# Table of Contents

Executive Summary ................................................................. iii
Contributors ................................................................................. v
Introduction .................................................................................. 1
Genomic Data: Overall Scope of the Data Science Challenges .................. 3
Critical Challenges in Genomics ....................................................... 3
Data Provenance, Collection, and Management ...................................... 3
Delineation of Phenotypes ................................................................. 5
Adjudication of Genomic Variants ..................................................... 7
Biostatistics and Bioinformatics ......................................................... 8
Data Sharing .................................................................................. 10
Bioethics and the Law ................................................................. 12
Key Recommendations for Genomics .............................................. 13
Conclusion .................................................................................... 14
Acknowledgements ....................................................................... 15
References ..................................................................................... 16
The National Consortium for Data Science ......................................... 18
Executive Summary

Society has been greatly impacted by the massive amounts of data that are now available for use. Immense streams of data are routinely collected across science, medicine, education, government, industry, social media, and entertainment. Yet our ability to efficiently and effectively use these “big data” sets is hampered by our fundamental lack of understanding of the science of data.

The interdisciplinary field of data science can provide solutions to the big data impasse by furthering our understanding of the production, management, and use of big data and by developing principles and theories to guide new discoveries and innovations.

The National Consortium for Data Science (www.data2discovery.org) was established in February 2013 as a public/private partnership of leading universities, governmental agencies, and businesses. The mission of the NCDS is to provide the foundation needed to advance the field of data science through research, education, and economic opportunity.

The Consortium has defined data science as the systematic study of the organization and use of digital data in order to accelerate discovery, improve critical decision-making processes, and enable a data-driven economy.

The Consortium aims to serve as a vital forum for the broad community of scientists, business leaders, funders, policy makers, and other stakeholders to identify challenges, solutions, and visions for the future of data science.

As part of its efforts to advance data science, the Consortium held the first annual Data Science Leadership Summit in April 2013. The theme of the summit was “Data to Discovery: Genomes to Health.” Genomics was chosen as the initial theme because the field offers some of the greatest challenges, as well as some of the greatest promises, in data science. The purpose of the summit was to gather together leaders in data science and genomics to discuss current challenges in genomics, brainstorm for solutions, and reach consensus on the most appropriate and technologically advanced recommendations to address the challenges.

Over 70 leaders in data science and genomics gathered at the summit to discuss critical challenges in six areas of genomics, as summarized below.

- **Data Provenance, Collection, and Management**
  Maintaining data provenance, without knowledge of how exactly data will be repurposed and reused, presents challenges to the collection and management of big data sets. These challenges are amplified by the technological complexities associated with data derived from multiple sources and incompatible or legacy data management systems. Trained data and information specialists are required to manage all aspects of the life cycle of genomic data.

- **Delineation of Phenotypes**
  The characterization of rich and complex phenotypes in the absence of standardization of data elements and harmonization of existing data sets presents significant challenges in the delineation of phenotypes. The lack of adequate tools and techniques to extract phenotype data from large data sets further complicates this matter.

- **Adjudication of Genomic Variants**
  The adjudication of variants is complicated by the lack of standards for both phenotype and variant data, evolving expert consensus on the functional impact of variants, and an absence of high quality, non-proprietary, clinical-grade variant databases. There is also a need for new analytical tools to evaluate rare variants, epistasis, and epigenetic influences.

- **Biostatistics and Bioinformatics**
  Challenges in biostatistics and bioinformatics are amplified on a large scale due to inadequate statistical models and software, insufficient computer processing power or unacceptable time delays when running complex models, heterogeneous and incomplete data sets, limited tools to analyze rare sequence variants and mutations along entire pathways or systems, and limited adoption of federated distributed data systems to promote data integration and sharing. There is also a need for a workforce of skilled analysts who are trained to handle large data sets, as well as effective Clinical Decision Support systems to guide interpretation of analytical results.

- **Data Sharing**
  Data sharing is complicated by the risk of re-identification of de-identified genomic samples and the lack of policies and procedures to protect against genomic breaches and to ensure data provenance as data is repurposed and reused. Incentives are needed to promote data sharing and encourage compliance with policies and procedures related to privacy, security, and provenance.

- **Bioethics and the Law**
  Unresolved bioethical issues abound in genomics, including those related to incidental findings, disclosure of
genetic test results, and privacy issues affecting sensitive populations. There also is a need for open discussions on the legal distinctions between physical property versus intellectual property versus informational property and between privacy versus confidentiality of genomic data.

In response to these challenges, the Leadership Summit participants identified key recommendations to move the field of genomics forward.

- **Foster interdisciplinary collaboration and coordination of efforts in genomics research.**

  A “consortium of consortia” and coordination of the efforts of individual groups will facilitate the development of standards and foster interdisciplinary collaboration.

- **Advance analytical approaches and tools for widespread adoption of standards and federated distributed data systems, harmonization of existing data sets, integrated analyses, data repurposing and reuse, and discovery science.**

  Standardization of data elements, harmonization of existing data sets, widespread adoption of federated distributed data systems, and the development of new analytical tools will enable integrated analyses of large data sets from multiple sources, foster data repurposing and reuse, and promote discovery science.

- **Promote data sharing while maintaining privacy, security, and provenance through incentives, new technical solutions, and cost-benefit analyses of different technological approaches to data sharing.**

  Incentive systems and the identification or development of cost-effective technical schemes for both data sharing and the incorporation of user-generated content will promote data sharing and compliance with policies and procedures for privacy, security, and provenance.

- **Develop automated, easy to use, stakeholder-driven, open source, Clinical Decision Support systems.**

  Clinical Decision Support systems that present genomic data in a simple, synthesized manner, reflect the perspectives of all stakeholders, and incorporate wiki-like capabilities and/or crowdsourcing will empower clinicians and other stakeholders to interpret and apply genomic findings and thus fully realize the power of personalized medicine.

- **Cultivate education and training initiatives in big data–based information technology, digital archiving, and analysis.**

  Education and training initiatives in information technology, digital archiving, and analysis of big data will cultivate the next generation of data science leaders. The incorporation of basic biostatistical concepts into existing training programs targeted at a broad range of specialties and stages of training and career development will enhance the ability of scientists and clinicians to more effectively interpret and apply genomic data.

- **Address bioethical and legal policies on issues such as incidental findings and the distinction between fair use and misuse of genomic data.**

  The National Consortium for Data Science is charged with leading open discussions on bioethics and the law as it relates to genomics, with an initial focus on incidental findings and the distinction between fair use and misuse of genomic data.

The National Consortium for Data Science is committed to advancing the field of data science, and the Leadership Summit was an initial step in that direction. The Consortium believes that the recommendations that arose from the summit will further advance the fields of genomics, specifically, and data science, more generally.
Contributors

**Contributing Authors:**
Stanley C. Ahalt, PhD, Director of RENCI and Professor of Computer Science, UNC-Chapel Hill
Chris Bizon, PhD, Senior Informatics Scientist, RENCI
Jim Evans, MD, PhD, Bryson Distinguished Professor, Departments of Genetics and Medicine, UNC-Chapel Hill, and Editor-in-Chief of Genetics in Medicine, the official journal of the American College of Medical Genetics
Yaniv Erlich, PhD, Principal Investigator and Andria and Paul Heafy Family Fellow, The Whitehead Institute
Geoffrey S. Ginsburg, MD, PhD, Director of Genomic Medicine, Duke Institute for Genome Sciences & Policy, Executive Director of the Center for Personalized Medicine, Duke Medicine, and Professor of Medicine and Pathology, Duke University Medical Center
Ashok Krishnamurthy, PhD, Deputy Director of RENCI
Leslie Lange, PhD, Professor of Genetics, UNC-Chapel Hill
Dan Maltbie, Chief Technology Officer, Annai Systems
Dan Masys, MD, Professor of Biomedical and Health Informatics, University of Washington
Charles Schmitt, PhD, Director of Informatics and Data Sciences, RENCI
Kirk Wilhelmsen, MD, PhD, Professor of Genetics and Neurology, UNC-Chapel Hill, Chief Domain Scientist for Genomics, RENCI

