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 continue 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 element (which TFs control its activity, how it influence target gene expression, etc.).	Cross-Topic Ideas
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 orgs 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

Nominator Name (Optional)	Timeframe for getting idea initiated and completed	Idea	Topic under which your idea falls under
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-environment interactions and rare regulatory variants. This would be a practically feasible “systems genetics” study of in vivo molecular phenotypes.	Topic 1: Discovery and interpretation of variation associated with human health and disease
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

Nominator Name (Optional)	Timeframe for getting idea initiated and completed	Idea	Topic under which your idea falls under
	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 context. 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 important “rare” variants from chip or low-pass sequencing with high accuracy.	Topic 1: Discovery and interpretation of variation associated with human health and disease
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

Nominator Name (Optional)	Timeframe for getting idea initiated and completed	Idea	Topic under which your idea falls under
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 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.	Topic 3: Predicting and validating functional consequences of 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+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 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 successful replication of studies and robust tangible results	Topic 3: Predicting and validating functional consequences of genome variation, including beyond single variants/genes

Nominator Name (Optional)	Timeframe for getting idea initiated and completed	Idea	Topic under which your idea falls under
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 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.	Topic 3: Predicting and validating functional consequences of 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 analysis 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

Nominator Name (Optional)	Timeframe for getting idea initiated and completed	Idea	Topic under which your idea falls under
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 health record datasets need to be better integrated.	Topic 4: Data resources , methods, technologies and 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

Nominator Name (Optional)	Timeframe for getting idea initiated and completed	Idea	Topic under which idea was submitted	Description	Key 1	Key 2
		NHGRI clustering loosely under Topic 1: Discovery and interpretation of variation associated with human health and disease				
		Genome "Dark Matter": SVs, TE's, Satellites				
	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	Transposable elements (human): biochem and functional char	Transposable elements	Transposable elements
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	Transposable elements: detection, function, tech dev	Transposable elements	Transposable elements
Bing Ren	Mid (5 years)	Functions of Transposable elements in normal biology and disease pathogenesis.	Cross-Topic Ideas	Transposable elements: function in disease	Transposable elements	Transposable elements
	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	SV: Seg dup contribution to phenotypic heterogeneity	Structural variants	Segmental duplications
	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	SV: Identify complex alleles and link to known variants (imputation)	Structural variants: Link complex variation to known variants	Imputation
	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	SV: Improve methods to detect, and find associations to phenotypes and link to known variants	Structural variants: detection, association, imputation	Improve methods
AF split		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	SV: General detection	Structural variants	Structural variation detection
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	Satellite DNA: Detection, transmission, mutation, contribution to phenotype	Satellite DNAs: variation, mutation, transmission, phenotype	Satellite sequences
		Large-Scale Sequencing, Variants, and Genomic Architecture				
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	Genetic architecture: cellular traits, fine mapping, link variants with genes	Variant identification: cellular traits; proving causality	Genetic architecture
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	Large-scale sequencing: diverse populations	Sequencing to add population diversity	Large-scale sequencing
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	Mendelian disease: find noncoding variation and modifiers	Noncoding variation and modifiers: Identification in Mendelian disease	Large-scale sequencing (Mendelian)
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	Large-scale sequencing: rare variants in pedigrees	Variant identification: rare variant ID and characterization	Large-scale sequencing (pedigrees)

