Imputation
Next Generation Sequencing
Imputation and Sequencing

Gonçalo Abecasis

University of Michigan School of Public Health

Imputation For Related Individuals

- Family members share large segments of chromosomes
- If we genotype many related individuals, we will effectively be genotyping a few chromosomes many times
- Propagate genotypes obtained in genome wide association study to related individuals
- Propagation can be based just on genetic relationships …
- … but will work better if we first identify shared chromosomal regions in each family using a subset of markers

Burdick et al, *Nat Genet,* 2006
Chen and Abecasis, AJHG, 2007

Relatedness in The Context of GWAS

- When analyzing family samples …

- FOR INDIVIDUALS WITH KNOWN RELATIONSHIPS
  - Impute genotypes in relatives, who may be completely untyped
  - Imputation works through long shared stretches of chromosome

- But the majority of GWAS that use “unrelated” individuals…

Relatedness in The Context of GWAS

- When analyzing family samples …

- FOR INDIVIDUALS WITH KNOWN RELATIONSHIPS
  - Impute genotypes in relatives, who may be completely untyped
  - Imputation works through long shared stretches of chromosome

- But the majority of GWAS that use “unrelated” individuals…

- FOR INDIVIDUALS WITH UNKNOWN RELATIONSHIPS
  - Impute observed genotypes in relatives
  - Imputation works through short shared stretches of chromosome

Observed Genotypes

Observed Genotypes

. . . . A . . . . A . . . . A . . .

. . . . G . . . . C . . . . A . . .

Study Sample

Reference Haplotypes

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

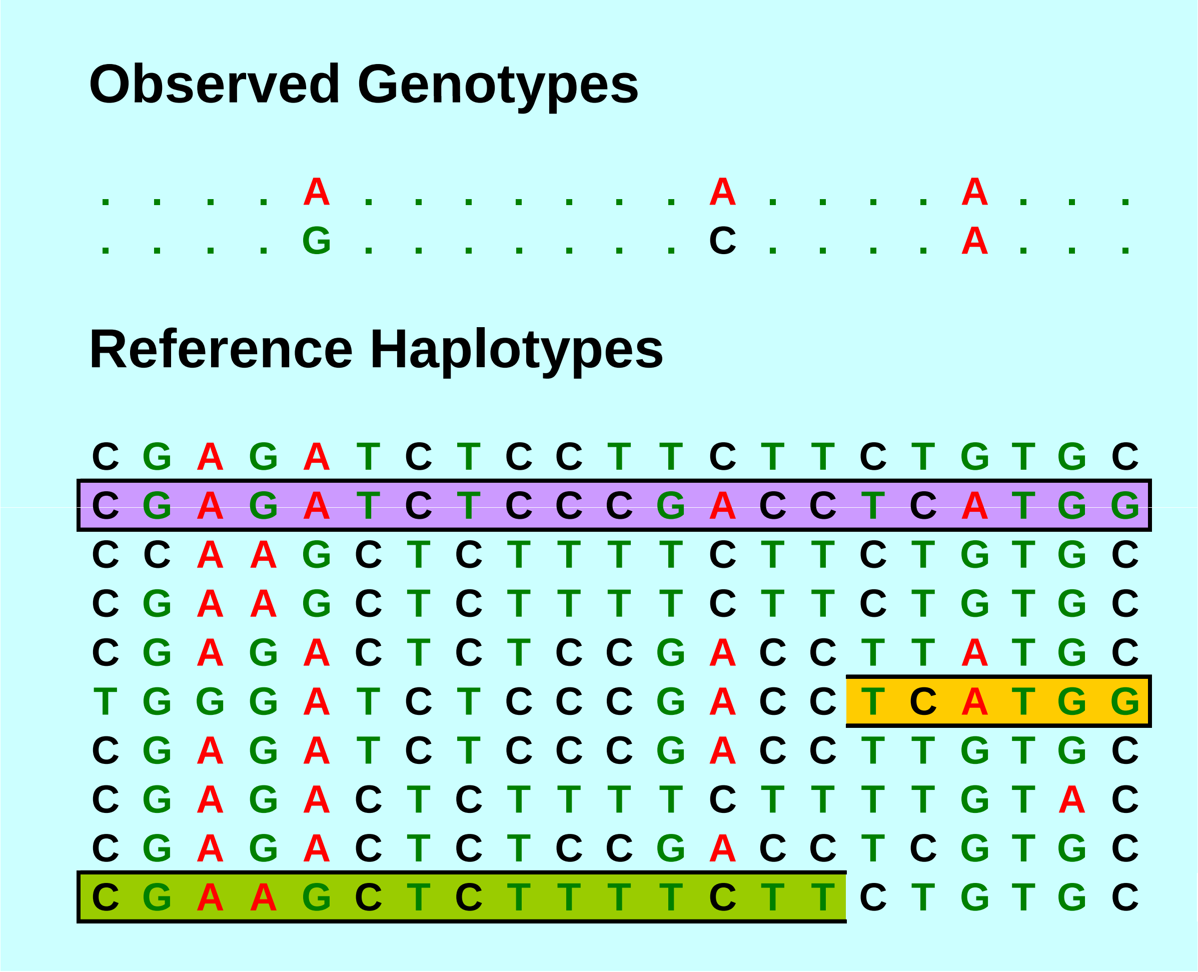
C G A G A C T C T T T T C T T T T G T A C

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

HapMap

Identify Match Among Reference



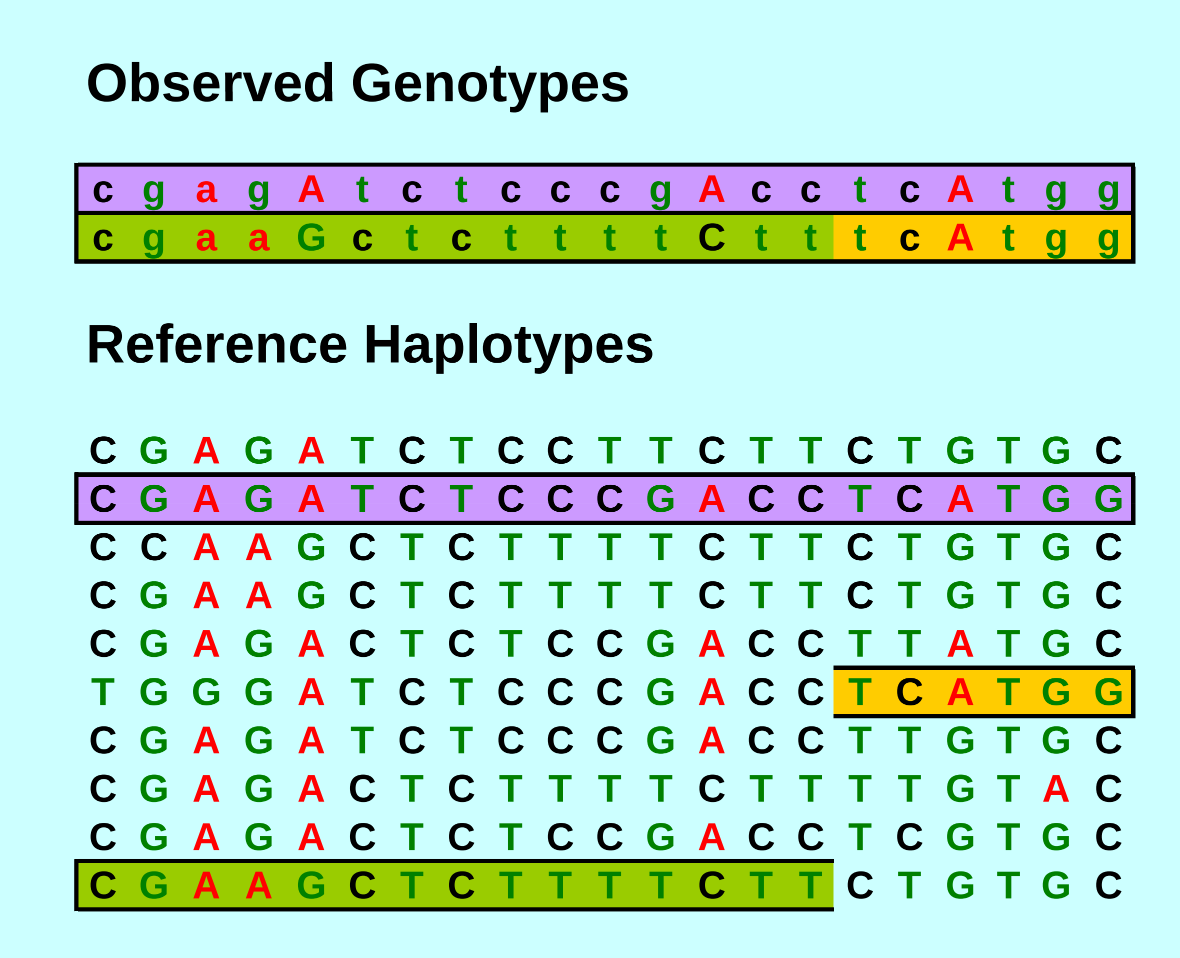
Observed Genotypes

```
. . . . A . . . . . . . A . . . . A . . .
. . . . G . . . . . . . C . . . . A . . .
```

