Progress in human genetics: GWAS and beyond

Mark Daly
Chief, Analytic and Translational Genetics Unit
Massachusetts General Hospital & Senior Associate Member Broad Institute

Image courtesy of Bang Wong

Genetic findings are the first step... where do we need to go to advance the science of therapeutics?

Identify target  Chemical screening  Animal models  Clinical trials

More often discussed... but less of a primary focus

Genetics can in some cases provide individualized medical insights
- diagnostics in cases of severe genetic disorders
- identification of individuals more or less likely to benefit from specific therapeutic interventions
- prediction of individuals at risk for severe idiosyncratic adverse drug responses

Vastly oversold for years - technological advances are bringing what has long been a futuristic vision much closer to reality

The scope of the challenge:
Within each cell:
2 copies of the genome
23 chromosomes
~20,000 genes
3.2B letters of DNA
Millions of polymorphic sites

Where are the differences that contribute to heritable disease?

"The molecular targets of all of today’s approved psychiatric drugs are the same as the targets of their pre-1960 prototypes (Table 2), and their mechanisms of action are not understood beyond a few initial molecular events."

<table>
<thead>
<tr>
<th>Drug class</th>
<th>Prototype compound</th>
<th>Molecular target(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mood stabilizer</td>
<td>Lithium (Li)</td>
<td>GSK3β, ionotrol 1-phosphatase*</td>
</tr>
<tr>
<td>Antipsychotic drugs</td>
<td>Chlorpromazine</td>
<td>Dopamine D, receptor</td>
</tr>
<tr>
<td>Antidepressants</td>
<td>Imipramine, Imipramine</td>
<td>Monoamine oxidase, NA, and 5-HT transporters</td>
</tr>
<tr>
<td>Benzodiazepine receptor agonists</td>
<td>Chlordiazepoxide</td>
<td>GABA receptor, benzodiazepine site</td>
</tr>
</tbody>
</table>

*Although much research favors GSK3β glycogen synthase kinase 3 as the relevant target of Li, the drug mechanism of action remains uncertain.
The origins: Mendel

- Mendel first recognized the discrete units of inheritance and that variation could in these units was transmissible and resulted in phenotypic differences.

9 Round-Yellow, 3 Round-green, 3 wrinkled-Yellow, 1 Wrinkled-green

The first challenge: LINKAGE

- Some pairs of phenotypes were not passed on independently – violating Mendel’s second rule.
- WHY?

Chromosomal Linkage

- Genes on the same chromosome are passed along in tandem unless meiotic cross-over occurs.
- Here, the genes of interest are separated by 3 cM, indicating about a 3% chance of recombination during meiosis.

Linkage mapping (1913)

I suddenly realized that the variations in strength of linkage, already attributed by Morgan to differences in the spatial separation of genes, offered the possibility of determining sequences in the linear dimension of a chromosome. I went home and spent most of the night (to the neglect of my undergraduate homework) in producing the first chromosome map, which included the sex linked genes y, w, v, m, and r, in the order and approximately the same relative spacing that they still appear on the standard maps.

— Alfred Sturtevant “A History of Genetics”

‘Mendelian’ diseases travel predictably and consistently in families

Dominant transmission

Family-based linkage analysis

Saw dramatic successes in the 1980s-90s for the localization of genes underlying countless Mendelian disorders: Huntington’s, CF, DMD, early onset forms of breast cancer, Alzheimer’s, diabetes…

Thousands of diseases or traits caused by mutations in a single gene (e.g., Huntington’s, CF, muscular dystrophy)
Mendelian disease genetics

- genotype → disease state

Linkage analysis and positional cloning are powerful because genetic risk factors are highly “penetrant”

Despite the fact that common diseases, like their rare counterparts, are highly heritable...

Sullivan, Daly, O’Donovan 2012

for most of our careers, human genetics has brought little to the table...

Explanations not difficult to come by

Many genes vs. 1
Incomplete Penetration

We suggest that evolutionary changes in anatomy and way of life are more often based on changes in the mechanisms controlling the expression of genes than on sequence changes in proteins. We therefore propose that regulatory mutations account for the major biological differences between humans and chimpanzees.” — King & Wilson. Science. April, 1975.

