NIH National Human Genome Research Institute

Report of NHGRI Strategic Planning Workshop:

"From Genome to Phenotype: Genomic Variation Identification, Association, and Function in Human Health and Disease" January 22-24, 2019

Executive Summary

As part of its 'Genomics2020' strategic planning efforts the National Human Genome Research Institute (NHGRI) convened a workshop to gather community input to: 1) help develop a strategic approach for the next 5-10 years for studies of genomic variation, function and association; and 2) discuss research directions that NHGRI should pursue. The workshop was videocast and can be viewed on Genome TV.

At a high level, there was strong support for the idea that relating genomic variation to phenotype should remain a priority research area for NHGRI for the next 5-10 years. There was strong support for mapping and functionally characterizing regulatory elements and genes, as well as performing large-scale perturbation experiments.

Strong support was voiced for a continued emphasis on technology development and computational genomics. This included improved methods for nucleic acid sequencing and synthesis, as well as the development of methods to analyze difficult genomic regions and haplotypes. There was an emphasis on the need for new and improved epigenomics, transcriptomics and multi-omics assays. Better methods are also needed for performing high-throughput studies of genome function. There was enthusiasm for improved modeling, as well as enhanced visualization and usability from data resources. There was also excitement about using computational approaches to identify the most informative experiments for interrogating the high-dimensional space of genome function analysis.

Strong support was voiced for an increased emphasis on applying functional genomics at scale, which would accelerate understanding of how genomic variants influence genome function. Suggestions included establishing a function for every gene, establishing a function for every regulatory element, and assessing 1% of known genomic variants for function. Suggestions also included a single-cell project to catalog association between DNA sequence variation and transcriptomic variation across individuals and across tissues/organs, as well as better comprehensive predictive modeling based on systematically collected functional genomics data.

Continued identification of disease variant associations was strongly supported; with NHGRI viewed as having a role in funding the analysis and interpretation of these data, especially methods development and establishing standards for data. Although opinions varied, continued NHGRI direct funding of large-scale genome sequencing was generally not seen as high priority. Debate centered on whether community-produced data would become abundant and freely shared for research or whether NHGRI-supported data generation was required for research purposes. Suggested projects included large-scale genome sequencing studies for complex diseases in collaboration with other funders, characterization of the genomic architecture of Mendelian and complex traits, a catalog of somatic variation, as well as the development of resources, methods, and approaches that would help gain a functional and mechanistic understanding of genomic variants, including results from both association and clinical studies.

Overall, this workshop highlighted that this is a critical and exciting time for genome sciences. Advances in technology, genomic sequencing, functional genomics, and computational genomics have created new opportunities for approaches to the characterization of genes, genomic elements, and genomic variants in relationship to human biology, health, and disease. The breadth and depth of topics covered solidify the need for NHGRI to define its role at The Forefront of Genomics. This will require careful consideration of what activities NHGRI should directly support, as opposed to activities NHGRI should strive to catalyze and influence through strategic partnerships and collaborations with other NIH Institutes/Centers and other funders.

Report of NHGRI Strategic Planning Workshop:

"From Genome to Phenotype: Genomic Variation Identification, Association, and Function in Human Health and Disease"

January 22-24, 2019

Table of Contents

Executive Summary....1

Main Report....3

Part I – Setting the Stage: Vision discussions to imagine what the field can be a decade from now....3

Part II – Scientific Issues: The current state-of-the-art to identify challenges and opportunities....4

Genomic Variant Discovery and Association....4

Genomic Variant and Genome Function....7

Focus Discussion 1: What can NHGRI do to facilitate bridging molecular and organismal phenotype?....9

Focus Discussion 2: Bridging Day 1 and Day 2 – Connecting discussions about variation, function and phenotype....10

Part III & IV – What should NHGRI do?: Integrating recommendations; placing in context and prioritizing. 10

Box 1: Highest Priority Topic Suggestions....12

Conclusions....12

Main Report

On January 22-24, 2019, the National Human Genome Research Institute (NHGRI) convened a workshop as part of its 'Genomics2020' strategic planning process. The goals of the workshop were for NHGRI to receive community input to: (1) help the Institute develop a strategic approach to find and characterize genomic variants as well as the genomic elements in which they reside or that they affect; (2) understand the effects of those variants on human health and disease; and (3) identify research projects, resources, and knowledge that NHGRI should pursue, as well as consideration of key ongoing efforts that should be continued or revised. The workshop was developed in a collaborative manner by an organizing committee comprised of members of the research community (led by Joseph Ecker, Eimear Kenny, Sharon Plon, Katherine Pollard, and Jay Shendure) and NHGRI staff, in concert with NHGRI leadership (see also Appendix 11: Acknowledgements). The workshop was videocast live and can be viewed on Genome TV, providing many details beyond the scope of this report.

