The Gene, Environment Association Studies Consortium (GENEVA): Maximizing the Knowledge Obtained from GWAS by Collaboration Across Studies of Multiple Conditions

Marilyn C. Cornelis,[1]* Arpana Agrawal,[2] John W. Cole,[3] Nadia N. Hansel,[4] Kathleen C. Barnes,[4] Terri H. Beaty,[5] Siiri N. Bennett,[6] Laura J. Bierut,[2] Eric Boerwinkle,[7] Kimberly F. Doheny,[8] Bjarke Feenstra,[9] Eleanor Feingold,[10] Myriam Fornage,[11] Christopher A. Haiman,[12] Emily L. Harris,[13] M. Geoffrey Hayes,[14] John A. Heit,[15] Frank B. Hu,[1] Jae H. Kang,[16] Cathy C. Laurie,[6] Hua Ling,[8] Teri A. Manolio,[17] Mary L. Marazita,[18] Rasika A. Mathias,[4] Daniel B. Mirel,[19] Justin Paschall,[20] Louis R. Pasquale,[16] Elizabeth W. Pugh,[8] John P. Rice,[2] Jenna Udren,[6] Rob M. van Dam,[1] Xiaojing Wang,[18] Janey L. Wiggs,[16] Kayleen Williams,[6] and Kai Yu[21] for the GENEVA Consortium

[1]*Harvard School of Public Health, Boston, Massachusetts*
[2]*Department of Psychiatry, Washington University School of Medicine, Saint Louis, Missouri*
[3]*School of Medicine, University of Maryland, Baltimore, Maryland*
[4]*Johns Hopkins University School of Medicine, Baltimore, Maryland*
[5]*Bloomberg School of Public Health, Johns Hopkins University, Baltimore, Maryland*
[6]*Collaborative Health Studies Coordinating Center, University of Washington, Seattle, Washington*
[7]*Human Genetics Center, University of Texas Health Science Center at Houston, Houston, Texas*
[8]*Center for Inherited Disease Research, Johns Hopkins University School of Medicine, Baltimore, Maryland*
[9]*Department of Epidemiology Research, Statens Serum Institut, Copenhagen, Denmark*
[10]*Department of Human Genetics, University of Pittsburgh, Pittsburgh, Pennsylvania*
[11]*Institute of Molecular Medicine, University of Texas, Houston, Texas*
[12]*Keck School of Medicine, University of South California, Los Angeles, California*
[13]*National Institute of Dental and Craniofacial Research, US National Institutes of Health (NIH), Bethesda, Maryland*
[14]*Feinberg School of Medicine, Northwestern University, Chicago, Illinois*
[15]*Division of Hematology, Mayo Clinic, Rochester, Minnesota*
[16]*Harvard Medical School, Boston, Massachusetts*
[17]*National Human Genome Research Institute, NIH, Bethesda, Maryland*
[18]*Department of Oral Biology, University of Pittsburgh, Pittsburgh, Pennsylvania*
[19]*Broad Institute of MIT and Harvard, Boston, Massachusetts*
[20]*National Center for Biotechnology Information, NIH, Bethesda, Maryland*
[21]*National Cancer Institute, NIH, Bethesda, Maryland*

Genome-wide association studies (GWAS) have emerged as powerful means for identifying genetic loci related to complex diseases. However, the role of environment and its potential to interact with key loci has not been adequately addressed in most GWAS. Networks of collaborative studies involving different study populations and multiple phenotypes provide a powerful approach for addressing the challenges in analysis and interpretation shared across studies. The Gene, Environment Association Studies (GENEVA) consortium was initiated to: identify genetic variants related to complex diseases; identify variations in gene-trait associations related to environmental exposures; and ensure rapid sharing of data through the database of Genotypes and Phenotypes. GENEVA consists of several academic institutions, including a coordinating center, two genotyping centers and 14 independently designed studies of various phenotypes, as well as several Institutes and Centers of the National Institutes of Health led by the National Human Genome Research Institute. Minimum detectable effect sizes include relative risks ranging from 1.24 to 1.57 and proportions of variance explained ranging from 0.0097 to 0.02. Given the large number of research participants ($N > 80{,}000$), an important feature of GENEVA is harmonization of common variables, which allow analyses of additional traits. Environmental exposure information available from most studies also enables testing of gene-environment interactions. Facilitated by its sizeable infrastructure for promoting collaboration, GENEVA has established a unified framework for genotyping, data quality control, analysis and interpretation. By maximizing knowledge obtained through collaborative GWAS incorporating environmental exposure information, GENEVA aims to enhance our understanding of disease etiology, potentially identifying opportunities for intervention. *Genet. Epidemiol.* 2010. © 2010 Wiley-Liss, Inc.

