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

Appendix 1: Agenda

Hilton Washington DC/Rockville Hotel & Executive Meeting Center, Rockville, MD 20852

Meeting will be recorded and videocast: genome.gov/genometvlive/

Meeting materials: event.capconcorp.com/wp/2019-nhgri/

Workshop Goals:

Participants will help NHGRI develop a strategic approach for the next 5-10 years to significantly advance the state-of-the-art in our ability to find and characterize genomic variants, the genomic elements in which they reside or that they affect and understand the effects of those variants on human health and disease.

We seek recommendations for projects, community resources, knowledge, and research directions that NHGRI should pursue, as well as consideration of key ongoing efforts that should be continued or revised.

We also seek recommendations for mid-scale projects that could be initiated in the next year (for example, ideas that build directly on the existing ENCODE Project and the Genome Sequencing Programs) as well as longer-term projects that may require pilot projects, development of technologies or resources and establishment of collaborations, before they could begin.

Workshop Rationale:

Relating genomic variation to human phenotype is a central issue in genomics.

There are several ways to consider this problem, but the elements are: identifying genomic variants; associating them with regulatory elements, genes and phenotypes; and understanding gene, regulatory element and variant function. Within this, or added to it, would be an understanding of epistasis and environmental influences on genotype.

There are now many individual examples (genes/phenotypes) where we have a reasonably detailed understanding of the relationship between variant and disease– these have defined the paradigm. But these so far mostly represent simpler cases: Mendelian (i.e., very strong effect) traits; diseases where the physiology is straightforward (e.g., inborn errors of metabolism); and some cancers.

But the aim of human genomics is to solve this problem at scale for all human traits that have an inherited component, and in so doing, gain biological insight into the nature of inherited disease, insight into mechanisms of variant, regulatory element and gene function, and ultimately to provide a rational foundation for clinical applications. To develop approaches to do this, NHGRI seeks to understand the state-of-the-art, gaps in the field (knowledge, methods, data and resources), find better ways to integrate the information from the separate elements, and identify promising new approaches to address the general problem.

This workshop is one of a series of activities devoted to strategic planning for NHGRI https://www.genome.gov/27570607/strategic-planning-overview/. The recommendations forthcoming from the workshop discussions will inform the NHGRI "2020 Vision for Genomics" Strategic Planning efforts.

Meeting Agenda:

The workshop is divided into four parts:

I. Setting the Stage: "Vision" discussions to imagine what the field can (or must) be a decade from now.

II. Scientific Issues: The current state-of-the-art to identify challenges and opportunities:
- Day 1: variant discovery and association with traits
- Day 2: functional element discovery and characterization, and the interpretation of how variants effect function
- Breakout sessions will extend the discussions, raising specific examples that NHGRI could pursue (e.g., specific projects, methods, knowledge, resources, data, etc.).

III. What should NHGRI do?: Integration of the recommendations from the preceding days; placing them in a wider context; identifying specific things NHGRI should do.

IV. Synthesis and Prioritization: Prioritization among the projects, directions, resources, etc. recommended during the workshop.

Tuesday, January 22, 2019

Location for all sessions excluding Breakout sessions will be in the Plaza Ballroom.

Please note that this meeting will be **live-streamed** and permanently archived.

8:00 – 8:30 a.m.	**Registration**	
8:30 – 8:40 a.m.	**Welcome and Introduction**	Eric Green
8:40 – 9:00 a.m.	**Statement of Meeting Goals**	Elise Feingold Adam Felsenfeld

Part I: Setting the Stage

	Visions of the Future *What will the field of genomics look like in 5-10 years? How will it get there?*	
9:00 – 9:15 a.m.	**Vision Talk 1**	Jay Shendure
9:15 – 9:30 a.m.	**Vision Talk 2**	Judy Cho

9:30 – 9:45 a.m.	**Vision Talk 3**	Emma Farley
9:45 – 10:45 a.m.	**Discussion**	**Moderators**: John Stamatoyannopoulos Heidi Rehm Tuuli Lapplainen
10:45 – 11:00 a.m.	**Break**	

Part II: Scientific Issues

11:00 – 11:05 a.m.	**Deliverables for Part II**	NHGRI Staff Meeting Advisors
11:05 – 11:25 a.m.	**NHGRI's Current Approach to "Variant to Function to Disease"**	NHGRI Staff
11:25 – 12:10 p.m.	**Current State of the Art in Variant Discovery and Association**	Amit Khera Laura Bierut Jonathan Haines
12:10 – 12:55 p.m.	**Moderated Discussion** *How should NHGRI approach these scientific issues in the future? What are the important questions to ask? Where and how will NHGRI get samples/data?*	**Moderator**: Gonçalo Abecasis
12:55 – 2:00 p.m.	**Lunch**	
2:00 – 2:15 p.m.	**Breakout Session Charge**	NHGRI Staff Meeting Advisors
2:15 – 3:15 p.m.	**Day 1 Breakout Sessions (concurrent)**	

	Breakout 1 (*Roosevelt Room*) *How much more sequencing, if any, is needed to study Mendelian and common disease, and what should NHGRI do in this area? Why?*	**Co-Chairs**: Eric Boerwinkle Nancy Cox
	Breakout 2 (*Madison Room*) *How and why to approach structural variation and other "hard to measure" variation?*	**Co-Chairs**: Charles Lee Karen Miga
	Breakout 3 (*Jefferson Room*) *How and why to approach more complex features- e.g., GxE, epistasis?*	**Co-Chairs**: Andrew Clark Eimear Kenny
3:15 – 3:45 p.m.	**Break**	
3:45 – 5:15 p.m.	**Breakout Summaries and Discussion**	Breakout Session Co-Chairs **Moderator**: Lon Cardon
5:15 – 5:30 p.m.	**Summation of Day 1 and Prep for Day 2**	NHGRI Staff
5:30 p.m.	**Group Photo**	

Wednesday, January 23, 2019

Location for all sessions excluding Breakout sessions will be in the Plaza Ballroom.

8:00 – 8:15 a.m.	**Introduction: Variant and Genome Function**	NHGRI Staff
8:15 – 9:00 a.m.	**Current State of the Art: Functional Data, Analysis and Interpretation**	Hugo Bellen Lea Starita Timothy Reddy
9:00 – 10:00 a.m.	**Moderated Discussion**	**Moderator**: Brenton Graveley

	How should we be doing this into the future? What resources, new technologies and computational capabilities do we need to generate and make use of functional data? What is needed to overcome barriers to performing functional studies at scale?	
10:00 – 10:15 a.m.	**Break**	
10:15 – 10:30 a.m.	**Breakout Session Charge**	NHGRI Staff Meeting Advisors
10:30 – 11:30 a.m.	**Day 2 Breakout Sessions (concurrent)**	
	Breakout 4 (*Roosevelt Room*)	**Co-Chairs**:
	Identification and characterization of all genes and regulatory elements	Ross Hardison Bing Ren
	Breakout 5 (*Madison Room*)	**Co-Chairs**:
	Determining the functional consequences of variants acting individually and in combination	Kelly Frazer Wendy Chung
	Breakout 6 (*Jefferson Room*)	**Co-Chairs**:
	Accurate prediction of the regulatory consequences of variants, and modeling gene regulation	Trey Ideker Christina Leslie
11:30 – 12:30 p.m.	**Lunch**	
12:30 – 2:00 p.m.	**Breakout Reports and Discussion** *10 minutes each plus one hour discussion*	Breakout Session Co-Chairs **Moderator**: Joseph Ecker
2:00 – 2:15 p.m.	**Break**	

2:15 – 3:00 p.m.	**Focus Discussion 1** *What can NHGRI do to facilitate bridging molecular and organismal phenotype?*	**Moderators**: Len Pennacchio Barbara Wold Andrea Califano
3:00 – 3:45 p.m.	**Focus Discussion 2** *Bridging Day 1 and Day 2: Connecting discussions about variation, function and phenotype*	**Moderators**: Richard Myers Katherine Pollard Daniele Fallin

Part III: What should NHGRI do?

3:45 – 4:00 p.m.	**Brief recap of Part III deliverables: Brainstorming for NHGRI-supported Activities** **How to Achieve the Science – Four Topics (below)** *What can NHGRI do to address the major recommendations? What insights, capabilities, policies, initiatives, collaborations, alliances, etc. should we pursue?*	NHGRI Staff Meeting Advisors
4:00 – 4:45 p.m.	**Topic 1: Discovery and Interpretation of Variation Associated with Human Health and Disease**	**Panel**: Barbara Stranger Anshul Kundaje David Valle **Moderator**: Michael Boehnke
4:45 – 5:30 p.m.	**Topic 3: Predicting and Characterizing Functional Consequences of Genome**	**Panel**: Nadav Ahituv Dana Crawford

	Variation, Including Beyond Single Variant/Gene	Neville Sanjana **Moderator**: Jonathan Pritchard

Thursday, January 24, 2019

Location for all sessions excluding Breakout sessions will be in the Plaza Ballroom.

8:00 – 8:15 a.m.	**Day 3: Introduction**	NHGRI Staff
8:15 – 9:00 a.m.	**Topic 4: Data Resources, Methods, Technologies and Computational Capabilities**	**Panel**: Christina Leslie Marylyn Ritchie Jason Buenrostro **Moderator**: Aviv Regev
9:00 – 9:45 a.m.	**Topic 2: Addressing Basic Research Questions that Anticipate Clinical Needs**	**Panel**: Howard Chang Stephen Chanock Howard Jacob **Moderator**: Sharon Plon
9:45 – 10:15 a.m.	**Break**	

Part IV: Synthesis and Prioritization

10:15 – 12:15 p.m.	**Recommendations for NHGRI Priorities** *Moderated discussion on future NHGRI priorities for specific*	**Meeting Co-chairs &** **Meeting Advisors**: Jay Shendure Joseph Ecker

	initiatives, data sets, knowledge, capabilities	Sharon Plon Katherine Pollard
12:15 – 1:00 p.m.	**Summation of the Meeting**	NHGRI Staff
1:00 p.m.	**Main Meeting Adjourns**	

Appendix 2: Participant List

Goncalo Abecasis
University of Michigan
goncalo@umich.edu

Adam R. Abate
University of California, San Francisco
arabate@gmail.com

Anjene Addington
National Institute of Mental Health, NIH
anjene.addington@nih.gov

Andrew Adey
Oregon Health & Sciences University
adey@ohsu.edu

Nadav Ahituv
University of California, San Francisco
nadav.ahituv@ucsf.edu

Erez Aiden
Baylor College of Medicine/Rice University
erez@erez.com

Orli G. Bahcall
Nature
o.bahcall@us.nature.com

Gill Bejerano
Stanford University
bejerano@stanford.edu

Hugo J. Bellen
Howard Hughes Medical Institute/Baylor College of Medicine
hbellen@bcm.edu

Laura J. Bierut
Washington University School of Medicine
laura@wustl.edu

Mike Boehnke
University of Michigan
boehnke@umich.edu

Jef D. Boeke
Institute for Systems Genetics, NYU
jef.boeke@nyulangone.org

Eric Boerwinkle
UTHealth School of Public Health
eric.boerwinkle@uth.tmc.edu

Vence L. Bonham
National Human Genome Research Institute, NIH
bonhamv@mail.nih.gov

Laurie A. Boyer
Massachusetts Institute of Technology
lboyer@mit.edu

Lawrence Brody
National Human Genome Research Institute, NIH lbrody@mail.nih.gov

Jason Buenrostro
Harvard University
jason_buenrostro@harvard.edu

Carol J. Bult
The Jackson Laboratory
carol.bult@jax.org

Shawn Burgess
National Human Genome Research Institute, NIH
burgess@mail.nih.gov

Eileen W. Cahill
National Human Genome Research Institute, NIH
eileen.cahill@nih.gov

Andrea Califano
Columbia University
ac2248@cumc.columbia.edu

Lon R. Cardon
BioMarin Pharmaceutical Inc.
lon.cardon@bmrn.com

Gemma Carvill
Northwestern University
gemma.carvill@northwestern.edu

Lisa Chadwick
National Human Genome Research Institute, NIH
lisa.chadwick@nih.gov

