Title
Genetic Simulation Tools for Post-Genome Wide Association Studies of Complex Diseases

Journal
Genetic Epidemiology, 39(1)

ISSN
0741-0395

Authors
Chen, Huann-Sheng
Hutter, Carolyn M
Mechanic, Leah E
et al.

Publication Date
2015

DOI
10.1002/gepi.21870

Peer reviewed

Published in final edited form as:
Genet Epidemiol. 2015 January ; 39(1): 11–19. doi:10.1002/gepi.21870.

Genetic Simulation Tools for Post-Genome Wide Association Studies of Complex Diseases

Huann-Sheng Chen[#1], **Carolyn M. Hutter**[#2], **Leah E. Mechanic**[#3], **Christopher I. Amos**[4], **Vineet Bafna**[5], **Elizabeth R. Hauser**[6], **Ryan D. Hernandez**[7], **Chun Li**[8], **David A. Liberles**[9], **Kimberly McAllister**[10], **Jason H. Moore**[11], **Dina N. Paltoo**[12], **George J. Papanicolaou**[13], **Bo Peng**[14], **Marylyn D. Ritchie**[15], **Gabriel Rosenfeld**[1], **John S. Witte**[16], **Elizabeth M. Gillanders**[#3], and **Eric J. Feuer**[#1]

[1] Surveillance Research Program, Division of Cancer Control and Population Sciences, National Cancer Institute, NIH, Bethesda, MD 20892

[2] Division of Genomic Medicine, National Human Genome Research Institute, NIH, Bethesda, MD 20892

[3] Epidemiology and Genomics Research Program, Division of Cancer Control and Population Sciences, National Cancer Institute, NIH, Bethesda, MD 20892

[4] Division of Community, Family Medicine, Dartmouth College, Lebanon, NH 03755

[5] Department of Computer Science and Engineering, University of California, San Diego, La Jolla, CA 92093

[6] Duke Molecular Physiology Institute, Duke University, Durham, NC 27710

[7] Department of Bioengineering and Therapeutic Sciences, University of California, San Francisco, CA 94143

[8] Department of Biostatistics, Vanderbilt University, Nashville, TN 37235

[9] Department of Molecular Biology, University of Wyoming, Laramie, WY 82071

[10] Susceptibility and Population Health Branch, National Institute of Environmental Health Sciences, NIH, Research Triangle Park, NC 27709

[11] Department of Genetics, Dartmouth College, Lebanon, NH 03755

[12] Office of Director, National Institutes of Health, Bethesda, MD 20892

[13] Division of Cardiovascular Sciences, Prevention and Population Sciences Program, National Heart, Lung, and Blood Institute, National Institutes of Health, Bethesda, MD 20892

[14] Department of Bioinformatics and Computational Biology, University of Texas MD Anderson Cancer Center, Houston, TX 77030

Address Correspondence to: Leah Mechanic, Ph.D., M.P.H. Program Director Division of Cancer Control and Population Sciences National Cancer Institute Epidemiology and Genomics Research Program Host Susceptibility Factors Branch 9609 Medical Center Drive, Room 4E104 Bethesda, MD 20892 mechanil@mail.nih.gov.

Authors have no conflicts of interest to declare.

[15] Department of Biochemistry and Molecular Biology, Pennsylvania State University, University Park, PA 16802

[16] Department of Epidemiology and Biostatistics, University of California, San Francisco, CA 94107

[#] These authors contributed equally to this work.

Abstract

Genetic simulation programs are used to model data under specified assumptions to facilitate the understanding and study of complex genetic systems. Standardized data sets generated using genetic simulation are essential for the development and application of novel analytical tools in genetic epidemiology studies. With continuing advances in high-throughput genomic technologies and generation and analysis of larger, more complex data sets, there is a need for updating current approaches in genetic simulation modeling. To provide a forum to address current and emerging challenges in this area, the National Cancer Institute (NCI) sponsored a workshop, entitled "Genetic Simulation Tools for Post-Genome Wide Association Studies of Complex Diseases" at the National Institutes of Health (NIH) in Bethesda, Maryland on March 11-12, 2014. The goals of the workshop were to: (i) identify opportunities, challenges and resource needs for the development and application of genetic simulation models; (ii) improve the integration of tools for modeling and analysis of simulated data; and (iii) foster collaborations to facilitate development and applications of genetic simulation. During the course of the meeting the group identified challenges and opportunities for the science of simulation, software and methods development, and collaboration. This paper summarizes key discussions at the meeting, and highlights important challenges and opportunities to advance the field of genetic simulation.

Keywords

genetic simulation; rare variants; next generation sequencing; complex phenotypes; computational resources

Introduction

Genetic simulation, computer modeling of genetic data under specified assumptions, had been widely used to study the impact of historical demographic and genetic factors on the genetic composition of present human populations; to develop and validate statistical methods to detect susceptibility genes for human genetic diseases; and to determine the most powerful study designs and statistical tests for identifying putative causal variants for traits and diseases (for review [Liu, et al. 2008; Ritchie and Bush 2010]). Simulation is important to allow for the evaluation of methods and investigation of models in a controlled, *in silico* experiment where the researcher can directly vary conditions, an element that is otherwise void in human population based studies of genetics.

The landscape of epidemiology, and genetic epidemiology specifically, is being transformed by the availability of affordable high throughput sequencing and development of other "omic" technologies [Lam, et al. 2013]. These technologies create an opportunity to explore

more fully the genome and investigate novel hypotheses about contributors to complex diseases. Moreover, these technologies result in an explosion of data [Grossman and White 2012]. The current bottleneck in research is no longer the generation of large scale genetic data, but the availability of computational tools to effectively analyze the data [Green and Guyer 2011] as well as the means to compare and contrast new tools. With this diversity of new, more complex data types and larger datasets, come new challenges identifying the ideal analytical methods to study these hypotheses. Genetic simulations play a critical role in developing these methods.