Complex traits

- Instead of one gene determining a disease or trait, many genes each exert a small influence
- None by themselves can cause or explain the disease or trait fully – but together with environmental influences combine to define an individual outcome

Most common diseases work this way

Fundamental Change 1: Collaborative, public-private projects have defined and made publicly available all shared human genetic variation

The SNP Consortium
1999-2002

HapMap Project
2002-2007

1000 Genomes Project
2009- present
Fundamental change 2:
Technological advances accelerate discovery

Inflammatory Bowel Disease:
Crohn’s Disease (CD) & Ulcerative Colitis (UC)
Chronic, heritable inflammatory diseases of the gastrointestinal system.
- 2000: 0 genes known
- 2005: 2 genes known

Performing a Genomewide Association Study

Best practices are key
- Technical QC
  - Removal of failed SNPs, samples
- Genetic QC
  - Mendelian segregation and HWE
  - Estimating relatedness, gender
  - Population structure
- Analysis-based QC
  - Do initial runs of test statistics show inflation, biases towards missing data, specific allele frequencies

Testing for association
- Most straightforward: compare proportion of each SNP allele in cases and controls

<table>
<thead>
<tr>
<th>SNP</th>
<th>Allele A</th>
<th>Allele G</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cases</td>
<td>22</td>
<td>976</td>
</tr>
<tr>
<td>Controls</td>
<td>68</td>
<td>932</td>
</tr>
</tbody>
</table>

Chi-sq = 24.5, p=7.3 x 10^-7

Simplest tests (single marker regression, chi-square) rule the day – association results requiring arcane statistics/complex multi-marker models are often less reliable

Multiple Testing
- In linkage, p = .001 (.05 / ~50 chromosomal arms) considered potentially significant
- In GWAS, we’re performing O(10^6) tests that are largely independent
  - Each study has hundreds of p<.001 purely by statistical chance (no real relationship to disease)
  - “Genome-wide significance” often set at p=5x10^-8 (= .05 / 1 million tests)
Reversing the curse: the story of GWAS

1996: Risch and Merikangas propose that a p-value of $5 \times 10^{-8}$ (equivalent to a p-value of 0.05 after a Bonferroni correction for 1 million independent tests) is a conservative threshold for declaring significant association in a genome-wide study.

2008: 3 groups publish empirically derived estimates based on dense genome-wide maps of common DNA and estimated appropriate dense-map numbers to be in the range of $2.5$ to $7.2 \times 10^{-8}$

Replication is key

- Don’t believe a report of association from a single study
  - Even with strict quality control there are artifacts that can affect 1 every thousand or ten thousand SNPs and escape notice
  - Strict genomewide significance generally not dramatically exceeded (if it’s reached at all in a single study)

Genomewide Association

‘Manhattan’ plot

Q-Q plot

Fine Mapping

ca. 2006 – the beginning of the era of the genome-wide association study
## Linkage vs. Association

NOD2: low-frequency, strong risk variants
IL23R: low-frequency, strong protective variant
ATG16L1: common associated variant

<table>
<thead>
<tr>
<th>Locus</th>
<th>Frequency</th>
<th>Odds-ratio</th>
<th>ASSOCIATION cases to achieve GWS</th>
<th>LINKAGE Pedigrees to achieve signif.</th>
</tr>
</thead>
<tbody>
<tr>
<td>NOD2 (3 coding SNPs)</td>
<td>5%</td>
<td>3.0</td>
<td>435</td>
<td>1400</td>
</tr>
<tr>
<td>IL23R (Arg381Gln)</td>
<td>7%</td>
<td>0.33</td>
<td>817</td>
<td>~30,000</td>
</tr>
<tr>
<td>ATG16L1 (Thr300Ala)</td>
<td>50%</td>
<td>1.4</td>
<td>1360</td>
<td>~40,000</td>
</tr>
</tbody>
</table>

### Combining studies yields greater power

Opportunity: by combining three published studies, we reap the power of an 8000 sample GWAS

Nearly all progress in GWAS has been the result of multiple study meta-analysis
Similar story across many complex diseases!

