

Whole Genome Sequencing

Biostatistics 666

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, ~50 genes now identified for type 2 diabetes.
- Techniques for genetic analysis are changing rapidly
 - What are some of the potential benefits and challenges for replacing genotyping with sequencing in complex trait studies?

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

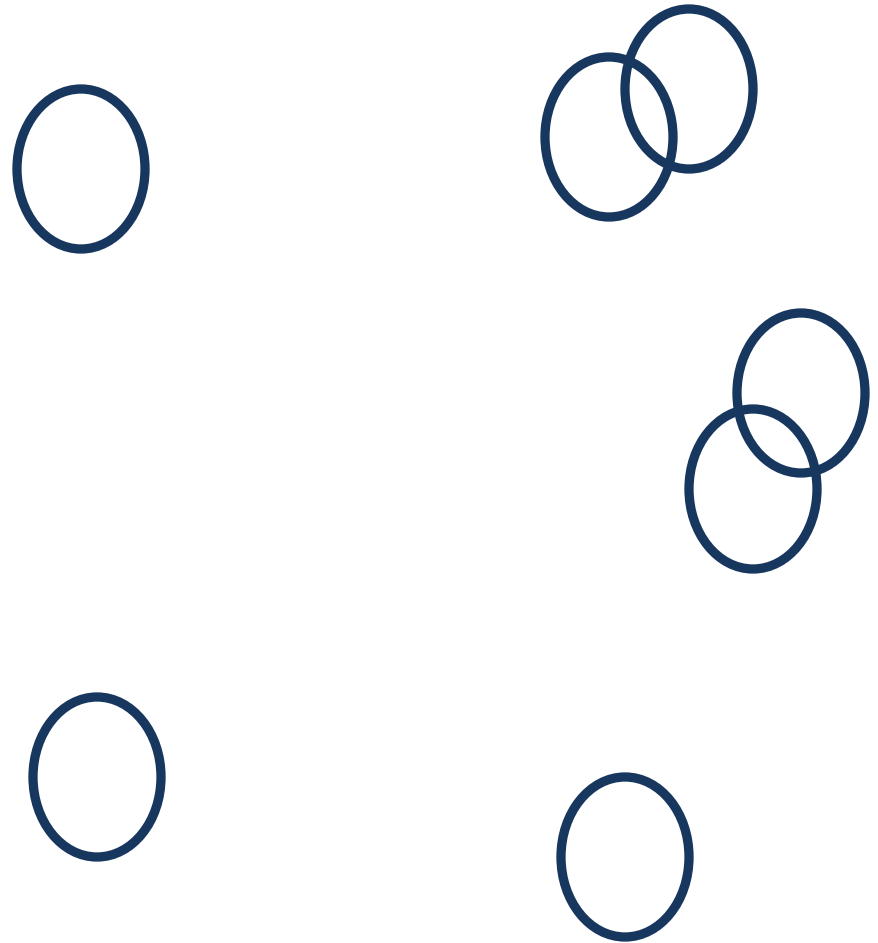
Kathiresan et al, *Nat Genet*, 2008, 2009

Evidence for Multiple Variants Per Locus

Example from Lipid Biology

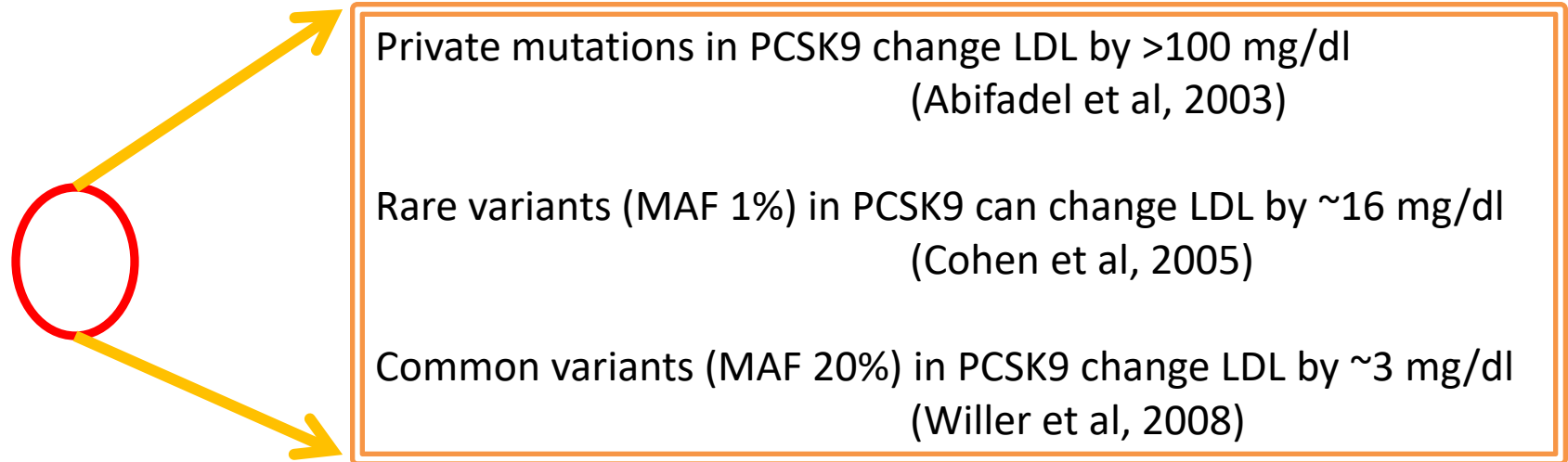
For several loci, there is clear evidence for independently associated common variants – even among markers typed in GWAS.

