

After the Genotyping is Done:
**Sequencing the Genomes of
Thousands of Individuals**

Gonçalo Abecasis

University of Michigan School of Public Health

Genomewide Association Studies

- Survey 100,000 – 1,000,000 SNPs in a large samples
- An effective way to skim the genome and ...
- ... find common variants associated with a trait of interest
- Rapid increase in number of known complex disease loci
 - For example, >25 loci now implicated in type 2 diabetes
- Often reveal (unexpected) connections between traits
- Still, many questions remain unanswered.

Questions that are Hard to Answer with Genomewide Association Data...

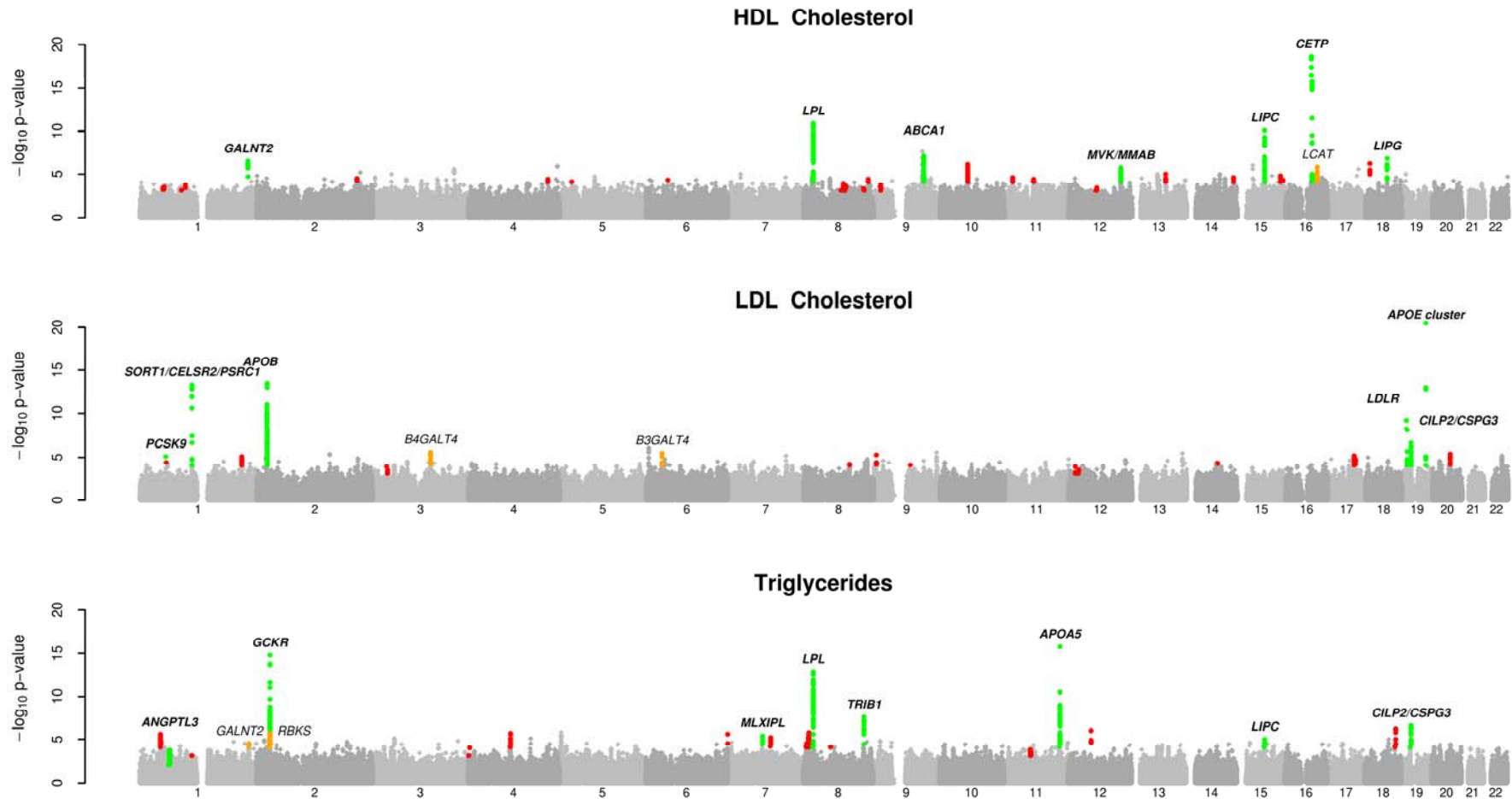
- What is the contribution of each identified locus to a trait?
 - Likely that multiple variants, common and rare, will contribute
- What is the mechanism? What happens when we knockout a gene?
 - Most often, the causal variant will not have been examined directly
 - Rare coding variants will provide important insights into mechanisms
- What is the contribution of structural variation to disease?
 - These are hard to interrogate using current genotyping arrays.
- Are there additional susceptibility loci to be found?
 - Only subset of functional elements include common variants ...
 - Rare variants are more numerous and thus will point to additional loci

What Is the Total Contribution of Each Locus?

