

Low Pass Sequencing

Gonçalo Abecasis

University of Michigan School of Public Health

Genomewide Association Studies

- Survey 500,000 SNPs in a large sample
- 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, SardiNIA project has >25 publications and counting!
- Still, many questions remain unanswered.

Questions that Might Be Answered With Complete Sequence Data...

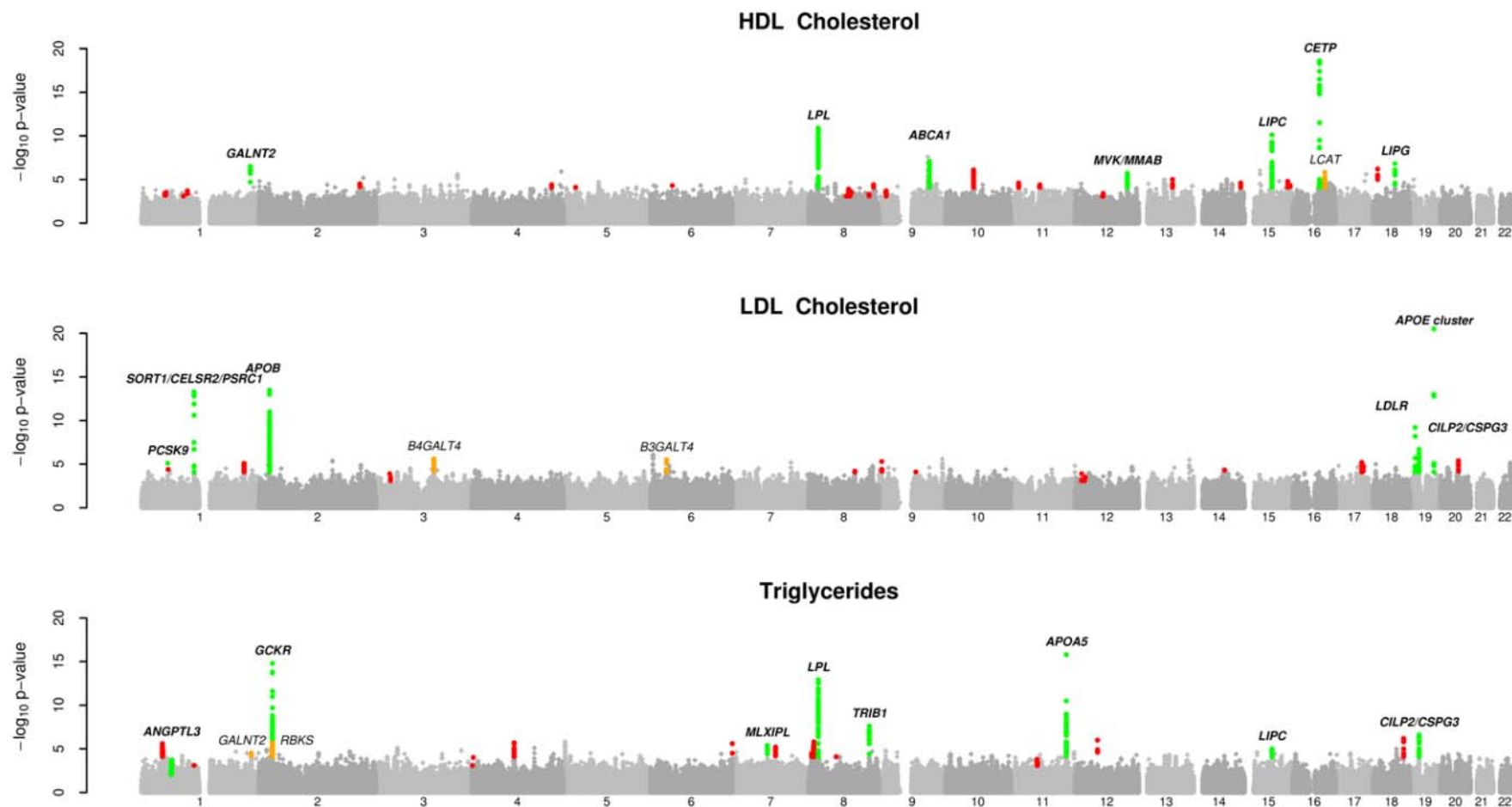
- What is the contribution of each identified locus to a trait?
 - Likely that multiple variants, common and rare, will contribute
- What is the mechanism? What happens when we knockout a gene?
 - Most often, the causal variant will not have been examined directly
 - Rare coding variants will provide 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