Simulated datasets provide a powerful resource for researchers, but challenges remain in advancing the field to meet the needs of an ever expanding array of new technology and data. Genomic data types vary in structure and type, e.g. haploid and diploid sequences, sex chromosomes, mitochondrial DNA, single nucleotide polymorphisms (SNPs), microsatellite markers, insertions, deletions, inversions, large indels, structural variations, and copy number variations in DNA and RNA sequence variation or variation in protein sequence. Simulated traits and/or phenotypes include, but are not limited to, disease status, and quantitative traits. Along with genomic data and phenotypic outcomes, simulation models may include environmental factors or exposures. To assist users and developers in comparing different simulators and selecting the one which is most appropriate for the scientific question being asked, the NCI created the Genetic Simulation Resources (GSR) website [Peng, et al. 2013], a catalogue of genetic simulation programs where programs are described using a series of standardized attributes.

One approach for advancing the science of simulation in a systematic manner is to establish forums for collaboration among simulation modelers [Mechanic, et al. 2012]. To provide one such forum, the National Cancer Institute (NCI) sponsored a workshop, entitled “Genetic Simulation Tools for Post-Genome Wide Association Studies of Complex Diseases” in Bethesda, Maryland on March 11-12, 2014. The aim of the meeting was to bring a broad spectrum of population geneticists, genetic epidemiologists, and computational scientists together to evaluate and advance the use of simulation models for genomic studies in epidemiologic analysis studies of complex diseases. Specifically, the goals of the workshop were: (i) identify opportunities, challenges and resource needs for the development and application of genetics simulation models; (ii) improve the integration of tools for simulation models and analysis of simulated data; (iii) foster collaborations to facilitate development and applications of genetic simulations. During the meeting, the group discussed challenges and opportunities in the science of simulation and development of software and methods for genetic simulation. In addition, the group highlighted opportunities to foster collaboration to facilitate addressing challenges for the scientific discipline and development of methods and software. This paper summarizes the key discussions at the meeting and opportunities to advance the field (Table I).

The Science of Genetic Simulation

State of the Science

Genetic simulation is a useful tool for improving our understanding of the genetic basis of cancer and other complex diseases [Hoban, et al. 2011; Ritchie and Bush 2010]. In contrast

to experimental data, where the "truth" is unknown, simulated data sets are created with defined attributes. These *in silico* data sets generate data under specific assumptions, and can be used to validate statistical methods and compare the power of different methods. Simulations can also be used to evaluate conditions, e.g. evolutionary history, that could have given rise to current observations in genetic data. By performing simulations under different conditions, inference can be drawn based on parameters and assumptions for simulations that best matches the empirical data. In addition, simulations may be used to examine how modifications to a system change the attributes of the datasets, looking both forward and backwards over time. Finally, simulations are useful in situations where data are unavailable, or too expensive to obtain empirically, such as genotypes in large pedigrees or studies of rare diseases. Given the broad range of applications, the complexity of the models and the breadth of scientific knowledge needed, genetic simulation should not be viewed as simply a useful tool in genetic studies, but should also be viewed as a developing scientific discipline of its own. Like any scientific discipline, it needs to progress systematically, learning from mistakes and setting standards for good practices. For example, the field of mathematical and simulation models that facilitate estimation of health care decisions has recently published its second report of good modeling research practices [Caro, et al. 2012].

To generate simulated genetic variation in humans, several approaches are used, including: (1) backwards (coalescent) approaches, where the ancestral conditions are modeled from the observed present conditions; (2) forward time simulations, where initial conditions are specified and simulation proceeds forward in time allowing for population pressures; (3) sideways simulation, which uses existing data and resamples it; (4) theoretical simulation, which simulates genotypes and phenotypes from theoretical distribution; (5) gene dropping, which passes ancestral genotypes along a fixed pedigree; (6) phylogenetic simulation, which evolves one or more genomic sequences along a fixed or dynamically generated phylogenetic tree. In a study examining the number of publications, or applications, which cited simulation programs in the GSR catalogue, backward and forward methods appeared to be the most commonly used simulations, representing about 57% and 15% of applications. Sideways, theoretical, and gene dropping methods excel in certain application areas and were also frequently used for simple home-made simulations. Phylogenetic simulation was utilized less frequently for genetic epidemiologic studies [Peng, et al. 2014]. Each simulation method has pros and cons, therefore it is important to consider which method is most appropriate for the scientific question being asked and the type of data structures being simulated [Hoban, et al. 2011; Liu, et al. 2008; Ritchie and Bush 2010].

Knowledge Gaps and Opportunities

Population-level simulations of cancer phenotypes have been used to study characteristics of interventions, such as tobacco control policies, and screening tests. For example, the Cancer Intervention and Surveillance Modeling Network (CISNET), uses micro-simulations and comparative modeling to improve understanding of cancer control interventions in prevention, screening and treatment (http://cisnet.cancer.gov/). However, to date the incorporation of genetics into CISNET models has been limited and there are relatively few applications that fully integrate detailed phenotype simulation with genetic data. In part, this

is because the joint simulation of genotypes and phenotypes can be very complicated, particularly when considering the genetic architecture of complex diseases. One example of joint simulation of genotypes and phenotypes is the dataset produced for Genetic Analysis Workshop 16 (GAW16) meeting [Kraja, et al. 2009]. For GAW 16, empirical patterns of SNP data from the Framingham Heart Study were used for simulations that jointly modeled SNPs, longitudinal cardiovascular phenotypes and environmental factors including diet, smoking and medication use. In another study, various genetic architectures were simulated under multiple disease models, or different phenotypes, for a complex disease [Agarwala, et al. 2013]. Another example of joint genotype-phenotype simulation used a model of quantitative traits subject to natural selection to explore potential contribution of rare variation to phenotypes [Thornton, et al. 2013]. Importantly, the nature of complex traits requires the development of simulation models capable of incorporating the biological heterogeneity of phenotypes, including longitudinal outcomes, time-dependent variables, environmentally modified traits and endophenotypes (Table I, 1.1).

In addition to improved joint modeling of genotype and phenotype, it is also important to consider how to best incorporate next generation sequencing technology and other “-omics” into models and genetic simulations [Mechanic, et al. 2012]. We need to incorporate realistic probability models of human disease into large-scale genomic data, expanding the scope of variation to chromosome wide and genome-wide simulations. We cannot assume that simulations of small numbers of genomic variants (i.e. range of 20 SNPs) [Ritchie and Bush 2010] will apply in these large-scale settings (i.e. range of 500,000 SNPs, or over 2,000,000 rare variants). Such an interpretation ignores the possibility that the type I error rate and power may change as the number of polymorphisms and samples increase and the computational burden of performing the method on a much larger data set. Most of our discoveries in genetic epidemiology, including GWAS findings for cancer [Hindorff, et al. 2011], follow standard genetic models under dominant, recessive and additive effects. However, the true genetic models underlying these traits likely involve epistatic effects, including gene-gene and gene-environment interactions, pairwise and higher order interactions and other complexities. For example, the GAMETES program was developed to simulate epistatic genetic models [Urbanowicz, et al. 2012].