Evidence that
Multiple Variants Will be Important

Evidence for Multiple Variants Per Locus

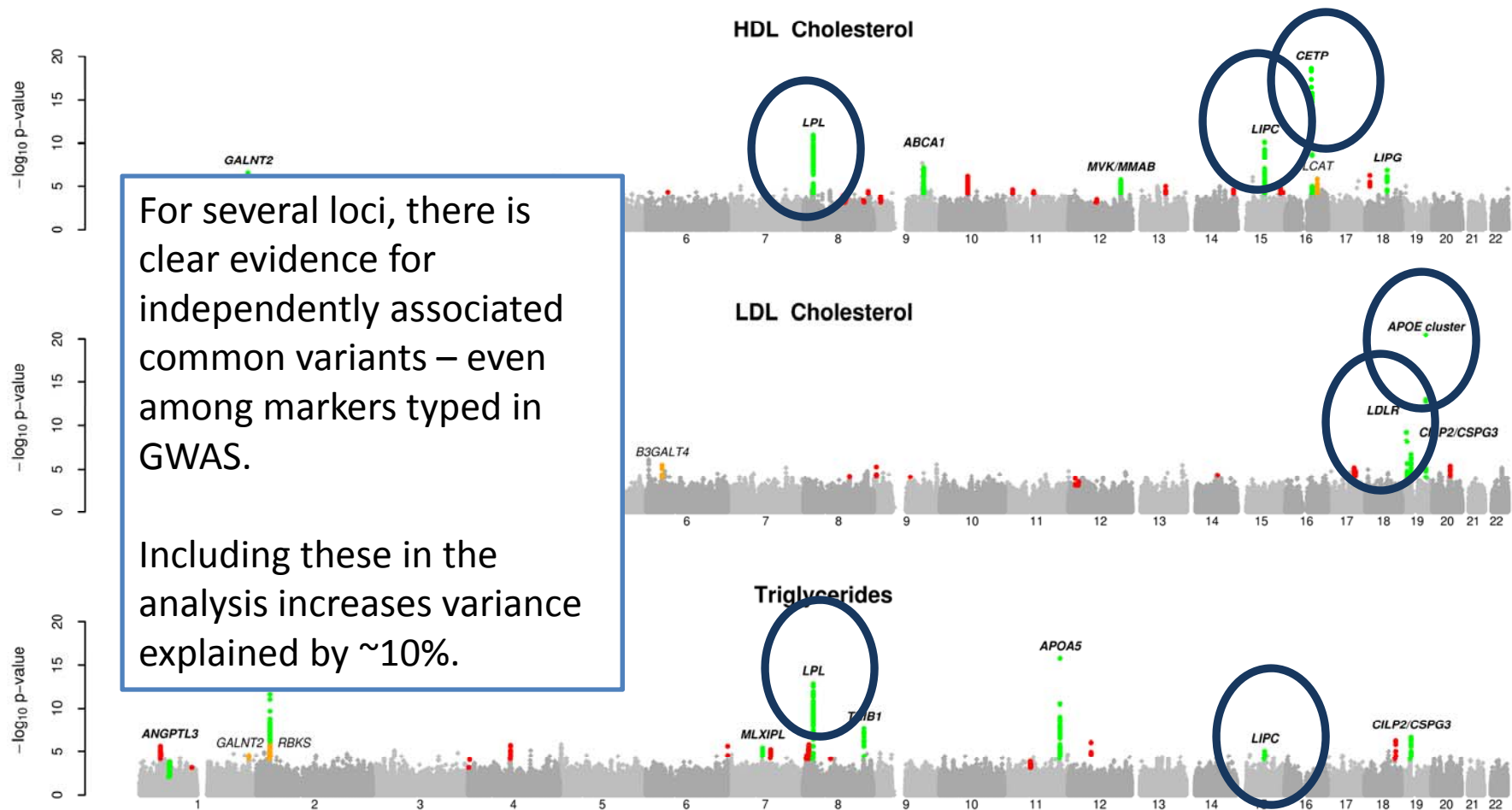
Example from Lipid Biology



Willer et al, *Nat Genet*, 2008
Kathiresan et al, *Nat Genet*, 2008, 2009

Evidence for Multiple Variants Per Locus

Example from Lipid Biology



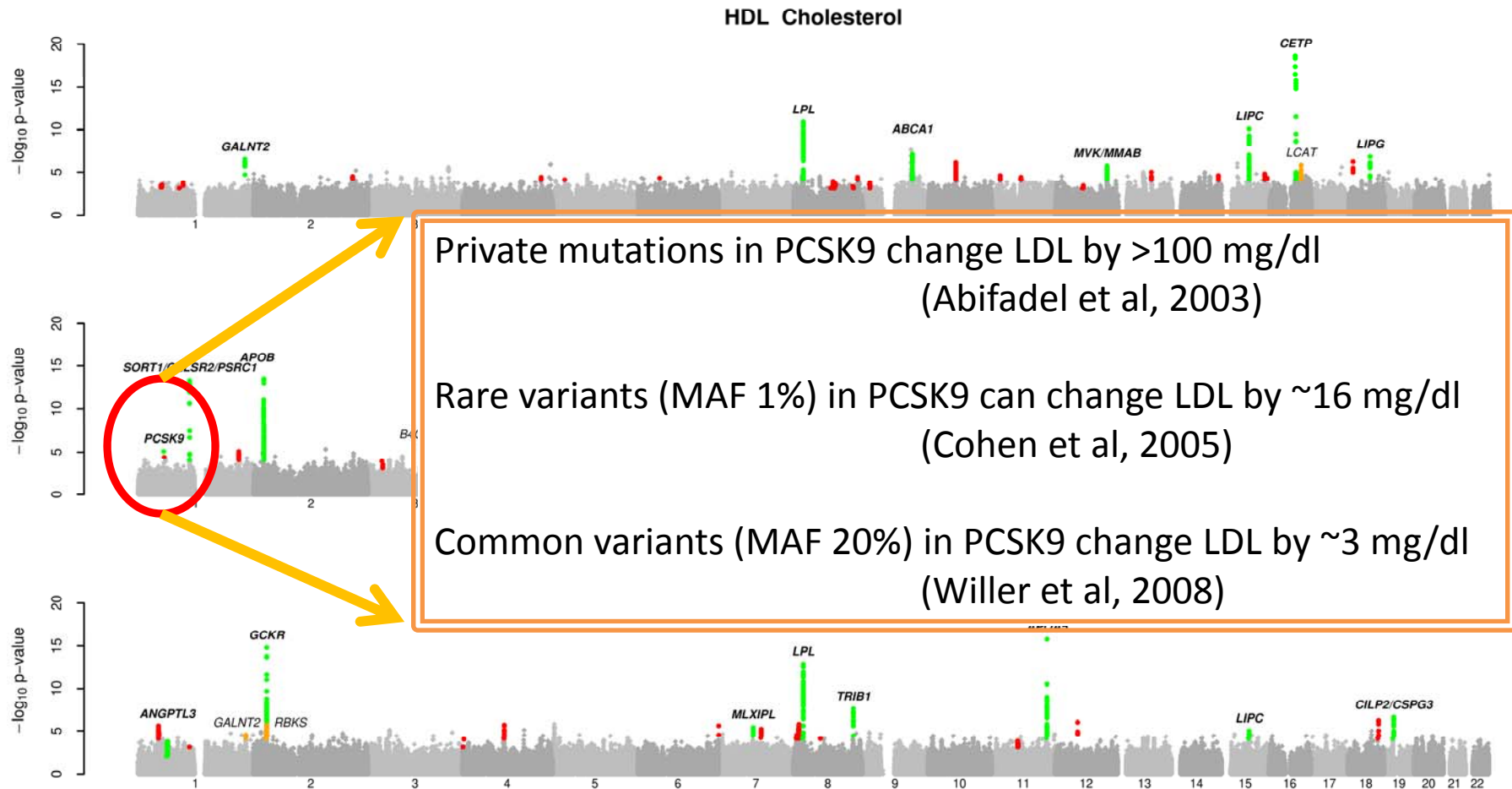
For several loci, there is clear evidence for independently associated common variants – even among markers typed in GWAS.

