BMC Genetics

Supplement Preface

Genetic Analysis Workshop 14: microsatellite and single-nucleotide polymorphism marker loci for genome-wide scans

Joan E Bailey-Wilson*[1], Laura Almasy[2], Mariza de Andrade[3], Julia Bailey[4], Heike Bickeböller[5], Heather J Cordell[6], E Warwick Daw[7], Lynn Goldin[8], Ellen L Goode[9], Courtney Gray-McGuire[10], Wayne Hening[11], Gail Jarvik[12], Brion S Maher[13], Nancy Mendell[14], Andrew D Paterson[15], John Rice[16], Glen Satten[17], Brian Suarez[16], Veronica Vieland[18], Marsha Wilcox[19], Heping Zhang[20], Andreas Ziegler[21] and Jean W MacCluer[2]

Address: [1]Inherited Disease Research Branch, National Human Genome Research Institute, National Institutes of Health, Baltimore, Maryland 21224, USA, [2]Department of Genetics, Southwest Foundation for Biomedical Research, San Antonio, Texas, USA, [3]Mayo Clinic, Rochester, Minnesota, USA, [4]University of California Los Angeles, Los Angeles, California, USA, [5]Universitat Gottingen, Germany, [6]Department of Medical Genetics, University of Cambridge, UK, [7]University of Texas M.D. Anderson Cancer Center, Houston, Texas, USA, [8]National Cancer Institute, National Institutes of Health, Bethesda, Maryland, USA, [9]Mayo Clinic College of Medicine, Rochester, Minnesota, USA, [10]Case Western Reserve University, Cleveland, Ohio, USA, [11]Johns Hopkins School of Medicine, Baltimore, Maryland, USA, [12]University of Washington, Seattle, Washington, USA, [13]Center for Craniofacial and Dental Genetics, University of Pittsburgh, Pittsburgh, Pennsylvania, USA, [14]State University of New York at Stony Brook, Stony Brook, New York, USA, [15]The Hospital for Sick Children, Toronto, Canada, [16]Washington University School of Medicine, St. Louis, Missouri, USA, [17]Centers for Disease Control, Atlanta, Georgia, USA, [18]University of Iowa, Iowa City, Iowa, USA, [19]Boston University, Boston, Massachusetts, USA, [20]Yale University, New Haven, Connecticut, USA and [21]Universitat de Luebeck, Luebeck, Germany

Email: Joan E Bailey-Wilson* - jebw@mail.nih.gov; Laura Almasy - almasy@darwin.sfbr.org; Mariza de Andrade - mandrade@mayo.edu; Julia Bailey - jbailey@npih.mednet.ucla.edu; Heike Bickeböller - hbickeb@gwdg.de; Heather J Cordell - heather.cordell@cimr.cam.ac.uk; E Warwick Daw - warwick@request.mdacc.tmc.edu; Lynn Goldin - goldinl@mail.nih.gov; Ellen L Goode - egoode@mayo.edu; Courtney Gray-McGuire - mcguire@darwin.epbi.cwru.edu; Wayne Hening - WAHeningMD@aol.com; Gail Jarvik - pair@u.washington.edu; Brion S Maher - brion@pitt.edu; Nancy Mendell - nmendell@notes.cc.sunysb.edu; Andrew D Paterson - andrew.paterson@utoronto.ca; John Rice - john@zork.wustl.edu; Glen Satten - gas0@cdc.gov; Brian Suarez - bks@themfs.wustl.edu; Veronica Vieland - veronica-vieland@uiowa.edu; Marsha Wilcox - mwilcox@bu.edu; Heping Zhang - heping.zhang@yale.edu; Andreas Ziegler - ziegler@imbs.uni-luebeck.de; Jean W MacCluer - jean@darwin.sfbr.org

* Corresponding author

Preface

This supplement to *BMC Genetics* contains the proceedings of the Genetic Analysis Workshop 14 (GAW14), which was held September 7–10, 2004, in Noordwijkerhout, The Netherlands. These workshops have been held since 1982 and now are held biennially. They serve as a forum for statisticians, epidemiologists, geneticists, and other scientists interested in these fields to introduce novel statistical methods and to evaluate and compare novel and existing methods. At each GAW, an existing dataset is selected, and a set of simulated data is devised such that statistical questions of wide and current interest may be addressed. These data are made available to scien-

tists worldwide who then report the results of their analyses of these data at the GAW meeting. GAW attendees must submit an analysis of one of these datasets, or be a workshop organizer or a dataset provider. The purpose of these workshops is to allow the comparison of statistical methodologies for genetic epidemiology using the same, well-described datasets. More information about GAW, including details of upcoming workshops, may be found at http://www.gaworkshop.org.