**Discussion Contributors:**
Pankaj Agarwal, MS, Bioinformatician, Bioinformatics Shared Resource, Duke Cancer Institute, Duke University
Kareem Aggour, MS, Senior Computer Engineer at General Electric
Bruce Aronow, PhD, Co-Director of Computational Medicine and Professor, University of Cincinnati Department of Pediatrics, Cincinnati Children’s Hospital
Brian Athey, PhD, Professor and Inaugural Chair of the Department of Computational Medicine and Bioinformatics, University of Michigan
Randy Avent, PhD, Associate Vice Chancellor of Research Development and Professor of Computer Science, North Carolina State University
Jonathan Berg, PhD, Assistant Professor, Department of Genetics, UNC-Chapel Hill
Geoffrey Bock, PSM, MBA, Associate Manager of Research Program Development and Support, The Hamner Institutes for Health Sciences
Chris Brown, PhD, Vice President for Research and Graduate Education, University of North Carolina General Administration
Robert Calderbank, PhD, Dean of Natural Sciences and Professor of Electrical & Computer Engineering, Duke University
Thomas Carsey, PhD, Director of the Odum Institute and Pearsall Distinguished Professor of Political Science, UNC-Chapel Hill
Jonathan Crabtree, Assistant Director of Archive and Information Technology, Odum Institute, UNC-Chapel Hill
Robert Darnell, MD, PhD, President of the Board of Directors and Scientific Director at the New York Genome Center
Allen Deary, PhD, Senior Advisor at the National Institute of Environmental Health Sciences
Michael De Leo, Corporate Consulting Engineer at Cisco Systems
Richard Durbin, PhD, Joint Head of Human Genetics and Head of Genome Informatics Group, Wellcome Trust Sanger Institute
Lee Elizabeth Edsall, PhD, Research Programmer at the University of California at San Diego
Barbara Entwisle, PhD, Vice Chancellor for Research and Kenan Distinguished Professor of Sociology, UNC-Chapel Hill
David Fellinger, Chief Scientist at DataDirect Networks
David Fenske, PhD, Dean of the College of Information Science and Technology and Isaac L. Auerbach Professor of Information Science, Drexel University
Mike Folk, PhD, President and Chief Executive Officer at The HDF Group
Terry Gaasterland, PhD, Director and Professor of Computational Biology and Genomics at the Scripps Genome Center
Tim Gabel, MBA, Executive Vice President of Social, Statistical, and Environmental Sciences at RTI International
Cynthia Gibas, PhD, Associate Professor of Bioinformatics and Genomics, UNC-Charlotte
Dmitry Gordenin, PhD, Senior Associate Scientist at the National Institute of Environmental Health Sciences
Russ Gyurek, MBA, Director of Innovation, Office of the Chief Technology Officer, Cisco Systems
Ed Hammond, PhD, Director of the Duke Center for Health Informatics at the Duke Translational Medicine Institute, Director of Applied Informatics at Duke Health Technology Solutions, and Professor of Medicine, Duke University
Dave Hinds, PhD, Principal Scientist, Statistical Genetics, 23 and Me
Ronald Hines, PhD, Associate Director for Health at the National Health and Environmental Effects Research Laboratory, Environmental Protection Agency
Don Hobart, JD, Associate Vice Chancellor for Research, UNC-Chapel Hill
Dan Janies, PhD, Carol Grotnes Belk Professor of Bioinformatics and Genomics, UNC-Charlotte
Tevfik Kosar, PhD, Associate Professor and Director of the Data Intensive Distributed Computing Laboratory, Department of Computer Science & Engineering, SUNY at Buffalo
Hye-Chung Kum, PhD, Research Assistant Professor of Computer Science, UNC-Chapel Hill
Shannon Lasater, MPA, Project Manager at SAS Institute
Jason Lieb, PhD, Professor of Genetics and Director of the Carolina Center for Genome Sciences, UNC-Chapel Hill
David Maret, MBA, Client Representative, Public Sector, IBM
Lawrence Mays, MD, Professor and Chair of the Department of Bioinformatics and Genomics, UNC-Charlotte
Howard McLeod, PharmD, Fred Eshelman Distinguished Professor of Pharmacy and Director of the Center for Pharmacogenomics and Individualized Therapy, UNC-Chapel Hill
Rodolfo Milito, PhD, Senior Technical Leader at Cisco Systems
John Moore, Senior Director of Advanced Initiatives & Strategic Planning, MCNC
John Quackenbush, PhD, Professor and Director of the Center for Cancer Computational Biology at the Dana-Farber Cancer Institute, Harvard University
Michael Resnick, PhD, Head of the Chromosome Stability Group at the National Institute of Environmental Health Sciences
Rose Ritts, PhD, Director of the Office of Licensing & Ventures and Director of the Duke Innovative Health Institute Health Initiative, Duke University
Bob Schnabel, PhD, Research Assistant Professor in the Division of Animal Sciences, University of Missouri
Jim Siedow, PhD, Vice Provost for Research and Professor of Biology, Duke University
Richard Smith, PhD, Director of the Statistical and Mathematical Sciences Institute
Teresa Squire, PhD, Senior Program Analyst of Science and Technology Development, North Carolina Biotechnology Center
Dan Stanzione, PhD, Deputy Director of the Texas Advanced Computing Center, University of Texas at Austin
James Taylor, PhD, Assistant Professor of Computational Biology & Informatics, Emory University
Jessica Tenenbaum, PhD, Associate Director for Bioinformatics, DTMI Biomedical Informatics Core, Duke University
David Threadgill, PhD, Professor and Head of the Genetics Department, North Carolina State University
Richard Woychick, PhD, Deputy Director of the National Institute of Environmental Health Sciences
Fred Wright, PhD, Professor of Biostatistics, Statistical & Computational Genome Analysis, UNC-Chapel Hill
David Yazdani, Senior Solutions Architect at Cisco Systems

How to reference this paper
Introduction

Our world is undergoing a massive transformation due to the widespread availability of “big data.” Science, medicine, education, government, industry, social media, and entertainment—all are experiencing radical changes as a result of these massive data sets. Immense streams of data are routinely collected by scientists during the course of experimentation, health care providers documenting electronic medical records, educators collecting data on student performance, students manipulating large data sets as part of their curriculum, governmental agencies tracking citizen behavior in the interest of local and national security, businesses tracking online behavior and customer transactions, investors tallying the risks and rewards of financial investments, and children and adults participating in social media forums and online games. These data collections hold tremendous potential to solve problems, improve health and quality of life, reform our education system, and create economic growth. Yet our ability to use these massive data sets in a cost-efficient and -effective manner is hampered by our fundamental lack of understanding of the science of data (Horvitz and Mitchell, 2010; Agrawal et al., 2011; Boyd and Crawford, 2011; Manyika et al., 2011) (Figure 1).

The highly interdisciplinary field of data science can provide the solution to the big data impasse through the systematic study of data to further our understanding of the production, management, and use of data and to develop principles and theories that can guide new discoveries and innovations and thus foster economic activity and the health, education, and well-being of people around the world.

The National Consortium for Data Science1 was established in February 2013 as a public/private partnership of leading universities, governmental agencies, and businesses. The mission of the NCDS is to provide the foundation needed to advance the field of data science through research, education, and economic opportunity.

The Consortium has defined data science as the systematic study of the organization and use of digital data in order to accelerate discovery, improve critical decision-making processes, and enable a data-driven economy.

The Consortium aims to serve as a vital forum to: engage a broad community of data science experts to define challenges, solutions, and visions for the future of data science; facilitate frequent, close interchange between data scientists and business leaders, funders, policy makers, and scientists in data-intensive domains; coordinate research priorities in data science; promote a focus on data science at multiple levels, including local, state, national, and international; support the development of technical, ethical, and policy standards for data; provide leadership in the advancement of data science; and cultivate the next generation of data science leaders through education and training programs in data science.

Organizations that committed to the establishment of the Consortium have been designated as Founding Members (see page 18) and share the responsibility of overseeing and directing the work of the Consortium through their participation in various committees and working groups. Membership fees support the Consortium’s activities.

As part of its efforts to advance data science, the Consortium established an Annual Data Science Leadership Summit. The first summit was held on April 23 and 24, 2013 in Chapel Hill, North Carolina, and the theme was “Data to Discovery: Genomes to Health”—a theme that was chosen

---

because the field of genomics offers some of the greatest challenges, as well as some of the greatest promises, in data science. The purpose of the summit was to gather together top leaders in data science and genomics to discuss current challenges in genomics, brainstorm for solutions, and reach consensus on the most appropriate recommendations to address the challenges.

Over 70 leading scientists, geneticists, industry leaders, and governmental representatives gathered from across the country to participate in the summit.

The summit focused on six critical challenges in genomics, drawn from across the data science spectrum (Figure 2): (1) Data Provenance, Collection, and Management; (2) Delineation of Phenotypes; (3) Adjudication of Genomic Variants; (4) Biostatistics and Bioinformatics; (5) Data Sharing; and (6) Bioethics and the Law.

This White Paper summarizes these challenge areas and defines the key recommendations developed by the summit participants to move the field of genomics—indeed, the field of data science—forward.

Figure 2: Data Science Challenges in Genomics. (Image provided courtesy of RENCI).
The field of genomics has benefitted from numerous advances in data science. For example, we now have unprecedented access to massive streams of rich genomic data from multiple sources, ranging from myriad species to many thousands of individual human genomes; the cost of collecting/managing/storing these massive data sets has decreased significantly, while computing power and memory storage capacity have increased exponentially; and major advances have been made in analytical tools and algorithms, including machine learning, natural language processing, and informatics (Horvitz and Mitchell, 2010; Kahn, 2011).

Yet while we are now capable of quickly amassing large volumes of genomic data at a relatively modest cost (Kahn, 2011; Manyika et al., 2011), we do not have adequate analytical tools, technologies, and policies to federate, process, synthesize, distill, and apply genomic data sets at a cost comparable to what is required to generate the data.

