Module 4: Population and Disease Genetics

- **L13: Population genetics:**
  - Measuring and understanding human variation
- **L14: Disease association mapping:**
  - Molecular basis of human phenotypic variation and disease
- **L15: Quantitative trait mapping:**
  - Intermediate phenotypes bridging the genotype-phenotype gap
- **L16: Heritability:**
  - Whole-genome disease association beyond top hits

Today: Human genetics and GWAS

1. Genetics intro: Mendel, human traits, Linkage
   - Mendelian traits: linkage, genetic maps, family studies
   - Complex traits: polygenicity, environment, continuous

2. Genome-wide association studies (GWAS)
   - Study design: QC, Chi-Sq, mult testing, replication, QQ
   - Fine mapping, linkage vs. association, combine studies

3. Interpretation: Individual loci vs. global signals
   - Loci: mechanism, assays, allelic, patients, mice, cells
   - Sys: networks, biases, stats, PPI, expression, diagnostics

4. Next Gen: exome, genome, medical sequencing
   - Common, low-freq, rare, private, somatic; diagnostics
   - Autism: polygenic, carriers; case/control, trios; de novo

Mode 1: Informing therapeutic development

How can genetics research help meet outstanding medical challenges?

- Most therapies developed through traditional approaches fail to have efficacy
- Therapeutic development desperately needs insights into human biology...
- Genetics provides a validated means to these insights
Mode 2: Personalized genomic medicine

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

Age-Related Macular Degeneration

Three bad and two good alleles

Where are the differences that contribute to heritable disease?

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

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

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, Alzheimers, diabetes…

Mendelian disease genetics

Linkage analysis and positional cloning are powerful because genetic risk factors are highly “penetrant”
Today: Human genetics and GWAS

1. Genetics intro: Mendel, human traits, Linkage
   - Mendelian traits: linkage, genetic maps, family studies
   - Complex traits: polygenicity, environment, continuous

2. Genome-wide association studies (GWAS)
   - Study design: QC, Chi-Sq, mult testing, replication, QQ
   - Fine mapping, linkage vs. association, combine studies

3. Interpretation: Individual loci vs. global signals
   - Loci: mechanism, assays, allelic, patients, mice, cells
   - Sys: networks, biases, stats, PPI, expression, diagnostics

4. Next Gen: exome, genome, medical sequencing
   - Common, low-freq, rare, private, somatic; diagnostics
   - Autism: polygenic, carriers; case/control, trios; de novo

Linkage revolution in medical genetics did not generally apply to complex traits

The first complex trait mapping: Altenburg and Muller (1920)

Hermann Muller

Polygenicity proved a fundamental barrier...

...as did the incomplete genotype-phenotype relationship

"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.

Many Mendelian traits can sum to a continuous distribution

Ronald Fisher (1918)
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

15 years ago

• First human genome was just being completed
• Only a handful of genes had been discovered for any common disease

Glazier, Nadeau and Aitman, Science 2002

Today: Human genetics and GWAS

1. Genetics intro: Mendel, human traits, Linkage
   – Mendelian traits: linkage, genetic maps, family studies
   – Complex traits: polygenicity, environment, continuous

2. Genome-wide association studies (GWAS)
   – Study design: QC, Chi-Sq, mult testing, replication, QQ
   – Fine mapping, linkage vs. association, combine studies

3. Interpretation: Individual loci vs. global signals
   – Loci: mechanism, assays, allelic, patients, mice, cells
   – Sys: networks, biases, stats, PPI, expression, diagnostics

4. Next Gen: exome, genome, medical sequencing
   – Common, low-freq, rare, private, somatic; diagnostics
   – Autism: polygenic, carriers; case/control, trios; de novo

Three major elements turned the tide

Genome Resources
Technology
Collaboration

Fundamental Change 1: Understanding the Genome

• Sequencing of the human genome
• Understanding and cataloging variation in the genome (HapMap, 1000 Genomes Project)

Fundamental change 2: Technological advances accelerate discovery

New technologies to examine genomes...

Arrays of millions of DNA variants (2005-present)
Sequencing the entire human genome (2010)

Enable us to compare the genomes of...