Like linkage, early GWAS efforts in SCZ did not conclusively define risk loci

Fundamental Change 3: Collaboration rather than competition is the key

PGC SCZ v2: Genomewide association in schizophrenia with 37,000 cases

More than 100 distinct regions of associated to schizophrenia!!!

Psychiatric Genomics Consortium (PGC)

300+ investigators
80 institutions
20 countries

ac·a·dem·ic \\a·ka-`de-mik\

1: a member of an institution of learning

2: having no practical importance: not involving or relating to anything real or practical
**Academic**

1: a member of an institution of learning
2: having no practical importance; not involving or relating to anything real or practical

“All science is either physics or stamp collecting” – Ernest Rutherford

---

**How do we move beyond stamp collecting to actionable biomedical insights useful in therapeutic development?**

Genetics needs to tell us what molecular perturbations are likely to be safe and effective, and the cell type those changes need to happen in.

---

**From genetics to therapies**

- Discovery of disease genes
- Genetic and environmental factors in patients
- Identification of disease pathways driven by human disease genes
- Modeling disease pathways in animals and patients
- Testing therapeutic hypotheses with small molecules in animals and patients
- Discovery and optimization of new medicines to improve clinical outcomes
- Clinical testing in patients
- Healthcare delivery

**Effective Therapeutic Interventions**

---

**General Challenge**

- Most associations do not identify specific genes and causal mutations – just pointers to small regions with causal influences on disease.
- In order to develop and act on a therapeutic hypothesis, we must go much further

**Key Questions**

- Which gene is connected to disease?
- What biological process is thereby implicated?
- What is the cellular context in which that process acts and is relevant to disease?
- What are the specific functional alleles which perturb the process and promote or protect from disease?

---

**Parallel tracks**

- **Bottom up**
  - Traditional approach of pursuing an individual association result to articulate the molecular impact of risk variation and how it acts to promote disease
- **Top down**
  - Utilize the complete set of GWAS results to identify common pathways/processes implicated in disease pathogenesis

**Early GWAS results implicate ATG16L1, IRGM**

2007

Multiple GWAS studies identify common Thr300Ala variant of **ATG16L1** as associated with CD

Noncoding variation at **IRGM**, later traced to 20 kb promoter CNV

n.b., prior to first GWAS, 0 references on Crohn’s and autophagy
Autophagy & Crohn’s disease

Discovery of disease genes 2007

Modeling disease pathways in animals and patients

Testing therapeutic hypotheses with small molecules in animals and patients

Clinical testing in patients

Effective Therapeutic Interventions

Healthcare delivery

Autophagy & Crohn’s Disease

Genetic and environmental factors in patients

Identification of disease pathways driven by human disease genes

Human cellular models of specific allele effects

Animal models of specific human disease alleles

Compounds

Hypothesis: Many associated genes implicate limited number of pathways

Proof: Statistically significant excess connectivity of genes in GWAS regions

Computational tools enable the integration of ‘human genetic screens’ with other genome-scale screening data

Proteins Encoded in Genomic Regions Associated with Immune-Mediated Disease Physically Interact and Suggest Underlying Biology

Proteins Suggest Underlying Biology

GRAIL plot from Franke et al 2010

Building Networks

Associated Regions

Direct Network

InWeb Network

Indirect Network

Evaluating Significance

examples of potential confounders

• Potential biases in what is discovered
  – large genes have greater chance of being hit by de novo CNVs, mutations
  – GWAS hits are to larger than average LD regions with higher than average gene density

• Potential biases in data we’re integrating with
  - Genes preferentially expressed in the brain are much larger than average
  - Certain families of genes have been more extensively studied in PPI experiments
  - Regulatory elements similarly found at higher density in gene regions

Evaluating Significance

Empirical Null Distribution

Repeat full permutation 50,000 times

...keep moving labels until the network has been fully permuted
PPI Networks identify specific genes and pathways

