The Author(s) *BMC Proceedings* 2016, **10**(Suppl 7):19
DOI 10.1186/s12919-016-0007-z

BMC Proceedings

PROCEEDINGS Open Access

Genetic Analysis Workshop 19: methods and strategies for analyzing human sequence and gene expression data in extended families and unrelated individuals

Corinne D. Engelman[1*], Celia M. T. Greenwood[2], Julia N. Bailey[3], Rita M. Cantor[4], Jack W. Kent Jr[5], Inke R. König[6], Justo Lorenzo Bermejo[7], Phillip E. Melton[8], Stephanie A. Santorico[9], Arne Schillert[10], Ellen M. Wijsman[11], Jean W. MacCluer[5] and Laura Almasy[12]

From Genetic Analysis Workshop 19
Vienna, Austria. 24-26 August 2014

Abstract

Genetic Analysis Workshop 19 provided a platform for developing and evaluating statistical methods to analyze whole-genome sequence and gene expression data from a pedigree-based sample, as well as whole-exome sequence data from a large cohort of unrelated individuals. In this article we present an overview of the data sets, the GAW experience, and summaries of the contributions arranged into nine methodological themes.

Introduction

This supplement to *BMC Proceedings* contains the proceedings of the Genetic Analysis Workshop 19 (GAW19), which was held August 24–27, 2014 in Vienna, Austria. The GAWs began in 1982 and are now held every two years. They provide a forum for statisticians, epidemiologists, geneticists, bioinformaticians, and other scientists interested in identifying genetic effects on complex diseases to evaluate and compare novel and existing statistical methods. Prior to each GAW, topics are chosen based on suggestions from previous attendees, an existing data set(s) is selected, and a set of simulated data is devised such that statistical questions of wide and current interest may be addressed. These data sets are made available to any researcher who requests them. The same data sets are provided to all researchers, thus facilitating the discussion and comparison of methods. After the GAW organizers release the data sets, researchers analyze the data and prepare a manuscript to submit to the workshop. Participation in the workshop is open to anyone who submits a manuscript, provides data, or participants in the workshop organization. More information about the GAWs, including details on upcoming meetings, can be found at http://www.gaworkshop.org.

Genetic Analysis Workshop 19

The family dataset provided for GAW18 was used again in GAW19 with a few small corrections. New data for GAW19 included gene expression profiles for the family data set and a relatively large data set of unrelated individuals. As in past years, a simulated phenotype data set was also provided. A brief description of the data sets follows while a more detailed description can be found in Blangero et al. [1].

A family data set was provided by the Type 2 Diabetes Genetic Exploration by Next-Generation Sequencing in Ethnic Samples [T2D-GENES] Consortium. It included data from 20 Mexican American families from San Antonio, Texas, USA, with whole genome sequence information on 464 individuals. The data set also included dense single nucleotide polymorphisms (SNPs) on 959

* Correspondence: cengelman@wisc.edu
[1]Department of Population Health Sciences, School of Medicine and Public Health, University of Wisconsin, 610 Walnut Street, 707 WARF, Madison, WI 53726, USA
Full list of author information is available at the end of the article

individuals, including the 464 sequenced subjects whose genotypes served as the input for the imputation procedure. Genotype data were provided for odd numbered autosomes only, and contained sequence data, data from a genome-wide Illumina chip containing almost 500 K SNPs, and variant dosages from the Merlin-based imputation procedure. Gene expression was measured in a subset of 647 individuals using peripheral blood mononuclear cells (PBMCs) collected at the first examination and an Illumina chip. The phenotype data were longitudinal measurements of systolic and diastolic blood pressure, sex, age, year of examination, use of antihypertensive medication and tobacco smoking.

A data set of unrelated individuals was also provided by the T2D-GENES Consortium. It included 1943 Hispanic individuals (1021 T2D cases and 922 controls) with whole-exome sequence data. For this data set, only samples and variants passing extensive quality control were provided. As with the family data set, only genotype data for odd numbered autosomes were provided. The phenotypic data included the same basic traits as the family data set, but were available only at a single time point.

