Next Generation Analytic Tools for Large Scale Genetic Epidemiology Studies of Complex Diseases

Leah E. Mechanic,[1]* Huann-Sheng Chen,[2] Christopher I. Amos,[3] Nilanjan Chatterjee,[4] Nancy J. Cox,[5] Rao L. Divi,[1] Ruzong Fan,[6] Emily L. Harris,[7] Kevin Jacobs,[8] Peter Kraft,[9] Suzanne M. Leal,[10] Kimberly McAllister,[11] Jason H. Moore,[12] Dina N. Paltoo,[13] Michael A. Province,[14] Erin M. Ramos,[15] Marylyn D. Ritchie,[16] Kathryn Roeder,[17] Daniel J. Schaid,[18] Matthew Stephens,[19,20] Duncan C. Thomas,[21] Clarice R. Weinberg,[22] John S. Witte,[23] Shunpu Zhang,[2] Sebastian Zöllner,[24] Eric J. Feuer,[2] and Elizabeth M. Gillanders[1]

[1]*Epidemiology and Genetics Research Program, Division of Cancer Control and Population Sciences, National Cancer Institute, NIH, Bethesda, Maryland*
[2]*Surveillance Research Program, Division of Cancer Control and Population Sciences, National Cancer Institute, NIH, Bethesda, Maryland*
[3]*Department of Epidemiology, The University of Texas, MD Anderson Cancer Center, Houston, Texas*
[4]*Biostatistics Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, NIH, Bethesda, Maryland*
[5]*Section of Genetic Medicine, Department of Medicine, The University of Chicago, Chicago, Illinois*
[6]*Department of Statistics, The Texas A&M University, College Station, Texas*
[7]*Translational Genomics Research Branch, Division of Extramural Research, National Institute of Dental and Craniofacial Research, NIH, Bethesda, Maryland*
[8]*Division of Cancer Epidemiology and Genetics, National Cancer Institute, NIH, Bethesda, Maryland*
[9]*Department of Epidemiology, Harvard School of Public Health, Boston, Massachusetts*
[10]*Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, Texas*
[11]*Susceptibility and Population Health Branch, National Institute of Environmental Health Sciences, NIH, Research Triangle Park, North Carolina*
[12]*Dartmouth College, Institute for Quantitative Biomedical Sciences, Lebanon, New Hampshire*
[13]*Division of Heart and Vascular Diseases, National Heart, Lung and Blood Institute, National Institutes of Health, Bethesda, Maryland*
[14]*Division of Statistical Genomics, Washington University School of Medicine, St. Louis, Missouri*
[15]*Office of Population Genomics, National Human Genome Research Institute, NIH, Bethesda, Maryland*
[16]*Center for Systems Genomics, Biochemistry and Molecular Biology, Pennsylvania State University, University Park, Pennsylvania*
[17]*Department of Statistics, Carnegie Mellon University, Pittsburgh, Pennsylvania*
[18]*Division of Biomedical Statistics and Informatics, Mayo Clinic, Rochester, Minnesota*
[19]*Department of Human Genetics, The University of Chicago, Chicago, Illinois*
[20]*Department of Statistics, The University of Chicago, Chicago, Illinois*
[21]*Preventive Medicine (Division of Biostatistics), Keck School of Medicine, University of Southern California, Los Angeles, California*
[22]*Biostatistics Branch, National Institute of Environmental Health Sciences, NIH, Research Triangle Park, North Carolina*
[23]*Department of Epidemiology and Biostatistics and Institute of Human Genetics, University of California San Francisco, San Francisco, California*
[24]*Department of Biostatistics, University of Michigan, Ann Arbor, Michigan*

Over the past several years, genome-wide association studies (GWAS) have succeeded in identifying hundreds of genetic markers associated with common diseases. However, most of these markers confer relatively small increments of risk and explain only a small proportion of familial clustering. To identify obstacles to future progress in genetic epidemiology research and provide recommendations to NIH for overcoming these barriers, the National Cancer Institute sponsored a workshop entitled "Next Generation Analytic Tools for Large-Scale Genetic Epidemiology Studies of Complex Diseases" on September 15–16, 2010. The goal of the workshop was to facilitate discussions on (1) statistical strategies and methods to efficiently identify genetic and environmental factors contributing to the risk of complex disease; and (2) how to develop, apply, and evaluate these strategies for the design, analysis, and interpretation of large-scale complex disease association studies in order to guide NIH in setting the future agenda in this area of research. The workshop was organized as a series of short presentations covering scientific (gene-gene and gene-environment interaction, complex phenotypes, and rare variants and next generation sequencing) and methodological (simulation modeling and computational resources and data management) topic areas. Specific needs to advance the field were identified during each session and are summarized. *Genet. Epidemiol*. 36 : 22–35, 2012.

Key words: gene-gene interactions; gene-environment interactions; rare variants; next generation sequencing; complex phenotypes; simulations; computational resources

Leah E. Mechanic and Huann-Sheng Chen; Eric J. Feuer and Elizabeth M. Gillanders contributed equally to this work.

*Correspondence to: Leah Mechanic, Division of Cancer Control and Population Sciences, National Cancer Institute, Epidemiology and Genetics Research Program, Host Susceptibility Factors Branch, 6130 Executive Boulevard, Room 5126, Bethesda, MD 20892 (Rockville, MD for express mail). E-mail: mechanil@mail.nih.gov

Published online in Wiley Online Library (wileyonlinelibrary.com/journal/gepi).
DOI: 10.1002/gepi.20652

INTRODUCTION

In the past 5 years, genetic epidemiology studies have progressed from investigating single variants in candidate genes to interrogating millions of variants in genome-wide association studies (GWAS). Over the past several years, GWAS have succeeded in identifying hundreds of genetic markers associated with common diseases. However, most of these markers confer relatively small increments of risk and, collectively, explain only a small proportion of familial clustering.

There are many possible explanations for the low attributable risk observed from current discoveries. Analysis of GWAS data has focused almost exclusively on detecting single SNP effects, while the bulk of susceptibility may reside in more subtle configurations that require accurate modeling of gene-gene or gene-environment interactions. Imprecise phenotyping or genetic heterogeneity may also contribute to difficulties in detecting genetic associations. In addition, findings have been limited to variants represented on commercially available genotyping arrays. Therefore, the contribution of other classes of genetic variation to complex diseases, including less common or rare variants and structural variants, has not been well-studied [Eichler et al., 2010; Lander, 2011; Manolio et al., 2009]. Each of these explanations presents numerous study design and analytical challenges to overcome. Moreover, with the advent of even higher density genotyping platforms and next generation sequencing technologies, new opportunities are emerging for identifying even larger numbers of genetic variants associated with disease. The tsunami of data generated using these technologies has created tremendous challenges for data management, storage, and interpretation. Despite these challenges, the National Institutes of Health (NIH) have funded few grants to specifically develop new analytic tools for genetic epidemiology studies.

To identify obstacles to future progress in genetic epidemiology research and provide recommendations to NIH for overcoming these barriers, the National Cancer Institute (NCI) sponsored a workshop, entitled "Next Generation Analytic Tools for Large-Scale Genetic Epidemiology Studies of Complex Diseases" on September 15–16, 2010. As the analytical challenges apply to multiple disease phenotypes, planning and organization of the workshop was performed by a trans-NIH steering committee, including members from the NCI (Division of Cancer Control and Population Sciences and Division of Cancer Epidemiology and Genetics), National Institute of Dental and Craniofacial Research, National Institute of Environmental Health Sciences (NIEHS), National Heart Lung and Blood Institute (NHLBI), and National Human Genome Research Institute. The goal of the workshop was to facilitate discussions on (1) statistical strategies and methods to efficiently identify genetic and environmental factors contributing to the risk of complex disease; and (2) how to develop, apply, and evaluate these strategies for the design, analysis, and interpretation of large-scale complex disease association studies in order to guide NIH in setting the future agenda in this area of research. To best represent a variety of view points and perspectives, the workshop brought together staff from NIH Institutes and included experts in fields of biostatistics, genetics, statistical genetics, genetic epidemiology, epidemiology, and computer science (Supplemental Table I).

The workshop was organized as a series of short presentations covering scientific (gene-gene and gene-environment interaction, complex phenotypes, and rare variants and next generation sequencing) and methodological (simulation modeling and computational resources and data management) topic areas, providing an overview of the state of the science. These presentations were followed by two breakout sessions on each topic area to discuss a series of questions prepared in advance of the meeting. Nilanjan Chatterjee (NCI Division of Cancer Epidemiology and Genetics) and Peter Kraft (Harvard School of Public Health) described challenges for studying gene-gene and gene-environment interactions. Nancy Cox (University of Chicago) and Matthew Stephens (University of Chicago) emphasized the importance of adequately appreciating the complexity of phenotypes in genetic analyses and presented alternative approaches to account for this complexity in analysis. John Witte (University of California, San Francisco) and Kathryn Roeder (Carnegie Mellon University) discussed different methods for analysis of rare and uncommon variants. Christopher Amos (The University of Texas M.D. Anderson Cancer Center) and Michael Province (Washington University School of Medicine) commented on the utility of simulation modeling, different methods and models, and potential pitfalls of simulation studies. Jason Moore (Dartmouth Medical School) and Marylyn Ritchie (Vanderbilt University) described the challenges of working with large-scale genetic epidemiology data sets and opportunities for the future using novel analytic and computational methods. Daniel J. Schaid (Mayo Clinic) summarized the topics discussed in the workshop and suggested several approaches for addressing challenges during his keynote address. After the breakout sessions, short reports were presented to the entire group by the discussion leaders (Gene-Gene and Gene-Environment Interactions, Clarice Weinberg, NIEHS; Complex Phenotypes, Duncan C. Thomas, Keck School of Medicine, University of Southern California; Rare Variants and Next Generation Sequencing, Sebastian Zöllner, University of Michigan School of Public Health; Simulations, Suzanne M. Leal, Baylor College of Medicine; Computational Resources and Data Management, Kevin Jacobs, NCI Division of Cancer Epidemiology and Genetics).