Part I of the workshop sought input on where the field should be a decade from now. Part II (the main body of the workshop) sought input on challenges and opportunities in genomic variant discovery and association with traits; functional element discovery and characterization; and the interpretation of how genomic variants affect genome function. Part III sought input on integrating workshop recommendations, placing them in context, and identifying specific roles for NHGRI. The workshop concluded with Part IV, an initial sampling of community input on prioritization of the projects and resources that were recommended during this workshop (See Appendix 1: Agenda).

The workshop participants spanned a breadth of scientific expertise and community roles. Representation included those with expertise in technology development, genomic variation and population genetics, functional genomics, human diseases, clinical studies, computational biology and informatics, and model organisms. The participants were also selected across the range of career stages that represented a diversity of roles, including established independent investigators, early stage investigators, members of various consortia, NHGRI Advisory Council members, and NIH intramural scientists and extramural scientific staff (see also Appendix 2: Participant List).

Part I – Setting the Stage: Vision discussions to imagine what the field can be a decade from now

The goal of Part I of the workshop was to establish a vision for the state of this field in 10 years, thus setting the stage for the remainder of the workshop to consider how to achieve that vision. Eric Green, NHGRI Director, and NHGRI Program Directors (Elise Feingold and Adam Felsenfeld) welcomed participants, reviewed the workshop goals and organization, and invited community input on bold ideas on new projects, resources methods, and technologies.

This was followed by presentations outlining visions for genomics from Jay Shendure, Judy Cho, and Emma Farley. One vision was an emphasis on basic studies of the genome driven by innovation in technology and computational approaches. This innovation could lead to the generation of truly comprehensive catalogs of population-scale human genomic variation, a comprehensive understanding of rare genetic diseases, global molecular atlases for human and key model organisms, as well as models that enable accurate variant-effect prediction and a functional understanding of the genome. Another vision was improved utility of genomics in the clinic, driven by infrastructure supporting multidisciplinary teams. This infrastructure could lead to integration of genomics into healthcare in order to generate improved biomarkers for efficient screening, appropriate genetic testing aligned with patient choices, better outcomes through earlier and more appropriate interventions, and causal (rather than correlative) understanding of the relationship between genotype and phenotype. A third vision was a better understanding of the association between genotype and phenotype, driven by better approaches to manipulate genomes and measure phenotypes at the scale that genomes are currently being sequenced. Testing the impact of all genomic variants (alone, as combinations, or on different genetic backgrounds) on genome function and phenotype could lead to understanding the function of every human gene and regulatory element at the cellular, organoid, and organism level, and ultimately to the translation of this information to improved diagnostics and therapeutics.

Discussion was moderated by Heidi Rehm and Tuuli Lappalainen. They began by reminding participants that there is a continuum (not a dichotomy) from rare to common disease, as well as between protein-coding and regulatory functions, and suggested that to understand genetic associations, the full range of models (cellular, organoid, and animal model as well as large-scale human data) will need to be used. Both engineered perturbations and naturally-occurring genomic variation are informative. They suggested the community should strive to routinely diagnose about 80% of Mendelian cases using whole exome sequencing/whole genome sequencing (WES/WGS; workshop attendees indicated the current diagnostic rate is on the order of 25-30%). They challenged the community to try to understand 95% of all single nucleotide variants (SNV) in at least 100 clinically actionable genes (today, most variants are classified as Variants of Unknown Significance (VUS)). They envisioned a future with whole-genome sequences for about 0.1% of living humans across populations, addressing the current inadequate representation of populations and variant classes. They also highlighted the need to learn the best approaches to determine polygenic risk scores for common diseases for ethnically diverse populations, and how best to use these scores in the clinic to improve health care.

One discussion theme was improved understanding of genome structure to support functional studies. High-quality, more complete, haplotype-resolved genome sequences should be a technology-development goal and also become the standard for genomic studies. A corollary was a plea to stop ignoring repeats in sequence assemblies and analyses. Additional support for comparative genomics was also raised.

A second theme was a better understanding of genome function. Currently, we do not understand, or only have a misleading or incomplete understanding, of what most genes do. However, elucidation of function relies on understanding phenotype, which is much more complicated than genotype. There was enthusiasm for perturbational maps (such as high-throughput mutagenesis of genomic regions followed by phenotypic analysis); while some thought we are ready to measure some phenotypes at scale, others were concerned over the potentially vast number of experimental conditions that would need to be interrogated to cover dimensions such as cell fate, variant, and genetic background. There was also interest in profiling genomic variants in different genetic backgrounds and environments, perhaps using model organisms. One suggestion for bridging the gap between molecular and organismal phenotypes was to begin adopting a systems or network view of function. It was suggested that in the conflict between comprehensiveness and complexity, complexity will win because we don't have the technology to collect or analyze data at the scale required to be truly comprehensive; we should accept that and identify useful but achievable goals.

A third theme was improvements in informatics and data science. More data (systematically collected, using multiple modalities) are needed to develop better predictive models; current models are lacking. Better support and infrastructure would help address the need for testing and sharing software. Better computational tools, such as natural language processing, could be applied to the medical literature to learn phenotypic associations. It was also suggested that we need to learn to track return-on-investment from research efforts and to establish a baseline prior to new launching efforts, so as to learn which ideas are successful.

