Workshop Report
Future Opportunities for Genome Sequencing and Beyond:
A Planning Workshop for the National Human Genome Research Institute
July 28-29, 2014

Executive Summary:

On July 28th and 29th, 2014, the National Human Genome Research Institute (NHGRI) convened a workshop to discuss scientific questions and opportunities that can be substantively addressed by large-scale genomics studies, and to consider options for future NHGRI programs in this area.

The workshop highlighted several key scientific opportunities for attention. First, using genomic sequencing to determine genetic variants underlying human disease and healthy traits, including both Mendelian conditions and complex diseases, continues to be an important activity, and will need to be addressed at scale. Second, improving our understanding of the impact of variants through functional genomics studies is critical to inform gene-disease relationships. Third, comparative and evolutionary genomic sequencing is still needed to inform the prioritization and interpretation of genetic variants. Fourth, translating genomics to medical practice will require a critical evaluation of the clinical utility of sequencing and approaches to clinical implementation. Finally, NHGRI should foster a "virtuous cycle" between discovery and clinical applications to accelerate our understanding of genotype-phenotype relationships, and their translation to genomic medicine. NHGRI needs to continue to support a critical mass of activity in all of these areas, while facilitating coordination across domains. Moving forward on these areas will require synergistic work amongst NHGRI program areas that are now largely distinct. Moreover, NHGRI may need to redefine boundaries and connections with other Institutes or funders as projects become more relevant to specific diseases and research domains. These opportunities make for an exciting, as well as a challenging, time to be evaluating future directions for large-scale sequencing efforts.

NHGRI has been a leader in creating ontologies and standards, and should continue to lead in this area. Given the large amount of sequencing occurring in the world, NHGRI could undertake steps to catalog different sequencing efforts, and to ensure that data are shared in truly useful ways. Given that raw data cannot always be shared, it is also important to improve interoperability and collaboration. Through these types of efforts the collective sequence data become a much more powerful resource for everyone.

Given the large number of research questions identified at the workshop, NHGRI should focus on ways to be trailblazing and catalytic. The Institute cannot tackle all questions in genomics, and should instead define and develop exemplar projects. Such projects should provide models, methods and resources that can influence and enhance the research being done in the larger scientific community. Further, given the scope of work involved in elucidating the genetic basis of human disease, NHGRI needs to develop partnerships that integrate and fully engage relevant scientific experts, funding sources and communities.

Main Report:

On July 28- 29, 2014, the National Human Genome Research Institute (NHGRI) convened a workshop to obtain opinions from the scientific community on future directions for the NHGRI Genome Sequencing Program (GSP; http://www.genome.gov/10001691). A participant list is in Appendix 1 (App. 1) and presentations are available on the NHGRI website (http://www.genome.gov/27558042). The previous evaluative workshop for the GSP was in 2009. The intervening five years have seen numerous advances in genomics: the continued plummeting cost of genome sequencing has multiplied the scale and scope of questions that can be addressed; clinical sequencing is expected soon to dwarf sequence data generated in research settings; novel methods for functional studies are emerging, several of which are primed for large-scale efforts. This changing and complicated landscape makes for an opportune time to sharpen thinking and identify new opportunities in alignment with the 2011 NHGRI Strategic Plan.

The current GSP programs are the Large-Scale Genome Sequencing and Analysis Centers (LSAC), the Centers for Mendelian Genomics (CMG), the Clinical Sequencing Exploratory Research (CSER) program and the Genome Sequencing Informatics Tools (GS-IT) program. Traditionally, NHGRI GSP programs are characterized as large-scale, highly-managed, consortia-oriented projects that generate community resources and advance technology while focusing on scientifically and medically relevant research areas. Prioritizing new opportunities for such projects will ensure that NHGRI continues to lead in the development of important resources, technologies, laboratory methods, data analyses and other genomic approaches. Fostering advances in these areas is critical for advancing our understanding of genomic biology and improving human health.

The objectives of the meeting were to:

1. Identify the scientific questions and opportunities that can be substantially addressed by large-scale genomics studies, starting with genome sequencing but also considering other genomic technologies.
2. Consider options for future NHGRI programs that would address these questions and opportunities.