Affecteds Controls

Populations of cases and controls
Family members with and without disease
Best practices are key

• Technical QC
  – Removal of failed SNPs, samples

• Genetic QC
  – Mendelian segregation and Hardy Weinberg Equilibrium
  – 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>rs11209026</td>
<td>Cases</td>
<td>Controls</td>
</tr>
<tr>
<td></td>
<td>22</td>
<td>976</td>
</tr>
<tr>
<td></td>
<td>68</td>
<td>932</td>
</tr>
<tr>
<td>Chi-sq = 24.5, p=7.3 x 10⁻⁷</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Simplest tests (single marker regression, $\chi^2$) rule the day. Association results requiring arcane statistics. Complex multi-marker models are often less reliable.

Multiple Testing

• In linkage, $p = .001 (0.05 / \sim 50$ 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 $p=5\times10^{-8} (=0.05 / 1$ 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 $2.5 \text{ to } 7.2 \times 10^{-8}$
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

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

_Example – associated SNP with MAF = 0.20_

A Genome-Wide Association Study Identifies IL23R as an Inflammatory Bowel Disease Gene


Genome-wide association study identifies new susceptibility loci for Crohn’s disease and implicates autophagy in disease pathogenesis


Schizophrenia GWAS: Number of significant loci

- 3,500 cases  0 loci
- 10,000 cases  5 loci
- 35,000 cases  62 loci!
- 65,000 cases  265 loci!

Similar inflection point found in every complex trait!

<table>
<thead>
<tr>
<th>Adult height</th>
<th>Crohn’s</th>
<th>Schizophrenia</th>
</tr>
</thead>
<tbody>
<tr>
<td>(per 5000/5000)</td>
<td>(per 1000/1000)</td>
<td>(per 3000/3000)</td>
</tr>
<tr>
<td>1x</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>2x</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>3x</td>
<td>7</td>
<td>5</td>
</tr>
<tr>
<td>9x</td>
<td>68</td>
<td>62</td>
</tr>
<tr>
<td>18x</td>
<td>180</td>
<td>-</td>
</tr>
</tbody>
</table>

Significantly associated regions (p < 5e-08)

Larger samples lead to new biological insights

- Proof that Schizophrenia is a heritable, medical disorder
- Genetic architecture similar to non-brain diseases and traits
- Many genes recognition of key pathways and processes
  - Voltage-gated calcium channels (CACNA1C, CACNA1D, CACNA1I, CACNB2)
  - Proteins interacting with FMRI, fragile X gene
  - Neuron organization: Postsynaptic density, dendritic spine heads
  - Enhancers: brain (angular gyrus, inferior temporal lobe), immune

Published Genome-Wide Associations through 03/2011, 1,319 published GWA at p<5x10^-8 for 221 traits

NHGRI GWA Catalog
www.genome.gov/GWASStudies
Goal of GWAS was to inform on the biology of disease in an actionable fashion

So how do we get there?

**Challenge of interpretation**

- 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?

**Genomic medicine: challenge and promises**

- Disease mechanism
- New target genes
- New therapeutics
- Personalized medicine

**Genomic medicine: challenge and promises**

- Relevant cell type
- Target genes
- Causal variant
- Upstream regulator
- Relevant pathways
- Intermediate phenotypes

**The promise of genetics**

- Disease mechanism
- New target genes
- New therapeutics
- Personalized medicine

**The challenge of mechanism**

- 90+% disease hits non-coding
- Target gene not known
- Causal variant not known
- Cell type of action not known
- Relevant pathways not known
- Mechanism not known

**The remedy**

- Annotate non-coding genome (ENCODE/Roadmap)
- Link enhancers to regulators and target genes
- Elucidate intermediate molecular and cellular phenotypes

**The deliverables**

- Relevant cell type
- Target genes
- Causal variant
- Upstream regulator
- Relevant pathways
- Intermediate phenotypes
Epigenomic mapping across 100+ tissues/cell types

Diverse tissues and cells
- Adult tissues and cells (brain, muscle, heart, digestive, skin, adipose, lung, blood...)
- Fetal tissues (brain, skeletal muscle, heart, digestive, lung, cord blood...)
- ES cells, iPSCs, differentiated cells (meso/endo/ectoderm, neural, mesench...)