Part II – Scientific Issues: The current state-of-the-art to identify challenges and opportunities

The goal of Part II was to review the current state of the science and to use that information for identifying opportunities to overcome current challenges. This part of the workshop first considered genomic variant discovery and association, followed by consideration of genomic variant and genome function (each with a cycle of presentations, discussions, and breakouts). This part concluded with two Focus Discussions to illuminate specific topics that NHGRI anticipated would require additional attention during the workshop.

Genomic Variant Discovery and Association

Part II began with brief summaries (Adam Felsenfeld, Elise Feingold) of NHGRI and NIH projects illustrating the current approaches to "Variant to Function to Disease" (See also: Appendix 3: Related Project Summaries).

Examination of genomic variant discovery and association began with vignettes from different perspectives from Amit Khera, Laura Bierut, and Jonathan Haines. Clinically meaningful polygenic risk estimates appear to be a generalizable approach across diseases, with potential to teach us about disease biology and enable disease prediction. One current challenge to following up the success of genome-wide association studies GWAS is conducting functional studies at the same scale as GWAS with respect to the number of loci tested. Another challenge is that genetic and environmental effects can "overlap" (for example, a genetic susceptibility to smoking behavior can create an environment that causes lung cancer). Concerns across study approaches include limited access to biological samples and limited information on environmental exposure. Translating knowledge of disease risk alleles to therapeutic targets also remains a major challenge. Challenges for data interpretation include lack of common mapping strategies or ontologies, difficulty of working with data from different sources, and moving from loci to genes to function.

Discussion, moderated by Gonçalo Abecasis, focused on NHGRI's approach to these scientific issues by identifying the important questions to ask, then considering how to obtain the needed samples and data.

One question raised was how to make sure we study the entire genome? The difficulty of finding genomic variants in repetitive regions or identifying structural variation with short-read sequences was raised. There was enthusiasm for using long-read sequencing to address the problems, and it was noted that some of the most difficult genomic regions are also the most polymorphic. A related question was how to identify target genes that are affected by a particular variant? This question is important for understanding the impact of non-protein-coding variants.

Another question was what are the best emerging analysis tools for both variant association and imputation? In particular, there was interest in assessing the utility of polygenic risk scores and in deriving polygenic risk scores for additional diseases across populations if the scores are shown to be of high clinical value. A related point was the need for clearer guidelines on what is actionable; it was suggested that current polygenic risk scores appear to have relatively low specificity compared to cancer diagnostic tests. Better tools for haplotype analysis was also raised in this discussion. A related question was how will clinicians be trained to understand polygenic risk scores? Without that training, they will not be useful in the clinical setting.

When considering obtaining samples and data, workshop participants noted that there is a large number of existing and emerging studies that have enrolled participants whose samples and data can be used for disease and phenotype association studies. They also noted the importance of continuing to develop new methods, including those for family-based study designs. One concern was regulatory issues revolving around consents. There was also concern about genomics causing an increase in health disparities. Limitations in existing studies were raised. For example, the Million Veterans Project is racially diverse, yet nearly all male, while genomic and descriptive data for the *All of Us* Research Program are not yet available. It was also acknowledged that some of the concerns in using existing studies (e.g., Framingham Heart) pre-date genomics, thus it is necessary to acknowledge and be aware of such limitations where they exist. A related concern is that current FDA rules might bias experiments towards late-stage disease, where it may be too late to see benefit from therapies and lifestyle changes indicated by genomic diagnostics. Prospectively, NHGRI could influence policy discussions through continued research on ethical, legal, and social implications (ELSI).

The idea of a large-scale molecular phenotyping project was raised, which could collect metabolomic and transcriptomics data from tens of thousands of individuals. This was described as combining analysis of standing genomic variation in people with analysis of synthetic variation in animals followed by careful phenotyping of molecular events closely linked to the variants, such as regulation of nearby genes. There was also support for a project to catalog associations between genetic variants and gene expression across a range of cell types in a cohort of individuals, though it was also recognized that a focus on gene expression only tests one way a variant might matter. Aside from this, little was said during this session about employing modalities beyond DNA sequence (e.g. transcriptome, epigenome, nuclear architecture, and functional genomics).

Lon Cardon introduced the breakout discussion session to the entire group. See Appendix 4: Breakout Session Guidance, and Appendix 5: Breakout Session Assignments.

Suggestions from Breakout 1, "How much more sequencing, if any, is needed to study Mendelian and common disease, and what should NHGRI do in this area? Why?", led by Eric Boerwinkle and Nancy Cox, included novel technologies and software tools to increase the quality and extent of genome data, inclusion of more diverse participants in association studies with a focus on high-impact diseases and co-funding, bridging rare Mendelian and common polygenic diseases, and proximal multi-omics using genetic variation as well as varied contexts. They also had a suggestion for a "next phase" human genome-like project: Annotate all genome elements through complete sequencing, combined with analysis incorporating comparative genomics, molecular perturbations, as well as naturally occurring and engineered variation.