Aravinda Chakravarti
NYU Langone Health
aravinda.chakravarti@nyulangone.org

Howard Chang
Stanford University School of Medicine
howchang@stanford.edu

Stephen J. Chanock
National Cancer Institute, NIH
chanocks@mail.nih.gov

Judy H. Cho
Icahn School of Medicine at Mount Sinai
judy.cho@mssm.edu

Clement Y. Chow
University of Utah
cchow@genetics.utah.edu

Wendy K. Chung
Columbia University
wkc15@cumc.columbia.edu

Andrew Clark
Cornell University
ac347@cornell.edu

Barak Cohen
Washington University
cohen@wustl.edu

Jonah Cool
Chan Zuckerberg Initiative
jcool@chanzuckerberg.com

Nancy J. Cox
Vanderbilt University Medical Center
nancy.j.cox@vumc.org

Mark Craven
University of Wisconsin
craven@biostat.wisc.edu

Dana Colleen Crawford
Case Western Reserve University
dana.crawford@case.edu

Brandi Davis-Dusenbery
Seven Bridges
brandi@sbgenomics.com

Megan Y. Dennis
University of California, Davis
mydennis@ucdavis.edu

Valentina di Francesco
National Human Genome Research Institute, NIH
vdifrancesco@mail.nih.gov

Joseph Robert Ecker
The Salk Institute for Biological Studies
ecker@salk.edu

Evan E. Eichler
University of Washington/Howard Hughes Medical Institute
eee@gs.washington.edu

Alvaro Encinas
National Human Genome Research Institute, NIH
alvaro.encinas@nih.gov

Dani Fallin
Johns Hopkins University
dfallin@jhu.edu

Emma Kirsten Farley
University of California, San Diego
efarley@ucsd.edu

Elise Feingold
National Human Genome Research Institute, NIH
feingole@mail.nih.gov

Adam Felsenfeld
National Human Genome Research Institute, NIH
adam_felsenfeld@nih.gov

Colin F. Fletcher
National Human Genome Research Institute, NIH
colin.fletcher@nih.gov

Kelly A. Frazer
University of California, San Diego
kafrazer@ucsd.edu

Molly J. Gasperini
University of Washington
gasperim@uw.edu

Tina Gatlin
National Human Genome Research Institute, NIH
gatlincl@mail.nih.gov

Kelly Gebo
All of Us Research Program, NIH
kelly.gebo@nih.gov

Mark B. Gerstein
Yale University
piaa@gersteinlab.org

Richard A. Gibbs
Baylor College of Medicine
agibbs@bcm.edu

Daniel Gilchrist
National Human Genome Research Institute, NIH
daniel.gilchrist@nih.gov

Kerry E. Goetz
National Eye Institute, NIH
goetzke@nei.nih.gov

Brenton R. Graveley
UConn Health
graveley@uchc.edu

Eric D. Green
National Human Genome Research Institute, NIH
ben.ryan@nih.gov

Jonathan Lee Haines
Case Western Reserve University, School of Medicine
ess13@case.edu

Ira Hall
Washington University
ihall@wustl.edu

Ross C. Hardison
Penn State University
rch8@psu.edu

Chuan He
The University of Chicago/Howard Hughes Medical Institute
chuanhe@uchicago.edu

Stephanie Hicks
Johns Hopkins Bloomberg School of Public Health
shicks19@jhu.edu

Martin Hirst
University of British Columbia
mhirst@bcgsc.ca

Carolyn M. Hutter
National Human Genome Research Institute, NIH
carolyn.hutter@nih.gov

Trey Ideker
University of California, San Diego
tideker@ucsd.edu

Nilah Ioannidis
Stanford University
nilah@stanford.edu

Howard J. Jacob
Abbvie
howard.jacob@abbvie.com

Erich Jarvis
Rockefeller University
ejarvis@rockefeller.edu

Mark Johnston
University of Colorado School of Medicine
mark.johnston@ucdenver.edu

Lynn Jorde
University of Utah School of Medicine
lbj@genetics.utah.edu

Monica J. Justice
Hospital for Sick Children
monica.justice@sickkids.ca

Robert W. Karp
National Institute of Diabetes and Digestive and Kidney Diseases, NIH
karpr@mail.nih.gov

Nicholas Katsanis
Duke University
katsanis@cellbio.duke.edu

Dave Kaufman
National Human Genome Research Institute, NIH
dave.kaufman@nih.gov

Eimear Kenny
Icahn School of Medicine at Mount Sinai
eimear.kenny@mssm.edu

Amit V. Khera
Massachusetts General Hospital/Broad Institute
avkhera@broadinstitute.org

Donna Krasnewich
National Institute of General Medical Sciences, NIH
dkras@mail.nih.gov

Anshul Kundaje
Stanford University
akundaje@stanford.edu

Pui-Yan Kwok
University of California, San Francisco
pui.kwok@ucsf.edu

Eric Lander
Broad Institute
lander@broadinstitute.org

Tuuli Lappalainen
New York Genome Center/Columbia University
tlappalainen@nygenome.org

Brittany Nicole Lasseigne
HudsonAlpha Institute for Biotechnology
blasseigne@hudsonalpha.org

Ryan M. Layer
University of Colorado, Boulder
ryan.layer@gmail.com

Charles Lee
The Jackson Laboratory for Genomic Medicine
charles.lee@jax.org

Christina Leslie
Memorial Sloan Kettering Cancer Center
cleslie@cnio.mskcc.org

Yun Li
University of North Carolina, Chapel Hill
yunli@med.unc.edu

Xihong Lin
Harvard University
xlin@hsph.harvard.edu

Mathieu Lupien
Princess Margaret Cancer Centre
mlupien@uhnres.utoronto.ca

Daniel MacArthur
Harvard University/Broad Institute/Massachusetts General Hospital
sartemik@broadinstitute.org

Laura A. Mamounas
National Institute of Neurological Disorders and Stroke, NIH
mamounas@ninds.nih.gov

Allison Mandich
National Human Genome Research Institute, NIH
allison.mandich@nih.gov

Tom Maniatis
Columbia University
tmaniatis@nygenome.org

Teri Manolio
National Human Genome Research Institute, NIH
manolio@nih.gov

Gabor T. Marth
University of Utah
gabor.marth@gmail.com

Eric Mendenhall
University of Alabama in Huntsville
eric.mendenhall@uah.ed

Karen Hayden Miga
University of California, Santa Cruz
khmiga@soe.ucsc.edu

Marilyn Miller
National Institute on Aging, NIH
millerm@nia.nih.gov

Karen Mohlke
University of North Carolina, Chapel Hill
mohlke@med.unc.edu

Alvaro Monteiro
Moffitt Cancer Center
alvaro.monteiro@moffitt.org

Ali Mortazavi
University of California, Irvine
ali.mortazavi@uci.edu

John Mudgett
Regional Pharma Consulting, LLC/ William Paterson University
jmudgett2@gmail.com

Richard Myers
HudsonAlpha Institute for Biotechnology
rmyers@hudsonalpha.org

Benjamin Michael Neale
Massachusetts General Hospital/Broad Institute
bneale@broadinstitute.org

Stefanie Nelson
National Cancer Institute, NIH
stefanie.nelson@nih.gov

John R. Nelson
GE Global Research
john.nelson@ge.com

Mukul Nerurkar
National Human Genome Research Institute, NIH
mukul.nerurkar@nih.gov

Elaine Ann Ostrander
National Human Genome Research Institute, NIH
eostrand@mail.nih.gov

Michael Pagan
National Human Genome Research Institute, NIH
michael.pagan@nih.gov

Kiara Palmer
National Human Genome Research Institute, NIH
kiara.palmer@nih.gov

Melissa A. Parisi
Eunice Kennedy Shriver National Institute of Child Health and Human Development, NIH
parisima@mail.nih.gov

William J. Pavan
National Human Genome Research Institute, NIH
bpavan@mail.nih.gov

Michael J. Pazin
National Human Genome Research Institute, NIH
michael.pazin@nih.gov

Dana Pe'er
Sloan Kettering Institute
peerster@gmail.com

Len Pennacchio
Lawrence Berkeley Laboratory
LAPennacchio@lbl.gov

Adam Phillippy
National Human Genome Research Institute, NIH
adam.phillippy@nih.gov

Sharon Emma Plon
Baylor College of Medicine
splon@bcm.edu

Katherine S. Pollard
University of California, San Francisco
kpollard@gladstone.ucsf.edu

Rudy Pozzatti
National Human Genome Research Institute, NIH
pozzattr@exchange.nih.gov

Jonathan Karl Pritchard
Stanford University
pritch@stanford.edu

Erin Ramos
National Human Genome Research Institute, NIH
erin.ramos@nih.gov

Soumya Raychaudhuri
Harvard Medical School/Broad Institute/Brigham and Women's Hospital
soumya@broadinstitute.org

Tim Reddy
Duke University
tim.reddy@duke.edu

Aviv Regev
Broad Institute
nnewman@broadinstitute.org

Heidi Rehm
Massachusetts General Hospital
hrehm@mgh.harvard.edu

Bing Ren
University of California, San Diego
biren@ucsd.edu

Marylyn D. Ritchie
University of Pennsylvania
marylyn@pennmedicine.upenn.edu

Laura Lyman Rodriguez
National Human Genome Research Institute, NIH
rodrigla@nih.gov

Meru Sadhu
National Human Genome Research Institute, NIH
meru@nih.gov

Neville Sanjana
New York Genome Center/New York University
nsanjana@nygenome.org

John Satterlee
National Institute on Drug Abuse, NIH
satterleej@nida.nih.gov

Lorjetta Schools
National Human Genome Research Institute, NIH
lorjetta.schools@nih.gov

Charlene A. Schramm
National Heart, Lung, and Blood Institute, NIH
schrammc@nih.gov

Lynn M. Schriml
University of Maryland School of Medicine
lschriml@som.umaryland.edu

Julie A. Segre
National Human Genome Research Institute, NIH
jsegre@nhgri.nih.gov

Beth Shapiro
University of California, Santa Cruz/Howard Hughes Medical Institute
bashapir@ucsc.edu

Jay Shendure
University of Washington
shendure@uw.edu

Stephen Thomas Sherry
National Center for Biotechnology Information, NIH
sherry@ncbi.nlm.nih.gov

Adam C. Siepel
Cold Spring Harbor Laboratory
asiepel@cshl.edu

Michael William Smith
National Human Genome Research Institute, NIH
smithmw@mail.nih.gov

John Stamatoyannopoulos
Altius Institute
jstam@altius.org

Lea M. Starita
University of Washington
lstarita@uw.edu

Taylorlyn Stephan
National Human Genome Research Institute, NIH
taylorlyn.stephan@nih.gov

Barbara E. Stranger
University of Chicago
bstranger@medicine.bsd.uchicago.edu

Ryan Tewhey
The Jackson Laboratory
ryan.tewhey@jax.org

Sarah Tishkoff
University of Pennsylvania
tishkoff@pennmedicine.upenn.edu

Jennifer L. Troyer
National Human Genome Research Institute, NIH
troyerj@mail.nih.gov

David Valle
Institute of Genetic Medicine, Johns Hopkins University School of Medicine
dvalle@jhmi.edu

Benjamin F. Voight
University of Pennsylvania
bvoight@pennmedicine.upenn.edu

Simona Volpi
National Human Genome Research Institute, NIH
simona.volpi@nih.gov

Ting Wang
Washington University School of Medicine
twang@wustl.edu

Meni Wanunu
Northeastern University
wanunu@neu.edu

Kris Wetterstrand
National Human Genome Research Institute, NIH
wettersk@mail.nih.gov

Anastasia L. Wise
National Human Genome Research Institute, NIH
anastasia.wise@nih.gov

Barbara Wold
Caltech Division of Biology and Biological Engineering
woldb@caltech.edu

