biostatistics, 1540 Alcazar Street, CHP-220, Los Angeles, California 90033, USA, [7]Human Genetics Unit, Indian Statistical Institute, Kolkata 700018, India, [8]Department of Health Sciences Research, Mayo Clinic, 200 First Street Southwest, Rochester, Minnesota 55905, USA, [9]Duke University, Durham, North Carolina 27710 USA, [10]Department of Genetics, Southwest Foundation for Biomedical Research, P.O. Box 760549, San Antonio, Texas 78245, USA, [11]Division of Biostatistics and Epidemiology, Cincinnati Children's Hospital Medical Center, 3333 Burnet Avenue, Mail Code 5041, Cincinnati, Ohio 45229, USA, [12]INSERM, U.563, University Paul-Sabatier, CPTP, Toulouse F-31300, France, [13]Institut für Medizinische Biometrie und Statistik, Universität zu Lübeck, Universitätsklinikum Schleswig-Holstein, Campus Lübeck, Maria-Goeppert-Strasse 1, 23562 Lübeck, Germany and [14]Johnson & Johnson Pharmaceutical Research and Development, 1125 Trenton-Harbourton Road, Titusville, New Jersey 08560, USA

E-mail: L Adrienne Cupples* - adrienne@bu.edu; Joseph Beyene - joseph@utstat.toronto.edu; Heike Bickeböller - hbickeb@gwdg.de; E Warwick Daw - warwick@wustl.edu; M Daniele Fallin - dfallin@jhsph.edu; W James Gauderman - jimg@usc.edu; Saurabh Ghosh - saurabh@isical.ac.in; Ellen L Goode - egoode@mayo.edu; Elizabeth R Hauser - Elizabeth.Hauser@duke.edu; Anthony Hinrichs - THinrichs@wustl.edu; Jack W Kent Jr - jkent@sfbrgenetics.org; Lisa J Martin - Lisa.Martin@chmcc.org; Maria Martinez - maria.martinez@toulouse.inserm.fr; Rosalind J Neuman - rneuman@wustl.edu; Michael Province - mprovince@wustl.edu; Silke Szymczak - silke.szymczak@imbs.uni-luebeck.de; Marsha A Wilcox - mwilcox@its.jnj.com; Andreas Ziegler - ziegler@imbs.uni-luebeck.de; Jean W MacCluer - jean@sfbrgenetics.org; Laura Almasy - almasy@sfbrgenetics.org
*Corresponding author

from Genetic Analysis Workshop 16
St Louis, MO, USA 17-20 September 2009

Published: 15 December 2009

BMC Proceedings 2009, **3**(Suppl 7):S1 doi: 10.1186/1753-6561-3-S7-S1

This article is available from: http://www.biomedcentral.com/1753-6561/3/S7/S1

Introduction

This supplement to *BMC Proceedings* contains the proceedings of Genetic Analysis Workshop (GAW) 16, which was held September 17-20, 2008, in St. Louis, Missouri, USA. Initiated in 1982, the GAWs are now held in even-numbered years with the purpose of evaluating strategies for detecting genetic effects of complex diseases, thought to be the result of the joint effects of environmental and genetic factors. Each GAW meeting begins with the distribution of datasets that those who attend the Workshop use for the purpose of developing and/or evaluating statistical methods. These datasets are

jointly chosen for the next Workshop through a discussion of those attending the meeting and the GAW Advisory Committee. At most Workshops, GAW has included a set of simulated datasets, so that researchers can examine the behavior of statistical methods when knowing the answer. A primary goal of the Workshops is to focus discussion on specific topics of interest and areas of methodological concern. The datasets are generally available to any researcher who requests them. Each person who desires to attend the Workshops must participate in the evaluation of at least one of the distributed datasets, investigating novel approaches or comparing emerging and existing methods. Participants also include those who have provided the data or participate in the Workshop organization. More information about GAW, including details of upcoming Workshops, may be found at http://www.gaworkshop.org.