Including these in the analysis increases variance explained by ~10%.

Willer et al, *Nat Genet*, 2008
Kathiresan et al, *Nat Genet*, 2008, 2009

Evidence for Multiple Variants Per Locus

Example from Lipid Biology



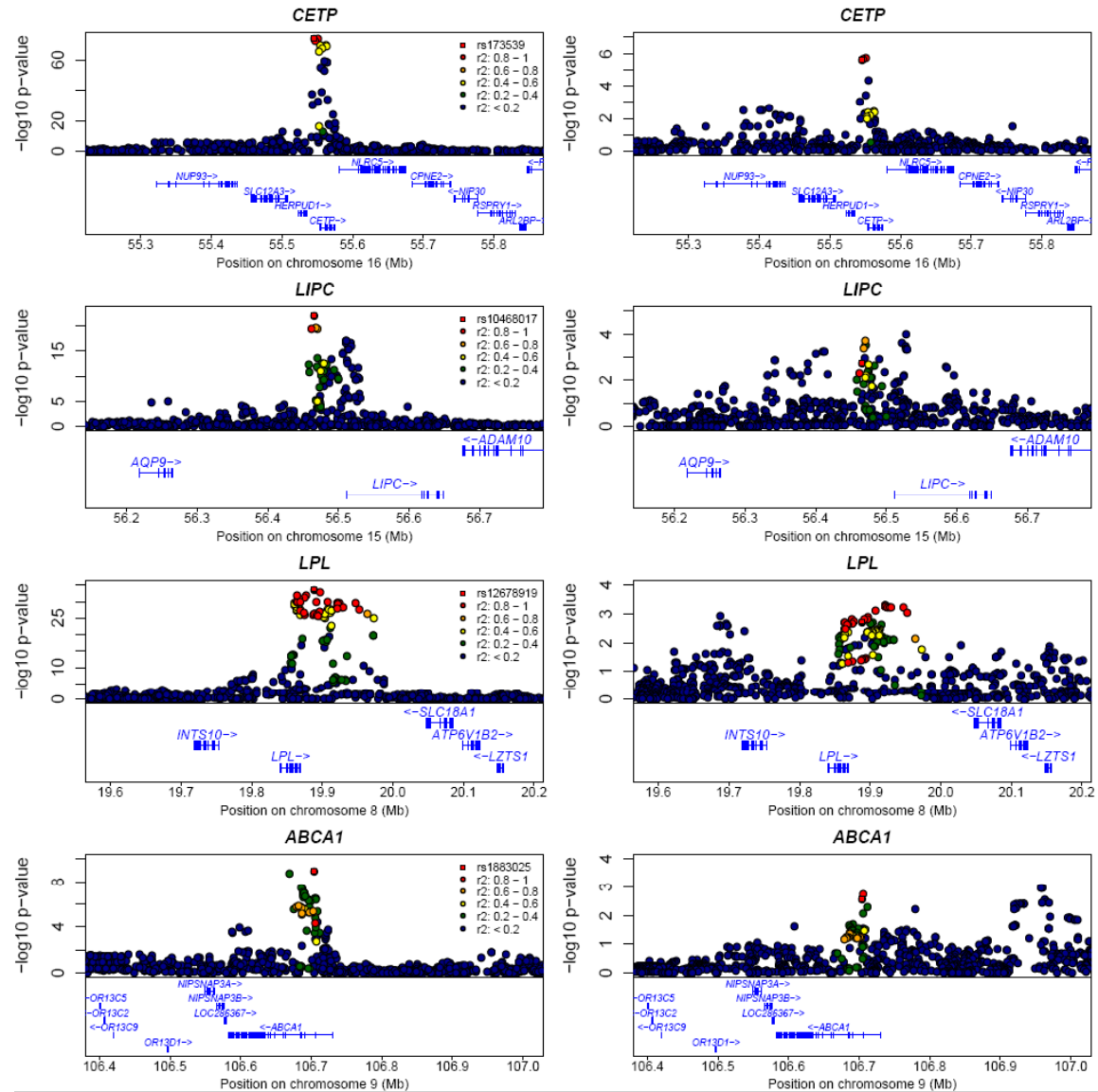
Willer et al, *Nat Genet*, 2008
Kathiresan et al, *Nat Genet*, 2008, 2009

Coronary Artery Disease and LDL-C associated alleles

Gene	LDL-C p-value	Frequency CAD cases	Frequency CAD ctrls	CAD p-value	OR
<i>APOE/C1/C4</i>	3.0×10^{-43}	.209	.184	1.0×10^{-4}	1.17 (1.08-1.28)
<i>APOE/C1/C4</i>	1.2×10^{-9}	.339	.319	.0068	1.10 (1.02-1.18)
<i>SORT1</i>	6.1×10^{-33}	.808	.778	1.3×10^{-5}	1.20 (1.10-1.31)
<i>LDLR</i>	4.2×10^{-26}	.902	.890	6.7×10^{-4}	1.29 (1.10-1.52)
<i>APOB</i>	5.6×10^{-22}	.830	.824	.18	1.04 (0.95-1.14)
<i>APOB</i>	8.3×10^{-12}	.353	.332	.0042	1.10 (1.03-1.18)
<i>APOB</i>	3.1×10^{-9}	.536	.520	.028	1.07 (1.00-1.14)
<i>PCSK9</i>	3.5×10^{-11}	.825	.807	.0042	1.13 (1.03-1.23)
<i>NCAN/CILP2</i>	2.7×10^{-9}	.922	.915	.055	1.11 (0.98-1.26)
<i>B3GALT4</i>	5.1×10^{-8}	.399	.385	.039	1.07 (0.99-1.14)
<i>B4GALT4</i>	1.0×10^{-6}	.874	.865	.051	1.09 (0.98-1.20)

CAD data from Wellcome Trust Case Control Consortium; LDL-C loci from Willer et al (2008)

