BMC Proceedings

PROCEEDINGS Open Access

Identifying rare variants from exome scans: the GAW17 experience

Saurabh Ghosh[1*], Heike Bickeböller[2], Julia Bailey[3], Joan E Bailey-Wilson[4], Rita Cantor[5], Robert Culverhouse[6], Warwick Daw[7], Anita L DeStefano[8], Corinne D Engelman[9], Anthony Hinrichs[10], Jeanine Houwing-Duistermaat[11], Inke R König[12], Jack Kent Jr[13], Nan Laird[14], Nathan Pankratz[15], Andrew Paterson[16], Elizabeth Pugh[17], Brian Suarez[10], Yan Sun[18], Alun Thomas[19], Nathan Tintle[20], Xiaofeng Zhu[21], Andreas Ziegler[12], Jean W MacCluer[13], Laura Almasy[13]

From Genetic Analysis Workshop 17
Boston, MA, USA. 13-16 October 2010

Abstract

Genetic Analysis Workshop 17 (GAW17) provided a platform for evaluating existing statistical genetic methods and for developing novel methods to analyze rare variants that modulate complex traits. In this article, we present an overview of the 1000 Genomes Project exome data and simulated phenotype data that were distributed to GAW17 participants for analyses, the different issues addressed by the participants, and the process of preparation of manuscripts resulting from the discussions during the workshop.

Introduction

This supplement of *BMC Proceedings* contains the proceedings of Genetic Analysis Workshop 17 (GAW17), which was held October 13–16, 2010, in Boston, Massachusetts, USA. The Genetic Analysis Workshops began in 1982 and are now held in even-numbered years. They provide a forum for investigators interested in identifying genetic effects on complex diseases to evaluate and compare novel and existing statistical methods. The purpose of these workshops is to allow the comparison of statistical methods for genetic epidemiology using common, well-described data sets. Before each workshop, topics are chosen, one or more existing data sets are selected, and a set of simulated data is created that permits investigation of current questions of broad interest in statistical genetics. These data are made available to any scientists who request them, and their analyses of these data are presented at the workshop. Participation in the workshop is open to anyone who submits an analysis of one of these data sets, provides data, or participates in workshop organization. More information about the Genetic Analysis Workshops, including details on upcoming meetings, can be found at http://www.gaworkshop.org.

Genetic Analysis Workshop 17

The backdrop of GAW17 was the failure of genome-wide association studies (GWAS) to identify a set of single-nucleotide polymorphisms (SNPs) that could jointly explain a substantial proportion of the heritability in the trait for many common diseases. There is an increasing belief that the common variant/common disorder paradigm, which forms the basis for GWAS, may not be the appropriate model for describing complex disorders. An alternative paradigm is that the "missing heritability" can be explained by rare variants that cannot be identified using GWAS.

The major focus of GAW17 was the statistical challenges that arise in association analyses of exome scan data composed of real sequence information on a large number of genes from the 1000 Genomes Project and simulated phenotypes. The primary objective was to evaluate existing methods and develop novel methods to identify rare variants that modulate the phenotypes. There were two data sets: one on 697 unrelated individuals and the other on the same number of individuals

* Correspondence: saurabh@isical.ac.in
[1]Human Genetics Unit, Indian Statistical Institute, Kolkata 700018, India
Full list of author information is available at the end of the article

BioMed Central

© 2011 Ghosh et al; licensee BioMed Central Ltd. This is an open access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

distributed in 8 extended families. In the family data 202 founders were chosen at random from the set of 697 unrelated individuals. All the individuals were modeled on subjects from the 1000 Genomes Project; their genotypes were obtained from the sequence data available in that database, and their phenotypes were simulated to produce a disease trait and related quantitative risk factors influenced by multiple genes.

SNP genotypes were obtained from the sequence alignment files provided by the 1000 Genomes Project for their pilot3 study (http://www.1000genomes.org). The UnifiedGenotyper method from the Genome Analysis Toolkit (GATK) package (http://www.broadinstitute.org/gsa/wiki/index.php/The_Genome_Analysis_Toolkit) was used for the detection of SNPs and for the calling of SNP genotypes. Because the 1000 Genomes Project genotypes were not phased and because some genotypes were missing as a result of incomplete sequence coverage in some individuals, the program fastPHASE (http://depts.washington.edu/uwc4c/express-licenses/assets/fastphase/) was used to infer missing genotypes and haplotypic phase. In the family data set, the program CHRSIM [1] was used to drop the phased founder genotypes throughout the rest of the pedigree. For each of the 24,487 autosomal SNPs identified in 3,205 genes, the information provided included the chromosome and base-pair location, the name of the gene in which it was located, whether the SNP was synonymous or nonsynonymous, and the minor allele frequency. For the family data set, fully informative markers were generated at each gene and were used to compute identity-by-descent scores at each gene location.