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



Evidence for Multiple Variants Per Locus

Example from Lipid Biology

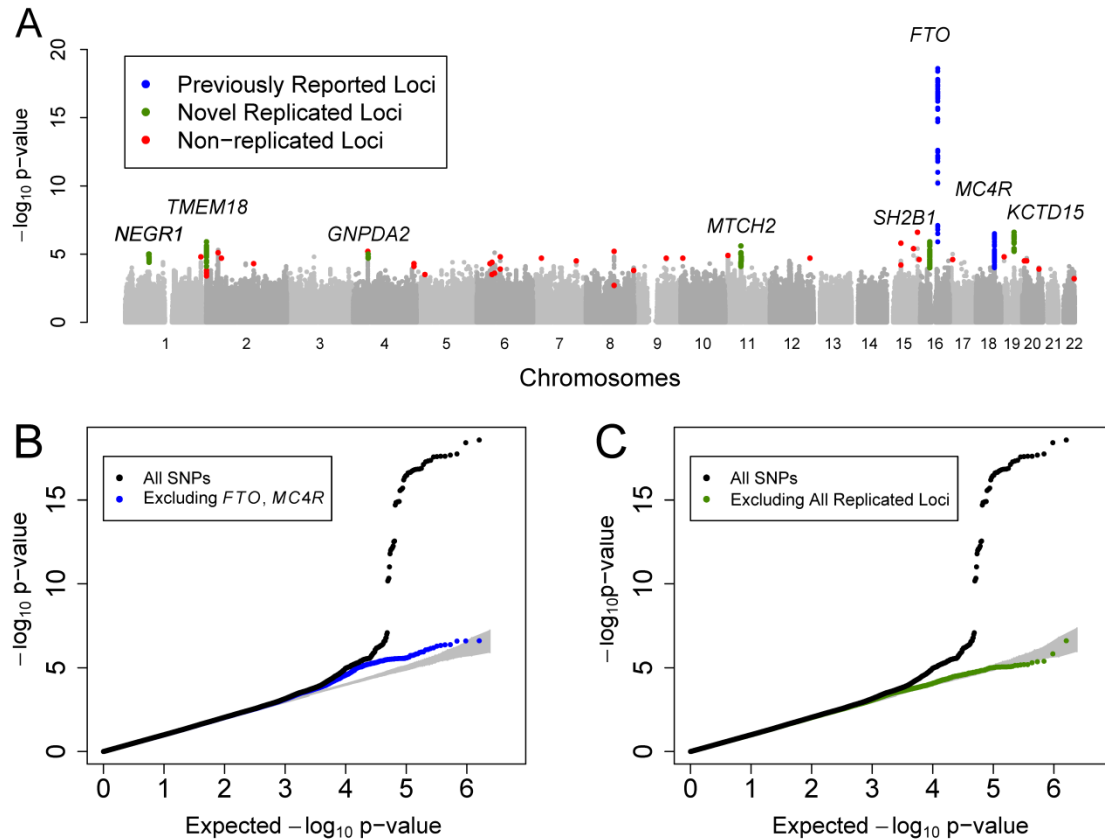


What is The Contribution of Structural Variants?