Suggestions from Breakout 2 "How and why to approach structural variation and other "hard to measure" variation?", led by Charles Lee and Karen Miga, included improved methods to manipulate high-molecular-weight DNA, new sequencing technologies beyond the current standards, improved reference genome approach, structural variant detection using telomere-to-telomere phased assemblies of diploid genomes, and methods to map epigenomics/transcriptomics data to the entire genome, including regions that are currently difficult to sequence.

Suggestions from Breakout 3 "How and why to approach more complex features- e.g., gene by environment, epistasis?", led by Andrew Clark and Eimear Kenny, included computational approaches for dimensional reduction, integrating familial, lifestyle, and environmental exposures into polygenic risk score models, investigating gene-gene interactions in the context of evolution, investigating the extent of conservation of gene regulatory networks across populations, and identifying suppressor variants using disease-free individuals who are homozygous for loss of function (LOF) mutations.

For a complete list of breakout suggestions see Appendix 6: List of All Project Suggestions from Breakout Sessions).

Breakout Sessions 1-3 Discussion:

One theme of this discussion was how much DNA sequencing is being done, and should be done, with and without NHGRI support; divergent views were expressed without a consensus view for prioritization. Some pointed out that companies and other initiatives are doing large-scale DNA sequencing, raising the suggestion that there was no need for NHGRI to pay for more. However, it was also pointed out that NHGRI-supported DNA sequencing was especially useful compared to that from other data generators, especially with respect to conditions for data sharing and diversity of participants. Some suggested NHGRI could have a bigger impact by transitioning to creating infrastructure and platforms for sharing (Electronic Health Records (EHR), clinic, and research data). Others were concerned about whether the data generated by these other sources would ultimately be shared for research.

Another theme was the need for better analysis tools and platforms, and the group expressed enthusiasm for a continued NHGRI role in this area. One idea was to create an analysis commons, including novel tool development, hardening, and sharing. Another idea was to facilitate comparisons of tools and algorithms. A concern was that transformative software and approaches are needed, which requires more creative ways of connecting programmers with data experts.

Participants noted value in NHGRI supporting the generation of reference genome sequences and research to establish a better understanding of structural variation, including variation in repetitive sequences. This could potentially be structured as an encyclopedia of repeat elements. NHGRI might also have a role in improving our understanding of patterns of mutational accumulation and the resulting bias on genome interpretation. It was noted that genomes are not stable over time in the soma, and this is not currently studied in a systematic manner other than the extreme case of somatic changes in cancer. Generation of high-quality genome sequence from samples that have already been extensively characterized in other NIH projects using functional genomics was also suggested.

The goal of achieving an understanding of genome function was equated with moving the field of genomics from identifying correlations towards an understanding of mechanism and causal relationships between variation and outcomes. It was suggested this mechanistic focus could perhaps lead to predictive genomics theories, akin to predictive theories in other areas of science.

The final theme emerging in this discussion focused on gene interactions and on the value of data generation beyond DNA sequence. Some noted they were surprised there was little mention of epigenomics and transcriptomics in the early parts of this workshop. Mendelian disorders were suggested to be a great place to start investigating gene-by-gene as well as gene-by-environment studies, and to potentially learn generalizable multi-omics approaches and generalizable rules about interactions. The importance of standards, standard processing, and quality metrics was highlighted when moving beyond DNA to RNA and epigenetics. Technology development to allow deployment of epigenomics at scale was suggested to be needed to realize its full promise. One suggestion for a large NHGRI project was to use perturbations (including naturally occurring genetic variation) with an overlay of deep molecular phenotyping (cell, model organism, and human) to completely annotate the function of all genes as well as non-genic sequences. It was suggested that NHGRI also develop analysis paradigms that other NIH Institutes/Centers or funding agencies could generalize to their systems of interest.

Genomic Variant and Genome Function

Examination of genomic variant and genome function began with vignettes from different perspectives from Hugo Bellen, Lea Starita, and Tim Reddy, all of whom recognized the current bottleneck is interpretation of the functional role of variants. Challenges to using model organism include the fact that many genes and variants remain to be characterized, phenotypic complexity is not fully assessed in simple, high-throughput assays, and the model organism culture is different from the physician culture. Successes showed great examples of progress being made to understand and classify VUS in disease genes and how model organisms help identify new disease genes. Clinicians are often convinced of the utility of the model organism data once they use it. Multiplex high-throughput protein-coding variant analysis is emerging as a promising approach to improve genomic variant interpretation, although it is subject to the availability of suitable assays for function and whether these functions are predictive of disease. This approach is also improved by the presence of a sufficiently large training set of variants known to be pathogenic or benign. High-throughput non-coding variant analysis provides opportunities for improved diagnostics, disease risk scores, and better identification of therapeutic targets. Better technology is needed to improve scale, targeting, and study design as well as to connect regulatory elements to genes.