“The $1,000 genome, the $100,000 analysis?” —Elaine Mardis, Musing on Genome Medicine,

Critical Challenges in Genomics

Data Provenance, Collection, and Management

Discussion Moderator: Dan Maltbie, Chief Technology Officer, Anniai Systems

Dan Maltbie is a Silicon Valley entrepreneur, with over thirty years of experience. Mr. Maltbie has held positions at GRiD Systems, DiviCom, Woven Systems, Juniper Networks, Hewlett-Packard, and other technology-oriented businesses. He has contributed to the development of many networking and high performance computing systems for genomics and a variety of other applications.

Overview: Predicting data repurposing and reuse, while maintaining data provenance, presents challenges to the collection and management of big data sets. These challenges are amplified by the technological complexities associated with data derived from multiple sources and incompatible or legacy data management systems. Trained data and information specialists are required to manage all aspects of the life cycle of genomic data.

Data collection and management are fundamental to genomics research and clinical application, yet traditional technologies and tools become less effective on a large scale. Incompatible standards, formats, and versions complicate the merging, integration, and reuse of data from multiple sources; this problem is evident even with small data sets but becomes tremendously complex when dealing with large data sets. A related important issue is whether data sets derived using different sequencing platforms can be legitimately combined. These issues are exacerbated by industry, as competition among external vendors encourages the differentiation and incompatibility of systems (Digital Standards Organization [Digistan], 2013).

Another challenge is that accepted ontologies and terminologies, if adopted at all, are evolving, thus necessitating continual updates to database systems and extensive database annotation. Often, these issues are addressed only by those groups specifically funded to maintain long-term data collections, such as the U.S. Library of Congress (www.loc.gov/index.html) and the National Center for Biotechnology Information (http://www.ncbi.nlm.nih.gov/).

Data repurposing and reuse are further complicated by the fact that it is difficult or impossible to predict how data will be reused. Frequently, the details of the data collection process, which typically are required to reuse the data, are not always transmitted with the data. Data sets derived from multiple sources often are collected without agreements in place to ensure provenance for the primary data sources or broad informed consent to allow data reuse. Existing consent forms used to collect genetic information do not adequately address reuse issues, and the regulatory infrastructure to ensure the ethical reuse of data is lacking.

One critical challenge is the absence of a theoretical foundation for differentiating between when standard technologies and formats are essential to facilitate the transfer, integration, and reuse of data from multiple sources and platforms and when standard technologies and formats are not essential because of the availability or development of better tools to enable the integration of heterogeneous data. In other words, while standards can facilitate data integration and sharing, they can also present barriers to the discovery of new data types, reuse of existing data, and development of new technologies; thus, understanding when standards are needed and when they are not is crucial (Figure 3).
Tracking data and metadata derived from multiple sources and systems, while maintaining privacy and security policies and procedures, can be extremely difficult on a large scale, and automating these processes is complicated. In addition, the costs involved in data storage and the maintenance of audit trails present a major barrier to the use of big data by small research groups and laboratories. The cost barrier is often unappreciated because of the availability of low-cost and relatively high-capacity memory devices such as flash drives, but the real world costs associated with enterprise-level data storage and curation of massive data sets are significant and involve much more than disk space, including the need for thousands of disk drives, back-up systems to cover drive failures, and substantial personnel time. A survey by the journal Science on the availability and use of data among the journal’s peer reviewers found that >50% of 1700 respondents regularly use data sets ≥1 gigabytes in size (and ~20% use data sets ≥100 gigabytes in size), yet 80.3% of respondents reported that they did not have sufficient funding for data curation in their laboratory or research group (Science Staff, 2011).

For biological samples, long-term preservation becomes a problem because DNA and RNA are subject to degradation over time. The storage of biological samples presents additional challenges in that freezers can malfunction, alarms may not be responded to, and back-up generators often are not purchased or installed due to cost constraints. The costs associated with long-term storage of biological samples are more difficult to estimate than those associated with digital storage because the costs and benefits of long-term storage of biological samples are unclear. For instance, failure rates for tape- and disk-based archival data are understood well enough to include reliability of data recovery as a standard parameter in selecting service-level agreements for data storage. Failure rates for stored biological samples are unknown.

In addition, future improvements in sequencing technology, analytical approaches, and digital storage capabilities may mitigate the need for long-term storage of biological samples. Also, the specifics of the informed consent process used to establish past collections can significantly affect the extent to which biological samples can be reused, but reuse needs are typically unknown when these collections are established. Another caveat is that the benefits of long-term storage may never be realized if biological samples are found to be useless due to insufficient or inaccurate processing, annotation, or tracking (Ginsburg et al., 2008). A recent high-profile case of a misidentified cancer cell line demonstrates the devastating ramifications associated with this problem; most notably, the publication of false results (ScienceDaily Staff, 2012).

“...just as DNA mismatch has exonerated death row inmates, DNA mismatch showing that a cell line doesn’t match its label can call into question perhaps decades worth of research done using the cells.”
—Staff, ScienceDaily, 2012

Existing cyberinfrastructures are often inadequate, lacking sufficient storage space, long-term maintenance and curation capabilities, and the ability to share or transfer massive data sets. Further, the energy demand required to support adequate cyberinfrastructure is staggering and associated not only with high costs but huge environmental impacts. Compounding the problem is the need for back-up energy sources to protect against the loss of data or data access due to power failures. A recent analysis of energy use by data centers, conducted by McKinsey & Company on behalf of The New York Times, found that only 6% to 12% of the electricity powering servers was used to support computation; the remainder went largely to cooling servers and keeping them idle but ready for use (Glanz, 2012).

“Worldwide, the digital warehouse uses about 30 billion watts of electricity, roughly equivalent to the output of 30 nuclear power plants...Data centers in the United States account for one-quarter to one-third of that load...” —James Glanz, New York Times, 2012

Legacy systems are often upgraded in an ad hoc manner that hinders the adoption of and compatibility with newer technologies. In addition, storage media often are not updated as technologies evolve, and software is frequently custom-made to run on servers optimized for a given application. These servers are then either shut down or not updated as tech-
nology advances, resulting in data that, even if stored on up-to-date storage media, cannot be read because the software cannot be run. This problem can have significant implications, as evidenced by a recent article in the journal Science, in which researchers at the Max Planck Institute for Physics detail the negative impact that the loss of “old” experimental data had on the field of particle physics (Curry, 2011). The authors estimate that the computing and storage resources required to properly curate their data after the experiment had ended would have cost only 1% of the total research budget, but the return would have continued indefinitely. While the costs of generating data in genomics and particle physics aren’t comparable, the fundamental problems related to data storage and reuse are similar across fields.

Finally, today’s information technologists are often inadequately trained to handle massive data sets. A survey of 752 business executives, conducted by the Economist Intelligence Unit, found that 41% of respondents report a lack of skilled employees as a key barrier to processing data more rapidly, compared to 32% of respondents who view a lack of technology as a key barrier (Giles, 2012).

Whereas “genotype” refers to an organism’s genetic code, “phenotype” refers to an organism’s observed characteristics or traits, including physical properties, developmental profile, and behavior.

An organism’s phenotype is influenced in large part by its genotype, although the environment, including one’s own behaviors, strongly affects phenotypes in ways that are poorly understood at present. Indeed, the fields of epigenomics and epidemiology focus on environmental influences on phenotype (Berger et al., 2009).

A major challenge in the delineation of phenotypes is that the characterization of complex phenotypes is quite difficult. A phenotypic trait such as standing height is easy to define, but genetic studies rarely apply a uniform methodology to measure height (e.g., shoes on or off, subject maximally extended or relaxed). Other traits such as cardiovascular disease or hypertension can be quite difficult to define, with expert consensus difficult to achieve. For diseases even more complex such as psychiatric disorders (e.g., autism, depression), one rate-limiting step in identifying associated genomic variants is the definition of the phenotype itself (Fornito and Bullmore, 2012). Indeed, descriptors and vocabularies can vary within a given field, particularly for complex disorders that are influenced by sociological, cultural, psychological, and environmental factors (e.g., pain, substance abuse).

A related challenge is that it is exceedingly difficult, if not impossible, to fully capture the richness of a phenotype with a finite set of variables. For example, consider a tumor. A pathologist can use a microscope to observe and describe a tumor over multiple focal planes and three-dimensional space. Capturing the full description of the pathologist’s report on that tumor with a quantitative data set presents an enormous challenge in data science. High-dimensional data on endophenotypes or biomarkers (i.e., quantifiable, heritable, biological traits that are associated with a given phenotype) may help overcome these challenges, but the analytical tools to validate and use these types of data are still under development (Vrieze et al., 2012).

In addition, standard ontologies and terminologies are not fully developed and are necessarily evolving, which hampers efforts to completely characterize and compare genomic variants mapped to complex traits. Even if standard ontologies and terminologies are implemented, it is unclear whether data derived from different measurement systems (e.g., office versus home blood pressure measurements, self-reported versus biomarker-defined) can be legitimately compared or combined. Additional sources of variability include variable names (e.g., hypertension, HTN, high blood pressure), measurement units (e.g., mg versus mg/kg versus ml), and multiple measurements for a given variable (with or without a master list of measurement dates). Medication-related variables are often the most challenging to deal with in this context, as medication can be categorized as current or recent use (yes/no), coded numerically (with a master list that may or may not be available), or provided a text name, which may be listed as a Brand name or generic name.