Genetic Analysis Workshop 16

Genetic Analysis Workshop 16 focused its efforts on the evaluation of genome-wide association studies of large genomic chip datasets containing hundreds of thousands genotypes from single-nucleotide polymorphisms (SNPs). There were three problem datasets, two consisting of data from ongoing studies and one simulated. All three datasets consisted of phenotypic and genome-wide SNP scan data. Problem 1 data came from studies of rheumatoid arthritis (RA), Problem 2 included genotypic and phenotypic data from the Framingham Heart Study (FHS), and Problem 3 consisted of simulated phenotypic data using the pedigrees and genotypic data provided to GAW16 by the Framingham Heart Study. Each of these datasets is described in more detail in Amos et al. [1], Cupples et al. [2], and Kraja et al. [3]. Data for Problems 2 and 3 required an application to the database for Genotypes and Phenotypes (dbGaP) at the National Center for Biotechnology Information [4], which processed applications through the National Heart, Lung and Blood Institute, and distributed the data. To apply, researchers needed to have an eRA Commons account, to obtain Institutional Review Board approval, to ensure security of the data and to sign a data, distribution agreement in conjunction with an institutional signing official.

Problem set 1

Data for Problem 1 was derived from a genome-wide study of RA. SNP genotype data were provided for 868 cases and 1,194 controls that had been assayed using an Illumina 550 k platform. The cases were independent individuals who had met the American College of Rheumatology criteria for RA. Four hundred forty-five cases came from a single member of sibling sets that were studied as a part of the North American Rheumatoid Arthritic Consortium (NARAC) because they had at least one additional sibling with rheumatoid arthritis; an additional 423 independent cases were included and were not selected for family history. The cases were recruited from across the United States and are predominantly of Northern European origin. The controls, derived from the New York Cancer Project, were enrolled in the New York metropolitan area and are somewhat enriched for individuals of Southern European or Ashkenazi Jewish ancestry compared with cases. Phenotypic data were also provided for *DRB1* alleles, which were classified according to the RA shared epitope, levels of anti-cyclic citrullinated peptide, and levels of rheumatoid factor IgM.

Problem set 2

Data for Problem 2 derived from a genome-wide scan conducted in Framingham Heart Study participants through the SNP Health Association Resource (SHARe). More detail describing this effort is included at the dbGaP [5]. Genotype data collected using Affymetrix 500 k (250 k Nsp and 250 k Sty) and 50 k gene centric platforms were provided for 6,848 participants with 6,621 in 766 pedigrees of three generations and 227 unrelated individuals. Phenotypic data for 7,130 participants were available for the first four examinations from the Original Cohort (recruited from 1948 to 1952) and Offspring Cohort (recruited from 1971 to 1975) and one examination for the Generation 3 Cohort (recruited from 2002 to 2005). These examinations were chosen because participants were approximately the same adult ages. Data included were demographics (sex and age), height, weight, and traditional risk factors for coronary heart disease (blood pressure and hypertension, diabetes and blood glucose, smoking, alcohol, and lipid levels). Additional data included, when appropriate, were age at onset of coronary heart disease, age at onset of diabetes, age at death, and age at last contact.

Problem set 3

Phenotypic data for Problem 3 were simulated, using the pedigrees and genotypes from Problem 2. The simulated data were derived from a model emulating lipid traits and their relationships to cardiovascular disease. Two hundred simulated replicates were provided for GAW16. For each replicate there were 6,476 subjects in families from the FHS, with their actual genotypes for Affymetrix 550 k SNPs and simulated phenotypes. The total number of subjects and pedigree structures differed from those in Problem 2, because between the times that simulation began and data were made available, additional FHS participants provided consent for use of their data. Simulated phenotypes at three visits, 10 years apart, were generated for Problem 3. Up to six "major" genes

influencing variation in high- and low-density lipoprotein cholesterol (HDL, LDL), and triglycerides (TG), and 1,000 "polygenes" were simulated for each trait. All polygenes act independently and have additive effects. A group of 39 polygenes influencing HDL were clustered on chromosome 11; otherwise, the polygenes for each trait were randomly distributed throughout the genome. At each simulated visit, individuals in the upper tail of the LDL distribution were designated as medicated. The proportion of subjects that are medicated increased across visits at 2%, 5%, and 15%. Coronary artery calcification (CAC) was simulated using age, lipid levels, and CAC-specific polymorphisms. The risk of myocardial infarction before each visit was determined by CAC and its interactions with smoking and two genetic loci. Smoking was simulated to be commensurate with rates reported by the Centers for Disease Control. The full model for these simulated data is included in Kraja et al. [3].