Nominator Name (Optional)	Timeframe for getting idea initiated and completed	Idea	Topic under which idea was submitted	Description	Key 1	Key 2
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 important "rare" variants from chip or low-pass sequencing with high accuracy.	Topic 1: Discovery and interpretation of variation associated with human health and disease	SNP imputation resources; optimize imputation panels for major populations/ancestry groups. To impute rare clinically important variants from cheap data	Variant imputation resource: rare variants, population diversity	Large-scale sequencing: Variant imputation
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	SV imputation: (include mobile element insertions) improved imputation resources (multiple HQ genomes) to impute SVs >1%, and establish limits of imputation.	Variant imputation resource: structural variants	Large-scale sequencing: Structural variation
	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	Large-scale seq in trios: de novo rates	de novo mutation rates	Large-scale sequencing (trios)
		Polygenic Risk and Genomic Architecture				
	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	Cross-Topic Ideas	Polygenic risk: relation to genomic architecture of disease. What variants (rare strong? Epistasis?) underlie tails of distribution?	Polygenic risk	Genomic architecture
		NHGRI clustering loosely under Topic 2: Addressing basic research questions that anticipate clinical needs				
		Basic Genomics Anticipating Clinical Needs				
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	Linking mol and org phenotypes	Translational Models	Epigenetics
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	Creation of translational models	Translational Models	Disease
	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 successful replication of studies and robust tangible results	Topic 3: Predicting and validating functional consequences of genome variation, including beyond single variants/genes	Standardization of pt derived iPSC studies.	iPSC	Standardization
	Mid (5 years)	proof of concept studies for n=1 genomic medicine	Topic 2: Addressing basic research questions that anticipate clinical needs	Personal genomic medicine (n of 1): proof-of-concept	Personal omics	Genomic medicine
	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	Animal models	Translational models/Humanized mouse	Model orgs
	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	Animal models	Translational models/Complex trait variation	Model orgs
		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	Model org: match model tissue to human tissue to select right model for right disease	Tissue models	Model orgs

Nominator Name (Optional)	Timeframe for getting idea initiated and completed	Idea	Topic under which idea was submitted	Description	Key 1	Key 2
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	Pathogenicity prediction: standardize	Pathogenicity prediction: standardize	Misc
Neville Sanjana	Mid (5 years)	Off-target gene editing. We must develop new (and unbiased) ways to detect off-target activity of genome editing.	Topic 2: Addressing basic research questions that anticipate clinical needs	Gene editing; off-target	Gene editing	Misc
		Policy/Process				
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	Mandate phenotypic data included in dbGaP.	Policy	dbGaP
	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	Process: define ROI metrics for efforts in this area	Process	Process
	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	Process: grantees and users evaluate lessons and value. Publicize	Process	Process
	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	Data privacy	Genomic data privacy	Misc
		NHGRI clustering loosely under Topic 3: Predicting and validating functional consequences of genome variation, including beyond single variants/genes				
		Function/Phenotype of Every or Sub-Set of Nucleotides/Variants/Mutations Associated with Disease				
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	Tech Dev for FC of variants in animal models, organoids	HT Variant function	Model orgs, organoids
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	Systematic determination of impact of NC variation on gene regulation/disease	HT reporter/CRISPR assays for noncoding function	Tech Dev
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	Function of every variant/NT individually	HT SNV assays	multi-omics
	Long (10+ years)	Assign a phenotype to every base in the genome. This was a "challenge statement" a few back. What happened?	Topic 3: Predicting and validating functional consequences of genome variation, including beyond single variants/genes	Function of every nucleotide/variant	Comprehensive mutagenesis	Nucleotide/Variant function
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	Function of every variant/nucleotide individually and in combination	Comprehensive mutagenesis	Nucleotide/Variant function

