Molecular Phenotypes of Null Alleles in Cells (MorPhiC)

Adam Felsenfeld, Colin Fletcher, Ajay Pillai, Stephanie Morris

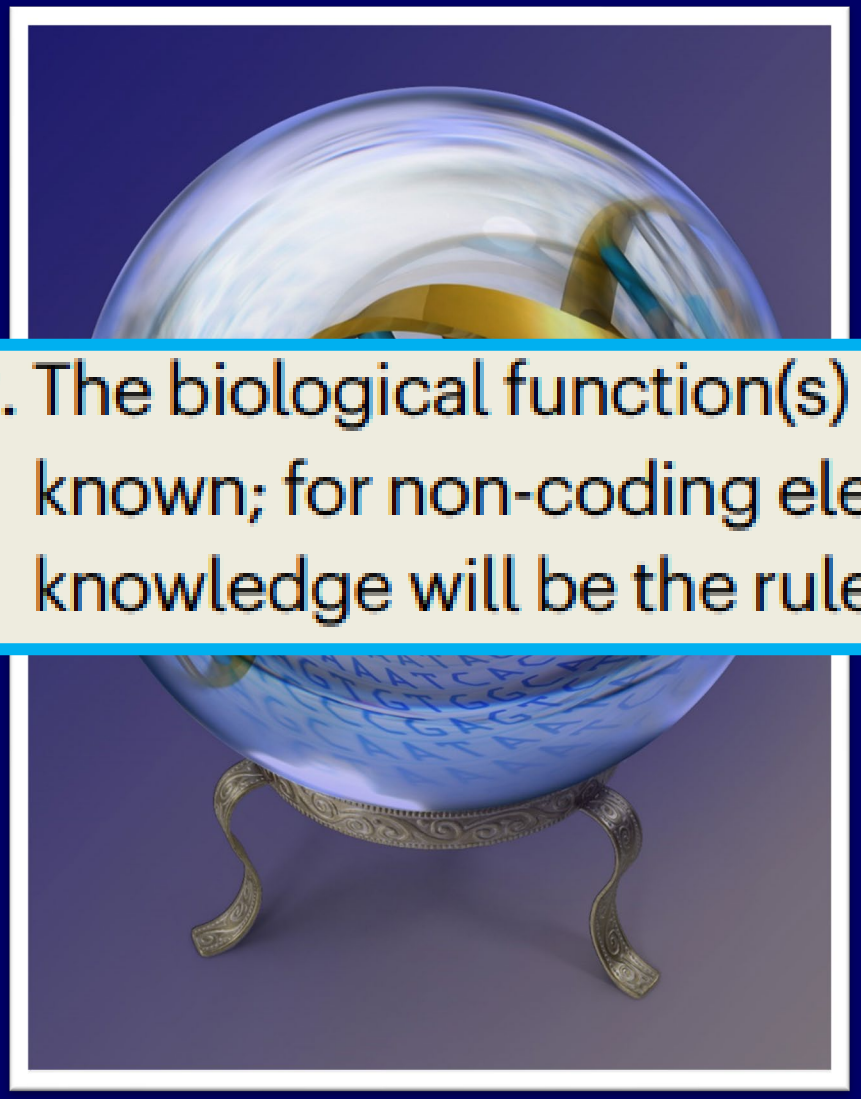
Concept Clearance

National Council for human Genome Research, February 22, 2021

National Human Genome Research Institute

The Forefront of Genomics

NHGRI Bold Predictions



Bold predictions for human genomics by 2030

Some of the most impressive genomics achievements, when viewed in retrospect, could hardly have been imagined ten years earlier. Here are ten bold predictions for human genomics that might come true by 2030. Although most are unlikely to be fully attained, achieving one or more of these would require individuals to strive for something that currently seems out of reach. These predictions were crafted to be both inspirational and aspirational in nature, provoking discussions about what might be possible at The Forefront of Genomics in the coming decade.

1. Generating and analysing a complete human genome sequence will be routine for any research laboratory, becoming

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.

associated phenotypic information for millions of human participants will be regularly featured at school science fairs.

6. The regular use of genomic information will have transitioned from boutique to mainstream in all clinical settings, making genomic testing as routine as complete blood counts.
7. The clinical relevance of all encountered genomic variants will be readily predictable, rendering the diagnostic designation 'variant of uncertain significance (VUS)' obsolete.
8. An individual's complete genome sequence along with informative annotations will, if desired, be securely and readily accessible on their smartphone.
9. Individuals from ancestrally diverse backgrounds will benefit equitably from advances in human genomics.
10. Breakthrough discoveries will lead to curative therapies involving genomic modifications for dozens of genetic diseases.

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

NHGRI Strategic Vision: Compelling Genomics Research Projects in Biomedicine

- **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

NIH NHGRI

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

Phase 2 (2026)

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

NIH NHGRI

MorPhiC

- **Why null alleles?**
- **Why molecular and cellular assays on multicellular systems?**

NIH NHGRI

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.

- **1000 protein coding genes. Priorities: No known function? Disease genes? Arbitrary? KO mouse exists?**

- **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.