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

Evidence that Copy Number Variants Important

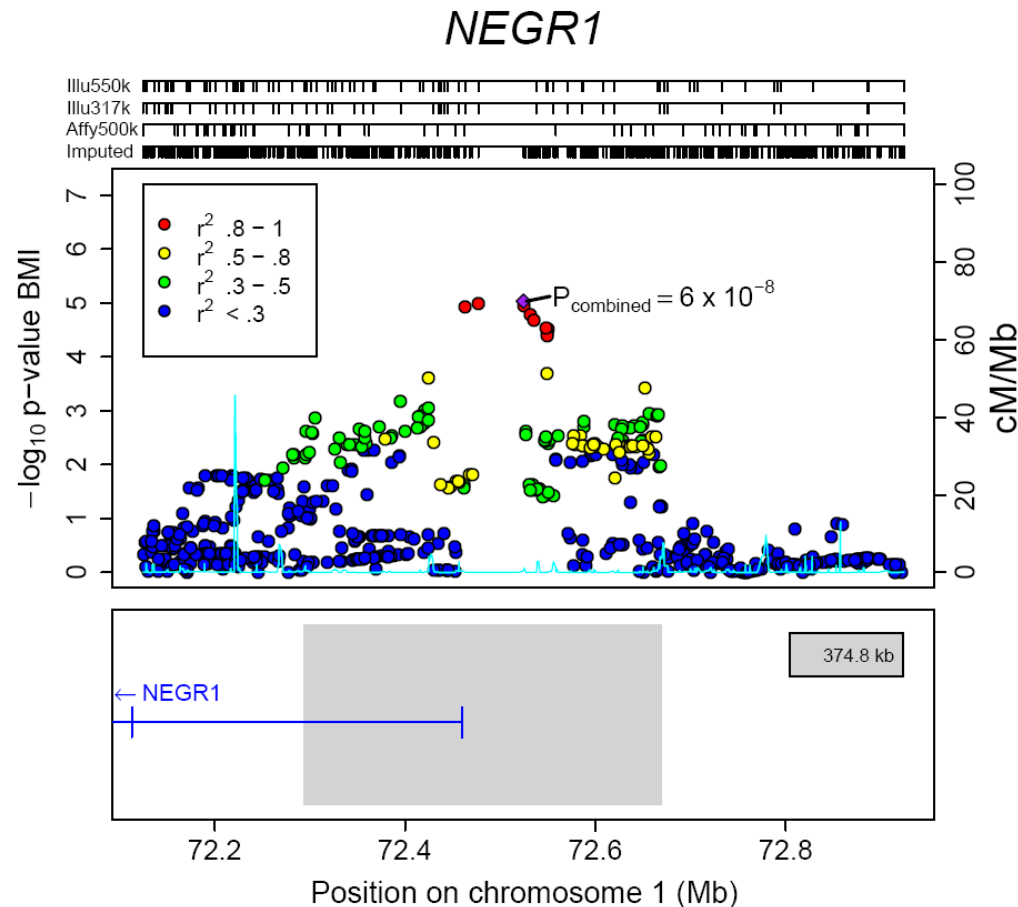
Example from Genetics of Obesity



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

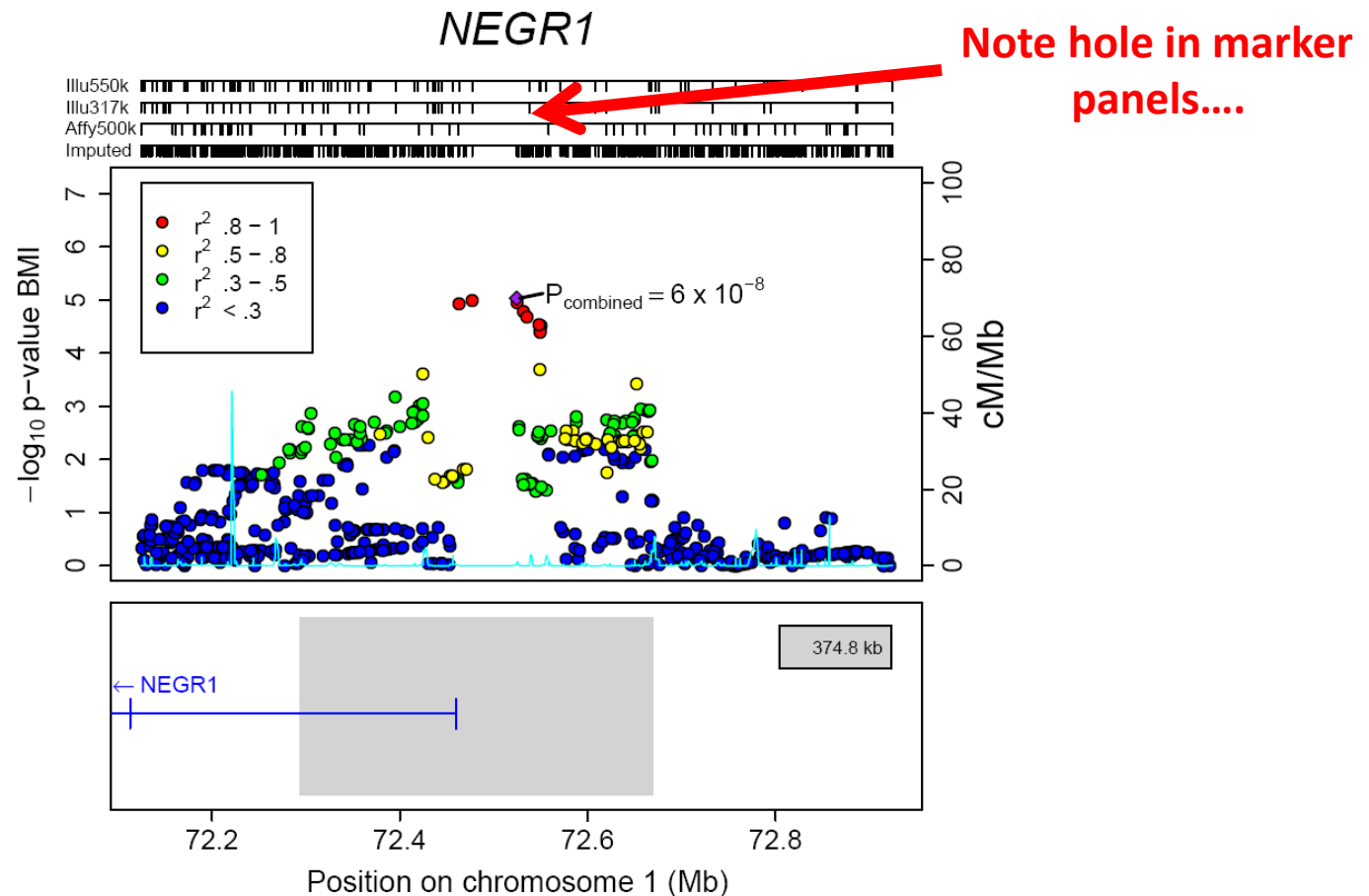
Evidence that Copy Number Variants Important