Discussion of genomic variant and genome function presentations, led by Brenton Graveley, started from the idea that the greatest need in genomics is to understand how the genome functions.

The first theme was how studies of genome functions should be done in the future. Model organisms are a great opportunity for this research. The concern was raised that sometimes peer review underestimates the value of model organism systems, and scientists working on humans forget over time what was first learned in model systems. There was agreement that a community perception changes were required for collaborations between physicians and basic scientists (especially for model organism studies); sometimes this simply requires direct exposure, allowing physicians to see the usefulness of model organism data, and for basic scientists to better understand the use of model organisms to address disease mechanisms and clinical applications. Continued NHGRI support for model organism databases was also identified as important.

The second theme centered around what resources, new technologies, and computational capabilities are needed to generate and make full use of functional data. Participants raised the value of NHGRI supporting technology development. One specific suggestion was cell recorders that could report what genes were expressed and/or what elements were active. There was also advocacy to make longer synthetic DNA more accurately and with lower cost. Technology for genome editing, epigenome editing, and manipulating RNA isoforms was raised. Development and application of appropriate high-throughput phenotyping approaches would be critical for these assays. Given the large number of possible experimental conditions arising from considering several dimensions such as cell type, genetic background, and environmental condition it would be important to think carefully about the most informative phenotypic assays; perhaps a focus on molecular phenotypes would be the way to start. Multimodal measurements were highlighted because they facilitate data integration. Continued data collection to learn the sequences that regulators interact with was requested. There was also interest in measuring protein abundance, as cells might buffer against fluctuation in RNA amounts, and also interest in variation that changes substrates for kinases, proteases, etc. New technology is needed to allow application of these assays to repeat sequences in order to include this often-ignored portion of genome. It was also proposed we require a much better understanding of gene function, moving towards a network view, and incorporating the idea that many genes have more than one function.

Two new resource needs for standardized, shared data were identified. First, a resource to allow others to benefit from the use of data generated from reporter, genome editing, and epigenome editing assays. Second, a resource to enable understanding of transposable elements.

The final theme was consideration of barriers to performing functional studies at scale. Consideration of how elements work together was identified as important, yet the huge space of combinatorics is daunting. It was suggested that we should be comprehensive when practical, but to remember that the actual goal is to achieve a comprehensive understanding, which may not require testing every possible condition. One option is for NHGRI to model approaches in the context of particular biological problems to increase the uptake by others for their systems. There was support for continued work to learn transcription factor-DNA and RNA binding protein-RNA interaction motifs and how they are affected by genomic variation and cellular context. It was also noted that association studies often collect only blood for DNA extraction; if they could also collect clinical biopsies that could support other work.

The breakout session reports were introduced by Joseph Ecker. See Appendix 4: Breakout Session Guidance, and Appendix 5: Breakout Session Assignments.

Suggestions from Breakout 4 "Identification and characterization of all genes and regulatory elements", led by Bing Ren and Ross Hardison, included near-term projects such as an atlas of cis-regulatory regions, assessment of function in the genome and epigenome (human and model organisms), reference epigenomes from 1000 cell types, and technology development. Longer-term ideas included scalable methods for phenotyping, scalable methods to perturb genomes, end-to-end complete human genome sequences, full genome sequences for a million species, and full epigenomes in single cells.

Suggestions from Breakout 5 "Determining the functional consequences of variants acting individually and in combination", led by Kelly Frazer and Wendy Chung, included a single-cell eQTL study of people, predictive modeling of genome function based on systematic data sets, methods/resources to aggregate and integrate human data and link to model organism data. Technology goals included better DNA synthesis, better nucleic acid sequencing, and multi-omics assays on the same biosamples/single cell.

Suggestions from Breakout 6 "Accurate prediction of the regulatory consequences of variants, and modeling gene regulation", led by Trey Ideker and Christina Leslie, included modeling ideas, such as predicting the trajectory of genome function over time, a map of mechanisms that connect genomic variation to a complex disease, methodologies to identify principles of gene regulation causal variants for disease, and approaches to model cell-cell interactions. Data and technology suggestions included sufficiently abundant and robust data to train causal models, systematic measurement of locations and interactions for proteins of common diseases, and modeling to identify most informative data types for further investment. Trade-offs among simple, interpretable models versus complex, highly accurate predictive models were discussed, with the group prioritizing some level of interpretability.

For a complete list of breakout suggestions see Appendix 6: List of All Project Suggestions from Breakout Sessions).

Breakout Session 4-6 Discussion began with the high-level observation that science has changed such that phenotyping now lags behind DNA sequencing, and the NIH is perhaps better prepared for the old landscape; science needs phenotyping centers today. Another observation was the tension between whether it is better to study genomic variants first or function first; perhaps we should be open to both approaches, depending on the question and assays. It was also suggested that pleiotropy is a source of important information and could be appropriate for NHGRI given that it is not disease-specific. More technology development is needed, given that many of these functional assays and phenotyping methods could be improved.