For GAW14, the overarching theme was comparison of microsatellite and single-nucleotide polymorphism (SNP) markers for genome-wide scans and the statistical

methods that can best exploit the information provided in such scans for linkage and association studies. Two datasets were available for GAW14 participants to analyze. As is traditional at GAWs, one of these was a simulated dataset and one consisted of data from an actual human study. Attempts were made in the data simulation to mirror many of the characteristics of the real dataset, including a map of microsatellite markers plus a denser map of SNP markers that were available for fine-mapping. Both datasets are discussed briefly below, and more detailed descriptions can be found in Edenberg et al. [1] and Greenberg et al. [2].

The Collaborative Study on the Genetics of Alcoholism (COGA) generously donated extensive family data on alcoholism, related phenotypes, pertinent covariates and a set of previously analyzed genome-scan microsatellite marker genotypes for use in GAW14. COGA is a nine-site national collaborative study funded by the National Institute on Alcohol Abuse and Alcoholism (NIAAA) and the National Institute on Drug Abuse (NIDA) with the primary goal of identifying and characterizing genes that affect the susceptibility to develop alcohol dependence and related phenotypes. COGA has been committed to sharing data with the scientific community to expedite progress in understanding alcoholism and related phenotypes. COGA also provided data to GAW11, and has created an archival database of these families, with both phenotypic data and immortalized cell lines; these data are accessible to investigators for further study through NIAAA http://www.niaaa.nih.gov/extramural/proj coga.htm. Genome-wide linkage scans, using microsatellite markers, have been performed on both an initial dataset of 105 multigenerational pedigrees and a replication dataset with 157 multigenerational pedigrees. In a departure from GAW tradition, extensive new SNP genotyping was performed on DNA provided by COGA for the previously genotyped families. Illumina, Affymetrix, and the Center for Inherited Disease Research (CIDR) performed this work and donated these data to GAW14 and to COGA. A subset of 143 informative families, selected from the initial and replication datasets, were genotyped for these SNP markers. A total of 1,350 COGA samples were genotyped by Affymetrix for 11,560 SNPs from their GeneChip Mapping 10 K array and by Illumina for their Linkage III Panel containing 4,763 SNP markers. These data are described in detail in Edenberg et al. [1].

The simulated data were designed to have similarities to the real dataset. It was assumed that a complex trait such as alcoholism might have both genetic and environmental risk factors. It was further assumed that such a complex trait might be defined/measured in a variety of different ways by different investigators, have associated endophenotypes that are common in the general population, and

is likely to be not one disease but a heterogeneous collection of clinically similar, but genetically distinct, entities. Disease characteristics and parameters were constant throughout all the simulations, but four different "studies" were simulated used varying ascertainment schemes based on differing assumptions about disease characteristics. One of the studies contained multiplex two and three generation pedigrees with at least four affected members. The simulated disease was a psychiatric condition with many associated behaviors (endophenotypes), almost all of which were genetic in origin. The underlying disease model contained four major genes and two modifier genes. The four major genes interacted with each other to produce three different "phenotypes", which were themselves heterogeneous. The population parameters were calibrated so that the major genes could be discovered by linkage analysis in most datasets. The association evidence was more difficult to calibrate but was designed to find statistically significant association in 50% of datasets. Some linkage disequilibrium between marker loci was simulated around some of the genes and also in areas without disease genes. Data distributed to participants contained about 1,000 SNPs and 400 microsatellite markers. Data obtainable via the internet consisted of a finer 10,000 SNP map, which also contained data on controls. These data are described in detail in Greenberg et al. [2].