Michael C. Zody
New York Genome Center
mczody@nygenome.org

Appendix 3: Related Project Summaries

Project Title	Project Acronym	Project URL	URL for Data	Project Summary	Sample Types	Assays/Data Types	Funding Agencies	Status
Functional Genomics								
Encyclopedia of DNA Elements	ENCODE	https://www.genome.gov/encode	https://www.encodeproject.org	Data collection and integrative analysis of human and mouse epigenomic and trancrilptomic data, including reference maps	Human and Mouse	Epigenomics Cell lines Transcriptomics Functional Genomics	NHGRI	Active
International Human Epigenome Consortium	IHEC	http://www.ihec-epigenomes.org/	https://epigenomesportal.ca/ihec/	Data collection and reference maps of human epigenomes for key cellular states relevant to health and diseases	Human	Transcriptomics Epigenomics	Consortium of projects funded by member nations	Active
Reference Epigenome Mapping Centers	REMC	http://www.road mapepigenomics.org/	https://www.encodeproject.org/matrix/?type=Experiment&award.project=Roadmap	Data collection, integrative analysis and a resource of human epigenomic data	Human (healthy)	Transcriptomics Epigenomics	NIH Common Fund	Completed
PsychENCODE	PsychEN CODE	http://grants.nih.gov/grants/guide/rfa-files/RFA-MH-14-020.html		Data collection and integrative analysis of human neural epigenomic, genomic, transcriptomic and proteomic data	Human	WES WGS Transcriptomics Epigenomics Proteomics	NIMH	Active
Functional Annotation of the Mammalian Genome	FANTOM	http://fantom.gsc.riken.jp		Data collection of CAGE transcriptomic data and data analysis to annotate human and mouse functional elements.	Human and Mouse	Transcriptomics CAGE	RIKEN	Active

4D Nucleome	4DN	https://commonfund.nih.gov/4Dn ucleome	https://www.4dnucleome.org	To understand the principles behind the organization of the nucleus in space and time, the role nuclear organization plays in gene expression and cellular function, and how changes in the nuclear organization affect normal development as well as various diseases.	Human	Multi-omics Cell lines Imaging	NIH Common Fund	Active
Genomics of Gene Regulation	GGR	https://www.genome.gov/27561317/genomics-of-gene-regulation/	https://www.encodeproject.org/matrix/?type=Experiment&award.project=	Determine how to develop predictive gene regulatory network models from genomic data	Human	Transcriptomics Epigenomics	NHGRI	Completed
Genotype-Tissue Expression Project	GTEx	http://www.gtexportal.org/home/		Data collection and analysis of variation in human gene expression, across individuals, and across >30 tissues from the same subjects	Human (healthy)	WGS WES Transcriptomics	NIH Common Fund	Active
Library of Integrated Network- based Cellular Signatures	LINCS	https://commonf und.nih.gov/LINCS/		Data collection and analysis of molecular signatures describing how different cell types respond to perturbing agents	Human	Transcriptomics Phosphoproteomics Cell lines Imagining Epigenomics	NIH Common Fund	Active
International Cancer Genome Consortium	ICGC	http://www.icgc. org/		Data collection and analysis of genomic, transcriptomic and epigenomic changes in 50	Human (tumor and normal)	WGS WES	Consortium of projects funded by	Active

				different tumor types (includes TCGA samples)		Transcriptomics Epigenomics	member nations	
The Cancer Genome Atlas	TCGA	http://cancergenome.nih.gov/		Data collection and analysis of genomic, transcriptomic, and epigenomic changes in ~30 different tumor types, and repository for DNA and RNA sequence data	Human (tumor and normal)	WGS WES Proteomics Transcriptomics Epigenomics	NHGRI, NCI	Completed
Non-Coding Variants Program	NoVa	https://www.genome.gov/27564944/noncoding-variants-program-nova/		Development of computational approaches to interpret sequence variation in non-coding regions, and assessment of approaches through targeted data collection	Various	Functional assays	NHGRI, NCI, NIDA	Active
Knockout Mouse Phenotyping Program	KOMP2	https://commonfund.nih.gov/KOMP2/		Data collection for standardized phenotyping of a genome-wide collection of mouse knockouts; member of International Mouse Phenotyping Consortium (IMPC)	Mice	Phenotypic	NIH Common Fund	Active
HubMap	HubMap	https://commonfund.nih.gov/hubmap		to facilitate research on single cells within tissues by supporting data generation and technology development to explore the relationship between cellular organization and function, as well as variability in normal tissue organization at the level of individual cells	Human (Healthy)	Transcriptomics Phosphoproteomics Imaging Epigenomics	NIH Common Fund	Active
Human Cell Atlas	HCA	https://www.humancellatlas.org/		To create comprehensive reference maps of all human cells—the fundamental units of life—as a basis for both understanding human health and diagnosing,	Human	Multiple	Investigator-orgainized effort	Active

				monitoring, and treating				
Toxicant Exposures and Responses by Genomic and Epigenomic Regulators of	TaRGET II	https://targetepigenomics.org	https://dcc.targetepigenomics.org	Multiple -omics measures of cellular response to toxicants.	Human	Multiple	NIEHS	Active
Extracellular RNA Communication	ERC	https://commonfund.nih.gov/exrna		to establish fundamental biological principles of extracellular RNA secretion, delivery, and impact on recipient cells; to describe exRNAs in human biofluids and the extent to which non-human exRNAs are present; to test clinical utility of exRNAs; and to provide a data and a resource repository for the community at-large.	Human		NIH Common Fund	Active
Sequencing for Variant Discovery and Association								
NHGRI Genome Sequencing Program (including multiple co-funding sources)	GSP	http://gsp-hg.org/	https://www.ncbi.nlm.nih.gov/gap	i. Exomes in Mendelian disease for gene/variant discovery-- resolve as many Mendelian disease as possible; families; ii. Well-powered exome and genome studies in common, multiple complex diseases, multiple designs (case/control, family, etc.). Understand genomic architecture of common disease.	Human	WES WGS Analysis	NHGRI, NHLBI, NEI, NIMH	Active
Trans-Omics for Precision Medicine	TOPMed	https://www.nhlbiwgs.org/	https://www.ncbi.nlm.nih.gov/gap	Genomes in case/control designs related to cardiovascular phenotypes. Additional -omics data added.	Human	WES WGS Proteomics Metabolomics Analysis	NHLBI	Active

Alzheimer's Disease Sequencing Project	ADSP	https://www.niagads.org/adsp/content/home	https://www.niagads.org/adsp/content/home	Exomes (case/control) and genomes (families) in AD	Human	WES WGS Analysis	NIA	Active
Type 2 Diabetes Genes	T2DGene s	www.type2diabetesgenetics.org/	www.type2diabetesgenetics.org/	Exomes and genomes in T2D	Human	WES WGS Analysis	NIDDK	Completed
Population Architecture using Genomics and Epidemiology	PAGE	https://www.pagestudy.org/	https://www.pagestudy.org/	Genotyping disease risk variants in diverse non-European populations	Human	Genotype	NHGRI	Completed
Electronic Medical Records and Genomics	eMERGE	https://www.genome.gov/27540473/electronic-medical-records-and-genomics-emerge-network/		Type disease-associated variants in patients at scale; integrate with electronic medical records, assess utility of variant data.	Human	Genotype Targeted sequence EHR	NHGRI	Active
Gabriella Miller Kid's First	GMKF	https://commonfund.nih.gov/kidsfirst	https://kidsfirst drc.org/	Childhood cancers and structural birth defects; families and small cohorts	Human	WGS Analysis	NIH Common Fund	Active
Biobank/ Large Cohort Sequencing								
All of Us	AoU/PMI	https://allofus.ni h.gov	https://www.researchallofus.o rg/	To create a national resource of over a million Americans' health information. AoU aims to oversample in underrepresented communities (racial/ethnic minorities, women, etc.). The program will sequence whole genomes and generate genotype data; collect health/lifestyle/environmental information. The program is part of the U.S.'s Precision Medicine Initiative	Human	WGS Genotype Transcriptomics EMR	NIH Common Fund	Active
UK BioBank	UKBB	https://www.ukbiobank.ac.uk/		Aims to improve prevention, diagnosis, and treatment of a wide range of serious and life- threatening illnesses. Follows the	Human	Samples EMR Imaging Genotyping WES	Wellcome Trust (with funding from Welsh, Scottish, British governments	Active

				health/well- being of over 500,000 volunteer participants and provides updated health information to qualified researchers.		Environmental Wearables	and other non-profits)	
Saudi Human Genome Project	SHG	https://www.saudigenomeprogram.org/en/	http://genomics.saudigenomeprogram.org/en/researchers/database-access/	An effort to solve genetic diseases and apply personalized medicine in Saudi Arabia by sequencing more than 100,000 individuals. The program offers free DNA sequencing and pre-marital genetic screening to Saudi residents.	Human	WES Genotyping	Saudi Arabian government	Active
NIMH Repository and Genomics Resource	NIMH-RGR	https://www.nimhgenetics.org/		Provides a collection of over 150,000 well-characterized, high quality patient and control samples from a range of mental disorders (i.e. autism, epilepsy, schizophrenia, etc.). Centers receive, process and store biomaterials, distribute data to qualified researchers	Human	Stem cells Phenotypic data Cell lines Biopsies WGS WES	NIMH	Active
Australian Genomics Health **Futures Mission**		http://www.healt h.gov.au/internet/budget/publishing.nsf/Content/budget2018- factsheet65.htm		A 10-year national approach to addressing clinical genomics in Australia using a cohort of 200,000 individuals. The program will invest in: new and expanded studies for rare diseases, cancers, and complex disease; clinical trials and collaboration opportunities and career pathways; ELSI	Human	Multiple	Australian government	Active

Estonia Personalized Medicine Programme (100,000 genomes)	PMP	https://www.geenivaramu.ee/en/about-us/estonian-biocentre		Recruit 100,000 additional participants for the Estonian Biobank and genotype them (genome- wide) and provide personalized reports in the national e-health portal. People of other nationalities (ex. Russian, Ukrainian, etc.) are represented in the initial cohorts. Secondary research includes: ancient DNA, population genetics, mitochondrial/Y-chromosome, and cellular stress	Human	WGS Genotype EMR	Estonian government (Estonia Ministry of Social Affairs/National Institute for Health Development/Estonia n Genome Center at University of Tartu)	Active
Japan Initiative on Rare and Undiagnosed Diseases	IRUD	https://www.ame d.go.jp/en/program/IRUD/		National research consortium that connects over 400 hospitals with 34 IRUD clinical and analysis centers. Clinical and genetic data is gathered for each case and findings are shared with committees, patient-matching sites, and with qualified investigators within and outside	Human	WES WGS EMR Clinical trials Pharmacogenomics	Japan Agency for Medical Research and Development	Active
France Medecine Genomique	NA			Ten year plan to construct a medical and industrial system to introduce precision medicine into the French healthcare system.	Human	WGS	Alliance Aviesan	Active
BGI Million Chinese Genomes	NA	https://www.bgi. com/us/company/news/bgi-	https://db.cng. org/cmdb	To sequence the genes of 1 million Chinese residents, including individuals across the country and of all ethnicities. The main	Human	WGS WES NIPT Transcriptomics	China's Ministry of Science and Technology	Active

		publishes-largest-ever-genomic-study-chinese-population-discoveries-140000-genomes-		goal is to understand how Chinese people "transform from health to disease", how the environment and interactions between genes and external factors influence the country's health.				
Miscellaneous Related								
Human Genome Reference Program	HGRP	https://www.genome.gov/pages/about/nachgr/september2018agendadocuments/sept2018council_hg_reference_program.pdf		The next iteration of NHGRI's support for the human genome reference assembly (formally GRC) is an upcoming program called the Human Genome Reference Program. The HGRP will sequence additional high-quality haplotype-resolved genomes from diverse populations; improve and implement the next generation of reference representations; develop new technology and bioinformatic tools to use on the reference	Human (and outside-supported mice, zebrafish, chicken)	Genomic transcripts Bioinformatic tools	NHGRI (previously included Wellcome-Sanger)	Active
Brain Research through Advancing Innovative Neurotechnologies Initiative	BRAIN	https://www.braininitiative.nih.gov/		Aims to revolutionize our understanding of the human brain by developing new technology that images, tracks, and visualizes brain cells, circuits, neural activity, and other integrated neurological approaches.	Human	Imaging Transcriptomics Epigenomics Proteomics Metabolomics Physiological Measurements ELSI	NIH	Active
Clinical Genome Resource	ClinGen	https://www.clinicalgenome.org/		A genomic knowledge base (resource) that defines clinical relevance of genes and variants for use in precision medicine and research. Consortium shares	Human	EMR Phenotypic WES WGS	NHGRI, NICHD	Active

				genomic and phenotypic data, standardizes clinical annotations, develops machine-learning approaches to improve calling; and disseminates the knowledge and resources				

Appendix 4: Breakout Session Guidance

Guidance to Breakout Session Participants

Breakout sessions have two assigned co-chairs and assigned participants. *Please participate in your assigned session.*

Breakout co-chairs and participants are asked to propose and discuss specific ideas for what NHGRI should do in the next 5-10 years. These ideas can be in the form of knowledge to be gained, resources and/or capabilities to be created, and/or specific projects.