In developing complex genetic models, it is important to consider the population genetic forces that have shaped modern human populations [Tennessen, et al. 2012]. Demography and natural selection have both had dramatic impacts on the patterns of genetic diversity, and on the frequency and spectrum of functional variants [Lohmueller 2013; Maher, et al. 2012]. For example, rare variants show different patterns and associations, compared to more common variants [Coventry, et al. 2010]. This will impact both the analysis methods [Zuk, et al. 2014], but also the simulation methods needed for studies of rare variation and other genetic variation (Table I, 1.2). Moreover, as we begin to model new technologies, we need to consider their error distributions. For example, with next generation sequencing, a realistic simulator should mimic how the data is generated, including library preparation, alignment and variant calling. ASAP is a program that incorporates these variables [Torstenson, et al. 2013]. Much remains to be learned about the characteristics of error distributions at each of these stages; simulations could help us understand how error rates from sequencing will impact genetic epidemiology studies (Table I; 1.3).

Much emphasis in genetic simulation research has been on germline SNPs, but other types of genetic variation are implicated in complex diseases. While simple models for small insertions and deletions can be simulated jointly with SNPs [Hernandez 2008] other types of structural variants (SVs) are increasingly being shown to impact disease risk. SVs arise under different mechanisms [Xing, et al. 2009], and multiple theories have been proposed to explain the complex genomic rearrangements observed in cancer, including chromothripsis [Stephens, et al. 2011] and breakage-fusion-bridge [Zakov, et al. 2013]. In order to understand the mechanistic basis of SVs, as well as the relationship between SVs and disease, we need to simulate the frequency, size and underlying causes for SVs. We also need to simulate related genetic factors, such as fragile regions, 3-D chromosome conformation and mapping artifacts, which add complexity to studies of SVs and SV disease associations. In addition to germline DNA variation, to better elucidate the underlying biological mechanisms of complex diseases and leverage improved molecular phenotyping approaches, simulation tools will be needed to inform analysis and interpretation of these other molecular data types. Simulations should be expanded to more complex models beyond DNA variation, including transcription factor networks, epigenetic and epigenomic factors, gene-expression, pathways and protein sequences (Table I, 1.4).

Many of these knowledge gaps contribute to the use of overly simplified or unrealistic simulation models. As noted above, it is important to consider higher order epistatic relationships and population genetic factors (Table I; 1.2, 1.5). BioSim is an example of a more complex simulator that incorporates underlying biological, pathologic and pharmacological processes (Moore, J.M. personal communication, 2014). At a biological level, realistic models can include a description of a genotype-phenotype map wrapped in a population genetic context. For example, nonsynonymous SNPs can be described by their effects on protein stability and protein-protein interaction using expectations from physical chemistry coupled to population genetic descriptions of fixation probabilities (for example [Grahnen and Liberles 2012; Liberles, et al. 2012]). Another example is using systems biochemistry to generate explicit definitions of phenotype that may ultimately relate to both human and infectious disease [Savageau and Fasani 2009], moving further from common assumptions of the independence of action of individual SNPs.

Despite the need for incorporation of more realism and complexity into simulation models to capture underlying biological complexity, if a model is too complicated it may pose issues for implementation and interpretation. It is important to ensure that increased complexity is actually moving simulations toward the true underlying biology, rather than adding unnecessary complexity. One strategy is to consider Approximate Bayesian Computation with relevant summary statistics [Beaumont, et al. 2002]. Another concern with the use of complex models is identifiability, both in a statistical sense that parameters are identifiable and in a biological sense, that inference about biological hypotheses is identifiable [Liberles, et al. 2013]. In addition, as we consider large-scale genetic variation, efficiency of the simulation becomes more critical. If a model is too complex, it may be computationally intensive in terms of time and storage needs. This is a particular challenge for next generation sequencing data, which may require simulation and analysis of terabytes of data. Data complexity may be reduced by only analyzing regions of interest, or focusing only on

parts of the data generation and analysis process. However, this strategy requires *a priori* knowledge regarding the critical regions or processes. A key challenge is identifying which aspects of the model are most important to create realistic simulation models for a given disease outcome and scientific question. The field will benefit from coordinated efforts to address this challenge, rather than rely on overly simplistic models and/or proliferation of under-evaluated complex methods.

Meeting attendees identified and discussed these areas of knowledge gaps for genetic simulation and suggested specific research priorities. First, the group recognized the importance of joint simulation of genotypes and realistically complex phenotypes. Second, it was noted that simulation of rare variants is an important area for current study. Given the increasing number of genetic epidemiology studies using whole genome and whole exome sequencing [Helgason, et al. 2013; Morrison, et al. 2013; Tennessen, et al. 2012], there is an opportunity for simulation work in this area to have a large impact on the field. Lastly, the group suggested future work in the area of RNA sequencing (RNAseq). Laboratory methods for RNAseq are still being developed and thus a better understanding of the underlying complexity of RNAseq data is needed before it will be ready for large scale simulations that can appropriately evaluate models and new analytic methods.

Software and Methods Development

Current Challenges

Several challenges were discussed related to genetic simulation software and development of methods. The issues discussed ranged from end-users selecting or developing the appropriate simulator for evaluation of analytical methods and the challenge of comparing evaluations when using different simulators, comparability of simulators, to lack of consistent documentation of the simulation programs themselves and application of these programs by end-users.