Overlap of AMD and HDL-C Loci

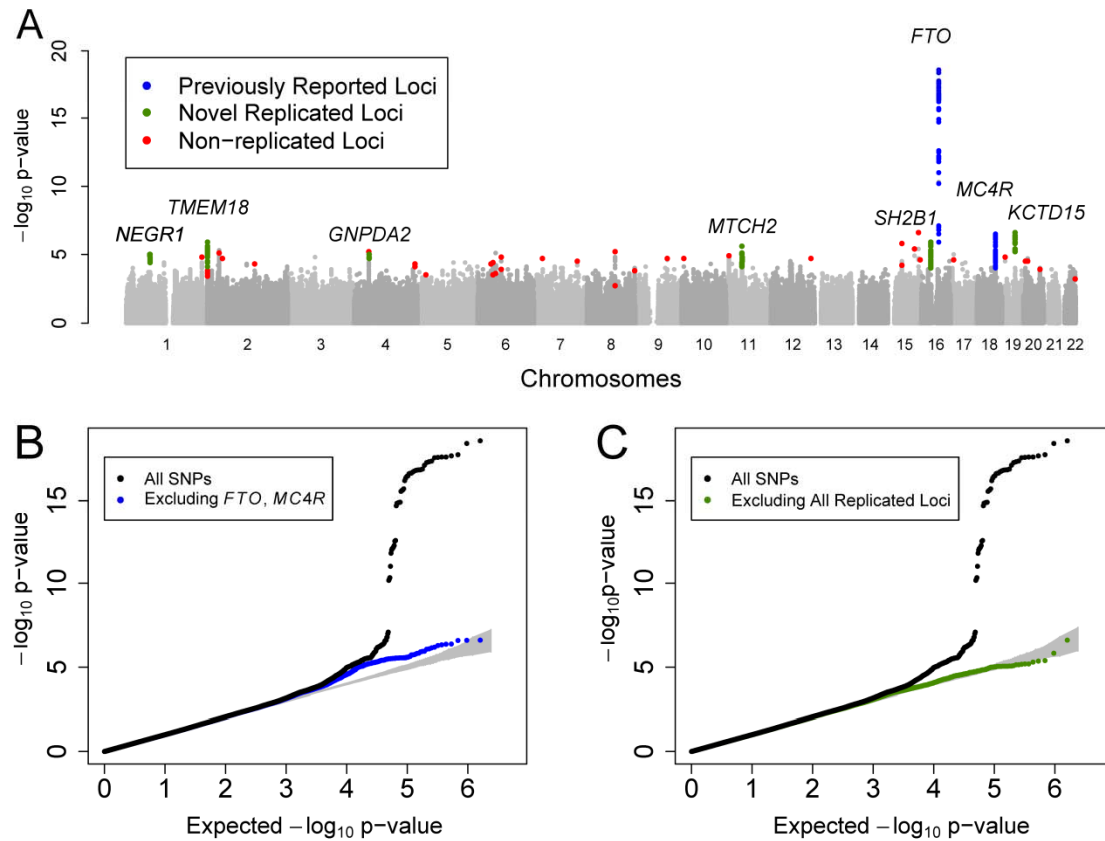


What is The Contribution of Structural Variants?

Current Arrays Interrogate
1,000,000s of SNPs,
but 100s of Structural Variants

Evidence that Copy Number Variants Important

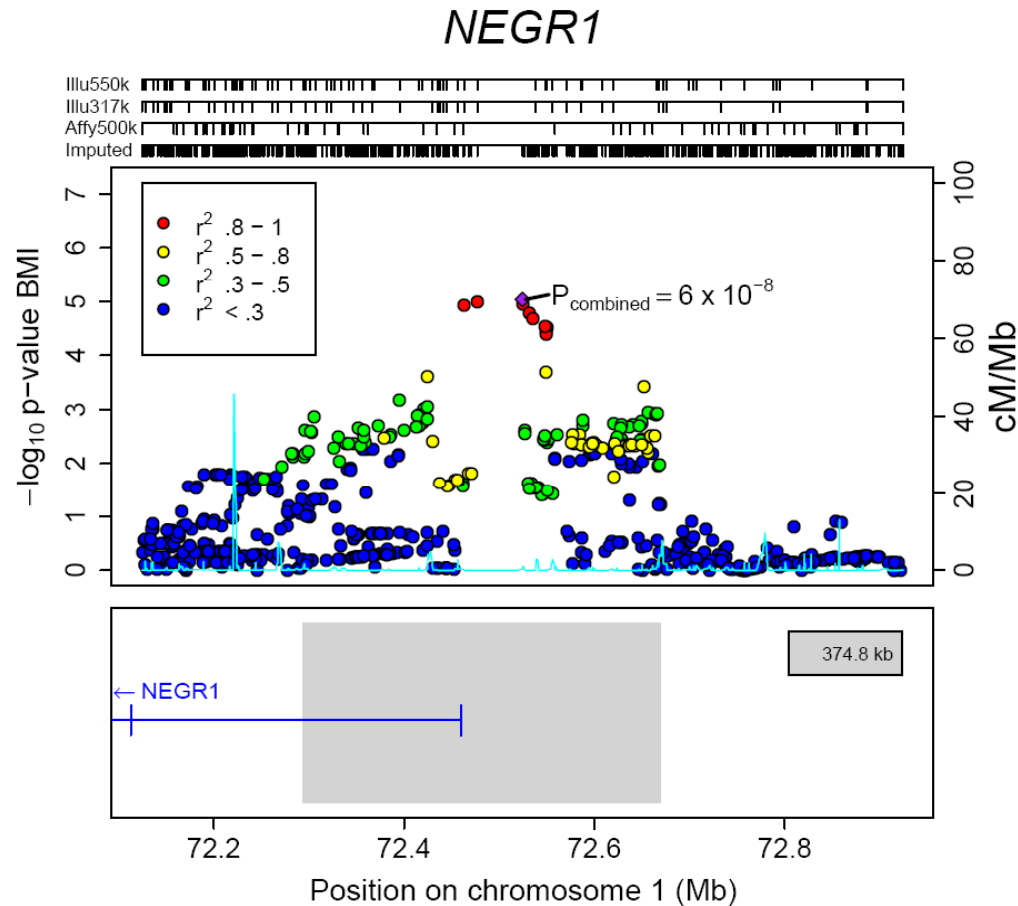
Example from Genetics of Obesity



Seven of eight confirmed BMI loci show strongest expression in the brain...