- For each, consider: What is its goal? Why is it important? Why should NHGRI undertake it?
- If NHGRI has an existing commitment in this area already, should that change? How (e.g., in the way we approach the question)? Why?
- If the proposed idea(s) are likely to connect to topics of other breakout sessions, how will they connect, and where are the synergies between them?
- If relevant, how should NHGRI's efforts fit into the larger ecosystem of other national and international efforts?

Some questions above will apply more to some breakout topics than others.

We ask co-chairs to come prepared to begin the discussion. They should review the ideas the meeting participants provided prior to the meeting, and solicit additional ideas and explanations, justifications, areas of synergy, etc. from the participants.

Following the breakout session, co-chairs will have 10 minutes per breakout group to report back to all the workshop participants with a summary and highlights of the discussion including, e.g., key ideas and recommendations; items of consensus; areas of disagreement. Following reports from each of the 3 breakout groups, the discussion moderators should lead a broader discussion to identify overlapping ideas across the breakout sessions and ideas of high interest to the group, and also make sure to ask the workshop participants if any relevant items appeared not to be covered.

The Day 1 Breakouts (sessions 1,2,3) will cover a set of potentially related topics on variant discovery:

1) How much more sequencing, if any, is needed to study Mendelian and common inherited disease, both in general and in the context of diverse populations? How should these activities be organized? *What should NHGRI do in this area? Why?*

2) How and why to approach structural variation and other "hard to measure" variation? What technologies and approaches are needed?

3) How and why to approach new complex features - e.g., GxE, epistasis?

The Day 2 Breakouts (sessions 4,5,6) will cover a set of topics on assessing genome function each from a different perspective:

4) What is needed to identify all regulatory elements (enhancers, promoters, insulators, RNA stability, etc.) as well as genes (including protein-coding isoforms, lncRNAs, smORFs)? How can we characterize the function of all genes and regulatory elements in different biological contexts?

Topics may include:

- Technologies, strategies, and tools to do this; opportunities to advance these.
- Which are important and why?
- What should "characterizing" genes entail?
- What should "characterizing" regulatory elements entail?
- How much more is there to do? Can this task be "finished"? Is there a definable stopping point?
- Should comparative genomics be used to address these questions?

5) What is needed to determine the functional consequences of variants, both individually as well as combinations of variants and, ultimately, all variants in a genome?

- Experimental systems:
    - Experimental systems for different biological contexts. What makes them appropriate?
    - What needs to be developed to understand variant function for known disease variants (e.g., model organism assays)? For understanding the potential effect of any variant (e.g., MPRA assays in organoids)?
    - How can we bring the relevant biological context to the assessment of variant function (e.g., cell type or tissue type)?
    - What are the best systems for understanding both coding and noncoding variation?
- Experimental and computational approaches
    - What new capabilities, data, and tools are needed, especially to move beyond one variant at a time?
- How do we solve the problem that this experimental space is potentially so large? (e.g., brute force development of ultra-high throughput methods; sensible prioritization; better fundamental understanding?

6) What is needed (experimental data and computational analysis) to accurately predict the consequences of genetic variants on biological structure and function? How do network perspectives fit into this? What is the role of data integration?

Topics may include:

- How do we understand the action of coding and noncoding variants in combination (cis and trans)?
- How do we ultimately relate findings about genes, regulatory elements and variants in them to biological phenoptypes?
- Computational modelling; AI/ML; sparse data
- What data types are needed for what purpose? What are the most useful data types, and how do we validate them?

Appendix 5: Breakout Session Assignments
Breakout Session Assignments – Day 1

Breakout 1	Breakout 2	Breakout 3
(*Roosevelt Room*)	(*Madison Room*)	(*Jefferson Room*)
Co-Chairs:	**Co-Chairs:**	**Co-Chairs:**
Eric Boerwinkle & Nancy Cox	**Charles Lee & Karen Miga**	**Andy Clark & Eimear Kenny**
Abecasis, Gonçalo	Abate, Adam	Ahituv, Nadav
Addington, Anjene	Adey, Andrew	Bahcall, Orli
Aiden, Erez Lieberman	Bierut, Laura	Bellen, Hugo
Bejerano, Gil	Boeke, Jef	Boehnke, Mike
Buenrostro, Jason	Bult, Carol	Bonham, Vence
Carvill, Gemma	Burgess, Shawn	Califano, Andrea
Chadwick, Lisa	Cahill, Eileen	Chakravarti, Aravinda
di Francesco, Valentina	Chang, Howard	Cho, Judy
Fallin, Danielle	Chanock, Stephen	Chow, Clement
Felsenfeld, Adam	Chung, Wendy	Cohen, Barak
Gasperini, Molly	Craven, Mark	Cool, Jonah
Gatlin, Tina	Crawford, Dana	Ecker, Joe
Graveley, Brent	Davis-Dusenbery, Brandi	Farley, Emma
Haines, Jonathan	Dennis, Megan	Gibbs, Richard
Hall, Ira	Eichler, Evan	Gilchrist, Daniel
Hardison, Ross	Fletcher, Colin	Goetz, Kerry
Hirst, Martin	Frazer, Kelly	Ideker, Trey
Jacob, Howard	Gebo, Kelly	Kaufman, Dave
Jorde, Lynn	Gerstein, Mark	Khera, Amit
Justice, Monica	He, Chuan	Kundaje, Anshul
Karp, Robert	Hicks, Stephanie	Labosky, Trish
Katsanis, Nico	Ionnidis, Nilah	Leslie, Christina
Lander, Eric	Jarvis, Erich	Lin, Xihong
MacArthur, Dan	Johnston, Mark	Lupien, Mathieu
Mamounas, Laura	Krasnewich, Donna	Marth, Gabor
Mandich, Allison	Kwok, Pui	Miller, Marilyn
Mohlke, Karen	Lappalainen, Tuuli	Monteiro, Alvaro

Myers, Rick
Nelson, Stefanie
Pagan, Michael
Pavan, Bill
Pennacchio, Len
Pozzatti, Rudy O.
Reddy, Tim
Schramm, Charlene
Segre, Julie
Shapiro, Beth
Sherry, Steve
Stamatoyannopoulos, John
Starita, Lea
Stranger, Barbara
Tishkoff, Sarah
Valle, David
Wang, Ting
Wold, Barbara
Zody, Mike

Lasseigne, Brittany
Layer, Ryan
Li, Yun
Maniatis, Tom
McAllister, Kimberly
Mendenhall, Eric
Mudgett, John
Muenke, Maximilian
Nelson, John
Parisi, Melissa
Phillippy, Adam
Raychaudhuri, Soumya
Ren, Bing
Schriml, Lynn
Smith, Michael
Tewhey, Ryan
Troyer, Jennifer
Voight, Ben
Volpi, Simona
Wanunu, Meni

Mortazavi, Ali
Ostrander, Elaine
Pazin, Mike
Pe'er, Dana
Plon, Sharon
Pollard, Katie
Pritchard, Jonathan
Ramos, Erin
Regev, Aviv
Rehm, Heidi
Ritchie, Marylyn
Sadhu, Meru
Sanjana, Neville
Satterlee, John
Shendure, Jay
Siepel, Adam
Stephan, Taylorlyn
Wetterstrand, Kris
Wise, Anastasia

Breakout Session Assignments – Day 2

Breakout 4	Breakout 5	Breakout 6
(*Roosevelt Room*)	(*Madison Room*)	(*Jefferson Room*)
Co-Chairs:	**Co-Chairs:**	**Co-Chairs:**
Ross Hardison & Bing Ren	**Wendy Chung & Kelly Frazer**	**Trey Ideker & Christina Leslie**
Abate, Adam	Abecasis, Gonçalo	Adey, Andrew
Ahituv, Nadav	Addington, Anjene	Boehnke, Mike
Aiden, Erez Lieberman	Bellen, Hugo	Buenrostro, Jason
Bahcall, Orli	Bierut, Laura	Bult, Carol
Bejerano, Gil	Boeke, Jef	Burgess, Shawn
Boerwinkle, Eric	Bonham, Vence	Cho, Judy
Cahill, Eileen	Califano, Andrea	Chow, Clement
Carvill, Gemma	Cardon, Lon	Cohen, Barak
Chadwick, Lisa	Chakravarti, Aravinda	Cool, Jonah
Cox, Nancy	Chang, Howard	Craven, Mark
Dennis, Megan	Chanock, Stephen	di Francesco, Valentina
Fallin, Danielle	Clark, Andy	Eichler, Evan
Farley, Emma	Crawford, Dana	Felsenfeld, Adam
Feingold, Elise	Davis-Dusenbery, Brandi	Gasperini, Molly
Fletcher, Colin	Gebo, Kelly	Gerstein, Mark
Gatlin, Tina	He, Chuan	Gilchrist, Daniel
Gibbs, Richard	Ionnidis, Nilah	Goetz, Kerry
Graveley, Brent	Johnston, Mark	Haines, Jonathan
Hall, Ira	Justice, Monica	Hicks, Stephanie
Hirst, Martin	Kaufman, Dave	Jacob, Howard
Jarvis, Erich	Kenny, Eimear	Jorde, Lynn
Katsanis, Nico	Lander, Eric	Karp, Robert
Krasnewich, Donna	Lee, Charles	Khera, Amit
Kwok, Pui	Li, Yun	Kundaje, Anshul
Labosky, Trish	MacArthur, Dan	Lappalainen, Tuuli
Layer, Ryan	Mamounas, Laura	Lasseigne, Brittany

Lupien, Mathieu
Marth, Gabor
Mendenhall, Eric
Mudgett, John
Muenke, Maximilian
Parisi, Melissa
Plon, Sharon
Pozzatti, Rudy O.
Sadhu, Meru
Sanjana, Neville
Satterlee, John
Shapiro, Beth
Shendure, Jay
Sherry, Steve
Stamatoyannopoulos, John
Tewhey, Ryan
Troyer, Jennifer
Voight, Ben
Volpi, Simona
Wanunu, Meni
Wold, Barbara

Mandich, Allison
McAllister, Kimberly
Miga, Karen
Mohlke, Karen
Myers, Rick
Nelson, John
Ostrander, Elaine
Pavan, Bill
Pazin, Mike
Pennacchio, Len
Reddy, Tim
Rehm, Heidi
Schramm, Charlene
Smith, Michael
Starita, Lea
Stephan, Taylorlyn
Stranger, Barbara
Tishkoff, Sarah
Valle, David
Wise, Anastasia

Lin, Xihong
Maniatis, Tom
Miller, Marilyn
Monteiro, Alvaro
Mortazavi, Ali
Nelson, Stefanie
Pagan, Michael
Pe'er, Dana
Phillippy, Adam
Pollard, Katie
Pritchard, Jonathan
Ramos, Erin
Raychaudhuri, Soumya
Regev, Aviv
Ritchie, Marylyn
Schriml, Lynn
Segre, Julie
Siepel, Adam
Wang, Ting
Wetterstrand, Kris
Zody, Mike

Appendix 6: Project Suggestions from Breakout Sessions

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?