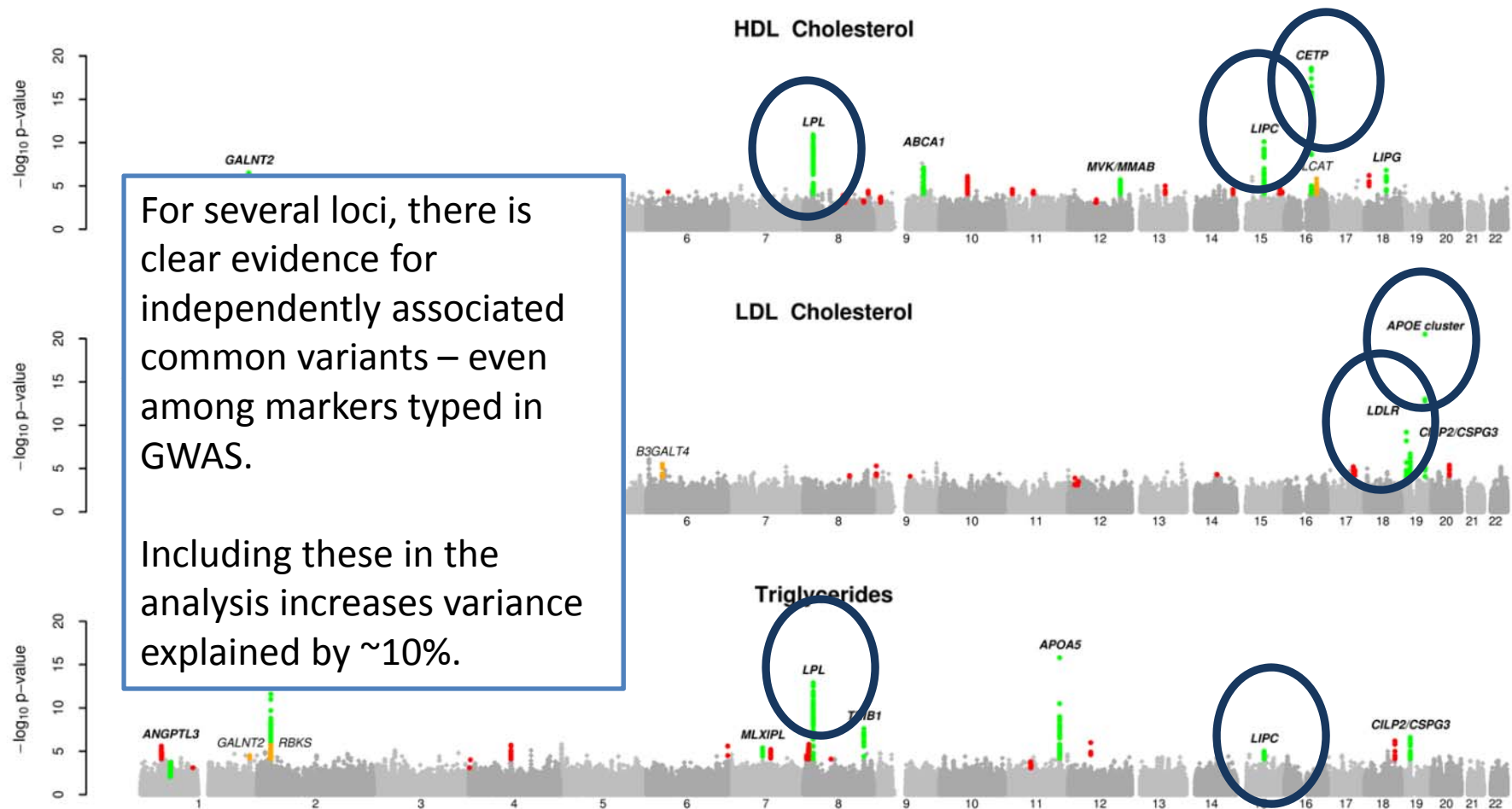
Example from Lipid Biology



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

Evidence for Multiple Variants Per Locus

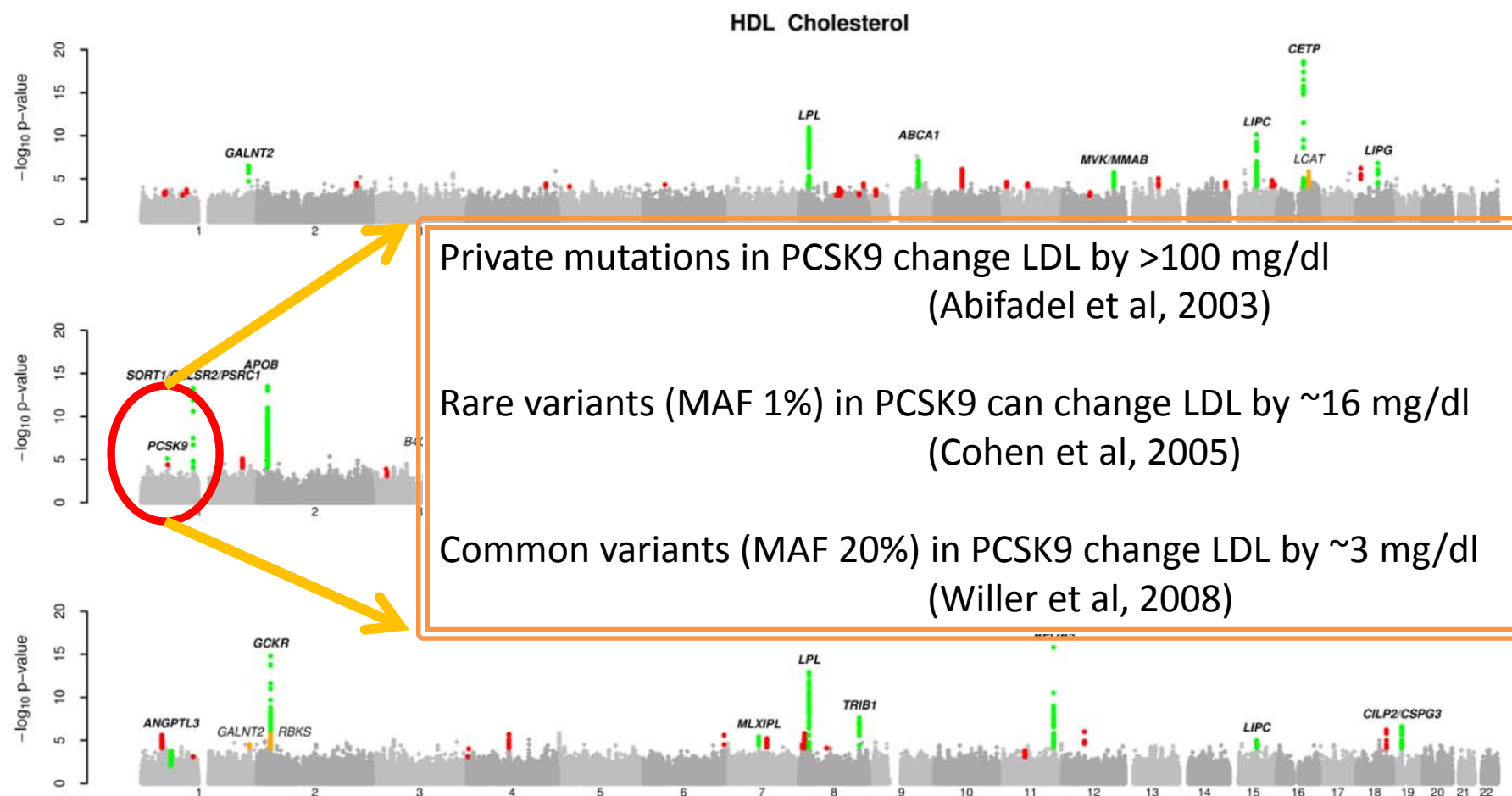
Example from Lipid Biology



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

Evidence for Multiple Variants Per Locus

Example from Lipid Biology



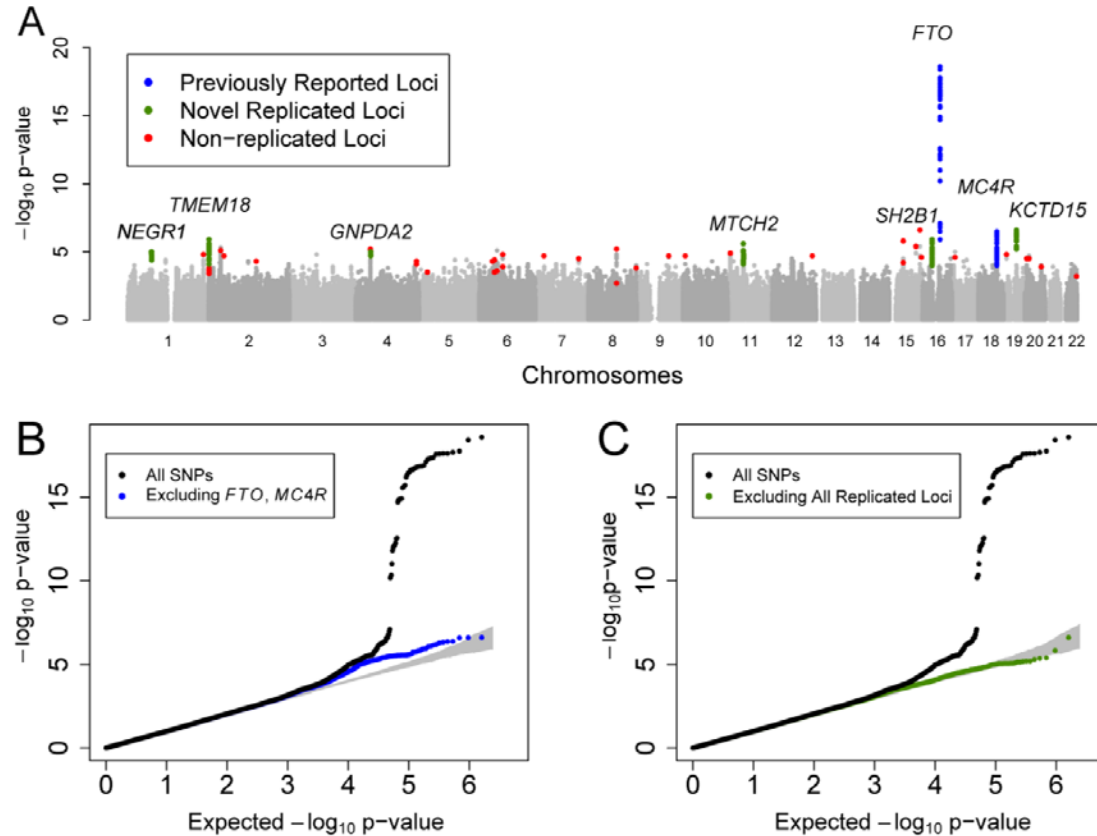
Willer et al, *Nat Genet*, 2008
Kathiresan et al, *Nat Genet*, 2008, 2009