The absence of adequate publicly available documentation further hampers efforts to interpret data. For instance, in
the Framingham Heart Study (http://www.framinghamheartstudy.org/), the trait “asthma” is associated with numerous different variables, each with varying descriptions and data collection dates (Table 1). A secondary user of these data would not know how to use the data without explicit assistance from someone intimately involved with the data collection process. The availability of a data dictionary might help, but only if it was complete and the secondary user knew how to use it, which is not typical.

Without standardization of data elements and variables and harmonization of existing data sets, the search for phenotypic data in large data sets can be extremely difficult, and adequate search tools are not currently available. Often, phenotype data elements are “hidden” among large amounts of irrelevant information, with cryptic variable names and insufficient or absent documentation to aid in the interpretation of data elements. Without adequate documentation on the data collection process, a secondary user will not understand the meaning of such data, and details such as variable naming schemes and terminology used in documentation can differ greatly across studies or sample subsets. This often leads to large amounts of data being discarded. Even if a given variable is standardized across studies, other important phenotype data that are related to that variable (e.g., date of collection, assay type) may be missing, which also leads to the loss of data. Likewise, if subject/sample identification numbers vary across studies or between genotype and phenotype data sets, then linkage of multiple data sets for an individual subject may be impossible. A related challenge is that documentation isn’t always available on how data were extracted and manipulated (e.g., transformation, categorization), which greatly influences the interpretation of variables.

Further complicating these issues is that there is a need to support variant correlations with already defined phenotypes, while also catalyzing 21st century discovery science. In particular, there is a need for automated tools and approaches to extract phenotypic data elements from data sources such as electronic medical records and other heterogeneous data (e.g., data from geospatial, epidemiologic, and billing databases) and correlate those data with patterns in genomic and other molecular data sets in order to facilitate the discovery of new, previously unrecognized phenotypes.

Thus, the dual challenge is to concurrently develop standards for existing phenotype variables (and the analytical tools required to extract those variables from large data sets) and also create new methods and tools to extract and correlate novel observable patterns (i.e., phenotype in its broadest sense) with genomic and other forms of high throughput laboratory data. The latter activity will inevitably result in new phenotype variables that will then also need to be standardized.

<table>
<thead>
<tr>
<th>Phenotype</th>
<th>Study</th>
<th>Data Set</th>
<th>Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Asthma</td>
<td>FHS</td>
<td>CARE_EX1_25_V1_0108</td>
<td>B128</td>
<td>D203-006-ASTHMA</td>
</tr>
<tr>
<td>Asthma</td>
<td>FHS</td>
<td>CARE_EX1_35_V1_0108</td>
<td>C93</td>
<td>WHEEZING OR ASTHMA IN INTERIM</td>
</tr>
<tr>
<td>Asthma</td>
<td>FHS</td>
<td>CARE_EX1_35_V1_0108</td>
<td>C94</td>
<td>WHEEZING OF LONG DURATION</td>
</tr>
<tr>
<td>Asthma</td>
<td>FHS</td>
<td>CARE_EX1_35_V1_0108</td>
<td>C95</td>
<td>SEASONAL WHEEZING</td>
</tr>
<tr>
<td>Asthma</td>
<td>FHS</td>
<td>CARE_EX1_35_V1_0108</td>
<td>C96</td>
<td>WHEEZING WITH RESPIRATORY INFECTIONS</td>
</tr>
<tr>
<td>Asthma</td>
<td>FHS</td>
<td>CARE_EX1_45_V1_0108</td>
<td>C374</td>
<td>CDX-ASTHMA</td>
</tr>
<tr>
<td>Asthma</td>
<td>FHS</td>
<td>CARE_EX1_45_V1_0108</td>
<td>D117</td>
<td>WHEEZING OR ASTHMA</td>
</tr>
<tr>
<td>Asthma</td>
<td>FHS</td>
<td>CARE_EX1_45_V1_0108</td>
<td>D118</td>
<td>WHEEZING-LONG DURATION</td>
</tr>
<tr>
<td>Asthma</td>
<td>FHS</td>
<td>CARE_EX1_45_V1_0108</td>
<td>D119</td>
<td>WHEEZING SEASONAL</td>
</tr>
<tr>
<td>Asthma</td>
<td>FHS</td>
<td>CARE_EX1_45_V1_0108</td>
<td>D120</td>
<td>WHEEZING-WITH RESPIRATORY INFECTIONS</td>
</tr>
<tr>
<td>Asthma</td>
<td>FHS</td>
<td>CARE_EX1_45_V1_0108</td>
<td>D360</td>
<td>CDX-ASTHMA</td>
</tr>
<tr>
<td>Asthma</td>
<td>FHS</td>
<td>CARE_EX1_55_V1_0108</td>
<td>E339</td>
<td>WHEEZING OR ASTHMA</td>
</tr>
<tr>
<td>Asthma</td>
<td>FHS</td>
<td>CARE_EX1_55_V1_0108</td>
<td>E340</td>
<td>TYPE OF WHEEZING OR ASTHMA</td>
</tr>
<tr>
<td>Asthma</td>
<td>FHS</td>
<td>CARE_EX1_55_V1_0108</td>
<td>E446</td>
<td>CDX-ASTHMA</td>
</tr>
<tr>
<td>Asthma</td>
<td>FHS</td>
<td>CARE_EX1_65_V1_0108</td>
<td>F309</td>
<td>WHEEZING OR ASTHMA</td>
</tr>
<tr>
<td>Asthma</td>
<td>FHS</td>
<td>CARE_EX1_65_V1_0108</td>
<td>F641</td>
<td>CDX-ASTHMA</td>
</tr>
<tr>
<td>Asthma</td>
<td>FHS</td>
<td>CARE_EX1_75_V1_0108</td>
<td>G127</td>
<td>ASTHMA IN INTERIM</td>
</tr>
<tr>
<td>Asthma</td>
<td>FHS</td>
<td>CARE_EX1_75_V1_0108</td>
<td>G128</td>
<td>WHEEZING IN CHEST FOR LAST 12 MONTHS</td>
</tr>
<tr>
<td>Asthma</td>
<td>FHS</td>
<td>CARE_EX1_75_V1_0108</td>
<td>G418</td>
<td>CDX-ASTHMA</td>
</tr>
<tr>
<td>Asthma</td>
<td>FHS</td>
<td>CARE_EX1_75_V1_0108</td>
<td>G670</td>
<td>RESP-EVER HAD ASTHMA</td>
</tr>
<tr>
<td>Asthma</td>
<td>FHS</td>
<td>CARE_EX1_75_V1_0108</td>
<td>G671</td>
<td>RESP-IN 12 MOS, HAD ASTHMA ATTACK</td>
</tr>
<tr>
<td>Asthma</td>
<td>FHS</td>
<td>CARE_EX1_75_V1_0108</td>
<td>G672</td>
<td>RESP-CURRENTLY TAKING MEDS FOR ASTHMA</td>
</tr>
<tr>
<td>Asthma</td>
<td>FHS</td>
<td>CARE_EX1_15_V4_0108</td>
<td>G3A407</td>
<td>CDX-ASTHMA</td>
</tr>
<tr>
<td>Asthma</td>
<td>FHS</td>
<td>CARE_EX1_15_V4_0108</td>
<td>G3A59</td>
<td>EVER HAD ASTHMA</td>
</tr>
<tr>
<td>Asthma</td>
<td>FHS</td>
<td>CARE_EX1_15_V4_0108</td>
<td>G5A40</td>
<td>ASTHMA-RECEIVED MEDICAL TREATMENT</td>
</tr>
<tr>
<td>Asthma</td>
<td>FHS</td>
<td>CARE_EX1_15_V4_0108</td>
<td>G5A41</td>
<td>WHEEZING OR ASTHMA</td>
</tr>
<tr>
<td>Asthma</td>
<td>FHS</td>
<td>CARE_EX1_15_V4_0108</td>
<td>G5A42</td>
<td>ASTHMA-WHAT AGE DID IT START</td>
</tr>
<tr>
<td>Asthma</td>
<td>FHS</td>
<td>CARE_EX1_15_V4_0108</td>
<td>G5A43</td>
<td>ASTHMA-WHAT AGE DID IT STOP</td>
</tr>
<tr>
<td>Asthma</td>
<td>FHS</td>
<td>CARE_EX1_15_V4_0108</td>
<td>G5A44</td>
<td>CDX-ASTHMA</td>
</tr>
<tr>
<td>Asthma</td>
<td>FHS</td>
<td>CARE_RESP1_65_V1_0108</td>
<td>RQ014</td>
<td>EVER HAD ASTHMA</td>
</tr>
<tr>
<td>Asthma</td>
<td>FHS</td>
<td>CARE_RESP1_65_V1_0108</td>
<td>RQ015</td>
<td>MONTHS</td>
</tr>
<tr>
<td>Asthma</td>
<td>FHS</td>
<td>CARE_RESP1_65_V1_0108</td>
<td>RQ016</td>
<td>ASTHMA</td>
</tr>
<tr>
<td>Asthma</td>
<td>FHS</td>
<td>CARE_SLEEP1_1999S_0108</td>
<td>ASTH121</td>
<td>Ph had attack of asthma in last 12 months?</td>
</tr>
<tr>
<td>Asthma</td>
<td>FHS</td>
<td>CARE_SLEEP1_1999S_0108</td>
<td>ASTHMA15</td>
<td>MD said ph had asthma?</td>
</tr>
<tr>
<td>Asthma</td>
<td>FHS</td>
<td>CARE_SLEEP1_1999S_0108</td>
<td>IST01</td>
<td>Inhaled steroids for asthma</td>
</tr>
<tr>
<td>Asthma</td>
<td>FHS</td>
<td>CARE_SLEEP1_1999S_V1_0108</td>
<td>ASTHMA15</td>
<td>MD said ph had asthma?</td>
</tr>
<tr>
<td>Asthma</td>
<td>FHS</td>
<td>CARE_SLEEP1_1999S_V1_0108</td>
<td>ASTHMA15</td>
<td>MD said ph had asthma?</td>
</tr>
<tr>
<td>Asthma</td>
<td>FHS</td>
<td>CARE_SLEEP1_1999S_V1_0108</td>
<td>IST01</td>
<td>Inhaled steroids for asthma</td>
</tr>
<tr>
<td>Asthma</td>
<td>FHS</td>
<td>CARE_SLEEP1_2003S_0108</td>
<td>IST02</td>
<td>INHALED STEROIDS FOR ASTHMA</td>
</tr>
<tr>
<td>Asthma</td>
<td>FHS</td>
<td>CARE_SLEEP1_2003S_0108</td>
<td>OAIA2</td>
<td>(LEUKOTRIENE RECEPTOR ANTAGONISTS AND INHIBITORS OF Lipo-OXIDASE)</td>
</tr>
<tr>
<td>Asthma</td>
<td>FHS</td>
<td>CARE_SLEEP1_2003S_0108</td>
<td>h201d</td>
<td>have asthma?</td>
</tr>
<tr>
<td>Asthma</td>
<td>FHS</td>
<td>CARE_SLEEP1_2003S_0108</td>
<td>h201e</td>
<td>Current asthma: do you still have asthma?</td>
</tr>
</tbody>
</table>