The workshop was intended as a forum for strategic input for future directions, rather than a referendum on the current GSP. NHGRI was particularly interested in examples of grand opportunities for a flagship NHGRI program; input on what the Institute is not doing, but should be doing; and discussion of the balance and benefits of large-scale, consortia-based pursuits. The largest emphasis was given to discussion of disease variant discovery -- areas associated with the current LSAC and CMG programs -- although the workshop also included discussions about clinical sequencing, informatics and analysis, and functional genomics, all areas of high interest to other NHGRI programs (e.g., GS-IT, CSER, Functional Genomics, and others). Areas associated with clinical sequencing, and CSER in particular, will require additional longer-term planning, encompassing broader consideration of other NHGRI genomic medicine and ethical, legal and social implications (ELSI) activities. Genome informatics, embodied in the current GSP by GS-IT, also requires additional planning involving other efforts in data science and informatics, including the NHGRI bioinformatics portfolio and the Big Data to Knowledge (BD2K) initiative (see App. 2 for a brief description and web links for National Institutes of Health [NIH] and NHGRI programs mentioned in this summary).

The meeting agenda (App. 3) was divided into three primary sections. Part 1 was a series of presentations and discussion centered on the science of the current NHGRI GSP, including the consequences should NHGRI not pursue these areas. In Part 2, break-out groups discussed goals and tactics for "Big Challenges" that NHGRI could pursue in the next five years in four scientific areas: i) genetic architecture of health and disease; ii) integrating genomic variant discovery with functional analysis; iii) clinical genome sequencing; and iv) comparative and evolutionary genomics. Part 3 focused discussion on options for NHGRI to implement organized, well-coordinated scientific efforts addressing the ideas raised earlier in the meeting. There were also five "Challenge Talks"(summarized in App. 4) where speakers addressed the charge: "If you had ~$10-20 M per year for ~5 years, what highly significant, innovative idea would you

pursue, perhaps grounded in genome sequencing, involving large-scale data production and analysis, but stretching the current NHGRI boundaries?"

At the outset of the workshop, Dr. Eric Green, Director of NHGRI, noted that the workshop agenda was designed to elicit ambitious ideas, and recognized there are numerous opportunities and paths forward. NHGRI is small compared to the rest of NIH, and the world of genomics. In order to maximize impact of its future large-scale genomics projects, NHGRI will need to develop models for partnerships, co-funding and cost-sharing. Dr. Green also noted that NHGRI will incorporate feedback from this workshop in programmatic discussions in the overall context of the NHGRI Strategic Plan and larger Extramural Research Program, in consultation with the National Advisory Council on Human Genome Research (NACHGR).

This workshop summary is organized in three sections: 1) the scientific areas listed above, focusing on goals, tactics and recommendations, 2) programmatic suggestions for moving from ideas to implementation, and 3) additional cross-cutting themes and recommendations raised in the large group discussions.

Section 1: Goals and Recommendations by Scientific Area

Genetic Architecture of Health and Disease at Scale

As noted in the NHGRI Strategic Plan, "Genomic sequencing could be used to determine the genetic variation underlying the full spectrum of disease from Mendelian to common complex diseases." Reaching this goal also requires a thorough understanding of the full spectrum of allele frequencies and types of variation.

"Discovering Variants Conferring Risk for Common Diseases" presented by Michael Boehnke

- In the next five years we should aim to advance our understanding of the genetic basis of human disease and use that knowledge to improve human health.
- Genome-wide association studies (GWAS) have provided substantial experience and success in understanding common variants associated with disease; however, the variants identified explain only a portion of disease heritability and the biological mechanism for many of these findings is still unknown.
- Large sample sizes are essential to detect the full spectrum of genetic variants associated with a single common disease. A reasonable starting estimate is sample sizes of 25,000 cases and 25,000 controls.
- Careful selection of appropriate and novel study designs is also an important way to maximize power.
- Many human diseases command our attention due to frequency, severity or cost burden. Prioritization for a limited number of exemplar diseases could include: large numbers of well-phenotyped, broadly consented individuals; organized investigator groups with a track record of data sharing and collaboration; and financial support from other funders.
- Failure of NHGRI to continue in this area would be a lost opportunity. More fragmented efforts would slow progress, and lead to less data and information sharing, including less interoperability of the data.