There are a large number of genetic simulation programs available for research purposes. The GSR catalogue currently describes 93 different programs for genetic simulation, with 13 additional programs pending entry [Genetic Simulation Resource 2014]. Despite such a large number of genetic simulation programs available, many seem to be rarely utilized. An evaluation of five recent issues of the journal Genetic Epidemiology, noted that only 8 out of 36 articles which included simulated data, used existing genetic simulators, or simulators already catalogued by GSR [Peng, et al. 2014]. These results may suggest that only a small fraction of the scope and diversity of genetic simulations are covered by existing simulators. It may also suggest that investigators often create a new simulator, as opposed to using an existing tool, resulting in potential replication of effort and redundancy of tools. The availability of a large number of programs makes it a challenge for end-users to select the appropriate tool for their research.

Several factors are likely contributing to the proliferation of programs. Every simulation design may have different requirements. For example, simulators must consider data structure (e.g. SNPs, next generation sequencing, RNA sequencing), study design and pedigree structure (e.g. case-control, case-only, trios, extended families), and population

characteristics (e.g. homogenous, admixed, non-random mating, different selection pressures). Therefore, several different programs may be required to address multiple scientific needs. Moreover, when researchers develop their own programs, it is easier to control all aspects of the simulation and develop a simulator which precisely addresses research requirements. With such a large number of distinct simulators, results obtained using these different simulators are difficult to compare. However, it is also possible that existing simulation programs could be used more broadly.

Another factor contributing to the large number of simulators is that many genetic simulation programs, including those catalogued within GSR, are no longer actively maintained. The field of genetic epidemiology has evolved rapidly over the past several years with a focus shifting from linkage studies, to candidate gene association studies, to genome wide association studies (GWAS) and now using next generation sequencing technologies in families and population based studies [Mechanic, et al. 2012]. The shift in research focus results in tools that are outdated or no longer of interest to the investigator or research community. Another reason for the lack of maintenance is that a software package may be replaced by better alternatives, including improved design or implementation. In addition, authors of software packages may have graduated or left the academic position where they created the package. Furthermore, maintenance of simulation programs is time consuming and difficult to have supported through traditional research grant funding. It is often considered easier to obtain funding for creation of a simulator as compared to maintaining the program. Therefore, many researchers may lack the motivation and resources to maintain older software packages.

Another challenge in using existing genetic simulation programs is limited documentation [Mechanic, et al. 2012] and user support. The documentation of simulation programs often uses implicit and domain-specific terms and assumptions, fails to include sufficient examples regarding the application of the simulator, and lacks a detailed description of simulation methods. Without sufficient documentation, users other than the program creator may have difficulty using the programs. This can lead users to develop their own tool, as described above. More importantly, lack of transparency about assumptions of simulation programs, could lead to inaccurate conclusions when using these programs. There is a need in the field to understand the breadth and appropriateness of genetic simulation models, specifically describing the limitations of a model (i.e. situations under which a model no longer holds, describing the parameter space and assumptions), and improved communications regarding these assumptions. Notably, there is a dearth of standards for documentation of genetic simulation programs. The GSR catalogue was designed to address issues with limited documentation and to improve transparency regarding model assumptions by characterizing the simulators according to well-defined features and incorporating user comments on software assumptions [Mechanic, et al. 2012; Peng, et al. 2013]. Nonetheless, more work is needed in this area and opportunities are outlined below.

Genetic simulations are often used to evaluate and compare analytical tools. However, there is a danger in creating a simulation specifically to test a new method, as results can be self-fulfilling prophecies due to assumptions embedded in the simulation [Mechanic, et al. 2012]. Moreover, creating a new simulator for the purposes of testing a method or cherry-picking

from the large number of different simulators may lead to unfair comparisons of analytical methods and potentially results in optimistic interpretations regarding method performance. Assessment of analytical methods should be based on assumptions that are aligned with the research question of interest; for example, we should not evaluate methods for rare variants under simulation assumptions appropriate for common variation. In genetic simulation of rare genetic variants, many models assume a large percentage of highly penetrant mutations, resulting in optimistic power assumptions. Even when we are using appropriate simulations, there is likely somewhere in the parameter space where the method does not work or performs poorly. Researchers need to identify and describe those situations where methods perform poorly as well as when methods perform best.

Opportunities for Methods and Software Development

Several of the challenges regarding comparability of genetic simulation programs and models may be addressed by development of an ontology to describe genetic simulation (Table I, 2.1). An ontology is a shared or controlled vocabulary which describes items or concepts within a domain, and can represent relationships between terms in the domain [Chen, et al. 2013]. Creation and implementation of an ontology in the field of genetic simulation could help clarify the assumptions and methodologies incorporated into different genetic simulation tools. Moreover, a standardized vocabulary could facilitate comparisons between simulation programs, result in improved consistency in descriptions of simulation programs, and allow for easier integration of disparate programs. HuPSON [Gundel, et al. 2013] is an example ontology developed to describe computer simulations in human physiology, which may serve as a framework to consider a genetic simulation ontology. Development of such an ontology would require collaboration between genetic simulation and bioinformatics ontology communities and may foster further interactions between these research communities. A workshop focused on developing ontology for genetic simulation may facilitate these collaborations.

Another suggestion was to develop guidelines and standards for reporting on genetic simulation, including documentation of programs, description of programs in journal articles, and reporting by end-users when using simulation programs for applications (Table I; 2.2). One strategy is to outline best practices for documentation of simulators and provide standard templates for technical documentation, similar to the CISNET model profiles [Habbema, et al. 2006]. In addition, funding support for documentation could provide motivation for developers to improve documentation. As well as documentation of the software itself, guidance for reporting in publications for developers and end users of simulation programs was recommended. To ensure sufficient adoption of the guidelines, agreement by journal editorial boards is required. An example which may serve as a guide are the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) guidelines, where a checklist was created for authors, reviewers, journal editors, and readers with a goal of improving reporting in epidemiology studies [Vandenbroucke, et al. 2007]. Guidelines for reporting and documentation of genetic simulation may help educate reviewers, journal editors, and authors about the necessity of reporting limitations of genetic simulation models and results obtained using these tools. Follow up discussions are needed to develop the recommended guidelines for genetic simulation reporting. These suggestions

could facilitate transparency in model assumptions, assist users in selecting the appropriate programs, aid reviewers in evaluating use of genetic simulators, and facilitate replication and comparisons by other researchers.