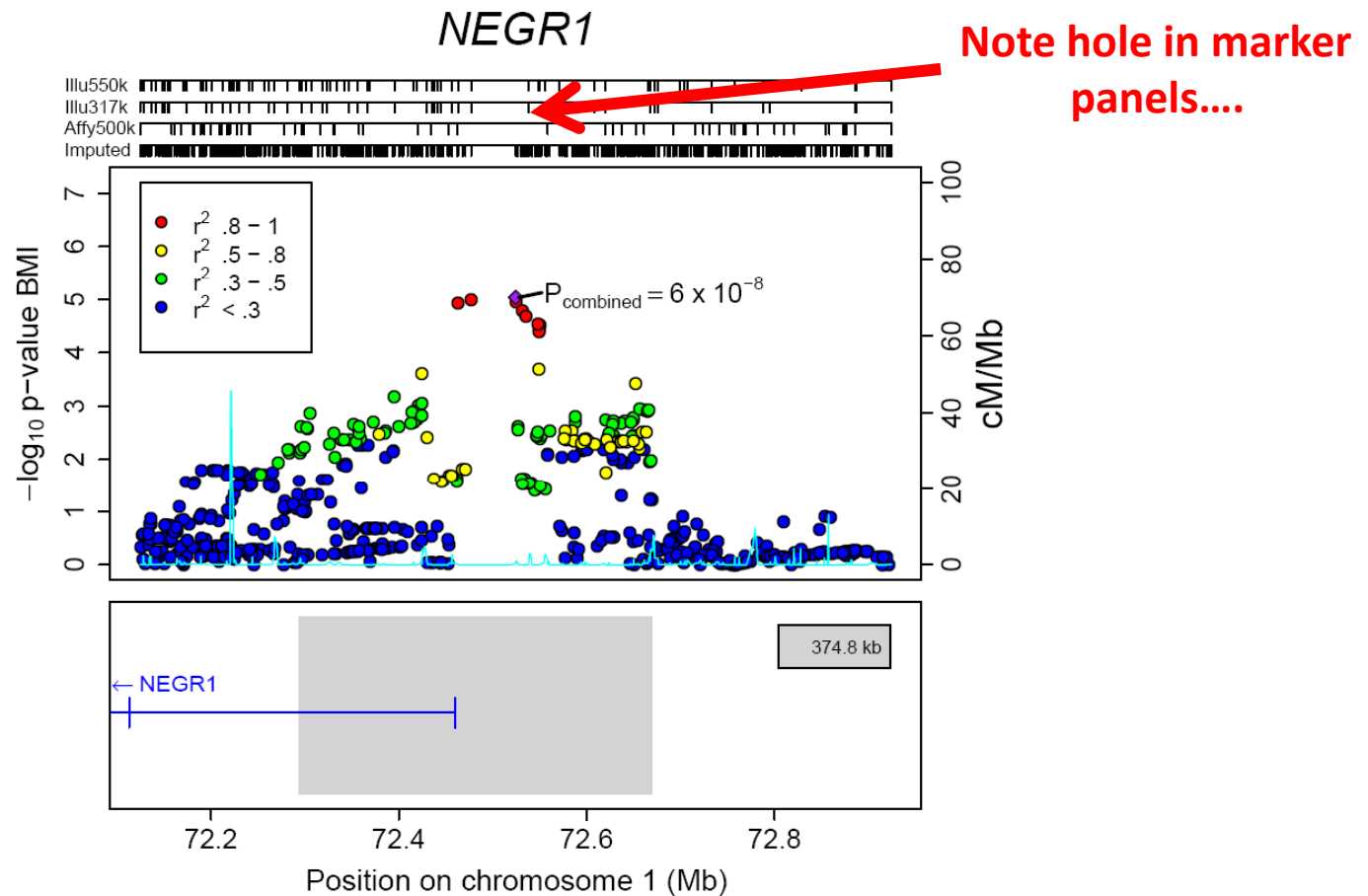
Evidence that Copy Number Variants Important