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

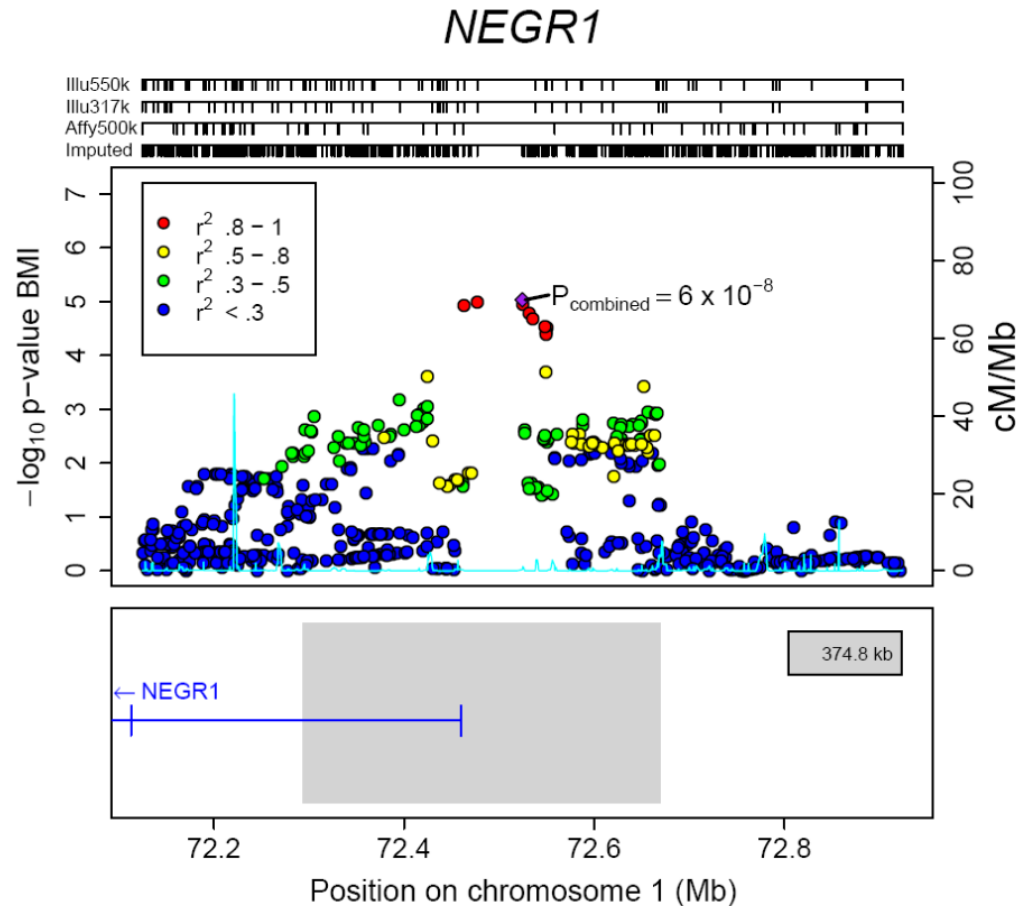


Seven of eight confirmed BMI loci show strongest expression in the brain...