Example from Genetics of Obesity

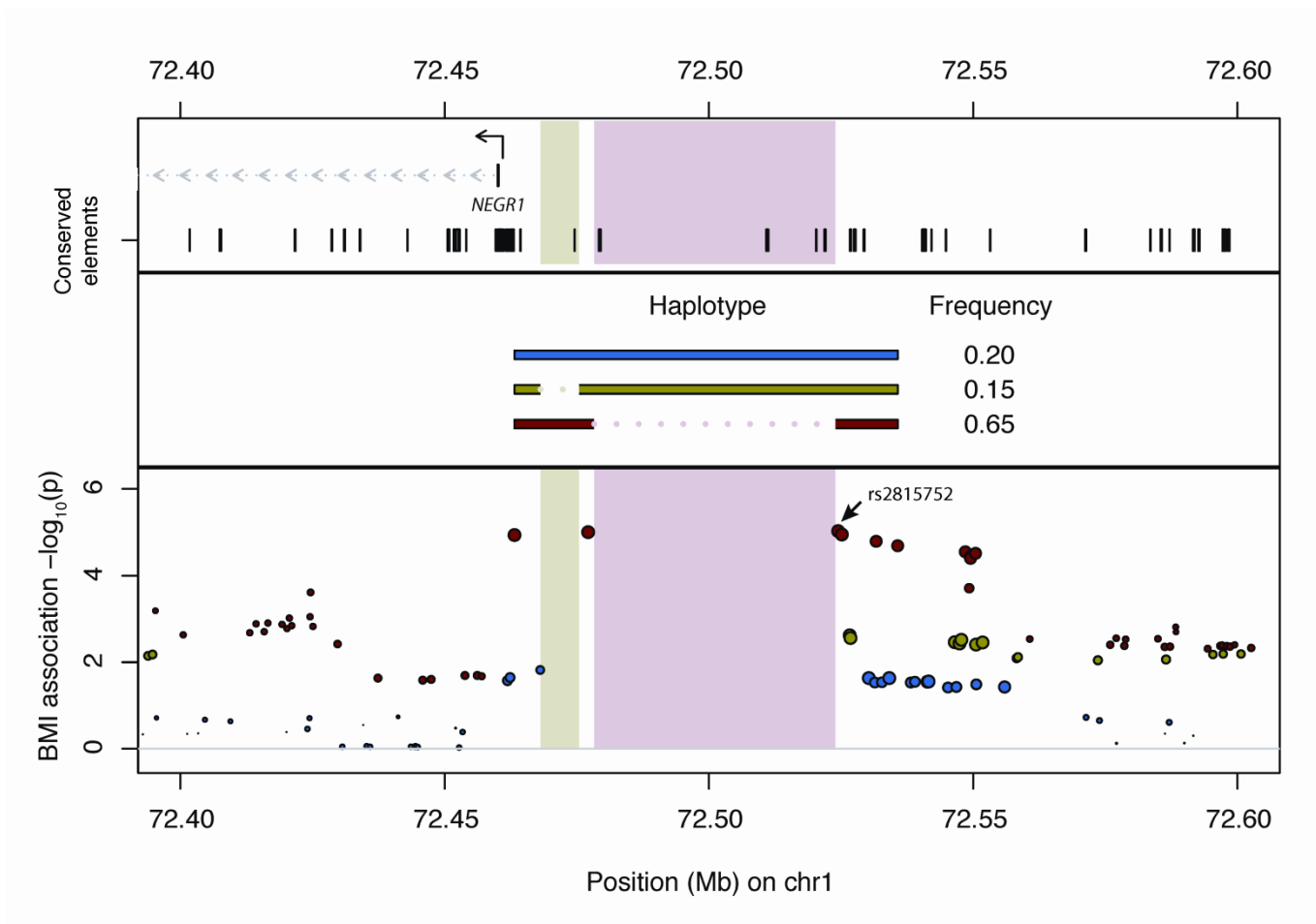


Evidence that Copy Number Variants Important

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

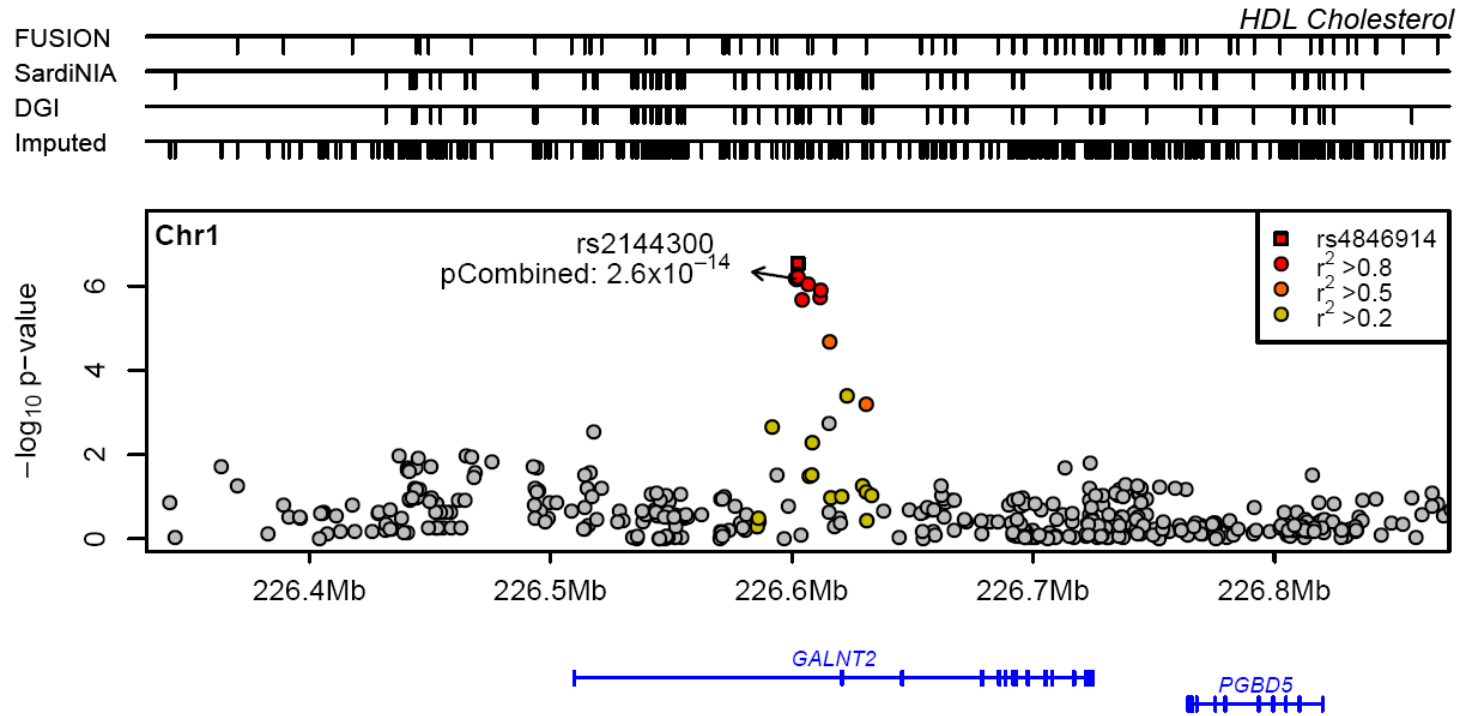
Recent Example from John Todd's Group

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

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 $\sim 30\%$

The Challenge

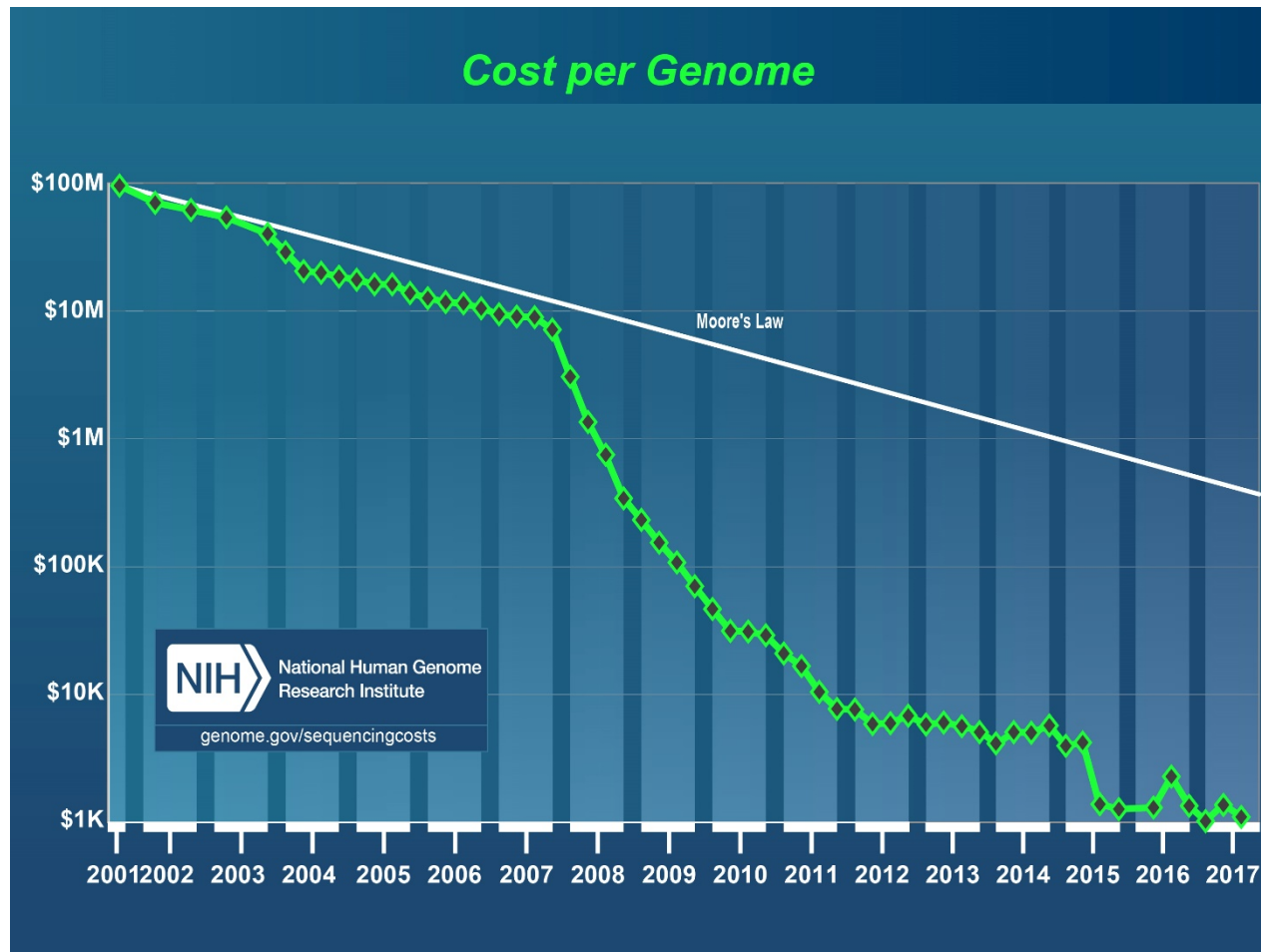
- Whole genome sequence data will greatly increase our understanding of complex traits
- Although a handful of genomes have been sequenced, this remains a relatively expensive enterprise
- Dissecting complex traits will require whole genome sequencing of 1,000s of individuals
- **How to sequence 1,000s 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
 - Illumina is currently the dominant technology (by far)
- Sequence data consist of millions or billions of short sequence reads with moderate accuracy
 - 0.5 – 1.0% error rates per base may be typical