Diverse epigenomic assays
- Histone modifications: H3K4me3, H3K4me1, H3K36me3
- H3K27me3, H3K9me3, H3K27/9ac
- 20+ more
- Open chromatin: DNA accessibility
  - WGBS, RRBS, MRE/MeDIP
- Gene expression
  - RNA-seq, Exon Arrays

Use resulting annotations and networks for GWAS

- Expand each GWAS locus using SNP linkage disequilibrium (LD)
  - Recognize relevant cell types: tissue-specific enhancer enrichment
  - Recognize driver TFs: enriched motifs in multiple GWAS loci
  - Recognize target genes: linked to causal enhancers

HaploReg: systematic mining of GWAS variants

- Start with any list of SNPs or select a GWA study
  - Mine ENCODE and Roadmap epigenomics data for hits
  - Hundreds of assays, dozens of cells, conservation, motifs
  - Report significant overlaps and link to info/browser
- Try it out: http://compbio.mit.edu/HaploReg

Ward, Kellis NAR 2011

Identifying disease-relevant cell types

- For every trait in the GWAS catalog:
  - Identify all associated regions at P-value threshold
  - Consider all SNPs in credible interval (R² ≥ 0.8)
  - Evaluate overlap with tissue-specific enhancers
  - Keep tissues showing significant enrichment (P < 0.001)
- Repeat for all traits (rows) and all cell types (columns)
Immune activation + neural repression in human + mouse

Epigenomics of AD progression

Inflammation as the causal component of Alzheimer’s disease

AD variants localize in immune cells, not neuronal

Immune Neuronal
time time

Immune activation precedes neuronal repression

Dissecting non-coding genetic associations

1. Tissue/cell type(s)
2. Target gene(s)
3. Causal nucleotide(s)
4. Upstream regulator(s)
5. Cellular phenotypes
6. Organismal phenotypes

2. Mechanistic dissection of a non-coding disease locus

• Identify cell type, causal SNP, regulator, targets, process
• Genome editing demonstrates variant causality
• Adipocyte browning drivers of obesity

This talk:
Apply these to the FTO locus

FTO Obesity Variant Circuitry and Adipocyte Browning in Humans

Melina Claussnitzer, Ph.D., Simon N. Daniel, Ph.D., Kyeong-Han Kim, Ph.D., Gerald Quon, Ph.D., Wouter Meuleman, Ph.D., Christine Haigler, M.S., Victoria Clark, M.S., Isabel S. Squizzato, M.S., Jasmine L. Rosado, Ph.D., Vijaya Purushottam, B.S., Natarajan Abendroth, M.S., Janelle Liu, B.S., Per Arne Swensson, Ph.D., T. Hsiang Hsu, Ph.D., Daniel J. Drucker, M.D., Gunnar Møller, M.D., Ph.D., Chi-Chung Hui, Ph.D., Hans Hauner, M.D., and Manolis Kellis, Ph.D.

September 3, 2015
Vol. 373, No. 10

FTO region: strongest association with obesity

• Associated with obesity, Type 2 Diabetes, Cardiovascular traits
• 89 variants in LD, spanning 47kb, intron 1 of FTO gene
• No protein-altering variants: regulatory role? Target gene, tissue?

1. Tissue: Chromatin states predict adipocyte function

Progenitors of white/beige adipocytes

12 kb super-enhancer
2. Targets: 3D folding and expr. genetics indicate IRX3+IRX5

Cohort of 20 homozygous risk and 18 homozygous non-risk individuals: Genotype-dependent expression?

Topological domains span 2.5Mb Implicate 8 candidate genes

eQTL targets: IRX3 and IRX5

Risk allele: increased expression (gain-of-function)

3. Causal SNP: motif enrichment + conservation: rs1421085

Regulatory motifs enriched in BMI GWAS hits

Regulatory motif combinations conserved across mammals

Causal nucleotide rs1421085: risk alters T to C, abolishes AT-rich motif

4. Regulator: Causality and epistasis of ARID5B repressor

Regulatory model: risk allele disrupts a repressor

Enhancer activity Cis/trans conditional analysis IRX3/5 expression

Steps 5-6. Does this circuitry actually lead to obesity?