- Fanconi anemia: 9 synthetic loci
- Rheumatoid arthritis: 27 loci
- Crohn’s disease: 25 loci

**Validation of PPI networks**
Further experimental support that the non-random networks are truly implicating the underlying genes

**Network genes are co-expressed**
Connected proteins are enriched for newly confirmed associated genes (p=6.5x10^{-4})

**Integrating Autoimmune Risk Loci with Gene-Expression Data Identifies Specific Pathogenic Immune Cell Subsets**
Xinshi Gu, Hyun Kim, Robert Plenge, Mark Daly, and Soumya Raychaudhuri
The American Journal of Human Genetics 85, 481–482, October 7, 2011

**ImmGen data set:**
223 murine immune cell subsets
Expression measured on 15,149 human homologs

Are human GWAS hits harboring loci significantly co-expressed in specific immune cell subsets?

**In fact they are!**

**Other opportunities:**
Cross-disease information

Genes coordinately associated to multiple disease are tightly functionally linked
Cotsapas et al, August 2011 PLoS Genetics

**Can genetics really take more substantive steps towards informing therapeutic development?**
True Complexity: moving from complex trait gene discovery to therapeutic hypotheses and models

- GWAS hits do not easily tell us specific genes and perturbations – let alone relevant pathways, complexes and cell types

- Figuring this out will require high-throughput genomics and functional biological screening just coming available
  - Even starting with the obvious (RNAseq and many flavors of ChIPseq in appropriate single cells/tissues, developmental timepoints)

- The simplest, supportable models (human cells, fish, mice, etc) then required
  - Not to fix mouse models of disease but to recapitulate, understand and reverse the specific molecular perturbations that cause human disease

- Then we can start thinking about chemical screening...

But we know genetics provides the substrate on which this can succeed...

Genome Sequencing is now a reality...

**Cost per Raw Megabase of DNA Sequence**

- Providing the capability to find the rare, stronger perturbations that provide therapeutic insight

...but gene discovery via sequencing is not easy

(Ref: all genetics meetings 2012-2013)

1000 cases of autism, 2500 cases of schizophrenia (and similar efforts in diabetes, MI) do not yield compelling novel gene identifications…much like the early GWAS

Computation and interpretation are the challenges

Two uses of NGS

- Exome/genome sequencing in rare and severe disease
- Exome/genome sequencing to complete the allelic architecture (and provide functional handles) at GWAS hits
Sifting signal from noise in exomes

- Every genome contains many rare, potentially functional variants
  - ~500 rare missense variants
  - ~100 LoF variants: ~20 homozygous, ~20 rare
  - ~100 rare variants in known disease genes
  - ~10 recessive disease-causing mutations (~1 LoF)
  - ~0.1 complete gene knockouts of well-preserved gene
  - sequencing errors

- In Mendelian disease patients we need to find 1-2 true causal mutations amidst this “noise” – in common disease patients, we do not even *know* if there is a severe mutation present in any case

**Exome Aggregation Consortium (ExAC): aggregating and calling 92,000 exomes**

<table>
<thead>
<tr>
<th>Consortium</th>
<th>Samples</th>
</tr>
</thead>
<tbody>
<tr>
<td>T2D (T2D-GENES, GoT2D, SIGMA)</td>
<td>16,167</td>
</tr>
<tr>
<td>Heart disease (Ottawa, ATVB, MiGen, PROMIS)</td>
<td>14,352</td>
</tr>
<tr>
<td>SCZ/Bipolar (multiple consortia)</td>
<td>12,361</td>
</tr>
<tr>
<td>The Cancer Genome Atlas (TCGA)</td>
<td>8,566</td>
</tr>
<tr>
<td>Autism (multiple consortia)</td>
<td>8,126</td>
</tr>
<tr>
<td>NHLBI-GO Exome Sequencing Project (ESP)</td>
<td>6,943</td>
</tr>
<tr>
<td>1000 Genomes Project</td>
<td>2,520</td>
</tr>
<tr>
<td>Inflammatory Bowel Disease</td>
<td>1,933</td>
</tr>
<tr>
<td>UK10K (autism/schizophrenia)</td>
<td>1,348</td>
</tr>
<tr>
<td>Northern Finnish Birth Cohort</td>
<td>965</td>
</tr>
<tr>
<td>Other (Mendelian, cancer)</td>
<td>18,515</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>91,796</strong></td>
</tr>
</tbody>
</table>