"Discovering the Genomic Bases of Mendelian Diseases" presented by Roderick McInnes

- Mendelian conditions are individually rare, but collectively they have a large impact on human health, affecting over 25 million Americans at a cost of $5 million per affected person over their lifetime.
- The goal of identifying the mutations underlying all Mendelian conditions - the great majority of which are loss-of-function - is the human equivalent of the mouse knockout project, and is of great importance to understanding of both human biology and genomic medicine.
- The NHGRI CMGs have demonstrated the power of high-throughput sequencing, using both phenotype driven and genotype driven approaches, for disease gene discovery. This work enables diagnostic and predictive testing, and improves our understanding of pleiotropy and genetic heterogeneity.

- Diagnosing Mendelian conditions at the molecular level is beneficial to patients and their families. Understanding the genetics and biology underlying the disease leads to more accurate genetic counseling and to potential improvements to prevention, prognosis and targeted therapies.
- Although the progress to date by the CMGs is impressive, the genetic basis of over 3,700 Mendelian conditions is still unknown. As many as 17,000 genes remain as candidates for Mendelian conditions.
- Understanding Mendelian conditions contributes to our understanding of complex diseases, and identifying genes for Mendelian conditions may lead to drug targets applicable to a broader range of patients.

Goals, opportunities and recommendations from the talks, break-out group and discussions included:

1. **Define the genotype-phenotype relationship underlying human diseases and healthy traits, as a means to illuminate pathophysiology as well as the fundamental molecular, cellular and organismal functions of all human genes and functional sequences.**
    - The role of NHGRI is to advance paradigms, develop and evaluate methods and tools, establish foundational resources, and foster partnerships with categorical Institutes and other relevant parties.
    - NHGRI cannot study all common diseases; instead efforts should focus on "exemplar" conditions that represent the spectrum of health and disease related phenotypes. Phenotypes should be selected to span a spectrum of disease classifications: rare and common disease; pediatric and adult conditions; and psychiatric, metabolic, developmental and infectious conditions. Consideration should be given to molecular, intermediate and clinical phenotypes. Ancestrally diverse populations should be included. Examples should be selected to enable testing and pioneering of multiple research paradigms, such as assessment of different study designs, the role of "other-omics", and comparison of exomes with whole genome approaches.
    - Large sample sizes are needed for exemplar studies of common diseases. It is not sufficient for power to be just adequate; the studies need high power for discovery of disease-associated variants within disease subtypes and across allelic spectra. It would be better to perform a smaller number of comprehensive, high- powered studies, rather than make compromises through a larger number of underpowered studies. Consideration should be given to novel designs and approaches to improve power (N.B. Nancy Cox presented an example of a novel design in her challenge talk. See App. 4).
    - NHGRI should continue to foster a critical mass of excellence for both Mendelian and complex diseases. Different approaches and expertise are needed so the programs should not be combined. However, the programs should also not be in silos, and should be coordinated and integrated to promote synergistic approaches and scientific advances.
    - Mendelian and complex diseases fall on a spectrum, and there is value to exploring the full diversity of genetic architecture, including diseases resulting from *de novo* genomic events, phenotypes with incomplete or variable penetrance, and conditions with genetic and environmental modifiers.
    - NHGRI needs to pioneer and support sequencing in longitudinal cohorts for both discovery and clinical sequencing. Ideally, sequencing will be in cohorts of well-phenotyped individuals who are followed over time for outcome and who can be recontacted for deeper phenotyping as new information emerges.
2. **Create and make widely available the knowledge base needed to interpret genome sequence variation in life science, drug discovery, clinical prediction and diagnosis.**
    - NHGRI should continue to serve as a leader in data aggregating and sharing. This will require an appropriate recognition of the needed investment in data science, bioinformatics and engineering along with an appropriate recognition of data security and privacy.
    - Capitalizing on the large amount of sequence data being generated throughout the world requires new database infrastructure models that promote data and information sharing, including sequence, phenotype and ongoing clinical information. Although there is strong need for action, some participants cautioned

against premature consensus on a single data commons model and instead supported a federated model that implements and evaluates multiple interoperable models. The Global Alliance for Genomics and Health (http://genomicsandhealth.org/) is working in this area.