Example from Genetics of Obesity

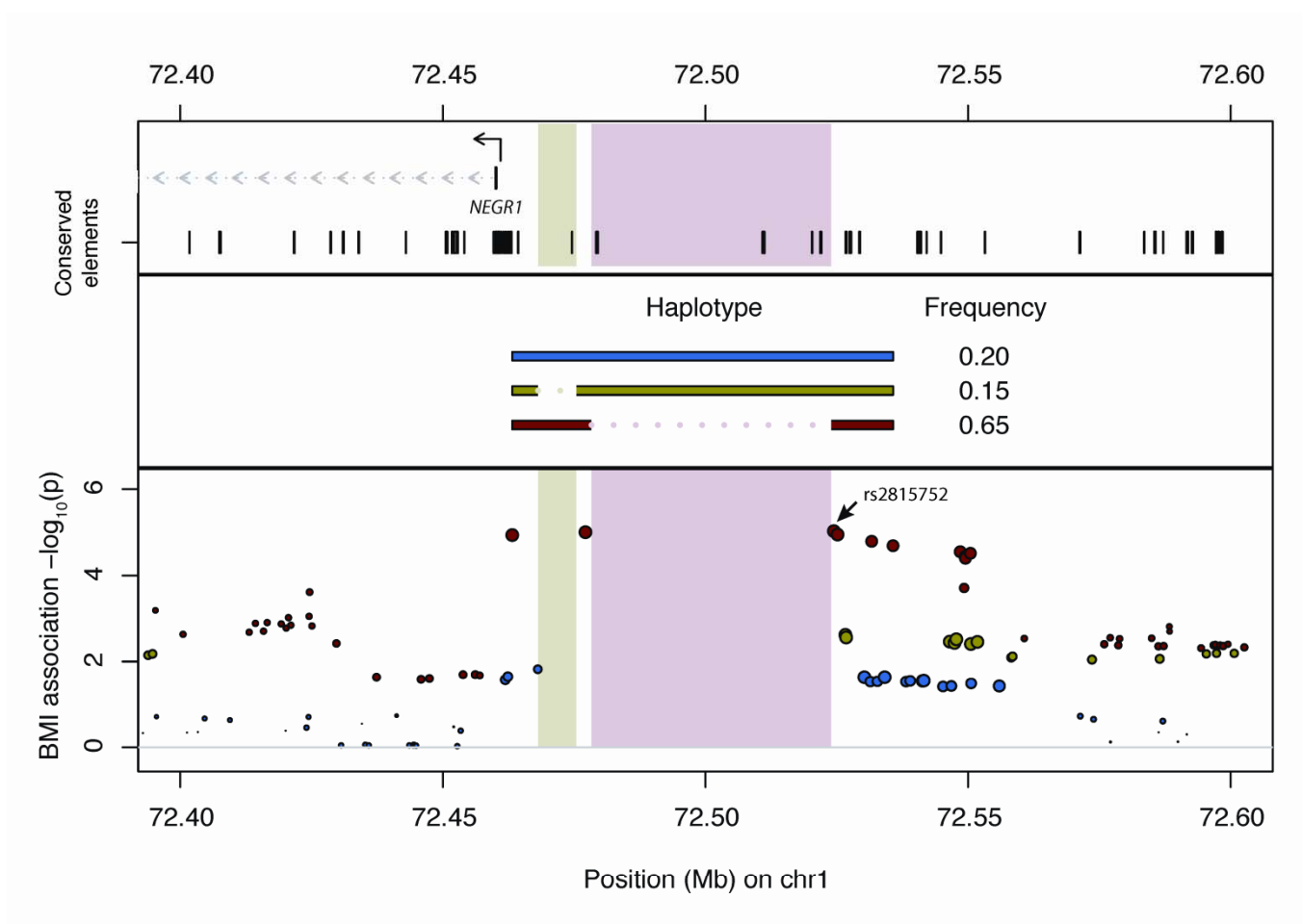


Evidence that Copy Number Variants Important

Example from Genetics of Obesity



Associated Haplotype Carries Deletion



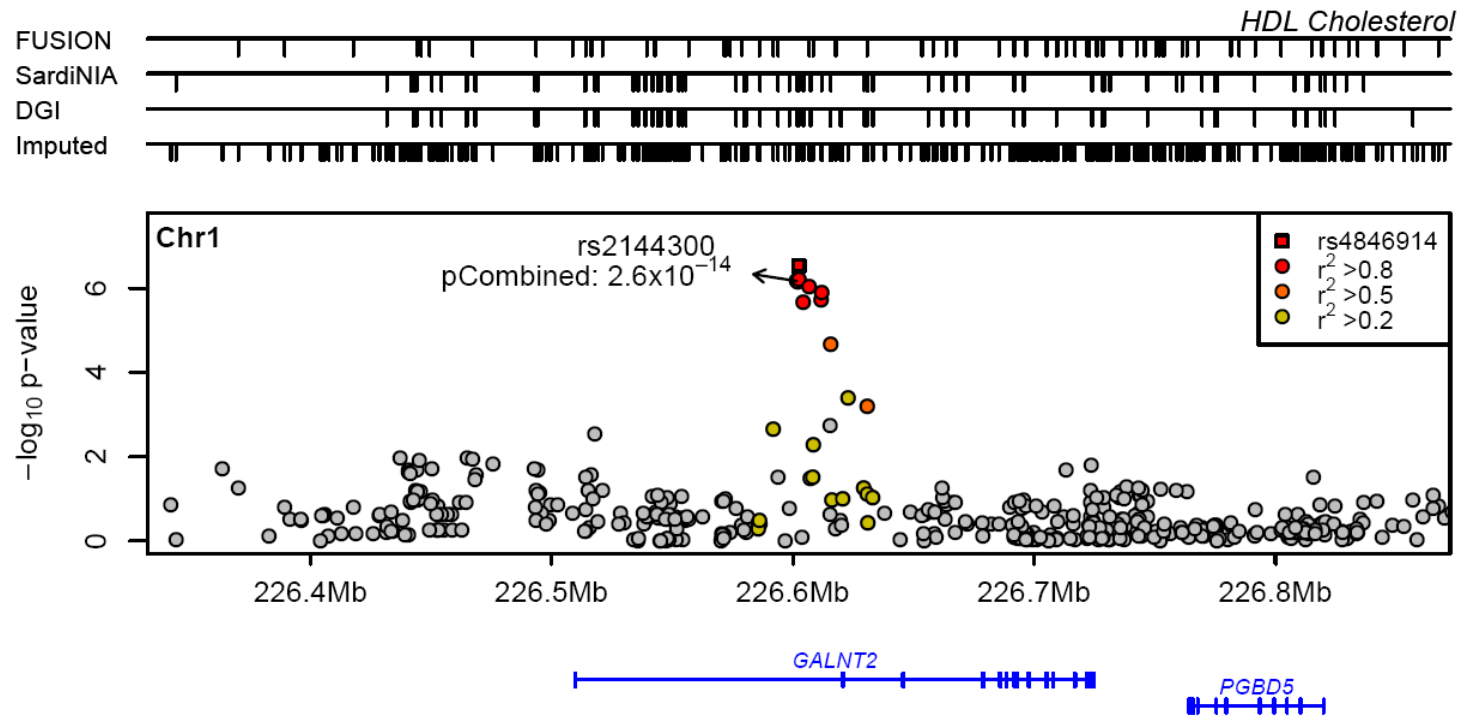
What is the Mechanism?

What Can We Learn From Rare Knockouts?

What We'd Like to Know

Recent Example from John Todd's Group

HDL-C Associated Locus



- GWAS allele with 40% frequency associated with ± 1 mg/dl in HDL-C
- *GALNT2* expression in mouse liver (Edmonson, Kathiresan, Rader)
 - Overexpression of *GALNT2* or *Galnt2* decreases HDL-C $\sim 20\%$
 - Knockdown of *Galnt2* increases HDL-C by $\sim 30\%$

Can Rare Variants Replace Model Systems?

Example from Type 1 Diabetes

- Nejentsev, Walker, Riches, Egholm, Todd (2009)
IFIH1, gene implicated in anti-viral responses, protects against T1D
Science **324**:387-389
- Common variants in IFIH1 previously associated with type 2 diabetes
- Sequenced IFIH1 in ~480 cases and ~480 controls
- Followed-up of identified variants in >30,000 individuals
- Identified 4 variants associated with type 1 diabetes including:
 - 1 nonsense variant associated with reduced risk
 - 2 variants in conserved splice donor sites associated with reduced risk
 - Result suggests disabling the gene protects against type 1 diabetes

Status of Whole Genome Sequencing: 1000 Genome Project Example

- NA12878 sequenced to 65x depth (33x Illumina, 20x SOLiD, 12x 454)
 - Parents also sequenced extensively
- Calls made at 90.5% of all sites in the reference genome (Q30)
 - Copy number variation (depth filter) excludes ~3% of genome
 - Uniqueness (mapping quality filter) excludes ~6% of genome
- 2,985,516 single nucleotide variants in NA12878
 - 93% of these variants are also seen in 1 or more the 180+ other sequenced individuals
- What portions of the genome have been re-sequenced?
 - 99.5% of HapMap III sites (with 99.93% concordance)
 - 98.0% of sites in MIR repeats
 - 98.0% of sites in L2 repeats
 - 91.6% of sites in protein coding exons
 - 78.1% of sites in L1 repeats
 - 70.9% of sites in Alu repeats
 - 28.3% of sites in segmental duplications (with an excess of SNPs!)