Subset of 63,658 “reference” samples

**A Systematic Survey of Loss-of-Function Variants in Human Protein-Coding Genes**

- 3.09 \( p < 10^{-3} \)
- 0.09 \( p < \)

**These data allow recognition of genes under exceptional constraint**

- 5-10% of genes with exceptionally reduced rates of missense variation
- heavily enriched for: OMIM dominant/X-linked hits autism/ID de novo LoF hits

**Using LoF deficiency to separate *de novo* LoF variants in cases and controls**

- Autism: 2297
- Unaffected: 510

- \( p = 1.2 \times 10^{-13} \)

**NGS in rare & severe disease**

- Child #1: M, 8 y, healthy
- Child #2: F
  - Severe diarrhea, day 3
  - Protein losing enteropathy, all known etiologies excluded
  - TG > 300 in first month
  - Albumin infusions, parenteral feeding required
  - Died at 17 months from recurrent septicemia
- Child #3: F
  - Identical postnatal phenotype as #2 starting on day 3
  - GI problems abate after 2y, TG treated with Questran
  - Now 3y, reasonably normal diet & growth

*Exome sequencing performed on parents and child #3*

Harland Winter, Elaine Lim (MGH)
Infantile onset of severe protein-losing diarrhea and hypertriglyceridemia

Exome sequencing identifies cause as homozygosity for rare splice variant of DGAT1

Estimated rate 1 in 10-100 million

Outcomes in two dimensions

- Personal
  - Understanding mutation may be instructive for guiding therapy and diet of surviving child
  - Permits potential genetic testing for this family

- Biological insight
  - First known human knockout of a gene provides fundamental insights into human biology
  - Explains severe adverse events in ongoing DGAT1 inhibition clinical trials

KO mice are not KO humans

Severe Mendelian disease points us at mutations that are bad (i.e., what NOT to do therapeutically)

Can we better target our genetic studies to deliver hypotheses on how we should intervene?

SCN9A (a.k.a. NaV1.7)

- Complete 2-hit loss-of-function = congenital indifference/insensitivity to pain
  - Cox et al, Nature 2006

- Different gain of function mutations = erythermalgia, small fiber neuropathy, paroxysmal extreme pain disorder
  - Proc Natl Acad Sci U S A. 2013 Sep 30. Discovery of a selective NaV1.7 inhibitor from centipede venom with analgesic efficacy exceeding morphine in rodent pain models.

Lipids: common & rare genetics point to targets

For purposes of presentation, loci are named according to a nearby gene of interest. In only a few cases is the causal gene yet proven.

Loci causing Mendelian dyslipidemic syndromes

Loci identified by GWAS

Loci targeted by lipid lowering therapies

Slide courtesy of S. Kathiresan

PCSK9 loss-of-function variant lowers LDL and protects against coronary artery disease

Outcome

Hazard Ratio for incident CHD: 0.50 (0.32 – 0.79)

P<0.003

1108 events/9524


As predicted by genetics – PCSK9 inhibition appears safe and effective

Learning to read our GWAS…

- **Fine-mapping**: largest possible genetic sample typed at high-density, imputed into 1000 Genomes variation patterns
- **Integration 1 (bottom-up)**: systematic application of eQTL, RNAseq, ChiPseq data at each locus
- **Integration 2 (top-down)**: implication of pathways and processes by analyzing all GWAS hits simultaneously using PPI data, cell-specific co-expression, etc.
- **Completing the picture**: exome sequencing to define effect of rare, high-impact functional variants – essentially obtaining the full allelic series at each locus

Overall summary 1

- 103 regions evaluated in fine mapping
- 6 regions top SNP is $10^{-5} > p > 10^{-6}$
- Assessment of multiple testing burden across the regions in this experiment suggests a -log10 $p > 5.8$ should be used to declare secondary signals - containing false positive rate across the experiment
Three Bayesian based approaches:
- Flat, Greedy
- Double exponential, LASSO & MCMC
- Final credible set