Individuals on the GAW mailing list of nearly 2,600 were notified through e-mail in Spring 2008 that data for the three Problems were available. A total of 183 groups requested GAW16 data: 124 for Problem 1 data and 59 for Problems 2 and 3 data, which needed to be accessed through dbGaP. In Summer 2008, 168 contributed papers were received describing analyses of these data sets. A book and CD containing these contributions plus descriptions of the data sets were distributed to GAW16 participants before the meeting in September.

The GAW16 participants included 240 individuals from all over the world, including Austria, Brazil, Canada, France, Germany, India, Korea, the Netherlands, Singapore, Spain, Taiwan, the United Kingdom, and the United States. The 168 contributions submitted to GAW16 were organized into 17 presentation groups of 7 to 18 papers each. These presentation groups were organized around the following themes: genome-wide association (GWA) for discrete traits; GWA for quantitative traits; multi-stage GWA strategies; haplotype-based analyses; controlling false-positive rates; multi-phenotype analyses; phenotype definition and development; quality control in GWA studies; machine learning; gene-gene interaction; gene-environment interaction; using gene expression, function, and pathways in GWA; combining information from linkage and association analyses; population and evolutionary genetics, including linkage disequilibrium patterns and population stratification; GWA analysis of longitudinal data; family-based GWA analyses; and gene- or region-based association analyses. Each presentation group was led by a person with previous GAW experience who facilitated group discussion, organized the group's oral presentation for the general GAW meeting, and took a lead in writing the group summary paper, which are published simultaneously with these proceedings in *Genetic Epidemiology* [6].

Members of presentation groups began interacting by e-mail and/or conference calls before GAW16, comparing and contrasting their approaches and results. Each presentation group met a full day at the Workshop, a first for GAW. During these meetings, they continued their discussions and finalized a group presentation, which was delivered to the full GAW16 audience during the general sessions on the subsequent two days. The group meetings were attended mostly by group participants, but were open to all GAW16 attendees. Seventy-two participants also contributed to poster sessions held during the general sessions. There also was a special general session on Novel Methods. Four papers submitted to GAW16 were selected before the meeting for presentation in this session because they had used or developed novel analytical approaches.

The 131 GAW contributions included in this issue of *BMC Proceedings* are a subset of the 168 contributions presented at GAW16. All contributions were peer-reviewed and selected on the basis of scientific merit.

The first three papers of these Proceedings describe the datasets. These are followed by the 131 individual GAW16 contributions organized by presentation group, and alphabetically by first author within each group. Additionally, in a supplement to the journal *Genetic Epidemiology*, published simultaneously with these Proceedings, a paper by each presentation group summarizes the contributions to that group and the lessons learned, comparing and contrasting contributions and describing their main themes and results. Overall, GAW16 generated many interesting discussions and some conclusions concerning appropriate approaches to the analysis of genome-wide association data. These discussions also highlighted areas in which further methodological development is needed.

Competing interests

The authors declare that they have no competing interests.

Acknowledgements

Literally hundreds of individuals donate their time and effort to make the Genetic Analysis Workshops possible, helping to select Workshop topics, providing real and simulated data sets to be distributed to Workshop participants, making local arrangements and staffing the registration desk at the Workshop, leading presentation groups, writing summary papers, reviewing manuscripts, and editing these proceedings.

The Genetic Analysis Workshops rely on the generous donation of existing data sets and the contributions of our colleagues who simulate

data for the Workshops. Many investigators contributed data to GAW16. Problem 1: Rheumatoid arthritis data were provided by Peter Gregersen, Christopher Amos, Wei Chen, Michael Seldin, Elaine Remmers, Lindsay Criswell, Kimberly Taylor, Annette Lee, Robert Plenge, and Daniel Kastner. Problem 2: Data from the Framingham Heart Study were provided by L. Adrienne Cupples, Nancy Heard-Costa, Monica Lee, Larry Atwood, and the Framingham Heart Study Investigators. Problem 3: The GAW16 simulated datasets were generated by Aldi Kraja, Robert Culverhouse, E. Warwick Daw, Jun Wu, Andrew Van Brunt, Michael Province and Ingrid Borecki.