Table 1. Variables Associated with “Asthma” in the Framingham Heart Study. (Table courtesy of Dr. Leslie Lange. The data were openly available at dbGaP, the database of Genotypes and Phenotypes, http://www.ncbi.nlm.nih.gov/gap).
Adjudication of Genomic Variants

Discussion Moderator: Martin Reese, PhD, Co-Founder, Chief Executive Officer, and Chairman of the Board, Omicia

Martin Reese is an internationally recognized expert in medical informatics and bioinformatics, with a track record of bringing grounded scientific knowledge to the corporate sector. Dr. Reese developed gene-finding algorithms for the Human Genome Project, organized the state-of-the-art Genome Annotation Assessment Project, and was a member of the Berkeley Drosophila Genome Project, which provided the essential proof-of-concept platform for Celera’s shotgun sequencing technology.

Overview: The adjudication of variants is complicated by the lack of standards for both phenotype and variant data, evolving expert consensus on the functional impact of variants, and an absence of high quality, non-proprietary, clinical-grade variant databases. There is also a need for new analytical tools to evaluate rare variants, epistasis, and epigenetic influences.

A genomic variant is defined as a change in the nucleotide base sequence as compared to the base sequence of a reference genome (Figure 4). Annotation of variants involves the assignment of biological function to genomic variants; it provides a systematic method for cataloging multiple sources of genomic data and integrating and synthesizing those data to determine biological function, especially with respect to the variant’s impact on clinical disease. The adjudication of variants refers to the process through which a consensus is reached regarding annotation of the functional consequences.

Challenges in variant annotation include the accuracy of variant calling to ensure that sequences are properly aligned and to determine the location of the variant in the genome (i.e., intergenic region, intron region, splice site, untranslated region, coding DNA sequence). Each human genome contains 1–4 million variants, most of which are likely to be biologically irrelevant. This presents enormous analytical and technical challenges to the identification of the few relevant variants for an individual patient. Challenges in variant annotation are amplified when the data are of low coverage or poor quality (e.g., missing data points). Indeed, high quality, clinical-grade variant databases are rare; most databases have incomplete data and unacceptable false positive and false negative rates. Problems related to false positives and false negatives become more pronounced when analyzing data from admixed populations because a given phenotype might be associated with population-specific, rare genomic variants (Mao et al., 2013).

In addition, the filtering of artifacts from variant databases becomes technologically difficult when dealing with massive data sets. Importantly, the functional significance of variants is quite difficult to interpret. For example, while it is straightforward to predict the change in peptide product due to a variant within a coding DNA sequence, the functional significance of the altered peptide product may not be clear. The functional significance of variants located in other genomic regions is even more difficult to predict. Clinical interpretation of variants is also complicated by the distinction between causation and correlation, as well as a publication bias that severely skews the variants that are published in the medical literature to those with biological significance.

Further, there’s a lack of expert consensus on the functional impact of known genomic variants, and the knowledge base (i.e., published literature, public databases, etc.) is constantly evolving and growing; thus, if a consensus is reached, new information will necessitate an evolving consensus, with unknown downstream effects.

Complicating these issues is the lack of standards for the storage of variant information. For example, the vcf (Variant Call Format) file format is the only standard that has been adopted by the genomics community as a convention...
for encoding metadata about variants. However, vcf files can be generated in different ways, which leads to issues with data integration of multiple data sources. In addition, variant data need to be captured in the context of phenotype data, but phenotype data are rarely collected in a standardized or harmonized manner, as described in the previous section. All of these issues complicate analytical approaches and make integrated analyses difficult to conduct.

While the advent of next-generation sequencing and its yield of large data sets has made the study of rare variants possible, this field of study is nascent, with disagreement among investigators as to the impact of rare variants on disease (Cirulli and Goldstein, 2010), as discussed in the next section.

Moreover, the mapping of genotypes to phenotypes is often based on individual SNPs. Only recently have experts begun to evaluate epistasis and the aggregate impact of non-allele genomic variants on gene function (Yandell et al., 2011), and epigenomic influences on variants are extremely difficult to factor in (Berger et al., 2009). Ironically, next-generation sequencing does a poor job at capturing large genomic variants and chromosomal abnormalities with known functional impact, such as Triple X Syndrome or Down Syndrome (Trisomy 21), thus necessitating the need to incorporate data from more traditional technologies, such as karyotyping, into any Clinical Decision Support system.

Lastly, issues related to intellectual property become relevant when researchers are asked to deposit variant data into a public database. Indeed, controversy continues regarding the patenting of genes and the right to create proprietary variant databases (PR Newswire Staff, 2012). In fact, testing for the BRAC1 and BRAC2 gene mutations, which are associated with an increased risk of breast and ovarian cancer, was until very recently under proprietary control by Myriad Genetics, at a cost of a few thousand dollars per patient (Myriad Genetics, 2013). This right was overturned on June 13, 2013, when the U.S. Supreme Court ruled 9-0 against the right to patent genes (NY Times Staff, 2013). While the issue of gene patents has been resolved for now, at least in the U.S., controversy related to intellectual (and informational) property rights and genomics likely will continue.

**Biostatistics and Bioinformatics**

**Discussion Moderator:** Kirk Wilhelmsen, MD, PhD, Professor of Genetics and Neurology, UNC-Chapel Hill, Chief Domain Scientist for Genomics, RENCI

Kirk Wilhelmsen is a practicing neurologist who combines his clinical expertise in neurology with his research interest in genetics. Dr. Wilhelmsen has mapped the loci for many diseases associated with simple modes of inheritance, including the first locus for the most common cause of non-Alzheimer’s dementia. His research focuses on the genetic mapping of susceptibility loci for complex behavioral genetic traits and non-Alzheimer’s dementia.

**Overview:** Challenges in biostatistics and bioinformatics are amplified on a large scale due to inadequate statistical models and software, insufficient computer processing power or unacceptable time delays when running complex models, heterogeneous and incomplete data sets, limited tools to analyze rare sequence variants and mutations along entire pathways or systems, and limited adoption of federated distributed data systems to promote data integration and sharing. There is also a need for a workforce of skilled analysts who are trained to handle large data sets, as well as effective Clinical Decision Support systems to guide interpretation of analytical results.

The techniques and tools of biostatistics and bioinformatics are applied in many areas of genomics, including data collection, variant annotation, gene mapping, and data security. Biostatistics and bioinformatics also play a critical role in the interpretation and clinical application of genomic data. A major challenge is that traditional tools and approaches do not scale for application to large-scale data due to inadequate statistical models (e.g., linear/nonlinear regression, correlational analyses) and/or software programs (e.g., MS Excel, SPSS, SAS), insufficient computer processing power, and unacceptable time delays when running complex models.

Real-time analysis is often difficult or impossible with large-scale data due to insufficient computer processing power and software algorithms. In particular, nonlinear algorithms become problematic even with modest data sets. Traditional visualization tools (e.g., 2-D graphs of summary statistics, tables) are static and do not link the visualization to the data; thus, they cannot be used effectively for the analysis of big data (Fox and Hendler, 2011).