- One goal is to understand the impact of loss-of-function variants at every gene, including genes that can be knocked out without noticeable adverse effects. Although work in model organisms remains important (see comparative genomics below), not all human genes have equivalents in other organisms, and some effects are specific to humans. This should also address variants that only show an effect when exposed to additional factors (gene modifiers, pharmacogenomics variants, gene-environment interactions, etc.).
- National and international "matchmaker" services can link patients (and their doctors) to other patients with similar genetic variants and phenotypes. Other services (e.g., GeneMatcher developed by the Baylor-Hopkins CMG, see https://genematcher.org/) can link researchers across consortia. Such linkages facilitate discovery and clinical applications. These services need to be straightforward for physicians to identify and interact with, and interoperable with other sequencing efforts including those sponsored by NHGRI.
- NHGRI should encourage additional resources that provide allele frequency data, standardized annotations, and other relevant aggregate information about human genome variants.

3. **Include a range of human diseases and of populations to expand discovery, define architecture, and broaden access and as a matter of social justice.**
    - NHGRI needs to continue to foster well powered studies across populations, races and ethnicities to minimize disparities in who benefits from discovery and clinical sequencing. Focusing on broadly consented individuals may limit diversity, so NHGRI needs to seek novel ways to ensure balance.
    - Studies of multiple populations spanning the diversity of humankind allow for a better understanding of the allelic spectrum and genetic architecture of disease. Rare variants in one population may be common or absent in other populations. By capitalizing on population and disease genetic architecture, we can better elucidate the relationship of specific variants and genes with disease.

Integrating Genomic Variant Discovery with Function

As discussed in the NHGRI Strategic Plan, continued acquisition of knowledge on genome function is valuable for understanding the biology of genomes and the genomic basis of disease. Integrating variant discovery with analysis of genome function can help with prediction of the following: causal variants based on identified tag variants; target genes and cell types based on disease associations; and mechanisms by which pathogenic variants act.

"Functional Genomics at Scale" presented by Joseph Ecker

- A long-term goal of functional genomics is to decipher the rules by which genes and gene networks are regulated and to understand how such regulation affects cellular function, development and disease.
- No existing *sequencing* programs are examining function at scale; however, several existing and upcoming NHGRI (ENCODE, GGR, FunVar) and NIH Common Fund (Roadmap Epigenomics Project, GTEx, LINCS) programs are making important contributions to our understanding of genome function (See App. 2).
- One challenge is that regulatory elements can act over long distances, implying that regulatory variants do not necessarily target the nearest gene. There is an opportunity to characterize long-range chromatin interactions and regulation at scale, to learn the connectivity between genes and their regulatory elements.
- Promising emerging genome editing approaches, such as CRISPR/Cas9, could allow for rapid and novel assessments of variant function for molecular and organismal phenotypes (including disease).
- A recent paper on the molecular basis for blond hair color (Guenther, et al. *Nature Genetics.* 2014) demonstrates current challenges to identifying causal function for non-coding variants, and how they can be overcome by combining genetic association studies, large-scale genome annotation projects such as ENCODE, detailed functional tests of enhancer activity, and animal models of the phenotype of interest.

Goals, opportunities and recommendations from the talk, break-out group and discussions included:

1. **Define the molecular, cellular, organ and organismal functions of coding and non-coding genome sequences as foundational for biology and interpretation of genomes.**
    - NHGRI, together with other NIH Institutes and Centers (ICs), should develop and deploy assays that faithfully report disease-relevant functions at the variant, gene and pathway levels. Prioritization should be given to both coding and non-coding regions. This information will provide foundational resources that integrate functional information with disease- and health- related variants.
    - Function should be considered at two scales: 1) the molecular/biochemical and cellular scale, which is closer to DNA variants and easier to assay at large scale; and 2) the organ and organismal scale, which is ultimately the most biologically relevant level but is not as easy to scale, systematize or characterize. We need to evaluate how different molecular, biochemical, and cellular assays capture organismal and clinical outcomes. Assays that do not correlate with the organismal or clinical outcomes may result in incorrect assessments of putative function and can mislead scientists utilizing functional resources.
    - NHGRI should consider both function-first approaches, which initially develop catalogs of function/elements for all possible sequences and then cross reference them with variants in disease studies as the latter are identified, and variant-first approaches, which start with disease susceptibility/association sequences and then characterize them functionally. (N.B. Jay Shendure's challenge talk is a hybrid approach, and David Haussler's a variant-first approach. See App. 3).
    - Computational methods should be developed to predict accurately the molecular functions of variation in non-coding and coding sequences, along with models that reflect cellular responses to perturbations.
2. **Develop and make widely available tools to manipulate genomic sequences at scale and experimentally characterize their impact, with the goal of developing generalizable methods for large-scale functional characterization of sequence variants in faithful models.**
    - No matter how large the sample sizes become in clinical databases and disease discovery cohorts, causal variants cannot be definitively identified solely on the basis of statistical associations. Supporting data will be needed to prioritize associations for further experimental testing to determine causality. NHGRI has the opportunity to lead the field in developing new ways to measure function and use the resulting information to inform disease discovery, perhaps resulting in better identification of clinically actionable variants.
    - NHGRI should raise the technical challenge on how to scale up the most important functional assays (i.e., take tests for small numbers of regions and scale them to "whole genome" and take tests for small numbers of individuals and scale them to whole populations), while maintaining assay validity. Large-scale assays should be benchmarked against deep and detailed functional studies in specific diseases or domains. NHGRI could facilitate this work by partnering with domain experts who understand the complexities and subtleties of these "gold standard" assays and diseases.
    - Initial development of large-scale assays and tools will likely focus on the molecular level, but should also consider how to scale to organ, organismal and clinical levels. This will be facilitated by the development of faithful models and assays that correlate with organismal and clinical outcomes in the relevant tissues and individuals. Some assays will provide more provisional information, and balanced expectations are needed when considering what specific assays can and cannot tell us about function at different levels.
    - The field currently has an unsophisticated view of how proteins interact with our genome and should work to improve our knowledge base in this area.
    - NHGRI could help foster cellular assays and other models that allow us to test how drugs and other environmental agents interact with our genome.
    - Personal genomics can be expanded to include personal functional genomics, where the functions of variants are directly measured in clinical settings.

3. **Systematically catalog molecular components and their interactions, across cell fates and cell states.**
    - Functional genomics is valuable beyond simple characterization of variants. High-throughput sequencing can be used to characterize cell types (N.B. ENCODE and Epigenomics Roadmap are doing this, and another example was given in Aviv Regev's talk on The Human Cell Atlas Project. See App. 4). Work in this area is applicable to multiple categorical Institutes and/or the Common Fund.
    - The catalog of regulatory elements is not complete. Additional profiling of regulatory data needs to be done in key tissues and cell types, with a focus on cellular contexts most relevant to human diseases.
    - Function should also be considered at the pathway and systems biology level. This requires incorporating concepts of pleiotropy and epistasis, as well as interactions of variants and genes, rather than focusing solely on the impact of individual variants in a single disease.
    - In addition, the group recognized that genomic assays are informative for genetic, environmental, gene by environment, and microbiome studies; however, they decided the NHGRI remit was probably limited to consideration of genetic effects, and did not further consider the other topics.

Clinical Genome Sequencing at Scale

As noted in the NHGRI Strategic Plan "Genomic discoveries will increasingly advance the science of medicine in the coming decades, as important advances are made in developing improved diagnostics, more effective therapeutic strategies, an evidence-based approach for demonstrating clinical efficacy, and better decision-making tools for patients and providers."

"Genome Sequencing for Clinical Care" presented by Dan Roden

- NHGRI, in partnership with the Common Fund and other NIH ICs, is undertaking a number of efforts to facilitate translation of genome sequencing into clinical care. This includes the IGNITE, NSIGHT, eMERGE, UDN, ClinGen, Clinical Center Genomics Opportunity and CSER programs (see App. 2).
- These efforts have different focuses (discovery and clinical), methodological approaches and patient populations. Common issues include integrating genomics into clinical practice and the electronic medical record (EMR), return of results, defining actionability for targeted and incidental findings, best practices for data sharing, and understanding longitudinal impacts on patients and research.
- A paradox of precision medicine is that sequencing data needs to be generated in large numbers of subjects to interpret what is seen in individual patients.
- The 6th Genomic Medicine meeting in January 2014 demonstrated the value of international collaboration.
- It is imperative for NHGRI to take coordinated action in this area to maximize the benefits and minimize the risks associated with clinical sequencing.