Key words: genome-wide association; complex disease; quantitative traits; gene-environment interaction; phenotype harmonization

Additional Supporting Information may be found in the online version of this article.
Contract grant sponsor: NIH Genes, Environment and Health Initiative (GEI); Contract grant numbers: U01HG04424; U01HG004438; Contract grant sponsor: NIDCR; Contract grant number: U01DE018993; Contract grant sponsor: NIH; Contract grant numbers: HHSN268200782096C; U01HG004422; HHSN268200782096C; U01HG004402; HHSN268200625226C; Z01CP010200; U01HG004729; U01HG004399; U01HG004726; U01HG004735; U01HG004415; HHSN268200782096C; U01HG004436; U01HG004423; U01HG004728; HHSN268200782096C; Contract grant sponsor: NIAAA; Contract grant number: U10AA008401; Contract grant sponsor: NIDA; Contract grant numbers: P01CA089392; R01DA013423; Contract grant sponsor: NHLBI; Contract grant numbers: N01HC55015; N01HC55016; N01HC55018; N01HC55019; N01HC-55020; N01HC-55021; N01HC55022; R01HL087641; R01HL59367; R01HL086694; Contract grant sponsor: NIH Roadmap for Medical Research; Contract grant number: UL1RR025005; Contract grant sponsor: NCI; Contract grant numbers: CA63464; CA54281; CA136792; Contract grant sponsor: NIDCR; Contract grant number: U01DE018903; Contract grant sponsor: NEI; Contract grant numbers: R01EY015473; R01EY015872; Contract grant sponsors: National Library of Medicine; Canadian Institutes of Health Research Fellowship; National Center for Research Resources (NCRR).

*Correspondence to: Marilyn C. Cornelis, Department of Nutrition, Harvard School of Public Health, 665 Huntington Avenue, Boston, MA 02115. E-mail: mcorneli@hsph.harvard.edu

Received 17 September 2009; Revised 26 November 2009; Accepted 14 December 2009
Published online in Wiley InterScience (www.interscience.wiley.com).
DOI: 10.1002/gepi.20492

INTRODUCTION

Genome-wide association studies (GWAS) have emerged as powerful approaches for identifying genetic variants influencing common, complex diseases and traits [Hunter et al., 2007; Sladek et al., 2007; Wellcome Trust Case Control Consortium, 2007; Yeager et al., 2007]. Most genetic loci discovered to date, however, account for only a small fraction of total phenotypic variation and most of the inherited component of risk remains unexplained. Some of this missing inherited risk, i.e. that proportion not attributable to variants identified to date, might be due to gene-environment (G × E) interactions that, when present, may adversely affect the ability to uncover risk loci [McCarthy and Hirschhorn, 2008]. Nearly all GWAS to date have concentrated on detecting and characterizing main effects and have not fully explored the potential role environmental factors play in modifying genetic risk [Clayton and McKeigue, 2001; Dempfle et al., 2008; Martinez, 2008]. Whether, and to what extent, the GWAS approach can be used to uncover these potential G × E interactions remains uncertain.

The formation of multiple consortia and collaborations has been crucial for success of the GWAS approach by increasing sample sizes, thereby increasing statistical power, enabling replication of findings from individual studies and establishing common methods of analysis [Manolio et al., 2007; Wellcome Trust Case Control Consortium, 2007]. In 2006, the United States Secretary of Health and Human Services initiated a NIH-wide program, the Genes, Environment and Health Initiative (GEI, http://www.gei.nih.gov/genetics/index.asp) which aims to accelerate understanding of genetic and environmental contributions to health and disease. There are two components to GEI: genetics and exposure biology. The genetics program includes a consortium for GWAS, as well as replication and fine-mapping studies, sequencing studies, functional studies, development of analytical methods and databases, and pilot clinical translation studies. The GWAS component, named the Gene, Environment Association Studies (GENEVA) consortium, was initiated in 2006 as a result of a series of requests for applications (RFAs) to support the establishment and work of a coordinating center (CC), genotyping centers (GCs) and study investigators (SI). The goals of the consortium are to (i) identify genetic variants associated with complex diseases and traits in initial genome-wide discovery studies; (ii) identify variations in gene-trait associations related to environmental exposures; and (iii) ensure the rapid sharing of data to the general scientific community.

Herein we describe the GENEVA consortium. We begin by outlining the organizational structure of GENEVA, including the global study management, GCs, CC and individual studies. We subsequently describe the operations of the consortium, including development of subcommittees and working groups designed to address the overall aims of this consortium. Finally, we discuss the potential contributions of GENEVA within the greater realm of genetic epidemiology, including an integrative, collaborative process for optimizing study methods.

ORGANIZATIONAL STRUCTURE OF GENEVA

The GENEVA Consortium consists of several NIH-based organizations and extramural participants. Key NIH participants include the Office of Population Genomics at the National Human Genome Research Institute (NHGRI), which directs the GENEVA program; National Institute of Dental and Craniofacial Research (NIDCR), which funds two of the GWAS; Program Officials from disease-relevant Institutes/Centers; and the National Center for Biotechnology Information (NCBI), which manages the database of Genotypes and Phenotypes (dbGaP), the data repository for GENEVA and other GWAS. The key academic participants include the individual studies and their investigators, the GCs, and the CC.