This meeting report summarizes the discussions held during the workshop and provides recommendations to facilitate future research in next generation of analytic tools for large-scale population studies. Specific recommendations to advance the field were identified during each session and are summarized in Table I.

GENE-GENE AND GENE-ENVIRONMENT INTERACTIONS

One session of this workshop was devoted to the challenges and issues associated with identifying gene-environment (G × E) and gene-gene (G × G) interactions in complex human diseases [for review: Cordell, 2009; Thomas, 2010]. As many GWAS of complex diseases are completed, there is a strong interest in the scientific community to follow up on these findings with analysis of G × E and G × G interactions (Recommendations 1.1, 1.2). There are multiple reasons for studying these types of interactions including integrating biological pathways and

TABLE I. Overview of recommendations to foster the next generation of research in genetic epidemiology of complex diseases

1. Gene-gene and gene-environment interactions
  1.1 Re-analysis of existing GWAS studies for G × G and G × E interactions
  1.2 Meta-analyses of existing studies for G × G and G × E interaction studies
  1.3 Development of methods and study designs that can better identify gene-gene (G × G) and gene-environment (G × E) interactions with user-friendly software
  1.4 Comparisons of study design models and methods for use with different data sets (e.g., case-control studies, family studies) for best approaches for G × G and G × E interaction discovery
  1.5 Improved methods of exposure assessment
  1.6 Incorporation of improved environmental measures into planning of long-term cohort studies
  1.7 Incorporation of repeated measures of exposures over time in cohort studies
  1.8 Reducing measurement error and identification of methods to handle sources of variability and optimal methods for sampling when using biomarkers
  1.9 Improved causal models to integrate biomarkers, genes, the environment and traits
  1.10 Harmonization of data types (especially environmental measures)
  1.11 Functional studies (e.g. mouse models, cell lines) for validation of interaction findings for genetic variants and environmental factors

2. Complex phenotypes
  2.1 Integration of many data types (e.g. genomics, metabolomics, gene expression, and epigenomics)
  2.2 Development of analysis methods that enable more efficient use of systems of related traits
  2.3 Increased investment in robust and accurate phenotype assessment
  2.4 Increased focus on identifying and validating intermediate phenotypes
  2.5 Building of pathway/network models to link inherited and acquired variation together in a unified framework
  2.6 Evaluation of the effectiveness of retrieving phenotypic data from EHRs compared to data collection methods used in population-based research studies
  2.7 Generation of one or more large and well-documented sets of study participants with extensive phenotype, genotype, and other “-omics” data to be used as a resource for developing and testing data integration methods

3. Rare variants and next generation sequencing
  3.1 Development and sharing of quality control (QC) methods for next generation sequencing
  3.2 Create a forum for establishing QC standards for next-generation sequencing
  3.3 Generation and sharing of standardized data sets for testing and evaluation of sequencing errors
  3.4 Development of improved methods to account for sequencing error (systematic and random)
  3.5 Improved methods to account for ambiguous functional information for variants, in particular in noncoding genomic regions
  3.6 Data to interpret accuracy of functional prediction algorithms for variants
  3.7 Estimation of population-specific variant frequencies
  3.8 Pilot sequencing studies to inform data analysis

4. Simulations
  4.1 Development and distribution of standardized simulated data sets
  4.2 Development of well-documented, modular, user-friendly, and open source simulation program(s)
  4.3 Availability of simulated data sets through a common repository, such as the database of Genotypes and Phenotypes (dbGaP)
  4.4 Creation of a web-based catalog of simulation programs and data sets
  4.5 Forums for collaboration among simulation modelers to advance the science of simulation in a systematic manner
  4.6 Funding for documentation of simulation programs and data sets
  4.7 Improved funding for simulation modeling

5. Computational resources and data management
  5.1 Regional shared computational clusters
  5.2 Training of graduate students and post-doctoral fellows in computer programming
  5.3 Increased funding and support for computational personnel
  5.4 More efficient data formats and data structures for data storage
  5.5 Conference to develop a consensus on best practices and methods to store, deliver, archive, and describe data
  5.6 Improved methods for combining data across multiple data sources and data types
  5.7 Development and sharing of standard quality control procedures for data combination and integration
  5.8 Development of new open-source, user-friendly analytical tools
  5.9 Establishment of new opportunities to support analytical tool development
  5.10 Conference to form consensus standards and identify formats needed for development of analytical software tools
  5.11 Improved annotation and curation of biological pathway databases
  5.12 Development of tools for data visualization
  5.13 Creation or identification of common, easily accessible, data sets for methods development
  5.14 Forum to share lessons learned regarding data management and analysis

TABLE I. Continued.

6. Overall
   - 6.1 Designate a set of GWAS (and/or sequencing) data sets available in dbGaP that are appropriate for methods development and testing, allowing performance of different methods to be more readily compared
   - 6.2 Sharing of lessons-learned for quality control and data analysis

understanding biological mechanisms that cause complex diseases, improving the power to discover underlying susceptibility loci, explaining heterogeneity across studies, and identifying susceptible subpopulations. Understanding G × E interactions involved in complex disease outcomes may improve performance of risk prediction models for disease prevention and treatment and is the basis of the field of pharmacogenomics [Thomas, 2010]. Despite the potential importance of these interaction studies, an open question is how much of the risk for trait or disease can be explained by G × G or G × E jointly, if joint effects are properly accounted for during analysis. Once concern is that empirical data suggest that interactions, even if they are present, may not explain much of the heritability [Hill et al., 2008]. Some recent reports suggested that multiplicative interactions, even if present, are likely to be of modest magnitude for complex diseases and may not be easily detectable using GWAS approaches [Ciampa et al., 2011; Milne et al., 2010].

An overarching theme of this session was to identify challenges for detecting and interpreting G × G and G × E interactions. It was noted that statistical interaction, per se, offers only circumstantial evidence of biological mechanism, i.e. statistical interaction is not the same as a biological interaction [Blot and Day, 1979; Rothman et al., 1980; Siemiatycki and Thomas, 1981; Thompson, 1991; Weinberg, 1986]. This distinction is due in part to the fact that detection of a statistical interaction refers to departure from a model on a particular scale [Thomas, 2004]. Moreover, in the context of GWAS data, only markers are measured and rather than the functional or causal variants themselves, which may not be ideal for testing mechanistic interactions. Even if the appropriate scale is selected, the power to detect interactions is often limited [Greenland, 1993]. However, participants argued that even in absence of statistical interaction, identifying the increased risk of disease with the combination of G and E remains of public health relevance. Participants agreed that although there are many different ways to statistically define G × G or G × E interactions, the key question is how to model joint effects when both multiple genes or multiple genes and multiple environmental factors influence risk. Most disease is "complex" in that multiple factors contribute to etiology, and the fundamental issue is not whether interactions exist, but how to jointly model the effects of multiple contributors to risk.

The presentations and breakout sessions focused on identifying novel study designs and strategies that would best detect G × G and G × E interactions, given the challenge of poor statistical power for detection (Recommendations 1.3, 1.4). Although very large sample sizes are generally necessary for identifying interactions, some methods, particularly case-only approaches, can be exploited to increase statistical power for detecting interactions, when the interacting factors can be assumed to be independently distributed in the underlying population [Piegorsch et al., 1994]. The gain in power, however, comes at the risk of major bias when the independence assumption is violated [Albert et al., 2001; Cornelis et al., 2011; Mukherjee et al., 2011; Thomas et al., 2011]. Recently, various types of hybrid methods have been proposed that can exploit the assumption of gene-environment independence to gain power, but are also robust to violation of the independence assumption [Chen et al., 2009; Li and Conti, 2009; Mukherjee and Chatterjee, 2008; Murcray et al., 2009].

Efficiency for studies of interaction can be enhanced by two-phase designs [Breslow and Chatterjee, 1998]. In a two-stage case-control design, the population can be oversampled for a rare exposure prior to genotyping to improve the statistical power to detect G × E. Designs that look for genetic effects in environmentally exposed or high-risk populations (e.g. individuals exposed to radiation for breast cancer treatment or BRCA1/2 carriers, respectively) can also be a powerful approach for detecting G × E interaction. Prior biological or pathway knowledge, if available, could also be incorporated into G × E scans and may increase power to detect associations. The formation of consortia can ultimately be used to achieve very large sample sizes, but this approach may also increase the heterogeneity of the study population.

Many of the methods described were focused on pairwise interactions, although it was acknowledged that more complex interactions are likely involved in the etiology of complex diseases. Methods for exploring higher order interactions have been developed including those for candidate genes or SNPs selected after prescreening [Kooperberg and Ruczinski, 2005; Ritchie et al., 2001] and in the context of GWAS discovery [Schwarz et al., 2010; Zhang and Liu, 2007]. There has been major computational progress toward examining these complex interactions, but challenges remain, including the need for even greater sample sizes to detect complex interactions compared with pairwise tests, and difficulty of replicating complex interactions in unique populations.