The Challenge

- Whole genome sequence data will greatly increase our understanding of complex traits
- Although a handful of genomes have been sequenced, this remains a relatively expensive enterprise
- Dissecting complex traits will require whole genome sequencing of 1,000s of individuals
- **How to sequence 1,000s of individuals cost-effectively?**

Next Generation Sequencing

Massive Throughput Sequencing

- Tools to generate sequence data evolving rapidly
- Commercial platforms produce gigabases of sequence rapidly and inexpensively
 - ABI SOLiD, Illumina Solexa, Roche 454, Complete Genomics, and others...
- Sequence data consist of thousands or millions of short sequence reads with moderate accuracy
 - 0.5 – 1.0% error rates per base may be typical

Shotgun Sequence Data



TAGCTGATAGCTAG**A**TAGCTGATGAGCCCGAT
ATAGCTAG**A**TAGCTGATGAGCCCGATCGCTGCTAGCTC
ATGCTAGCTGATAGCTAG**C**TAGCTGATGAGCC
AGCTGATAGCTAG**C**TAGCTGATGAGCCCGATCGCTG
GCTAGCTGATAGCTAG**C**TAGCTGATGAGCCCGA

Sequence Reads

5'-ACTGGTCGATGCTAGCTGATAGCTAG**C**TAGCTGATGAGCCCGATCGCTGCTAGCTCGACG-3'

Reference Genome

A/C

Predicted Genotype

Shotgun Sequence Data

Sequence Reads

5'-ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG-3'

Reference Genome

$$P(\text{reads} | A/A) = 1.0$$

$$P(\text{reads} | A/C) = 1.0$$

$$P(\text{reads} | C/C) = 1.0$$

Possible Genotypes

Shotgun Sequence Data

GCTAGCTGATAGCTAGCTAGCTAGCTGATGAGCCCGA

Sequence Reads

5'-ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG-3'

Reference Genome

$P(\text{reads} | A/A) = P(\text{C observed, read maps} | A/A)$

$P(\text{reads} | A/C) = P(\text{C observed, read maps} | A/C)$

$P(\text{reads} | C/C) = P(\text{C observed, read maps} | C/C)$

Possible Genotypes

Shotgun Sequence Data

GCTAGCTGATAGCTAGCTAGCTAGCTGATGAGCCCGA

Sequence Reads

5'-ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG-3'

Reference Genome

$$P(\text{reads} | A/A) = 0.01$$

$$P(\text{reads} | A/C) = 0.50$$

$$P(\text{reads} | C/C) = 0.99$$

Possible Genotypes

Shotgun Sequence Data



AGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTG
GCTAGCTGATAGCTAGCTAGCTAGCTGATGAGCCCGA

Sequence Reads

5'-ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG-3'

Reference Genome

$$P(\text{reads} | A/A) = 0.0001$$

$$P(\text{reads} | A/C) = 0.25$$

$$P(\text{reads} | C/C) = 0.98$$

Possible Genotypes

Shotgun Sequence Data

ATGCTAGCTGATAGCTAGCTAGCTAGCTGATGAGCC
AGCTGATAGCTAGCTAGCTAGCTGATGAGCCCGATCGCTG
GCTAGCTGATAGCTAGCTAGCTAGCTGATGAGCCCGA

Sequence Reads

5'-ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG-3'

Reference Genome

$$P(\text{reads} | A/A) = 0.000001$$

$$P(\text{reads} | A/C) = 0.125$$

$$P(\text{reads} | C/C) = 0.97$$

Possible Genotypes

Shotgun Sequence Data



ATAGCTAG**A**TAGCTGATGAGCCCCGATCGCTGCTAGCTC
ATGCTAGCTGATAGCTAG**C**TAGCTGATGAGCC
AGCTGATAGCTAG**C**TAGCTGATGAGCCCCGATCGCTG
GCTAGCTGATAGCTAG**C**TAGCTGATGAGCCCCGA

Sequence Reads

5'-ACTGGTCGATGCTAGCTGATAGCTAG**C**TAGCTGATGAGCCCCGATCGCTGCTAGCTCGACG-3'

Reference Genome

$$P(\text{reads} | A/A) = 0.00000099$$

$$P(\text{reads} | A/C) = 0.0625$$

$$P(\text{reads} | C/C) = 0.0097$$

Possible Genotypes

Shotgun Sequence Data



TAGCTGATAGCTAG**A**TAGCTGATGAGCCCGAT
ATAGCTAG**A**TAGCTGATGAGCCCGATCGCTGCTAGCTC
ATGCTAGCTGATAGCTAG**C**TAGCTGATGAGCC
AGCTGATAGCTAG**C**TAGCTGATGAGCCCGATCGCTG
GCTAGCTGATAGCTAG**C**TAGCTGATGAGCCCGA

Sequence Reads

5'-ACTGGTCGATGCTAGCTGATAGCTAG**C**TAGCTGATGAGCCCGATCGCTGCTAGCTCGACG-3'

Reference Genome

$$P(\text{reads} | A/A) = 0.00000098$$

$$P(\text{reads} | A/C) = 0.03125$$

$$P(\text{reads} | C/C) = 0.000097$$

Possible Genotypes

How Low Coverage Analysis Works...



TAGCTGATAGCTAG**A**TAGCTGATGAGCCCGAT

ATAGCTAG**A**TAGCTGATGAGCCCGATCGCTGCTAGCTC

ATGCTAGCTGATAGCTAG**C**TAGCTGATGAGCC

AGCTGATAGCTAG**C**TAGCTGATGAGCCCGATCGCTG

GCTAGCTGATAGCTAG**C**TAGCTGATGAGCCCGA

Sequence Reads

5'-ACTGGTCGATGCTAGCTGATAGCTAG**C**TAGCTGATGAGCCCGATCGCTGCTAGCTCGACG-3'

Reference Genome

$P(\text{reads} | A/A) = 0.00000098$

$P(\text{reads} | A/C) = 0.03125$

$P(\text{reads} | C/C) = 0.000097$

Making a genotype call requires combining sequence data with prior information.

Individual Based Prior: Every site has 1/1000 probability of varying or so.

Population Based Prior: Use frequency information from examining others at the same site.

Haplotype Based Prior: Examine other chromosomes that are similar at locus of interest.