Two hundred simulation replicates were carried out in both data sets. The genotypes were held fixed for all the replicates. Data on three quantitative phenotypes and a binary affection status phenotype were generated. Simulated data were also available on three covariates: Age, Sex, and Smoking status. A more complete description of the GAW17 data is provided by Almasy et al. [2].

The availability of the GAW17 data was announced by e-mail in the summer of 2010 to the more than 2,600 individuals on the Genetic Analysis Workshop mailing list. Two hundred four groups requested GAW17 data. One hundred sixty-six contributed papers were received that described analyses of the data sets. The GAW17 participants included 274 individuals from 19 countries: Australia, Austria, Belgium, Canada, China, Costa Rica, France, Germany, Hong Kong, India, the Netherlands, Singapore, South Korea, Spain, Switzerland, Taiwan, United Kingdom, United States, and US Virgin Islands. The 166 submitted contributions were organized into 15 presentation groups based on common methodological themes. The themes of the different presentation groups were genes with multiple rare variants (Group 1), identification of rare functional variants (Group 2), use of predicted function of gene or SNP (Group 3), identification or incorporation of gene-environment interactions (Group 4), comparison of unrelated and family data (Group 5), conditioning on known genes or variants (Group 6), scoring routines or aggregate effects (Group 7), multiple testing (Group 8), impact of linkage disequilibrium (Group 9), joint analyses of disease and risk factors (Group 10), incorporation of linkage information (Group 11), tagging of rare variants with common variants (Group 12), haplotype-based analyses (Group 13), regression and data mining methods for multiple rare variants (Group 14), and collapsing methods for rare variants (Group 15). Each presentation group was led by a person with previous Genetic Analysis Workshop experience. This person facilitated group discussions, organized the group's oral presentation to the general meeting, and took the lead in writing the group summary paper (published in *Genetic Epidemiology*).

Members of most presentation groups began interacting before GAW17 through e-mail and a discussion forum set up on the Genetic Analysis Workshop website, comparing and contrasting their approaches and results. Each presentation group also met at least once during the workshop, where they continued their discussions and finalized a group presentation that was delivered to the full GAW17 audience during the general sessions. The group meetings were attended mostly by group participants but were open to all GAW17 attendees. During poster sessions, 87 individual contributions were presented. The 119 GAW17 contributions included in this issue of *BMC Proceedings* are a subset of the 166 contributions presented at GAW17. All these papers have been peer-reviewed and were selected on the basis of scientific merit.

The first paper in this proceedings describes the data set provided to the participants of GAW17. This is followed by the 119 individual contributions organized by presentation group and alphabetically by first author within each group. In addition, in a forthcoming supplement to the journal *Genetic Epidemiology*, a paper by each presentation group summarizes the contributions to that group and a concluding paper on the lessons learned compares and contrasts the contributions and describes their main themes and results. Overall, GAW17 generated many interesting discussions and some conclusions concerning appropriate approaches for analyzing sequence data and identifying rare causal variants. These discussions also highlighted areas in which further methodological development is needed. A general summary of these overall GAW17 conclusions is provided by Wilson and Ziegler [3].

Acknowledgments

The Genetic Analysis Workshops would not succeed without the dedicated efforts of a huge number of individuals. These include those who help to select workshop topics, provide real and simulated data sets to be