Another challenge identified by the group was difficulty in measuring the environment. Given that environmental exposures drive the etiology of many complex diseases, how an exposure is measured, where in the population distribution an exposure falls, and when an exposure is measured in a population can have profound consequences for discovering and/or replicating G × E [Kraft and Hunter, 2009]. In addition, for several complex diseases, the critical time period for measuring exposure in unclear. Moreover, most current measurements of environmental factors have significant margins of error. Taken together, these challenges decrease the power to detect G × E.

The participants discussed different approaches to minimize measurement error in exposure assessment while controlling costs. One option is a two-stage design in which a subsample of the population is measured using a more accurate exposure test, the remainder is measured with a cost-efficient test, and a joint analysis of main and subsample data is performed. Alternatively, it may be cost-effective to use the less precise measure and make up the loss in power with increased enrollment. Clever multistage sampling, such as exposure stratified sampling [Breslow and Chatterjee, 1998] or countermatching [Langholz and Borgan, 1995], can also increase the efficiency of study while keeping costs down [Andrieu et al., 2001]. In addition, in case-control studies, panels of related exposures can be pooled for increased power and decreased cost. Meeting participants identified several needs including the development of improved methods of exposure assessment and incorporation of methods into long-term cohort studies (Recommendations 1.5–1.7).

Biomarkers can be used in epidemiology studies as a method to assess environmental exposure. These markers can be either direct measurements as biomarkers of exposure or measures of early effect [Vineis and Perera, 2007]. Exposure biomarkers are direct measurements of the toxin or chemical, whereas measures of early effect reflect the underlying biological process. An example of a measure of early effect may include epigenetic changes, which may result in changes of gene expression in response to environmental exposure [for review: Jirtle and Skinner, 2007]. Measurement error can be a particular challenge when using biomarkers because of sources of variability, such as heterogeneity across individuals in absorption and metabolism, systematic and random differences between levels in the substrate measured compared with the target organ, variability from handling of biospecimens, timing of collection (e.g. circadian, or seasonal variability in marker levels), and laboratory variability, among others. It was cautioned that studies using biomarkers as surrogates for exposures need to be carefully designed since genetics might influence the measurement of the exposures the biomarkers are representing (e.g. genetic variation might influence uptake or alter metabolism of a toxin). Some needs that were identified for measuring biomarkers included: reducing the measurement error, identifying methods to handle sources of variability, and identifying optimal methods for sampling when using biomarkers (Recommendation 1.8).

In addition to needs identified regarding biomarkers themselves, improved causal models are needed to integrate biomarkers, genes, the environment, and traits (Recommendation 1.9). Although establishing causation in human studies is notoriously difficult, fraught with hidden biases, the one-way causal direction from genes to phenotypes has been leveraged to build and evaluate causal models in genetic epidemiology based on Mendelian randomization [Bochud et al., 2008; Smith and Ebrahim, 2004]. Further work to disentangle the influence of genes on biomarkers, intermediate traits, and distant phenotypes might guide interpretations of genetic associations [Vansteelandt et al., 2009], and might guide further laboratory studies of genetic causation [Drake et al., 2006; Schadt et al., 2005].

Debate with some level of disagreement focused on what constitutes appropriate replication in G × E and G × G interaction studies. Some participants suggested that too much emphasis is placed on replication and it is possible that a truly functional variant may not replicate across different study populations for biological and statistical reasons. Under a model of complexity with significant G × G and G × E interactions, it could be a challenge to replicate between independent studies when those samples are drawn from different populations with different underlying environmental exposures and genetic modifiers. However, given that an observed association also can arise purely from chance in the context of genome-wide analysis, other participants cautioned against lowering the threshold of significance for G × G or G × E interactions, since doing so would increase false-positive associations. In addition, interaction findings may need to be prioritized by the interpretability of the observed joint effect and whether such a joint effect replicates in independent studies. It was pointed out that replication of some kind of interaction coefficient in an independent study is not easily interpretable unless investigators observe similar patterns of joint effects. This may be a challenge if the genetic marker is not the causative variant, as different patterns of linkage disequilibrium (LD) between populations could result in conflicting or even opposite detected interactions.

Other unique challenges associated with G × E and G × G interaction studies were discussed. The most predictive statistical or analytical methods for integrating environmental exposures over time are not clear, and this will need to be resolved since many exposures affect disease risk over long periods of time. Some participants suggested that geospatial technology should be utilized to measure some exposures over a continuous timeframe (e.g. using residential history to measure air pollution exposure over time). Harmonization of exposures across studies will increase our ability to form consortia to detect G × E interactions, but is extremely difficult retrospectively (Recommendation 1.10). Unique challenges also arise when exploring the role of G × G and G × E interactions for prenatal effects—two genomes are involved, imprinting effects may play a key role, and the timing of the exposure measurements is critical, particularly with regard to birth defects. Improved methods for determining how historical time-dependent exposures should be weighted over time could provide insights into disease latencies and mechanisms. The group also suggested that functional studies (e.g. mouse models, cell lines) are needed to validate interaction findings and complement population-based observations (Recommendation 1.11).

COMPLEX PHENOTYPES

Understanding the underlying biological mechanisms is a principal goal in the study of complex traits/diseases. The heterogeneous nature of many complex traits/diseases and the difficulty in obtaining and integrating multifaceted phenotypic information to characterize and subtype complex diseases create analytical challenges for genetic epidemiology studies. In recent years, technology has improved our ability to identify thousands of genetic variants associated with diseases or traits. It is now possible to annotate the genome with functional information about how genetic variants relate to expression, methylation, metabolite levels, or protein levels, further improving the ability to identify interesting genetic variants. Despite the large number of discoveries, translating these discoveries into biological insights remains challenging. There are few examples of insights into biological pathways that stand out [i.e. the complement system in macular degeneration [Klein et al., 2005] and autophagy in Crohn's disease [Rosenstiel et al., 2009] among the large number of SNP discoveries].

The impact of phenotype characteristics on the study of genetic variants was explored by examining the distribution of the number of SNP discoveries in GWAS by phenotype using the Long-Cox ratio [Cox, 2010]. The Long-Cox ratio is the ratio of the number of SNPs reproducibly associated with a complex trait by GWAS divided by the log of the number of genotypes (full GWAS plus followup) required when making these discoveries. By comparing the ratio for different phenotypes, several characteristics were observed. In phenotypes, such as Type 1 diabetes, Crohn's disease, and rheumatoid arthritis, more discoveries were made with less genotyping investment. It was suggested that these phenotypes may be more precisely defined and better distinguished from other closely related phenotypes, and involve single organ systems. In contrast, diseases that may be considered to have less precise phenotype definitions and involved multiple organ systems (e.g. hypertension and cardiovascular disease) had fewer discoveries with much greater genotyping investment. Consistent with this observation, association of genetic variants with specific clinical subtypes of breast cancer was observed [Broeks et al., 2011]. These results suggest that increased precision of phenotyping may reduce heterogeneity and improve power to detect genetic variants associated with disease.

Multidimensional and multivariate methods to integrate information from different data types and related phenotypes will be key to making progress (Recommendation 2.1). Integrating data types, including "omics" data such as DNA methylation, RNA expression, and protein expression, may lead to improved understanding of biological mechanisms. Endophenotypes and intermediate phenotypes, defined as traits that are heritable and associated with disease, may provide insight into biological mechanisms for complex diseases because these intermediate traits are more proximal to genetic variation or environmental exposure than the complex disease [Gottesman and Gould, 2003; Kendler and Neale, 2010]. However, experience thus far shows that the genetic relationship between a disease and its endophenotypes or intermediate traits may be complex, with potential differences in underlying genetic mechanisms [e.g. glucose traits and type 2 diabetes: De Silva and Frayling, 2010; cognitive performance and schizophrenia: Cirulli and Goldstein, 2010].

Despite the proliferation of data and information, much of the analysis has focused on simple, univariate analyses. Exploiting multivariate information to perform more complex analyses could improve statistical power and provide scientific insights, especially when phenotypes are related (e.g. height and weight). One method discussed was to test partitions of independent, dependent, or linear combinations of effects for the different phenotypic measures and using data to provide more or less support for particular partitions [Stephens, 2010]. Multivariate methods could be expanded to facilitate data integration and incorporate multiple levels of phenotypic information. However, improved methods are needed to effectively model these systems and fit topological graphs of complex relationships between phenotypic measures (Recommendation 2.2).

Numerous challenges regarding complex phenotypes were identified and discussed. In many cases to date, the relative investment in collecting robust and accurate phenotype and environmental exposure data has been small compared to the time and resources devoted to measuring and interpreting genomic data. If phenotypic heterogeneity reflects etiologic heterogeneity, measuring phenotypic subtypes, biomarkers, and intermediate phenotypes in addition to primary disease endpoints, can improve power to identify the underlying genetic risk factors contributing to disease. However, choosing the correct intermediates to measure and analyze can be difficult because of gaps in biological knowledge. For example, only limited numbers of intermediate phenotypes and biomarkers have been identified for cancer (e.g. colorectal polyps and prostate-specific antigen). It was emphasized that endophenotypes are more informative for prospective studies due to the risk of "reverse causation" in retrospective designs. Meeting participants recommend increased investment in robust phenotyping and emphasis on identifying and validating intermediate phenotypes (Recommendations 2.3–2.5). The group identified data types currently overlooked that may provide biological insights in complex phenotypes: Geographic Information Systems, animal model systems (e.g. zebrafish) where high-throughput assays may be performed, cell models and systems (e.g. induced pluripotent stem cells), and data regarding association between gene expression and genotype in cell lines to inform epidemiological research.