STUDY MANAGEMENT

Management of GENEVA is coordinated through NHGRI, NCBI, and the GENEVA Steering Committee. The GENEVA Program Official at the NHGRI Office of Population Genomics facilitates achievement of scientific goals and provides institutional oversight and guidance to the consortium. The GENEVA Steering Committee is composed of the Principal Investigators from the specific studies, the CC, the GCs, and the NHGRI Project Scientist. An External Scientific Panel (ESP), composed of senior scientists with expertise in G × E interactions and genome-wide association research, was established by NHGRI to provide scientific insight into the overall direction of GENEVA and advice on specific design issues. While not actively involved in GENEVA operations, the ESP advises the Steering Committee and NHGRI on the scientific directions of GENEVA, the soundness of its methods and approaches and, when necessary, potential alternative strategies.

PARTICIPATING STUDIES AND INVESTIGATORS

Table I provides an overview of the 14 participating studies. Studies are predominately case-control by design with variable sampling schemes and cover a wide spectrum of complex qualitative and quantitative phenotypes. Each study has quality phenotype and environmental exposure data available as a result of past funding opportunities. While some phenotypes have been represented in other primary studies or GWAS consortia, others such as the Oral Clefts, Dental Caries, Birth Weight and Premature Birth studies constitute the first and/or largest known GWAS of their trait to date. All phenotypes have important public health significance (e.g. high prevalence rates, potential treatment/management opportunities) and evidence for both a genetic and environmental component. Non-substance-related psychiatric disorders, breast and ovarian cancer are notable absentees, which is largely a consequence of their non-representative response to the RFA or incompatibility with the GEI guidelines. Nevertheless, some of these have already received considerable attention through previous initiatives [Hunter et al., 2007; Manolio et al., 2007]. Most studies have also proposed secondary phenotypes that either differ from or complement their primary outcome of interest. For example, the Lung Cancer, Lung Health and Alcohol Dependence studies all have plans to investigate smoking behavior and additional smoking-related phenotypes.

Although most GENEVA studies include persons of European-ancestry, the Alcohol Dependence, Coronary Heart Disease, Prostate Cancer, Birth weight/Maternal Glycemia, Premature Birth and Ischemic Stroke studies have large proportions of subjects of African descent. A significant number of Hispanics and/or Asians are also included in the Oral Clefts and Prostate Cancer studies. These study samples will enable sufficiently powered investigations of non-European-ancestral populations who have been under-represented in published GWAS.

Despite diverse phenotypic outcomes, designs and populations, all GENEVA studies have the common interest of incorporating environmental factors into their analyses, consistent with the goals of GEI. Indeed, each is well suited for these analyses, given their collection of extensive environmental exposure information. Table I provides a *sample* of variables available for each study that will also be provided through the controlled access process of dbGaP.

GENOTYPING CENTERS

The Center for Inherited Disease Research (CIDR) at Johns Hopkins University and the Broad Institute of Massachusetts Institute of Technology and Harvard University were selected as the two GCs. The GCs use highly trained staff, standardized protocols, robotics and integrated laboratory information management systems with in-house quality control (QC) assessments to provide cost-efficient, high-throughput, high quality genotyping capability.

COORDINATING CENTER

The Collaborative Health Studies Coordinating Center (CHSCC, Department of Biostatistics, University of Washington,) assists with genotype data cleaning, phenotype data organization and coordination of logistics and administration of the consortium. It also serves as an internal data repository for GENEVA and assists in cross-study phenotypic data harmonization. With guidance from the Steering Committee, NHGRI and the ESP, the CC provides leadership and management for administrative and scientific data management matters. While the SI are responsible for their own data analysis, the CC also provides statistical advice as needed.

COMMITTEES AND WORKING GROUPS

The GENEVA study structure promotes collaborative efforts and ongoing interactions among all participating studies. Since the initiation of GENEVA, various subcommittees and working groups have been established to address specific issues related to analysis, genotyping QC and assurance, phenotype harmonization, cross-study integration and other challenges inherent to collaborative studies. Through monthly teleconferences and in-person Steering Committee meetings held three times per year, this interacting network of teams has been crucial in addressing the consortium's aims.

ADDRESSING THE AIMS OF GENEVA

IDENTIFY GENETIC VARIANTS RELATED TO COMMON, COMPLEX DISEASES AND TRAITS

Since GENEVA studies a wide range of complex traits utilizing various study designs, the power to detect genetic effects will vary substantially. Given study-specific parameters (i.e. study design, sample size and baseline risk) and assuming a minor allele frequency of 0.3, all studies have 80% power to detect additive variant relative risks of >1.57 and proportions of variance explained for primary quantitative traits ranging from 0.0097 to 0.02 (Table I). Combining data on common outcomes and environmental measures across studies will also allow tests with greater power for even modest effect sizes. The availability of phenotype data common to multiple studies also provides a platform for exploring other, potentially novel, gene-trait associations. For example, anthropometric measures, such as height, weight and body mass index (BMI), are uniformly available across a majority of the studies and we anticipate genome-wide scan data for ~40,000 subjects for cross-study analysis of these traits. Assuming an effect allele frequency of 0.3, we are sufficiently powered to detect a marginal correlation coefficient of at least 0.0011 when BMI is the outcome of interest. Investigators are also looking across studies at smoking and alcohol consumption behavior, female reproductive history and oral health. GENEVA investigators also plan to assess genetic loci associated with novel traits such as caffeine consumption, physical activity and "wellness" (i.e. protection against disease). As one of the first GWAS of caffeine intake,