21st Century Sequencing Costs



<http://genome.gov/sequencingcostsdata>

Shotgun Sequence Reads

ACTGGTCGATGCTAGCTGATAGCTAGCTA
GCTGATGAGCCCGATCGCTGCTAGCTCG
AGCTGATAGCTAGCTAGCTGATGAGCCCGA
GAGCCCGATCGCTGCTAGCTCGACG

- Typical short read might be <50-150 bp long and not very informative on its own
- Reads must be arranged (*aligned*) relative to each other to reconstruct longer sequences

Base Qualities

Short Read Sequence
GCTAGCTGATAGCTAGCTGATGAGCCCGA

Short Read Base Qualities
30.30.28.28.29.27.30.29.28.25.24.26.27.24.24.23.20.21.22.10.25.25.20.20.18.17.16.15.14.14.13.12.10

- Each base is typically associated with a quality value
- Measured on a “Phred” scale, which was introduced by Phil Green for his Phred sequence analysis tool

$BQ = -\log_{10}(\epsilon)$, where ϵ is the probability of an error

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 can now handle tens of millions of reads per hour ...
- Analyzing these data without a reference human genome would require much longer reads or result in very fragmented assemblies

Read Alignment – Food for Thought

- Typically, all the words present in the genome are indexed to facilitate read mapping ...
 - What are the benefits of using short words?
 - What are the benefits of using long words?
- How matches do you expect, on average, for a 10-base word?
 - Do you expect large deviations from this average?

Mapping Quality

- Measures the confidence in an alignment, which depends on:
 - Size and repeat structure of the genome
 - Sequence content and quality of the read
 - Number of alternate alignments with few mismatches
- The mapping quality is usually also measured on a “Phred” scale
- Idea introduced by Li, Ruan and Durbin (2008) *Genome Research* **18**:1851-1858

Mapping Quality Definition

- Given a particular alignment \mathbf{A} , we can calculate

$$\begin{aligned} P(\mathbf{S}|\mathbf{A}, \mathbf{Q}) &= \\ &= \prod_i^n P(\mathbf{S}_i|\mathbf{A}, \mathbf{Q}) \\ &= \prod_i^n \left\{ \frac{1}{3} 10^{-\mathbf{Q}_i/10} \right\}^{I(S_i \text{ mismatch}|\mathbf{A})} \left\{ 1 - 10^{-\mathbf{Q}_i/10} \right\}^{I(S_i \text{ match}|\mathbf{A})} \end{aligned}$$

- Then, the mapping quality is:

$$MQ(\mathbf{S}|\mathbf{A}_{best}, \mathbf{Q}) = \frac{P(\mathbf{S}|\mathbf{A}_{best}, \mathbf{Q})}{\sum_i P(\mathbf{S}|\mathbf{A}_i, \mathbf{Q})}$$

- In practice, summing over all possible alignments is too costly and this quantity is approximated (for example, by summing over the most likely alignments).

Refinements to Mapping Quality

- In their simplest form, mapping qualities apply to the entire read
- However, in gapped alignments, uncertainty in alignment can differ for different portions of the read
 - For example, it has been noted that many wrong variant calls are supported by bases near the edges of a read
- Per base alignment qualities were introduced to summarize local uncertainty in the alignment

Per Base Alignment Qualities

Short Read

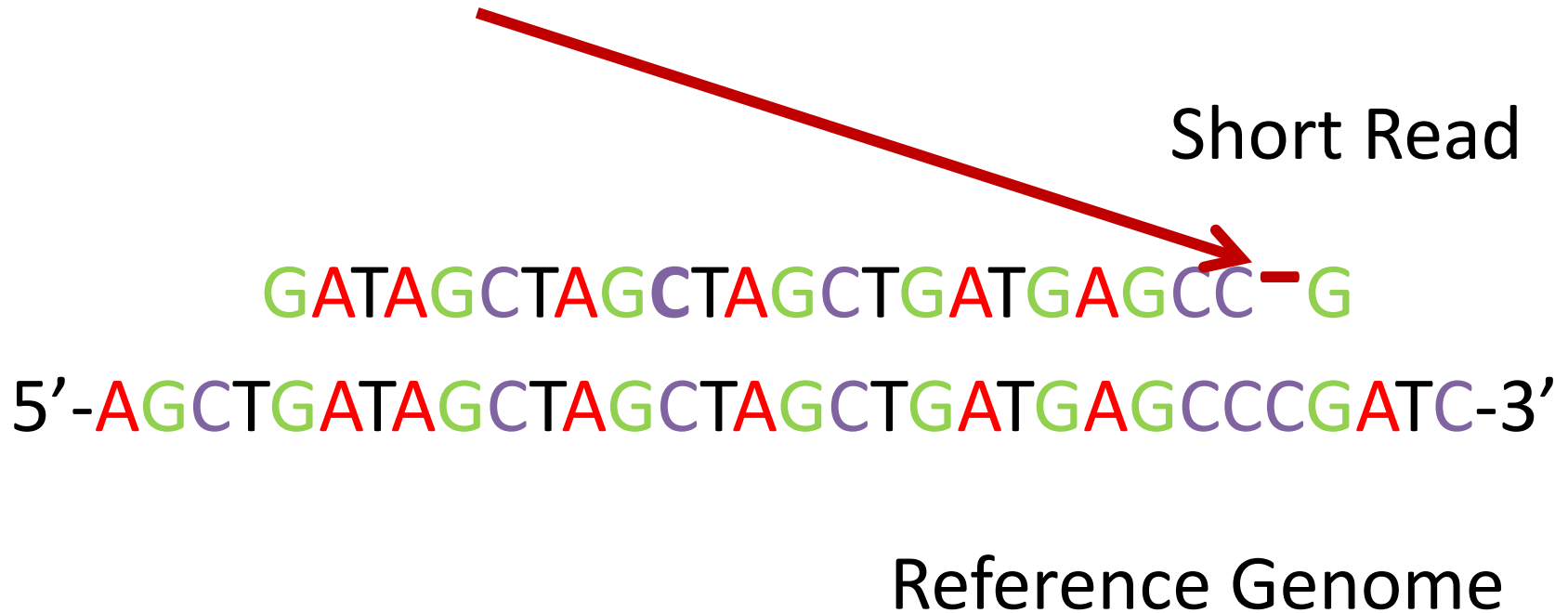
GATAGCTAGCTAGCTGATGA GCCG

5'-AGCTGATAGCTAGCTAGCTGATGAGCCCGATC-3'

Reference Genome

Per Base Alignment Qualities

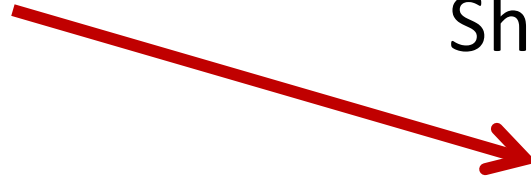
Should we insert a gap?