- Sequencing studies that span the continuum from rare Mendelian to common polygenic diseases.
- Inclusion of diverse participants in sequence-based studies, with a focus on high impact diseases and co-funding.
- Human Genome Project V2: annotate all genome elements using comparative genomics and molecular perturbations, natural and engineered variation
- Catalyze the development and hardening of novel technologies and software tools to increase the quality and depth of genomic data
- Proximal multi-omics (e.g., DNA methylation from long read sequence data) in cells, model organisms, and humans using natural and engineered variation in diverse contexts
- Technology development for better sequencing

Breakout 2: How and why to approach structural variation and other "hard to measure" variation?

- New/Improved methods to manipulate/deliver HMW DNA (increasing throughput and quality)
- New sequencing technologies outside of the established systems
- Rebooting the reference genome: Pan-genome references from more, higher quality, diverse individuals
- Telomere-to-telomere phased assemblies of diploid genomes
- New functional methods for difficult genomic regions (Hi-C, ENCODE, etc.)
- Improved methods for imputation to associate difficult genomic regions with function

Breakout 3: How and why to approach more complex features- e.g., GxE, epistasis?

- Projects that leverage wearable devices, incorporate environmental surveys (e.g., PhenX Toolkit) and geocoding to link EHRs, population datasets, biomarkers, and genetic data.
- New machine learning approaches to cope with dimension explosion in integrative analysis
- Polygenic risk score models that integrate familial, lifestyle, and environmental exposures
- Functional mapping projects that charts the genome, epigenome, post-transcription, cell signaling/inter-cellular interactions, perturbations, and etc. in 2-3 specific biological contexts.
- Haplotypic series project on a subset of suspected disease genes to determine regulatory functions, interaction with trans-factors, and gene dosage effects on disease risk.
- Apply network approaches to identify indirect causal interactions; validate with larger datasets and/or direct molecular assays
- Epistasis in the context of evolution: how much gene-gene interaction is expected in standing genetic variation, how does that change in an expanding population?
- Gene regulatory network studies in diverse populations to test inter-population divergence and conservation
- Analysis of individuals with homozygous LOF mutations and no disease phenotypes to identify large effect suppressor variants

Breakout 4: Identification and characterization of all genes and regulatory elements

Projects that are ready in the next few years

- Comprehensive atlas of cis elements in DNA and RNA
- Pilot project to thoroughly assess function in the genome and epigenome
- FUNCODE – choose subset (1%) of human genome, perturb and define functional consequences
- FUNCODE for model organisms

- Perturbations followed by organismal phenotypes in model organisms
- Knockout all genes in human haploid cells, phenotype
- 1000 epigenome project: 1000 cell types
- Universal vertebrate gene nomenclature
- Atlas of elements under selection (evolutionary analysis)
- Atlas of nuclear architecture

Projects that are designed for the longer term (5-10 years)

- Scalable methods for phenotyping
- Scalable methods to perturb DNA in human and model organisms
- End to end complete human genome
- Assemble full genome sequences for a million species
- Full epigenome in single cells
- Comprehensive catalog of protein function

Technology development

- Effective assays for molecular phenotypes, at single cell resolution (scRNA-seq, scATAC-seq, etc.)
- High throughput means for perturbing the genome (such as Perturb-seq) coupled with molecular phenotyping
- Informative experimental model systems
- Tools for monitoring temporal dynamics
- Epigenome mapping tools at high resolution and at scale
- Standards for evaluating phenotypic effects: What evidence is sufficient to consider a candidate regulatory element to be functional? What panel of cell lines, tissues, organoids, organisms will be sufficient to give confidence that the potential phenotype of mutated ccREs has been sufficiently evaluated?
- Machine learning approaches
- Resource development of gene editing/silencing/overexpression for high-throughput studies in model organisms
- In situ spatially resolved genomic profiling of solid tissues/organs (e.g. MERFISH)
- Technologies to synthesize long DNA (kilobases) economically and at scale.
- Complete and error-free genome assemblies

Breakout 5: Determining the functional consequences of variants acting individually and in combination

- Single-cell GTEx-like project: learn proximate effects of nucleotide changes in people; could include chromatin accessibility and/or metabolomics
- Predictive modeling of genome function based on systematic data sets
- Methods/resources to aggregate, integrate data: population allele frequency, clinical phenotypes, function; model organism data
- Constraint analysis of human genomes: include regulatory regions; link to phenotype data
- Identify genetic variants that affect RNA structure, nucleic acid modifications using new assays
- Determine correlation between RNA abundance and protein abundance/localization
- Technology for longer DNA synthesis (better, faster, cheaper)
- Technology for longer read nucleic acid sequencing (DNA, RNA)
- Technology for multi-omics assay (e.g. DNA, RNA) from a single biosample (single cells)

Breakout 6: Accurate prediction of the regulatory consequences of variants, and modeling gene regulation

- Predict trajectory of genome function from DNA sequence: genome-wide chromatin accessibility, enhancer activity, gene expression during perturbation time course
- Generate a comprehensive map of cellular mechanisms: focus on interpreting genetic variation underlying complex diseases
- Explore and develop modern machine learning methodologies (including deep learning, graphical models) for problems that elucidate principles of gene regulation and identify causal variants in disease
- Develop experimental systems (single cell) and computational methodologies to model cell-cell interactions and screen how genes, elements and variants modulate these interactions
- Move functional screening technologies and epigenomic assays into relevant biological systems and generate sufficiently robust, abundant data to train causal models
- Systematic measurement of protein locations and interactions for systems underlying common disease
- Use modeling to determine which data types (and biological systems) should be invested in further (versus less informative data or already abundant data)

Appendix 7: Ideas From Topic Sessions

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

1. **Complex Diseases:** To define a robust foundation for the genetic analysis of complex diseases (with sufficient disease cases), NHGRI should encourage, help organize, and provide support (in collaboration with other ICs and foundations) for large-scale genome sequencing studies for an exemplar set of 10 complex diseases that are expected to cover a range of genetic and environmental architectures. This goal should be accomplished over 5 years and expanded to other complex traits over 10 years. The diseases should be chosen based on factors including:
    - impact on public health
        - impact on health disparities (e.g. racial, ethnic, gender)
        - availability of large numbers of diverse, well phenotyped, broadly consented study participants
        - existence of a strong, highly collaborative disease genetics community
        - substantial support from the relevant NIH institute and/or other funders
        - availability or attainability of appropriate biobanked resources
    - combined with array-based genotyping of many more individuals beyond those sequenced who can augment sample size and association power.

2. **Mendelian disease gene identification** (10 year project): With the hypothesis that there will be one or more organismal phenotypes for every gene in the genome (current tally ~20%), NHGRI should support work with an emphasis on:

    - Sample recruitment from diverse worldwide populations with robust phenotypic information
    - WES and WGS to identify candidate causative variants
    - Analysis using enhanced human and model organism databases and expanded atlases of proximal phenotypes (transcriptome, proteome, metabolome) for an extensive set of genetic variants
    - Scalable functional testing of robust candidate variants in cellular and model organism systems
    - Expanded efforts to explore the genetic interface between Mendelian disease and common complex traits

3. **Perturbational and combinatorial functional genomics and microscopy at scale**. To obtain comprehensive causal trans-regulatory networks and high performance variant effect prediction models, NHGRI should fund efforts that:
    - stimulate development and use of scalable sequencing + microscopy-based technologies
    - For saturated marginal and combinatorial genetic and epigenetic perturbation of coding and non-coding elements optionally centered around transcription factors/RNA binding proteins/chromatin-modifiers/lncRNAs
    - Followed by high-throughput mapping of multi-modal molecular and cellular phenotypes
        - in diverse spatial, temporal, cellular contexts/environments, ancestry groups and model organisms relevant to disease.

4. **Functional genomics at scale: cellular and model organisms**. To interpret trait-associated genetic variation, NHGRI should support an expanded cell-type specific functional genomics project, modeled in part, after GTEx, for individual cell types. This would include:

- Collection of approximately 400 cell types from a human cohort of approximately 1,000 individuals, including sequencing of genomic DNA of all donors, plus mRNA and small RNA seq, ATAC-seq, proteomics, genome-wide epigenomics.
- Characterization of cis- and trans-eQTLs, and other xQTLs
- Considerations: Sex-balanced design, ancestry, developmental stage (both adult and prenatal?), cells within tissues vs sorted cell-types, living donors vs post-mortem donors?, pilot study of multi-omics on part of genome?

Beyond assessing function of naturally occurring genetic variation, NHGRI should support a functional genomics project to assess function of genome-edited sequence and targeted mutagenesis of target genes.

5. **Statistical and computational methods development**. NHGRI should encourage development and maintenance of statistical and computational methods and community benchmarking efforts to learn (for example)

    - Models that capture causal cis- and trans- networks from large-scale natural and artificial perturbation datasets
    - Models that can generalize across context (cell state, species, longitudinal axes, ancestry, genetic background, sex, cellular or organismal environments)
    - To effectively map sequence, molecular profiles and variants in repeat regions (transposable elements, satellite repeats)
    - Models that can effectively integrate diverse data modalities (sequencing and imaging)
    - Models that can integrate across variant allele frequency and variant classes
    - Models that are interpretable and/or effective hypothesis ranking engines to assist follow up experiments and data-driven experimental design

6. **Experimental methods development**. NHGRI should support experimental methods development in multiple areas including:
    - long-read sequencing technology to better define complex genomic regions and rearrangements, support RNA-seq haplotyping and epigenome profiling
    - adapting genomics assays to single cell assays
    - modifying/developing multi-omics assays that can work in concert on the same cells
    - genome/epigenome modifying tools, including delivery

7. **Interoperability and usability of data resources**. NHGRI should support development of a resource to integrate genetic variation and functional genetic information. This resource should collate into a database annotation of genetic variants:

    - Population-specific variant frequency
    - Annotation with context-specific predictive functional scores
    - Association with phenotypes
    - Connection to gene and functional element function
    - Relation to evolutionary information including selection and constraint

    NHGRI should incentivize usability of computational tools and models via:

- Interactive adaptive user interfaces and visualization specifically designed for discovery, exploration, and hypothesis queries in natural or structured language formats
- Smart dataset search and recommendation engines
- Seamless linking of version controlled and DOI supported raw data hubs -> processed, harmonized data hubs -> software hubs -> model zoos -> scientific literature

8. **As an integral part of the investigation of genetic diseases**, NHGRI should support the development of model organisms and model systems to identify the genetic architecture of Mendelian and complex traits and to evaluate candidate causative variants. NHGRI should further facilitate the functional and phenotypic analyses of genes and variants responsible for Mendelian and complex traits on non-inbred, complex genetic backgrounds to identify candidate orthologous human modifier genes and epistatic factors.

9. Given the large and rapidly growing need for genome scientists, for scientists and physicians using genome data, and members of the lay public interested in genomics, **NHGRI should substantially increase its budget for training and outreach**.

10. **Wide data sharing is critical in (genome) science.** NHGRI should seek whenever possible to focus large-scale genetic studies on deeply phenotyped, broadly consented individuals where this consent status is clearly established. Since this is not always possible, NHGRI also support development of privacy-preserving statistical/computational methods to allow as much information as possible to be shared even when individual-level data cannot be shared.

Topic 2: Addressing Basic Research Questions that Anticipate Clinical Needs

1. **Deep catalog of somatic events across many tissues** and many different ages of individuals. Perform colony mapping of materials (1000s….). Perform single cell analysis and integration with GTEX-analyses.
2. **scGETx** - "GTEX on Steroids" – warm autopsy of patients across multiple tissues, target individuals with high risk of disease. Comprehensive readout of RNA, and regulatory elements activity as a function of human genetic variation, both inherited and somatic at single cell level. Aim for 1000 individual subjects. Accessible tissues: blood, skin, gut
3. **Deep perturbation profiling.** Include proximal molecular readout of variants at every base in all exons, RNA elements, DNA regulatory elements. Prioritization: Disease SNPs, VUS in disease genes, HiC contacts, sites of RNA structure, RNA modifications.
4. **Single cell multiomic technology**. Multiple genomic modalities and perturb in single cells. Create the computational framework to interpret multi-omic dataset.
5. **Improved tools for data visualization** to improve data interpretation and usefulness for the community. Include communication of genomic data to clinicians and public for genomic literacy.
6. **Develop better tools for evidence to be used in medicine**. From existing clinical data, digital medicine and real world data. Define evidence codes such that diagnostic testing lab directors and physicians can use for data from model organisms to interpret variants to maximize the use of that data.
7. **Increase the computational biology workforce**.
8. **Plan to work up VUS for 25% of most important genes in human health**
9. **Deep phenotyping of individuals with biallelic or heterozygous loss of function alleles**. Identify through academic and commercial testing sources. Phenotype through the UDN network or similar efforts.