Reference Haplotypes

```
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
```

Phase Chromosome, Impute Missing Genotypes



Observed Genotypes

```
c g a g A t c t c c c g A c c t c A t g g
c g a a G c t c t t t t C t t t c A t g g
```

Reference Haplotypes

```
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
```

Implementation

- Markov model is used to model each haplotype, conditional on all others

- Gibbs sampler is used to estimate parameters and update haplotypes
  - Each individual is updated conditional on all others
  - In parallel to updating haplotypes, estimate “error rates” and “crossover” probabilities

- In theory, this should be very close to the Li and Stephens (2003) model

Does Imputation Really Work? Results from One Recent Assessment

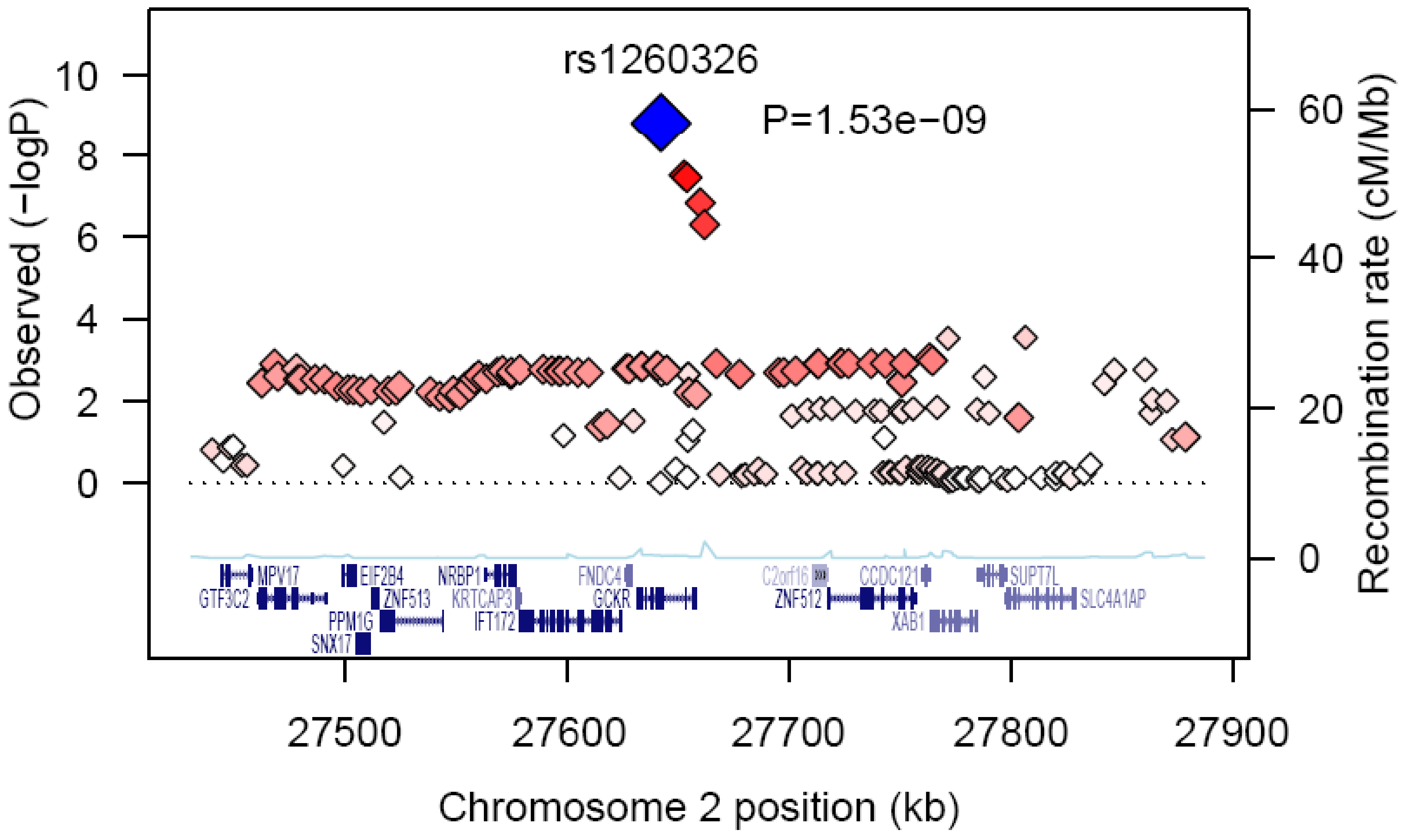
- Use 438,670 SNPs to impute 2.5M SNPs in GAIN psoriasis scan
  - Nair et al, *Nature Genetics*, in press

- Re-genotyped ~906,600 SNPs in 90 samples using the Affymetrix 6.0 chip.

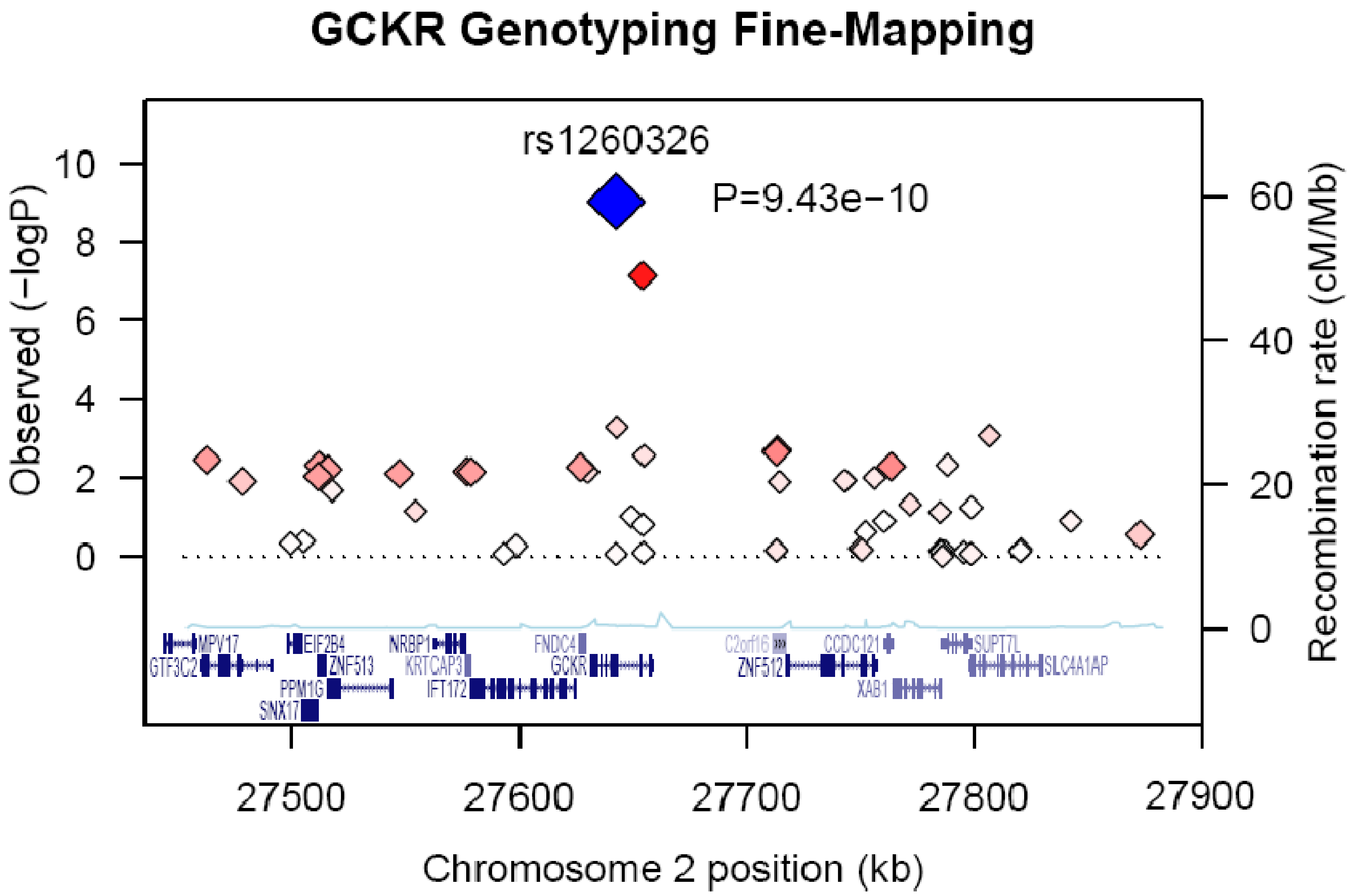
- Discrepancy rate of 1.80% per genotype (0.91% per allele).
  - 57,747,244 imputed genotypes compared with Affymetrix calls
  - 661,881 non-Perlegen SNPs present in the Affymetrix 6.0

- Average r^2 between imputed calls and Affymetrix calls was 0.93.
  - r^2 exceeded 0.80 for >90% of SNPs.