Per Base Alignment Qualities

**Compensate for Alignment Uncertainty
With Lower Base Quality**

Short Read



GATAGCTAGCTAGCTGATGAGCCG

5'-AGCTGATAGCTAGCTAGCTGATGAGCCCGATC-3'

Reference Genome

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
- Let's 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



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

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

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

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

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

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

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

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

ATGCTAGCTGATAGCTAGCTAGCTGATGAGCC
AGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTG
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

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

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

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

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

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

From Sequence to Genotype: 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

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

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

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

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

From Sequence to Genotype: 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.

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

From Sequence to Genotype: 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	Posterior(A/A) = <.001
$P(\text{reads} A/C) = 0.03125$	Prior(A/C) = 0.32	Posterior(A/C) = 0.999
$P(\text{reads} C/C) = 0.000097$	Prior(C/C) = 0.64	Posterior(C/C) = <.001

Population Based Prior: Use frequency information from examining others at the same site.
In the example above, we estimated $P(A) = 0.20$

From Sequence To Genotype: 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 $P(A) = 0.20$

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

Shotgun Sequence Data

Haplotype 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) = 0.81

Posterior(A/A) = <.001

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

Prior(A/C) = 0.18

Posterior(A/C) = 0.999

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

Prior(C/C) = 0.01

Posterior(C/C) = <.001

Haplotype Based Prior: Examine other chromosomes that are similar at locus of interest.
In the example above, we estimated that 90% of similar chromosomes carry allele A.

Shotgun Sequence Data

Haplotype 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

$P(\text{reads} | A/A) = 0.00000098$ $\text{Prior}(A/A) = 0.81$

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

$P(\text{reads} | A/C) = 0.03125$ $\text{Prior}(A/C) = 0.18$

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

$P(\text{reads} | C/C) = 0.000097$ $\text{Prior}(C/C) = 0.01$

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

Haplotype Based Prior: Examine other chromosomes that are similar at locus of interest.
In the example above, we estimated that 90% of similar chromosomes carry allele A.

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

Challenges with the basic approach ...



ACTGGTCGATGCTAGCTGATAGCTAGCTAGATGATGAGCCCCGATCGCTGCTAGCTCGACG
ACTGGTCGATGCTAGCTGATAGCTAGCTAGATGATGAGCCCCGATCGCTGCTAGCTCGACG
ACTGGTCGATGCTAGCTGATAGCTAGCTAGATGATGAGCCCCGATCGCTGCTAGCTCGACG
ACTGGTCGATGCTAGCTGATAGCTAGCTAGATGATGAGCCCCGATCGCTGCTAGCTCGACG
ACTGGTCGATGCTAGCTGATAGCTAGCTAGATGATGAGCCCCGATCGCTGCTAGCTCGACG
ACTGGTCGATGCTAGCTGATAGCTAGCTAGATGATGAGCCCCGATCGCTGCTAGCTCGACG
ACTGGTCGATGCTAGCTGATAGCTAGCTAGATGATGAGCCCCGATCGCTGCTAGCTCGACG
ACTGGTCGATGCTAGCTGATAGCTAGCTAGATGATGAGCCCCGATCGCTGCTAGCTCGACG
ACTGGTCGATGCTAGCTGATAGCTAGCTAGATGATGAGCCCCGATCGCTGCTAGCTCGACG
ACTGGTCGATGCTAGCTGATAGCTAGCTAGATGATGAGCCCCGATCGCTGCTAGCTCGACG
ACTGGTCGATGCTAGCTGATAGCTAGCTAGATGATGAGCCCCGATCGCTGCTAGCTCGACG
ACTGGTCGATGCTAGCTGATAGCTAGCTAGATGATGAGCCCCGATCGCTGCTAGCTCGACG
ACTGGTCGATGCTAGCTGATAGCTAGCTAGATGATGAGCCCCGATCGCTGCTAGCTCGACG
ACTGGTCGATGCTAGCTGATAGCTAGCTAGATGATGAGCCCCGATCGCTGCTAGCTCGACG
ACTGGTCGATGCTAGCTGATAGCTAGCTAGATGATGAGCCCCGATCGCTGCTAGCTCGACG
ACTGGTCGATGCTAGCTGATAGCTAGCTAGATGATGAGCCCCGATCGCTGCTAGCTCGACG

5' -ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTGATGAGCCCCGATCGCTGCTAGCTCGACG-3'

Challenges with the basic approach ...

[illegible]

5' - ACTGGT **C** GAT **G** CTA **G** CTA **G** A TAG CTA **G** CTA **G** A TAG CCA **C** GAT **C** GCT **G** CTA **G** CTA **C** GAC **G** - 3'

Challenges with the basic approach ...



```
CTAGATGATGAGCCCGATCGCTGCTAGCTC
AGATGATGAGCCCGATCGCTGCTAGCTCGA
GATGATGAGCCCGATCGCTGCTAGCTCGAC
AGATGATGAGCCCGATCGCTGCTAGCTCGA
ATGATGAGCCCGATCGCTGCTAGCTCGACG
GATGATGAGCCCGATCGCTGCTAGCTCGAC
AGATGATGAGCCCGATCGCTGCTAGCTCGA
GATGATGAGCCCGATCGCTGCTAGCTCGAC
GCTAGCTAGCTGATGAGCCCGATCGCTGCT
GATAGCTAGCTAGCTGATGAGCCCGCTCGC
AGCTAGCTGATGAGCCCGATCGCTGCTAGC
CTAGCTGATGAGCCCGATCGCTGCTAGCTC
GCTGATAGCTAGCTAGCTGATGAGCCCGAT
GATGCTAGCTGATAGCTAGCTAGCTGATGA
GTCGATGCTAGCTGATAGCTAGCTAGCTGA
TAGCTAGCTAGCTGATGAGCCCGATCGCTG
```

5' - ACTGGTCCGATGCTAGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG - 3'

Challenges with the basic approach ...



