Experimental Data Subgroup

Len Pennacchio
Jay Shendure
John Stamatoyannopoulos
Wendy Winckler
Experimental Data Subgroup

- **Goal:** evaluate methods by which investigators can query whether candidate variants have a biological effect
Motivation for Functional Analysis

- **GWAS** peak $\rightarrow$ causal variant(s)

- Clinical genetics $\rightarrow$ Functional consequences of variants of unknown significance

- “Functionalizing” poorly characterized genes of interest, *i.e.* developing functional assays

Need generic, accessible, high-throughput methods and resources to facilitate the functional analysis of both coding and regulatory variation.
Challenges

• Spectrum of experimental methods exist
• How to select the most appropriate method?
  – Type of variant
  – Context
  – Access to samples, reagents, and techniques
  – Throughput, time, and cost
  – Weight of evidence
• Caution required in interpretation—for both negative and positive results
## Selected Experimental Methods

<table>
<thead>
<tr>
<th>Techniques</th>
<th>Variant class</th>
<th>Experiment type</th>
<th>High-throughput?</th>
<th>Evidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>mouse or zebrafish knockin</td>
<td>Any</td>
<td>in vivo</td>
<td>no</td>
<td>strong</td>
</tr>
<tr>
<td>genome editing</td>
<td>Any</td>
<td>in vivo, in vitro</td>
<td>yes</td>
<td>strong</td>
</tr>
<tr>
<td>mouse or zebrafish knockout</td>
<td>LoF allele</td>
<td>in vivo</td>
<td>no</td>
<td>strong</td>
</tr>
<tr>
<td>cell culture shRNA knockdown</td>
<td>LoF allele</td>
<td>in vitro</td>
<td>yes</td>
<td>suggestive</td>
</tr>
<tr>
<td>cDNA complementation</td>
<td>LoF allele</td>
<td>in vitro</td>
<td>no</td>
<td>suggestive</td>
</tr>
<tr>
<td>splicing assay</td>
<td>Splicing</td>
<td>ex vivo, in vitro, in vivo</td>
<td>yes</td>
<td>strong</td>
</tr>
<tr>
<td>protein-specific biochemical or cellular assays</td>
<td>Protein-altering alleles</td>
<td>in vitro</td>
<td>no</td>
<td>suggestive</td>
</tr>
<tr>
<td>correlation with expression</td>
<td>Regulatory</td>
<td>ex vivo, in vitro</td>
<td>yes</td>
<td>suggestive</td>
</tr>
<tr>
<td>reporter construct</td>
<td>Regulatory</td>
<td>in vitro</td>
<td>yes</td>
<td>suggestive</td>
</tr>
</tbody>
</table>
Regulatory variants
Functional regulatory variation: Levels of evidence framework

Level 1: *in vivo* evidence from *in situ* models

1a *In situ* / whole locus model of strongly genetically implicated variant that precisely recapitulates the phenotype at the organismal level

1b *In situ* genome modification (e.g., genome editing / knock-in/out)

1c Whole-locus transgenic lines (e.g., YAC, BAC; single copy)

1d *In situ* measurement of gain/loss of regulatory protein binding directly coupled to *in vivo* gene product phenotype

1e *In situ* gain/loss of regulatory protein not coupled to gene product

Level 2: Evidence from artificial/condensed construct models

2a Standard transgenic animal

2b Stable transfection (integrated into genome)

2c Transient transfection (ex-genomic)

Level 3: Non-cellular assays (e.g., gel shifts)
Example: Level 1a

Hereditary Persistence of Fetal Hemoglobin (HPFH)
Scientific trajectory of variant characterization

1985

1992

1995
A point mutation in the \( ^\gamma \)-globin gene promoter in Greek hereditary persistence of fetal haemoglobin

Francis S. Collins*, James E. Mecherlli, Minoru Yamakawa, Julian Pan, Sherman M. Weissman & Bernard G. Forget†

Departments of Human Genetics, Internal Medicine, and Molecular Biophysics and Biochemistry, Yale University School of Medicine, 333 Cedar Street, New Haven, Connecticut 06510, USA

G to A substitution in the distal CCAAT box of the \( ^\gamma \)-globin gene in Greek hereditary persistence of fetal haemoglobin

Richard Gelinas*, Brian Endlich*, Carla Pfeiffer*, Mayumi Yagi† & George Stamatoyannopoulos†
A single point mutation is the cause of the Greek form of hereditary persistence of fetal haemoglobin

Meera Berry, Frank Grosveld & Niall Dillon

Received 27 May accepted 26 June 1992.