GCKR "In Silico" Fine-Mapping Using Imputation



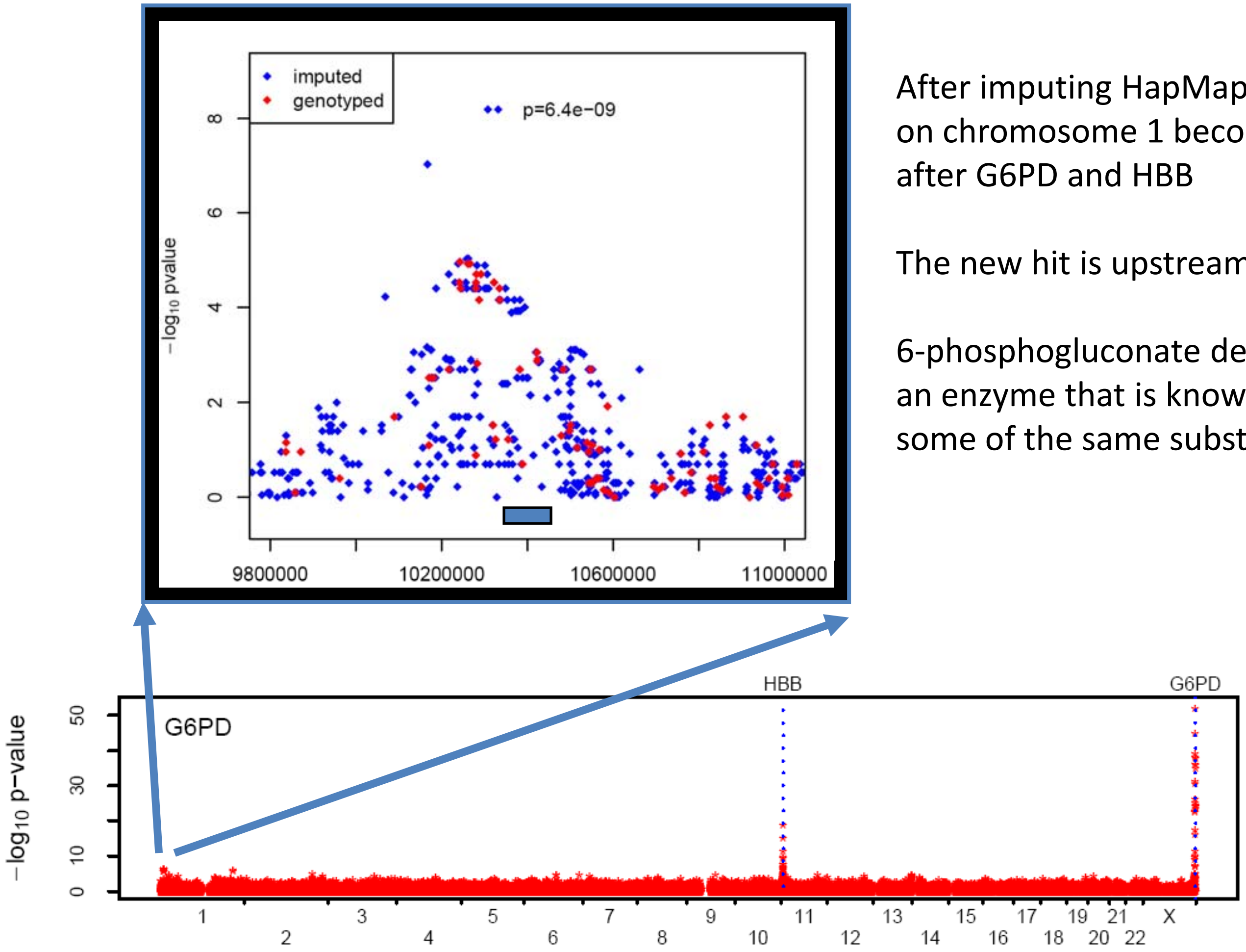
Association between triglycerides and GCKR
Sekar Kathiresan, DGI, see poster 32



Association between triglycerides and GCKR

Sekar Kathiresan, DGI, see poster 32

Sardinia G6PD Activity Example …

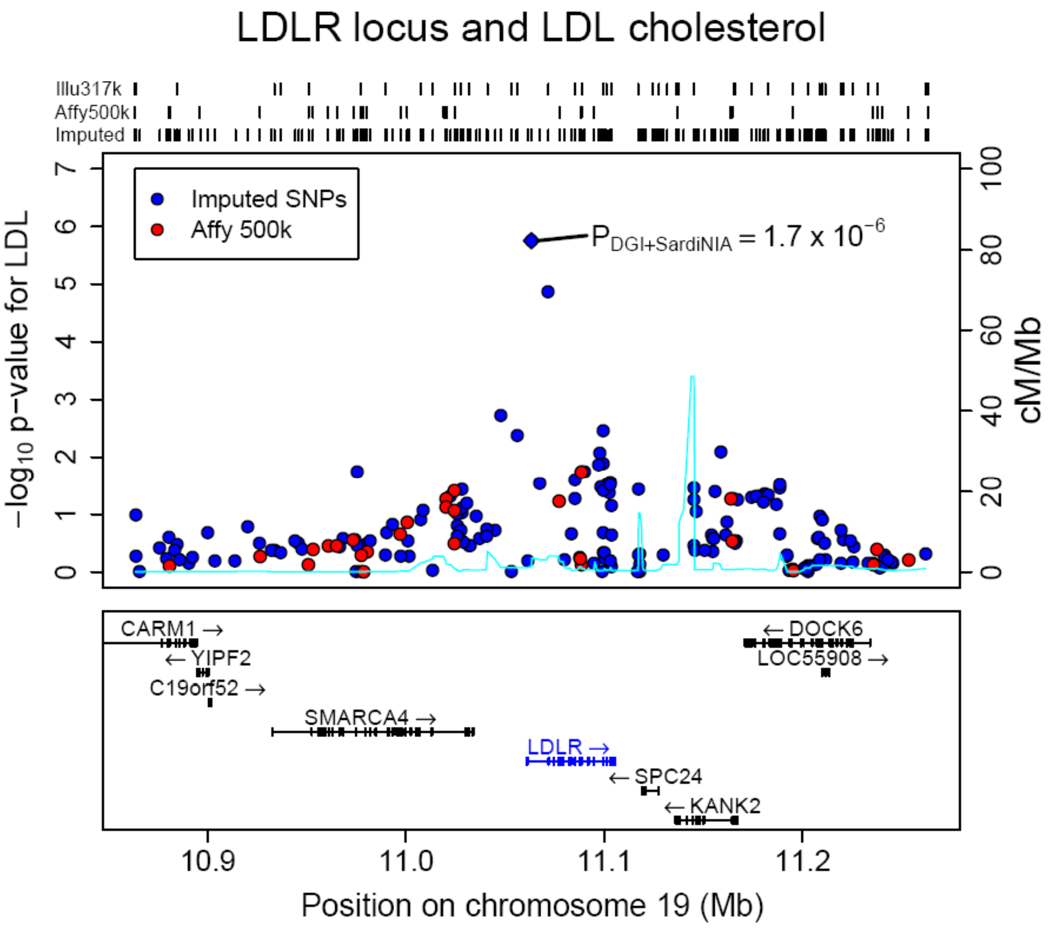


After imputing HapMap SNPs a region on chromosome 1 becomes top hit after G6PD and HBB

The new hit is upstream of 6PGD

6-phosphogluconate dehydrogenase is an enzyme that is known to metabolize some of the same substrates as G6PD