TABLE I. Characteristics of participating studies

PI	Institution	Primary outcome	Population	Study design	Samples	Platform	OR_G[a]	Key Environmental Exposures[b]
Phase 1 study investigators								
T. Beaty	Johns Hopkins U	Oral clefts	European American, European, Asian	Case-parent trios	2,664 cases 5,328 parents	Illumina 610-Quad	1.31	smoking alcohol prenatal vitamins
L. J. Bierut	Washington U	Alcohol/nicotine dependence	European American, African American	Case-control	2,012 cases 2,011 controls	Illumina 1M	1.37	age of onset of substance use traumatic experiences religiosity
E. Boerwinkle	U of Texas	Coronary heart disease	European American, African American	Cohort	2,285 cases 13,360 non-cases	Affy 6.0	1.25	smoking diet physical activity obesity
N. Caporaso	National Cancer Inst	Lung cancer	European, European American	Case-control	2,848 cases 2,915 controls	Illumina 550-Duo	1.30	smoking
F. B. Hu	Harvard U	Type 2 diabetes	European American, African American, Asian American	Nested case-control	2,891 cases 4,337 controls	Affy 6.0	1.27	obesity physical activity smoking alcohol diet
W. Lowe	Northwestern U	Birth weight/maternal glycemia	European, European American, Afro Caribbean, Mexican Hispanic	Mother-child pairs	3,550 mother 3,550 children	Illumina 610-Quad Illumina 1M	0.012 (R^2)	maternal uterine environment (various)
M. L. Marazita	U of Pittsburgh	Dental caries	European American	Cohort of families and unrelated individuals	4,073	Illumina 610-Quad	1.40	fluoride diet *S. mutans* family/behavioral traits
J. Murray	U of Iowa	Premature birth	European, African American	Mother-child pairs	6,300	Illumina 660W-Quad/Omni1-Quad	1.55	smoking prenatal vitamin alcohol

Phase 2 study investigators								
K. C. Barnes	Johns Hopkins U	Lung function decline in chronic obstructive pulmonary disease	European American	cohort	4,287	Illumina 660W Quad	0.0097 (R^2)	smoking
M. Fornage	U of Texas	Longitudinal blood pressure profiles	European American	Cohort	2,064	Affy 6.0	0.020 (R^2)	physical activity psychosocial factors smoking obesity
C. A. Haiman	U of Southern California	Prostate cancer	African American, Latino American, Japanese American	Nested case-control	4,400 cases 4,400 controls	Illumina 660W-Quad	1.24	smoking obesity physical activity alcohol diet
J. A. Heit	Mayo Clinic	Venous thrombosis	European American	Population-based case-control	1,300 cases 1,300 controls	Illumina 660W-Quad	1.57	obesity smoking major surgery hospitalization for acute medical illness trauma/fracture leg paresis exogenous hormone exposure pregnancy or postpartum stroke myocardial infarction prior cancer by cancer type
B. Mitchell.	U of Maryland	Ischemic stroke	European American, African American	Population-based case-control	920 cases 942 controls	Illumina Omni1-Quad	1.56	smoking migraine physical activity oral contraceptive-use
L. R. Pasquale	Harvard Medical School	Primary open–angle glaucoma	European American	Case-control	1,200 cases 1,200 controls	Illumina 610/660W-Quad	1.49	alcohol caffeine

[a] Multiplicative allelic relative risks detectable (or proportion of variation explained: R^2) with 80% power assuming 30% risk allele frequency, 0.1–30% (study specific) disease prevalence and type 1 error of 1E-08.

[b] Environmental exposure data to be deposited on dbGaP.

mega-analysis of data on 22,000 genome-wide scans will afford 80% power to detect additive genetic variants that explain marginal effects as small as 0.0019 while satisfying a type 1 error level of 1E-08.

The Analysis Subcommittee was formed to provide expert advice on shared analysis issues, such as the development of methods for within-study analysis for studies with significant ethnic variation, related individuals and/or longitudinal data. With the different genotyping platforms utilized and unique characteristics of each study, the Genotyping Subcommittee was established to streamline submission of samples for genotyping, establish standards for QC and serve as a liaison with the Analysis Subcommittee to tackle novel concerns arising from data-cleaning efforts on the genome-wide marker panels. To facilitate effective collaborations both within and outside GENEVA, the Imputation Working Group addresses methods of imputation, including choice of reference panel and how these imputed data should be distributed and analyzed.

In light of the opportunities for cross-study analysis of common traits, the Phenotype Harmonization Subcommittee identifies phenotypic measures of interest that are amenable to cross-study harmonization. The subcommittee formulates and implements strategies for successful meta-analysis and pooled analysis of individual participant data. A working group for each shared phenotype consists of representatives from each study contributing data as well as the CC and NIH. Challenges specific to cross-study analyses that need to be addressed include accounting for differences in population structure, study design, and environmental exposure and genotype assessment. Combining cohorts from different countries, or from different sites within the same country, will require investigating and addressing the problem of confounding due to population stratification [Campbell et al., 2005; Helgason et al., 2005; Seldin and Price, 2008]. Likewise, analyses that rely upon a common pool of controls, where the outcome or environment exposure of interest may not be universally available, must also be performed with considerable caution [Wellcome Trust Case Control Consortium, 2007]. A working group is actively pursuing the use of GENEVA samples as controls for genetic matching and will provide measurable insight on the impact this approach has on risk loci discovery.