Data “exploration” is typically the first step of analysis and provides insights into variable transformation (e.g., categorical versus continuous variables, logarithmic scaling) and the relationships among variables. With large data sets, this step is often difficult or impossible to fully complete, which can lead to situations in which variables are not adequately represented and/or important relationships are overlooked because they are not realized or anticipated a priori.

Data “cleaning” prior to analysis is often extremely grueling and time-consuming when performed on a large scale. Human verification and quality control become very difficult when dealing with large data sets, and this increases the likelihood of errors and omissions. Additionally, data cleaning often involves multiple operations, which raises questions related to whether data derived from intermediate steps should be saved or reproduced on an as-needed basis. If intermediate data are saved, then issues related to file naming and documentation become important to consider in order to track and audit the data through each step of analysis.
In addition, the complexity and interconnectedness of biological systems are increasingly recognized, and analyses are often based on integrated data sets from multiple sources, which can be challenging due to heterogeneous and incomplete data. Privacy issues become important considerations when working with data sets from multiple sources, as do issues related to provenance.

In the context of genomics, there has been a move away from the era of candidate gene studies to an era in which thousands to millions of variants can be identified for an individual patient and the causal one(s) then needs to be identified (Figure 5). This “N=1” approach is particularly difficult when one is looking for rare sequence variants causing Mendelian disease because each individual has thousands of rare sequence variants in their genome, each of which (alone or in combination) could alter gene function. The challenge is complicated by a bias toward looking for variants in the coding region of genes, primarily because these are easier to identify and interpret; however, for mutations with low penetrance, the disease-causing variant(s) may not be in the coding region. A further complication is that alterations in pathways and systems may be more important than a mutation in a single gene, so there is a need to map not only single gene mutations but mutations along an entire pathway or system. Environmental influences also are not well understood. Another challenge when dealing with rare variants is insufficient or absent phenotype data. For example, rare neurological diseases are often described using ill-defined terms, in part because neurologists have not reached consensus on the clinical features of rare neurological diseases.

Large data sets drawn from multiple sources of data can help overcome some of the challenges involved in identifying rare disease-causing variants in a given individual. However, this approach presents additional challenges. For example, technical approaches to federate data across disparate systems are often insufficient or not widely adopted. Sampling issues are also important to consider when using large data sets because sample populations for genomic studies are often homogeneous, thus rendering the possibility for incorrect inferences. Similarly, while large data sets enable risk assessment and diagnostics, correlation does not equal causation. Conditional probabilities with confidence intervals can suggest that a rare variant causes a specific disease, but this approach does not provide a definitive conclusion regarding the disease-causing rare variant(s) because a genomic hypothesis cannot be directly tested under controlled conditions, thus requiring that findings are validated in independent data sets. In addition, confidence intervals may be unacceptably large to draw any conclusions regarding the relationship between a rare variant and disease.

Large data sets may also help to identify the genetic architecture for common but complex traits (e.g., substance abuse). Heritability patterns for complex genetic traits are not always clear, with “missing” heritability customary. Large data sets may facilitate the identification of biologically important patterns within genomic data that may help to explain complex genetic traits or define new complex traits.

Challenges in the analysis of genomic data are complicated by the fact that analysts often are not adequately trained to work with massive data sets. A 2011 McKinsey Global Institute Report projects that by 2018, the

---

**Figure 5.** Classification of Disease-Causing Sequence Variants by Allele Frequency and Effect Size. (From Manolio et al., Nature 461:747-753, 2009. Reprinted with permission from Nature Publishing Group.)

---

**EFFECT SIZE**

- **HIGH**
- **INTERMEDIATE**
- **MODEST**
- **LOW**

**ALLELE FREQUENCY**

- **VERY RARE**
- **RARE**
- **LOW FREQUENCY**
- **COMMON**
United States will have a shortfall of 440,000 to 490,000 deep analysts with advanced training in data science, and an additional shortfall of 1.5 million data-savvy managers and general analysts (Manyika et al., 2011).

Finally, in terms of clinical application and Clinical Decision Support systems, there is a risk of overloading the already overworked and information-overloaded clinician with too much data and too little variant annotation and interpretation to make sense of the data. This is a very real possibility, particularly given today’s health care environment, with clinicians facing increased demands and information coupled with decreased time for direct patient contact. Further, while clinician assessment of genomic findings can overcome a lack of statistical power in genomic data sets, large amounts of data can diminish returns. Additionally, the current approach of positioning the health care provider as the primary decision-maker, in terms of health care choices, is no longer applicable in today’s world, where health-related information is freely and widely available. Patients, families, health care insurance agencies, and policy makers are playing an increasing role in health care delivery and will continue to do so. There is a pressing need to convey genomic information to all stakeholders in a manner that is fast and easy to understand and apply to patient care.

**Data Sharing**

**Discussion Moderator:** Yaniv Erlich, PhD, Principal Investigator and Andria and Paul Heafy Family Fellow, The Whitehead Institute

Yaniv Erlich is an expert in computational human genetics. Dr. Erlich’s research focuses on the development of new algorithms for high throughput sequencing, including compressed sensing approaches to identify rare genetic variations, new algorithms for personal genomics, and the use of Web 2.0 information for genetic studies. He is the recipient of the Harold M. Weintraub award and the IEEE/ACM-CS HPC award. He also was selected to serve on 2010 Tomorrow’s Genome Technology team.

**Overview:** Data sharing is complicated by the risk of re-identification of de-identified genomic samples and the lack of policies and procedures to protect against genomic breaches and to ensure data provenance as data is repurposed and reused. Incentives are needed to promote data sharing and encourage compliance with policies and procedures related to privacy, security, and provenance.

Numerous challenges relate to data sharing, including issues of privacy, security, and provenance. Because genomic data are often integrated with data from other sources, including administrative data and data from public databases, issues related to the de-identification/re-identification of data become prominent. Indeed, it is now possible to “hack” de-identified genomic data through the use of public databases. For instance, public genealogy databases (e.g., ysearch, www.ysearch.org; smgf, www.smgf.org) typically use the genotype of the Y chromosome to infer surnames, which are most often paternally inherited. A recent, highly publicized study demonstrated that the identity of de-identified male research subjects could be determined with a high degree of certainty by combining surname inferences derived from these recreational genealogy databases with additional data linked to each subject such as state of residence and age—data elements that are not classified as Protected Health Information (PHI) by the Health Insurance Portability and Accountability Act (HIPAA) or the Institutional Review Board (IRB) and so can be included in “de-identified” research databases (Gymrek et al., 2013; Figure 6, page 11).

Legal issues come into play when genomic information is shared with patients or other users of the data and the uncertainties and known problems with the data are not made clear. Data reuse is also problematic in that privacy policies often supersede efforts to share data for secondary analysis. For example, dbGaP does not permit authorized data users to submit secondary data derived from dbGaP data.

The real possibility of re-identification of genomic samples has generated arguments in favor of the open sharing of genomic data and a move away from informed consent, with the idea being that the benefits greatly outweigh the risks (McEwen et al., in press). Existing policies and procedures may need to be modified to better protect against the risk of re-identification of genomic samples; alternatively or in addition, new policies and procedures may need to be developed (Gymrek et al., 2013). Indeed, the true risks of genomic privacy breaches are likely unknown at present.

In 2005, a clever 15-year-old boy who was conceived via an anonymous sperm donor became the first person to use data from recreational genealogy databases and other public sources to breach anonymity and successfully identify his biological father. —Stein, The Washington Post, 2005

An additional challenge is that data are often maintained in “data silos,” with unresolved issues related to data provenance and sharing. Indeed, the academic incentive system discourages data sharing because of a focus on in-
individual advancement through publications and funding, rather than collective achievement through the dissemination of data (Howison and Herbsleb, 2011).

The proprietary interests of health care systems likewise discourage data sharing. While federally funded researchers are encouraged to deposit genomic data into databases such as dbGaP, breaches in embargo periods (during which time external research groups cannot access an individual researcher’s data) can be detrimental to an academic career (Holden, 2009). Publications and database documentation rarely describe in detail the data collection, processing, and analysis steps in order to enable the reproduction of quantitative results to support published conclusions.

A related challenge is that federal agency members, industry heads, scientific journal editors, and academic investigators have not partnered to develop the policies and infrastructure to guide and support the long-term maintenance of either primary data or secondary data sets. dbGaP is an exception, but as noted, it is not without problems, including issues related to data access and reuse.

In terms of security, remote access to genomic data is often necessary but difficult to secure and protect. The incentives for maintaining high standards in privacy are often not clear or fully appreciated. In this regard, the distinction between the use of genomic data by health care providers and covered entities (under HIPAA regulation; US DHHS, http://www.hhs.gov/ocr/privacy) versus non-covered entities (under IRB regulation; US DHSS, http://www.hhs.gov/ohrp/assurances) often is not well understood, particularly among clinicians who are used to accessing PHI as part of their customary patient care.

Even when privacy issues are fully understood, compliance and enforcement are difficult to maintain. For example, controlled access to data (i.e., authorized access, encryption) and a secure virtual workspace or network do not prevent data transfer away from the secure workspace (i.e., printing, downloading, email, etc.)—a common problem and a major source of security breach (Shoffner et al., 2013).