LDLR and LDL example



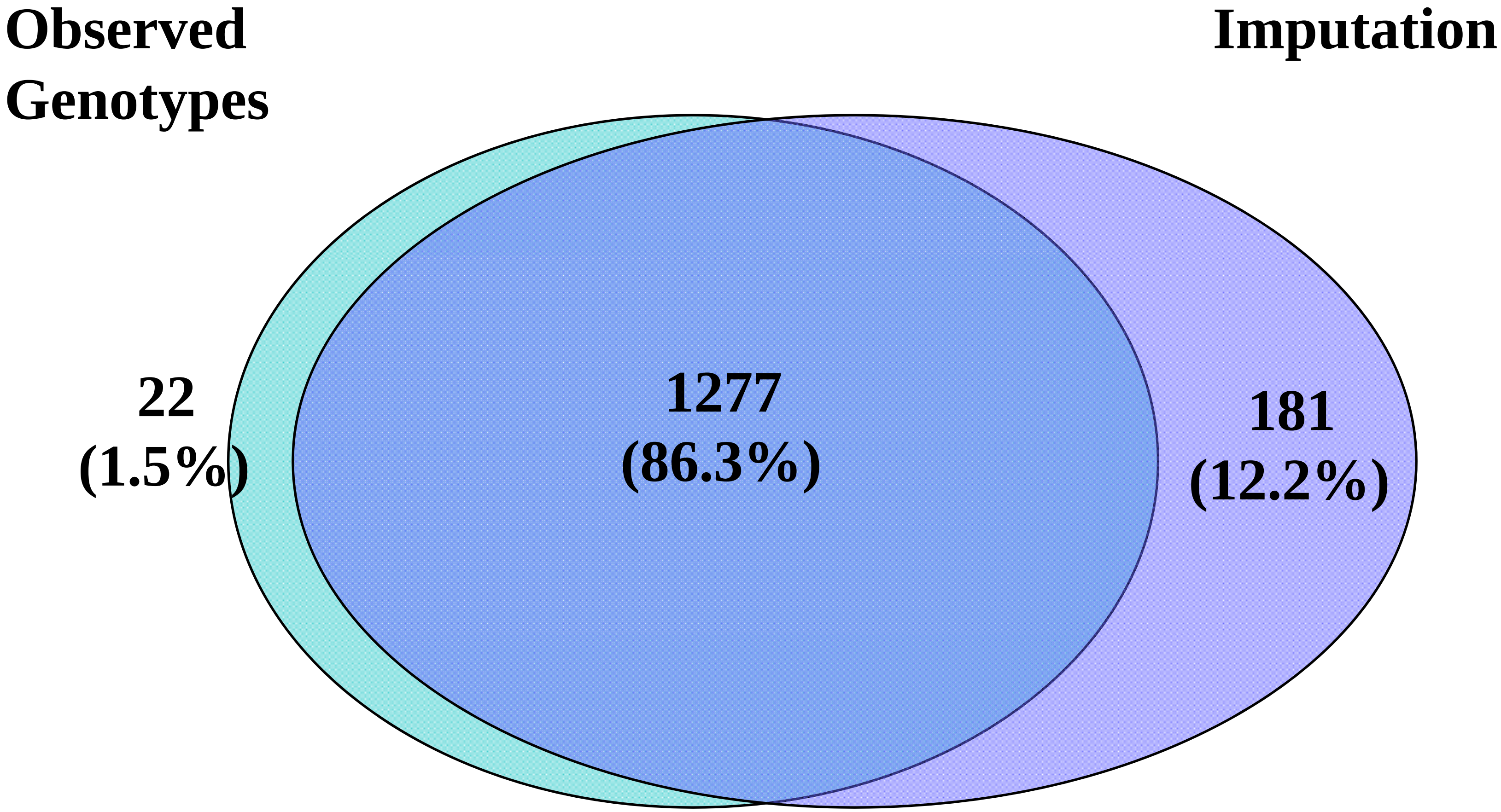
Impact of HapMap Imputation on Power

	Power	
Disease SNP MAF	tagSNPs	Imputation
2.5%	24.4%	56.2%
5%	55.8%	73.8%
10%	77.4%	87.2%
20%	85.6%	92.0%
50%	93.0%	96.0%

Power for Simulated Case Control Studies.
Simulations Ensure Equal Power for Directly Genotype SNPs.

Simulated studies used a tag SNP panel that captures
80% of common variants with pairwise $r^2 > 0.80$.

For eQTL Mapping, Imputation Increases Number of *cis* eQTL by ~10%



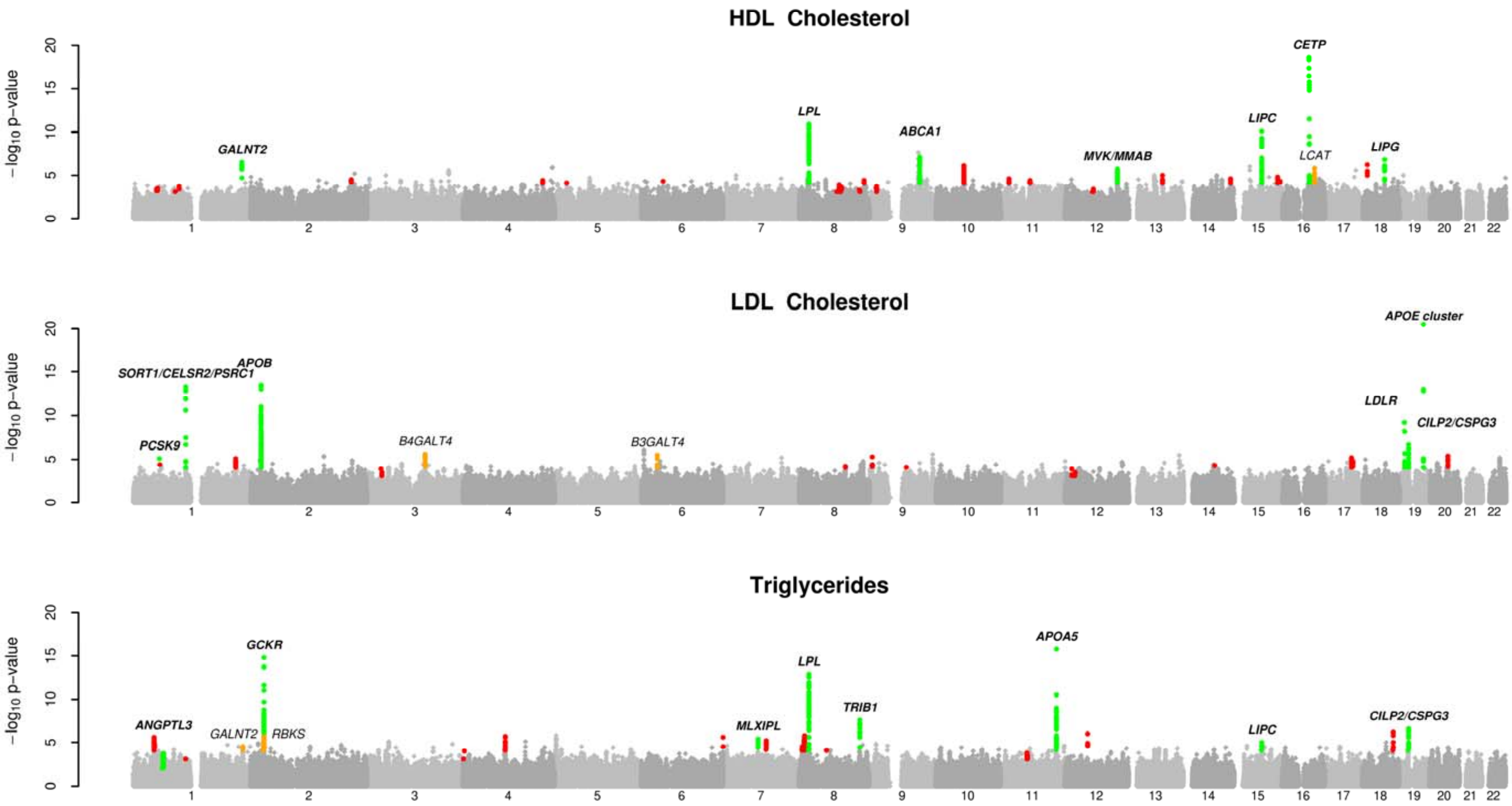
Combining Genomewide Studies: Cholesterol Levels Example

- HDL-Cholesterol, LDL-Cholesterol and Triglycerides
  - Strongly associated with risk of coronary artery disease
  - Important non-genetic factors include diet, statins, age
  - Several previously identified genes
  - Heritability 30-40%

- Our experiment
  - Examine 8,816 individuals from 3 genomewide scans
  - Scans used different marker platforms, combined with imputation
  - Individually, SardiNIA, FUSION and DGI scans had 1-3 hits
  - Confirm findings in >11,500 additional individuals