10. **Integrate familial, lifestyle, and environmental exposures into PRS** models to improve their prediction of disease phenotypes.
11. **Systematic assessment of off-target effects to advance use of genome engineering in medicine**.
12. **How can we detect larger structural variants using long-read sequencing or other technologies in the clinical setting?**
13. **Genomic privacy**: Develop appropriate tools to improve sharing of data from clinical labs and individuals while maintaining appropriate privacy issues.

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

- Cis-linking variants and regulatory elements to their target genes
- Gene function: what does every gene do?
- Trans-regulatory effects within cells: network effects and cellular phenotypes
- Effects of genes on development and organism phenotypes
- Understanding G x E

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

- Large scale perturbations in pools of cells, followed by high content, high resolution data
- Next-generation QTLs with better phenotype: single cell profiles, spatial assays in tissue, etc.), EHR
- Transferable gene regulatory or multi scale models from multi-layered data across biological systems

Technology development

- Test many variants with faster, cheaper, higher resolution, quantitative phenotypes
- Comprehensive data/sample resources with impact across domains
- Develop models from G to P that are either interpretable (non black box; mechanistic) and/or predictive (causal); ideally but not necessarily jointly
- Scaled accessible data platforms spanning genotypes and diverse phenotypes

From Audience

- Phased T2T assembly of at least one human genome. Comprehensive understanding of the genome: complete T2T reference genomes (data resources). First one should be non-European; should be African to max variation. The data would facilitate generation of synthetic human chromosomes. Should it be one or a panel?
- Use this to functional characterize these elements, ideally in a haploid cell line like Hap1, could be done in parallel
- Scaling: how to leverage the early complete ones for scaled application later on
- Technological opportunities for hard to sequence regions; What are the remaining barriers and needed resources? (to be laid out by leaders in field)
- Data residing across locations: hospitals, industry. Can NHGRI become a designated agent, or engage FNIH in this role, so that there is agency to port a patient's data (the patient would assign this right)
- Subcellular addition to the descriptions above. Nucleus and otherwise. Need cell biology data. See point 6 for justification.

- Molecular profiles following perturbation has limitations. (1) effect may not be seen in the data, multiple layers are redundant; (2) combos may be difficult to do comprehensively (3) pleiotropy; indirect distal effects difficult to distinguish from direct. Key: structural information
- Choose what to profile/perturb based on hypothesis from our current data (i.e. computational predictions); use structured knowledge
- When to perturb? Homeostasis buffers out the impacts. Needs to be done dynamically. Development, environmental stimulus.
- Model organisms. Innovate the existing model organisms. Building of models with substantial amount of human DNA. Models as tools to enable generic human biology. Genetic diversity and phenotyping.
- Training in genome literacy. Public, lay community. What genomes mean.
- We have limited understanding of non-disjunction and understanding the centromere is critical for this. Need to think about variation in the centromeres. Generate models
- Worry that full mechanistic model can be challenging absent modeling of behavior.
- Complexity of computational aspects. Transfer learning, or using the genetics to choose the PSR tails, put us in cond. Inference world…. We need to continue the commitment to CS!
- Sequence trios to understand mutational pattern (see above)
- Comparative genomics:

Appendix 8: Ideas From Synthesis And Prioritization Session

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

- Experimental methods development: sequencing technology, single cell and multi-omics assays, and - genome/epigenome modifying tools
- Functional genomics at scale: single-cell extended GTEx-like project, human and model organisms
- Statistical and computational methods development: data integration, networks, models (transferable, interpretable)
- Interoperability and usability of data resources: phenotypes, genes and elements; context-specific predictive scores, interactive interface, smart search/recommendation engines
- Perturbational and combinatorial functional genomics and microscopy: to obtain comprehensive causal trans-regulatory networks and high-performance variant effect prediction models
- Mendelian disease gene identification: samples from diverse populations, WES and WGS, functional testing of candidates, interface with common complex traits
- Identify the genetic architecture of Mendelian and complex traits and evaluate candidate variants: development of model organisms/systems, identify candidate orthologous human modifier genes
- Training and outreach: genome scientists, scientists and physicians using genome data, and members of the lay public interested in genomics
- Facilitate data usability: focus studies on deeply phenotyped, broadly consented individuals, and support development of privacy-preserving statistical/computational methods
- Genetic analysis of complex diseases: large-scale genome sequencing studies for an exemplar set of 10 complex diseases, covering a range of genetic and environmental architectures

Topic 2: Addressing Basic Research Questions that Anticipate Clinical Needs

- Deep perturbation profiling: proximal molecular readout of variants at every base in all exons, RNA elements, DNA regulatory elements
- Single cell GTEx-like project: target individuals with high disease risk of disease, comprehensive readout of RNA and regulatory element activity, 1000 individuals
- Improved tools for data visualization: improve data interpretation, communication to clinicians and general public
- Single cell multi-omic technology: multiple genomic measurements in single cells, computational framework to interpret
- Increase the computational biology workforce
- Approaches to detect larger structural variants in the clinical setting
- Deep catalog of somatic events: across many tissues and many individuals of different ages, single cell analysis
- Better tools for evidence to be used in medicine: define evidence codes such that diagnostic testing lab directors and physicians can use for data from model organisms/systems
- Deep phenotyping of individuals with heterozygous loss of function alleles
- Develop appropriate tools to improve sharing of data from clinical labs and individuals while maintaining appropriate privacy
- Characterize Variants of Unknown Significance: for 25% of most important genes in human health
- Integrate familial, lifestyle, and environmental exposures into PRS models to improve their prediction of disease phenotypes
- Systematic assessment of off-target effects to advance use of genome engineering in medicine

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

- Complete atlas of genes and their function: Genome-wide perturbations; cell lines, stem cells, organoids, animal models; transcriptome and open chromatin; proliferation, drug resistance, etc.
- 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 variants in organisms and organoids: single cell analysis, measure inter-cellular interactions, high throughput animal systems
- High-throughput measurement of trans-regulatory effects: all genes, diverse cell types, network readout, cell/tissue type specific effects
- 1000 epigenome project: 1000 cell types from 1000 individuals; deeply characterize epigenomes
- Atlas of nuclear architecture: genomics and imaging data; tools for researchers to use their studies
- Measure environment at the single-cell and organismal levels: high-throughput approaches; integrating/harmonizing existing datasets; genotyping of existing well-phenotyped datasets including EHRs

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

- Full human genome reference: telomere to telomere assemblies, diverse populations, haplotype resolved,
- Perturbations in pools of cells, high content, high resolution data collection: coding and non-coding elements, individually and in combinations, model driven experimental design of efficient perturbations
- Build interacting data platforms, variant portals and visualization platforms for large-scale genome processing, analysis, integration, and exploration via cloud computing.
- Build transferable gene regulatory or multi scale models from multi-layered data across biological systems.
- Resources for engineered models: cell systems, organoids, model organisms; include genetic diversity and high-resolution phenotyping
- Develop mechanistic structured and machine learning models that move from gene level to higher order: from variants to genes to networks to cell biology to predictive models
- Accessible, queryable biobanks with matching EHRs and consented tissues along with systematic phenotype extraction and mapping from EHRs in collaboration with clinical informatics experts
- Comparative genomics and epigenomics: up to 64K vertebrate genomes, end to end
- Perform next-generation human QTL study with high resolution, single cell and spatial profiles, and detailed EHR data
- Develop predictive models, including structured PRS in diverse populations, modeling genetics along with environment, multi-omics, other clinical features, and work to combine with mechanistic models
- Build benchmark datasets to compare methods, computational tools and pose challenge problems.
- Training in genome literacy for the general public, including the meaning of PRS

Appendix 9: Project Suggestions From Focus Sessions

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

- Catalog the function of all genes: perhaps based on crucial scientific questions, perhaps starting with genes that have some information may reveal more information
- An integrated view of cells (network/systems level): genomic research is typically done at the gene level
- Data annotation standards: common standards support findability, accessibility and interoperability, and is especially lacking for phenotype information
- Use cellular logic to understand how genetic variants alter phenotype at molecular and higher scales. Start with highly-integrated biology for a small number of distinct phenotypes, include modeling
- Develop approaches to determine the effects of combinations of genes: design comprehensive, systematic approaches; single cell data; random fluctuations might reveal dependencies not seen at higher levels
- Reducing complexity of the genotype to phenotype relationship: using experiments and modeling to balance choice of genes, elements, conditions, and assays to maximize understanding.
- Complex phenotypes are best studied in complex systems: For example, behavior requires more complex systems than human cells; flies and worms may not be sufficient for language and social communication
- Environmental and genetic information should be integrated: each informs the other
- Moving beyond statistics and GWAS into quasi-mechanistic genome-wide regulation.
- Developing quantitative phenotypic measurements beyond molecular assays: spatial resolution (MERFISH), cellular resolution
- Using technologies to perform gene knockouts or perturbations means you don't need to know the precise expected organismal phenotype. Reporter assays are built to see expected phenotypic results and how sequence encodes function

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

- Integrating functional genomics with individual genotypes/genomes and clinical phenotypes and matching cell lines/tissues/organisms with different assays and at different developmental stages
- Genomics and imaging research are increasingly integrated
- Start to distinguish the baseline driven by genetics to concepts related to exposure, by adding exposure-type measurements
- Interoperability comes from developing and enforcing standards, ideally before starting studies
- Machine learning and artificial intelligence on EHRs benefits both the medical and research communities.
- Investigators should be aware of projects such as the Physicians' Health Study (PHS) and the Nurses' Health Study (NHS): The valuable sequencing data could be brought to the community by asking them to share their data.
- PheWAS may be a next step for large-scale analyses and can be done using biobanks
- Continue and strengthen training, especially K08 for genome sciences
- NHGRI should continue to partner with other institutes: Existing cohorts are good resources for gene and variant discovery, other institutes could provide input on study design
- Big projects should consider an open-door policy to include potentially valuable opinions; independent investigators and consortia could both benefit from talking with each other
- Data utility depends on sharing conditions, robustness, linkage to phenotypes, and linkage to other data

Appendix 10: Bold Ideas (submitted before the workshop)

Nominator Name (Optional)	Timeframe for getting idea initiated and completed	Idea	Topic under which your idea falls under
Trey Ideker	Mid (5 years)	Projects to dramatically increase coverage of the current molecular pathway maps	Cross-Topic Ideas
Ben Voight	Mid (5 years)	Centralized repository where multiple biobanks can be queried for pheWAS rapidly, publicly, and integratively	Cross-Topic Ideas
	Mid (5 years)	Integration of high resolution structures generated by Cryo-EM with human phenotypic variation at scale	Cross-Topic Ideas
Andrea Califano	Mid (5 years)	Assembly of a multi-layer (transcriptional post-transcriptional and post-translational) model of regulation for specific tissue contexts to model the effect of genetic and epigenetic variants on human phenotypes	Cross-Topic Ideas
Andrea Califano	Mid (5 years)	Assemble a map of drug-induced perturbational profiles to recapitulate the functional role of human variants in mediating phenotypic outcomes	Cross-Topic Ideas
Andrea Califano		Creating a platform to assess the alignment of human tissue and model organism tissue on an objective basis to facilitate selection of model organisms to study the role of specific mutations in disease. This addresses a critical deficiency of current approaches due to the lack of objective criteria for the use of model organisms to elucidate the role of specific variants and mutations	Cross-Topic Ideas
John Mudgett	Short (next 18 months)	How to inform return on investment (ROI) metrics for the varied efforts under this umbrella, and promote the goals as returning on investment	Cross-Topic Ideas