Willer et al, *Nature Genetics*, 2009

Evidence that Copy Number Variants Important

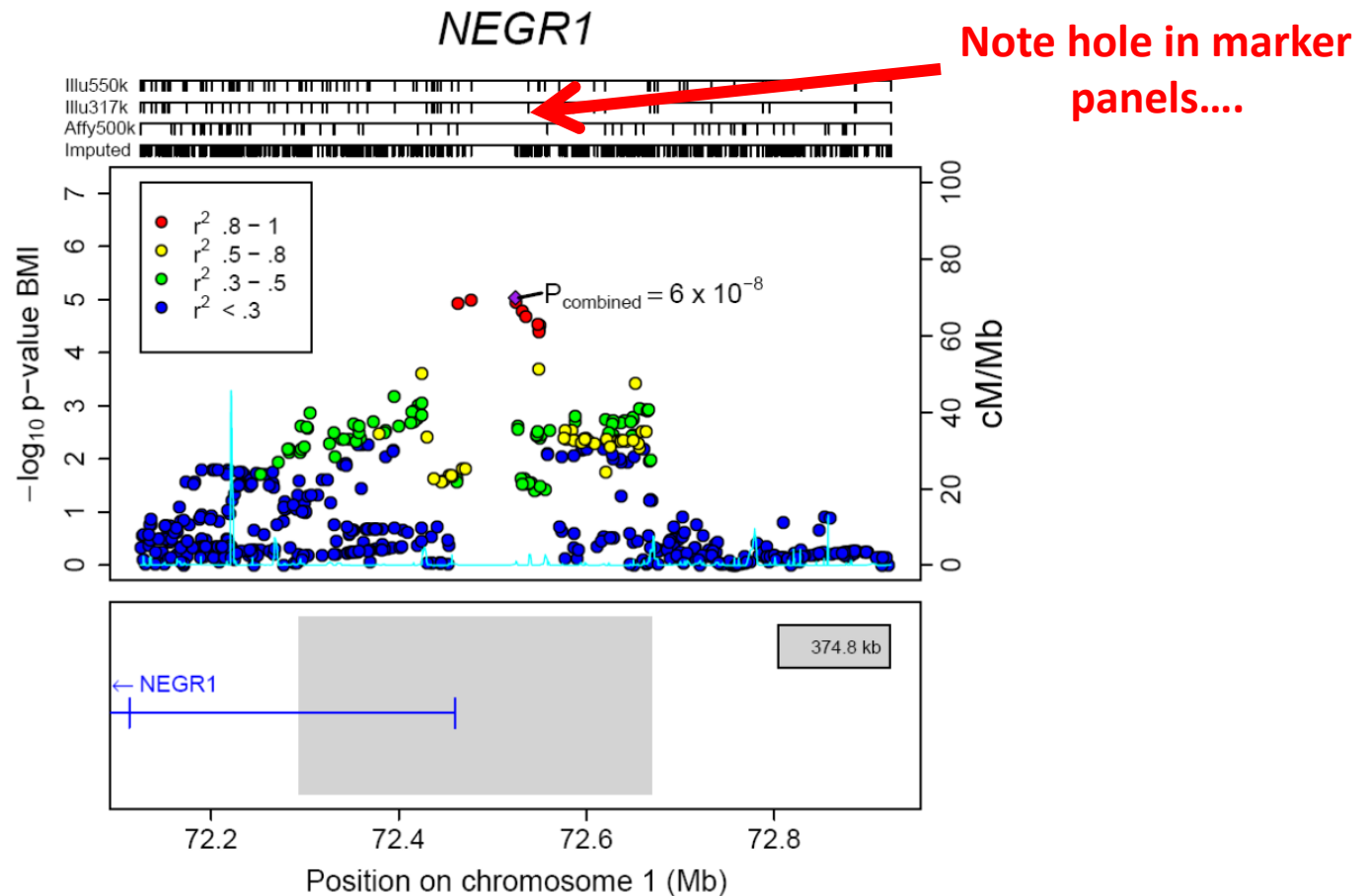
Example from Genetics of Obesity



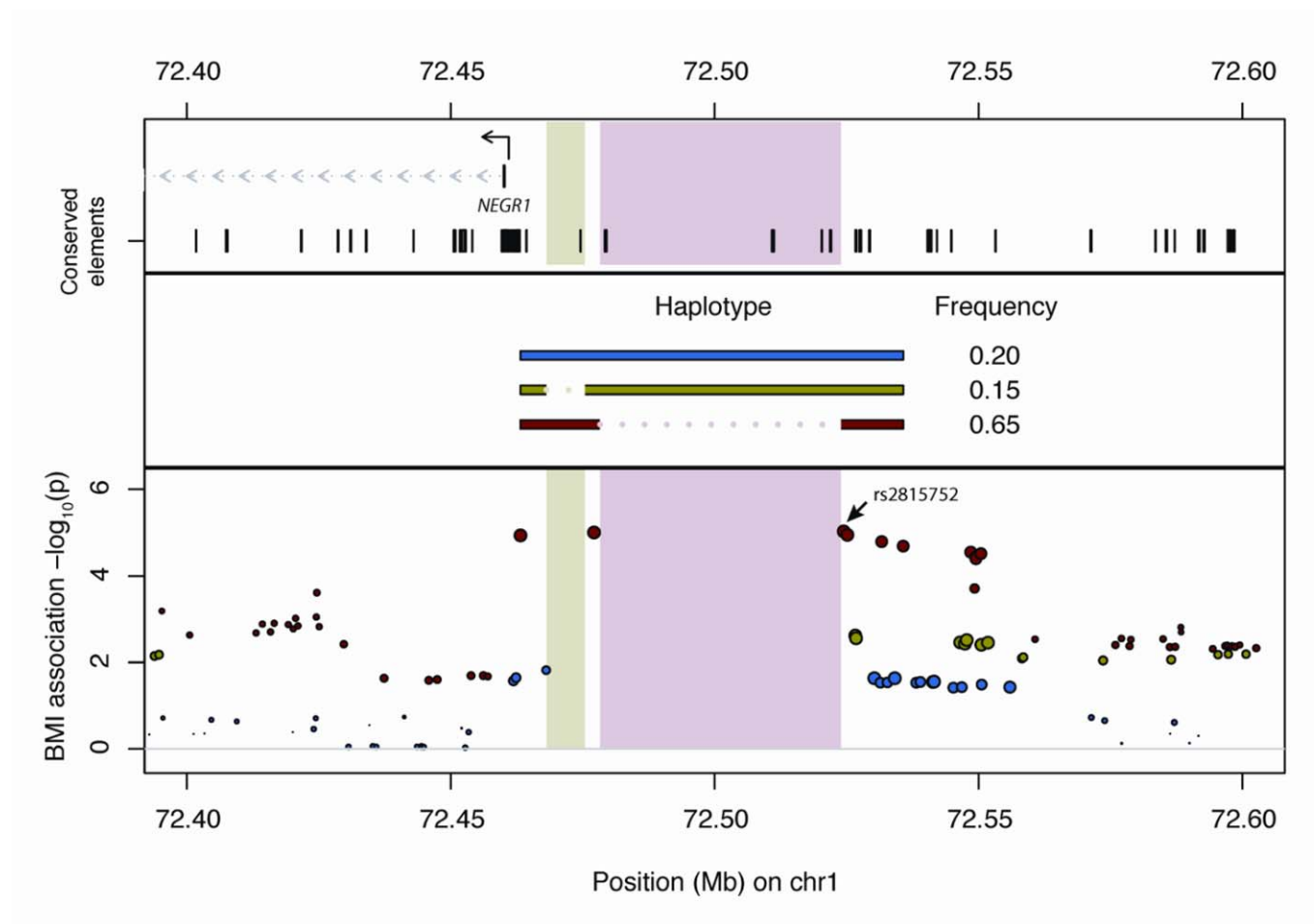
Willer et al, *Nature Genetics*, 2009

Evidence that Copy Number Variants Important

Example from Genetics of Obesity



Associated Haplotype Carries Deletion



Willer et al, *Nature Genetics*, 2009

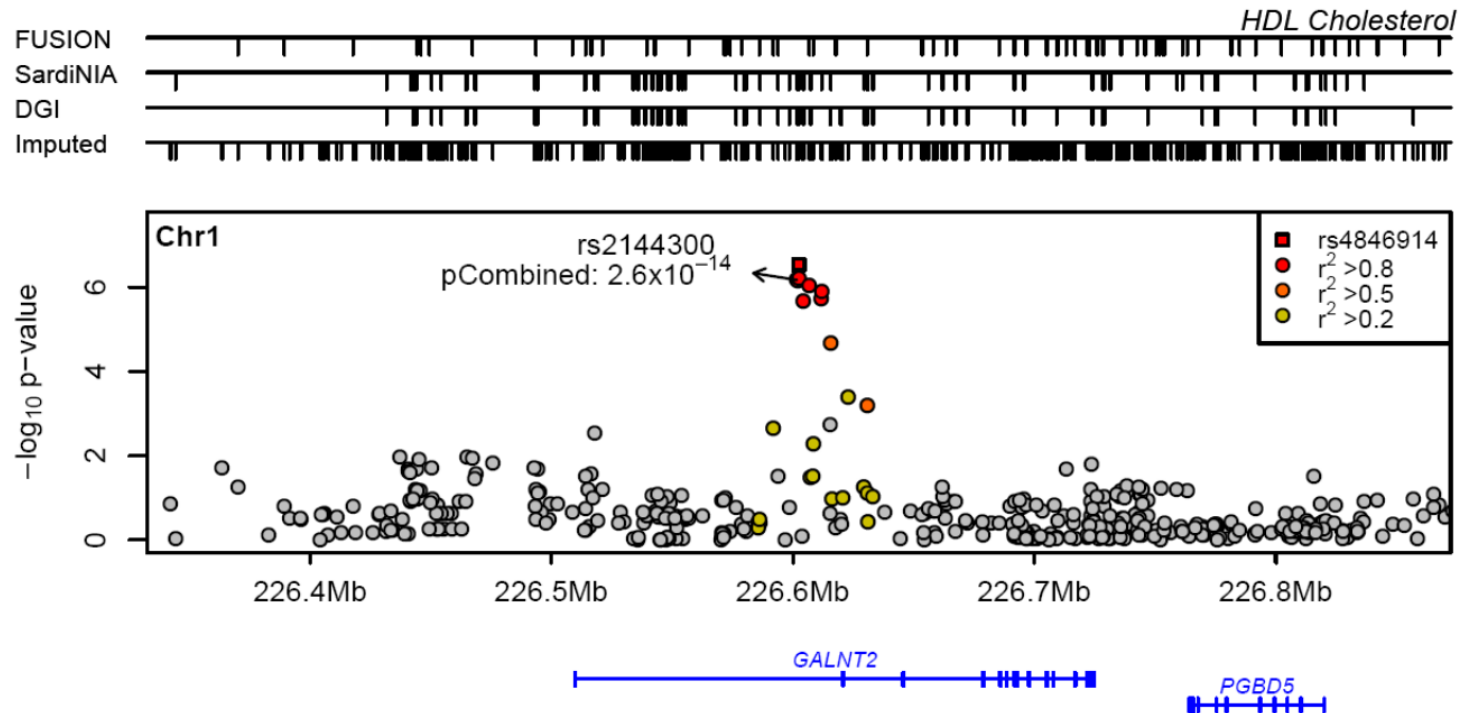
What is the Mechanism?