It was also noted that the cost of collecting extensive, well-measured phenotypes from large numbers of research participants is currently prohibitive. Therefore, different study designs and approaches were discussed. Two-phase study designs in which more detailed phenotypic information is collected in subsamples and analyzed jointly or one-stage study designs that sample a subset of individuals in more detail can help to reduce the costs of measuring phenotypes and exposures in large studies. Another approach is to perform a multistage design in which sampling is performed based on genetic and phenotype data to maximize efficiency. In this type of design, one would need to account for the bias in the sampling of the population. Biorepositories linked to Electronic Health Records (EHRs) contain a wealth of phenotypic data that could be exploited as an alternative method of collecting dense phenotype data. However, retrieving phenotypic data from EHRs is challenging and the comparative advantage of clinical biorepositories compared to data collection in population-based studies needs to be evaluated (Recommendation 2.6).

Another major challenge arises from the difficulties in integrating multiple data types, such as genotype, gene expression, microbiome, metabolomic, and methlyation data, is to both define a more etiologically homogenous population for analysis and better understand the biology contributing to disease risk and phenotypic variation. The group identified several opportunities to address these challenges. Ideally, one or more large and well-documented sets of study participants with extensive phenotype, genotype, and other -omics data could be generated and made broadly available as a resource for testing methods for data integration (Recommendation 2.7). The NIH Common Fund Project, Genotype-Tissue Expression (GTEx) (https://commonfund.nih.gov/GTEx/), is an example of such a resource. However, establishing such resources is complicated and expensive. Therefore, it might be more practical to integrate data using distinct studies or populations and use hierarchical modeling to integrate data from nonmatching data sets by making connections across models. For example, the association between genes and gene expression phenotypes could be examined in one data set and the model of this association then applied to data in a second population from which genes and phenotype data are obtained. It was also suggested that an increased emphasis should be placed on developing tools for data visualization, aligning data across disparate platforms, and providing increased funding for computational personnel.

To incorporate more precisely defined phenotypes or to better analyze complex phenotypes, larger sample sizes will be needed which result in challenges for integrating data across multiple distinct studies. While meeting participants agreed that there has been a shift toward consortia to achieve larger sample sizes, even more facilitation of interactions amongst investigators would be helpful. Another challenge to combining data from multiple different studies is the lack of uniformity in methods used for collecting phenotypic data, though activities such as the PhenX project (https://www.phenx.org/) are helping to address this issue. Finally, successful data integration across multiple studies and data sources will require additional information about individual data sets including detailed information about study design, phenotyping protocols, and quality control methods implemented.

RARE VARIANTS AND NEXT GENERATION SEQUENCING

Recent advances in next generation sequencing technologies have invigorated the search for rare variants involved in the etiology of complex traits. However, the power to detect an association with a single rare variant is low even in very large samples [for review: Bansal et al., 2010; Carvajal-Carmona, 2010; Cirulli and Goldstein, 2010]. In an effort to obtain increased statistical power from smaller sample sizes, many groups have investigated aggregating sets of rare variants into a single group and testing their collective frequency differences between cases and controls. Recently published studies show that power to detect rare variants can be greatly enhanced by collapsing variants in a target region, such as a gene or exon [Han and Pan, 2010; Li and Leal, 2008, 2010; Madsen and Browning, 2009; Morgenthaler and Thilly, 2007; Price et al., 2010].

However, it is very difficult to define which rare variants should be aggregated into a single group for analysis. One solution discussed at the meeting was to use empirical methods to determine the optimal weighting and aggregation scheme. In such cases, multiple aggregation schemes could be considered, evaluated based on some metric (e.g. minimum *P*-value), and corrected by permutation. A simulation study using sequence data from 18 candidate genes from the folate metabolism pathway was presented, which compared several different approaches to aggregating such variants. Two approaches to empirically determine the most efficient grouping of rare variants were described. The first considered multiple possible groupings leveraging functional elements and annotations (e.g. minor allele frequency, MAF, or protein coding function information) in a genomic region to collapse the variants together. The second was an agnostic "step-up" approach in which all possible subsets were considered and the grouping that best differentiates cases and controls was selected. Results of this simulation study showed that leveraging prior information to guide the collapsing of rare variants is advantageous only when information is quite accurate, but the step-up approach worked well across a broad range of plausible scenarios [Hoffmann et al., 2010].

In general, aggregation methods not only have increased power to detect associations but also have limitations. For example, most aggregation methods assume that all variation affecting phenotype acts in the same direction, but the effects of rare variants could be heterogeneous—some variants have no effect, some variants increase risk, and some variants are protective. Summing across variants with differential effects can obscure true associations. Therefore, statistical methods that allow for heterogeneity of SNP effects within a gene are needed. The C-α test statistic was presented as a novel approach, which is robust in the presence of a mixture of effects across a set of rare variants. Specifically, analysis of both simulated and case-control data showed that the C-α statistic had power comparable to other aggregation methods when all effects were in the same direction, but much greater power when protective and risk variants exist in the test set [Neale et al., 2011]. The HapMap project examining ENCODE regions has shown that rare variants are often unique to particular ethnic groups [International HapMap consortium, 2010; International HapMap Consortium, 2007; International HapMap, 2005], thus highlighting the potential for false-positive associations due to population stratification in rare variant association studies. Therefore, similar to all methods designed to discover rare variants the C-α test is sensitive to population stratification because its null assumes rare variants are equally likely in cases and controls; this could be controlled using standard strategies such as principal component analyses to match or control for ancestry.

A number of open questions surrounding the discovery of rare variants were discussed and meeting participants provided recommendations for advancing the study of rare variants. These included the implications and limitations of existing high coverage genotyping and sequencing technologies and how theory can incorporate any errors. The group agreed that sequencing, as opposed to ultra-high-density SNP chips, will be necessary to detect very rare variants and variants that only occur in cases. There was strong consensus that every subject in a next generation sequencing study should also be genotyped using a standard genome-wide chip for quality control (QC) purposes (including assessing and controlling for population stratification).

Participants recognized that assembly and base calling are difficult technological challenges, and the scientific community must proactively develop and implement standard QC measures (Recommendation 3.1). One approach the group supported was creating a forum to establish QC standards for next generation sequencing (Recommendation 3.2). It was noted that the ability to process sequencing data will likely improve in the future. Therefore, participants discussed storage of the raw BAM files. Failing to store these files reduces the utility of the stored data. However, storage costs for this data will be substantial. Several noted that these files should be stored until QC standards are established. At a minimum, the group suggested saving the BAM files for a few bench mark data sets. Publicly available data sets would be a valuable resource for testing error-checking methods (Recommendation 3.3). The group agreed that NIH would be wise to facilitate the generation and sharing of standard validation data sets. It would be helpful if each of the different sequencing platforms provided control or referent samples for each platform, and in particular, running each sequencing platform on a shared set of common samples would provide a means for cross-referencing strengths and weaknesses of each platform. Well-calibrated quality scores that evaluate the likelihood a polymorphism truly exists at a particular location and confidence in the genotype assignment could be used as a weight in a rare variants test. The challenge is obtaining well-calibrated scores similar to those generated by the phred/phrap software for Sanger sequencing reads. Error in sequencing studies can arise randomly as a result of low depth of coverage or systematically when sequencing particular local sequences (e.g. repetitive elements or homopolymer stretches). Further research is needed to develop improved methods to account for sequencing error, develop analytic methods that distinguish between random and systematic genotyping errors, and evaluate the impact of rare variant misclassification on genetic association studies (Recommendation 3.4).

Discussion also focused on the feasible models for rare variants and how population genetics models and existing GWAS and linkage data inform such models. Importantly, it was noted that the relative contribution of uncommon (MAF 0.5–5%), rare (MAF $<0.5\%$), and private (detected in single individuals) variants to disease risk is unclear. Participants agreed rare variants with much stronger associations or odds ratios (ORs) compared to those observed in studies of common variants (i.e. OR = 1–1.3) must be under selection pressure. Modeling of existing linkage study results can be informative for understanding potential architectures of rare variants for a given disease. Assuming that the variant is of intermediate frequency, some argued that in large multiplex families, any variants with ORs greater than 10 would already have been discovered by linkage analysis. Well-powered GWAS can provide an upper bound for the ORs and MAFs of rare variants. It is unclear whether most disease-associated rare variants will be located within exons, although GWAS results suggest that many disease-associated common variants are located outside of exons. Participants recommended that rare variant studies begin simply (e.g. Mendelian diseases and families) as a proof of principle. Until cost differentials between exome and whole-genome sequencing improve, focusing on exomes is sensible because exonic variation is easier to interpret than nonexonic variation. In approximately 12 months, NHLBI will have sequenced 10,000 exomes. These data will be tremendously informative for designing studies.