Sequence Based Genotype Calls

- **Individual Based Prior**
 - Assumes all sites have an equal probability of showing polymorphism
 - Specifically, assumption is that about 1/1000 bases differ from reference
 - If reads were error free and sampling Poisson ...
 - ... 14x coverage would allow for 99.8% genotype accuracy
 - ... 30x coverage of the genome needed to allow for errors and clustering

Sequence Based Genotype Calls

- **Individual Based Prior**
 - Assumes all sites have an equal probability of showing polymorphism
 - Specifically, assumption is that about 1/1000 bases differ from reference
 - If reads were error free and sampling Poisson ...
 - ... 14x coverage would allow for 99.8% genotype accuracy
 - ... 30x coverage of the genome needed to allow for errors and clustering
- **Population Based Prior**
 - Uses frequency information obtained from examining other individuals
 - Calling very rare polymorphisms still requires 20-30x coverage of the genome
 - Calling common polymorphisms requires much less data

Sequence Based Genotype Calls

- **Individual Based Prior**
 - Assumes all sites have an equal probability of showing polymorphism
 - Specifically, assumption is that about 1/1000 bases differ from reference
 - If reads were error free and sampling Poisson ...
 - ... 14x coverage would allow for 99.8% genotype accuracy
 - ... 30x coverage of the genome needed to allow for errors and clustering
- **Population Based Prior**
 - Uses frequency information obtained from examining other individuals
 - Calling very rare polymorphisms still requires 20-30x coverage of the genome
 - Calling common polymorphisms requires much less data
- **Haplotype Based Prior or Imputation Based Analysis**
 - Compares individuals with similar flanking haplotypes
 - Calling very rare polymorphisms still requires 20-30x coverage of the genome
 - Can make accurate genotype calls with 2-4x coverage of the genome
 - Accuracy improves as more individuals are sequenced

Simulation Results: Common Sites

- Detection and genotyping of Sites with MAF >5% (2116 simulated sites/Mb)
 - **Detected Polymorphic Sites: 2x coverage**
 - 100 people 2102 sites/Mb detected
 - 200 people 2115 sites/Mb detected
 - 400 people 2116 sites/Mb detected
 - **Error Rates at Detected Sites: 2x coverage**
 - 100 people 98.5% accurate, 90.6% at hets
 - 200 people 99.6% accurate, 99.4% at hets
 - 400 people 99.8% accurate, 99.7% at hets

Simulation Results: Rarer Sites

- Detection and genotyping of Sites with MAF 1-2% (425 simulated sites/Mb)
 - **Detected Polymorphic Sites: 2x coverage**
 - 100 people 139 sites/Mb detected
 - 200 people 213 sites/Mb detected
 - 400 people 343 sites/Mb detected
 - **Error Rates at Detected Sites: 2x coverage**
 - 100 people 98.6% accurate, 92.9% at hets
 - 200 people 99.4% accurate, 95.0% at hets
 - 400 people 99.6% accurate, 95.9% at hets

Shallow Sequencing Great in Simulations...

What About in Practice?

- **Predictions: Detection Rate, 2x coverage**

- 100 people 99.3% of sites with MAF > 5%
- 200 people 99.9% of sites with MAF > 5%
- 400 people >99.9% of sites with MAF > 5%

- **Predictions: Accuracy, 2x coverage**

- 100 people 98.5% accuracy, 90.6% at hets
- 200 people 99.6% accuracy, 99.4% at hets
- 400 people 99.8% accuracy, 99.7% at hets

- **Predictions: 60 individuals Matching Observed Depths**

- 91% accuracy at heterozygous sites
- 98% accuracy at homozygous sites

Whole Genome Sequencing Studies

- Suppose we could afford 2,000x data (6,000 GB)
- We could sequence 67 individuals at 30x

Sequencing of 67 individuals at 30x depth

Minor Allele Frequency	0.5 – 1.0%	1.0 – 2.0%	2.0 – 5.0%	>5%
Proportion of Detected Sites	59.3%	90.1%	96.9%	100.0%
Genotyping Accuracy	100.0%	100.0%	100.0%	100.0%
.... Heterozygous Sites Only	100.0%	100.0%	100.0%	100.0%
Correlation with Truth (r^2)	99.8%	99.9%	99.9%	100.0%
Effective Sample Size ($n \cdot r^2$)	67	67	67	67

Whole Genome Sequencing Studies

- Suppose we could afford 2,000x data (6,000 GB)
- We could sequence 1000 individuals at 2x

Sequencing of 1000 individuals at 2x depth				
Minor Allele Frequency	0.5 – 1.0%	1.0 – 2.0%	2.0 – 5.0%	>5%
Proportion of Detected Sites	79.6%	98.8%	100.0%	100.0%
Genotyping Accuracy	99.6%	99.5%	99.5%	99.8%
.... Heterozygous Sites Only	78.8%	89.5%	95.9%	99.8%
Correlation with Truth (r^2)	56.7%	76.1%	88.2%	97.8%
Effective Sample Size ($n \cdot r^2$)	567	761	882	978

Whole Genome Sequencing Studies

- Suppose we could afford 2,000x data (6,000 GB)
- We could sequence 1000 exomes at 100x
- How much enrichment of functional variants should we expect in exons?
 - For rare Mendelian variants, extreme enrichment ...
 - For common variants, enrichment appears mild ...
- Hybrid that combines deep exome re-sequencing and shallow examination of rest of genome may emerge

Sequencing Thousands of Genomes, Status

- 1000 Genomes Consortium will deliver ~2,000 genomes into public domain next year
 - Compared to <10 genomes by 2008 and ~180 human genomes examined so far in 2009
- Current technology imposes trade-off between number of individuals sequenced and completeness of each sequence
- What are the right questions to ask?

Acknowledgements

University of Michigan

- Mike Boehnke
- Cristen Willer
- **Wei Chen**
- **Paul Anderson**
- **Tom Blackwell**

- Paul Scheet (now MD Anderson)
- Serena Sanna (now Cagliari)
- Weimin Chen (now Virginia)
- **Yun Li (now North Carolina)**
- **Liming Liang (now Harvard)**

Around the World

- David Schlessinger
- Manuela Uda
- Antonio Cao
- Mark McCarthy
- **Karen Mohlke**
- **Sekar Kathiresan**
- **Joel Hirschhorn**
- Bill Cookson
- Mark Lathrop
- **Richard Durbin**
- **Anand Swaroop**
- Dwight Stambolian
- Albert Edwards
- (... many others ...)