The rheumatoid arthritis studies were supported by NIH grant AR44422, NIH contract N01-AR-7-2232, funding from Genome Canada and Associations AFP, Polyarctique-Groupe Taitbout and Rhumatisme et Travail. Funding for UK researchers was provided by the Arthritis Research Campaign.

Data for Problems 2 and 3 were distributed through the NIH repository dbGaP and applications for data access were reviewed by a committee at NHLBI. Many individuals at dbGaP and NHLBI contributed to data transfer, designing a GAW16 study page at the dbGaP web site [7], development of instructions for obtaining GAW16 Problem 2 and 3 data and the data use agreement, tracking data requests, and reviewing applications. We extend our thanks to Debbie Eng, Richard Fabsitz, Mike Feolo, Cashell Jaquish, Christopher O'Donnell, Susan Old, Mona Pandey, Steve Sherry, and Paul Sorlie.

Contributions to GAW16 were organized into discussion and presentation groups focused on various methodological and analytic themes. Twenty-six people generously volunteered to lead these groups, initiating interactions among group members before GAW16, leading group meetings at GAW16, organizing summary presentations for the larger GAW16 audience, serving as editors for the publication and peer review process for this volume, and taking responsibility for the preparation of a summary paper for *Genetic Epidemiology*. Being a group leader is a time consuming task and one that is critical to the success of the Workshops. As such, their efforts deserve special recognition. We are grateful to the following people who led the group discussions and preparation of summary presentations (in group numerical order): Ellen Goode, Duncan Thomas, Saurabh Ghosh, Rosalind Neuman, Elizabeth Hauser, Lisa Martin, Jack Kent, Marsha Wilcox, Andreas Ziegler, Silke Szymczak, Ping An, Michael Province, Corinne Engelman, Jim Gauderman, Nathan Tintle, Heike Bickeböller, Elizabeth Marchani, E. Warwick Daw, Ellen Wijsman, Tony Hinrichs, Brian Suarez, Berit Kerner, Danielle Fallin, Stacey Knight, Maria Martinez, and Joseph Beyene.

Useful comments and criticisms of the papers in this volume were provided by 120 scientific reviewers: Diana Abbott, Alexandre Alcais, Andrew Allen, Jennifer Asimit, Joan Bailey-Wilson, Jill Barnholtz-Sloan, Jenny Barrett, Terri Beaty, Lars Beckmann, Joanna Biernacka, Tim Bishop, Mike Boehnke, Stefan Böhringer, Jack Bowden, Alfonso Buil, Shelley Bull, Nicola Camp, Rita Cantor, Cheng Cheng, David Conti, Heather Cordell, Robert Culverhouse, Mariza De Andrade, Marcella Devoto, Guoqing Diao, Irina Dinu, Marie-Hélène Dizier, Duan Duan, Marie-Pierre Dubé, Frank Dudbridge, Priya Duggal, Josée Dupuis, Carol Etzel, Cathy Falk, Wenhong Fan, Mary Feitosa, Christine Fischer, Nora Franceschini, Brooke Fridley, Xiaoyi Gao, Emmanuelle Génin, Rodney Go, Lynn Goldin, Alisa Goldstein, Derek Gordon, Harald Göring, Courtney Grey-McGuire, Charles Gu, Peter Holmans, Bin Huang, Yifan Huang, Candace Kammerer, Xiayi Ke, Mehdi Keddache, Terri King, Alison Klein, Inke König, Peter Kraft, Lydia Kwee, Carl Langefeld, Jason Laramie, Juan P. Lewinger, Jing Li, Mingyao Li, Shuying Sue Li, Wentian Li, Yi-Ju Li, Jian'an Luan, Brion Maher, James Malley, Rasika Mathias, Nancy Mendell, Brackie Mitchell, Richard Morris, Alison Motsinger, Nandita Mukhopadhyay, Bertram Muller-Myhsok, Nora Nock, Kari North, Michael Nothnagel, Jeff O'Connell, Jurg Ott, Grier Page, V. Shane Pankratz, George Papanicolaou, Andrew Paterson, Ruth Pfeiffer, Silvano Presciuttini, Elizabeth Pugh, Dajun Qian, Evadnie Rampersaud, John Rice, Steve Rich, Marylyn Ritchie, Nancy Saccone, Glen Satten, Mike Schmidt, Silke Schmidt, Paola Sebastiani, Jemila Seid Hamid, Svati Shah, William Shannon, Janet Sinsheimer, Susan Slager, Anne Spence, Catherine Stein, Carolin Strobl, Yun Ju Sung, Alun Thomas, Duncan Thomas, John R. Thompson, David Tregouet, Ya-Yu Tsai, Grace Wahba, Kai Wang, Shuang Wang, Jessica Woo, Yun Joo Yoo, Lue Ping Zhao, Xiaofeng Zhu. We are grateful for their contributions.