distributed to workshop participants, make local arrangements and staff the registration desk at the workshop, lead presentation groups, write summary papers, review manuscripts, and edit these proceedings.
Contributions to GAW17 were organized into discussion and presentation groups focused on various methodological and analytic themes. Twenty-two people generously volunteered to lead these groups, initiating interactions among group members before GAW17, leading group meetings at GAW17, organizing summary presentations for the larger GAW17 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*. Their efforts deserve special recognition. We are grateful to the following people, who led the group discussions and prepared the summary presentations: Julia Bailey, Joan Bailey-Wilson, Heike Bickeböller, Rita Cantor, Rob Culverhouse, E. Warwick Daw, Anita DeStefano, Corrine Engleman, Anthony Hinrichs, Jeanine Houwing-Duistermaat, Jack Kent Jr, Inke König, Nan Laird, Nathan Pankratz, Andrew Paterson, Elizabeth Pugh, Brian Suarez, Yan Sun, Alun Thomas, Nathan Tintle, Xiaofeng Zhu, and Andreas Ziegler.
Useful comments and criticisms of the papers in this volume were provided by 125 scientific reviewers: Alexandre Alcais, Laura Almasy, Christopher Amos, Ping An, Allison Ashley-Koch, Beth Atkinson, Christy Avery, Marie-Claude Babron, Michael Badzioch, Raji Balasubramanian, M. Michael Barmada, Jenny Barrett, Saonli Basu, Justo Lorenzo Bermejo, Joanna Biernacka, Timothy Bishop, Michael Boehnke, Stefan Boehringer, Karl Broman, Sharon Browning, Alfonso Buil, Shelley Bull, Alexandre Bureau, William Bush, Andrea Callegaro, Nicola Camp, Daniel Chassmen, Wei-Min Chen, Ching-Yu Cheng, Erica Childs, Andy Collins, Heather Cordell, Karen Cuenco, Mariza de Andrade, Marcella Devoto, Marie-Pierre Dube, Priya Duggal, Josee Dupuis, Jeanette Eckel-Passow, Paul Eilers, Sarah Ennis, Cathy Falk, Dani Fallin, Cathy Fann, Christine Fischer, Nora Franceschini, France Gagnon, Xiaoyi Gao, Chad Garner, Emanuelle Genin, Lynn Goldin, Alisa Goldstein, Ellen Goode, Harald Göring, Celia Greenwood, Fangyi Gu, Johathan Haines, Elizabeth Hauser, Yuan Jiang, Suh-Hang Hank Juo, Cristina Justice, Xiayi Ke, Abbas Khalili, Alison Klein, Daniel L. Koller, Aldi Kradja, Peter Kraft, Ake Ku, Kenneth Lange, Carl Langfeld, Bingshan Li, Chun Li, Wentian Li, Liming Liang, Jian'an Luan, Brion Maher, Partha Pratim Majumder, James Malley, Lisa Martin, Maria Martinez, Brett McKinney, Nancy Mendall, Yan Meng, Brackie Mitchell, Andrew Morris, Alison Motsinger-Reif, Cassandra Mucray, Indranil Mukhopadhyay, Nandita Mukhopadhyay, Bertram Muller-Myhsok, Rosalind Neuman, Nora Nock, Kari North, Michael Nothnagel, Jeff O'Connell, George Papanicolaou, Andrew Paterson, Ruth Pfeiffer, Dajun Qian, Evadnie Rampersaud, John Rice, Steve Rich, Marylyn Ritchie, Andre Scherag, Audrey Schnell, Mary Sehl, Claire Simpson, Janet Sinsheimer, Anne Spence, Hans Stassen, Catherine M. Stein, Marc Suchard, Yun Ju Sun, Heejong Sung, Michael Swartz, Silke Szymczak, Duncan Thomas, Asuman Turkmen, Kai Wang, Shuang Wang, Ellen Wijsman, Marsha Wilcox, Mary Wojczynski, Qunyuan Zhang, Hongyu Zhou. 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 Genetic Analysis Workshops web site and mailing list; and preparation of the proceedings. The workshops could not succeed without her commitment and her enthusiasm. We also thank Selina Flores, Gene Hopstetter, Richard Polich, Rene Sandoval, Rudy Sandoval, and Gerry Vest, who helped with data distribution, communications with participants, and preparation of the online preworkshop volume; and Thomas Dyer, John Blangero, Juan Peralta, Joanne Curran, Jack Kent, and Jac Charlesworth, who worked on simulating the mini-exome data set for GAW17 and prepared the data for distribution. Mimi Braverman assisted with the editing of the GAW17 proceedings, and Maria Messenger and Malinda Mann formatted the articles and prepared the files.
Local arrangements for GAW17 required many hours of planning and organization. We are grateful to local organizer Adrienne Cupples as well as volunteers Yansong Cheng, Seung Hoan Choi, Wei Gao, Audrey Hendricks, Susan Hwang, Denver Lybarger, Alisa Manning, Julius Ngwa, Ed Olmo, Jing Wang, Zheng Xiang, Baiyun Yao, and Vivian Zhuang for welcoming us to Boston and for their efforts to ensure a successful workshop.
The GAW Advisory Committee, which has a rotating membership, has overall responsibility for long-term planning for the workshops. Its membership at the time of GAW17 included Laura Almasy (chair), Ingrid Borecki, Adrienne Cupples, Saurabh Ghosh, Elizabeth Hauser, Jean MacCluer, Maria Martinez, Glen Satten, Ellen Wijsman, John Witte, Xiaofeng Zhu, and Andreas Ziegler.
Continuous funding for the Genetic Analysis Workshops has been provided since 1982 by the National Institute of General Medical Sciences (NIGMS), through National Institutes of Health grant R01 GM31575 awarded to Jean MacCluer and Laura Almasy. This grant also provided scholarship funds to help defray travel costs for 40 graduate students and postdoctoral trainees attending GAW17. We wish to thank Donna Krasnewich for her interest in the Genetic Analysis Workshops and for her efforts as Program Director for the workshop grant at the time of GAW17. We are also grateful to Irene Eckstrand of NIGMS for her enthusiasm and interest in the workshops since they were first envisioned in 1981. The workshop's results, published in *BMC Proceedings* (2011), would not be possible without the support of these individuals and NIGMS.
We particularly thank Jean MacCluer, who envisioned the need for genetic analysis workshops and pursued and obtained funding for them. Her leadership has been indispensable to the success of the workshops.
As always, we wish to express our appreciation to the Genetic Analysis Workshop participants, without whose ongoing, enthusiastic support the workshops could not have enjoyed their continuing success.
This article has been published as part of *BMC Proceedings* Volume 5 Supplement 9, 2011: Genetic Analysis Workshop 17. The full contents of the supplement are available online at http://www.biomedcentral.com/1753-6561/5?issue=S9.