1. Establish relevant tissue/cell type: pre-adipocytes
2. Establish downstream target gene(s): IRX3 and IRX5
3. Establishing causal nucleotide variant: rs1421085
4. Establish upstream regulator causality: ARID5B
5. Establish cellular phenotypic consequences
6. Establish organismal phenotypic consequences

Expression analysis to recognize target processes

Search for genes co-expressed with IRX3 and IRX5 (n=20 indiv.)

Reflected in cellular phenotypes

Negative correlation: mitochondria Positive correlation: lipid storage

Risk carriers: increased mito Non-risk: increased adipocytes

Test model by systematic perturbations

Thermogenic stimuli (e.g. cold) Browning mitochondrial thermogenesis

Lean
Obese

ARID5B KD (obesity) ARID5B OE (anti-obesity)

IRX3, IRX5 knock-down (anti-obesity phenotypes)
IRX3, IRX5 overexpression (pro-obesity phenotypes)

C-to-T motif rescue (anti-obesity phenotypes) T-to-C motif disruption (pro-obesity phenotypes)

Risk allele: shift from dissipation to storage
IRX3+IRX5 expression impacts energy utilization

Risk individuals: IRX3/5 repression restores respiration, thermogenesis
Non-risk: IRX3/5 overexpression disrupts respiration, thermogenesis

Irx3 adipose repression: anti-obesity phenotypes in mice

54% reduced body weight
Resistance to high-fat diet

Increased energy dissipation
- No reduction in appetite
- No increase in exercise
- In thermoneutral conditions
- Day and night (not exercise)

Single-nucleotide editing reverses thermogenesis in humans

rs1421085 editing alters IRX3+IRX5 expression (500,000 and 1 million nucleotides away!)
rs1421085 editing restores thermogenesis

rs1421085 causality: C-to-T editing rescues IRX3/IRX5 expression, ARID5B repression, thermogenesis, developmental expression

Model: beige ↔ white adipocyte development

FTO obesity locus mechanistic dissection

1. Establish relevant tissue/cell type: pre-adipocytes
2. Establish downstream target gene(s): IRX3 and IRX5
3. Establishing causal nucleotide variant: rs1421085
4. Establish upstream regulator causality: ARID5B
5. Establish cellular phenotypic consequences: thermogenesis
6. Establish organismal phenotypic consequences: body weight

Today: Human genetics and GWAS

1. Genetics intro: Mendel, human traits, Linkage
   - Mendelian traits: linkage, genetic maps, family studies
   - Complex traits: polygenicity, environment, continuous
2. Genome-wide association studies (GWAS)
   - Study design: QC, Chi-Sq, mult testing, replication, QQ
   - Fine mapping, linkage vs. association, combine studies
3. Interpretation: Individual loci vs. global signals
   - Loci: mechanism, assays, allelic, patients, mice, cells
   - Sys: networks, biases, stats, PPI, expression, diagnostics
4. Next Gen: exome, genome, medical sequencing
   - Common, low-freq, rare, private, somatic; diagnostics
   - Autism: polygenic, carriers; case/control, trios; de novo
Bayesian fine-mapping: Predict causal variant and cell type

RIVIERA: multi-trait GWAS integration

Capture conserved elements

Combine GWAS+Epig to find new target genes/SNPs

Prioritize sub-threshold loci (<10^{-4})

Validate new enhancers:
- allelic activity, enh-prom looping

Machine learning predictive features

Validate new genes in hum/mou/zb

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

DAPPLE

GRAIL plot from Franke et al 2010

Evaluating Significance

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

Direct connectivity
p << 2x10^{-5}
p = 3x10^{-4}
p = 1.11x10^{-3}
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)

GWAS hits significantly co-expressed in specific immune cell subsets

Other opportunities: Cross-disease information

Molecular diagnostics in IBD

Molecular diagnostics flag patients with worst outcome

~Molecular~ diagnosis (based on GWAS SNPs & serologic biomarkers) concordant with GI dx: CD & UC patients can be distinguished accurately

>90% of patients correctly classified with >90% reliability

Black dots represent patients diagnosed with UC who later underwent colectomy and then developed full-blown Crohn’s disease
Today: Human genetics and GWAS