A related issue is that the owner of the data often determines the policies and procedures for privacy and security; however, if ownership is transferred, the new owners may not adopt the same policies and procedures. Analyses of costs, risks, and benefits of privacy policies and procedures are rarely considered before their implementation, but these are important considerations that factor strongly into any privacy and security plan (Shoffner et al., in press). A certain level of risk of a breach may be acceptable, for instance, if the cost, time, or impact on the speed of science is unacceptable or unreasonable with a suggested privacy and security plan.

Figure 6. Re-identification of Sequenced Subjects through Surname Inferences Derived from Genealogy Databases. The figure shows three pedigrees, for which recovered surnames and Internet resources, including genealogy databases, were used to trace distant patrilineal relatives (marked as arrows in red patrilineal lines) to sequenced de-identified subjects (filled black squares) for re-identification. (From Gymrek et al., Science 339:321-324, 2013. Reprinted with permission from AAAS).
Bioethical and legal issues are critical to evaluate when considering genomic data because genomic data are accruing more quickly than expert consensus is being reached on the interpretation of those data (McEwen et al., 2013). Perhaps more significantly, technological advances now permit the identification of detrimental genomic variants for clinical diseases for which there are no available cures or treatments. Whether this information should be shared with patients is a topic of considerable debate. For example, Alzheimer’s Disease, even the early-onset form, is typically asymptomatic until middle age (and generally much older); yet, genetic testing for increased risk of Alzheimer’s Disease is possible at any age (Brauser, 2012). The risks and benefits of sharing information regarding increased genetic risk need to be carefully considered for diseases such as this one that do not become symptomatic until later in life, do not have a cure, and have only minimally effective treatments. Also unclear is how to deal with incidental findings—evidence for increased genetic risk that is identified as part of a general screen for a different genetic disease (McEwen et al., 2013). If the decision is made to withhold a genetic test result from a patient, then the challenge becomes determining the technical steps that should be taken to ensure that the data are not inadvertently released. For example, the data could be excised from the original data set, excised from the final analysis set, or maintained in the final analysis set but masked from view by the physician.

Additional challenges in bioethics include the fact that the field of genomics is not without error, and false positives and false negatives are both possible. Conveying this possibility to patients can be difficult and may lead to either a false sense of security regarding genetic risk or additional medical testing and/or treatment, which may not be necessary and may even produce more harm. For instance, the incidence of prophylactic mastectomies is rising in part because of women’s fears when confronted with a genetic test result that indicates an increased risk (but not 100% certainty) of developing breast cancer (Bankhead, 2012).

Bioethical issues also arise when considering the sharing of patient data and patient privacy. Determining who should have access to a patient’s risk of developing a genetic disease can be difficult. For example, consider the unborn child. In a few years, expecting parents will be able to order, for ~$1000, a complete genetic transcript of their child’s DNA, with currently available information on genomic variants and associated phenotypes (Scientific American Staff, 2013). But should those results be revealed? After all, without proper guidance, uninformed parents could easily misinterpret such test results, leading to psychological distress, unnecessary abortions, increased health care costs, and a variety of other problematic outcomes.

Legal considerations add even greater complexity to any discussion of bioethics in genomics. There is a legal distinction between real property (e.g., a blood sample) versus intellectual property (e.g., gene patent) versus informational property (e.g., genetic code?). An individual has legal rights and protections regarding the first two types of property, but at present, the law does not provide protections for informational property rights. To date, Roe versus Wade (1973) remains the only legal case regarding privacy rights related to one’s body. Indeed, genomic data are commonly used without consent (or even an individual’s awareness) in criminal investigations. For example, a discarded cigarette butt is routinely used as a source of genomic DNA from a criminal suspect. This activity is protected by the Fourth Amendment, which states (as defined in part by extra-constitutional law) that an unreasonable search occurs only when the government intrudes upon a legally cognizable interest of an individual. In the example of a discarded cigarette butt, the cigarette butt and any DNA that is removed from it are considered “abandoned” property. If DNA is abandoned, then an individual gives up his/her legally cognizable interest and the Fourth Amendment no longer offers protection against a search for genomic information. The Supreme Court’s ruling in California versus Greenwood (1988)—that a search of abandoned personal possessions (e.g., cigarette butt) does not constitute an unreasonable search—has generally been upheld by courts of law. Perhaps surprisingly in this era of rapid advances in genomics, 21 states and Washington D.C. do not have any laws in
place regarding covert DNA collection; the remaining states have only context-specific statutes.

The Fourth Amendment of the U.S. Constitution: “The right of the people to be secure in their persons, houses, papers, and effects, against unreasonable searches and seizures, shall not be violated, and no warrants shall issue, but upon probable cause, supported by oath or affirmation, and particularly describing the place to be searched, and the persons or things to be seized.”

Distinctions also need to be made between privacy versus confidentiality versus security. These terms carry very different meanings in the context of the law than they do in the context of technology, research, and medicine. Very few legal protections surround an individual’s rights to privacy or security as they relate to genomic information. HIPAA Privacy Rule protections and regulatory requirements apply to genomic data that is acquired as a result of health care services, but they do not apply to genomic data arising in other contexts, such as participation in research outside of clinical settings or from direct-to-consumer genomic testing. HIPAA also permits the sharing of genomic data under the same terms that it permits the sharing of “conventional” health care records, including its use for research (US DHHS, http://www.hhs.gov/ocr/privacy).

Society as a whole has not had an active discussion on why it is problematic to make genomic sequences public information. In other realms of society, there is very little privacy or confidentiality, and it is acceptable to share personal information. For example, photos are routinely posted on Facebook; personal location is commonly broadcast using Waze; and thoughts are publicly shared via Twitter. Yet society overwhelmingly regards genomic information as private. Why genomic information should be treated differently than other forms of personal information is a question with many possible answers. Perhaps people view their DNA as somehow part of their individual. Perhaps they fear misuse of their genomic data. Misuse of genomic information (e.g., denial of health insurance coverage) should be prohibited, but whether that translates into a general right to privacy protected under the law is for genomic information is not clear. These and related bioethical and legal issues have yet to be fully vetted by stakeholders and society as a whole.

Key Recommendations for Genomics

After discussion of the many challenges inherent in genomics, the Leadership Summit participants focused on a few critical challenges and reached consensus on the key recommendations to address those challenges and thereby move the field forward.

Specifically, the Consortium recommends the actions listed below.

Recommendations to foster interdisciplinary collaboration and coordination of efforts in genomics

• Efforts to create a “Consortium of Consortia” need to be fostered to centralize the efforts and resources of competing or parallel groups and thereby eliminate redundancies and more quickly move the field forward (e.g. Worlwide Genome Web [Maltbie, 2013]).
• The activities of individual groups in the development of standards for variants and phenotypes (e.g., ClinVar, PhenX, EMERGE, CRVR) must be coordinated to allow for crowdsourcing, industry involvement, and international input.
• Greater collaboration is required among data scientists, information technologists, genomic researchers, and geneticists to further efforts at software development (e.g., hackathons) for research and clinical applications.

Recommendations to advance analytical approaches and tools for widespread adoption of standards and federated distributed data systems, harmonization of existing data sets, integrated analyses, data repurposing and reuse, and discovery science

• Phenotypic data elements need to be standardized and existing phenotype data sets must be harmonized to facilitate analysis and promote data sharing, repurposing, and reuse.
• New tools need to be developed to enable integrated analyses of data on genomic variants and phenotypes; these tools need to be capable of extracting phenotypic data elements from existing data sets for use in hypothesis-driven research and from real-world data sets for use in discovery science in order to identify new, previously unrecognized phenotypes patterns in data derived from genomic data sets and other heterogeneous types of data sets.
• New or recoded software algorithms and computational approaches need to be developed to better structure, manage, and analyze massive data sets and metadata over time as data are shared, analyzed, applied, and reused; funding programs such as the joint Big Data program sponsored by the National Institutes of Health and the National Science Foundation represent only an initial step in this direction.
• Federated distributed data systems (e.g., iRODS, Data Verse) must be adopted across research groups, with trust fabrics in place to ensure quality of and provenance over data and to facilitate analysis.
Recommendations to promote data sharing while maintaining privacy, security, and provenance through incentives, new technical solutions, and cost-benefit analyses of different technical approaches to data sharing

- Incentives to promote data sharing need to be developed for journals, academic institutions, hospitals, governmental agencies, and industries; yet provenance over data sets must be preserved even as data sharing is promoted.

- Incentives (e.g., bail bonds for access to genomic data, automated audit trails, extensive annotation of data sets) need to be implemented to ensure data provenance and compliance with any adopted data sharing policies and procedures.

- Exploration and further research are needed on the best technical solutions (e.g., homomorphic encryption, multi-party computing) for data sharing schemes and the development of new schemes to enable the sharing of user-generated content (e.g., PatientsLikeMe) for use in genomics and other areas of biomedicine and social science.

- A thorough cost-benefit analysis of controlled versus open access data sharing policies and procedures must be conducted, with the caveats that the impact of genomic privacy breaches may not be fully understood at present, the analysis may be case- or institution-specific, and in some instances it may be necessary to allow for the re-identification of patients (e.g., to promote the development of new technologies for data sharing).