John Mudgett	Short (next 18 months)	There should be a voice of customer effort to help define the value, pain points, and lessons learned as we go forward. Also, to gather some testimonials re impact of the NHGRI efforts	Cross-Topic Ideas
	Mid (5 years)	Delineate how metabolic adaptation impacts the epigenetic identity of the genome	Cross-Topic Ideas
Ting Wang	Mid (5 years)	A better understanding of sequences derived from transposable elements. They make up a large proportion of the human genome; numerous anecdotes exist supporting that many of them play critical roles in gene and genome regulation; yet their studies are much under-represented. More systematic and streamlined technology development and analysis are needed to tackle these sequences as well as their variations.	Cross-Topic Ideas
Ting Wang	Mid (5 years)	Variation and evolution of epigenomes, comparative epigenomics. Computational framework for epigenome comparison across species and individuals.	Cross-Topic Ideas
Bob Karp	Mid (5 years)	More powerful computational methods capable of identifying associations with genes of small effect size in samples as small as 100 individuals. There are many important phenotypes which are difficult to measure in larger numbers of people (e.g., complex physiological and behavioral tests, responses to controlled dietary or exercise interventions or other environmental perturbations).	Cross-Topic Ideas
Bing Ren	Long (10+ years)	A better catalog of functional elements in the human genome. Functional annotation of non-coding sequences continues to be a major challenge despite the annotation of millions of candidate cis elements. The key missing pieces include the cell type(s) each element is active in, the target gene(s) of the element, and biological function of the	Cross-Topic Ideas

		element (which TFs control its activity, how it influences target gene expression, etc).	
Bing Ren	Mid (5 years)	Functions of Transposable elements in normal biology and disease pathogenesis.	Cross-Topic Ideas
	Mid (5 years)	pipelines to test variants in simpler model organisms like Drosophila.	Cross-Topic Ideas
Mark Gerstein	Mid (5 years)	We suggest constructing a large publicly accessible database with appropriate privacy restriction that includes genotypes for individuals with a wide range of phenotypes (healthy and various diseases). This database will consist of molecular data (transcriptomics, proteomics, etc.), electronic health records and wearable activities. Such comprehensive and harmonized resource will allow researchers to share intermediate results, investigate disease mechanism and facilitate efficient publishing. Currently, this is a considerable challenge with many of the existing disease databases.	Cross-Topic Ideas
	Mid (5 years)	proof of concept studies for n=1 genomic medicine	Cross-Topic Ideas
Soumya Raychaudhuri	Mid (5 years)	Using single cell data to interpret common genetic variation	Cross-Topic Ideas
Karen Miga	Mid (5 years)	Advance epigenetic maps to include satellite DNAs and other repeat-rich region omitted from the reference genome. Data supports that these regions are bound to a myriad of transcription factors, and that their epigenetic/transcription regulation plays a role in cancer/aneuploidy, aging, and stress response. This needs to be an extension of the ENCODE project, with a focus on new methods (computational/experimental) and epigenetic targets that are specific to these regions.	Cross-Topic Ideas

Andy Clark	Short (next 18 months)	Which individuals end up in the tails of the PRS distribution? The fact that plots of PRS scores vs. mean phenotype are non-linear, with a strong arcing upward at the highest PRS scores, seems to be well established now. This result is at odds with the simple infinitesimal model and should raise all kinds of flags. Additive models flag these individuals, but their risk is underestimated. Why? Is it epistasis? Rare alleles of large effect?	Cross-Topic Ideas
Neville Sanjana	Mid (5 years)	Full transcriptome control. What is the minimal set of multiplexed genome engineering manipulations it would take to change the transcriptome of one cell to another cell? Can we computationally predict the best genes to target and experimentally validate these predictions?	Cross-Topic Ideas
Jay Shendure	Mid (5 years)	Large-scale mutagenesis of ENCODE and other cell lines (of genes, regulatory regions, etc., possibly tiling the entire genome) coupled to single cell phenotypic readouts (expression, chromatin accessibility, etc.)	Cross-Topic Ideas
Gill Bejerano	Mid (5 years)	Give computational genomics an equal seat at the table: Nothing in medical/experimental genomics makes sense, except in the light of genomic tool building. Open an NHGRI branch specializing in computational tool development. Hire POs with CS/genomics PhDs. Add a study section for genomic tool development. Develop calls where computational genomicists lead clinicians & experimentalists. Lead genomics into the 21st century.	Cross-Topic Ideas
Tuuli Lappalainen	Mid (5 years)	When likely causal proximal disease genes are identified, understanding the downstream cellular effects is a key bottleneck. Scaling up eQTL mapping to 10,000+ of individuals would answer this by mapping of trans-eQTLs and causal regulatory network effects, as well as gene-	Topic 1: Discovery and interpretation of variation associated with human health and disease

		environment interactions and rare regulatory variants. This would be a practically feasible "systems genetics" study of in vivo molecular phenotypes.	
Tuuli Lappalainen	2-3 years	eQTL mapping in specific cell types in hundreds of individuals is an obvious next step after GTEx and an extension of HCA. Unpublished work in GTEx has shown how cell type specific effects are absolutely essential for interpreting genetic regulatory effects, their tissue specificity/sharing, and interactions with e.g. age and sex, as well as improving the resolution for GWAS colocalization.	Topic 1: Discovery and interpretation of variation associated with human health and disease
Ben Voight	Mid (5 years)	Large-scale, trio-based whole genome resequencing across diverse ancestries to character rates of de novo mutation	Topic 1: Discovery and interpretation of variation associated with human health and disease
	Mid (5 years)	Biochemical and functional characterization of transposable elements in the human genome	Topic 1: Discovery and interpretation of variation associated with human health and disease
	Mid (5 years)	Quantifying the contribution of segmental duplication to phenotypic heterogeneity	Topic 1: Discovery and interpretation of variation associated with human health and disease
	Mid (5 years)	Identifying alleles of complex variation and linking to already known/existing variants (if possible)	Topic 1: Discovery and interpretation of variation associated with human health and disease

	Mid (5 years)	Identifying alleles of complex variation to discern how/if they impact phenotypes/disease. Methods need to be developed to affordably sequence these variants using long reads, which can be linked to already known variants and connected with phenotypes/disease of existing large consortia. These complex variants likely represent a significant proportion of genetic risk that is currently being overlooked in systematic genome-wide studies. To do this, new tools need to be developed (cheaper, higher throughput long read sequencing; improved bioinformatic approaches).	Topic 1: Discovery and interpretation of variation associated with human health and disease
	Mid (5 years)	A consortium that would perform and analyze saturated mutagenesis of every nucleotide (and subsequent pairwise mutagenesis) in putative reg. element and transcriptional unit across a diverse set of relevant contexts with multiple phenotypic outputs. Such a project would dramatically enhance our ability to learn models for prediction of pathogenic variants	Topic 1: Discovery and interpretation of variation associated with human health and disease
Soumya Raychaudhuri	Mid (5 years)	Defining the genetic architecture of cellular traits, fine-mapping variants, and proving causality.	Topic 1: Discovery and interpretation of variation associated with human health and disease
Karen Miga	Long (10+ years)	Satellite DNAs are known to vary considerably in the human population, yet little is known about the extent of this variability (transmission/de novo mutation rate) and the association of this variation with human disease/cellular function. New methods need to be developed to study the extent of variation in the population and across multigenerational pedigrees. Further, new tools/methods need to be developed to incorporate these novel variants into disease association/genomic medicine studies	Topic 1: Discovery and interpretation of variation associated with human health and disease

Tim Reddy	Long (10+ years)	The genetics effects on gene regulatory function. Many NHGRI genetics studies have associated non-coding variation with phenotypes; and many NHGRI studies have mapped regulatory elements in many contexts. Functional genomics studies across populations is now possible at scale, and a key opportunity to bridge genetics and genomics towards the goals of functionally connecting genotype with phenotype and disease.	Topic 1: Discovery and interpretation of variation associated with human health and disease
Tim Reddy	Mid (5 years)	Non-coding contributions to rare/Mendelian disease. Gene-sequencing efforts (many NHGRI funded) to genetically diagnose rare diseases often fail to identify gene mutations that explain Mendelian diseases. Further, such diseases have genetic modifier loci. These results indicate that Mendelian disease are more complex than previously thought. Focused studies integrating NHGRI-led rare-disease studies with genomic efforts could disruptive advances rare disease diagnosis.	Topic 1: Discovery and interpretation of variation associated with human health and disease
Gemma Carvill	Mid (5 years)	Greater diversity in population-scale sequencing, particularly in Africa building upon H3 Africa infrastructure	Topic 1: Discovery and interpretation of variation associated with human health and disease
Lynn Jorde	Mid (5 years)	Collection and analysis of family/pedigree data (allows analysis of multiple copies of rare variants in homogeneous environment).	Topic 1: Discovery and interpretation of variation associated with human health and disease
Michael Zody	Short (next 18 months)	Improved SNP imputation resources. Building on existing projects like CCDG and TOPMed, sequence ~100,000 additional genomes designed to optimize imputation panels for major ancestry groups and subgroups, with the goal of being able to impute clinically	Topic 1: Discovery and interpretation of variation associated with human health and disease

		important “rare” variants from chip or low-pass sequencing with high accuracy.	
Michael Zody	Mid (5 years)	Improved SV imputation resources, including mobile element insertion and sequence missing from the reference. Generate higher quality genomes in sufficient number to (a) determine what fraction of structural variation is imputable from SNPs and (b) build accurate imputation panels for all imputable SVs >1% (including gene copy variation).	Topic 1: Discovery and interpretation of variation associated with human health and disease
Jonathan Pritchard	Long (10+ years)	High throughput measurement of trans-regulatory networks: transcriptional networks, diverse forms of protein regulatory networks, signaling pathways etc. in many cell types. We are now at roughly the same point for trans-networks as we were for cis-regulation a decade ago—we know some general principles but very few specifics. Nonetheless, these are of central importance in connecting genetic variation to phenotypes. Dense measurement of trans networks is now tractable for the first time using emerging technologies for cellular perturbations and single cell measurements.	Topic 1: Discovery and interpretation of variation associated with human health and disease
Gill Bejerano	Mid (5 years)	Standardize genomic pathogenicity prediction: Pathogenicity prediction is a wild west. No benchmarks, no standards, no best practices. Flawed tool building, tool use, and tool comparison abound. NHGRI should standardize this field, enforce best practices, support benchmark development, and encourage friendly competitions to define, make clinically usable and improve the state of the art.	Topic 1: Discovery and interpretation of variation associated with human health and disease
John Mudgett	Mid (5 years)	Humanized mice (engineered, engrafted, and microbiome) - can there really be a human 'avatar' in our quest for translational models?	Topic 2: Addressing basic research questions that

			anticipate clinical needs
Andy Clark	all time scales	Animal models of complex trait variation – Many fundamental questions about the way that genetic variation maps into phenotypic variation are still addressed through animal studies. When human studies identify candidates, animal models can be the fastest and most convincing route to understanding mechanism. This is no time to abandon animal models!	Topic 2: Addressing basic research questions that anticipate clinical needs
Neville Sanjana	Mid (5 years)	Off-target gene editing and long-read sequencing. We must develop new (and unbiased) ways to detect off-target activity of genome editing. Beyond small indels (which has been the historic focus of the field), how can we detect larger structural variants using long-read sequencing?	Topic 2: Addressing basic research questions that anticipate clinical needs
Gill Bejerano	Mid (5 years)	Genomic privacy: Our ability to derive vast information from a person's genome grows rapidly. Genomic analysis is currently all or nothing: You often share your entire genome, to discover one or a handful relevant facts about it (e.g. Mendelian diagnosis). Cryptographic methods should be developed to exactly find these genomic nuggets without sharing the remaining 99.9999999% of the patient's genome.	Topic 2: Addressing basic research questions that anticipate clinical needs
Aravinda Chakravarti	Long (10+ years)	Although many types of networks exist, the one relevant for human genetic genotype-phenotype studies is the "Davidson" gene regulatory network (GRN) that includes DNA, RNA and protein components. They need to be cell-type specific. GRNs are modular, come in a limited set of architectures and are conserved. These networks are intrinsic to understanding which variation affects which components and how.	Topic 3: Predicting and validating functional consequences of genome variation, including beyond single variants/genes