Finally, the Cross-Study Integration Subcommittee was established to develop recommendations regarding efficient and streamlined cross-study data analysis, sharing data within GENEVA, and collaborating with other projects or consortia outside of GENEVA. The subcommittee develops recommendations for study-wide guidelines for issues such as disclosing individual-level findings that may be clinically significant, and for standardized publications policy for authorship and management of meta-analyses.

IDENTIFYING DIFFERENCES IN GENE-TRAIT ASSOCIATIONS RELATED TO ENVIRONMENTAL EXPOSURES

Most participating studies in GENEVA have collected extensive measures of environmental exposures and therefore have the opportunity to address the second aim of the consortium, which is to identify variations in gene-trait associations related to environmental exposures. Successfully meeting this aim could ultimately distinguish population subgroups potentially susceptible to the protective or adverse effects of these environmental exposures. Accounting for these G × E interactions might also improve our ability to identify additional risk loci.

SI have selected environmental exposures relevant to their primary outcome to be utilized in tests for G × E interactions (Table I). Designing a sufficiently powered study and locating an appropriate external study for replication are just two examples of major barriers to uncovering true interactions. When applying the standard logistic regression test for interaction, most individual studies will be limited to detecting interactions of large effect sizes. Nevertheless, new methods for G × E interaction testing have been and will continue to be developed to boost statistical power for detection while maintaining low type 1 error [Chatterjee and Carroll, 2005; Kraft et al., 2007; Murcray et al., 2009; Weinberg, 2009]. Methods based on logistic regression continue to dominate the field and generally test for interactions *specifically*, or main genetic associations allowing for heterogeneity in genetic effect across environment strata. Model-free or machine learning approaches in the context of GWAS are relatively new and currently computationally expensive. The performance of each method will vary with the distributional assumptions underlying the phenotypic outcome, the environment and their suspected interaction. The Analysis Subcommittee considers each approach and its strengths, limitations and feasibility for a particular scenario, and advises SI on the most appropriate method for their G × E interaction of interest. Each SI is responsible for data analysis and plans to replicate initial findings as outlined in their response to the RFA. G × E interactions will also be investigated in cross-study trait analysis; some of which are sufficiently powered even when applying a conservative test for interaction. For example, gene-smoking interactions for both BMI and caffeine consumption are highly anticipated and we will have 80% power to detect marginal R^2 for interaction effects as modest as 0.0013 and 0.002, respectively.

Thus far, little is known on how the traditional and recently proposed methods for testing G × E interactions perform in the context of GWAS and whether they can be applied to meta- or cross-study analysis for discovery purposes. The latter is especially important, because in order to achieve the sample sizes required to detect small to modest interaction effect sizes, a cross-study collaborative approach may be the only option. Unlike other consortia, GENEVA is well positioned to apply these methods and in doing so we will finally have a better measure of their performance. Caveats to their application and interpretation might also be uncovered which, in turn, will aid in further method development and optimization.

ENSURING THE RAPID SHARING OF DATA TO THE SCIENTIFIC COMMUNITY

To accelerate and facilitate the discovery of genetic variants related to health and disease, genotype (SNP calls), phenotype and exposure data from each of these studies will be shared with the scientific community through dbGaP's controlled access process when data cleaning is complete [Mailman et al., 2007]. Raw intensity data will also be made available to enable approved users to apply alternative genotype calling algorithms or for other method development purposes. Final data files from each study as well as supporting documents and data dictionaries are organized by the CC and are sent to NCBI, where they are deposited in dbGaP (http://www.ncbi.nlm.nih.gov/gap). A 1-year protected period for dissemination allows GENEVA investigators to analyze the data and report study results. During this period, individual-level and summary genotype data in dbGaP are available to authorized researchers outside of GENEVA, but they agree not to submit publications or make presentations using the data. To date, four of the GENEVA studies have genetic and phenotypic data publicly available on dbGaP. In addition to over 78 authorized data requests for *independent* analysis, over 35 studies and consortia have proposed *collaborations* with GENEVA investigators.

GENEVA'S POTENTIAL CONTRIBUTIONS TO ACCELERATING DEVELOPMENTS IN GENETIC EPIDEMIOLOGY

In parallel with meeting the aims outlined above, GENEVA will likely have an important impact on the design and conduct of future GWAS as well as the broader field of genetic epidemiology.

GENOTYPING QC AND ASSURANCE

GENEVA has formulated a precise work flow with accompanying protocols to effectively manage the extensive and diverse phenotypic and genotypic data from across studies. The overall flow of data from the SI to the GC, to NCBI and the CC, and then finally to data release on dbGaP is outlined in Supplementary Figure.