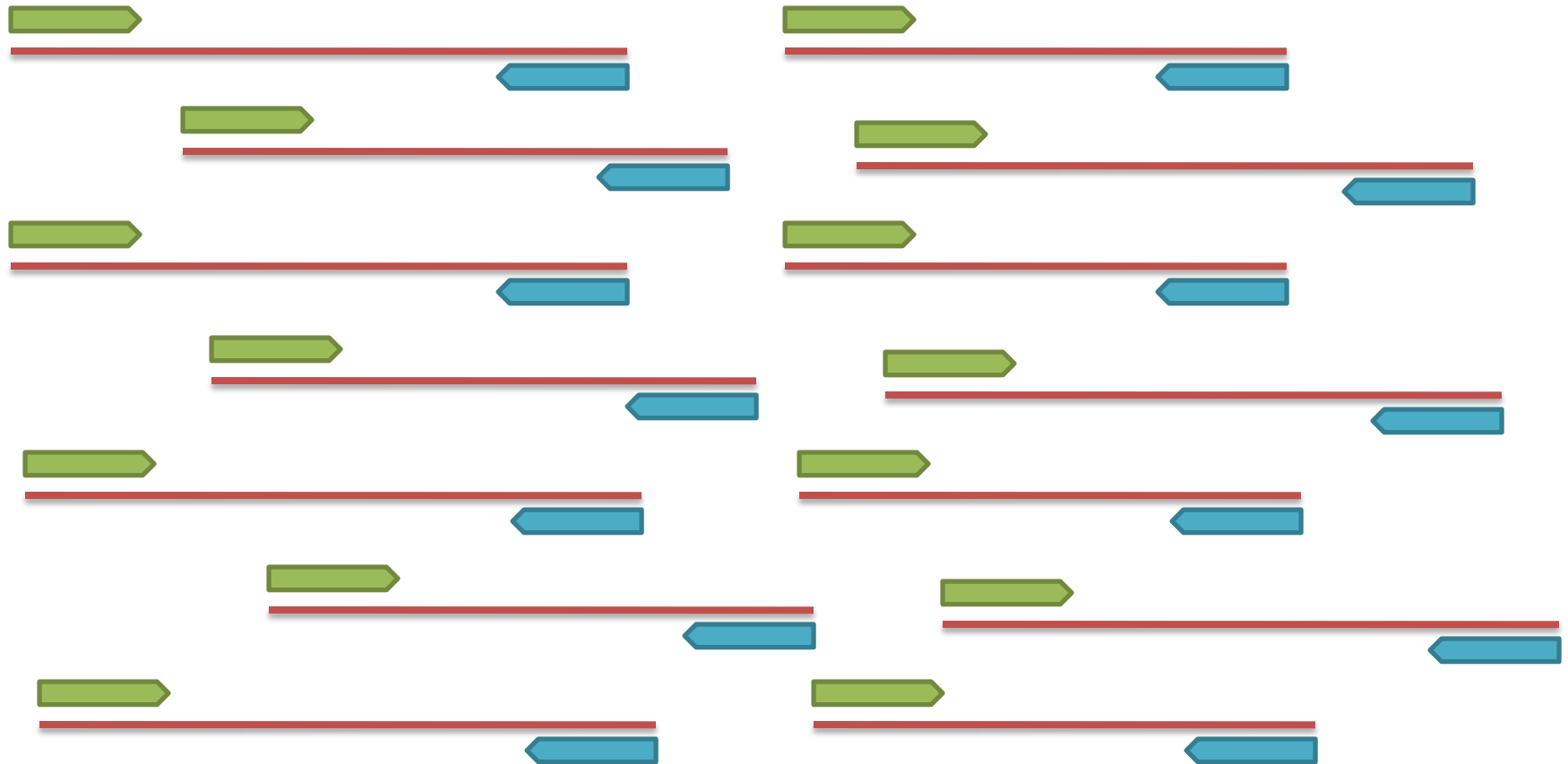
ACTAGTCGATGCTGGCTGATAGCTAGCTAGATGATGAGCCCGTTCGCTCCTAGCTCGACG
ACTAGTCGATGCTGGCTGATAGCTAGCTAGATGATGAGCCCGTTCGCTCCTAGCTCGACG
ACTAGTCGATGCTGGCTGATAGCTAGCTAGATGATGAGCCCGTTCGCTGCTAGCTCGACG
ACTAGTCGATGCTGGCTGATAGCTAGCTAGATGATGAGCCCGTTCGCTCCTAGCTCGACG
ACTAGTCGATGCTGGCTGATAGCTAGCTAGATGATGAGCCCGTTCGCTCCTAGCTCGACG
ACTAGTCGATGCTGGCTGATAGCTAGCTAGATGATGAGCCCGATCGCTGCTAGCTCGACG
ACTAGTCGATGCTGGCTGATAGCTAGCTAGATGATGAGCCCGTTCGCTCCTAGCTCGACG
ACTAGTCGATGCTAGCTGATAGCTAGCTAGATGATGAGCCCGTTCGCTCCTAGCTCGACG
ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG
ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG
ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG
ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG
ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG
ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG
ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG

5' - ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG - 3'

Variant Filtering

- Modern callers start with a candidate list of sites and annotate these ...
 - Likely good sites: variants in HapMap or Omni 2.5M arrays
 - Likely problematic sites: variants that deviate from HWE or don't segregate in multiple families
- Then, build a model that separates likely good sites from likely bad ones ...
 - SVM, VQSR, self-organizing maps,
- Possible features ...
 - What is the mapping quality of reads with the variant?
 - How many other differences in reads with the variant?
 - How many individuals are heterozygotes and homozygotes?
 - How many reads with the variant are on the forward and reverse strand?
 - What fraction of reads have the variant in heterozygotes?
 - ...

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

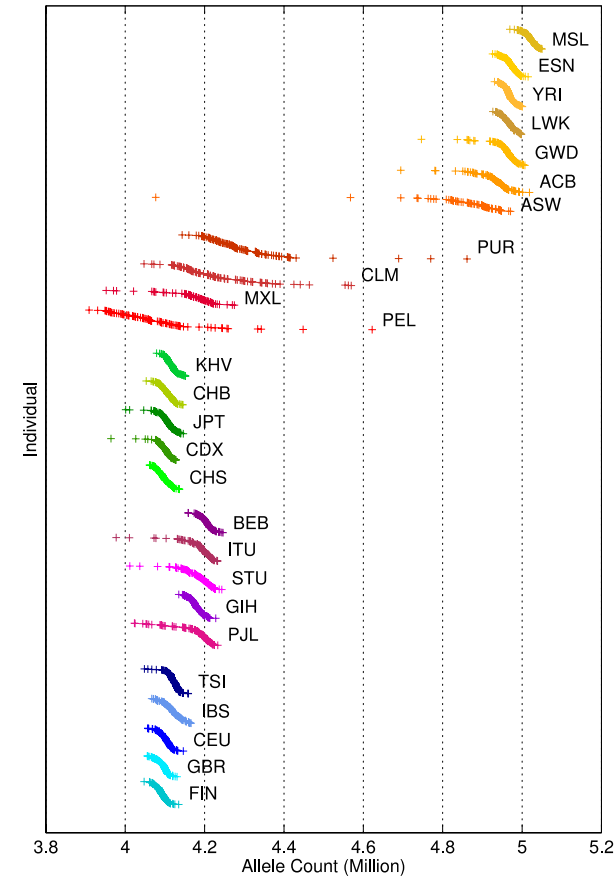
How Much Variation is There?