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 1 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

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 Reads



ACTGGTCGATGCTAGCTGATAGCTAGCTA
GCTGATGAGCCCGATCGCTGCTAGCTCG
AGCTGATAGCTAGCTAGCTGATGAGCCCGA
GAGCCCGATCGCTGCTAGCTCGACG

- Typical short read might be <25-100 bp long and not very informative on its own
- Reads must be arranged (*aligned*) relative to each other to reconstruct longer sequences

Read Alignment

GCTAGCTGATAGCTAGCTAGCTGATGAGCCCGA

Short Read (30-100 bp)

5'-ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG-3'

Reference Genome (3,000,000,000 bp)

- The first step in analysis of human short read data is to align each read to genome, typically using a hash table based indexing procedure
- This process now takes no more than a few hours per million reads ...
- Analyzing these data without a reference human genome would require much longer reads or result in very fragmented assemblies

Calling Consensus Genotype - Details

- Each aligned read provides a small amount of evidence about the underlying genotype
 - Read may be consistent with a particular genotype ...
 - Read may be less consistent with other genotypes ...
 - A single read is never definitive
- This evidence is cumulated gradually, until we reach a point where the genotype can be called confidently
- I will next outline a simple approach ...

Shotgun Sequence Data



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

Sequence Reads

5'-ACTGGTCGATGCTAGCTGATAGCTAG**C**TAGCTGATGAGCCCGATCGCTGCTAGCTCG**A**CG-3'

Reference Genome

A/C

Predicted Genotype

Shotgun Sequence Data

Sequence Reads

5'-ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG-3'

Reference Genome

$P(\text{reads} \mid \text{A/A, read mapped}) = 1.0$

$P(\text{reads} \mid \text{A/C, read mapped}) = 1.0$

$P(\text{reads} \mid \text{C/C, read mapped}) = 1.0$

Possible Genotypes

Shotgun Sequence Data

GCTAGCTGATAGCTAGCTAGCTAGCTGATGAGCCCGA

Sequence Reads

5'-ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG-3'

Reference Genome

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

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

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

Possible Genotypes

Shotgun Sequence Data

GCTAGCTGATAGCTAGCTAGCTAGCTGATGAGCCCGA

Sequence Reads

5'-ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG-3'

Reference Genome

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

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

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

Possible Genotypes

Shotgun Sequence Data


AGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTG
GCTAGCTGATAGCTAGCTAGCTAGCTGATGAGCCCGA

Sequence Reads

5'-ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG-3'

Reference Genome

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

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

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

Possible Genotypes

Shotgun Sequence Data

ATGCTAGCTGATAGCTAGCTAGCTAGCTGATGAGCC
AGCTGATAGCTAGCTAGCTAGCTGATGAGCCCGATCGCTG
GCTAGCTGATAGCTAGCTAGCTAGCTGATGAGCCCGA

Sequence Reads

5'-ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG-3'

Reference Genome


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

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

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

Possible Genotypes

Shotgun Sequence Data


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

Sequence Reads

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

Reference Genome

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

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

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

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

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

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

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

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

$$P(\text{Genotype}|\text{reads}) = \frac{P(\text{reads}|\text{Genotype})\text{Prior}(\text{Genotype})}{\sum_G P(\text{reads}|G)\text{Prior}(G)}$$

Combine these likelihoods with a prior incorporating information from other individuals and flanking sites to assign a genotype.

Ingredients That Go Into Prior

- Most sites don't vary
 - $P(\text{non-reference base}) \sim 0.001$
- When a site does vary, it is usually heterozygous
 - $P(\text{non-reference heterozygote}) \sim 0.001 * 2/3$
 - $P(\text{non-reference homozygote}) \sim 0.001 * 1/3$
- Mutation model
 - Transitions account for most variants ($C \leftrightarrow T$ or $A \leftrightarrow G$)
 - Transversions account for minority of variants

Shotgun Sequence Data

Individual Based Prior



TAGCTGATAGCTAG**A**TAGCTGATGAGCCCGAT

ATAGCTAG**A**TAGCTGATGAGCCCGATCGCTGCTAGCTC

ATGCTAGCTGATAGCTAG**C**TAGCTGATGAGCC

AGCTGATAGCTAG**C**TAGCTGATGAGCCCGATCGCTG

GCTAGCTGATAGCTAG**C**TAGCTGATGAGCCCGA

Sequence Reads

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

Reference Genome

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

$\text{Prior}(A/A) = 0.00034$

$\text{Posterior}(A/A) = <.001$

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

$\text{Prior}(A/C) = 0.00066$

$\text{Posterior}(A/C) = 0.175$

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

$\text{Prior}(C/C) = 0.99900$

$\text{Posterior}(C/C) = 0.825$

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

Shotgun Sequence Data

Individual Based Prior



TAGCTGATAGCTAG**A**TAGCTGATGAGCCCGAT

ATAGCTAG**A**TAGCTGATGAGCCCGATCGCTGCTAGCTC

ATGCTAGCTGATAGCTAG**C**TAGCTGATGAGCC

AGCTGATAGCTAG**C**TAGCTGATGAGCCCGATCGCTG

GCTAGCTGATAGCTAG**C**TAGCTGATGAGCCCGA

Sequence Reads

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

Reference Genome

$P(\text{reads} | A/A) = 0.00000098$ $\text{Prior}(A/A) = 0.00034$ $\text{Posterior}(A/A) = <.001$

$P(\text{reads} | A/C) = 0.03125$ $\text{Prior}(A/C) = 0.00066$ $\text{Posterior}(A/C) = 0.175$

$P(\text{reads} | C/C) = 0.000097$ $\text{Prior}(C/C) = 0.99900$ $\text{Posterior}(C/C) = 0.825$

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

Shotgun Sequence Data

Population Based Prior



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

Sequence Reads

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

Reference Genome

$P(\text{reads} | A/A) = 0.00000098$ **Prior(A/A) = 0.04** $\text{Posterior(A/A)} = <.001$

$P(\text{reads} | A/C) = 0.03125$ **Prior(A/C) = 0.32** $\text{Posterior(A/C)} = 0.999$

$P(\text{reads} | C/C) = 0.000097$ **Prior(C/C) = 0.64** $\text{Posterior(C/C)} = <.001$

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

In the example above, we estimated $P(A) = 0.20$

Shotgun Sequence Data

Population Based Prior



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

Sequence Reads

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

Reference Genome

$P(\text{reads} | A/A) = 0.00000098$ $\text{Prior}(A/A) = 0.04$

$\text{Posterior}(A/A) = <.001$

$P(\text{reads} | A/C) = 0.03125$ $\text{Prior}(A/C) = 0.32$

$\text{Posterior}(A/C) = 0.999$

$P(\text{reads} | C/C) = 0.000097$ $\text{Prior}(C/C) = 0.64$

$\text{Posterior}(C/C) = <.001$

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

In the example above, we estimated $P(A) = 0.20$

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

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

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

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

Combine these likelihoods with a prior incorporating information from other individuals and flanking sites to assign a genotype.