Goals, opportunities and recommendations from the talks, break-out groups and discussions included:

1. **Define clinical contexts in which genome sequencing improves patient outcomes, including clinical validity and utility, value and cost effectiveness.**
    - There was discussion of NHGRI's role as sequencing becomes more wide-spread in clinical practice. The general consensus was that NHGRI grants should not be paying for clinical services in the provision of routine care. Instead the Institute should support catalytic research that advances the translation of genetic and genomic findings into clinical settings, again providing a model to be built on by other groups. Action in this area is needed to reach the goal of improving human health.
    - NHGRI should support research that demonstrates whether, and in what situations, genome-scale testing improves health. Widely employed tests should have a rational basis, and additional knowledge, data and experience is needed to inform decisions made by funders and regulators.

- An evidence-based paradigm is needed to demonstrate clinical utility, cost-effectiveness, and the overall value of implementing genomic sequencing in clinical settings. Work in this area will require partnering with disease experts. It will also require different types of studies and study designs, including randomized trials, implemented in a variety of clinical settings and across diverse populations.
- NHGRI should complement work in diagnostic clinical sequencing with research that addresses the role of sequencing for prevention, screening and other public health applications.

2. **Improve technical platforms to enable rapid, robust detection of all clinically relevant variation in a single test.**
    - Clinical sequencing will benefit from improvements in accuracy, expansion of mutation types detected, and decreases in cost and turn-around time. The private sector is also driving innovation in this area.
    - Identifying and assessing RNA/transcriptome variation may be relevant for some conditions.
    - The spectrum of tissues undergoing clinical sequencing should be increased, including circulating cell-free DNA, single cells and samples that will help improve our understanding of non-cancer somatic variation.
3. **Leverage clinical sequencing data for research use, leading to a learning system that improves our understanding of variant and gene disease relationships.**
    - NHGRI needs to position itself to positively influence the large amount of sequencing that occurs, and is increasingly going to occur, outside of NHGRI's purview in both the public and private sectors. NHGRI can make targeted contributions by improving data sharing, developing and cataloging tools, and modeling "exemplar" studies focused on clinical utility and implementation.
    - NHGRI should foster the "virtuous cycle" between sequencing in clinical practice and in discovery research. Research drives discovery and creates tools that improve diagnosis and translation. Complementary to this, clinical sequencing and clinical data can be harnessed to drive novel discovery and knowledge integration. Flow between the two areas is needed to maximize both translation and discovery.
        - NHGRI should help develop a multi-use longitudinal cohort of all patients undergoing clinical genome-scale sequencing. Ideally this resource will allow targeted re-phenotyping of individuals based on their sequence results. Distributed platforms need to be developed that make data sharing as easy as possible for busy laboratories and doctors.
        - NHGRI should also consider other approaches, including sequencing in populations with EMR records to collect targeted and genome-wide sequence data in well-phentoyped individuals.
        - Data and knowledge aggregation, which is crucial for variant interpretation, will be facilitated by improving sharing of clinical data, but must be balanced with patient and privacy concerns.
    - Physicians, patients and families should be engaged in all aspects of clinical genomics and genomic medicine. This will require improved education of the public on the value of genomics and health.
    - NHGRI should explore using the Children's Oncology Group (COG) as a model for genomic medicine. COG is an NCI-sponsored clinical trials cooperative group, comprising over 200 institutions with multidisciplinary teams, that conducts a spectrum of clinical research and translational research.
4. **Define robust approaches to determine the pathogenicity of genomic variants using genetic, functional, and computational data in a statistically valid framework.**
    - If clinical sequencing is implemented at scale, we cannot continue to rely on manual curation.
    - NIH should facilitate the development of standards for clinical genomic sequencing, variant annotation, interpretation and clinical delivery.
5. **Identify effective and efficient methods for implementing sequencing into routine medical practice, and further refine an NHGRI agenda for implementation research in genomic medicine.**
    - NHGRI should help develop novel clinical decision support tools for ordering and applying genomic information (N.B. an example was given in Challenge Talk by Dan Masys, see App. 4). When possible, these tools should incorporate point-of-care education for physicians.

- To broaden the population impact, sequencing that is shown to have clinical utility should be incorporated into a wider set of clinical and socioeconomic settings.
- NHGRI should connect with professional societies to provide appropriate expertise, guidance and data at the early stages of consensus development for guidelines.