Given the large number and diversity of simulators and quantity of outdated programs, appraising the different tools could benefit end-users and foster the science of simulation. One suggested approach is to add a GSR-certification feature to the GSR catalogue (Table I, 2.3). Certification could be based on a defined checklist of features including whether programs were open source, had user-friendly implementation (or provided an installer for supported platforms), provided adequate documentation, and used standard data input and output formats. In addition, older programs which had not been updated could be tagged within or even removed from the resource. Simulation programs may also be evaluated based on the utilization of these tools or number of downloads. However, these criteria are not currently captured within the GSR catalogue. Code repositories, such as SourceForge (http://sourceforge.net/) and GitHub (http://github.com/), record numbers of downloads and could become sources for examining the utilization of different simulators.

In addition to appraising the different simulation tools, more efforts are needed to support the maintenance of simulation programs (Table I, 2.4). Researchers developing these programs could explore funding opportunities such as Small Business Innovation Research (SBIR) program applications or form collaborations with companies which may have more experience than academic researchers disseminating and maintaining software packages. Recent NIH initiatives, such as the NIH Big Data to Knowledge (BD2K) and the NCI Informatics Technology for Cancer Research (ITCR), emphasize sustainability and dissemination of software tools. Opportunities such as these could be explored by genetic simulation researchers. Moreover, encouraging software developers to deposit software into code repositories could provide options for other developers to support software maintenance and enable commenting on the software tools (Table I, 2.5).

It is important to properly recognize the contributions of the researchers who deposit software and code. Repositories need to include mechanisms for users to cite any software and code that they access, thereby giving appropriate credit to the developers. Furthermore, maintaining and updating software packages requires intellectual engagement and time. Research faculty, and those in equivalent positions, should be evaluated not only on their scientific publication of novel methods, but also on their depositing, updating and maintaining software packages. Academic evaluation of software developers will be facilitated by citation mechanisms that help demonstrate the usefulness of their software to the larger community.

Beyond documentation and maintenance, simulation developers should consider the requirements of end-users (Table I, 2.6). Many tools catalogued in GSR are primarily command line interfaces and challenging for non-programmers to implement. One strategy to engage end-users is to incorporate a graphical user interface (GUI). While GUIs are potentially useful for simpler genetic simulations, and can make analysis easier for less statistically sophisticated end-users, GUIs may not be feasible for more complicated simulators. Additional concerns with GUI implementations are that these could result in

careless use, without appropriate documentation of steps implemented by end-users. They also do not work for large-scale or batch analysis, and inhibit utilization of computer clusters required for advanced simulators. A common strategy to address this problem is to provide a command line interface in conjunction with a GUI that both generates and executes the command line code, allowing for advanced applications and facilitating reproducibility by all users.

Since the large number of different simulators makes it difficult to compare tools, another suggestion was for the genetic simulation community to recommend a core set of genetic simulation programs for the most common research questions (Table I, 2.7). The challenge with this strategy is that it is difficult to determine which simulators are optimal. Methods have advantages and disadvantages in specific situations, and no single simulator will be optimal in all circumstances. As an alternative, a small number of multi-use programs could be developed to support flexible, broad-based models (Table I, 2.8). This strategy could foster a diversity of models in a small number of high quality programs. Importantly, if the community advocates use of specific simulation tools, these must be widely available and maintain good documentation.

Instead of advocating for specific simulators, another strategy could be to create a comprehensive framework, or genetic simulation server, for the development and distribution of genetic simulation programs and data (Table I, 2.9). This framework would depend on development of an ontology for genetic simulation to facilitate creation of Application Program Interfaces (APIs). In this framework, several different simulation methods could be linked to an analysis server. Through a web interface, a user might then perform an analysis on the server to generate simulated data. Notably, the simulated data would be available to the user, but also stored for other users. This server could store simulated data sets, or even selected benchmark data sets. Users could search for different simulation tools or data sets using the web interface. Because all analyses are stored, it will be easier for researchers to compare results. Furthermore, the server could store common tasks as modules for all users, such as generating random errors for sequencing data. By establishing common APIs and directing input and output formats, it would be possible to add different modules or for other simulation tools to interoperate with this server. Many advantages to such an approach were recognized. However, several concerns were raised regarding this approach, including maintenance costs, storage costs, required processing power, and intellectual property issues.

A more comprehensive evaluation of simulation programs would require comparing these programs to established benchmarks. In a comparison of several coalescent simulation programs in their ability to model recombination hotspots in genomic sequences, two simulators failed to simulate 500 samples of 5MB sequences and the other three demonstrated varying accuracy in position and intensity of simulated recombination hotspots [Yang, et al. 2014]. In another study of simulators for GWAS, datasets simulated by 5 simulators vary greatly in linkage disequilibrium and minor allele frequency patterns [Xu, et al. 2013]. Therefore, it is unclear how results of subsequent analyses compare using different simulation programs. One possible approach to address is to create a test suite for each type of simulation to compare these programs. This type of comparative modeling

would be a substantial effort by the research community, but would help assure comparability among studies and improve the scientific value of simulation (Table I, 2.10).

While many of these suggestions would improve the utilization of genetic simulators, many end users would prefer to access the simulated data sets themselves. Benchmark simulated data sets would provide standards to compare analysis tools [Mechanic, et al. 2012]. The data sets could be created for typical applications and provided with detailed descriptions regarding the method of data simulation. This group recommended renewing efforts to identify benchmark datasets (Table I, 2.11). Benchmark simulations have been provided by the Genetic Analysis Workshop (GAW) [Ziegler, et al. 2011], but are often focused on specific simulations that have been the topics of particular GAW meetings. More researchers should be encouraged to make simulated datasets publicly available (Table I, 2.12), perhaps at the time of publication. The use of independently generated simulated datasets would facilitate fair comparison between statistical methods and help identify the strength and weakness of methods.

Fostering Collaboration

Opportunities to Advance Science and Methods of Genetic Simulation

Meeting participants recognized that fostering collaboration is critical to addressing challenges described regarding the science, software, and development of methods for genetic simulation. Moreover, many of the suggested opportunities described above would be best accomplished as teams, instead of individual groups working in silos.