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**TAGCTGATGAGCCCGATCGCTGCTAGCTCG**ACG**-3'

Reference Genome

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

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

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

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

Recipe For Imputation With Shotgun Sequence Data

- Start with some plausible configuration for each individual
- Use Markov model to update one individual conditional on all others
- Repeat previous step many times
- Generate a consensus set of genotypes and haplotypes for each individual

Silly Cartoon View of Shot Gun Data

[illegible]

Cartoon View of Shot Gun Data

c	G	a	G	A	t	c	T	c	C	t	T	c	T	t	c	t	g	T	G	c
C	g	A	g	a	t	C	T	C	C	C	g	a	c	C	t	c	a	t	g	g
C	C	A	a	G	c	t	C	T	t	t	t	c	t	t	c	t	g	T	G	c
c	g	a	a	g	c	t	C	T	T	T	t	C	t	t	c	t	g	t	g	c
c	g	a	g	a	c	T	c	t	C	c	g	A	C	C	t	t	A	T	G	c
t	g	g	g	a	t	C	t	C	C	c	G	A	C	C	t	C	A	t	G	G
C	G	A	g	A	t	c	t	c	c	c	G	a	C	c	t	T	g	T	g	c
c	g	a	g	a	c	t	C	t	T	t	T	c	t	t	t	t	g	t	A	c
C	G	a	g	A	c	t	C	T	c	c	g	a	c	C	T	c	G	t	g	c
C	G	A	A	g	c	T	c	t	T	t	T	c	T	t	C	T	g	t	G	C
c	G	A	g	A	T	C	t	c	C	t	T	c	T	T	c	t	g	t	G	c
c	g	A	g	a	t	c	t	c	C	C	g	A	C	c	T	C	A	T	G	g
c	c	A	a	G	c	t	C	t	T	T	t	c	t	T	c	T	G	t	G	C
C	G	A	a	g	c	T	c	t	T	t	t	c	T	T	c	T	g	t	G	C
c	g	a	G	A	C	t	C	t	c	c	g	a	c	c	t	t	a	T	G	c
T	g	g	g	a	T	c	t	C	c	c	g	a	C	C	t	c	a	t	g	g
c	g	a	G	A	T	C	t	C	C	c	G	a	c	C	T	T	g	t	G	C
c	g	a	G	A	c	T	c	T	T	t	T	c	T	T	t	T	g	t	a	c
c	G	A	G	a	c	T	c	T	c	c	G	A	c	c	T	C	G	t	g	C
c	g	A	A	g	c	T	c	t	t	t	t	c	t	t	c	t	g	t	G	c