A simulated data set of 200 phenotype replicates was provided for both the family and the unrelated data sets. It was based closely on the real data, with the family structure (for the family data set), sex, and age taken directly from the real data. Blood pressure, medication use, and tobacco smoking were generated anew for each replicate, using the distributional structure found in the real data. The simulated values of systolic and diastolic blood pressure were influenced by over 1000 variants in over 200 genes. In addition, a normally-distributed trait, Q1, was simulated that was not influenced by any genetic variants, but was correlated between family members (in the family data set). The simulation model is described in detail in Blangero et al. [1].

The availability of the GAW19 data was announced by email in Spring of 2014 to roughly 3500 individuals on the GAW mailing list. A total of 121 groups requested GAW19 data and 87 manuscripts were submitted to GAW organizers prior to the workshop. Submitting authors were asked to select a topic that their research was most aligned with to facilitate discussion before and during the workshop. This resulted in 9 discussion/presentation groups: gene expression (Group 1), machine learning and data mining (Group 2), variant collapsing approaches (Group 3), family-based approaches (Group 4), filtering variants and placing informative priors (Group 5), methods for joint analysis of multiple phenotypes (Group 6), longitudinal analyses (Group 7), pathway-based analyses (Group 8), population-based association (Group 9). The GAW19 participants included 115 individuals from 16 countries: Australia, Belgium, Canada, China, Egypt, Germany, Hong Kong, Japan, the Netherlands, Poland, South Korea, Spain, Taiwan, Turkey, the United Kingdom, and the United States of American.

At GAW19, all groups were led by a person with previous GAW experience. This person encouraged and organized the discussion and presentations prior to, during and after the workshop. Discussions largely started before the workshop and continued at the workshop within group meetings. Each discussion group, directed by the group leader, was also in charge of preparing a presentation of the issues discussed in the group and the conclusions drawn. These presentations were made to all GAW attendees in plenary sessions. There were also two poster sessions where individual contributions could be presented.

After the workshop, participants were given just over two months to revise and resubmit their manuscript for external peer review by experts in the field. The group leader typically served as associate editor for the group. To avoid possible conflicts of interest, articles to which the group editor contributed were reassigned to other groups for the peer-review process. Of the 79 manuscripts submitted after the workshop, 57 were accepted for publication in this issue of *BMC Proceedings*. The papers are organized according to the group they were in, preceded by the data description by Blangero et al. [1].

The nine GAW19 group leaders each summarized the contributions to their group and reviewed the relevant literature in short manuscripts published in a supplement to *BMC Genetics*. These 9 summary papers, with their short reviews of the state of each field, will provide a useful entry point to researchers working with genomic data. A summary of these papers follows.

The summary paper on family-based approaches led by Wijsman [2] provides a brief history of family-based genetic studies and describes how and why such studies are enjoying resurgence, partially due to the enrichment of rare causal genetic variants in such samples. The specific topics addressed in the contributions varied widely, from initial study design questions, through quality control to many aspects of data analysis, and found numerous benefits associated with studies of carefully selected related individuals.

Two groups discussed issues around leveraging external information. In the group led by Bailey [3], they provide an overview of concepts and commonly used approaches for annotating variants in the genome, as well as a survey of several principles that are used for filtering or restricting analysis to only a subset of the variants. The participants found, in general, that appropriate choices of filters or priors increases power, not only due to increasing the sizes of true signals, but also due to reducing the number of tests performed or the proportion of null tests. The topic of pathway

analysis was addressed in the group led by Kent [4]. Some participants in this group developed new approaches to pathway analysis, and many used the simulated data to assess performance. Many in the group experienced challenges in coping with the dimensionality of the data and, due to the imperfections of the required databases usually required for pathway analysis, the generalizability of identified pathways was a concern.