To minimize bias and spurious associations that may occur with using combined data from different studies and conducting a large number of statistical tests, GENEVA has built upon previous efforts [Chanock et al., 2007; Miyagawa et al., 2008] and provides an extensive guide for QC and quality assurance (QA) for users [Laurie et al., 2009; submitted]. The consortium has developed new approaches to (1) distinguish gender misidentification from sex chromosome aberrations, (2) detect autosomal chromosome aberrations that may affect genotype calling accuracy, (3) measure DNA quality, (4) infer relatedness through identity-by-descent estimates and (5) use duplicate concordance to filter SNP quality. Genotypic data are distributed to the entire project team for quality assessment, which occurs as a collaborative process led by the CC and involving the appropriate SI team, GCs, NHGRI, NCBI, and any interested GENEVA investigators or NIH staff who wish to listen in.

Given the diverse structures of the studies, data quality standards are decided for each study as part of the collaborative QC process. The CC prepares a detailed report regarding the outcomes of each measure. These reports are provided through the controlled access process of dbGaP, along with the unfiltered data set, a set of filters and a tool to apply the filters to create a filtered data set. Thus far, GENEVA has focused on SNPs, but the GCs have plans to implement QC/QA metrics for CNVs once common CNV maps are established and detection methods are more standardized.

IMPUTATION

The high-quality genotyping data produced by the QC/QA process will undoubtedly contribute to SNP imputation accuracy, which will be essential for successful cross-study integration. The Imputation Working Group is leading GENEVA efforts to impute all data in a uniform manner despite the differences in study designs, genotyping platforms and population structures. Choice of imputation software, HapMap build and population reference panel, available computational resources, and methods for incorporating quality scores and other metrics of accuracy and efficiency are among the many factors to be addressed.

CROSS-STUDY ANALYSES

GENEVA's proposed cross-study GWAS mirror the meta-analytical approach but with additional challenges anticipated. Some traits, including habitual alcohol and caffeine consumption, physical activity and sleeping behavior are difficult to define with many external factors influencing their measurement. Moreover, very little is known regarding the properties of the discovery process in cross-study analyses of GWAS-derived signals, especially for complex traits. Such approaches are susceptible to the same issues as in single studies pursuing agnostic associations, but have additional caveats to attend to; between-study heterogeneity being a particularly important one [Pereira et al., 2009].

Some heterogeneity in cross-study results is anticipated and it may be attributable to biases in the collection of exposure data, phenotype definition, participant selection, population structure, and various elements of the genotyping process [Ioannidis et al., 2007; Nakaoka and Inoue, 2009]. GENEVA's refined genotyping QC protocol should safeguard against some of these biases, but those pertaining to cross-study differences in study design will require special attention and are, therefore, addressed by individual phenotype harmonization working groups. Heterogeneity may also reflect genuine differences such as LD structure or environmental exposure diversity across populations [Nakaoka and Inoue, 2009]. The former may assist in pinpointing the causal variant and the latter may lead to hypothesis generation, complementing those already proposed by GENEVA investigators and those that might be pursued either within or outside the consortium. Thus, despite the challenges GENEVA anticipates through cross-study analyses, results should generate a new insight into the gene-trait association.

DATA MANAGEMENT

Advances in molecular biology have led to an astounding growth in information generated by the scientific community [Barnes and Gray, 2003]. This has intensified the need for efficient access to and management of large data sets to maximize their utility. The CC has currently implemented the use of the Network Common Data Form (netCDF) interface that allows one to create, access and share array-oriented data in a self-describing and por table form (http://www.unidata.ucar.edu/software/netcdf). As genotyping data accrues in GENEVA sophisticated bioinformatics tools for data management and knowledge expansion will be vital for integration with other components of the GEI Genetics Program such as development of relational databases to combine information on SNP annotation, putative function and biological pathways. These will complement and provide necessary support for novel loci uncovered in genetic association studies.

SUMMARY

Nearly all GWAS to date have concentrated on detecting and characterizing main effects of genes and have underemphasized the potential role the environment plays in modifying genetic risk [Clayton and McKeigue, 2001; Dempfle et al., 2008; Martinez, 2008]. This decreased attention may, in part, be due to the paucity of established methods for the study of G × E interactions in a GWAS context. GWAS present many common challenges in analysis and interpretation that are likely to have common solutions [Manolio et al., 2007; Wellcome Trust Case Control Consortium, 2007]. These solutions and the potential for combining phenotype and genotype data across studies to enhance statistical power are best developed through collaborative approaches, as demonstrated by the Wellcome Trust Case Control Consortium (WTCCC), Genetic Association Information Network (GAIN) and the Psychiatric Genetics Consortium [Manolio et al., 2007; Psychiatric GWAS Consortium Steering Committee, 2009; Wellcome Trust Case Control Consortium, 2007].

GENEVA is one of a handful of collaborative GWA programs that involve many different diseases and traits, rather than focusing on a single disease or related traits such as the Myocardial Infarction Genetics (MIGen), Diabetes Genetics Replication and Meta-analysis (DIAGRAM), or Tobacco and Genetics (TAG) consortia. Others we are aware of include GAIN and WTCCC; models upon which GENEVA has built and expanded to include larger numbers of secondary phenotypes and greater harmonization of these phenotypes across studies. The well-developed infrastructure of the GENEVA consortium, as well as its collection of studies with extensive environmental exposure data, enhances the benefit of collaborative work to further maximize knowledge obtainable through GWAS. Indeed, the goal of focusing on the combined role of genetics and environment will aid in development and application of new analytic methods to consider G × E interaction at a genome-wide level. GENEVA's initial efforts will focus on SNP analysis, yet it is actively pursuing the role of other forms of genetic variation, including CNVs. Moreover, sharing GENEVA's growing repository of data with the broader scientific community should accelerate identification of variants related to complex diseases and identify opportunities for developing effective interventions. In parallel to meeting the aims of the consortium, GENEVA is intended to provide and broadly disseminate analytical and bioinformatic approaches for use in the design and conduct of future GWAS.