Other study design considerations were discussed. Some approaches for selecting the study population may improve the likelihood of discovery of rare variants [for review: Cirulli and Goldstein, 2010]. These include sampling individuals with extreme phenotypes, family history of disease, extreme environmental exposure for G × E studies, or using genetic linkage or haplotype data in selecting study populations. In addition to proof of principle studies, two-stage designs were discussed in which the first stage is used to screen for rare variants and the second stage is used for replication. One challenge encountered with multistage designs (or replication) is filtering variants for consideration in the second stage. Several methods may be used including filtering based on data quality (e.g. consistency, coverage), prediction of function, and expected population frequencies from HapMap or dbSNP. In family-based designs, an additional approach for filtering may include using co-segregation information. Regarding functional prediction, the field needs better methods to account for ambiguous functional information and data informing the accuracy of predictions (Recommendations 3.5, 3.6). Population-specific variant frequencies and accounting for population stratification also are needed to use methods that filter based on variant frequencies (Recommendation 3.7).

There was much discussion about the evidence or standards needed to evaluate associations observed with rare variants and whether replication should be used as a "gold" standard. Participants felt strongly that technical replication, repeating the sequencing or genotyping on the identical samples, should be absolutely required. It was noted that technical replications should start from the original specimen because library preparation is often a source of genotyping error. Meeting participants felt that allele frequency should influence replication strategy; single rare variants cannot be replicated in an independent sample. Unlike the standard set with GWAS requiring replication of the specific genetic variant, significant aggregation tests with different rare variants within the same gene or region should be considered replication.

Genotype imputation, which allows for the evaluation of evidence for association at genetic markers that are not directly genotyped, is now an essential tool in the analysis of GWAS. Current challenges for imputation of rare variants were discussed. The low MAF of rare variants and their low LD with other variants make them difficult to impute. The meeting participants felt that current reference genomes available in HapMap are insufficient to impute variants with MAF $<2\%$ [Zawistowski et al., 2010]. The count of a variant in a reference panel matters; observing a variant five or more times is sufficient for imputation. Quality of haplotyping is important for imputation. For accumulation tests, information from poor imputation still may usefully contribute to test statistics. In a recent study, whole-genome sequencing was performed on a small subset of a population with GWAS data, followed by imputation for the larger population, and association analysis and replication by direct genotyping identified a rare variant associated with sick sinus syndrome [Zeggini, 2011].

Finally, meeting participants were asked to consider the utility of funding a small (pilot) sequencing study of rare variants, given our current knowledge of the genetic architecture of common diseases (i.e. it will differ depending on the disease). The group agreed that low power studies can be informative for data handling issues (Recommendation 3.7). NIH can facilitate sharing of data sets and should facilitate data access. However, using public controls can increase sample issues (e.g. heterogeneity). Cases and controls should be genetically matched prior to performing a sequencing study.

SIMULATIONS

Simulations are often performed to evaluate conditions that could give rise to current observations, to develop replications that permit comparisons between statistical methods for analysis, and to evaluate how changes to a system alter its attributes. There are three types of approaches to simulation studies in humans: (1) coalescent (or backwards) approaches, in which ancestral conditions are modeled from present observation; (2) forward-time simulations, in which initial conditions are specified and modeled forward in time; and (3) sideways simulations, in which conditions are modeled by resampling existing data. Coalescent models can provide rapid simulations for restricted models, forward-time models provide flexibility to study a broad range of questions but are computationally demanding, and sideways models can simulate data from very dense sets of markers but are limited to haplotypes and alleles contained in available data. Hybrid approaches that may also be used, for fitting a coalescent model then applying a forward-time simulation in which the allele frequencies of the disease locus are constrained by the coalescent, allows data from large regions to be simulated [Peng et al., 2007]. Depending on the problem and available data, each of these approaches offers distinct advantages and disadvantages [for review: Liu et al., 2008; Ritchie and Bush, 2010].

A danger of using simulations is that selecting the simulated data set for the purpose of testing a particular method can become a self-fulfilling prophecy, i.e. the method being tested is optimal using the particular data set simulated. For example, if the data set was generated using known candidate regions, testing statistical methods leveraging biological knowledge will perform better than agnostic methods. Other potential sources of bias in generating simulated data sets are poor random number generators.

Currently, there is a lack of benchmark simulation programs or data sets and established criteria for evaluation of programs or models. Therefore, it is difficult for investigators who are less familiar with simulation methodologies to appropriately select programs and data sets. It was noted that some researchers are using outdated simulated data sets generated with overly simplistic models because they are unaware of other data sources. Evaluating simulation programs and resulting data sets is often difficult because of inadequate documentation or lack of comparability between approaches. As a result of discussions during the workshop, participants provided the following recommendations.

One recommendation was the creation of a small number of generalizable benchmark data sets and/or a reference simulation program(s) for the user community (Recommendation 4.1). The characteristics of potential benchmark data sets were discussed. Given the large amount of sequencing and genotyping data that will be available shortly (e.g. 1000 Genomes, NHLBI 1000 exomes) and limited knowledge about the genome, some suggested using sideways approaches to generate simulated data sets. Meanwhile other meeting participants commented that this approach is limited because it is not possible to model natural selection using sideways approaches. The following characteristics of benchmark data sets were suggested: genotypes of European and African-American ancestry, gene-based variants in pathways, and G × G and G × E. Moreover, exome sequence data selected for simulation modeling should have parallel GWAS data available to more comprehensively model genetic structure in these data sets. Several phenotypic models, including quantitative and qualitative traits, with multiple replicates could be added to the real genotype data to create the benchmark data sets. These sets should be available as Variant Call Format format, or binary version of the Sequence Alignment/Map by request. Probability terms for genotypes should be incorporated into the models to account for genotyping error.

Access to the reference data sets was also discussed. To facilitate sharing of the benchmark sets, access should be through NIH's database of Genotypes and Phenotypes (dbGaP) or a similar resource with a simple application process. Since sideways simulations could be generated from exome sequencing data (or even whole-genome sequencing data), there may be data access issues. To avoid this issue, the exomes selected for the generation of the simulated data might be limited to those that do not require IRB approval for access, and would not be subject to strict data use limitations; such requests would still require review by an NIH Data Access Committee.

Although benchmark data sets would benefit methods development, such data sets will not meet all investigator needs and cannot feasibly capture all genetic and phenotypic models of interest. Therefore, the group suggested that NIH support the development of a simulation program or series of programs (Recommendation 4.2). Several approaches for generating simulation program(s) were discussed, including development by a single group or as a collaborative effort by multiple groups. The latter was believed to be more beneficial, but was likely to be more expensive. Regardless of the approach, the group emphasized that any reference simulation program(s) developed should be well-documented, open-source, modular, user-friendly, adaptable and scalable to allow maximum flexibility and future development. An alternative suggestion called for a centralized group of simulators to generate a series of reference scripts that would be provided for users. Such reference programs or scripts could be used to generate data sets for a broader set of purposes than a small set of benchmark simulated data sets. Moreover, reference programs would reduce issues associated with data storage because the data sets could simply be regenerated. Several features were identified as desirable in a reference program(s), but the group recommended that the programs should begin with a smaller, initial set of features. Some suggested features included allowing different approaches to simulation, and modeling quantitative and qualitative traits, interactions, ascertainment models, and correlation of phenotypes.

As the group discussed possible benchmark data sets and programs, it was noted that several simulated data sets are currently available that may be used for methods development, including all simulated data generated for the Genetic Analysis Workshops (GAW) workshops. Another opportunity to foster analytical methods development is to facilitate sharing of these existing data sets, even in the absence of developing the proposed benchmark data sets. Therefore, the group recommended that researchers and GAW be encouraged to deposit simulated data to dbGaP for distribution (Recommendation 4.3). The GAW16 data are already available through dbGAP and could serve as a model. Deposition of the data to a common location would facilitate the development of documentation standards for simulated data and would serve as a standard resource for simulated data sets.

It would also alleviate some of the distribution burden on the individual investigators who generated the data. A concern was raised about whether dbGaP would be able to handle all this additional data.

Because many of these recommendations will take time to implement, the group suggested that an immediate opportunity could be to create a web page that summarizes available simulation programs and data, including links to both programs and data (Recommendation 4.4). The website could include a blog for researchers to comment on the utility of various software tools for different purposes and aid in evaluation of the software or data sets.

Meeting participants supported further research into a diversity of approaches to simulation, as well as hybrid approaches, and noted that the science of simulation could be stimulated by a supported forum for comparison of the structure, assumptions, and results of models in a systematic manner (similar to what has been done in the Cancer Intervention and Surveillance Modeling Network for population modeling) (Recommendation 4.5). Such a group could also develop benchmark data sets, and distribute well-documented programs with guidelines to assist selection of the most appropriate approach for specific situations. Moreover, the group noted that limited opportunities exist for those developing simulations to work together and collaborate on common issues. This type of forum could supplement existing analysis and methods workshops such as GAW or the Pharmacogenomics Research Network.

Finally, the group discussed the need for improved documentation and standards for simulation modeling. Funding is needed both for the development of programs and for the documentation of simulated data sets and programs (Recommendations 4.6 and 4.7). Often documentation of the program is the least exciting aspect of developing the program and there is currently limited financial support for documentation. Standards for data set and program documentation are also lacking.