Tests of association between genetic data and phenotypes were discussed in several of the working groups. The summary paper from the population-based association group led by Bermejo [5] presents various new methods for testing association, as well as numerous strategies for coping with the large number of sequence-identified rare variants and how to decide on validation strategies. Particular attention was drawn to problems of estimation and convergence with sparse data, particularly when additional covariates are being explored. A useful table is provided that contains references for most key software and methods used by the group. In the summary paper from the rare variant tests group led by Santorico [6], an extensive review of methods is provided followed by carefully placing the 6 contributions into the resulting analytic framework. The discussion highlights the need for future extensions and generalizations of the concept of collapsing tests. Melton led a group that addressed tests of association with longitudinal phenotypes [7]. This paper describes recent publications on longitudinal data, the computational challenges, and benefits in power and understanding resulting from appropriate longitudinal analyses. Contributors to a group led by Schillert, developed and compared methods for analysis of multivariate phenotypes [8]. They frame the heterogeneous terminology and goals that are in use for analysis and interpretation of multivariate phenotypes. Although the approaches used by participants were diverse, all showed potential, both in terms of power and computational feasibility.

Cantor led the group studying methods for analysis of gene expression data [9]. The paper provides a quick yet broad overview of the ways in which gene expression data have typically been analyzed since high-throughput arrays became accessible, as well as important design and analysis issues. One of the key issues addressed by most group members was how to utilize gene expression measures taken from related individuals. Finally, a group on machine learning and data mining led by König [10] also addressed questions around data integration. Their paper is organized around key messages such as the benefits of integrating data of different types and the computational limitations. For each of the messages, a brief introduction presents some key references prior to introducing the work of the group.

Overall GAW19 generated many interesting discussions and some conclusions concerning the analysis of human sequence and gene expression data. These discussions also highlighted areas in which further methodological development is needed.

Acknowledgements

Numerous individuals contribute to GAW by helping select Workshop topics, providing data sets, conducting simulations, distributing data to the participants, leading discussion groups, overseeing the writing of group summaries, reviewing manuscripts, and preparing everything that needs to be done for the event management of the workshop and for the publishing process afterwards.

We are grateful to the T2D-GENES consortium for allowing GAW19 participants to use the whole-exome and -genome sequence, gene expression, and blood pressure data sets around which this Workshop was based. The T2D-GENES Consortium is supported by NIH grants U01 DK085524, U01 DK085584, U01 DK085501, U01 DK085526, and U01 DK085545. The other genetic and phenotypic data for GAW19 were provided by the San Antonio Family Heart Study and San Antonio Family Diabetes/Gallbladder Study, which are supported by NIH grants R01 HL0113323, P01 HL045222, R01 DK047482, and R01 DK053889. Additional Starr County genotype and phenotype data were supported by NIH grants R01 DK073541 and R01 HL102830. The VAGES study was supported by a Veterans Administration Epidemiologic grant. The FIND-SA study was supported by NIH grant U01 DK57295. The GAW is supported by NIH grant R01 GM031575.

The GAW19 discussion groups were led by Julia Bailey, Justo Lorenzo Bermejo, Rita M. Cantor, Jack W. Kent Jr., Inke R. König, Phillip Melton, Stephanie Santorico, Arne Schillert, and Ellen M. Wijsman. We are grateful to them for their work before, during, and after GAW19 in initiating, organizing and overseeing pre-workshop communication, group discussions, group presentations, and summary paper writing.