Taken together, GENEVA's efforts, in conjunction with replication, fine mapping, sequencing, and functional studies as well as database development, and clinical translation studies, will undoubtedly enhance our understanding of disease etiology and identify opportunities for treatment and prevention.

ACKNOWLEDGMENTS

Funding support for the GENEVA genome-wide association studies was provided through the NIH Genes, Environment and Health Initiative (GEI). Some studies also received support from individual NIH Institutes. Beaty (NIDCR: U01DE018993, NIH contract: HHSN268200782096C); Bierut (U01HG004422, NIAAA: U10AA008401, NIDA: P01CA089392, R01DA013423, NIH contract: HHSN268200-782096C); Boerwinkle (U01HG004402, NHLBI: N01HC55015, N01HC55016, N01HC55018, N01HC55019, N01HC-55020, N01HC-55021, N01HC55022, R01HL087641, R01HL59367, R01HL086694, NIH contract: HHSN268200625226C, NIH Roadmap for Medical Research: UL1RR025005); Caporaso (Z01CP010200); Fornage (U01HG004729); Hu (U01HG004399); Haiman (U01HG004726, NCI: CA63464, CA54281, CA136792); Heit (U01HG004735); Lowe (U01HG004415); Marazita (NIDCR:U01DE018903, NIH contract: HHSN268200-782096C); Mitchell (U01HG004436); Murray (U01HG004423); Pasquale [U01HG004728, NEI: R01EY015473, NEI: R01EY015872 (J. L. Wiggs)]. Assistance with phenotype harmonization and genotype cleaning, as well as with general study coordination, was provided by the GENEVA Coordinating Center (U01HG004446). Assistance with data cleaning was provided by the National Center for Biotechnology Information. Genotyping was performed at the Broad Institute of MIT and Harvard, with funding support from the NIH GEI (U01HG04424), and Johns Hopkins University Center for Inherited Disease Research, with support from the NIH GEI (U01HG004438) and the NIH contract "High throughput genotyping for studying the genetic contributions to human disease" (HHSN268-200782096C). Funding support was also provided, in part, by the Intramural Research Program of the NIH, National Library of Medicine. M. C. Cornelis is a recipient of a Canadian Institutes of Health Research Fellowship. A. Agrawal is supported by grants from the NIDA. J. W. Cole is supported by the Office of Research and Development, Medical Research Service and the Research Enhancement Award Program in Stroke, Department of Veterans Affairs; the American Heart Association, and the General Clinical Research Centers Program, National Center for Research Resources (NCRR), NIH. K. C. Barnes is supported in part by the Mary Beryl Patch Turnbull Scholar Program. L. J Bierut and J. P. Rice are listed as inventors on a patent (US 20070258898) held by Perlegen Sciences, Inc., covering the use of certain SNPs in determining the diagnosis, prognosis, and treatment of addiction. L. J. Bierut has acted as a consultant for Pfizer, Inc. in 2008.

The GENEVA consortium thanks the staff and participants of all GENEVA studies for their important contributions.

REFERENCES

Barnes MR, Gray IC. 2003. Bioinformatics for Geneticists. New Jersey: Wiley.

Campbell CD, Ogburn EL, Lunetta KL, Lyon HN, Freedman ML, Groop LC, Altshuler D, Ardlie KG, Hirschhorn JN. 2005. Demonstrating stratification in a European American population. Nat Genet 37:868–872.

Chanock SJ, Manolio T, Boehnke M, Boerwinkle E, Hunter DJ, Thomas G, Hirschhorn JN, Abecasis G, Altshuler D, Bailey-Wilson JE, Brooks LD, Cardon LR, Daly M, Donnelly P, Fraumeni Jr JF, Freimer NB, Gerhard DS, Gunter C, Guttmacher AE, Guyer MS, Harris EL, Hoh J, Hoover R, Kong CA, Merikangas KR, Morton CC, Palmer LJ, Phimister EG, Rice JP, Roberts J, Rotimi C, Tucker MA, Vogan KJ, Wacholder S, Wijsman EM, Winn DM, Collins FS. 2007. Replicating genotype-phenotype associations. Nature 447:655–660.

Chatterjee N, Carroll RJ. 2005. Semiparametric maximum likelihood estimation exploiting gene-environment independence in case-control studies. Biometrika 92:399–418.

Clayton D, McKeigue PM. 2001. Epidemiological methods for studying genes and environmental factors in complex diseases. Lancet 358:1356–1360.

Dempfle A, Scherag A, Hein R, Beckmann L, Chang-Claude J, Schafer H. 2008. Gene-environment interactions for complex traits: definitions, methodological requirements and challenges. Eur J Hum Genet 16:1164–1172.