1. Genetics intro: Mendel, human traits, Linkage
   – Mendelian traits: linkage, genetic maps, family studies
   – Complex traits: polygenicity, environment, continuous

2. Genome-wide association studies (GWAS)
   – Study design: QC, Chi-Sq, mult testing, replication, QQ
   – Fine mapping, linkage vs. association, combine studies

3. Interpretation: Individual loci vs. global signals
   – Loci: mechanism, assays, allelic, patients, mice, cells
   – Sys: networks, biases, stats, PPI, expression, diagnostics

4. Next Gen: exome, genome, medical sequencing
   – Common, low-freq, rare, private, somatic; diagnostics
   – Autism: polygenic, carriers; case/control, trios; de novo

Further progress: getting to functional alleles and rarer causes of disease

Can next-generation sequencing give us what we need?

Genome Sequencing is now a reality

Cost per Human Genome

<table>
<thead>
<tr>
<th>Year</th>
<th>Venter / capillaries</th>
<th>Watson / 454</th>
<th>African, Asian, Cancer pair</th>
<th>169 in Genbank</th>
</tr>
</thead>
<tbody>
<tr>
<td>2012</td>
<td>$1M</td>
<td>$10k</td>
<td>$100k</td>
<td>$100M</td>
</tr>
<tr>
<td>2008</td>
<td>$10k</td>
<td>$1M</td>
<td>$10M</td>
<td>$100k</td>
</tr>
<tr>
<td>2007</td>
<td>$100k</td>
<td>$10M</td>
<td>$100M</td>
<td>$100k</td>
</tr>
<tr>
<td>2006</td>
<td>$20,000k</td>
<td>$100M</td>
<td>$1M</td>
<td>$1M</td>
</tr>
<tr>
<td>2005</td>
<td>$200k</td>
<td>$100M</td>
<td>$100k</td>
<td>$100k</td>
</tr>
<tr>
<td>2004</td>
<td>$2,000k</td>
<td>$100M</td>
<td>$100k</td>
<td>$100k</td>
</tr>
<tr>
<td>2003</td>
<td>$20,000k</td>
<td>$100M</td>
<td>$100k</td>
<td>$100k</td>
</tr>
<tr>
<td>2002</td>
<td>$200,000k</td>
<td>$1M</td>
<td>$10M</td>
<td>$100k</td>
</tr>
<tr>
<td>2001</td>
<td>$2,000,000k</td>
<td>$100k</td>
<td>$100k</td>
<td>$100k</td>
</tr>
<tr>
<td>2000</td>
<td>$20,000,000k</td>
<td>$10M</td>
<td>$100k</td>
<td>$100k</td>
</tr>
<tr>
<td>1999</td>
<td>$200,000,000k</td>
<td>$100k</td>
<td>$100k</td>
<td>$100k</td>
</tr>
<tr>
<td>1998</td>
<td>$2,000,000,000k</td>
<td>$1M</td>
<td>$10M</td>
<td>$100k</td>
</tr>
<tr>
<td>1997</td>
<td>$20,000,000,000k</td>
<td>$100k</td>
<td>$100k</td>
<td>$100k</td>
</tr>
<tr>
<td>1996</td>
<td>$200,000,000,000k</td>
<td>$1M</td>
<td>$10M</td>
<td>$100k</td>
</tr>
<tr>
<td>1995</td>
<td>$2,000,000,000,000k</td>
<td>$10k</td>
<td>$100k</td>
<td>$100k</td>
</tr>
<tr>
<td>1994</td>
<td>$20,000,000,000,000k</td>
<td>$100k</td>
<td>$100k</td>
<td>$100k</td>
</tr>
<tr>
<td>1993</td>
<td>$200,000,000,000,000k</td>
<td>$1M</td>
<td>$10M</td>
<td>$100k</td>
</tr>
<tr>
<td>1992</td>
<td>$2,000,000,000,000,000k</td>
<td>$10k</td>
<td>$100k</td>
<td>$100k</td>
</tr>
<tr>
<td>1991</td>
<td>$20,000,000,000,000,000k</td>
<td>$100k</td>
<td>$100k</td>
<td>$100k</td>
</tr>
<tr>
