2. The biological function(s) of every human gene will be known; for non-coding elements in the human genome, such knowledge will be the rule rather than the exception.
NHGRI Strategic Vision: Compelling Genomics Research Projects in Biomedicine

- Comprehensive views of genes and regulatory elements
- Genetic architecture of human diseases and traits
- Enhancing diversity in genomics research
- Multi-omic studies in clinical settings
- Genomic learning healthcare systems

https://doi.org/10.1038/s41586-020-2817-4
Comprehensive views of genes and regulatory elements

“...an unprecedented opportunity to decipher the individual and combined roles of each gene and regulatory element. This must start with establishing the function of each human gene, including the phenotypic effects of human gene knockouts.”

https://doi.org/10.1038/s41586-020-2817-4
MorPhiC Long Term Goal

- Create a catalog of molecular and cellular phenotypes of null alleles for (essentially) all genes in human

- *In vitro* multicellular assays

- Consistency of assays

- Comprehensive with respect to genes

Phase 1 goal: 1000 genes
Phase 1 (2022) – Develop pipeline

- Start with 1000 genes
- Can high-quality data be produced at scale?
- Prospects for throughput and cost improvement
- ID and address cost/throughput/technical barriers
- Identify scientific challenges
- Produce initial high-quality data for analysis

Contingent on lessons of Phase 1, to be evaluated prior to a new Concept
MorPhiC

• Why null alleles?

• Why molecular and cellular assays on multicellular systems?
Benefits of Catalog

• Provide basic, consistent cellular/molecular information for all genes

• Fill gap between proximate molecular phenotypes, and anatomical/physiological phenotypes (from human disease, KOMP)

• Foothold to mechanism- scalable; Inform pathways
More Benefits…

• Combine with data from other “variant X function” and association studies

• Molecular and cellular phenotypes may be quantitative

• Good substrate for computation, including machine learning

• Other deliverables: cell lines, tools.
Challenges

• Can mutagenesis scale? Informative HT assays?

• Cell type specificity – assays have to capture phenotypes “enough”. Will not assay all relevant tissues for many genes

• Extreme pleiotropy/cell lethals may hinder interpretability

• How to deal with genetic background effects – diverse samples
Phase 1 Scope - Overview

Phase 1 to develop a pipeline, assess technical and analytical barriers to scale, address “Challenges”. To inform a potential Phase 2.


- Test multiple approaches for mutagenesis and assays. Strongly prefer multicellular systems (e.g. organoids)
Phase 1 Scope

- Develop standards/QC (eg, for mutation QC; compare assays, data formats, etc.)
- Diverse samples to understand genetic background effects
- Use data in analyses to inform production (best data types, data formats, applications, etc.)
- Develop data infrastructure
Phase 1 (2022) – Develop pipeline

- Start with 1000 genes
- Can high-quality data be produced at scale?
- Prospects for throughput and cost improvement
- ID and address cost/throughput/technical barriers
- Identify scientific challenges
- Produce initial high-quality data for analysis

• Get samples
• Engineer alleles
• Test and compare assays
• Characterize challenges (e.g. lethals, no phenotype, background effects)
• Data standards/dissemination/integration
• Analyses to develop use-cases and deliverables

Phase 2 (2026)

Contingent on lessons of Phase 1, to be evaluated prior to a new Concept
MorPhiC Structure

I. Data Production Centers: System and Assay Development (4-6 centers, up to $1.4M each)

• Choose Phase 1 genes/criteria; overlap; QC standards

• Test systems/creating KO (multiple, compare, variety of tissues)

• Test assays (multiple, overlap, compare, replicability)

• Data standards

• Work with other components on comparison analysis.
II. Data Analysis and Validation (2-3 awards, up to $0.5M each)

• Propose analyses that raise/address key scientific issues (e.g. imputation, pleiotropy, networks, cell type inference, etc.)