For genetic simulation tools to be used most effectively, more communication is required between simulation developers, end-users of simulation programs, and researchers from other communities (e.g. epidemiologists, physicians, engineers, statisticians and others) (Table I, 3.1). Increased engagement within and outside scientific disciplines could clarify assumptions and increase appropriate applications of simulation tools. As genetic epidemiology progresses with use of simulation, communication with the parallel fields of population genetics, comparative genomics, and molecular evolution, that are asking many of the same questions about the nature of the genotype-phenotype map will be important. These fields have developed with very different toolboxes, assumptions, and inference strategies and cross fertilization would be beneficial to both communities. Moreover, genetic simulators would likely benefit from a deeper understanding of how simulations are effectively used in other fields, including protein biology, physics and pharmacology. Increased collaboration and engagement with these fields would build bridges among groups with different skill sets and allow for sharing of expertise. Determining how to describe genetic simulation to other fields, could result in more appreciation by other research communities. These collaborations should start from the planning stages of a study could result in more appreciation and understanding of what simulation is and how it can be used. Importantly, the requirements and uses for simulated data need to be more broadly communicated, because experimental data is not a sufficient control for evaluating analytical methods. Further, by engaging end-users in the discussion, they will better understand when and how simulation tools may be used most appropriately and limitations of these tools. One strategy to improve engagement could be to create a consortium dedicated to fostering the

science of genetic simulation (Table I, 3.2). Another suggestion was to create an on-line forum for discussion about genetic simulation (Table I, 3.3).

CISNET may serve as a model for collaboration for the genetic simulation community. The approach innovated by the CISNET group is systematic comparative modeling with central questions to be addressed by groups collaboratively with a common set of inputs and outputs. Reproducibility of simulation results across models adds credibility to results and differences in results are used to identify potential areas for further study. The CISNET modeling network has interacted with several other agencies including United States Preventive Services Task Force (USPSTA), Agency for Health Care Research and Quality (AHRQ), and the Centers for Disease Control (CDC). Likewise, a goal for the genetic simulation community could be to increase interaction with broader scientific communities. In addition to using the CISNET network as a model, genetic simulation modelers could explore opportunities to link the genetic simulation community into CISNET to examine how genetic data informs CISNET models (Table I, 3.4).

Another strategy to foster the community of genetic simulations is a focus on education and training. Supporting the development of a curriculum in genetic simulation would be an investment in the future of genetic simulations, and would help set and shape expectations for simulation development. One suggestion was for a course to focus on simulation as a science and understanding the role and importance of simulations, and, distinct from simulation software programming itself, the assessment of simulation methods. The content of educational material could emphasize end-user needs to foster more interaction between end-users and developers. Importantly, these materials may contribute to more sophisticated knowledge of simulation methods by both developers and end-users. In addition, the course could highlight requirements for documentation of tools. The curriculum could be developed in the form of a massive open online course (MOOC), available to unlimited participation on the web. By making the educational materials accessible on line, they may more likely be adopted by the research community. Another suggestion was to create training based on the “Google Summer of Code” (http://www.google-melange.com/), where students work together to support open source development projects. At a simulation summer of code, students could be trained on programming and best practices for genetic simulation, accurate documentation, and evaluation of programs against benchmark test suite (Table I, 3.5).

Summary and Conclusions

Genetic simulations are essential for the study of the genetics of complex diseases. As the genetic epidemiology field develops new analytical methods to adapt to the ever-evolving data landscape, simulations and simulated data sets are critical for validation of these novel methods. Natural, biological data cannot be used for the validation and comparison of methods. While these approaches may be used to determine if a method finds what is already known, any other finding is subjective because the truth in biological systems is unknown.

During the course of the “Genetic Simulation Tools for Post-Genome Wide Association Studies of Complex Diseases” meeting, participants identified many challenges and

opportunities to advance the field and science of simulations. Several knowledge gaps in the science of simulations were identified as priorities, specifically recognizing the necessity of realism in simulation models while balancing the need for efficiency of implementation. The scientific priorities suggested were improved modeling of phenotypes, next generation sequencing and rare variant data, and RNA sequencing data. In addition, comparison and evaluation of different simulators and studies performed using these tools remains a challenge. While GSR is a first step to attempt to address this challenge, by providing detailed descriptions of genetic simulation programs, more work is needed including standards for reporting and documentation, appraising or evaluating genetic simulation programs, and identifying benchmark datasets.

Finally, recognizing the need for the genetic simulation community to work together with each other, end-users, and other fields, the main opportunity which may further the field is to foster collaborations, by forming groups to address common problems, educate developers and users of simulation programs and improving communications. Education of the next generation of genetic epidemiologists in the best practices of genetic simulation may be one of the best strategies to impact this discipline in the future.

Acknowledgements

We acknowledge the National Cancer Institute (NCI) Division of Cancer Control and Population Sciences (DCCPS) Surveillance Research Program (SRP) for providing funds to support this meeting.

References

Agarwala V, Flannick J, Sunyaev S, Altshuler D. Evaluating empirical bounds on complex disease genetic architecture. Nat Genet. 2013; 45(12):1418–27. [PubMed: 24141362]

Beaumont MA, Zhang W, Balding DJ. Approximate Bayesian computation in population genetics. Genetics. 2002; 162(4):2025–35. [PubMed: 12524368]

Caro JJ, Briggs AH, Siebert U, Kuntz KM. Modeling good research practices--overview: a report of the ISPOR-SMDM Modeling Good Research Practices Task Force-1. Med Decis Making. 2012; 32(5):667–77. [PubMed: 22990082]

Chen H, Yu T, Chen JY. Semantic Web meets Integrative Biology: a survey. Brief Bioinform. 2013; 14(1):109–25. [PubMed: 22492191]

Coventry A, Bull-Otterson LM, Liu X, Clark AG, Maxwell TJ, Crosby J, Hixson JE, Rea TJ, Muzny DM, Lewis LR, Wheeler DA, Sabo A, Lusk C, Weiss KG, Akbar H, Cree A, Hawes AC, Newsham I, Varghese RT, Villasana D, Gross S, Joshi V, Santibanez J, Morgan M, Chang K, Iv WH, Templeton AR, Boerwinkle E, Gibbs R, Sing CF. Deep resequencing reveals excess rare recent variants consistent with explosive population growth. Nat Commun. 2010; 1:131. [PubMed: 21119644]

Genetic Simulation Resource. Genetic Simulation Resources (GSR). 2014

Grahnen JA, Liberles DA. CASS: Protein sequence simulation with explicit genotypephenotype mapping. Trends in Evolutionary Biology. 2012; 4(1):e9.