A total of 64 individuals assisted in peer review of the papers in this volume: Chris Amos, Marie-Claude Babron, Joan Bailey-Wilson, Elizabeth M. Blue, Jenny Barrett, Sharon Browning, Shelley Bull, Gemma Cadby, Jenny Chang-Claude, Jac C. Charlesworth, Heather Cordell, Robert Culverhouse, L. Adrienne Cupples, Kiranmoy Das, Mariza deAndrade, Vincent P. Diego, Josée Dupuis, Michael P. Epstein, David Fardo, Christine Fischer, Nora Francheschini, France Gagnon, Lynn Goldin, Derek Gordon, Harald Göring, Audrey Hendricks, Jeanine Houwing-Duistermaat, Yijuan Hu, Iuliana Ionita-Laza, Yuan Jiang, Jack W. Kent Jr., Mark K. Kos, Jinghua Liu, Sharon Lutz, James Malley, Dörthe Malzahn, Maria Martinez, Braxton Mitchell, Ben Neale, Kari North, Michael Nothnagel, Andrew Paterson, Elizabeth Pugh, Steve Rich, Marylyn Ritchie, Mohammed H. Saad, Nancy Saccone, Glen A. Satten, Daniel Schaid, André Scherag, Mary Sehl, Kim Siegmund, Henner Simianer, Claire L. Simpson, Janet Sinsheimer, Eric Sobel, Hans Stassen, Yun Ju Sung, Silke Szymczak, Elizabeth A Thompson, Timothy Thornton, Nathan Tintle, Mengyuan Xu, and Peng Zhang.

A total of 19 individuals assisted in the peer review process for summary papers: Shelley Bull, Karim Oualkacha, Angelo Canty, Li Hsu, Jennifer Listgarten, Chantal Mérette, Laurent Briollais, Catherine Stein, Pingzhao Hu, Rafal Kustra, Aurélie Labbe, Joseph Beyene, Omar de la Cruz, Andrew Paterson, Duncan Thomas, John Witte, Jinko Graham, Lei Sun, and Andrew Morris. We are grateful to each of them for their constructive comments, criticisms, and feedback.

Beginning with GAW7 in 1991, Vanessa Olmo has been responsible for major aspects of Workshop organization. We are grateful to her for the many things she does that keep GAW running smoothly, which includes interacting with participants, organizers, editors, and publishers; coordinating data requests and data distribution; facilitating selection of Workshop sites and making local arrangements; maintaining the GAW web site and mailing list; and preparing many aspects of the GAW proceedings. Vanessa Olmo assisted with distribution of data, communication with participants and preparation of the pre-workshop volume. Ravindranath Duggirala, Tanya Teslovich, Xueling Sim, Sharon Fowler, Thomas Dyer, John Blangero, Juan Peralta, Marcio Almeida, and Jack Kent, worked on data simulation and preparation. Sophie Colunga assisted with the publication process, reviews and communications, while Malinda Mann typeset the articles for these proceedings.

The GAW Advisory Committee assists with planning for the GAWs, including selection of workshop sites and topics. At the time of GAW19, its members were: Laura Almasy (chair), Shelly Bull, Adrienne Cupples, Corinne Engelman, Jim Gauderman, Jeanine Houwing-Duistermaat, Inke König, Jean MacCluer, Andrew Patterson, Michael Province, and Ellen Wijsman.
Since 1982, GAW has been funded 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 assist graduate students and post-doctoral trainees attending GAW19. We would like to recognize Donna Krasnewich for her ongoing support and for her efforts as Program Director for the GAW grant at the time of GAW19. These proceedings, as well as the continued work of statistical genetic methods development through the collaborative format of the Genetic Analysis Workshops, would not be possible without her support and that of NIGMS. We are particularly grateful to Jean MacCluer, without her there would be no GAW.
As always, we wish to express our gratitude to the GAW participants, whose ongoing, enthusiastic support and vigorous scientific discussions are the very foundation of the Workshop.

Declarations

This article has been published as part of *BMC Proceedings* Volume 10 Supplement 7, 2016: Genetic Analysis Workshop 19: Sequence, Blood Pressure and Expression Data. Summary articles. The full contents of the supplement are available online at http://bmcproc.biomedcentral.com/articles/supplements/volume-10-supplement-7. Publication of the proceedings of Genetic Analysis Workshop 19 was supported by National Institutes of Health grant R01 GM031575.

Authors' contributions

JNB, RMC, CDE, CMTG, JWK, IRK, JLB, PEM, SAS, EMW, JWM, and LA participated in workshop organization and editing of the GAW19 Proceedings. CDE and CMTG drafted the text of this manuscript with contributions from LA. All authors read and approved the final manuscript.