- An average genome includes:
 - About 4M SNPs
 - About 500K indels
 - Hundreds or thousands of larger deletions
- Numbers are probably underestimates ...
- ... some variants are hard to call with short reads
- 1000 Genomes Project (2012) *Nature* **491**:56-65

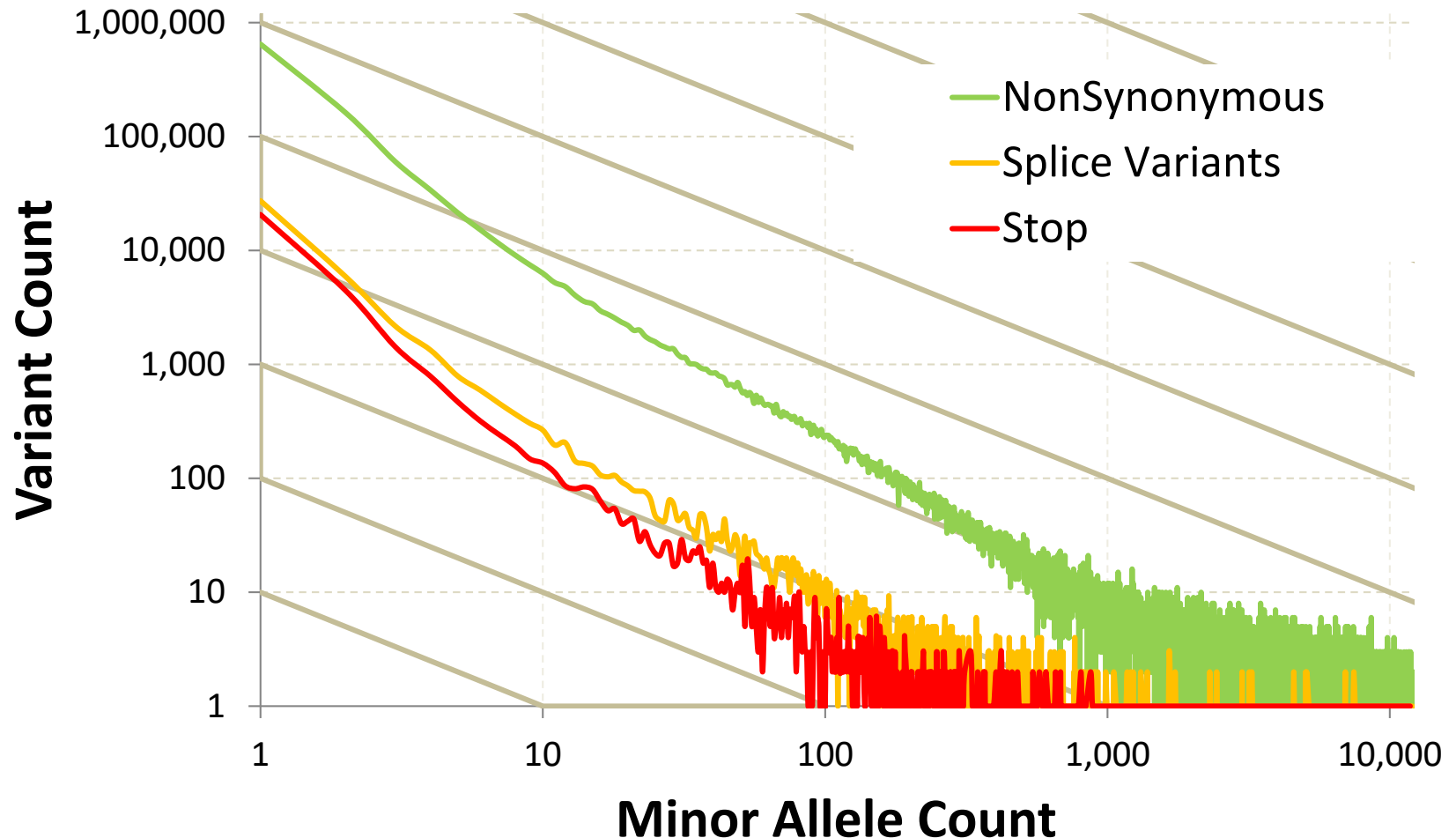
Variants per genome

(1000 Genomes Project)

Type	Variant sites / genome
SNPs	~3,800,000
Indels	~570,000
Mobile Element Insertions	~1000
Large Deletions	~1000
CNVs	~150
Inversions	~11



Allele Frequency Spectrum (After Sequencing 12,000+ Individuals)



How Much Variation is There?

(TOPMed 65K)

Variant Type	Category	# PASS	# FAIL	% dbSNP (PASS)	Known/Novel Ts/Tv (PASS)
SNP	All	438M	85M	22.9%	1.93 / 1.69
	Singleton	202M	24M	8.5%	1.23 / 1.54
	Doubleton	69M	8.8M	12.6%	1.61 / 1.74
	Tripletion ~ 0.1%	142M	24M	34.9%	2.23 / 1.99
	0.1% ~ 1%	13M	4.5M	98.2%	2.17 / 1.79
	1 ~ 10%	6.5M	2.9M	99.6%	1.82 / 1.75
	>10%	5.3M	2.0M	99.8%	2.11 / 1.88
Indels	All	33.4M	26.2M	20.1%	
	Singleton	15.7M	4.7M	10.1%	
	Doubleton	5.3M	1.8M	12.6%	
	Tripletion ~ 0.1%	10.7M	8.0M	26.7%	
	0.1% ~ 1%	2.8M	968K	88.9%	
	1 ~ 10%	432K	2.3M	98.5%	
	>10%	298K	1.4M	99.6%	

How Much Variation is There?

(TOPMed 65K – Coding Variation)

Type	Category	PASS Variants	% AC = 1	% AC ≤ 2	AF < 0.1%	AF < 1%
SNP	All	438M	46.1%	61.9%	94.2%	98.7%
	Synonymous	1.62M	42.9%	58.7%	94.5%	97.6%
	Missense	3.44M	47.7%	64.1%	96.8%	98.8%
	Stop Gain	103K	54.4%	71.3%	98.4%	99.5%
	Essential Splice	111K	54.2%	70.3%	96.8%	98.6%
Indels	All	33.4M	47.0%	62.8%	94.9%	98.8%
	Frameshift	97.0K	59.9%	76.0%	98.7%	99.6%
	Inframe	65.6K	48.6%	65.3%	97.5%	99.3%
	Ess. Splice & Others	12.7K	52.7%	68.8%	97.0%	98.8%

Summary

- Introduction to whole genome sequencing
 - Read mapping
 - Genotype calling
 - Analysis of structural variation
- Sequencing and the genetics of complex traits
 - Advantages and disadvantages versus genotyping
 - What sorts of things might we learn?

Recommended Reading

- The 1000 Genomes Project (2010) A map of human genome variation from population-scale sequencing. *Nature* **467**:1061-73