Nominator Name (Optional)	Timeframe for getting idea initiated and completed	Idea	Topic under which idea was submitted	Description	Key 1	Key 2
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	Impact of protein-coding rare variants on transcriptome	Comprehensive mutagenesis of every rare coding variant	Genome engineering
	Mid (5 years)	Integration of high resolution structures generated by Cryo-EM with human phenotypic variation at scale	Cross-Topic Ideas	Protein structure: CryoEM integrated with phenotypes (coding variants?)	Variant function (coding)	Protein structure
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	Measuring impact of disease-assoc mutations.	HT variant function	multi-omic readouts
	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	HT/saturation mutagenesis: regulatory seq, with output	Variant function	HT/comprehensive mutation profiles
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	Small molecule perturbation	Variant function	Perturbation profiles
Soumya Raychaudhuri	Mid (5 years)	Using single cell data to interpret common genetic variation	Cross-Topic Ideas	Variant function: single cell	Variant function	Variant profiles: single cell
	Mid (5 years)	pipelines to test variants in simpler model orgs like Drosophila.	Cross-Topic Ideas	Variant function: Model org variant testing pipeline	Variant function	Model orgs
	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	HT/saturation mutagenesis (genes and reg regions): single cell -omics readouts	Variant function	HT mutation profiles
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	Cell-type specific eQTLs	Variant function: Cell-type specific elements	eQTLs
		Epigenetics/omics				
	Mid (5 years)	Delineate how metabolic adaptation impacts the epigenetic identity of the genome	Cross-Topic Ideas	Epigenetics: Effect of metabolic adaptation	Epigenetics	Metabolism
	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	Epigenetic maps of satellite, other repeat sequences	Epigenetics	Satellite sequences
	Mid (5 years)	Variation and evolution of epigenomes, comparative epigenomics. Computational framework for epigenome comparison across species and individuals.	Cross-Topic Ideas	Epigenomics: Comparative epi	Epigenetics	Model orgs
		Linking Variants and Regulatory Elements with 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 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.	Topic 3: Predicting and validating functional consequences of genome variation, including beyond single variants/genes	Linking regulatory elements and genes	enhancer mutagenesis	enhancer function
	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 context. 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	Connect (noncoding) disease variants with regulatory elements (genetics of gene reg); across populations	Variant function (noncoding)	Population studies
		Gene Regulatory Pathways and Networks				

Nominator Name (Optional)	Timeframe for getting idea initiated and completed	Idea	Topic under which idea was submitted	Description	Key 1	Key 2
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	Cell-type specific gene regulatory networks	Gene Regulatory Networks	Cell-type specific networks
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	Predicting impact of variants/mutations using networks; need more data to create better networks	Gene Regulatory Networks	Genetic perturbations
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+trans regulatory architecture of the cell	Topic 3: Predicting and validating functional consequences of genome variation, including beyond single variants/genes	perturbations to find causal cis and trans regulators	cis/trans regulators	Perturbation multi-omics profiles
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	Trans-regulatory networks (transcriptional, protein, signaling); single cell, perturbations	Trans-regulatory Networks	Perturbations
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-environment interactions and rare regulatory variants. This would be a practically feasible "systems genetics" study of in vivo molecular phenotypes.	Topic 1: Discovery and interpretation of variation associated with human health and disease	Regulatory networks: Trans eQTLs	Trans eQTLs	Regulatory Networks
Trey Ideker	Mid (5 years)	Projects to dramatically increase coverage of the current molecular pathway maps	Cross-Topic Ideas	Molecular pathway maps	Pathway maps	Pathway models
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	Model: integrate multi-omic tissue specific data with variant and phenotype.	Pathway maps	Pathway models
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	"Transcriptome control": engineer transcriptome to change one cell to another.	Pathway engineering	Transcriptional control
		Gene X Environment				
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	Principles of how genome/variants mediates response to environmental exposures.	GxE	GxE
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 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.	Topic 3: Predicting and validating functional consequences of genome variation, including beyond single variants/genes	Designs that accurately and at scale quantify GxE in humans.	GxE	GxE
		NHGRI clustering loosely under Topic 4: Data resources , methods, technologies and computational capabilities				
		Technology Development				
		DNA Synthesis				
	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	Tech Dev for long pieces of custom DNA	DNA synthesis	Tech Dev
		DNA/RNA Sequencing				
	Mid (5 years)	Direct sequencing of RNA (not cDNA) technology development	Topic 4: Data resources , methods, technologies and computational capabilities	Direct RNA sequencing	RNA sequencing	Tech Dev