Competing interests

All authors declare that they have no competing interests.

Author details

[1]Department of Population Health Sciences, School of Medicine and Public Health, University of Wisconsin, 610 Walnut Street, 707 WARF, Madison, WI 53726, USA. [2]Lady Davis Institute for Medical Research, Jewish General Hospital, 3755 Côte Ste. Catherine, Montreal, QC H3T 1E2, Canada. [3]Department of Epidemiology, University of California Los Angeles Fielding School of Public Health, Box 951772, Los Angeles, CA 90095, USA. [4]Department of Human Genetics, David Geffen School of Medicine at UCLA, 695 Charles E. Young Dr, South, Los Angeles, CA 90024-7088, USA. [5]Department of Genetics, Texas Biomedical Research Institute, PO Box 760549, San Antonio, TX 78245-0549, USA. [6]Institut für Medizinische Biometrie und Statistik, Universität zu Lübeck, Universitätsklinikum Schleswig-Holstein, Campus Lübeck, Lübeck, Germany. [7]Statistical Genetics Group, Institute of Medical Biometry and Informatics, University of Heidelberg, Im Neuenheimer Feld 305, 69120 Heidelberg, Germany. [8]Centre for Genetic Origins of Health and Disease, University of Western Australia, Perth, WA, Australia. [9]Department of Mathematical & Statistical Sciences, University of Colorado-Denver, PO Box 173364, Denver, CO 80204, USA. [10]Institut für Medizinische Biometrie und Statistik, Universität zu Lübeck, Ratzeburger Allee 160, 23562 Lübeck, Germany. [11]Department of Medicine, Department of Biostatistics, Division of Medical Genetics, University of Washington, Seattle, WA 98195, USA. [12]South Texas Diabetes and Obesity Institute, University of Texas Rio Grande Valley, San Antonio, TX 78229, USA.

Published: 18 October 2016

References

1. Blangero J, Teslovich TM, Sim X, Almeida MA, Jun G, Dyer TD, Johnson M, Peralta JM, Manning AK, Wood AR, et al. Omics-squared: Human genomic, transcriptomic and phenotypic data for Genetic Analysis Workshop 19. BMC Proc. 2015;9 Suppl 8:S2.
2. Wijsman EM. Family-based approaches: design, imputation, analysis, and beyond. BMC Genet. 2015;16 Suppl 3:S9.
3. Friedrichs S, Malzahn D, Pugh EW, Almeida MA, Liu XQ, Bailey JN. Filtering genetic variants and placing informative priors based on putative biological function. BMC Genet. 2015;16 Suppl 3:S4.
4. Kent Jr JW. Pathway-based analyses. BMC Genet. 2015;16 Suppl 3:S5.
5. Bermejo JL. Above and beyond state-of-the-art approaches to investigate sequence data: summary of methods and results from the population-based association group at the Genetic Analysis Workshop 19. BMC Genet. 2015;16 Suppl 3:S1.
6. Santorico SA, Hendricks AE. Progress in methods for rare variant association. BMC Genet. 2015;16 Suppl 3:S7.
7. Chiu Y-F, Lee C-Y, Hsu F-C. Multipoint association mapping for longitudinal family data: an application to hypertension phenotypes. BMC Proc. 2015;9 Suppl 8:S43.
8. Schillert A, Konigorski S. Joint analysis of multiple phenotypes: summary of results and discussions from the Genetic Analysis Workshop 19. BMC Genet. 2015;16 Suppl 3:S8.
9. Cantor RM, Cordell HJ. Gene expression in large pedigrees: analytic approaches. BMC Genet. 2015;16 Suppl 3:S2.
10. König IR, Auerbach J, Gola D, Held E, Holzinger ER, Legault M-A, Sun R, Tintle N, Yang H-C. Machine learning and data mining in complex genomic data-a review on the lessons learned in Genetic Analysis Workshop 19. BMC Genet. 2015;16 Suppl 3:S6.