ACKNOWLEDGMENTS: This paper is dedicated to the memory of B.M. We thank G. Myers and K. Marckes for assistance with phylogenetic analysis of the Irish National Centre for Bioinformatics for their facilities. J. Howe for independent attempts at cultivating blood samples from animal 2358; Seromorphics, Inc. (Atlanta, GA) for blood specimens from subject 73314; R. Kestebren, P. Kutz and D. Ho for discussions; D. Decker and M. Vison for technical assistance; and C. Dowes and A. J. Nathanson for manuscript preparation. This work was supported by grants from the NIH, the US Army Medical Research Acquisition Activity, the Life and Health Insurance Medical Research Fund, and the Birmingham Center for AIDS Research. Q.M.S. is a REW Scholar in the Biomedical Sciences.

In normal humans the fetal stage-specific γ-globin genes are silenced after birth and not expressed in the adult. Exceptions are seen in cases of hereditary persistence of fetal haemoglobin (HPFH). These are clinically important because the elevated levels to establish a large number of bred lines. When the wild-type γβ minilocus was introduced into fertilized mouse eggs, five transgenic mice were obtained. Southern blots showed that two of the founders were mosaic (31 and 36) and that all contained the intact γδ minilocus, albeit at different copy numbers (Table 1, and data not shown). S1 nuclease protection analysis showed that the γ-globin gene expression was suppressed in adult mice (Fig. 1a, b). In contrast, the human β-globin gene was expressed at this stage at levels comparable to those observed for the mouse β-major-globin genes5 (Fig. 1b; Table 1). The suppression of the wild-type γ-globin gene is in agreement with results obtained when a minilocus containing only the γ-globin gene is introduced into mice6. Repeated phlebotomy increases the number of reticulocytes, but even under those conditions the γ-globin gene remains suppressed (Fig. 1b). When the −117 mutant γβ minilocus was introduced into mice, nine transgenic mice were obtained and Southern blots showed that they con-
Use of yeast artificial chromosomes (YACs) in studies of mammalian development: Production of \( \beta \)-globin locus YAC mice carrying human globin developmental mutants

(developmental regulation/transgenic mice/hereditary persistence of fetal hemoglobin/\( \delta \beta \)-thalassemia)

KENNETH R. PETERSON*,†, QI LIANG LI*, CHRISTOPHER H. CLEGG*,‡, TATSUO FURUKAWA*, PATRICK A. NAVAS*, ELIZABETH J. NORTON*, TYLER G. KIMBROUGH*, and GEORGE STAMATOYANNOPoulos*§

*Division of Medical Genetics, Department of Medicine, University of Washington, Seattle, WA 98195; †Bristol-Myers Squibb Pharmaceutical Research Institute, Princeton, NJ 08543; ‡Department of Genetics, University of Washington, Seattle, WA 98195

Communicated by Stanley M. Garrity, University of Washington, Seattle, WA, March 8, 1995

**ABSTRACT** To test whether yeast artificial chromosomes (YACs) can be used in the investigation of mammalian development, we analyzed the phenotypes of transgenic mice carrying two types of \( \beta \)-globin locus YAC developmental mutants: (i) mice carrying a \( G \rightarrow A \) transition at position −117 of the \( A_{\gamma} \) gene which is responsible for the \( G \) form of developmental regulation of gene expression in transgenic mice (1). Our data show that the genes of the \( \beta \)-globin locus YAC (\( \beta \)-YAC) are correctly regulated during development in the mouse (1), thus demonstrating the usefulness of the YAC/transgenic mouse system.

In this work, we test whether YACs can be used for the
Scientific trajectory of variant characterization

1985

1992

1995

Structure of a human β-globin locus YAC
A single point mutation in a 273kb single copy YAC, functionally profiled across development
Example: Level 1d
Alpha thalassemia

A Regulatory SNP Causes a Human Genetic Disease by Creating a New Transcriptional Promoter


We describe a pathogenic mechanism underlying a variant form of the inherited blood disorder α-thalassemia. Association studies of affected individuals from Melanesia localized the disease trait to the telomeric region of human chromosome 16, which includes the α-globin gene cluster, but no molecular defects were detected by conventional approaches. After resequencing and using a combination of chromatin immunoprecipitation and expression analysis on a tiled oligonucleotide array, we identified a gain-of-function regulatory single-nucleotide polymorphism (rSNP) in a non-genic region between the α-globin genes and their upstream regulatory elements. The rSNP creates a new promoter-like element that interferes with normal activation of all downstream α-like globin genes. Thus, our work illustrates a strategy for distinguishing between neutral and functionally important rSNPs, and it also identifies a pathogenetic mechanism that could potentially underlie other genetic diseases.