Comparative and Evolutionary Genomics

As stated in the NHGRI Strategic Plan: "Ultimately, human biology must be understood in the context of evolution." Evolutionary processes are foundational to understanding variation in relation to human biology and disease.

Breakout chairs Andrew Clark and Evan Eichler highlighted significance and accomplishments in this area:

- Evolution and population genetics can provide an unbiased framework for the discovery and prioritization of genomic regions for genotype-phenotype correlations.
- NHGRI has been a trailblazer in comparative genomics, fostering increased scientific expertise, computational methods, community resources and collaborative consortia.
- Sequencing and alignment of primate and vertebrate genomes has led to identification of >3 million evolutionarily conserved elements and improved understanding of the origin of human and mammalian lineages.
- The Human Genome Reference Consortium has improved the human reference sequence; however, further work is needed for complex and intermediate sized structural variation and for representing human genome variation at all scales in a manner that is integrated with the reference genome.

Goals, opportunities and recommendations from the break-out groups and discussions included:

1. **Produce high quality de novo sequencing and assembly of genomes.**
   - NHGRI should advance sequence technologies to enable assembly of a novel genome for $10K at a quality that exceeds the current human reference assembly.
   - NHGRI should apply advanced and developing sequencing and mapping technologies to obtain ~50 human reference quality or better genomes that adequately sample known human diversity.
   - NHGRI should continue working toward a phased "telomere to telomere" genome assembly that provides a comprehensive assessment of all genetic variation, including intermediate-size and complex structural variation. Understanding the portions of the human genome and types of variation that is missed in production level mapping to the current reference genome, and overcoming this limitation, has large medical and human biology implications.
2. **Understand specific genomic changes in human and primate linages.**
   - NHGRI should sequence genomes from multiple primate species. This will allow inference of the origins, constraints and novelty of human and other primates' lineage-specific genome attributes.
   - Detailed evolutionary knowledge of our genome and that of our nearest evolutionary neighbors will be important for interpreting model organism studies, and this information can be cross referenced with functional characteristics of variants (see genome function discussion, above).
3. **Obtain nucleotide-level resolution of every conserved element in humans.**
   - In addition to primates, NHGRI should comprehensively sequence 200 non-primate mammals. Analysis of these data will allow us to understand origins, conservation and lineage-specific constraints on genomic elements down to the single base pair level across the mammalian lineage.
   - This information will be useful for the interpretation and characterization of disease-associated variants, and to obtain a more precise foundational understanding of the genomes of mammalian model organisms.
4. **Leverage model organisms for functional genomics.**

- Model organisms are still of enormous utility for understanding context dependence of variant functions. Sequencing of reference panels of model organisms will accelerate functional characterization, and improve our understanding of genotype/phenotype associations.
- Model organisms allow for the examination of identical genotypes in different environments, and facilitate studies of gene-environment interactions. The National Institute of Environmental Health Sciences (NIEHS) and National Science Foundation (NSF) are potential partners in this area.
- The scientific community should broaden the model organisms used in functional studies, since there is importance in considering a wide variety and diversity of models. Primates offer unique opportunities to address scientific questions relevant to human health, but also have inherent challenges.

5. Develop the informatics infrastructure needed to assemble, display and compare multiple genomes between and within species.

- Sequencing efforts in comparative genomics need to be organized as a resource for the full community and should be coordinated, so that investigators learn from one another and minimize duplication of effort.
- NHGRI should contribute to the development of better algorithms and interactive interfaces that help users assemble genomes from raw data, align them to each other, and understand and manipulate comparative genome data. Such interfaces include browsers that can provide representation of rearrangements and paralog/copy number differences between and within species and interfaces that can translate from differences at the DNA level to differences at the RNA, protein and higher functional levels.

Section 2: Moving from Ideas to Implementation

General

- NHGRI should continue to consider consortia models where grants have individual aims, but members form working groups and identify common elements and goals from the start of the program.
- NHGRI plays a critical role in keeping ELSI embedded in all of this activity.
- Although it is important to take action, we also need to caution against premature consensus, and should document and evaluate the strengths and weaknesses of different options.
- NHGRI has a demonstrated history of strong project leadership. Developing large-scale projects involving partnerships and cost-sharing with other entities will require significant program management and guidance. NHGRI should continue to value, support, recruit and retain excellent program officers.