COMPUTATIONAL RESOURCES AND DATA MANAGEMENT

Advances in next generation sequencing technologies and their use in discovering genetic variations associated with complex diseases are generating enormous amounts of data. The major bottleneck in genome sequencing is no longer data generation, but the computational challenges around data analysis, display, and integration of disparate data types [Green and Guyer, 2011]. As GWAS data have stretched informatics capacity, meeting the storage and analytical needs for next generation sequencing data will be even more challenging. The average next generation sequencing experiment generates terabytes of data [Zhang et al., 2011]. In fact, the rate of increase in DNA sequencing and genotyping capacity is outstripping the rate of increase in disk storage [Richter and Sexton, 2009; Stein, 2010]. In some cases, the expense of storing and archiving raw data is greater than repeating the experiment, which is not ideal given finite biospecimen resources and the need to evaluate changes in base-calling methods over time. In a review of informatics requirements for next generation sequence data, some needs outlined included scalable, dense, and inexpensive disk storage systems, high-performance disk storage systems, archival storage systems, improved software, data analysis tools and increased staffing to handle the large increase in data [Richter and Sexton, 2009].

Adequate computation infrastructure is necessary to analyze and manage large-scale data sets. Participants strongly agreed that there is a need for investigator(s)-led regional or statewide shared computing clusters, to support small- to medium-sized laboratories in particular (Recommendation 5.1). The group specified that researchers, rather than administrators, should design these clusters because of their better understanding of analytical needs. Exploiting the power of graphical processing units (GPUs) may enhance computational power [Sinnott-Armstrong et al., 2009; Greene et al., 2010] for some analytical studies. However, to effectively use clusters and GPU technologies (or GPGPU/FPGA), an increased emphasis on cluster-friendly software and conversion of applications for GPU computing is required. Overall, meeting participants noted that computing power, relative to other challenges discussed, is less of a limitation because they felt that the technology is already moving in this direction. However, it was noted that while the goal of the $1,000 genome sequencing platform is likely within reach, this does not include the cost of data analysis or storage, which some suggested could be substantial [Mardis, 2010; Pennisi, 2011]. Finally, given the computational demands of scientific research, the group advocated for improved and increased training of graduate students and postdoctoral fellows in computer programming either through training programs or other grant support and increased support of computational personnel (Recommendations 5.2 and 5.3).

Meeting participants agreed that impending issues surrounding data storage and networking were most daunting as the field migrates to denser genotyping and sequencing platforms and more complex analyses. Historical methods of data storage, e.g. relational databases, are no longer adequate for the next generation of studies. Some newer models to improve storage capacity being explored include compressed and binary formats for data, hybrid database architectures (e.g. row/column oriented designs or chunking formats), virtualization of data, and cloud computing. Participants identified needs for cost-effective and increased data storage capacity, including new, more efficient data structures and formats (Recommendation 5.4). In addition to new formats and structures, participants concluded that there is a need for establishing standardized data formats for data storage and networking. One approach for developing standards discussed was organizing a conference of multiple different interested groups to reach a consensus on the best methods to store, deliver, archive, describe, and distribute data (Recommendation 5.5).

In addition to issues regarding data storage, combining data sets from multiple sites and sources creates further challenges that require careful attention to QC assessment. Best practices for QC of GWAS data were recently published based on lessons learned from the eMERGE network and Gene Environment Association Studies (GENEVA) program [Laurie et al., 2010; Turner et al., 2001]. Integrating GWAS and sequencing data with other data sources, e.g. omics data, also leads to data management challenges. These additional -omic data sets are quite large. Improved methods for combining meta-dimensional data from multiple data sources (e.g. dbGaP, gene ontology databases) and data types (e.g. DNA, RNA, protein and clinical data) are needed along with standard QC procedures (Recommendations 5.6 and 5.7).

Analytical software tools are needed to efficiently manage large genotype and sequence data sets, including the ability to efficiently subset, merge, annotate, harmonize data, perform variant calling, run standard QC tasks, fit standard models to test for relationships and population structure, and detect association with phenotypes. Meeting participants advocated for the development of user-friendly, and ideally open-source, tools available for the research community to accommodate next generation sequencing data and more complex forms of data analysis (Recommendation 5.8). The group outlined needs for baseline tools and environments for core libraries, data management, QC, data analysis, and methods development. These tools would ideally be extendable to more sophisticated tasks and models. It was suggested that a new funding mechanism should be used to support such development (Recommendation 5.9). As well as funding support, the group suggested that NIH should coordinate the development of these resources by organizing a conference among multiple groups to form consensus standards and formats that are needed (Recommendation 5.10). The group emphasized that more collaboration is needed among computer scientists, statisticians, and biologists to more effectively leverage biological knowledge and interpret the vast amounts of data obtained.

As described during the session on complex phenotypes, most genetic epidemiology studies focused on a "one SNP at a time" approach, ignoring the complexity of disease pathways. Biological pathways are typically nonlinear and a linear assumption may hinder our ability to detect complex relationships. Therefore, some participants suggested that analytical approaches that handle different models may improve our ability to detect genetic variants or pathways with important roles in disease [Moore et al., 2010]. Some alternatives to traditional linear models mentioned included symbolic modeling of epistasis [Moore et al., 2007], computational evolution systems [Moore et al., 2008], and logic regression [Kooperberg et al., 2001]. Strengths of these approaches are that they are not bound by assumptions of the underlying linear model. However, some cautioned that testing such a large number of models, or even millions of models, may be hindered by a large false-positive rate (i.e. there is a low prior probability that any individual model is correct). This limitation may be partially addressed by providing biological knowledge to limit the model space of the search, i.e. limiting the number of potential models. Improved high-throughput biological systems, as currently being assessed in the National Toxicology Program's toxicogenomics program (http://ntp.niehs.nih.gov/?objectid = 7E6CAEBD-BDB5-82F8-F8C29152153B80B1), to test different models may be used to add to the knowledge base for analysis. Additional analytical approaches may be found in the computer science fields, such as quantum computing or immune systems.

Leveraging biological knowledge was suggested as an approach for managing and analyzing large complex genetic data sets. Biological annotation may be implemented to guide searches for interactions or as an approach for combining rare variants obtained by sequencing. However, many different databases exist (e.g. Kegg, Biocarta, Ensemble, Entrez, Gene Ontology), and these databases often use different genome builds and have inconsistencies in gene nomenclature. Another challenge for using these databases relates to the uncertainty in current pathway knowledge and ontologies. The uncertainty in these models within existing databases should be quantified and accounted for within analysis. Importantly, a large percentage of known genes do not map to functional annotation within these biological functional databases. Therefore, improved annotation is needed along with other sources of biological information to better assign pathways. Combining data based on biological pathway information is an emerging field—consistent, reliable, and well-curated annotation resources are needed (Recommendation 5.11).

Data visualization is an integral part of scientific discovery; this is challenging for large data sets. One novel method of visualization described was the 3D Heat Map which enables exploration of high-throughput data using an interactive medium that allows addition of information or annotation to the map [Moore et al., 2011]. The group discussed ways to utilize the emerging 3D visualization tools in genomic epidemiology. The majority agreed that innovative methods for visualizing data may lead to novel scientific discoveries and emphasized the need for baseline charts and visualizations which require no scripting, but could be used for a variety of applications and improve current capabilities for visualizing increasingly complex data sets (Recommendation 5.12).

Sharing genomics data sets must be facilitated by information technology, which must also enforce data access protections based on the guidance of IRBs, data access committees, legal requirements, and institutional policies. Participants debated the best practices for data sharing and discussed road blocks to accessing existing databases or datasets. The accessibility of data housed in dbGaP was discussed. Some meeting participants believed that although access to dbGaP is straightforward to users of the resource, the process may be a barrier to data sharing among nonusers. For example, in computer science fields, data sets often are made available by simply clicking on links. It was also suggested that some simple data sets could be made available or identified for sharing to allow researchers to use comparable sets for methods development research (Recommendation 5.13). Additionally, the group opined that, at present, computing architectures, software development, data cleaning methods, etc., are not generally published, despite being critically important. Participants suggested that the field should support a mechanism to publish and/or share this type of knowledge, perhaps through additional analytical conferences (Recommendation 5.14).

SUMMARY AND CONCLUSIONS

In addition to recommendations made regarding specific topic areas, several common themes became apparent over the course of the workshop. Meeting participants were excited about the opportunities for discovery using the avalanche of data now available (or arriving soon), but were cautious about the challenges ahead in regards to data management and interpretation. It was frequently noted that although the costs for laboratory genotyping assays have declined, substantial costs for data management and analysis remain. These areas are frequently under-represented or under-appreciated in grant applications and study sections. Some also cautioned that generating more data will not lead to new biological insights; instead, improved analytical methods may be more beneficial.

An emphasis in many discussions was that all these different data types and discoveries must be evaluated in a biological context. However, leveraging functional or mechanistic information is only as good as the science behind the annotations. Our understanding of the underlying biology is incomplete and future work should address methods to supplement current knowledge.

The current approach for analyzing much of the existing genetic epidemiology data based on identifying a list of SNPs reaching a *P*-value threshold also was discussed; this approach may not be optimal for studies of complex diseases. Several different methods were discussed, including multivariate methods to model related phenotypes. An alternative method could be to focus on model selection (e.g. by grouping SNPs into genes, genes into pathways, or environmental exposures) with the goal of replicating the model [Zhou et al., 2010]. Many other fields (e.g. computer science, physics, and operations research) have been working on approaches to analyzing high-dimensional data, and enhanced communication with other fields will foster future research.

As analytical approaches become more complex and integrate phenotypic data, multiple genotypes, environmental exposures, and less common or rare variants, replication of observed associations becomes more difficult. Therefore, several discussions were centered on the standards and requirements for replication.