Characterizing regulatory element function was one theme. One idea was for a pilot project ("FUNCODE") to study 1% of regulatory elements modeled on the pilot phase of ENCODE. Such a pilot could combine -omics and imaging phenotyping, learn how to improve scale, pilot new assays, and begin to understand combinatorics and long-range regulation. Another idea was to elucidate which approaches are most informative for understanding biology by characterizing a set of genomic variants using several systems and asking which systems find the expected effects. An alternative approach could attempt to learn how deep an understanding about regulation is needed to understand a biological event (such as differentiation), which could help to develop generalizable approaches. The idea of investigating the extremes of gene expression and genomic variants relative to health and disease was also raised to learn about the boundaries of typical biology.

Characterizing gene function was another theme. It was suggested that systematic data collection was needed for a more accurate understanding. One idea was to first find a biochemical, cellular, or organismal phenotype for 90% of genes, then identify medically relevant variants in those genes, and then use multi-omic assays to understand the genome to phenotype correlation. Another idea was identifying and phenotyping individuals with homozygous null alleles, which will provide information for a subset of genes.

One theme was an emphasis on supporting research in more physiological systems. This could be primary cells instead of cell lines or culture models for building tissues and organs. Readily accessible samples from humans were also raised. Finally, the importance of cell-cell interactions was raised, which could be important in studying cells in physiological states as well as understanding how the interactions program cell states.

Another theme was the need for better statistical approaches. For example, normalizing -omics and EHR data at scale is sufficiently complex that better algorithms are needed. Methods to correct for batch effects and reference sets of samples for modeling batch effects were raised. Finally, a better understanding of the sources of error and how to track it, as well as the need for involving statisticians in experimental design prior to data collection, was recognized. There are also substantial concerns around consent and platform issues when working with large-scale private data in the cloud.

A final theme was the need for intuitive, efficient data visualization. It is no longer practical to open all data in a browser in this era of thousands of samples and hundreds of data types. Dimensional reduction of data was suggested as one approach to support deriving inferences from data. Web-browser improvements were also suggested. Visualization of single-cell data is another unsolved problem. Finally, the idea of visualization tools that could suggest what regions/samples were most interesting was raised.

Focus Discussion 1: What can NHGRI do to facilitate bridging molecular and organismal phenotype?

Focus Discussion 1 “What can NHGRI do to facilitate bridging molecular and organismal phenotype?” asked participants to address and critique two approaches: relating genome variation to disease status or other observed human phenotypes, often without including data collection or analysis of molecular phenotypes (expression, proteomic, etc.); and studying molecular phenotypes, often without confirming that perturbations of those phenotypes actually impact health or disease at the organismal level. (Moderators: Len Pennacchio, Barbara Wold, Andrea Califano). The session began with an introductory presentation and was followed by group discussion (see Appendix 9: Project Suggestions From Focus Sessions).

The proposed goal was to understand how to progress from “only” cataloging and making genotype-phenotype associations (either molecular or organismal) towards a more mechanistic interpretation and understanding. One step on this path could be to increase attention on expanded phenotyping applied at the same throughput as DNA sequencing studies; resources could be made available by decreasing attention to catalog completeness for the sake of completeness. Another step would be to move beyond the artificial limitation of genomic-only (i.e., DNA and RNA sequencing) assays and include proteomic, metabolomic, and high-content imaging measurements. For example, extending ENCODE-like activities to include post-transcriptional and post-translational signal transduction would be a valuable addition. Finally, learning to identify and model the cellular logic supporting conversion of genomic variants into phenotypes could be a transformative and generalizable approach that could be adopted by others.

One anchor project could be to study the function of all genes (though there were divergent views among participants on whether to include non-coding genes), all transcribed regions, or all regulatory regions. This could be accomplished through genetic perturbations and analysis of standing genomic variation. With these technologies at scale, it might be possible to study enhancers and learn which of them are redundant. Tracking the effects of variation at the single-cell level may be important. It was noted that a project to consider natural genomic variation in humans might need tens of thousands of individuals; one limitation is that by averaging over many individuals we might miss gene-by-gene interactions, as with GWAS today. It was raised that the phenotype associated with null alleles often captures a subset of gene function, while missing other interesting gene functions. Pleiotropy was raised as another phenotyping challenge. One suggestion was that getting as close as possible to the relevant context (e.g., cell fate or cell state) would perhaps help in determining the appropriate phenotype(s) to consider.

NHGRI has been at the forefront of developing standard ontologies that would facilitate this work, and continued support is required. NHGRI could help to identify the appropriate role of model organisms in learning to understand complex phenotypes. NHGRI could also undertake a project to develop quantitative metrics to characterize the success of various model systems (e.g., model organisms, primary cells, organoids, people, and iPSC derivatives).