<td>1990</td>
<td>$200,000,000,000,000,000k</td>
<td>$1M</td>
<td>$10M</td>
<td>$100k</td>
</tr>
<tr>
<td>1989</td>
<td>$2,000,000,000,000,000,000k</td>
<td>$10k</td>
<td>$100k</td>
<td>$100k</td>
</tr>
<tr>
<td>1988</td>
<td>$20,000,000,000,000,000,000k</td>
<td>$100k</td>
<td>$100k</td>
<td>$100k</td>
</tr>
<tr>
<td>1987</td>
<td>$200,000,000,000,000,000,000k</td>
<td>$1M</td>
<td>$10M</td>
<td>$100k</td>
</tr>
<tr>
<td>1986</td>
<td>$2,000,000,000,000,000,000,000k</td>
<td>$10k</td>
<td>$100k</td>
<td>$100k</td>
</tr>
<tr>
<td>1985</td>
<td>$20,000,000,000,000,000,000,000k</td>
<td>$100k</td>
<td>$100k</td>
<td>$100k</td>
</tr>
<tr>
<td>1984</td>
<td>$200,000,000,000,000,000,000,000k</td>
<td>$1M</td>
<td>$10M</td>
<td>$100k</td>
</tr>
<tr>
<td>1983</td>
<td>$2,000,000,000,000,000,000,000,000k</td>
<td>$10k</td>
<td>$100k</td>
<td>$100k</td>
</tr>
<tr>
<td>1982</td>
<td>$20,000,000,000,000,000,000,000,000k</td>
<td>$100k</td>
<td>$100k</td>
<td>$100k</td>
</tr>
<tr>
<td>1981</td>
<td>$200,000,000,000,000,000,000,000,000k</td>
<td>$1M</td>
<td>$10M</td>
<td>$100k</td>
</tr>
<tr>
<td>1980</td>
<td>$2,000,000,000,000,000,000,000,000,000k</td>
<td>$10k</td>
<td>$100k</td>
<td>$100k</td>
</tr>
<tr>
<td>1979</td>
<td>$20,000,000,000,000,000,000,000,000,000k</td>
<td>$100k</td>
<td>$100k</td>
<td>$100k</td>
</tr>
<tr>
<td>1978</td>
<td>$200,000,000,000,000,000,000,000,000,000k</td>
<td>$1M</td>
<td>$10M</td>
<td>$100k</td>
</tr>
<tr>
<td>1977</td>
<td>$2,000,000,000,000,000,000,000,000,000,000k</td>
<td>$10k</td>
<td>$100k</td>
<td>$100k</td>
</tr>
<tr>
<td>1976</td>
<td>$20,000,000,000,000,000,000,000,000,000,000k</td>
<td>$100k</td>
<td>$100k</td>
<td>$100k</td>
</tr>
<tr>
<td>1975</td>
<td>$200,000,000,000,000,000,000,000,000,000,000k</td>
<td>$1M</td>
<td>$10M</td>
<td>$100k</td>
</tr>
<tr>
<td>1974</td>
<td>$2,000,000,000,000,000,000,000,000,000,000,000k</td>
<td>$10k</td>
<td>$100k</td>
<td>$100k</td>
</tr>
<tr>
<td>1973</td>
<td>$20,000,000,000,000,000,000,000,000,000,000,000k</td>
<td>$100k</td>
<td>$100k</td>
<td>$100k</td>
</tr>
<tr>
<td>1972</td>
<td>$200,000,000,000,000,000,000,000,000,000,000,000k</td>
<td>$1M</td>
<td>$10M</td>
<td>$100k</td>
</tr>
<tr>
<td>1971</td>
<td>$2,000,000,000,000,000,000,000,000,000,000,000,000k</td>
<td>$10k</td>
<td>$100k</td>
<td>$100k</td>
</tr>
<tr>
<td>1970</td>
<td>$20,000,000,000,000,000,000,000,000,000,000,000,000k</td>
<td>$100k</td>
<td>$100k</td>
<td>$100k</td>
</tr>
</tbody>
</table>

Computation and interpretation are the challenges

NGS publications in human genetics 2011

- Mendelian/extremely rare disorders 35%
- Cancer genomics 15%
- Review/perspective articles 25%
- Analysis Methods 15%
- Everything else… 10%(de novo or inherited mutations in chronic, complex disease, model systems)