• Reveal data and design issues (validations, statistical analyses, utility for different uses)

• Test integration with other functional data sets

• Identify community resource deliverables
III. Data Resource (1 award, up to $1.5M/year, starting smaller and ramping)

• Receive, wrangle, annotate, present data for consortium and community use

• Lead data formats discussion

• Integrate data from Data Production Centers

• Enable community use

• Pursue opportunities to integrate with similar/complementary resources; work towards API compatibility
Relationship to other projects

- All “variant/perturbation X molecular function studies” (e.g. IGVF, GTEx/dGTEx)
- Disease association studies (Mendelian/MGRC, common disease)
- Clinical studies/resources that interpret variants (ClinGen, UDN)
- KO studies (KOMP, gnomAD)

*MorPhiC data should be aggregated/integrated with data from other “variant/perturbation X function” studies*
Summary

Long term goal is catalog of molecular and cellular phenotypes of null alleles for all human genes *in vitro*

Phase 1 to develop a pipeline, assess barriers to scale, address challenges. 1000 genes. 4 years.

Structure

- 4-6 Data Production centers (UM1; $7M/year total)
- Three Analysis Centers (U01; $1.5M/year total)
- Data Resource (U24; ramp to $1.5M/year total)
Thanks to many colleagues for extensive input on ideas and presentation

Ajay Pillai
Colin Fletcher
Stephanie Morris
Carolyn Hutter

Mike Pazin
Larry Brody
Lisa Brooks
Lisa Chadwick
Elise Feingold
Dan Gilchrist
Jen Troyer
Heidi Sofia
Questions?
MorPhiC Data Flow

Prod → Data Resource → Community
Prod → Data Resource → Analysis Centers
Prod → Data Resource → Community
Prod → Data Resource → Analysis Centers
<table>
<thead>
<tr>
<th>Project</th>
<th>Allele types</th>
<th>Number of genes</th>
<th>Number of alleles/gene</th>
<th>Molec. phe</th>
<th>Cell. phe</th>
<th>Multicellular in vitro assays (eg organoids)</th>
<th>Organismal phe</th>
<th>Human disease phe</th>
</tr>
</thead>
<tbody>
<tr>
<td>ENCODE</td>
<td>Existing in cell lines</td>
<td>Genome-wide (not just genes)</td>
<td>Mostly NA</td>
<td>Yes, rich</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>GTEx</td>
<td>Existing in human</td>
<td>Undefined</td>
<td>One</td>
<td>RNAseq downstream</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>RVAS (Human Mendelian and complex)</td>
<td>Existing in human</td>
<td>Eventually all as long-term goal for Mendelians</td>
<td>One-few per gene, eventually (esp. Mendelians)</td>
<td>No (only as follow-up)</td>
<td>No</td>
<td>No</td>
<td>Yes, often/usually = disease phenotype</td>
<td>Yes</td>
</tr>
<tr>
<td>IGVF FOA</td>
<td>Coding and noncoding, as proposed</td>
<td>Not all; may be sparse wrt. genes</td>
<td>Potentially many</td>
<td>cis and downstream possible. Probably very rich.</td>
<td>Possible</td>
<td>Possible</td>
<td>Possible</td>
<td>No</td>
</tr>
<tr>
<td>KOMP</td>
<td>KO’s</td>
<td>All</td>
<td>1-few/gene</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes, mouse</td>
<td>No</td>
</tr>
<tr>
<td>MorPhiC</td>
<td>KOs</td>
<td>All</td>
<td>One</td>
<td>Yes, rich</td>
<td>Yes</td>
<td>Yes for many</td>
<td>No</td>
<td>No</td>
</tr>
</tbody>
</table>
Phase 1 (2022)
- 1000 protein coding genes
- Can high-quality data be produced at scale?

Phase 2 (2026)
- Contingent on Phase 1
- to be evaluated prior to a new Concept

Phase 1: Develop Pipeline

Phase 2: Scale-up Production

Catalog of All Genes