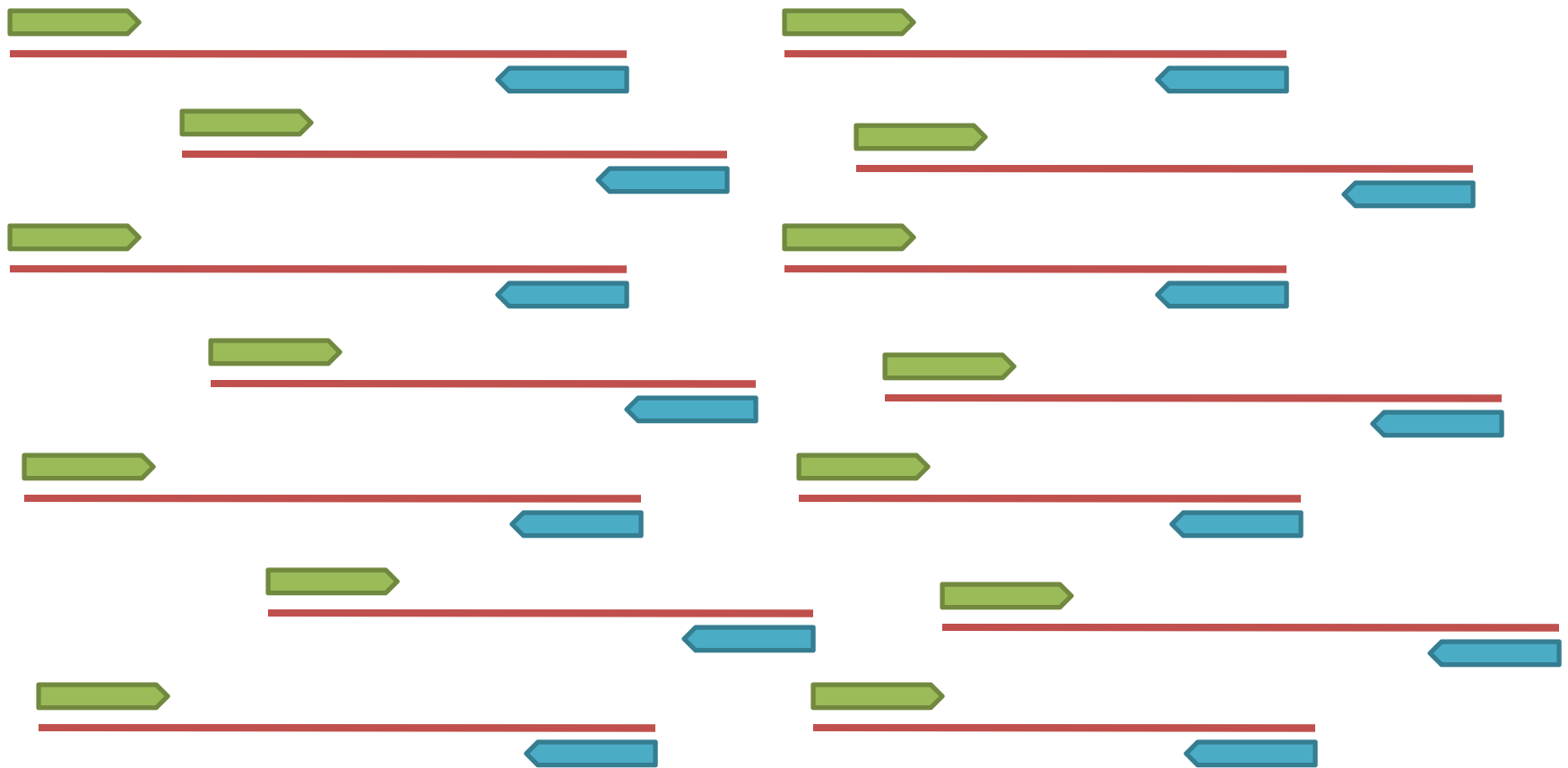
Simulation Results: Common Sites

- 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

Paired End Sequencing



Population of DNA fragments of known size (mean + stdev)
Paired end sequences

Paired End Sequencing

Paired Reads



Initial alignment to the reference genome



Paired end resolution

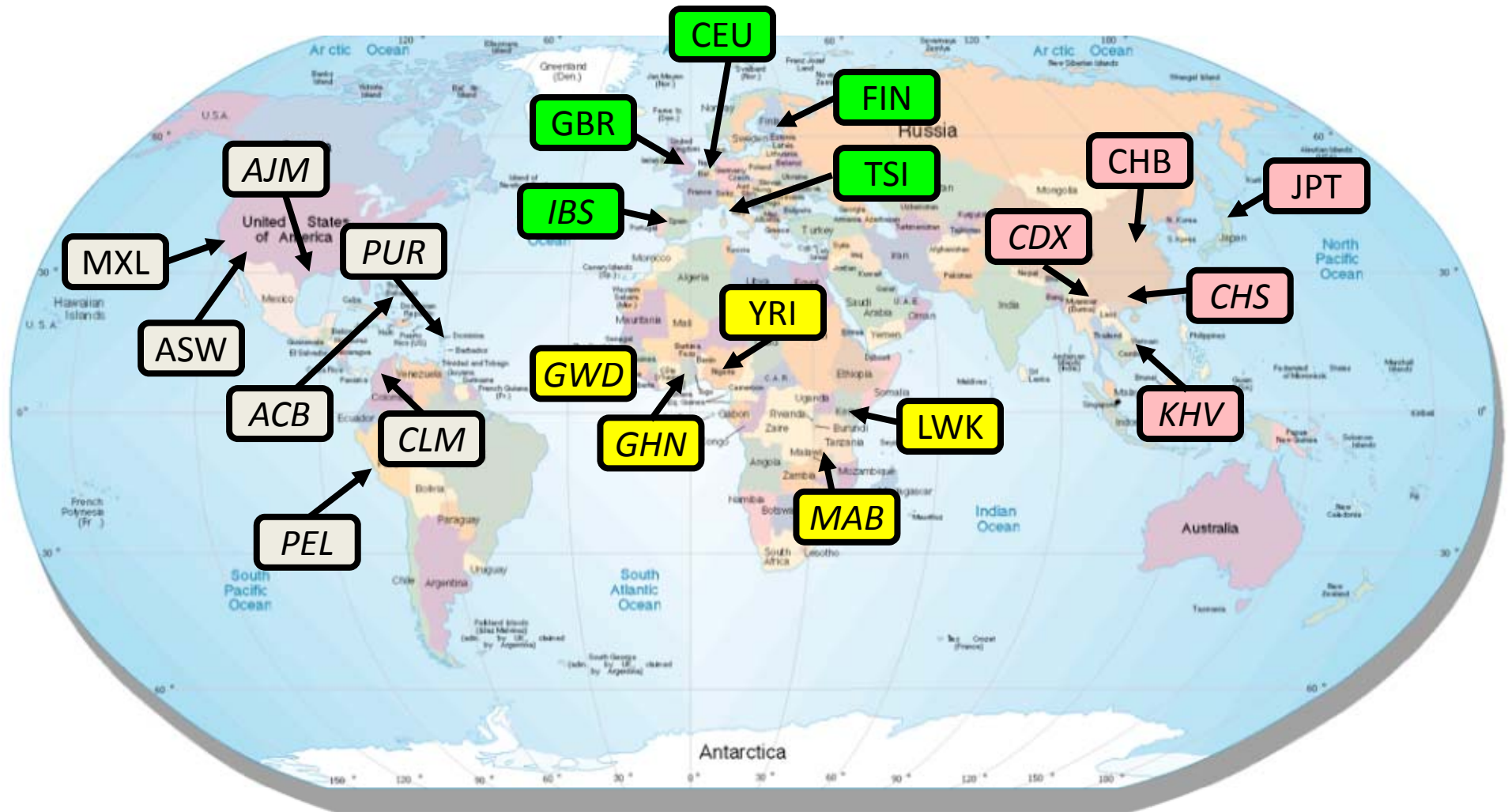


Detecting Structural Variation

- Read depth
 - Regions where depth is different from expected
 - Expectation defined by comparing to rest of genome ...
 - ... or, even better, by comparing to other individuals
- Split reads
 - If reads are longer, it may be possible to find reads that span the structural variation
- Discrepant pairs
 - If we find pairs of reads that appear to map significantly closer or further apart than expected, could indicate an insertion or deletion
 - For this approach, “physical coverage” which is the sum of read length and insert size is key
- De Novo Assembly

1000 Genomes Project: Initial Analysis of Pilot Datasets

1,135 samples at 4x in 2009/10
(this will later expand to 2,000 samples)



Major population groups comprised of subpopulations of ~100 individuals each

1000 Genomes Project: Pilots

- Pilot 1: 2-4x coverage of 180 people
- Pilot 2: 20x coverage of 2 trios
- Pilot 3: targeted sequencing of 1000 genes

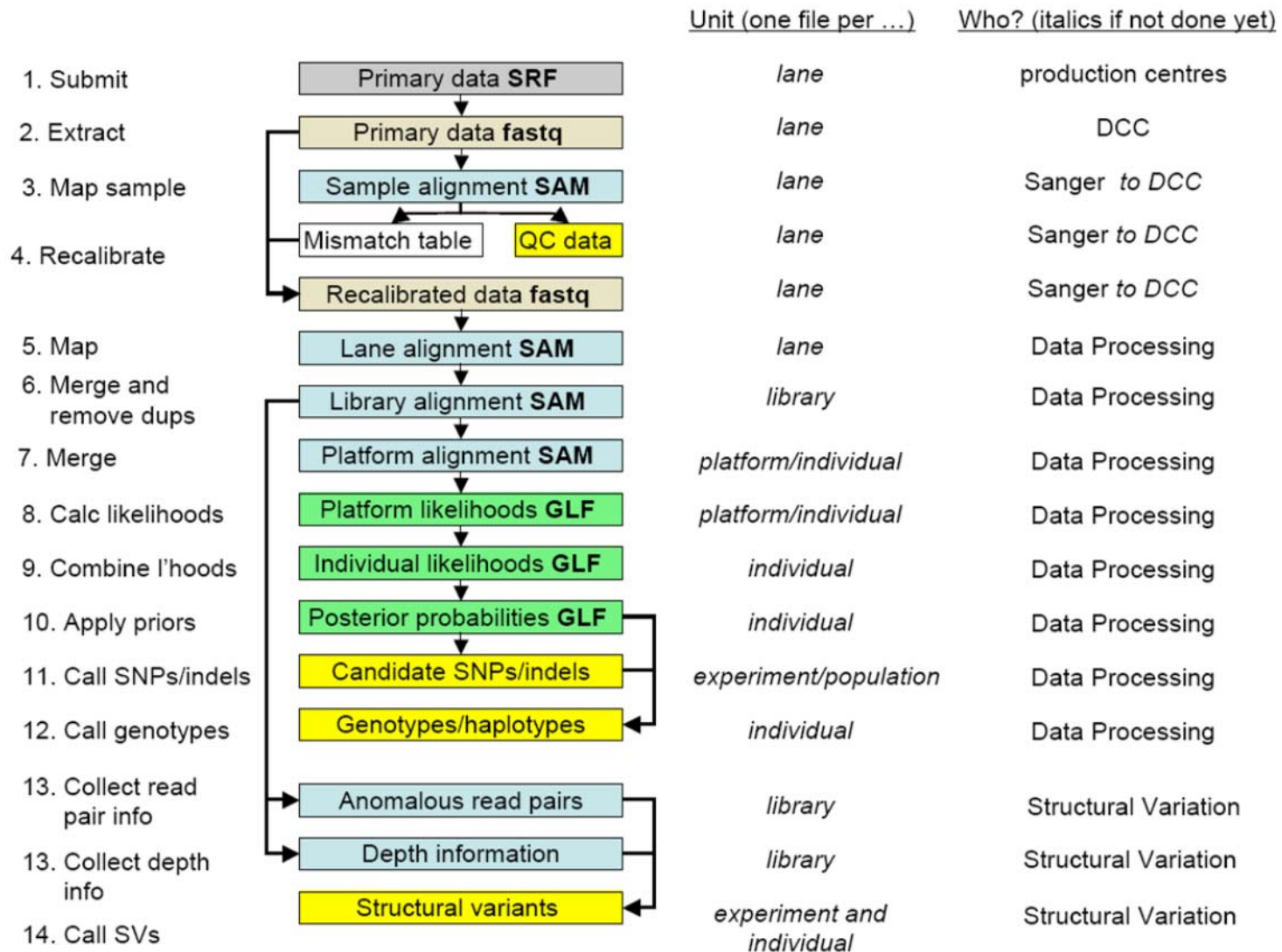
- December 2008: Initial trio analysis (including 340 Gb of sequence)
- January 2009: Initial analysis of low coverage samples (576 Gb)

- 11,479,146 unique SNPs
 - 6,405,006 SNPs already in dbSNP 129
 - 5,074,140 new SNPs deposited into dbSNP