Focus Discussion 2: Bridging Day 1 and Day 2 – Connecting discussions about variation, function and phenotype

Focus Discussion 2: "Bridging Day 1 and Day 2: Connecting discussions about variation, function and phenotype" considered the growing need for potentially complementary NHGRI (and related NIH) programs to avoid becoming "siloed" in order achieve scientific goals more efficiently. The session (moderated by Katie Pollard, Dani Fallin, and Rick Myers) began with an introductory presentation of a long, yet partial, list of active and completed projects from NHGRI as well as the NIH Common Fund and other NIH Institutes/Centers that examined genomic variation, function, and phenotypes. Examples of successful connections were noted among programs focused on gene discovery and characterization in humans, such as collaborations among both the NHGRI Centers for Mendelian Genomics (CMG) and the NIH Common Fund Undiagnosed Disease Network (UDN) with model organism programs such as the NIH Common Fund Knockout Mouse Phenotyping (KOMP2) program (see also Appendix 9: Project Suggestions From Focus Sessions).

During the discussion, it was noted that interoperability can break down silos; interoperable projects require planning and coordination of standards and metadata, which increase the value of the outputs. Distributed data that are not interoperable hinder use for endeavors such as training machine-learning predictors. It was suggested that NHGRI could support meetings to bring people from projects with different scientific focus into the same room to coordinate. Suggestions were made for increasing interoperability with other areas as well. For example, studies of gene-by-environment interactions could be enhanced by increased collaborations with NIEHS or by increased sequencing DNA from participants in existing cohort studies with rich environmental exposure data (e.g., health professionals' study and nurses' study). Similarly, proposed studies using EHR data would benefit from improved partnerships with clinicians.

Part III & IV – What should NHGRI do?: Integrating recommendations; placing in context and prioritizing.

The goal of Part III was to seek input on what NHGRI can do at The Forefront of Genomics to address the major recommendations that arose earlier in the workshop. Each topic session began with brief presentations of project ideas inspired by the workshop or ideas submitted prior to the workshop (see also Appendix 10: Bold Ideas), followed by group discussion. Part III is not presented here in detail because the Part III lists were superseded by the Part IV lists, which are discussed below. (For a complete list of ideas from Part III see also Appendix 7: Ideas from Topic Sessions).

Topic 1 "Discovery and Interpretation of Variation Associated with Human Health and Disease" began with presentations of project ideas from Michael Boehnke, Barbara Stranger, Anshul Kundaje, and David Valle, followed by group discussion. Topic 2: "Addressing Basic Research Questions that Anticipate Clinical Needs" began with presentations of project ideas from Sharon Plon, Stephen Chanock, Howard Chang, and Howard Jacob, followed by group discussion. Topic 3: "Predicting and Characterizing Functional Consequences of Genome Variation, Including Beyond Single Variant/Gene" began with presentations of project ideas from Jonathan Pritchard, Neville Sanjana, Nadav Ahituv, and Dana Crawford, followed by group discussion. Topic 4: "Data Resources, Methods, Technologies and Computational Capabilities" began with presentations of project ideas from Aviv Regev, Christina Leslie, Jason Buenrostro, and Marylyn Ritchie, followed by group discussion.

The goal of Part IV was to seek initial input on prioritization of the projects and resources that were recommended during Part III. Participants were asked to consider these factors when prioritizing: What is most valuable to advance genomics? Is NHGRI uniquely positioned to take this on? What are long-term ideas and what are ideas that could be started in the next year?

Feedback was collected through electronic polling of participants in the room, followed by a Moderated discussion.

Strengths of this approach included everyone could provide input (rather than a sampling of a comments by a subset of participants) and discussion provided an opportunity to consider whether participants agreed with the voting results and to elucidate why projects were ranked in the observed order. Caveats included the lists were unlikely to be optimal due to rapid lumping and splitting (which led to some ideas being vague), participants had a short time to consider and vote, voting was done only within topic areas rather than for the entire workshop, and the appropriate balance of investigator versus consortium projects was not considered. Recognizing those limitations, the group did consider the results for in each topic area. Ideas receiving the strongest support for each topic are described in Box 1: Highest Priority Topic Suggestions. For complete lists of suggestions considered for each topic see Appendix 8: Ideas From Synthesis And Prioritization Session.

In discussion of the polling results, the group agreed that integrating functional genomic approaches with direct assessment of the human population was a key area to consider. The group found it notable and telling that offering training opportunities polled high even though the discussion time it received at the workshop was small. Finally, the only topic that focused on high-throughput genome sequencing for studying common diseases ended up at the bottom of the Topic 1 list. Similarly sequencing to understand somatic variation (Topic 3 list) also scored poorly. There was the sense that projects supported by other NIH Institutes and Centers might continue to include a lot of disease-specific sequencing. However, participants noted that functional genomics would need to include genome sequencing, as do a number of other the suggestions considered at this workshop.