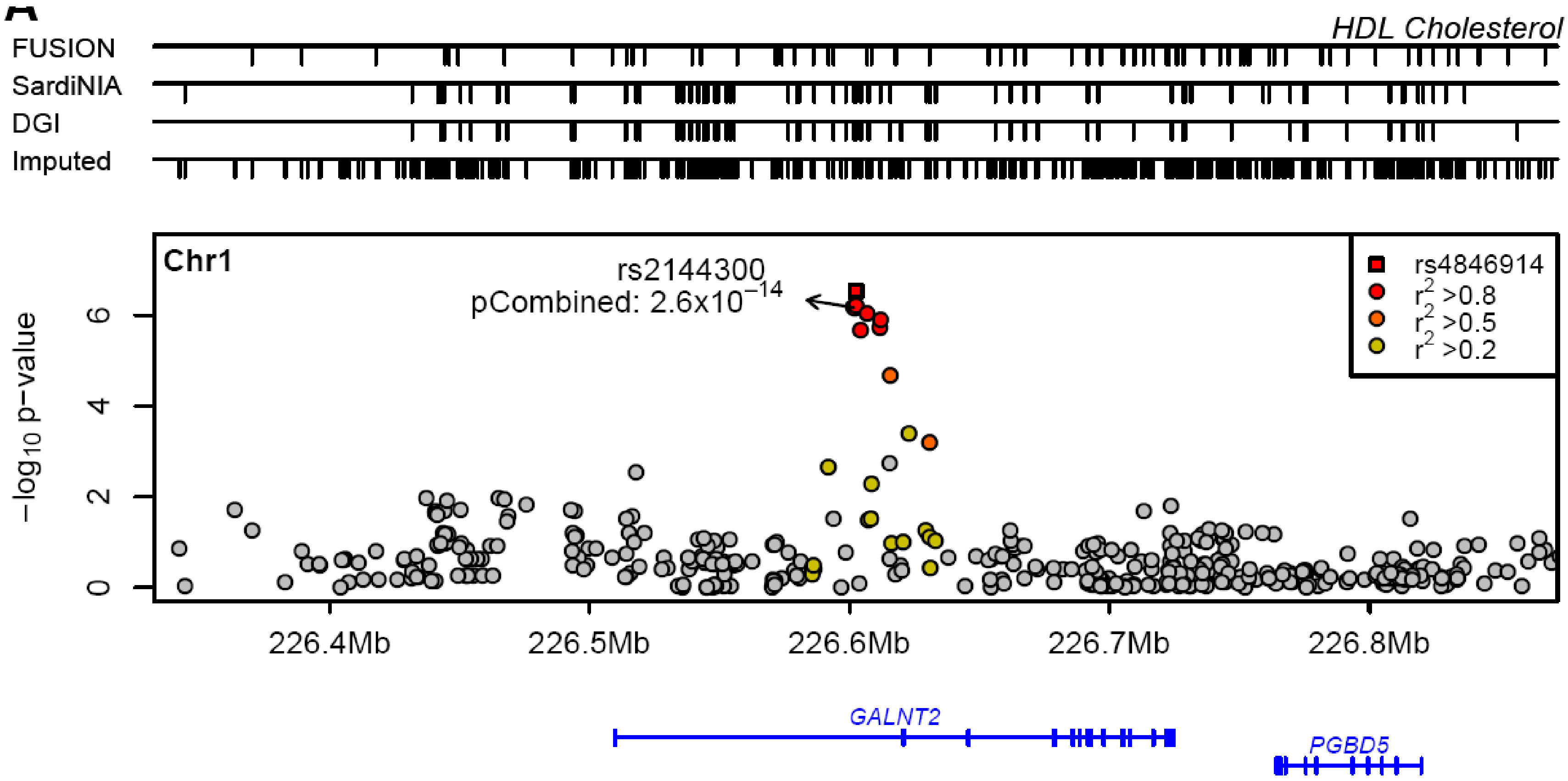
- Identified a total of 18 loci associated with cholesterol at $p < 5x10^{-8}$