John Mudgett	Mid (5 years)	Address the relevance and ontologies of epigenetics between translational models (murine) and human pathologies/disease states	Topic 3: Predicting and validating functional consequences of genome variation, including beyond single variants/genes
Nadav Ahituv	Long (10+ years)	Functional characterization of every nucleotide change and combination of nucleotides changes in the human genome.	Topic 3: Predicting and validating functional consequences of genome variation, including beyond single variants/genes
Nadav Ahituv	Mid (5 years)	Developing high-throughput functional characterization tools for nucleotide variants in animal models and organoids.	Topic 3: Predicting and validating functional consequences of genome variation, including beyond single variants/genes
John Mudgett	Mid (5 years)	Phenocopying genetic based human disease and pathologies in translational models - lessons learned and overcoming barriers	Topic 3: Predicting and validating functional consequences of genome variation, including beyond single variants/genes
	Mid (5 years)	For clinical whole genome sequencing to be useful, we must understand the vast noncoding genome. Though over >1,000,000 candidate regulatory elements have been biochemically annotated, few are validated	Topic 3: Predicting and validating functional consequences of

		and paired to their target genes. We propose the perturbation of every candidate regulatory elements in the human genome, followed by phenotyping the expression of every gene in all relevant cell types.	genome variation, including beyond single variants/genes
Anshul Kundaje	Mid (5 years)	Coordinated efforts to profile multiple molecular and cellular phenotypes in stimulated, perturbed conditions with temporal dynamics. Such datasets will be critical to learn causal cis and trans regulatory architecture of the cell	Topic 3: Predicting and validating functional consequences of genome variation, including beyond single variants/genes
	Long (10+ years)	Assign a phenotype to every base in the genome. This was a "challenge statement" a few years back. What happened?	Topic 3: Predicting and validating functional consequences of genome variation, including beyond single variants/genes
Katie Pollard	Long (10+ years)	Predicting the effects of mutations / genetic perturbations using cellular networks. To do so, we need better network data (genetic, physical, regulatory interactions) and novel models for how mutations propagate and interact given a network with missing data / uncertainties / errors.	Topic 3: Predicting and validating functional consequences of genome variation, including beyond single variants/genes
	Short (next 18 months)	creating a set of standards for patient-derived iPSC studies. Almost every group focused on translating genetic discoveries to better treatment options is focused on using patient-derived iPSC models. However, no guidelines in terms of numbers of biological and technical replicates (#patients, clones, differentiations) and appropriate controls exist, this will be (and is) a major limitation of	Topic 3: Predicting and validating functional consequences of genome variation, including beyond

		successful replication of studies and robust tangible results	single variants/genes
Tim Reddy	Mid (5 years)	Greatly expanding understanding of how the human genome mediates environmental responses. We know that many diseases involve both genetic and environmental effects. While several institutes support research on specific environmental exposures, NHGRI is uniquely positioned to support research on broader principles of environmental responses; and how those responses vary across genetic variation and cell type.	Topic 3: Predicting and validating functional consequences of genome variation, including beyond single variants/genes
Tim Reddy	Long (10+ years)	Technology development to make translating genetic associations into disease mechanisms routine. Recent development in high-throughput reporter assays and CRISPR-based genome/epigenome editing make it conceivable that we could be able to systematically determine how non-coding genetic variation (alone or in combination) alters gene regulation and causes diseases. NHGRI is uniquely positioned to be the leader in making this vision a reality.	Topic 3: Predicting and validating functional consequences of genome variation, including beyond single variants/genes
Michael Zody	Long (10+ years)	RNA-Seq (and maybe ATAC-Seq and other functional seq) of cells containing disease-associated mutations (native, edited, or model organism derived) to directly assay regulatory and splicing function of potential non-coding mutations for a range of different types of putatively causal non-coding mutations.	Topic 3: Predicting and validating functional consequences of genome variation, including beyond single variants/genes
Andy CLark	Long (10+ years)	Genotype x phenotype interaction – This is the area with the biggest mismatch between human and model/agricultural organism research in complex traits. GxE is universal in the latter, and quite often swamps major	Topic 3: Predicting and validating functional consequences of

		effects, and yet it gets totally insufficient concern in humans. We are badly in need of designs that accurately and at scale quantify GxE in humans.	genome variation, including beyond single variants/genes
Neville Sanjana	Long (10+ years)	Population scale genome editing. Can we understand the effect of every protein-coding rare variant on the transcriptome? I propose precise genome engineering of every rare variant in 5 cell lines from genetically diverse donors paired with a post-editing RNA-sequencing readout.	Topic 3: Predicting and validating functional consequences of genome variation, including beyond single variants/genes
Jay Shendure	Mid (5 years)	Functional measurements for ~9M potential SNVs (~0.1% of all possible SNVs) (aggregate across range of methods for mutation (e.g. MPRA, DMS, CRISPR, etc.) and phenotyping (e.g. growth, expression, protein stability, single cell assays, etc.))	Topic 3: Predicting and validating functional consequences of genome variation, including beyond single variants/genes
Gill Bejerano	Short (next 18 months)	Phenotype dbGaP: dbGaP can become much more useful if it only contained detailed (pre-diagnostic) phenotypic information per deposited patient. This rule should be enforced by NHGRI. The burden of adhering to it can be greatly alleviated by tools like ClinPhen (Deisseroth, 2018) that automate the extraction of HPO terms (non PHI information) from genetics free text. Large community gains guaranteed.	Topic 3: Predicting and validating functional consequences of genome variation, including beyond single variants/genes
	Short (next 18 months)	The field needs cost effective methods for generating long pieces of custom DNA at scale (1-2 kb or longer).	Topic 4: Data resources, methods, technologies and computational capabilities

	Long (10+ years)	"Halting the upcoming train wreck": Dealing with reproducibility, data access, and optimal use of data given that the thousands of data sets already available create tens of thousands of analyses that each use different combinations of data and slightly different methods. How do we know what is "real" and "right"? A problem only to get worse in the next 5+ years.	Topic 4: Data resources, methods, technologies and computational capabilities
	Mid (5 years)	Direct sequencing of RNA (not cDNA) technology development	Topic 4: Data resources, methods, technologies and computational capabilities
Aravinda Chakravarti	Short (next 18 months)	Although EHRs are increasingly used in genomic studies the phenotypes are used in a very bland and naïve manner. We need better methods to identify and quantify various phenotypes with assessment of their accuracy and trends. Additionally, we need better methods to extract medication and treatment data in a quantitative (dose-response dependent) manner with assessment of compliance.	Topic 4: Data resources, methods, technologies and computational capabilities
John Mudgett	Mid (5 years)	Data Integration between human and translational model efforts	Topic 4: Data resources, methods, technologies and computational capabilities
Ting Wang	Short (next 18 months)	What do we do with legacy genomic data? Centers and labs have generated and will continue to generate large data. Some of the data will become legacy. Should we keep them around? If so who's footing the bill? It is not just storage problem, but a data architecture problem.	Topic 4: Data resources, methods, technologies and computational capabilities

	Mid (5 years)	Long read, accurate, affordable DNA sequencing	Topic 4: Data resources, methods, technologies and computational capabilities
Anshul Kundaje	Mid (5 years)	We need a revolution in user interfaces and search engines for maximizing the utility of the massive bolus of genomics data	Topic 4: Data resources, methods, technologies and computational capabilities
Anshul Kundaje	Mid (5 years)	Investment in a unified public repository of predictive models for genomics (model zoos) analogous to data repositories (e.g. GEO/SRA) and publication repositories (pubmed). Ultimately these three types of repos should be interlinked. Will dramatically accelerate scientific throughput and improve reproducibility	Topic 4: Data resources, methods, technologies and computational capabilities
	Mid (5 years)	Methods to analyze and interpret structural variation	Topic 4: Data resources, methods, technologies and computational capabilities
Katie Pollard	Mid (5 years)	Discover unique genetic features (elements, motifs, domains, genes) across diseases by building and dissecting feature importance in computational models. Other institutes are unlikely to support this cross-phenotype/disease big data approach.	Topic 4: Data resources, methods, technologies and computational capabilities
Bing Ren	Short (next 18 months)	Better phenotypic using single cell genomic assays.	Topic 4: Data resources, methods, technologies and computational capabilities

Karen Miga	Mid (5 years)	Support to develop new sequencing technologies and validation methods to complete high-resolution maps of repeat-rich regions that span human peri/centromeres, subtelomeres, and acrocentric short arms	Topic 4: Data resources, methods, technologies and computational capabilities
Tim Reddy	Long (10+ years)	Developing a functional encyclopedia of gene regulatory elements. ENCODE has made major contributions by annotating the locations of regulatory elements across the human genome. Systematically determining the function of those elements (both alone and in terms of their effects on target genes) by leveraging the power/experience of an NHGRI consortium would be of immense value.	Topic 4: Data resources, methods, technologies and computational capabilities
	Long (10+ years)	Improvements in high-throughput live-cell imaging, many of the genomic based functional studies require lysing cells to capture a snapshot of cell state after perturbation of a locus/loci, improvements in live-cell imaging and novel 'read-out' approaches will provide a more dynamic view of response, this will be key for driving drug discovery and pharmacological responses over time.	Topic 4: Data resources, methods, technologies and computational capabilities
	Long (10+ years)	Improvements in high-throughput live-cell imaging, many of the genomic based functional studies require lysing cells to capture a snapshot of cell state after perturbation of a locus/loci, improvements in live-cell imaging and novel 'read-out' approaches will provide a more dynamic view of response, this will be key for driving drug discovery and pharmacological responses over time.	Topic 4: Data resources, methods, technologies and computational capabilities
Lynn Jorde	Mid (5 years)	Solutions addressing the challenges of siloed datasets -- multiple genomics, transcriptomics, proteomics, and electronic	Topic 4: Data resources, methods, technologies and

		health record datasets need to be better integrated.	computational capabilities
Lynn Jorde	Mid (5 years)	Software pipelines for large-scale genome processing and analysis via cloud computing. Essentially the “software” for the “hardware” provided by the AnVIL project.	Topic 4: Data resources, methods, technologies and computational capabilities
Michael Zody	Mid (5 years)	Multi-modal dynamic data visualization: We will see new types of data, and new needs and uses for existing data. There is a critical need for exploratory visualization tools that enable researchers to develop hypotheses; to combine data across projects; to visualize data across data types, across projects, and across time and space; and to synthesize new cohorts for further research.	Topic 4: Data resources, methods, technologies and computational capabilities
Gill Bejerano	Mid (5 years)	GenoNLP: Biomedical texts – health records, PubMed papers, textbooks - hold troves of unstructured information directly relevant to the relationship between genomic variation, function and human health. Tapping this treasure trove requires tool development that standard Computer Science Natural language processing research (of far simpler texts) will not provide. NHGRI can lead “genoNLP” into a golden age of great value.	Topic 4: Data resources, methods, technologies and computational capabilities

Appendix 11: Acknowledgements

Workshop Organizing Committee leaders:

External: Joseph Ecker, Eimear Kenny, Sharon Plon, Katherine Pollard, Jay Shendure

NHGRI: Eric Green, Carolyn Hutter, Adam Felsenfeld, Elise Feingold, Lisa Brooks, Eileen Cahill, Colin Fletcher, Allison Mandich, Elaine Ostrander, Bill Pavan, Mike Pazin, Erin Ramos, Lorjetta Schools, Mike Smith, Taylorlyn Stephan, Kris Wetterstrand

NHGRI AV: Alvaro Encinas, Kiara Palmer, Mukul Nerurkar, Ernesto del Aguila