Nominator Name (Optional)	Timeframe for getting idea initiated and completed	Idea	Topic under which idea was submitted	Description	Key 1	Key 2
	Mid (5 years)	Long read, accurate, affordable DNA sequencing	Topic 4: Data resources , methods, technologies and computational capabilities	Long read, accurate, affordable DNA sequencing	DNA sequencing	Tech Dev
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	Sequencing rich repeat regions	DNA sequencing	Tech Dev
		Structural variation				
	Mid (5 years)	Methods to analyze and interpret structural variation	Topic 4: Data resources , methods, technologies and computational capabilities	Methods to analyze and interpret structural variation	SV	Tech Dev
		Phenotyping				
Bing Ren	Short (next 18 months)	Better phenotypic using single cell genomic assays.	Topic 4: Data resources , methods, technologies and computational capabilities	Better phenotypic using single cell genomic assays.	Phenotyping	Tech Dev
	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	Improvements in high-throughput live-cell imaging	Live cell imaging	Tech Dev
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	better methods to extract phenotypic, medication and treatment information from EHRs	EHRs	Phenotypic data
		Data Sets/Resources				
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	Developing a functional encyclopedia of gene regulatory elements.	Functional element annotation	Data resource
	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	HT/saturation mutagenesis (genes and reg regions): single cell -omics readouts	Nucleotide function	Data resource
	Long (10+ years)	A better catalog of functional elements in the human genome. Functional annotation of non-coding sequences continue 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 element (which TFs control its activity, how it influence target gene expression, etc.).	Cross-Topic Ideas	Functional element annotation (non-coding): Cell-type specific, connect to target genes, TF binders of	Functional element annotation	Data resource
Ben Voight	Mid (5 years)	Centralized repository where multiple biobanks can be queried for pheWAS rapidly, publicly, and integratively	Cross-Topic Ideas	BioBank central query resource	Biobank	Data resource/Biobank
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	Large-scale data resource: biobank (seq, transcriptomics, proteomics, EHR, wearable trackers, etc.)	Biobank	Data Resource/Biobank
		Data Integration				
John Mudgett	Mid (5 years)	Data Integration between human and translational model efforts	Topic 4: Data resources , methods, technologies and computational capabilities	Data Integration between human and translational model efforts	Translational models	Data Integration
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	Improved tools for Natural Language Processing to capture relevant but unstructured information	Natural Language Processing	Computational Tool Development

Nominator Name (Optional)	Timeframe for getting idea initiated and completed	Idea	Topic under which idea was submitted	Description	Key 1	Key 2
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	Discover unique genetic features by building and dissecting feature importance in computational models	Functional element annotation	Computational modeling
		Data Quality, Accessibility and Visualization				
		Data Quality				
	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 analysis 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	Data reproducibility, access, and optimal use	Data Reproducibility	Data Utility
		Data Accessibility				
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	Software pipelines for large-scale genome processing and analysis via cloud computing	Data Analysis	Cloud computing
Lynn Jorde	Mid (5 years)	Solutions addressing the challenges of siloed datasets -- multiple genomics, transcriptomics, proteomics, and electronic health record datasets need to be better integrated.	Topic 4: Data resources , methods, technologies and computational capabilities	Integration of multi-omics and E.H.R. datasets	Data Integration	E.H.R.
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	Improved user interfaces and search engines for maximizing data utility	User Interface	Search engines
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	Long-term plan for legacy data	Data Storage	Data Architecture
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	Unified public repository of predictive models for genomics analogous to data and publication repositories	Predictive models	Data Integration
		Data Visualization				
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	Multi-modal dynamic data visualization	Data Integration	Data Visualization
		Computational Tool Dev				
	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	Computational tool development: new HG branch; dedicated study section; FOAs with CS leading experimentalists/clinicians	Computational tool dev	Process
	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	Association analysis: computational methods to improve power	Association analysis: computational tools	Comp tools