What do we learn from meta-analysis? Combined Lipid Scan Results



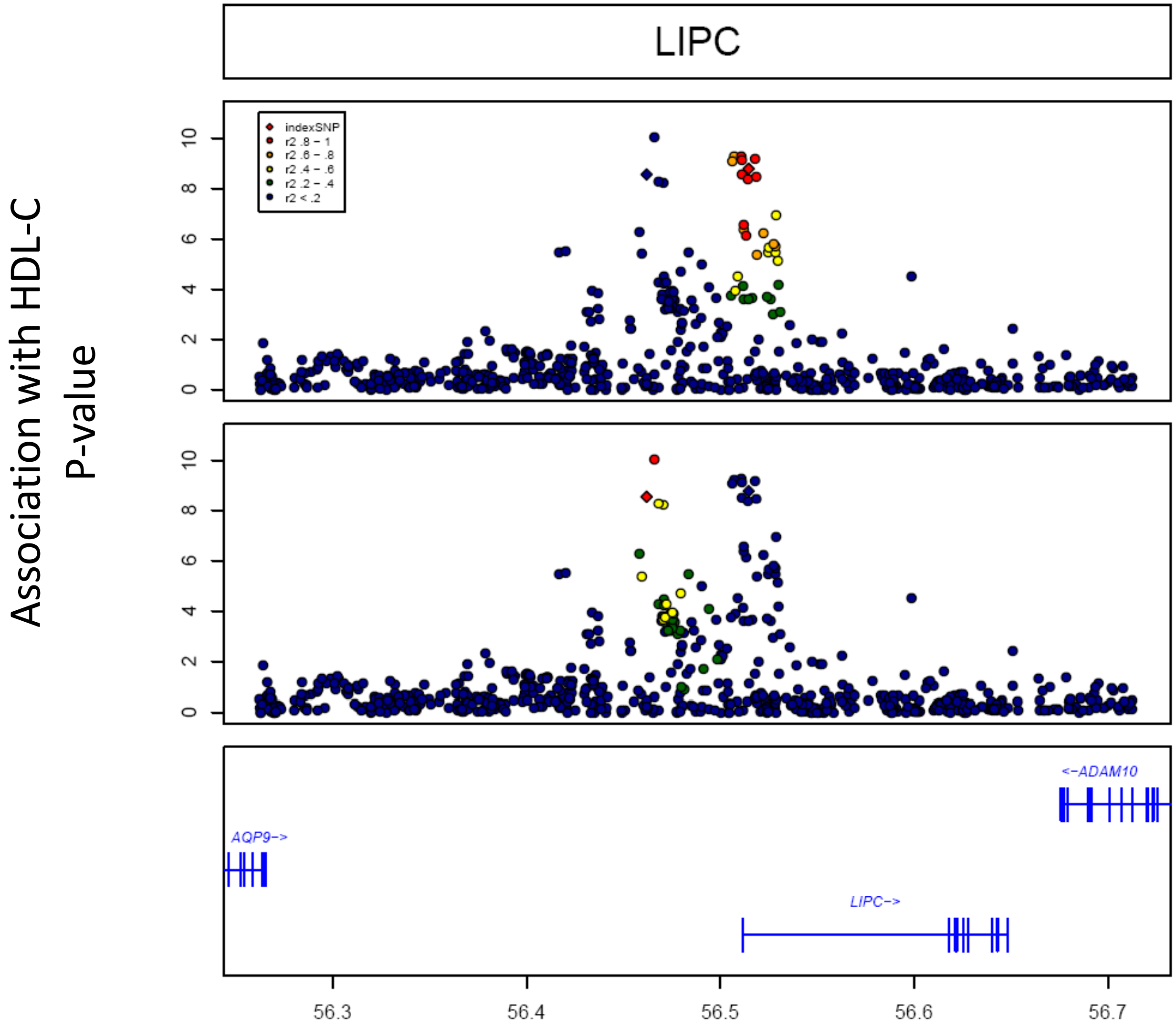
Willer et al, *Nat Genet*, 2008

New HDL Locus



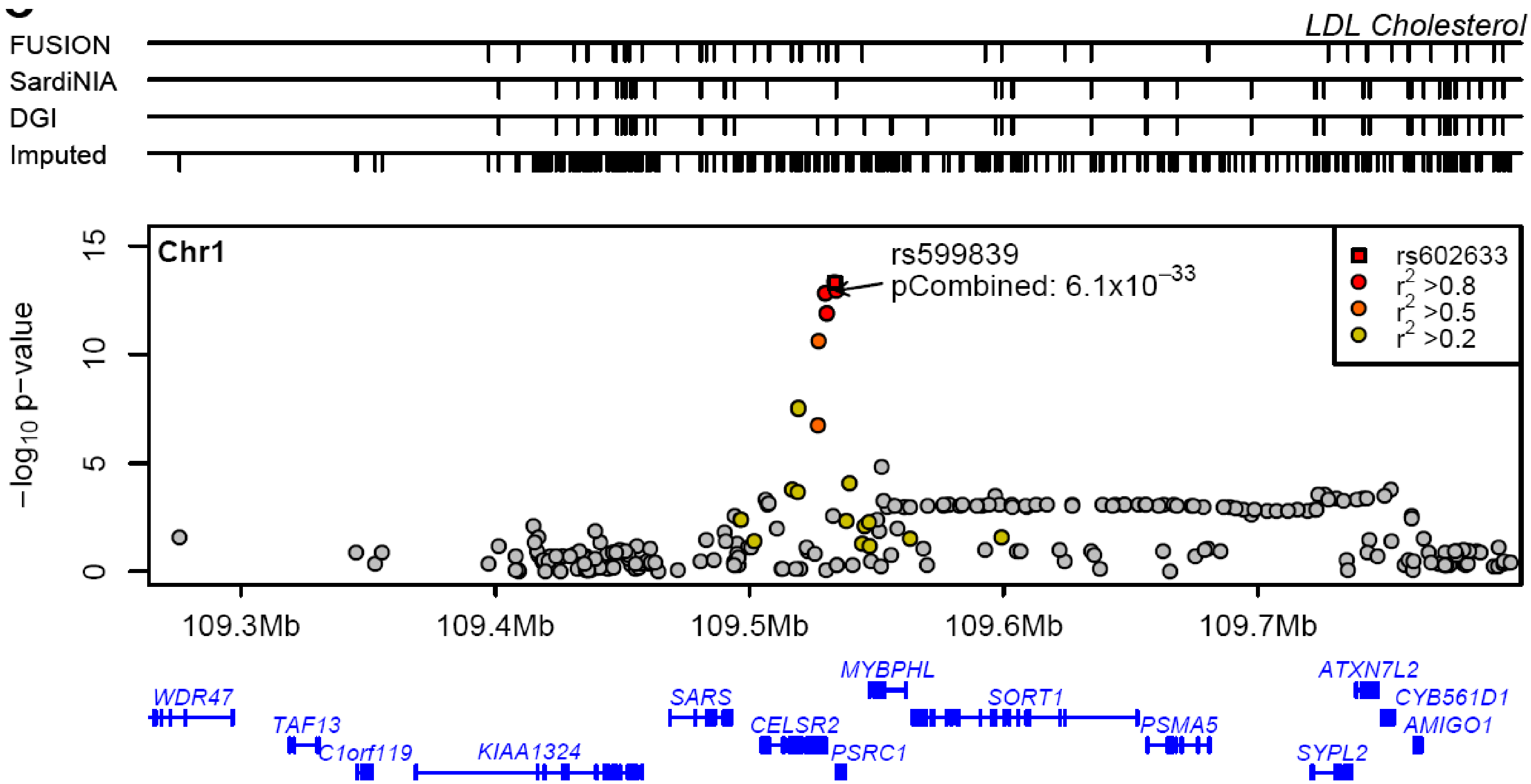
Willer et al, *Nat Genet*, 2008

New HDL Signal For An Old Locus …



What happens when we contrast results with related traits?

New LDL Locus,
Previously Associated with CAD



Comparison with Related Traits: Coronary Artery Disease and LDL-C Alleles

Gene	LDL-C p-value	Frequency CAD cases	Frequency CAD ctrls	CAD p-value	OR
APOE/C1/C4	**$3.0x10^{-43}$**	**.209**	**.184**	**$1.0x10^{-4}$**	**1.17 (1.08-1.28)**
APOE/C1/C4	**$1.2x10^{-9}$**	**.339**	**.319**	**.0068**	**1.10 (1.02-1.18)**
SORT1	**$6.1x10^{-33}$**	**.808**	**.778**	**$1.3x10^{-5}$**	**1.20 (1.10-1.31)**
LDLR	**$4.2x10^{-26}$**	**.902**	**.890**	**$6.7x10^{-4}$**	**1.29 (1.10-1.52)**
APOB	**$5.6x10^{-22}$**	**.830**	**.824**	**.18**	**1.04 (0.95-1.14)**
APOB	**$8.3x10^{-12}$**	**.353**	**.332**	**.0042**	**1.10 (1.03-1.18)**
APOB	**$3.1x10^{-9}$**	**.536**	**.520**	**.028**	**1.07 (1.00-1.14)**
PCSK9	**$3.5x10^{-11}$**	**.825**	**.807**	**.0042**	**1.13 (1.03-1.23)**
NCAN/CILP2	**$2.7x10^{-9}$**	**.922**	**.915**	**.055**	**1.11 (0.98-1.26)**
B3GALT4	**$5.1x10^{-8}$**	**.399**	**.385**	**.039**	**1.07 (0.99-1.14)**
B4GALT4	**$1.0x10^{-6}$**	**.874**	**.865**	**.051**	**1.09 (0.98-1.20)**

Data from WTCCC; Willer et al, *Nature Genetics*, 2008

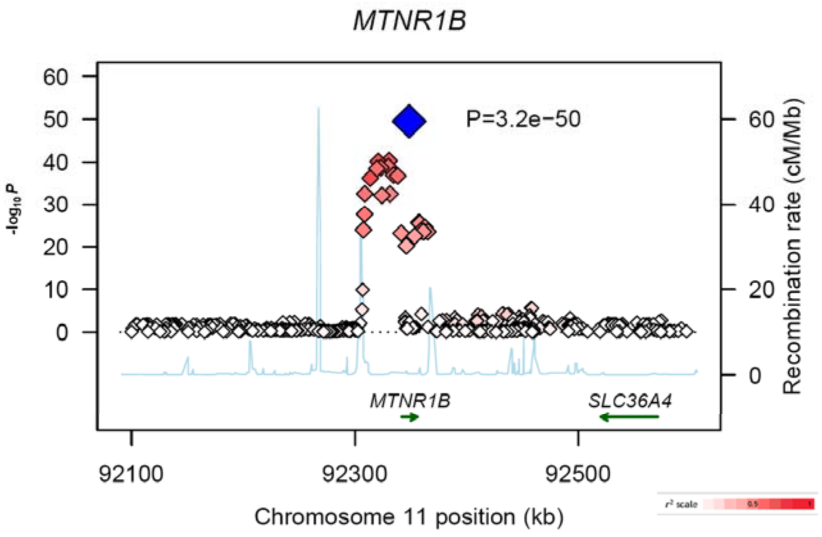
MTNR1B influences glucose levels in non-diabetics and is a new T2D locus

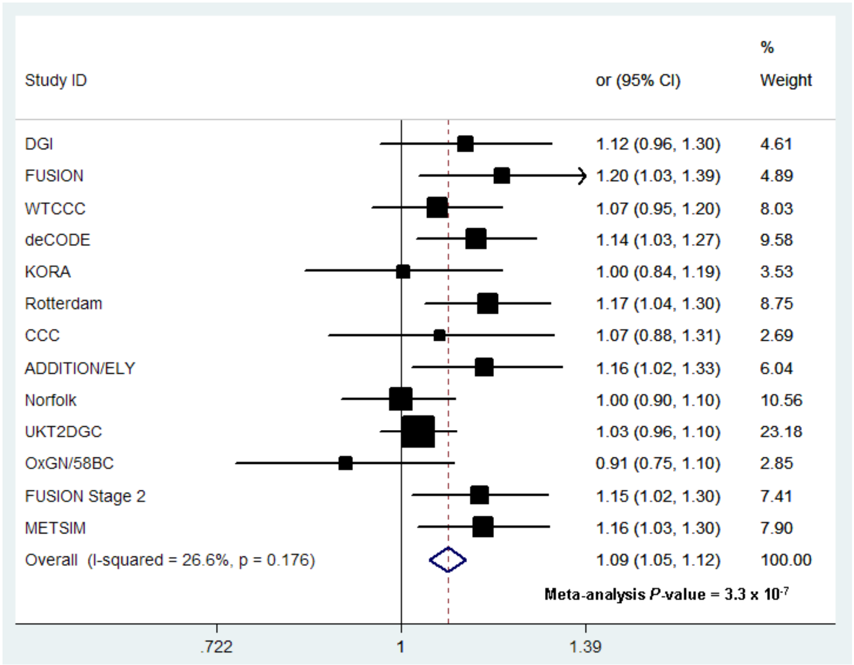
Association with glucose,

36,000 non-diabetics

Association with diabetes,

18,000 cases vs. 64,000 controls





Study ID	or (95% CI)	% Weight
DGI	1.12 (0.96, 1.30)	4.61
FUSION	1.20 (1.03, 1.39)	4.89
WTCCC	1.07 (0.95, 1.20)	8.03
deCODE	1.14 (1.03, 1.27)	9.58
KORA	1.00 (0.84, 1.19)	3.53
Rotterdam	1.17 (1.04, 1.30)	8.75
CCC	1.07 (0.88, 1.31)	2.69
ADDITION/ELY	1.16 (1.02, 1.33)	6.04
Norfolk	1.00 (0.90, 1.10)	10.56
UKT2DGC	1.03 (0.96, 1.10)	23.18
OxGN/58BC	0.91 (0.75, 1.10)	2.85
FUSION Stage 2	1.15 (1.02, 1.30)	7.41
METSIM	1.16 (1.03, 1.30)	7.90
Overall (I-squared = 26.6%, p = 0.176)	1.09 (1.05, 1.12)	100.00