Green ED, Guyer MS. Charting a course for genomic medicine from base pairs to bedside. Nature. 2011; 470(7333):204–13. [PubMed: 21307933]

Grossman RL, White KP. A vision for a biomedical cloud. J Intern Med. 2012; 271(2):122–30. [PubMed: 22142244]

Gundel M, Younesi E, Malhotra A, Wang J, Li H, Zhang B, de Bono B, Mevissen HT, Hofmann-Apitius M. HuPSON: the human physiology simulation ontology. J Biomed Semantics. 2013; 4(1):35. [PubMed: 24267822]

Habbema JD, Schechter CB, Cronin KA, Clarke LD, Feuer EJ. Modeling cancer natural history, epidemiology, and control: reflections on the CISNET breast group experience. J Natl Cancer Inst Monogr. 2006; (36):122–6. [PubMed: 17032902]

Helgason H, Sulem P, Duvvari MR, Luo H, Thorleifsson G, Stefansson H, Jonsdottir I, Masson G, Gudbjartsson DF, Walters GB, Magnusson OT, Kong A, Rafnar T, Kiemeney LA, Schoenmaker-Koller FE, Zhao L, Boon CJ, Song Y, Fauser S, Pei M, Ristau T, Patel S, Liakopoulos S, van de Ven JP, Hoyng CB, Ferreyra H, Duan Y, Bernstein PS, Geirsdottir A, Helgadottir G, Stefansson E, den Hollander AI, Zhang K, Jonasson F, Sigurdsson H, Thorsteinsdottir U, Stefansson K. A rare nonsynonymous sequence variant in C3 is associated with high risk of age-related macular degeneration. Nat Genet. 2013; 45(11):1371–4. [PubMed: 24036950]

Hernandez RD. A flexible forward simulator for populations subject to selection and demography. Bioinformatics. 2008; 24(23):2786–7. [PubMed: 18842601]

Hindorff LA, Gillanders EM, Manolio TA. Genetic architecture of cancer and other complex diseases: lessons learned and future directions. Carcinogenesis. 2011; 32(7):945–54. [PubMed: 21459759]

Hoban S, Bertorelle G, Gaggiotti OE. Computer simulations: tools for population and evolutionary genetics. Nat Rev Genet. 2011; 13(2):110–22. [PubMed: 22230817]

Kraja AT, Culverhouse R, Daw EW, Wu J, Van Brunt A, Province MA, Borecki IB. The Genetic Analysis Workshop 16 Problem 3: simulation of heritable longitudinal cardiovascular phenotypes based on actual genome-wide single-nucleotide polymorphisms in the Framingham Heart Study. BMC Proc 3 Suppl. 2009; 7:S4.

Lam TK, Spitz M, Schully SD, Khoury MJ. Drivers" of translational cancer epidemiology in the 21st century: needs and opportunities. Cancer Epidemiol Biomarkers Prev. 2013; 22(2):181–8. [PubMed: 23322363]

Liberles DA, Teichmann SA, Bahar I, Bastolla U, Bloom J, Bornberg-Bauer E, Colwell LJ, de Koning AP, Dokholyan NV, Echave J, Elofsson A, Gerloff DL, Goldstein RA, Grahnen JA, Holder MT, Lakner C, Lartillot N, Lovell SC, Naylor G, Perica T, Pollock DD, Pupko T, Regan L, Roger A, Rubinstein N, Shakhnovich E, Sjolander K, Sunyaev S, Teufel AI, Thorne JL, Thornton JW, Weinreich DM, Whelan S. The interface of protein structure, protein biophysics, and molecular evolution. Protein Sci. 2012; 21(6):769–85. [PubMed: 22528593]

Liberles DA, Teufel AI, Liu L, Stadler T. On the need for mechanistic models in computational genomics and metagenomics. Genome Biol Evol. 2013; 5(10):2008–18. [PubMed: 24115604]

Liu Y, Athanasiadis G, Weale ME. A survey of genetic simulation software for population and epidemiological studies. Hum Genomics. 2008; 3(1):79–86. [PubMed: 19129092]

Lohmueller KE. The impact of population demography and selection on the genetic architecture of complex traits. 2013 arXiv:1306.5261 [q-bio.PE].

Maher MC, Uricchio LH, Torgerson DG, Hernandez RD. Population genetics of rare variants and complex diseases. Hum Hered. 2012; 74(3-4):118–28. [PubMed: 23594490]

Mechanic LE, Chen HS, Amos CI, Chatterjee N, Cox NJ, Divi RL, Fan R, Harris EL, Jacobs K, Kraft P, Leal SM, McAllister K, Moore JH, Paltoo DN, Province MA, Ramos EM, Ritchie MD, Roeder K, Schaid DJ, Stephens M, Thomas DC, Weinberg CR, Witte JS, Zhang S, Zollner S, Feuer EJ, Gillanders EM. Next generation analytic tools for large scale genetic epidemiology studies of complex diseases. Genet Epidemiol. 2012; 36(1):22–35. [PubMed: 22147673]

Morrison AC, Voorman A, Johnson AD, Liu X, Yu J, Li A, Muzny D, Yu F, Rice K, Zhu C, Bis J, Heiss G, O'Donnell CJ, Psaty BM, Cupples LA, Gibbs R, Boerwinkle E. Whole-genome sequence-based analysis of high-density lipoprotein cholesterol. Nat Genet. 2013; 45(8):899–901. [PubMed: 23770607]

Peng B, Chen H, Mechanic LE, Racine B, Gillanders EM, Feuer EJ. Genetic data simulators and their applications: an overview. Genet Epidemiol. 2014 (submitted).