The human α-globin cluster, located at the telomeric region of chromosome 16 (16p13.3), includes an embryonic gene (ζ), two minor α-like genes (δ and θ), two α genes (α2 and α1), and two pseudogenes (αψ1 and αψ2) (1, 2). Previously described molecular defects could be found. The pattern of inheritance suggested that individuals with HbH disease are homozygotes for a codominant defect, referred to here as (αα)T, causing α-thalassemia with a predicted genotype of (αα)T/(αα)T (table S1). Linkage to a variable number of tandem repeats (VNTR) (6) located ~8.5 kb from the α-globin genes (Fig. 1), we found that all individuals with the (αα)T mutation shared a common VNTR allele (fig. S1), demonstrating that this is a cis-linked defect. Further association studies, using known SNPs, showed that the (αα)T haplotype extends from the 16p telomere, with loss of association immediately downstream of the α-globin cluster (coordinate 168,467 in Fig. 1) defining the centromeric border of the region containing the cis-acting mutation. We estimated that the frequency of the (αα)T defect in the island population is ~0.04 (fig. S1).

We therefore resequenced the (αα)T haplotype by isolating bacterial artificial chromosomes (BACs) from a library constructed from the peripheral blood DNA of patient L with the Melanesian type of HbH disease [(αα)T/(αα)T]. BACs spanning the α-globin cluster and the surrounding ~213 kb of DNA (coordinates 21,059 to 234,236) were sequenced (DQ431198), and we identified 283 SNPs and/or sequence differences (Fig. 1) by comparison with the current wild-type sequence (National Center for Biotechnology Information database build 35, coordinates 1 to 223878), consistent with estimates of the frequency of SNPs throughout the genome (7). This now presented a sit-
Example 2:
Mouse Knock-ins
Mouse Site-Specific Integration
**KNOCK-IN MOUSE STUDIES:** Introduce Human Mutation into Mouse Gene

**Examples**

**CAG Repeat Expansions Introduced into:**

1) **Huntingtin gene- Short Repeat: Nuclear Inclusion Body Formation in Striatal Neurons**


2) **Huntingtin gene- Long Repeat: Neurological Abnormalities**


3) **Spinocerebellar Ataxia Type 1 Gene- Motor Coordination Defects**


**Point Mutation Introduced into:**

1) **Presenilin-1 Gene- Single Amino Acid Change Causes Hippocampus Neuron Sensitivities**

Reproducible Association between Human APOA5 Common Variation and Plasma Triglyceride Levels

Do these Haplotypes Affect APOA5 Gene Product Levels in Vivo?

Pennacchio et al, 2001 Science
Generation of Site-Specific Single-Integrant Haplotype Transgenes

Precise docking at HPRT

Common

Noncoding, 5UTR (Kozak) (7 change)

S19W (putative signal peptide) (sole change)

Compare APOA5:
- mRNA Levels in Liver
- Protein Levels in Plasma
Generation of Site-Specific Single-Integrant Haplotype Transgenes

APOA5 HAP1

APOA5 HAP2

APOA5 HAP3

Common

Noncoding, 5'UTR (Kozak) (7 change)

S19W (putative signal peptide) (sole change)

Liver mRNA

Liver mRNA

Liver mRNA

Plasma Protein

Plasma Protein

Plasma Protein

70% Decrease

S19W is likely responsible for TG Association

Example 3:
9p21 and Coronary Artery Disease

9p21 common risk variant:
increases CAD risk by 30%
>20% of population homozygous

(2007, J Cohen and K Stefansson Labs)

Dysregulation of Cdkn Genes in the Heart

Increased Aortic Smooth Muscle Cell proliferation

Supports CAD risk interval harbors distant-acting regulatory function(s)

Visel et al. Nature 2010
In Vitro Studies: Empowered by ENCODE

One Third of Variants found in this Element
Several Effect Putative STAT1 Binding Site