Merged and integrated for conservative (union) credible set definition

First Crohn’s GWAS (2006)
- 950 cases, 980 controls

Today’s fine-mapping

IL23R locus – 2013 fine mapping:
- 5 conditionally independent alleles – all associated to both CD & UC
- 3 SNP credible set including R381Q
- 4 SNP cred. set - rs2019262 (non-coding)
- 1 SNP cred. set - V362I (rare, protective)
- 2 SNP cred. set – G149R (rare, protective)
- 24 SNP credible set (non-coding)

Summary: 139 independent concordant signals (blue/green)

1-variant credible sets
(Yes, Virginia, GWAS can be wrestled down to conclusive causal variants – sometimes)

- 200 bp into intron of TNFSF6B
- 5 kb downstream of GPR35
- 500 bp from TSS of JAK2, massive ENCODE peak
- Intronic to IL2RA (MEF2A/MEF2C binding site)
- RELA/NFKB binding site 40 kb upstram IKZF1
- 10 kb from TSS of NNX2-3, also LINCO1475 inbetween SNP and NNX2-3
- Intron LRRK2
- 5kb downstream from HNF4A
- 4 kb from TSS of PRDM1

8 coding, 10 non-coding – highly enriched for coding versus remaining GWAS results
Refining genetic signals: Example SMAD3

**2010: GWAS**
CD scans implicates region on chromosome 15 containing multiple genes {SMAD3, AAGAB, IQCH}

**2012-2013: fine-mapping**
Primary signal refined to 5 SNP set – top SNP is 43% likely to be causal

Most likely Crohn’s disease SNP in SMAD3 disrupts a conserved AP1 site

**2010: GWAS**
CD scans implicates region on chromosome 15 containing multiple genes {SMAD3, AAGAB, IQCH}

**2012-2013: fine-mapping**
IIBDGC immunochip project refines association to SMAD3 noncoding functional element

A minority of autoimmunity SNPs create or destroy canonical transcription binding site motifs

<table>
<thead>
<tr>
<th>Gene</th>
<th>Observed</th>
<th>Expected</th>
<th>P-value</th>
<th>Association</th>
</tr>
</thead>
<tbody>
<tr>
<td>HLA-DRB1</td>
<td>10</td>
<td>20</td>
<td>0.01</td>
<td>SVM</td>
</tr>
<tr>
<td>CASP15</td>
<td>5</td>
<td>10</td>
<td>0.05</td>
<td>SVM</td>
</tr>
<tr>
<td>CASP15</td>
<td>2</td>
<td>4</td>
<td>0.02</td>
<td>SVM</td>
</tr>
<tr>
<td>CASP15</td>
<td>2</td>
<td>4</td>
<td>0.02</td>
<td>SVM</td>
</tr>
<tr>
<td>CASP15</td>
<td>2</td>
<td>4</td>
<td>0.02</td>
<td>SVM</td>
</tr>
</tbody>
</table>

Autoimmunity SNPs are enriched at TF binding sites
Immune TF motifs are strongly enriched in the 100-nucleotides adjacent to autoimmunity SNPs

Comparison of GWAS hits with blood eQTL hits

Integration of eQTL data

- Examine all significant regional cis-eQTLs ($p < p_{\text{threshold}}$)
- Regenerate eQTL association, conditional on association to index SNP(s) from IBD fine-mapping
- Assess by permutation significance of the $p$-value drop – setting regional $p < 0.05$

Path is not always so straight...but possible

Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease

Path is not always so straight...but possible

Splice variant in CARD9 (MAF ~ 0.5%) causes premature truncation of protein and strongly protects against the development of Crohn's disease ($p < 10^{-16}$)

Examples

Path is not always so straight...but possible

Protective form of CARD9 does not activate inflammatory signals

Missing segment of CARD9 required for binding to TRIM62

Ubiquitination by TRIM62 required for CARD9 activation