Peng B, Chen HS, Mechanic LE, Racine B, Clarke J, Clarke L, Gillanders E, Feuer EJ. Genetic Simulation Resources: a website for the registration and discovery of genetic data simulators. Bioinformatics. 2013; 29(8):1101–2. [PubMed: 23435068]

Ritchie MD, Bush WS. Genome simulation approaches for synthesizing in silico datasets for human genomics. Adv Genet. 2010; 72:1–24. [PubMed: 21029846]

Savageau MA, Fasani RA. Qualitatively distinct phenotypes in the design space of biochemical systems. FEBS Lett. 2009; 583(24):3914–22. [PubMed: 19879266]

Stephens PJ, Greenman CD, Fu B, Yang F, Bignell GR, Mudie LJ, Pleasance ED, Lau KW, Beare D, Stebbings LA, McLaren S, Lin ML, McBride DJ, Varela I, Nik-Zainal S, Leroy C, Jia M, Menzies A, Butler AP, Teague JW, Quail MA, Burton J, Swerdlow H, Carter NP, Morsberger LA, Iacobuzio-Donahue C, Follows GA, Green AR, Flanagan AM, Stratton MR, Futreal PA, Campbell PJ. Massive genomic rearrangement acquired in a single catastrophic event during cancer development. Cell. 2011; 144(1):27–40. [PubMed: 21215367]

Tennessen JA, Bigham AW, O'Connor TD, Fu W, Kenny EE, Gravel S, McGee S, Do R, Liu X, Jun G, Kang HM, Jordan D, Leal SM, Gabriel S, Rieder MJ, Abecasis G, Altshuler D, Nickerson DA, Boerwinkle E, Sunyaev S, Bustamante CD, Bamshad MJ, Akey JM. Evolution and functional impact of rare coding variation from deep sequencing of human exomes. Science. 2012; 337(6090):64–9. [PubMed: 22604720]

Thornton KR, Foran AJ, Long AD. Properties and modeling of GWAS when complex disease risk is due to non-complementing, deleterious mutations in genes of large effect. PLoS Genet. 2013; 9(2):e1003258. [PubMed: 23437004]

Torstenson ES, Li B, Li C. ASAP: an environment for automated preprocessing of sequencing data. BMC Res Notes. 2013; 6:5. [PubMed: 23289815]

Urbanowicz RJ, Kiralis J, Sinnott-Armstrong NA, Heberling T, Fisher JM, Moore JH. GAMETES: a fast, direct algorithm for generating pure, strict, epistatic models with random architectures. BioData Min. 2012; 5(1):16. [PubMed: 23025260]

Vandenbroucke JP, von Elm E, Altman DG, Gotzsche PC, Mulrow CD, Pocock SJ, Poole C, Schlesselman JJ, Egger M. Strengthening the Reporting of Observational Studies in Epidemiology (STROBE): explanation and elaboration. Epidemiology. 2007; 18(6):805–35. [PubMed: 18049195]

Xing J, Zhang Y, Han K, Salem AH, Sen SK, Huff CD, Zhou Q, Kirkness EF, Levy S, Batzer MA, Jorde LB. Mobile elements create structural variation: analysis of a complete human genome. Genome Res. 2009; 19(9):1516–26. [PubMed: 19439515]

Xu Y, Wu Y, Song C, Zhang H. Simulating Realistic Genomic Data With Rare Variants. Genetic Epidemiology. 2013; 37(2):163–172. [PubMed: 23161487]

Yang T, Deng HW, Niu T. Critical assessment of coalescent simulators in modeling recombination hotspots in genomic sequences. BMC Bioinformatics. 2014; 15:3. [PubMed: 24387001]

Zakov S, Kinsella M, Bafna V. An algorithmic approach for breakage-fusion-bridge detection in tumor genomes. Proc Natl Acad Sci U S A. 2013; 110(14):5546–51. [PubMed: 23503850]

Ziegler A, Ghosh S, Dyer TD, Blangero J, MacCluer J, Almasy L. Introduction to genetic analysis workshop 17 summaries. Genetic Epidemiology. 2011; 35(S1):S1–S4. [PubMed: 22128048]

Zuk O, Schaffner SF, Samocha K, Do R, Hechter E, Kathiresan S, Daly MJ, Neale BM, Sunyaev SR, Lander ES. Searching for missing heritability: designing rare variant association studies. Proc Natl Acad Sci U S A. 2014; 111(4):E455–64. [PubMed: 24443550]

Table I

Opportunities to Advance Field of Genetic Simulation

The Science of Genetic Simulation	
1.1	Development of simulation models capable of incorporating the biological heterogeneity of phenotypes, including longitudinal outcomes, time-dependent variables, environmentally modified traits and endophenotypes.
1.2	Incorporation of more complex genetic models into simulations, including interactions, population genetic models, and differing patterns of association.
1.4	Consideration of sources of error in modeling data from new technologies
1.4	Development of genetic simulation models for other types of genetic variation (e.g. structural variation) and other types of variation (e.g. epigenetics, gene expression)
1.5	Need for increased realism and complexity in genetic simulations, balanced with efficiency and implementation requirements
Software and Methods Development	
2.1	Creation of an ontology for genetic simulation
2.2	Development of guidelines and standards for reporting on genetic simulation, including documentation of programs, description of programs in journal articles, and reporting by end-users when using simulation programs for applications
2.3	Certification of genetic simulation tools based on defined checklist including whether programs were open source, user-friendly implementation (or provide an installer for supported platforms), provide adequate documentation, and use standard data input and output formats
2.4	Increased support for maintenance of genetic simulation programs
2.5	Encouragement of deposition of software into code repositories
2.6	Consideration of requirements of end-users including incorporation of graphical user interface and appropriate documentation
2.7	Selection and recommendation of a core set of genetic simulation programs for the most common research questions
2.8	Development of a small number of multi-use programs to support flexible, broad-based models for genetic simulation
2.9	Creation of comprehensive framework, or genetic simulation server for development and sharing of genetic simulation programs and data
2.10	Comparative modeling using genetic simulation programs using a defined test suite to compare programs and data obtained using these programs
2.11	Identification and promotion of common data sets for comparison of analysis tools
2.12	Encouragement of making simulated data sets public available
Fostering Collaboration	
3.1	Increased communication between simulation developers, end-users of simulation programs, and researchers from other communities
3.2	Formation of a consortium dedicated to fostering the science of genetic simulation
3.3	Creation of an on-line forum for discussion about genetic simulation
3.4	Exploration of opportunities to interact with Cancer Intervention and Surveillance Modeling Network (CISNET) consortium and other communities
3.5	Development and support of a curriculum in genetic simulation to train researchers in genetic simulation, programming and best practices, including accurate documentation and evaluation of programs.