Disclaimer: very rough tally based on my going through pages and pages of PubMed listings, YMMV
**What’s in a single exome?**

- ~20,000 DNA variants in/near protein coding DNA
- ~100 rare missense variants
- ~100 loss-of-function variants (~20 rare or private)

A few things are a bit more interpretable…but not absolute slam dunks…

- ~1 de novo variant per exome (only 5% LoF)
- <5% chance that an individual has a complete knockout of even a single gene where LoF mutations are rare in general

**CD: insights about CARD9 suggest therapeutic hypothesis**

**Cell model**

- Risk variant
- Protective variant
- Drug leads that mimic the protective variant may reduce inflammation

**Therapeutic hypothesis**

- CD

**NGS in rare & severe cases**

- **Child #1:** M, 8 y, healthy
  - 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 #2:** 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)
1 in 100 million genetic deficiency

- ~20,000 variants in or near protein coding regions…
- …only a single instance in which the child had inherited two rare mutations in the same gene, one from each parent.

Exon | Intron
TTGCC gtagga

splice donor site disrupted GT -> GC

Only 1 in 10,000 individuals are carriers of this mutation.

Confirmation of sequencing results

Outcomes

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

- Biological
  - First known human knockout of a gene provides fundamental insights into human biology

Today: Human genetics and GWAS

1. Genetics intro: Mendel, human traits, Linkage
   - Mendelian traits: linkage, genetic maps, family studies
   - Complex traits: polygenicity, environment, continuous

2. Genome-wide association studies (GWAS)
   - Study design: QC, Chi-Sq, mult testing, replication, QQ
   - Fine mapping, linkage vs. association, combine studies

3. Interpretation: Individual loci vs. global signals
   - Loci: mechanism, assays, allelic, patients, mice, cells
   - Sys: networks, biases, stats, PPI, expression, diagnostics

4. Next Gen: exome, genome, medical sequencing
   - Common, low-freq, rare, private, somatic; diagnostics
   - Autism: polygenic, carriers; case/control, trios; de novo

Mosaicism: Somatic mutation burden in disease from scRNA

- 4925 cells
- 37 individuals
- Mutations: Glial >> neuronal
- Disease neurons >> healthy neurons
- Burden differs in disease: olig >> neurons
- Mutational burden: males >> female indiv.
- Cell-type-specific mosaic mutation burden/gene
- Enriched: vesicle trafficking, Golgi, interm. filament • Mosaic hits in glial cells (dividing)

Moving genomics into diagnostics

- NGS technologies are now achieving accuracies and costs comparable to Sanger-based diagnostic screening
- Reality: the same test will sometimes yield immediately clinically interpretable results and in others, simply potential clues that require further research to interpret
- Challenge: how can we make these technologies and tools available broadly in hospitals throughout the world
Today: Human genetics and GWAS

1. Genetics intro: Mendel, human traits, Linkage
   - Mendelian traits: linkage, genetic maps, family studies
   - Complex traits: polygenicity, environment, continuous

2. Genome-wide association studies (GWAS)
   - Study design: QC, Chi-Sq, mult testing, replication, QQ
   - Fine mapping, linkage vs. association, combine studies

3. Interpretation: Individual loci vs. global signals
   - Loci: mechanism, assays, allelic, patients, mice, cells
   - Sys: networks, biases, stats, PPI, expression, diagnostics

4. Next Gen: exome, genome, medical sequencing
   - Common, low-freq, rare, private, somatic; diagnostics
   - Autism: polygenic, carriers; case/control, trios; de novo

Tackling extreme heterogeneity

ARRA Autism Sequencing Collaboration
- Collaboration among expert sequencing centers with experienced autism genetics labs
- Co-funded by NIMH and NHGRI
  - Particular thanks to Thomas Lehner and Adam Felsenfeld
- Investigators:
  - Joe Buxbaum (MSSM), Mark Daly (Broad/MGH), Bernie Devlin/Kathryn Roeder (Pittsburgh/CMU), Richard Gibbs/Eric Boerwinkle (Baylor), Jerry Schellenberg (Penn), Jim Sutcliffe (Vanderbilt)