Recommendations to develop automated, easy to use, stakeholder-driven, open source, Clinical Decision Support systems

- Clinical Decision Support systems need to be developed such that genomic information is presented in a manner that is quick and easy to synthesize and apply in the context of personalized patient care (e.g., visual analytics, computable documentation); these systems need to be automatically updated as new genomic information becomes available.

- Clinical Decision Support systems must reflect and incorporate the perspectives and needs of not only the clinician but also the patient, family, and other stakeholders (e.g., policy makers, health care insurance providers, industry leaders).

- Clinical Decision Support systems should be open source and developed with wiki-like capabilities and/or through crowd-sourcing.

Recommendations to cultivate education and training initiatives in big data–based information technology, digital archiving, and analysis

- New education and training initiatives need to focus on big data–based information technology, digital archiving, and analysis.

- A data archivist should be included as an integral part of research study teams to assist with the standardization of data elements, harmonization of existing data sets, and curation of data.

- Existing training programs in medicine, social science, and biology need to incorporate training on basic biostatistical concepts in data analysis and visualization.

- Training should be emphasized and enabled across a broad spectrum of specialties and stages of training and career development.

Recommendations to address bioethical and legal policies on issues such as incidental findings and the distinction between fair use and misuse of genomic data

- The National Consortium for Data Science should serve as the central forum for discussions related to bioethics and the law in the context of genomics.

- An open dialogue needs to take place on the bioethics of incidental genomic findings; this discussion needs to involve all key stakeholders (e.g., genomic researchers, clinicians, health insurance representatives, health policy makers, patient advocacy groups).

- “Fair use” and “misuse” need to be defined in relation to the sharing of genomic data and reflect the perspectives of all key stakeholders.

Conclusion

The National Consortium for Data Science is committed to broadly advancing the field of data science. The Annual Data Science Leadership Summit, this year focused on genomics, represents the Consortium’s first effort to bridge communities in academics, industry, and government and convene top leaders in data science in order to identify challenges and recommendations to move the field forward. Widespread adoption of the recommendations put forth by the summit participants will ensure advances in genomics and data science. Several of the recommendations will require a significant amount of funding, and the Consortium encourages funding agency engagement on how to best support the recommended initiatives. Other recommendations will require changes in cultural norms and open discussions on sensitive issues. We expect future Leadership Summits to likewise stimulate energetic discussion and lead to the development of new recommendations in other areas of data science. The National Consortium for Data Science supports these and other initiatives in data science and encourages the involvement of local, state, national, and international entities.
Acknowledgements

The National Consortium for Data Science acknowledges the generous support of the following Leadership Summit sponsors: North Carolina Biotechnology Center; RENCI; RTI; UNC-Chapel Hill; UNC-Charlotte; and UNC General Administration. We also acknowledge the following persons for organizing and planning the Leadership Summit: Karen Green; Elyza Haley; David Knowles; Chris Lenhardt; and Jennifer Resnick. We value the detailed speaker notes provided by Kerry Dorr and Lindsay Rizzardi of the Curriculum in Genetics and Molecular Biology, Joshua Welch of the Department of Computer Science, and Zhuzhu Zhang of the Curriculum in Bioinformatics and Computational Biology at UNC-Chapel Hill. Finally, we acknowledge Josh Coyle and Joe Hope for providing multimedia coverage of the summit and graphics support, Patrick Hardison for providing fact-checking and proofreading, Karamarie Fecho, PhD, for providing writing support, Karen Green for providing editorial and design support, and RENCI for providing funding for this support.
References


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


The National Consortium for Data Science (NCDS) is dedicated to advancing the field of data science through the development of education and training initiatives, the provision of a unique research infrastructure to test new theories and techniques in data science, the translation of research innovations into economic gain, and the organization of tasks forces, summits, and workshops to guide public policy through thought leadership.

**Founding Member Institutions:**

- Cisco
- Drexel University
- Duke University
- GE
- The Hamner Institutes for Health Sciences
- IBM
- MCNC
- National Institute for Environmental Health Science
- NetApp
- North Carolina State University
- Renaissance Computing Institute (RENCI)
- RTI International
- SAS
- Texas A & M University
- UNC-Chapel Hill
- UNC Charlotte
- UNC General Administration
- US Environmental Protection Agency

**Leadership Summit Speakers:**

**Facilitators:**

- Stanley C. Ahalt, PhD, Director of RENCI and Professor of Computer Science, UNC-Chapel Hill
- Jim Evans, MD, PhD, Bryson Distinguished Professor, Departments of Genetics and Medicine, Director of the Cecil G. Sheps Center for Health Services Research, UNC-Chapel Hill

**Keynote Address:**

- Eric Green, MD, PhD, Director of the National Human Genome Research Institute, Acting Associate Director for Data Science, National Institutes of Health

**Evening Address:**

- John Wilbanks, Chief Commons Officer at Sage Bionetworks, Senior Fellow at the Ewing Mation Kauffman Foundation

**Session Speakers:**

- Yaniv Erlich, PhD, Principal Investigator and Andria and Paul Heafy Family Fellow, The Whitehead Institute
- Nita Farahany, JD, PhD, Professor of Genome Sciences & Policy, Duke University
- Leslie Lange, PhD, Professor of Genetics, UNC-Chapel Hill
- Dan Maltbie, Chief Technology Officer, Annai Systems
- Martin Reese, PhD, Chief Executive Officer, Omicia
- Kirk Wilhelmmsen, MD, PhD, Professor of Genetics and Neurology, UNC-Chapel Hill; Chief Domain Scientist for Genomics, RENCI

**Leadership Summit Planning Committee:**

- Stanley C. Ahalt, PhD, Director of RENCI and Professor of Computer Science, UNC-Chapel Hill
- Jim Evans, MD, PhD, Bryson Distinguished Professor, Departments of Genetics and Medicine, Director of the Cecil G. Sheps Center for Health Services Research, UNC-Chapel Hill
- Geoffrey S. Ginsburg, MD, PhD, Director of Genomic Medicine, Duke Institute for Genome Sciences & Policy, Executive Director of the Center for Personalized Medicine, Duke Medicine, and Professor of Medicine and Pathology, Duke University Medical Center
- Jason Lieb, PhD, Professor of Biology and Director of the Carolina Center for Genome Sciences, UNC-Chapel Hill
- Dan Masys, MD, Professor of Biomedical and Health Informatics, University of Washington
- Howard McLeod, PharmD, Fred Eshelman Distinguished Professor, Eshelman School of Pharmacy, Director of the Institute for Pharmacogenomics and Individualized Therapy, Professor, Department of Medicine, UNC-Chapel Hill
- Charles Schmitt, PhD, Director of Informatics and Data Sciences, RENCI
- David Threadgill, PhD, Professor and Chair of the Department of Genetics, NC State University
- Kirk Wilhelmmsen, MD, PhD, Professor, Department of Genetics and Neurology, UNC-Chapel Hill; Chief Domain Scientist for Genomics, RENCI
Leadership Summit White Paper Editorial Board:

Stanley C. Ahalt, PhD, Director of RENCI and Professor of Computer Science, UNC-Chapel Hill

Jim Evans, MD, PhD, Bryson Distinguished Professor, Departments of Genetics and Medicine, UNC-Chapel Hill, and Editor-in-Chief of Genetics in Medicine, the official journal of the American College of Medical Genetics

Yaniv Erlich, PhD, Principal Investigator and Andria and Paul Heafy Family Fellow, The Whitehead Institute

Nita Farahany, JD, PhD, Professor of Genome Sciences & Policy, Duke University

Geoffrey S. Ginsburg, MD, PhD, Director of Genomic Medicine, Duke Institute for Genome Sciences & Policy, Executive Director of the Center for Personalized Medicine, Duke Medicine, and Professor of Medicine and Pathology, Duke University Medical Center

Leslie Lange, PhD, Professor of Genetics, UNC-Chapel Hill

Jason Lieb, PhD, Professor of Biology and Director of the Carolina Center for Genome Sciences, UNC-Chapel Hill

Dan Maltbie, Chief Technology Officer, Annai Systems

Dan Masys, MD, Professor of Biomedical and Health Informatics, University of Washington

Howard McLeod, PharmD, Fred Eshelman Distinguished Professor, Eshelman School of Pharmacy, Director of the Institute for Pharmacogenomics and Individualized Therapy, Professor, Department of Medicine, UNC-Chapel Hill

Martin Reese, PhD, Chief Executive Officer, Omicia

Charles Schmitt, PhD, Director of Informatics and Data Sciences, RENCI

Teresa Squire, PhD, Senior Program Analyst, Science and Technology Development, NC Biotechnology Center

David Threadgill, PhD, Professor and Chair of the Department of Genetics, NC State University

John Wilbanks, Chief Commons Officer at Sage Bionetworks, Senior Fellow at the Ewing Marion Kauffman Foundation

Kirk Wilhelmsen, MD, PhD, Professor of Genetics and Neurology, UNC-Chapel Hill, Chief Domain Scientist for Genomics, RENCI

For additional information, please contact:

Stanley C. Ahalt, PhD
Renaissance Computing Institute
Telephone: 919-445-9640
Fax: 919-445-9669
Email: ahalt@renci.org