In the spring of 2004, the availability of the GAW14 data was announced by e-mail to the more than 2,000 individuals on the GAW mailing list. A total of 129 groups requested GAW14 data. The COGA data were requested by 95 groups and the simulated data by 88 groups; 88 groups requested both datasets. In the summer of 2004, 183 contributed papers were received describing analyses of these datasets. A book and CD containing these contributions plus papers describing the datasets were distributed to workshop participants.

A total of 232 individuals from 14 countries attended GAW14. Attendees included investigators from four continents: Asia, Australia, Europe, and North America. The 183 contributions submitted to GAW14 were organized into 18 presentation groups of 7 to 15 papers each. The papers were grouped based on common methodological themes. Because the datasets were similar in many important aspects, and could be used to explore similar analytical problems, most presentation groups were assigned with no regard to the dataset analyzed. Thus, many of the groups of papers include analyses of both the simulated data and the COGA data. Within each presentation group, a co-author with previous GAW experience was asked to serve as group leader to facilitate group discussion, organize an oral presentation for the group, and take the lead in writing the group summary papers that will be published in *Genetic Epidemiology*. The 18 presentation groups were

organized around common methodological issues: comparisons of SNPs versus microsatellite markers in linkage studies, integration of SNPs and microsatellites in linkage and association studies, linkage mapping methods, quantitative trait mapping, fine mapping, methods for generating haplotypes and identifying haplotype-tagging SNPs, approaches to dealing with linkage disequilibrium, methods for association mapping, methods for case-control analysis and multivariate analyses, applications of these methods to analysis of alcoholism, smoking and related traits in the COGA data, methods for data mining, methods for dealing with genetic heterogeneity, methods for detecting gene × gene interaction, approaches to dealing with genotyping errors, pedigree errors and missing data, and methods to test for parent of origin, genomic imprinting, mitochondrial and X-linked effects in genome scans. Each presentation group met individually during the workshop and members of most groups communicated beforehand to begin comparing and contrasting the approaches taken and the results obtained by group members. At GAW14, many groups used part of their group meeting time to allow individual investigators to present their work. These group meetings were mostly attended by group participants but were open to all GAW14 attendees. From this process, each group developed an oral presentation, summarizing and synthesizing the work of the individual papers, which was delivered to the full workshop audience during the general sessions. Individual contributions were also presented in the form of 59 posters displayed during four poster sessions.

The 163 papers included in this supplement to *BMC Genetics* are a subset of the 183 contributions presented at GAW14. All of these papers have been reviewed for scientific merit. The proceedings begin with two papers describing the two datasets, followed by the 163 individual GAW14 contributions organized by presentation group and alphabetically by first author within each group. In addition to the individual papers in this volume, each presentation group has a summary in a forthcoming supplement to the journal *Genetic Epidemiology* in which the present manuscripts are compared and contrasted and the important themes and results from each group are described. Novel methods for linkage and association analyses using SNPs and microsatellites were developed, compared to existing methods, and applied to the GAW data, resulting in many interesting conclusions concerning appropriate approaches to the analysis of these sorts of data and also concerning areas of methodological development that are currently needed in this field.

Acknowledgements

References

1. Edenberg HJ, Bierut LJ, Boyce P, Cao M, Cawley S, Chiles R, Doheny KF, Hansen M, Hinrichs T, Jones K, Kelleher M, Kennedy GC, Liu G, Marcus G, McBride C, Murray SS, Oliphant A, Pettengill J, Porjesz B, Pugh EW, Rice JP, Rubano T, Shannon S, Steeke R, Tischfield JA, Tsai YY, Zhang C, Begleiter H: **Description of the data from the Collaborative Study on the Genetics of Alcoholism (COGA) and single-nucleotide polymorphism genotyping for Genetic Analysis Workshop 14.** *BMC Genet* 2005, **6(Suppl 1):**S2.
2. Greenberg DA, Zhang J, Shmulewitz D, Strug LJ, Zimmerman R, Singh V, Marathe S: **Construction of the model for the Genetic Analysis Workshop 14 simulated data: genotype-phenotype relationships, gene interaction, linkage, association, disequilibrium, and ascertainment effects for a complex phenotype.** *BMC Genet* 2005, **6(Suppl 1):**S3.