Discovery and Clinical Sequencing

- For complex disease NHGRI should develop programs that have flexibility, but at the same time, have clear goals and scientific boundaries. Programs cannot be too vague, and need structure that allows for meaningful evaluation. In considering which disease and studies to include in large-scale variant discovery projects, NHGRI needs to develop a system that is nimble and responsive to ensure that large sequencing efforts are partnered with projects that have the right types of samples, and the right scientists engaged in all aspects of the project. This should be a transparent process that enables access and includes community input.
- The amount of exome and genome data generated in research and clinical settings is expected to continue to increase significantly. NHGRI cannot expect to be a primary driver of sequence production and capacity. However, the Institute can take a catalytic lead in setting standards, improving and implementing new test methods, disseminating and integrating information, and serving as a model for other groups. A suggestion is for NHGRI to fund multiple pilot activities to explore how to organize capabilities and resources.
- Consideration should be given to implementing foundational resources that make sequencing broadly useful for discovery and clinical applications. NHGRI needs to enhance public awareness of the process, progress and success of our programs. As one example, projects "building" for the future should be complemented by projects that can have fruitful products on a shorter timescale. Demonstrating impact and showing objective measures of progress are needed to maintain support of the rest of the community.

- The CMGs and CSER projects demonstrate advantages of diversifying sequencing capacity. These centers foster the development of specialized scientific expertise that complements more large-scale efforts.

Functional Genomics

- Large-scale centers with generalized capacity for assessing variant function are likely premature, because the most universally valuable data types and methods to take to scale are not yet known. For functional studies, there may be an advantage to leveraging existing consortia with expertise in specific tools, assays and approaches. Activity -- particularly standards, quality control, resource development and data sharing -- can be managed without being prematurely or overly prescriptive.
- NHGRI should foster new and emerging methods for assessing function. This might best be achieved through R01s attempting creative technology development, and may be a good place to partner with the National Institute of General Medical Sciences (NIGMS).

Section 3: Other Discussions, Cross-Cutting Themes and Recommendations

- The broad scope of challenges and opportunities raised at this workshop requires disease- and domain-specific knowledge and expertise beyond that of the NHGRI community. Further, NHGRI does not have the funds or capacity to tackle all of the topics discussed. For this reason, it is important that NHGRI facilitate partnerships with other Institutes, funders (i.e., the Patient-Centered Outcomes Research Institute (PCORI)) and the private sector, both nationally and internationally.
- NHGRI needs to take the title of the workshop to heart and expand beyond genome sequencing. This will include functional studies, as discussed, as well as opportunities for large-scale efforts in epigenomics and metabolomics.
- NHGRI should strive to put the "W" back in whole genome sequencing (WGS). This includes considering how the WGS as done today differs from perfection (i.e., phased telomere to telomere contiguity), articulating what is missing in current technologies, and specifying what could be enabled by alternative technologies. This topic is sometimes overlooked, since it is viewed as solved, or passé, but is an area that could benefit from more investment and creativity. Cases who currently go undiagnosed through routine sequencing for Mendelian disorders and other clinical phenotypes should be assessed with respect to the potential role of structural variants and other missing sequence data.
- NHGRI programs all require advances and investment in bioinformatics, biocomputing and data science. Needs in this area should be considered in the design of research programs and funding opportunity announcements. Projects need to have the right size and balance of expertise, along with sufficient resources and support, in bioinformatics, biocomputing and data science.
- Modern shared Application Programming Interface design methods will enable interoperability of data across projects. Genomics could learn from other fields, regarding how to achieve interoperability both within genomics, as well as beyond genomics to other disciplines.
- NHGRI should give sufficient attention to phenotype, including studying well-phenotyped individuals who are consented for follow-up phenotyping as needed. Phenotypic information should be collected in ways that can be shared and combined across studies.
- NHGRI needs to invest in more training, including increasing diversity in the people doing the research; providing opportunities for PhDs to get training in clinical activities; training medical students, fellows and practicing clinicians; increasing training in computational, statistical and quantitative genomics; and keeping training grants at a sufficient size to allow for the critical mass needed to propel success.
- Human disease is a perturbation of systems. Systems biology approaches should be incorporated into studies. The role of genetic variants needs to be appreciated in the larger context of the cell, the individual and the environment, which can be achieved through continued studies of gene-gene and gene-environment interactions and pathways.