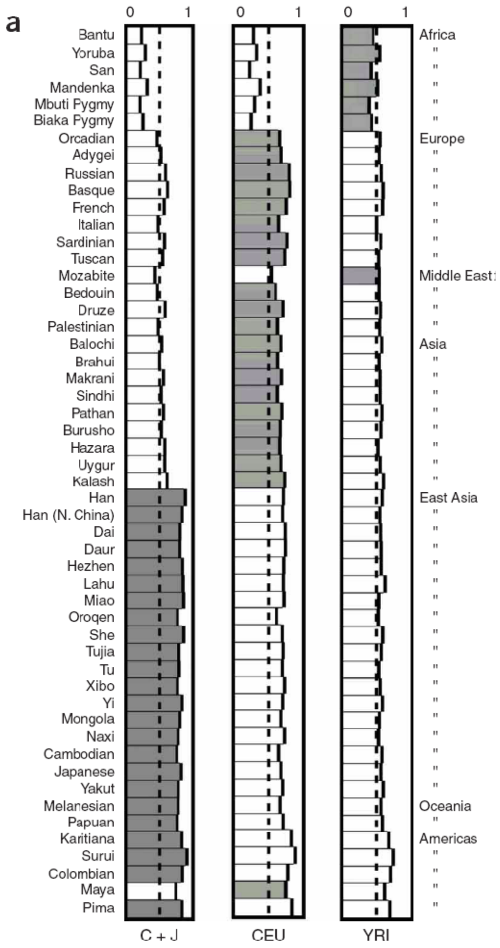
Meta-analysis *P*-value = 3.3 x 10^{-7}

Prokopenko et al, *Nature Genetics*, 2009

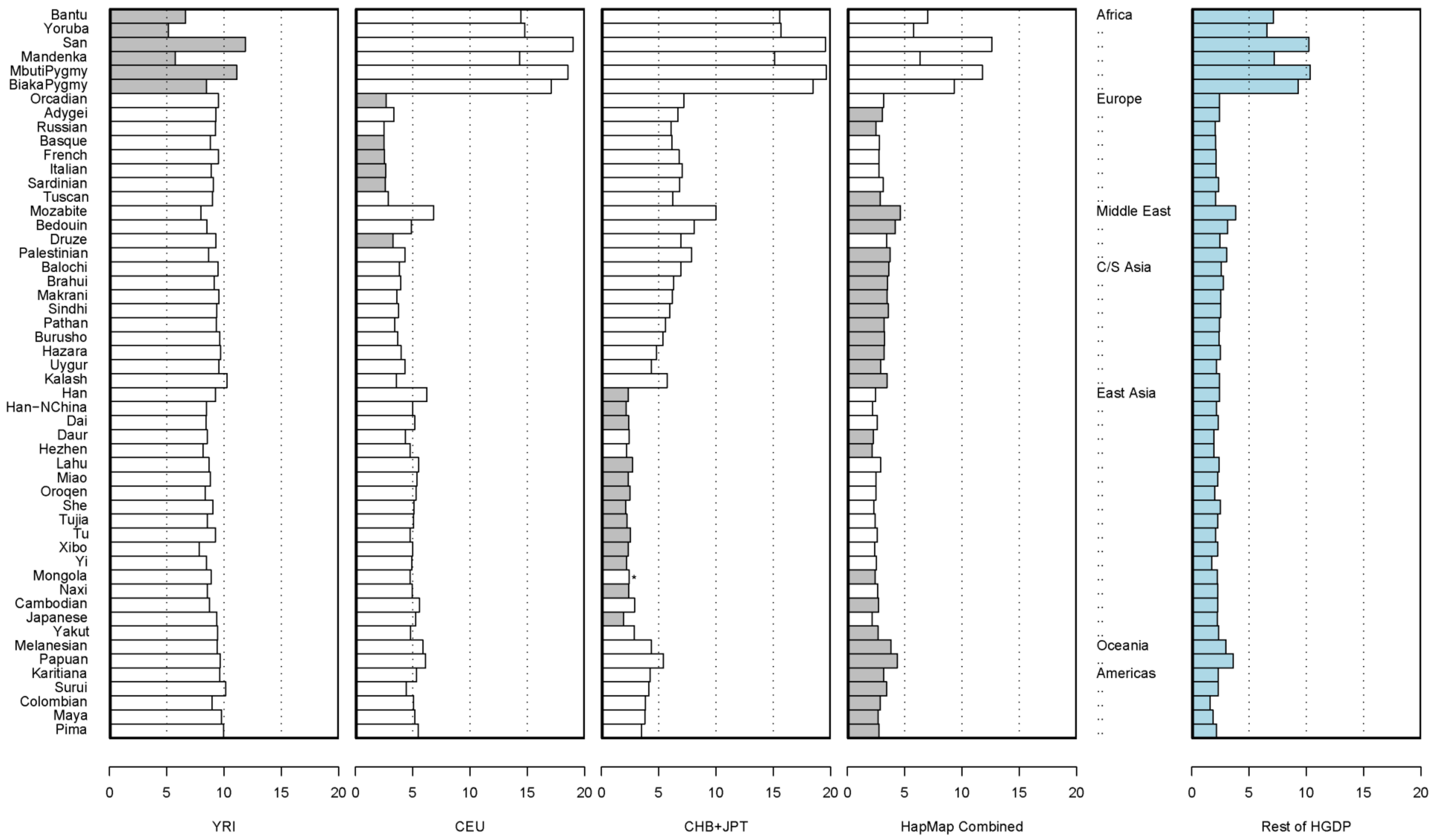
Does Imputation Work Across Populations?

- Conrad et al. (2006) dataset
- 52 regions, each ~330 kb
- Human Genome Diversity Panel
  - ~927 individuals, 52 populations
- 1864 SNPs
  - Grid of 872 SNPs used as tags
  - Predicted genotypes for the other 992 SNPs
  - Compared predictions to actual genotypes

Tag SNP Portability



Percentage of Alleles Imputed Incorrectly



(Evaluation Using ~1 SNP per 10kb in 52 x 300kb regions For Imputation)

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 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’-ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG-3’

Reference Genome

A/C

Predicted Genotype

Shotgun Sequence Data

Sequence Reads

★

5’-ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG-3’

Reference Genome

P(reads|A/A)= 1.0

P(reads|A/C)= 1.0

P(reads|C/C)= 1.0

Possible Genotypes

Shotgun Sequence Data

★

GCTAGCTGATAGCTAG**C**TAGCTGATGAGCCCGA

Sequence Reads

5’-ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG-3’

Reference Genome

P(reads|A/A)= P(C observed, read maps |A/A)

P(reads|A/C)= P(C observed, read maps |A/C)

P(reads|C/C)= P(C observed, read maps |C/C)

Possible Genotypes

Shotgun Sequence Data

GCTAGCTGATAGCTAGCTAGCTGATGAGCCCGA

Sequence Reads

5’-ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG-3’

Reference Genome

P(reads|A/A)= 0.01

P(reads|A/C)= 0.50

P(reads|C/C)= 0.99

Possible Genotypes

Shotgun Sequence Data

★

AGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTG

GCTAGCTGATAGCTAGCTAGCTGATGAGCCCGA

Sequence Reads

5’-ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG-3’

Reference Genome

P(reads|A/A)= 0.0001

P(reads|A/C)= 0.25

P(reads|C/C)= 0.98

Possible Genotypes

Shotgun Sequence Data

★

ATGCTAGCTGATAGCTAG**C**TAGCTGATGAGCC

AGCTGATAGCTAG**C**TAGCTGATGAGCCCGATCGCTG

GCTAGCTGATAGCTAG**C**TAGCTGATGAGCCCGA

Sequence Reads

5’-ACTGGTCGATGCTAGCTGATAGCTAG**C**TAGCTGATGAGCCCGATCGCTGCTAGCTCGACG-3’

Reference Genome

P(reads|A/A)= 0.000001

P(reads|A/C)= 0.125

P(reads|C/C)= 0.97

Possible Genotypes

Shotgun Sequence Data

★

ATAGCTAGATAGCTGATGAGCCCGATCGCTGCTAGCTC

ATGCTAGCTGATAGCTAGCTAGCTGATGAGCC

AGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTG

GCTAGCTGATAGCTAGCTAGCTGATGAGCCCGA

Sequence Reads

5’-ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG-3’