This Element:
1) binds STAT1 in vitro
2) binding lowers CDKN2B RNA in vitro
3) STAT1 occupancy is less in Risk CAD LCL

9p21 DNA variants associated with Coronary Artery Disease impair IFNγ signaling response
Kelly Frazer et al. Nature 2011
Future of Experimental Data

• 1000+ GWAS peaks $\rightarrow$ causal variant(s)

• Clinical genetics $\rightarrow$ Functional consequences of variants of unknown significance

• Facilitating genetics $\rightarrow$ biology

• High-throughput or massively parallel methods for assessing the functional consequences of observed and potential variation
Functional assessment of observed regulatory variation in situ allelic occupancy
Conservation is an imperfect guide to regulatory function.
Massively parallel functional assessment of potential regulatory variants

- Tail vein injection
- RNA from liver
- PCR and sequencing of transcribed tags
- Complex library of tagged enhancer haplotypes

Activity Profile
• All possible mutations assayed in one experiment
• Distribution of effect sizes for regulatory mutations (i.e. establishing null distribution)
Key points

• Experimental data can be very useful!
  – Identifying causal gene / variant(s)
  – Variants of (unknown $\rightarrow$ known) significance
  – Genetics $\rightarrow$ biological understanding

• Subjective exercise: no experiment is perfect
  – Demonstrating experimental effect $\neq$ causation
  – Failure to show effect $\neq$ non-causation
  – Multiple lines of evidence better

• Need for more high-throughput approaches
Discussion Questions

1. Feedback on accuracy, completeness and organization of experimental methods table?

2. Feedback on proposed levels of evidence?

1. How should experimental data be weighted relative to genetic analysis?

2. How should editors and reviewers be guided to think about experimental data in the context of manuscripts?
## Selected Experimental Methods

<table>
<thead>
<tr>
<th>Techniques</th>
<th>Variant class</th>
<th>Experiment type</th>
<th>High-throughput?</th>
<th>Evidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>mouse or zebrafish knockin</td>
<td>Any</td>
<td>in vivo</td>
<td>no</td>
<td>strong</td>
</tr>
<tr>
<td>genome editing</td>
<td>Any</td>
<td>in vivo, in vitro</td>
<td>yes</td>
<td>strong</td>
</tr>
<tr>
<td>mouse or zebrafish knockout</td>
<td>LoF allele</td>
<td>in vivo</td>
<td>no</td>
<td>strong</td>
</tr>
<tr>
<td>cell culture shRNA knockdown</td>
<td>LoF allele</td>
<td>in vitro</td>
<td>yes</td>
<td>suggestive</td>
</tr>
<tr>
<td>cDNA complementation</td>
<td>LoF allele</td>
<td>in vitro</td>
<td>no</td>
<td>suggestive</td>
</tr>
<tr>
<td>splicing assay</td>
<td>Splicing</td>
<td>ex vivo, in vitro, in vivo</td>
<td>yes</td>
<td>strong</td>
</tr>
<tr>
<td>protein-specific biochemical or</td>
<td>Protein-altering alleles</td>
<td>in vitro</td>
<td>no</td>
<td>suggestive</td>
</tr>
<tr>
<td>cellular assays</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>correlation with expression</td>
<td>Regulatory</td>
<td>ex vivo, in vitro, in vivo</td>
<td>yes</td>
<td>suggestive</td>
</tr>
<tr>
<td>reporter construct</td>
<td>Regulatory</td>
<td>in vitro</td>
<td>yes</td>
<td>suggestive</td>
</tr>
</tbody>
</table>
Functional regulatory variation: Levels of evidence framework

**Level 1: in vivo evidence from in situ models**

1a *In situ* / whole locus model of strongly genetically implicated variant that precisely recapitulates the phenotype at the organismal level

1b *In situ* genome modification (e.g., genome editing / knock-in/out)

1c Whole-locus transgenic lines (e.g., YAC, BAC; single copy)

1d *In situ* measurement of gain/loss of regulatory protein binding directly coupled to *in vivo* gene product phenotype

1e *In situ* gain/loss of regulatory protein not coupled to gene product

**Level 2: Evidence from artificial/condensed construct models**

2a Standard transgenic animal

2b Stable transfection (integrated into genome)

2c Transient transfection (ex-genomic)

**Level 3: Non-cellular assays** (e.g., gel shifts)