The group thought the poll results did not reflect their enthusiasm with respect to “writing” genomes; this topic was embedded in the technology development topic and perhaps should have been separate idea. Some topics that were discussed earlier in the workshop appeared not to be included in the polling topics: for example, there was little mention of human participant diversity, evolution, and population genetics. Similarly, standards and quantification of error appeared to be absent as separate topics.

Box 1: Highest Priority Topic Suggestions

Topic 1 "Discovery and Interpretation of Variation Associated with Human Health and Disease"

- Experimental methods development
- Functional genomics at scale: cellular and model organisms
- Statistical and computational methods development
- Interoperability and usability of data resources
- Perturbational and combinatorial functional genomics and microscopy at scale

Topic 2 "Addressing Basic Research Questions that Anticipate Clinical Needs"

- Deep perturbation profiling, proximal molecular readout of variants at every base
- Single cell GTEx-like project
- Improved tools for data visualization
- Single cell multi-omic technology (wet lab and dry lab)
- Increase the computational biology workforce
- Approaches to detect larger structural variants in the clinical setting

Topic 3 "Predicting and Characterizing Functional Consequences of Genome Variation, Including Beyond Single Variant/Gene"

- Complete atlas of genes and their function
- Complete atlas of elements under selection or cis-regulatory elements and variants across cell types and conditions; computational models for interpretation and visualization
- Training opportunities for engaging undergraduate students with an emphasis on underrepresented groups in STEM
- High-throughput measurement of genomic variants in organisms and organoids

Topic 4 "Data Resources, Methods, Technologies and Computational Capabilities"

- Complete telomere-to-telomere sequences of human genomes
- Large-scale perturbations in pools of cells, followed by high-content, high-resolution data collection
- Interacting data platforms, variant portals, and visualization platforms for large-scale genome processing, analysis, integration, and exploration via cloud computing
- Transferable gene regulatory or multi-scale models from multi-layered data across biological systems

Conclusions

Participants were in strong agreement that NHGRI should continue to support research on identifying genomic variants, associating them with phenotypes of interest, and characterizing their influence on genome function. As noted earlier, one conclusion of this workshop is the current bottleneck in the field is characterizing and understanding the function of variants; no longer are finding variants and associating them with phenotypes the slow steps. Phenotyping and functional characterization now lag behind sequencing, and science needs progress on this front; yet today the NIH is perhaps better prepared for the old landscape and more linear approach.

Overall the workshop highlighted 3 scientific areas as well as a cross-cutting approach that appeared to merit prioritization. Many of the highest-priority topic suggestions from Box 1, and other parts of the workshop discussion, can be readily placed in this framework.

Technology Development and Informatics:

- Improved DNA/RNA sequencing, including longer reads and detection of modified bases
- Improved high-throughput functional genomics methods, including gene editing, epigenome editing, base editing, and reporter assays; single cell approaches; spatial resolution
- Improved multi-omic approaches, including single cell (wet lab and dry lab)
- Complete telomere-to-telomere sequences of human genomes
- Transferable gene regulatory or multi-scale models from multi-layered data across biological systems
- Interacting data platforms, variant portals, and visualization platforms for large-scale genome processing, analysis, integration, and exploration via cloud computing
- Enhanced interoperability and usability of data resources
- Increase the computational biology workforce

Functional genomics:

- Deep perturbation profiling, proximal molecular readout of variant function at every base
- High-throughput functional genomics using cellular, organoid, and organismal models
- High-throughput perturbational and combinatorial functional genomics and microscopy
- Complete atlas of genes and their function
- Complete atlas of elements under selection, cis-regulatory elements and variants across cell types and conditions
- Computational models for interpretation and visualization of element, variant, and gene function
- Large-scale perturbations in pools of cells, followed by high-content, high-resolution data collection

Disease variant associations:

- Single cell GTEx-like project
- Approaches to detect larger structural variants in the clinical setting
- New methods to analyze association studies, including analyses of association data produced by others
- Characterization of genomic architecture, including both gene-by-gene and gene-by-environment interactions

Model organisms (cross-cutting approach):

- Experimental systems to test observations from human studies; for example, characterizing the function of candidate variants associated with disease
- Experimental systems to learn the general principles and develop methods for application to human studies; examples include gene-by-gene interactions, gene-by-environment interactions, and connecting molecular phenotypes to organismal phenotypes

Overall, this workshop highlighted that this is a critical and exciting time for genome sciences. Advances in technology, genome sequencing, functional genomics, and computational genomics have created new opportunities for approaches to the characterization of genes, genomic elements, and genomic variants in relationship to human biology, health, and disease. This workshop was a valuable source of community input to NHGRI, and the Institute is grateful to all who participated. As one part of its ‘Genomics2020’ strategic planning process, NHGRI has heard from external scientists working in a variety of areas and across a range of career stages and institutions. The breadth and depth of topics covered solidify the need for NHGRI to define its role at The Forefront of Genomics. This will require careful consideration of what activities NHGRI should directly support, versus areas NHGRI should catalyze and influence through strategic partnerships and collaborations with other NIH Institutes/Centers and other funders.