Since GAW7 in 1991, Vanessa Olmo has had major responsibility for all aspects of Workshop organization. Over the years, as the Workshops have increased in size and complexity, she has taken on greatly increased responsibilities. She has primary responsibility for Workshop logistics, including interaction with participants, organizers, editors, and publisher; data distribution; site selection and liaison with local organizers; maintenance of the GAW web site, wiki, and mailing list; collation and distribution of pre-GAW papers; and preparation of the proceedings. The GAWs could not succeed without her commitment and her enthusiasm. We also thank Selina Flores who helped with data distribution, communications with participants, and preparation of the pre-GAW volume; and 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 the 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, while Maria Messenger and Malinda Mann typeset the articles. 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 ensure a successful GAW.

The GAW Advisory Committee, which has a rotating membership, has overall responsibility for long-term planning for the GAWs. Its membership at the time of GAW16 included Laura Almasy (chair), 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.

Continuous funding for the GAWs has been provided since 1982 by the National Institute of General Medical Sciences (NIGMS), through grant R01 GM31575 to Jean MacCluer and Laura Almasy. This grant also provided scholarship funds to help defray travel costs for 34 graduate students and post-doctoral trainees attending GAW16. We wish to thank Richard Anderson of NIGMS for his interest in GAW and for his efforts as Program Director for the GAW grant at the time of GAW16 and also Donna Krasnewich, who has recently taken over these duties. We are particularly grateful to Irene Eckstrand of NIGMS for her enthusiasm and interest in the GAWs since they were first envisioned in 1981. The GAWs

would not be possible without the support of these individuals and NIGMS.

We particularly thank Jean MacCluer, who envisioned the need for GAWs and pursued and obtained funding for them. Her leadership has been indispensable to the success of the GAWs.

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.

This article has been published as part of *BMC Proceedings* Volume 3 Supplement 7, 2009: Genetic Analysis Workshop 16. The full contents of the supplement are available online at http://www.biomedcentral.com/1753-6561/3?issue=S7.

References

1. Amos CI, Chen WV, Seldin MF, Remmers E, Taylor KE, Criswell LA, Lee AT, Plenge RM, Kastner DL and Gregersen PK: **Data for Genetic Analysis Workshop 16 Problem 1, association analysis of rheumatoid arthritis data.** *BMC Proc* 2009, **3(suppl 7):**S2.
2. Cupples LA, Heard-Costa N, Lee M, Atwood LD and for the Framingham Heart Study Investigators: **Genetic Analysis Workshop 16 Problem 2: The Framingham Heart Study data.** *BMC Proc* 2009, **3(suppl 7):**S3.
3. Kraja AT, Culverhouse R, Daw EW, Wu J, Van Brunt A, Province MA and 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.
4. National Center for Biotechnology Information: **Database for Genotypes and Phenotypes.** http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000128.v2.p2.
5. National Center for Biotechnology Information: **Database for Genotypes and Phenotypes. Framingham SNP Health Association Resource (SHARe). Study Accession: phs000007.v7.p4.** http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs0000 07.v7.p4.
6. MacCluer JW, Cupples LA and Almasy L: **Genetic Analysis Workshop 16: approaches to analysis of genome-wide data.** *Genet Epidemiol in press.*
7. National Center for Biotechnology Information: **Database for Genotypes and Phenotypes. GAW16 Framingham and Simulated Data. Study Accession: phs000128.v2.p2 -.** http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000128.v2.p2.