Fundamental Challenges
- Extreme polygenicity
  - If rare variants in a particular gene confer risk, it is likely only a small fraction of cases carry them
- Extremely high background rate of neutral variation
  - >300,000 variants found ‘on target’ to date

Data collected to date
- Phase 1: 920 cases and 892 controls
  - Match autism cases to NIMH Controls on PCA distance using common variation
  - Sequencing split between Broad and Baylor
  - Pair samples through NGS process
  - Production recently completed, analysis ongoing
- Phase 2: 174 trios
  - Sequencing performed at all consortium sites
  - Many more in process

Nonsense mutations specific to cases or controls
- Many nonsense mutations seen in ~1% of cases and no controls – however nearly just as many seen uniquely in controls
- Likely some are relevant – but which ones?!?!
What about de novo variation?

945 trios and counting...

<table>
<thead>
<tr>
<th></th>
<th>OBSERVED</th>
<th>EXPECTED</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>de novo</td>
<td>de novo</td>
</tr>
<tr>
<td></td>
<td>per trio</td>
<td>per trio</td>
</tr>
<tr>
<td>Affected</td>
<td>0.360</td>
<td>0.270</td>
</tr>
<tr>
<td>Synonymous</td>
<td>0.043</td>
<td>0.024</td>
</tr>
<tr>
<td>Missense</td>
<td>0.156</td>
<td>0.086</td>
</tr>
<tr>
<td>Loss of Function</td>
<td>0.213</td>
<td>0.268</td>
</tr>
<tr>
<td>Unaffected</td>
<td>0.012</td>
<td>0.012</td>
</tr>
<tr>
<td>Synonymous</td>
<td>0.007</td>
<td>0.007</td>
</tr>
<tr>
<td>Missense</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Loss of Function</td>
<td>0.087</td>
<td>0.087</td>
</tr>
</tbody>
</table>

• Comparisons to well-calibrated mutation model and to observed rates in unaffecteds
• Vast majority of de novo missense variants and half the de novo LoF are unrelated to autism risk

Most compelling results becoming convincing...

<table>
<thead>
<tr>
<th>Gene</th>
<th># de novo LoF</th>
<th># de novo missense</th>
<th>p-value</th>
<th># LoF cases</th>
<th># LoF control</th>
<th># LoF ESP</th>
</tr>
</thead>
<tbody>
<tr>
<td>KATNAL2</td>
<td>2</td>
<td>0</td>
<td>9.96E-06</td>
<td>3</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>DYRK1A</td>
<td>2</td>
<td>0</td>
<td>2.16E-05</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>POGZ</td>
<td>2</td>
<td>0</td>
<td>7.62E-05</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>SCN2A</td>
<td>2</td>
<td>1</td>
<td>1.33E-04</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>CHD8</td>
<td>2</td>
<td>0</td>
<td>2.17E-04</td>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Appropriate threshold for genome-wide significance roughly 1x10^-6

.05 / (Ngenes * Ntests)

Ntests here set at 2 (all LoF, all LoF+missense)

What have we learned thusfar?

5% of cases carry a large CNV or chromosomal abnormality
5% of cases have two copies of an important gene disabled
~10% of cases carry a high risk spontaneous point mutation

Nearly all cases carry a high-risk inherited genetic background defined by common and rare DNA variants

All of these forms of genetic variation are producing fundamental genetic insights into the biology of autism

For all of these ‘bins’, only a few of the specific factors have been identified – we need to increase our efforts to both explain the biology and provide insights to families

Summary: Human genetics and GWAS

1. Genetics intro: Mendel, human traits, Linkage
   - Mendelian traits: linkage, genetic maps, family studies
   - Complex traits: polygenicity, environment, continuous

2. Genome-wide association studies (GWAS)
   - Study design: QC, Chi-Sq, mult testing, replication, QQ
   - Fine mapping, linkage vs. association, combine studies

3. Interpretation: Individual loci vs. global signals
   - Loci: mechanism, assays, allelic, patients, mice, cells
   - Sys: networks, biases, stats, PPI, expression, diagnostics

4. Next Gen: exome, genome, medical sequencing
   - Common, low-freq, rare, private, somatic; diagnostics
   - Autism: polygenic, carriers; case/control, trios; de novo