Other themes included encouraging the use of common data sets, developing QC standards, and encouraging enhanced collaboration with other fields for future methods development. The use of common data sets, either identified in dbGaP or another resource, would facilitate the development and evaluation of methods and standards (Recommendation 6.1). Moreover, the group supported improved sharing of lessons learned or best practices and QC procedures not typically published, either through workshops or small meetings (Recommendation 6.2).

Finally, the workshop participants frequently noted challenges for future investigators. The group recommended improved computational training and support for graduate students and postdoctoral fellows, as these skills are essential for next generation studies. Importantly, as the current research environment is focused on team science and collaboration, the group suggested that there was a need for an improved system for recognizing the contribution of young investigators who are team players.

REFERENCES

Albert PS, Ratnasinghe D, Tangrea J, Wacholder S. 2001. Limitations of the case-only design for identifying gene-environment interactions. Am J Epidemiol 154:687–693.

Andrieu N, Goldstein AM, Thomas DC, Langholz B. 2001. Counter-matching in studies of gene-environment interaction: efficiency and feasibility. Am J Epidemiol 153:265–274.

Bansal V, Libiger O, Torkamani A, Schork NJ. 2010. Statistical analysis strategies for association studies involving rare variants. Nat Rev Genet 11:773–785.

Blot WJ, Day NE. 1979. Synergism and interaction: are they equivalent? Am J Epidemiol 110:99–100.

Bochud M, Chiolero A, Elston RC, Paccaud F. 2008. A cautionary note on the use of Mendelian randomization to infer causation in observational epidemiology. Int J Epidemiol 37:414–416.

Breslow NE, Chatterjee N. 1998. Design and analysis of two-phase studies with binary outcomes. Appl Stat 48:457–468.

Broeks A, Schmidt MK, Sherman ME, Couch FJ, Hopper JL, Dite GS, Apicella C, Smith LD, Hammet F, Southey MC, et al. 2011. Low penetrance breast cancer susceptibility loci are associated with specific breast tumor subtypes: findings from the Breast Cancer Association Consortium. Hum Mol Genet 20:3289–3303.

Carvajal-Carmona LG. 2010. Challenges in the identification and use of rare disease-associated predisposition variants. Curr Opin Genet Dev 20:277–281.

Chen YH, Chatterjee N, Carroll RJ. 2009. Shrinkage estimators for robust and efficient inference in haplotype-based case-control studies. J Am Stat Assoc 104:220–233.

Ciampa J, Yeager M, Amundadottir L, Jacobs K, Kraft P, Chung C, Wacholder S, Yu K, Wheeler W, Thun MJ, et al. 2011. Large scale exploration of gene-gene interactions in prostate cancer using a multi-stage genome-wide association study. Cancer Res.

Cirulli ET, Goldstein DB. 2010. Uncovering the roles of rare variants in common disease through whole-genome sequencing. Nat Rev Genet 11:415–425.

International HapMap Consortium TIH. 2010. Integrating common and rare genetic variation in diverse human populations. Nature 467:52–58.

Cordell HJ. 2009. Detecting gene-gene interactions that underlie human diseases. Nat Rev Genet 10:392–404.

Cornelis MC, Tchetgen Tchetgen EJ, Liang L, Qi L, Chatterjee N, Hu FB, Kraft P. 2011. Gene-environment interactions in genome-wide association studies: a comparative study of tests applied to empirical studies of type 2 diabetes. Am J Epidemiol.

Cox NJ. 2010. Complex traits. University of Chicago.

De Silva NM, Frayling TM. 2010. Novel biological insights emerging from genetic studies of type 2 diabetes and related metabolic traits. Curr Opin Lipidol 21:44–50.

Drake T, Schadt E, Lusis A. 2006. Integrating genetic and gene expression data: application to cardiovascular and metabolic traits in mice. Mamm Genome 17:466–479.

Eichler EE, Flint J, Gibson G, Kong A, Leal SM, Moore JH, Nadeau JH. 2010. Missing heritability and strategies for finding the underlying causes of complex disease. Nat Rev Genet 11:446–450.

Gottesman II, Gould TD. 2003. The endophenotype concept in psychiatry: etymology and strategic intentions. Am J Psychiatry 160:636–645.

Green ED, Guyer MS. 2011. Charting a course for genomic medicine from base pairs to bedside. Nature 470:204–213.

Greene CS, Sinnott-Armstrong NA, Himmelstein DS, Park PJ, Moore JH, Harris BT. 2010. Multifactor dimensionality reduction for graphics processing units enables genome-wide testing of epistasis in sporadic ALS. Bioinformatics 26:694–695.

Greenland S. 1993. Basic problems in interaction assessment. Environ Health Perspect 101:59–66.

Han F, Pan W. 2010. A data-adaptive sum test for disease association with multiple common or rare variants. Hum Hered 70:42–54.

Hill WG, Goddard ME, Visscher PM. 2008. Data and theory point to mainly additive genetic variance for complex traits. PLoS Genet 4:e1000008.

Hoffmann TJ, Marini NJ, Witte JS. 2010. Comprehensive approach to analyzing rare genetic variants. PLoS ONE 5:e13584.

International HapMap Consortium, Frazer A, Ballinger DG, Cox DR, Hinds DA, Stuve LL, Gibbs RA, Belmont JW, Boudreau A, Hardenbol P, Leal SM, Pasternak S, Wheeler DA, Willis TD, Yu F, Yang H, Zeng C, Gao Y, Hu H, Hu W, Li C, Lin W, Liu S, Pan H, Tang X, Wang J, Wang W, Yu J, Zhang B, Zhang Q, Zhao H, Zhao H, Zhou J, Gabriel SB, Barry R, Blumenstiel B, Camargo A, Defelice M, Faggart M, Goyette M, Gupta S, Moore J, Nguyen H, Onofrio RC, Parkin M, Roy J, Stahl E, Winchester E, Ziaugra L, Altshuler D, Shen Y, Yao Z, Huang W, Chu X, He Y, Jin L, Liu Y, Shen Y, Sun W, Wang H, Wang Y, Wang Y, Xiong X, Xu L, Waye MM, Tsui SK, Xue H, Wong JT, Galver LM, Fan JB, Gunderson K, Murray SS, Oliphant AR, Chee MS, Montpetit A, Chagnon F, Ferretti V, Leboeuf M, Olivier JF, Phillips MS, Roumy S, Sallée C, Verner A, Hudson TJ, Kwok PY, Cai D, Koboldt DC, Miller RD, Pawlikowska L, Taillon-Miller P, Xiao M, Tsui LC, Mak W, Song YQ, Tam PK, Nakamura Y, Kawaguchi T, Kitamoto T, Morizono T, Nagashima A, Ohnishi Y, Sekine A, Tanaka T, Tsunoda T, Deloukas P, Bird CP, Delgado M, Dermitzakis ET, Gwilliam R, Hunt S, Morrison J, Powell D, Stranger BE, Whittaker P, Bentley DR, Daly MJ, de Bakker PI, Barrett J, Chretien YR, Maller J, McCarroll S, Patterson N, Pe'er I, Price A, Purcell S, Richter DJ, Sabeti P, Saxena R, Schaffner SF, Sham PC, Varilly P, Altshuler D, Stein LD, Krishnan L, Smith AV, Tello-Ruiz MK, Thorisson GA, Chakravarti A, Chen PE, Cutler DJ, Kashuk CS, Lin S, Abecasis GR, Guan W, Li Y, Munro HM, Qin ZS, Thomas DJ, McVean G, Auton A, Bottolo L, Cardin N, Eyheramendy S, Freeman C, Marchini J, Myers S, Spencer C, Stephens M, Donnelly P, Cardon LR, Clarke G, Evans DM, Morris AP, Weir BS, Tsunoda T, Mullikin JC, Sherry ST, Feolo M, Skol A, Zhang H, Zeng C, Zhao H, Matsuda I, Fukushima Y, Macer DR, Suda E, Rotimi CN, Adebamowo CA, Ajayi I, Aniagwu T, Marshall PA, Nkwodimmah C, Royal CD, Leppert MF, Dixon M, Peiffer A, Qiu R, Kent A, Kato K, Niikawa N, Adewole IF, Knoppers BM, Foster MW, Clayton EW, Watkin J, Gibbs RA, Belmont JW, Muzny D, Nazareth L, Sodergren E, Weinstock GM, Wheeler DA, Yakub I, Gabriel SB, Onofrio RC, Richter DJ, Ziaugra L, Birren BW, Daly MJ, Altshuler D, Wilson RK, Fulton LL, Rogers J, Burton J, Carter NP, Clee CM, Griffiths M, Jones MC, McLay K, Plumb RW, Ross MT, Sims SK, Willey DL, Chen Z, Han H, Kang L, Godbout M, Wallenburg JC, L'Archevêque P, Bellemare G, Saeki K, Wang H, An D, Fu H, Li Q, Wang Z, Wang R, Holden AL, Brooks LD, McEwen JE, Guyer MS, Wang VO, Peterson JL, Shi M, Spiegel J, Sung LM, Zacharia LF, Collins FS, Kennedy K, Jamieson R, Stewart J. 2007. A second generation human haplotype map of over 3.1 million SNPs. Nature 449:851–861.

Jirtle RL, Skinner MK. 2007. Environmental epigenomics and disease susceptibility. Nat Rev Genet 8:253–262.