- May 2009: Updated trio analysis (700 Gb)
- May 2009: Updated analysis of low coverage samples (1.9 Tb)

- <ftp://ftp.1000genomes.ebi.ac.uk/>
- <ftp://ftp-trace.ncbi.nih.gov/1000genomes/>

1000 Genome Projects: Data Processing



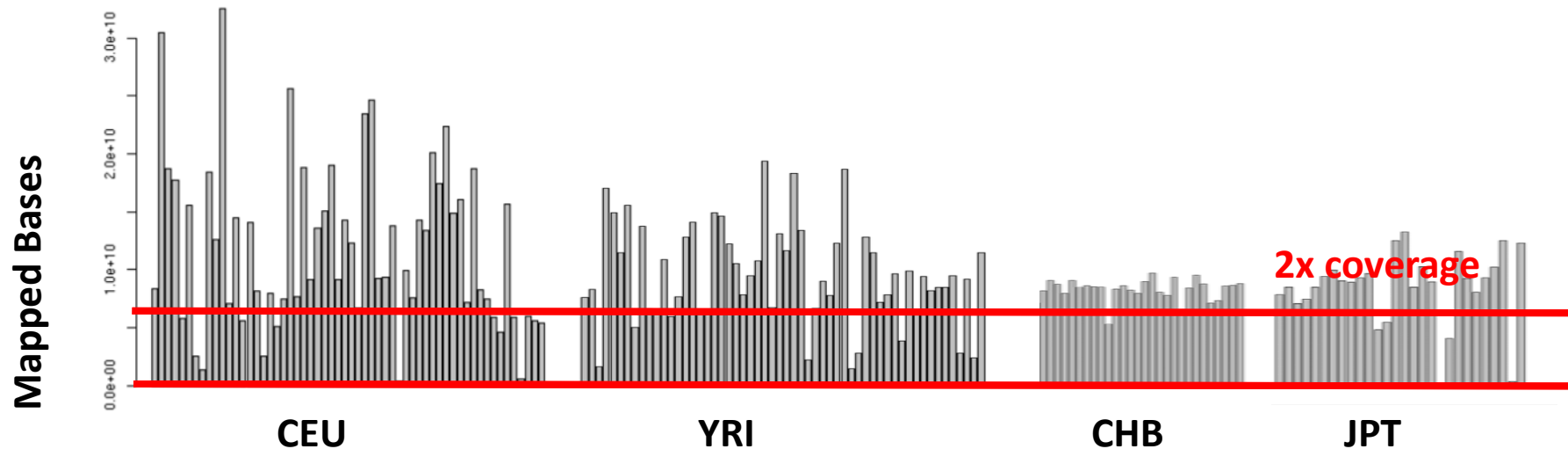
Slide courtesy Richard Durbin

1000 Genomes Project: Deeply Sequenced Trio (CEU)

- NA12878 (child) sequenced to 65x depth (33x Illumina, 20x SOLiD, 12x 454)
 - Parents sequenced to 26x, 33x
- Calls made at 90.5% of all sites in the reference genome (Q30)
 - Depth filter excludes ~3% of genome
 - Map quality filter excludes ~6% of genome
- 2,985,516 non-reference calls in NA12878
- Where are calls being made?
 - 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!)

Individuals Sequenced at Low Depth

- In addition to the two trios, sequence data now available for 178 individuals
- These samples typically have 2-4x sequence depth and are, individually, less informative
- However, combined analyses of the sample set can be very informative



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%

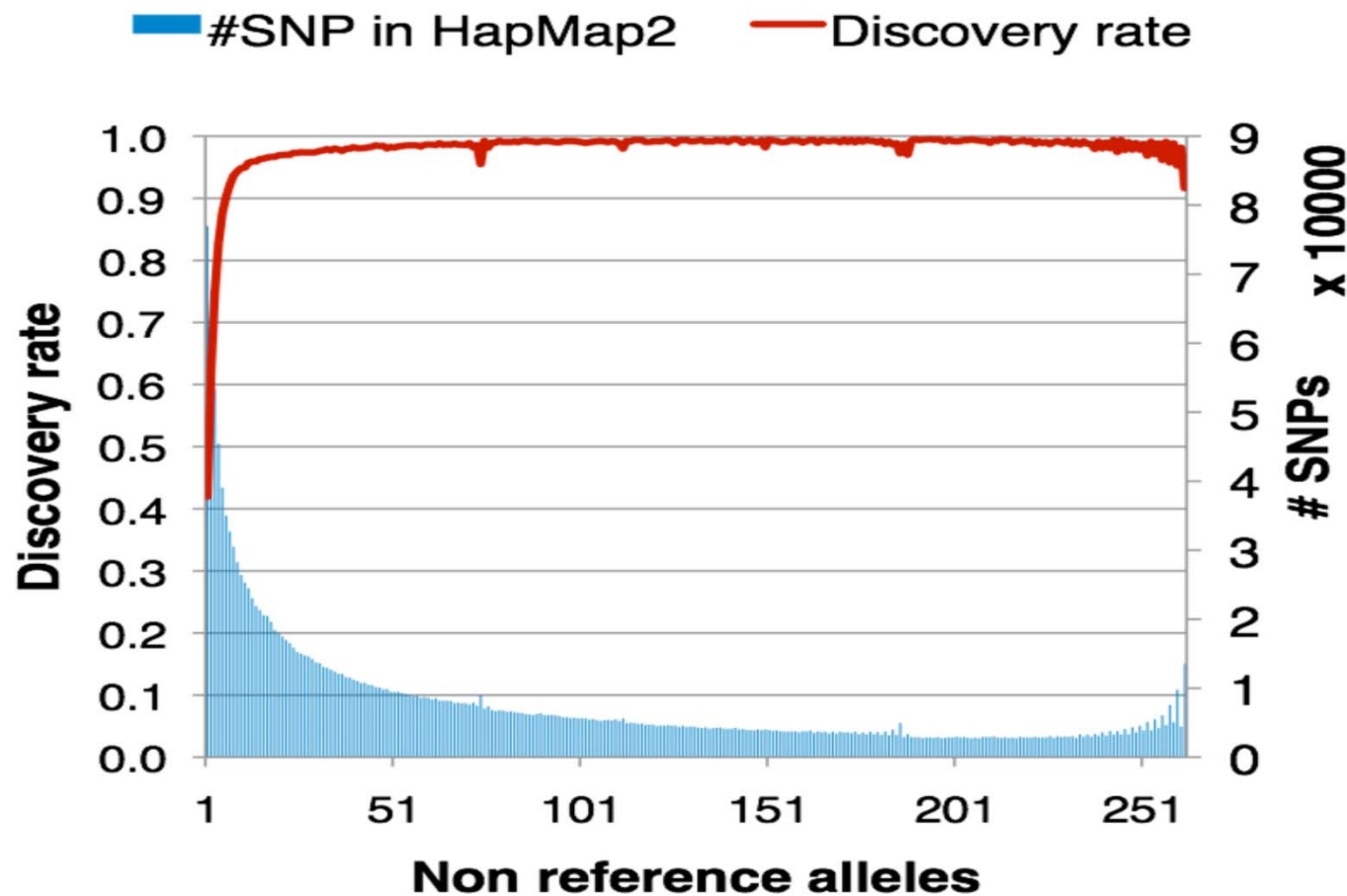
- **Predictions: Accuracy, 2x coverage**

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

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

- 91% accurate at heterozygous sites (Actual: 92%)
- 98% accurate at homozygous sites (Actual: 98%)

Discovering Most Alleles That Occur >5 Times in Sequenced Samples



Implications for 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

Implications for 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