Helgason A, Yngvadottir B, Hrafnkelsson B, Gulcher J, Stefansson K. 2005. An Icelandic example of the impact of population structure on association studies. Nat Genet 37:90–95.

Hunter DJ, Kraft P, Jacobs KB, Cox DG, Yeager M, Hankinson SE, Wacholder S, Wang Z, Welch R, Hutchinson A, Wang J, Yu K, Chatterjee N, Orr N, Willett WC, Colditz GA, Ziegler RG, Berg CD, Buys SS, McCarty CA, Feigelson HS, Calle EE, Thun MJ, Hayes RB, Tucker M, Gerhard DS, Fraumeni Jr JF, Hoover RN, Thomas G, Chanock SJ. 2007. A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer. Nat Genet 39:870–874.

Ioannidis JP, Patsopoulos NA, Evangelou E. 2007. Heterogeneity in meta-analyses of genome-wide association investigations. PLoS One 2:e841.

Kraft P, Yen YC, Stram DO, Morrison J, Gauderman WJ. 2007. Exploiting gene-environment interaction to detect genetic associations. Hum Hered 63:111–119.

Mailman MD, Feolo M, Jin Y, Kimura M, Tryka K, Bagoutdinov R, Hao L, Kiang A, Paschall J, Phan L, Popova N, Pretel S, Ziyabari L, Lee M, Shao Y, Wang ZY, Sirotkin K, Ward M, Kholodov M, Zbicz K, Beck J, Kimelman M, Shevelev S, Preuss D, Yaschenko E, Graeff A, Ostell J, Sherry ST. 2007. The NCBI dbGaP database of genotypes and phenotypes. Nat Genet 39:1181–1186.

Manolio TA, Rodriguez LL, Brooks L, Abecasis G, Ballinger D, Daly M, Donnelly P, Faraone SV, Frazer K, Gabriel S, Gejman P, Guttmacher A, Harris EL, Insel T, Kelsoe JR, Lander E, McCowin N, Mailman MD, Nabel E, Ostell J, Pugh E, Sherry S, Sullivan PF, Thompson JF, Warram J, Wholley D, Milos PM, Collins FS. 2007. New models of collaboration in genome-wide association studies: the Genetic Association Information Network. Nat Genet 39:1045–1051.

Martinez FD. 2008. Gene-environment interaction in complex diseases: asthma as an illustrative case. Novartis Found Symp 293:184–192; discussion 192–197.

McCarthy MI, Hirschhorn JN. 2008. Genome-wide association studies: potential next steps on a genetic journey. Hum Mol Genet 17:R156–R165.

Miyagawa T, Nishida N, Ohashi J, Kimura R, Fujimoto A, Kawashima M, Koike A, Sasaki T, Tanii H, Otowa T, Momose Y, Nakahara Y, Gotoh J, Okazaki Y, Tsuji S, Tokunaga K. 2008. Appropriate data cleaning methods for genome-wide association study. J Hum Genet 53:886–893.

Murcray CE, Lewinger JP, Gauderman WJ. 2009. Gene-environment interaction in genome-wide association studies. Am J Epidemiol 169:219–226.

Nakaoka H, Inoue I. 2009. Meta-analysis of genetic association studies: methodologies, between-study heterogeneity and winner's curse. J Hum Genet 54:615–623.

Pereira TV, Patsopoulos NA, Salanti G, Ioannidis JP. 2009. Discovery properties of genome-wide association signals from cumulatively combined data sets. Am J Epidemiol 170:1197–1206.

Psychiatric GWAS Consortium Steering Committee. 2009. A framework for interpreting genome-wide association studies of psychiatric disorders. Mol Psychiatry 14:10–17.

Seldin MF, Price AL. 2008. Application of ancestry informative markers to association studies in European Americans. PLoS Genet 4:e5.

Sladek R, Rocheleau G, Rung J, Dina C, Shen L, Serre D, Boutin P, Vincent D, Belisle A, Hadjadj S, Balkau B, Heude B, Charpentier G, Hudson TJ, Montpetit A, Pshezhetsky AV, Prentki M, Posner BI, Balding DJ, Meyre D, Polychronakos C, Froguel P. 2007. A genome-wide association study identifies novel risk loci for type 2 diabetes. Nature 445:881–885.

Weinberg CR. 2009. Less is more, except when less is less: studying joint effects. Genomics 93:10–12.

Wellcome Trust Case Control Consortium. 2007. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447:661–678.

Yeager M, Orr N, Hayes RB, Jacobs KB, Kraft P, Wacholder S, Minichiello MJ, Fearnhead P, Yu K, Chatterjee N, Wang Z, Welch R, Staats BJ, Calle EE, Feigelson HS, Thun MJ, Rodriguez C, Albanes D, Virtamo J, Weinstein S, Schumacher FR, Giovannucci E, Willett WC, Cancel-Tassin G, Cussenot O, Valeri A, Andriole GL, Gelmann EP, Tucker M, Gerhard DS, Fraumeni Jr JF, Hoover R, Hunter DJ, Chanock SJ, Thomas G. 2007. Genome-wide association study of prostate cancer identifies a second risk locus at 8q24. Nat Genet 39:645–649.