Author details
[1]Human Genetics Unit, Indian Statistical Institute, Kolkata 700018, India. [2]University of Göttingen Medical School, Göttingen, 37073, Germany. [3]Department of Epidemiology, David Geffen School of Medicine, University of California, Los Angeles, CA 90095 USA; and Research Service, VA GLAHS Epilepsy Center of Excellence, Epilepsy Genetics/Genomics Laboratories, Los Angeles, CA, USA. [4]National Human Genome Research Institute, National Institutes of Health, Baltimore, MD 21224, USA. [5]Department of Human Genetics, David Geffen School of Medicine, University of California, Los Angeles, CA 90095, USA. [6]Department of Internal Medicine, Washington University School of Medicine, St Louis, MO 63110, USA. [7]Division of Statistical Genomics, Washington University School of Medicine, St Louis, MO 63110, USA. [8]School of Public Health, Boston University, Boston, MA 02118, USA. [9]School of Medicine and Public Health, University of Wisconsin, Madison, WI 53726, USA. [10]Department of Psychiatry, Washington University School of Medicine, St Louis, MO 63110, USA. [11]Department of Medical Statistics and Bioinformatics, Leiden University Medical Center, Leiden 2300RC, The Netherlands. [12]Institut für Medizinische Biometrie und Statistik, Universität zu Lübeck, Universitätsklinikum Schleswig-Holstein, Campus Lübeck, Lübeck, Germany. [13]Department of Genetics, Texas Biomedical Research Institute, San Antonio, TX 78245, USA. [14]Department of Biostatistics, Harvard School of Public Health, Boston, MA 02115, USA. [15]Department of Laboratory Medicine and Pathology, University of Minnesota, Minneapolis, MN 55104, USA. [16]Division of Biostatistics, University of Toronto, Toronto, ON M5S 3G4, Canada. [17]Center for Inherited Disease Research, Johns Hopkins University, Baltimore, MD 21202, USA. [18]Department of Epidemiology, Emory University, Atlanta, GA 30322, USA. [19]Department of Epidemiology, University of Utah, Salt Lake City, UT 84108, USA. [20]Department of Math and Computer Science, Dordt College, Sioux Center, IA 51250, USA. [21]Department of Epidemiology and Biostatistics, Case Western Reserve University School of Medicine, Cleveland, OH 44106, USA.

Competing interests
The authors declare that there is no conflict of interests.

Published: 29 November 2011

References
1. Speer M, Terwilliger JD, Ott J: **A chromosome-based method for rapid computer simulation.** *Am J Hum Genet* 1992, **52**:A202.
2. Almasy LA, Dyer TD, Peralta JM, Kent JW Jr, Charlesworth JC, Curran JE, Blangero J: **Genetic Analysis Workshop 17 mini-exome simulation.** *BMC Proc* 2011, **5(suppl 9)**:S2.

3. Wilson AF, Ziegler A: **Lessons learned from Genetic Analysis Workshop 17: transitioning from genome-wide association studies to whole-genome statistical genetic analysis.** *Genet Epidemiol* **X(suppl X)**:XX-XX.

doi:10.1186/1753-6561-5-S9-S1
Cite this article as: Ghosh *et al.*: Identifying rare variants from exome scans: the GAW17 experience. *BMC Proceedings* 2011 **5**(Suppl 9):S1.

Submit your next manuscript to BioMed Central and take full advantage of:

- **Convenient online submission**
- **Thorough peer review**
- **No space constraints or color figure charges**
- **Immediate publication on acceptance**
- **Inclusion in PubMed, CAS, Scopus and Google Scholar**
- **Research which is freely available for redistribution**

Submit your manuscript at
www.biomedcentral.com/submit

BioMed Central