Kendler KS, Neale MC. 2010. Endophenotype: a conceptual analysis. Mol Psychiatry 15:789–797.

Klein RJ, Zeiss C, Chew EY, Tsai J-Y, Sackler RS, Haynes C, Henning AK, SanGiovanni JP, Mane SM, Mayne ST, et al. 2005. Complement factor H polymorphism in age-related macular degeneration. Science 308:385–389.

Kooperberg C, Ruczinski I. 2005. Identifying interacting SNPs using Monte Carlo logic regression. Genet Epidemiol 28:157–170.

Kooperberg C, Ruczinski I, LeBlanc ML, Hsu L. 2001. Sequence analysis using logic regression. Genet Epidemiol 21:S626–S631.

Kraft P, Hunter D. 2009. The challenge of assessing complex gene-gene and gene-environment interactions. In: Khoury M, Bedrosian S, Gwinn M, Higgins J, Ioannidis J, Little J, editors. Human Genome Epidemiology, 2nd edition. New York: Oxford University Press.

Lander ES. 2011. Initial impact of the sequencing of the human genome. Nature 470:187–197.

Langholz B, Borgan Ø 1995. Counter-matching: a stratified nested case-control sampling method. Biometrika 82:69–79.

Laurie CC, Doheny KF, Mirel DB, Pugh EW, Bierut LJ, Bhangale T, Boehm F, Caporaso NE, Cornelis MC, Edenberg HJ, et al. 2010. Quality control and quality assurance in genotypic data for genome-wide association studies. Genet Epidemiol 34:591–602.

Li D, Conti DV. 2009. Detecting gene-environment interactions using a combined case-only and case-control approach. Am J Epidemiol 169:497–504.

Li B, Leal SM. 2008. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. Am J Hum Genet 83:311–321.

Liu DJ, Leal SM. 2010. A novel adaptive method for the analysis of next-generation sequencing data to detect complex trait associations with rare variants due to gene main effects and interactions. PLoS Genet 6:e1001156.

Liu Y, Athanasiadis G, Weale ME. 2008. A survey of genetic simulation software for population and epidemiological studies. Hum Genomics 3:79–86.

Madsen BE, Browning SR. 2009. A groupwise association test for rare mutations using a weighted sum statistic. PLoS Genet 5:e1000384.

Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, et al. 2009. Finding the missing heritability of complex diseases. Nature 461:747–753.

Mardis ER. 2010. The $1,000 genome, the $100,000 analysis? Genome Med 2:84.

Milne R, Gaudet M, Spurdle A, Fasching P, Couch F, Benitez J, Arias Perez JI, Zamora MP, Malats N, dos Santos Silva I, et al. 2010. Assessing interactions between the associations of common genetic susceptibility variants, reproductive history and body mass index with breast cancer risk in the breast cancer association consortium: a combined case-control study. Breast Cancer Res 12:R110.

Moore JH, Barney N, Tsai CT, Chiang FT, Gui J, White BC. 2007. Symbolic modeling of epistasis. Hum Hered 63:120–133.

Moore JH, Andrews PC, Barney N, White BC. 2008. Development and evaluation of an open-ended computational evolution system for the genetic analysis of susceptibility to common human diseases. Lect Notes Comput Sci 4973:129–140.

Moore JH, Asselbergs FW, Williams SM. 2010. Bioinformatics challenges for genome-wide association studies. Bioinformatics 26:445–455.

Moore J, Cowper Sal.Lari R, Hill D, Hibberd P, Madan J. 2011. Human microbiome visualization using 3D technology. Pac Symp Biocomput 16:154–164.

Morgenthaler S, Thilly WG. 2007. A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: a cohort allelic sums test (CAST). Mutat Res/Fundam Mol Mech Mutagen 615:28–56.

Mukherjee B, Chatterjee N. 2008. Exploiting gene-environment independence for analysis of case–control studies: an empirical bayes-type shrinkage estimator to trade-off between bias and efficiency. Biometrics 64:685–694.

Mukherjee B, Ahn J, Gruber SB, Chatterjee N. 2011. Testing gene-environment interaction in large scale case-control association studies: possible choices and comparisons. Am J Epidemiol.

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

Neale BM, Rivas MA, Voight BF, Altshuler D, Devlin B, Orho-Melander M, Kathiresan S, Purcell SM, Roeder K, Daly MJ. 2011. Testing for an unusual distribution of rare variants. PLoS Genet 7:e1001322.

Peng B, Amos CI, Kimmel M. 2007. Forward-time simulations of human populations with complex diseases. PLoS Genet 3:e47.

Pennisi E. 2011. Human genome 10th anniversary. Will computers crash genomics? Science 331:666–668.

Piegorsch WW, Weinberg CR, Taylor JA. 1994. Non-hierarchical logistic models and case-only designs for assessing susceptibility in population-based case-control studies. Stat Med 13:153–162.

Price AL, Kryukov GV, de Bakker PIW, Purcell SM, Staples J, Wei L-J, Sunyaev SR. 2010. Pooled association tests for rare variants in exon-resequencing studies. Am J Hum Genet 86:832–838.

Richter BG, Sexton DP. 2009. Managing and analyzing next-generation sequence data. PLoS Comput Biol 5:e1000369.

Ritchie MD, Bush WS. 2010. Genome simulation: approaches for synthesizing in silico datasets for human genomics. In: Jay CD, Jason HM, editors. Advances in Genetics. New York: Academic Press, p 1–24.

Ritchie MD, Hahn LW, Roodi N, Bailey LR, Dupont WD, Parl FF, Moore JH. 2001. Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. Am J Hum Genet 69:138–147.

Rosenstiel P, Sina C, Franke A, Schreiber S. 2009. Towards a molecular risk map—recent advances on the etiology of inflammatory bowel disease. Semin Immunol 21:334–345.

Rothman KJ, Greenland S, Walker AM. 1980. Concepts of interaction. Am J Epidemiol 112:467–470.

Schadt EE, Lamb J, Yang X, Zhu J, Edwards S, GuhaThakurta D, Sieberts SK, Monks S, Reitman M, Zhang C, et al. 2005. An integrative genomics approach to infer causal associations between gene expression and disease. Nat Genet 37:710–717.

Schwarz DF, König IR, Ziegler A. 2010. On safari to Random Jungle: a fast implementation of Random Forests for high-dimensional data. Bioinformatics 26:1752–1758.

Siemiatycki J, Thomas DC. 1981. Biological models and statistical interactions: an example from multistage carcinogenesis. Int J Epidemiol 10:383–387.

Sinnott-Armstrong N, Greene C, Cancare F, Moore J. 2009. Accelerating epistasis analysis in human genetics with consumer graphics hardware. BMC Res Notes 2:149.

Smith GD, Ebrahim S. 2004. Mendelian randomization: prospects, potentials, and limitations. Int J Epidemiol 33:30–42.

Stein LD. 2010. The case for cloud computing in genome informatics. Genome Biol 11:207.

Stephens M. 2010. A Unified Framework for Testing Multiple Phenotypes for Association with Genetic Variants. Washington DC: The American Society of Human Genetics.

International HapMap Consortium. 2005. A haplotype map of the human genome. Nature 437:1299-1320.

Thomas D. 2004. Statistical Methods in Genetic Epidemiology. New York: Oxford University Press.

Thomas D. 2010. Gene-environment-wide association studies: emerging approaches. Nat Rev Genet 11:259–272.

Thomas DC, Lewinger JP, Murcray CE, Gauderman WJ. 2011. GE-whiz! Ratcheting gene-environment studies up to the whole genome and the whole exposome. Am J Epidemiol.

Thompson WD. 1991. Effect modification and the limits of biological inference from epidemiologic data. J Clin Epidemiol 44:221–232.

Turner S, Armstrong LL, Bradford Y, Carlson CS, Crawford DC, Crenshaw AT, de Andrade M, Doheny KF, Haines JL, Hayes G, et al. 2001. Quality Control Procedures for Genome-Wide Association Studies. New York: Wiley.

Vansteelandt S, Goetgeluk S, Lutz S, Waldman I, Lyon H, Schadt EE, Weiss ST, Lange C. 2009. On the adjustment for covariates in genetic association analysis: a novel, simple principle to infer direct causal effects. Genet Epidemiol 33:394–405.

Vineis P, Perera F. 2007. Molecular epidemiology and biomarkers in etiologic cancer research: the new in light of the old. Cancer Epidemiol Biomarkers Prev 16:1954–1965.

Weinberg CR. 1986. Applicability of the simple independent action model to epidemiologic studies involving two factors and a dichotomous outcome. Am J Epidemiol 123:162–173.

Zawistowski M, Gopalakrishnan S, Ding J, Li Y, Grimm S, Zöllner S. 2010. Extending rare-variant testing strategies: analysis of noncoding sequence and imputed genotypes. Am J Hum Genet 87:604–617.

Zeggini E. 2011. Next-generation association studies for complex traits. Nat Genet 43:287–288.

Zhang Y, Liu JS. 2007. Bayesian inference of epistatic interactions in case-control studies. Nat Genet 39:1167–1173.

Zhang J, Chiodini R, Badr A, Zhang G. 2011. The impact of next-generation sequencing on genomics. J Genet Genomics 38:95–109.

Zhou H, Sehl ME, Sinsheimer JS, Lange K. 2010. Association screening of common and rare genetic variants by penalized regression. Bioinformatics 26:2375–2382.