MorPhiC Structure

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

MorPhiC Structure

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**

NIH NHGRI

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

NIH NHGRI

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)

NIH NHGRI

Thanks to many colleagues for extensive input on ideas and presentation

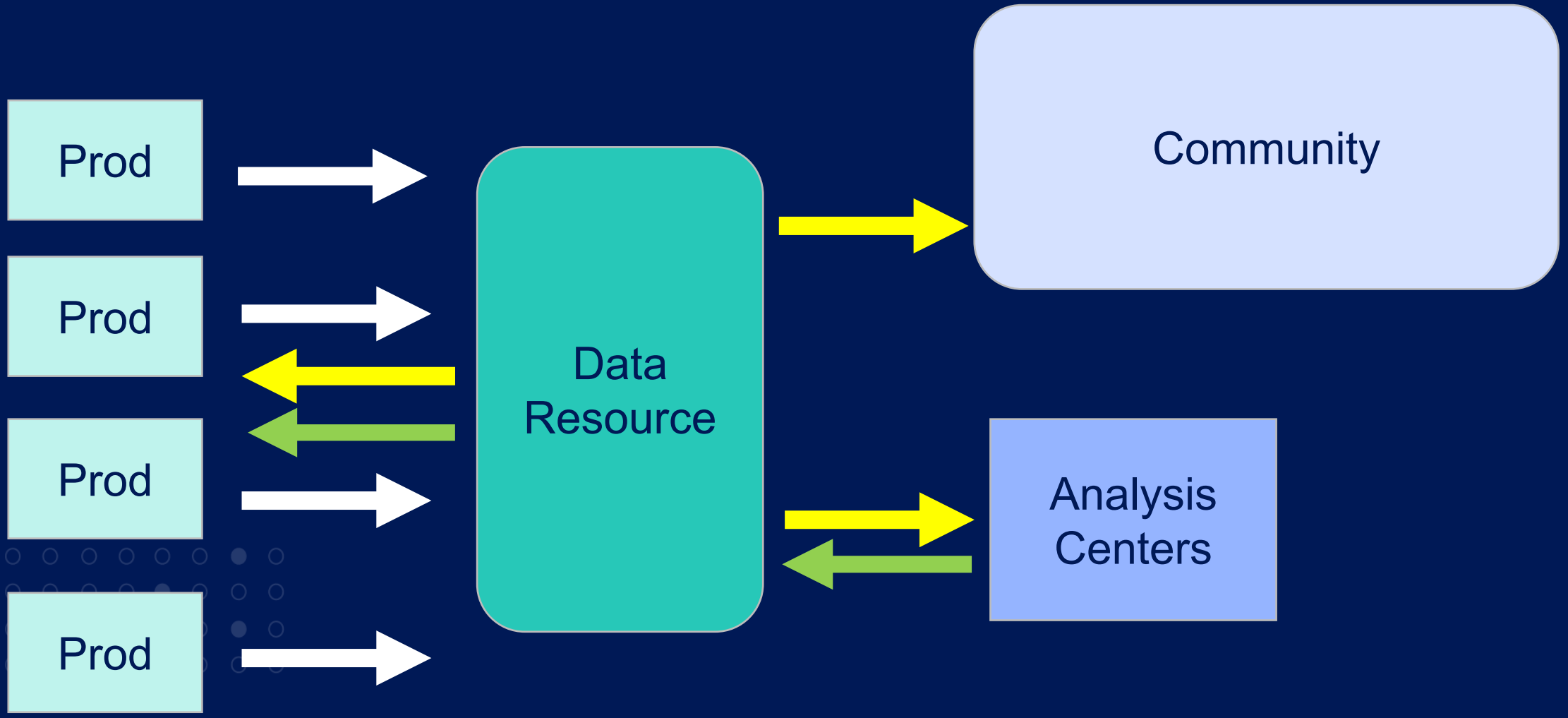
Ajay Pillai
Colin Fletcher
Stephanie Morris
Carolyn Hutter

Mike Pazin
Larry Brody
Lisa Brooks
Lisa Chadwick
Elise Feingold
Dan Gilchrist
Jen Troyer
Heidi Sofia

NIH
NHGRI

Questions?

MorPhiC Data Flow



NIH NHGRI

Project	Allele types	Number of genes	Number of alleles/gene	Molec. phe	Cell. phe	Multicellular in vitro assays (eg organoids)	Organismal phe	Human disease phe
ENCODE	Existing in cell lines	Genome-wide (not just genes)	Mostly NA	Yes, rich	No	No	No	No
GTEx	Existing in human	Undefined	One	RNAseq downstream	No	No	No	No
RVAS (Human Mendelian and complex)	Existing in human	Eventually all as long-term goal for Mendelians	One-few per gene, eventually (esp. Mendelians)	No (only as follow-up)	No	No	Yes, often/usually = disease phenotype	Yes
IGVF FOA	Coding and noncoding, as proposed	Not all; may be sparse wrt. genes	Potentially many	cis and downstream possible. Probably very rich.	Possible	Possible	Possible	No
KOMP	KO's	All	1-few/gene	No	No	No	Yes, mouse	No
MorPhiC	KOs	All	One	Yes, rich	Yes	Yes for many	No	No

Phase 1 (2022)

-1000 protein coding genes
- Can high-quality data be produced at scale?

Phase 1: Develop Pipeline

Phase 2: Scale-up Production

Catalog of All Genes

Phase 2 (2026)

-Contingent on Phase 1
- to be evaluated prior to a new Concept

NIH NHGRI