Reference Genome

P(reads|A/A)= 0.00000099

P(reads|A/C)= 0.0625

P(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’-ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG-3’

Reference Genome

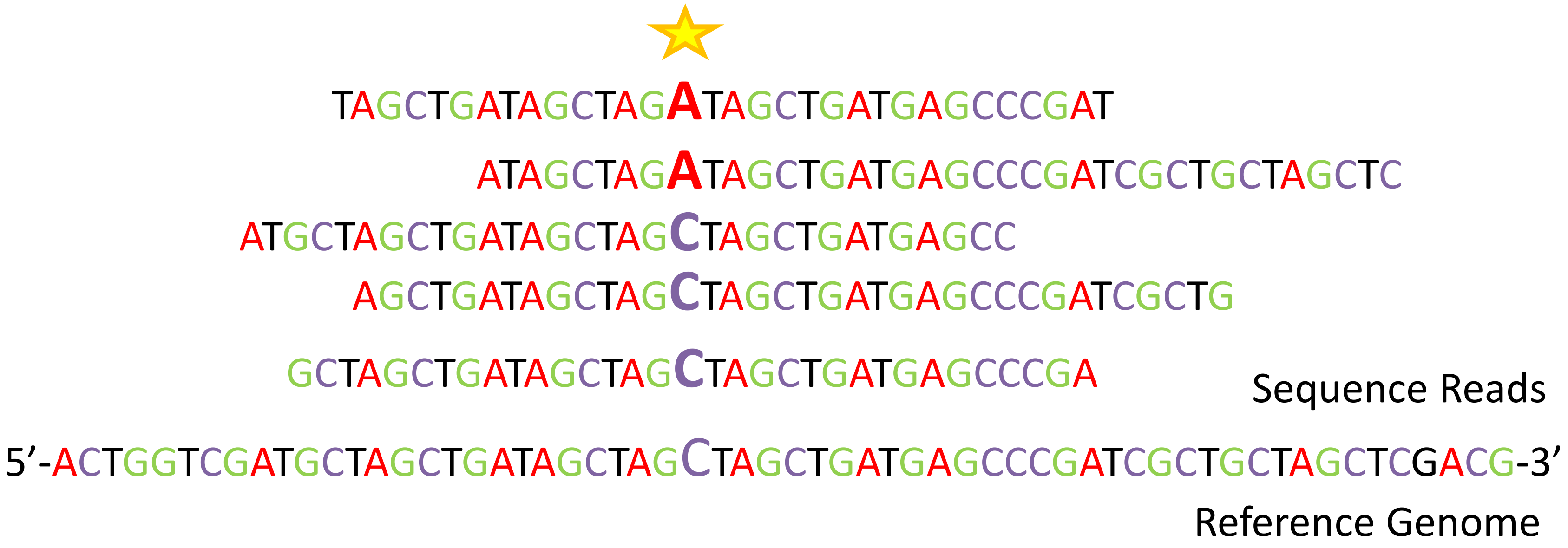
P(reads|A/A)= 0.00000098

P(reads|A/C)= 0.03125

P(reads|C/C)= 0.000097

Possible Genotypes

Shotgun Sequence Data



TAGCTGATAGCTAGATAGCTGATGAGCCCGAT

ATAGCTAGATAGCTGATGAGCCCGATCGCTGCTAGCTC

ATGCTAGCTGATAGCTAGCTAGCTGATGAGCC

AGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTG

GCTAGCTGATAGCTAGCTAGCTGATGAGCCCGA

Sequence Reads

5’-ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG-3’

Reference Genome

P(reads|A/A)= 0.00000098

P(reads|A/C)= 0.03125

P(reads|C/C)= 0.000097

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

P(allele observed, read maps|Genotype)

- Consider a site with possible alleles G/C
  - Let G be the reference allele

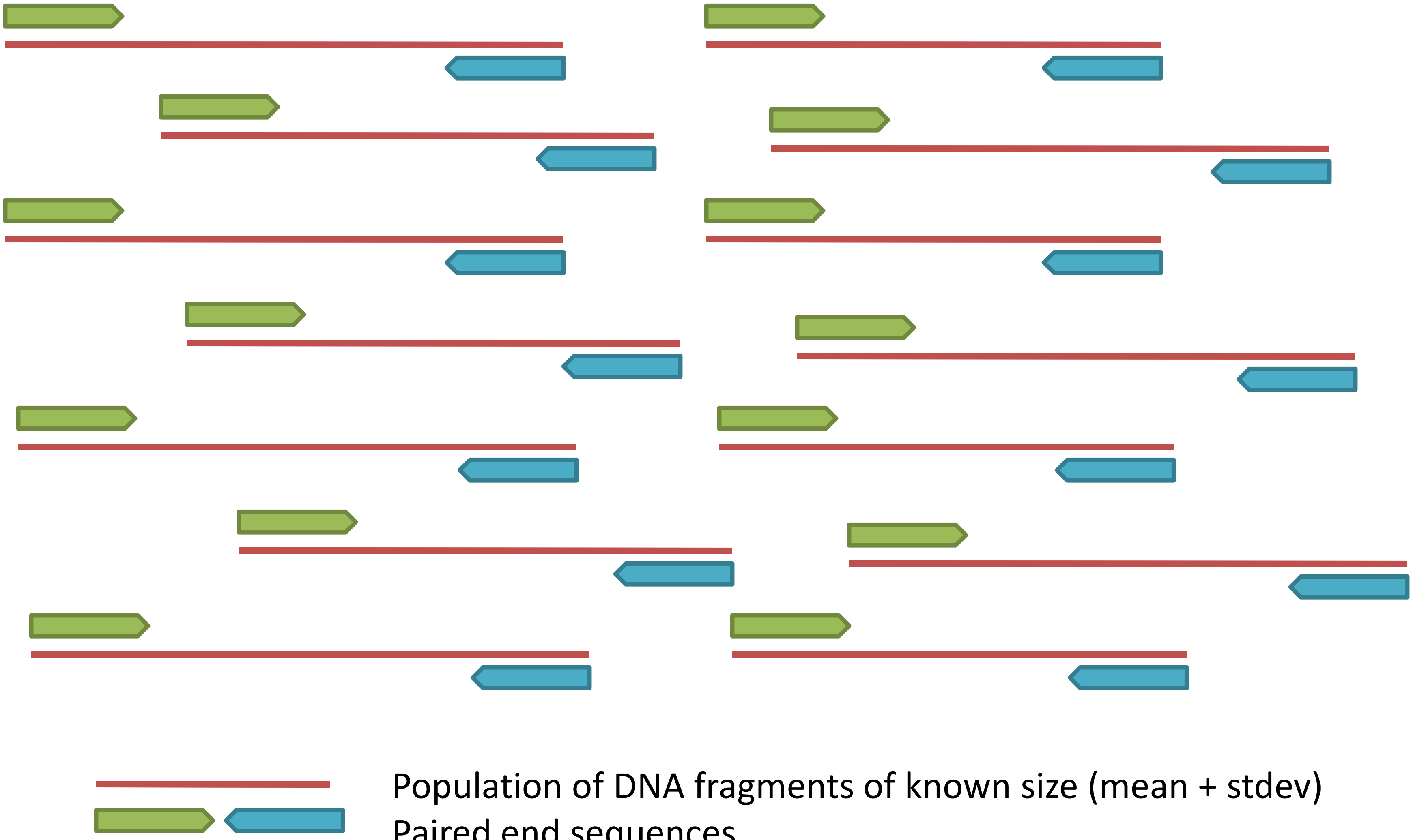
- Assume reads with $<k$ mismatches to reference can be mapped

- True genotype is G/G
  - G observed: $(1-\varepsilon)$ P(rest of read has $<k$ mismatches)
  - C observed: (ε) P(rest of read has $<k-1$ mismatches)
- True genotype is C/C
  - G observed: (ε) P(rest of read has $<k$ mismatches)
  - C observed: $(1-\varepsilon)$ P(rest of read has $<k-1$ mismatches)
- True genotype is G/C
  - G observed: ½ P(rest of read has $<k$ mismatches)
  - C observed: ½ P(rest of read has $<k-1$ mismatches)

Next Generation Sequencing: Key Parameters

- Read length
  - Longer reads can reach more of the genome
- Paired end libraries
  - Reads can be sequenced in pairs with known separation (e.g. 200 +/- 20 bp)
  - Increases “length”
  - Allows sequencing of repetitive regions
- Per base accuracy
- Read depth

Paired End Sequencing

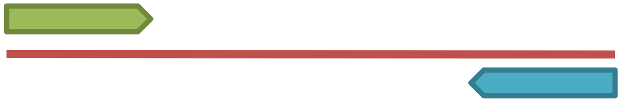


Population of DNA fragments of known size (mean + stdev)

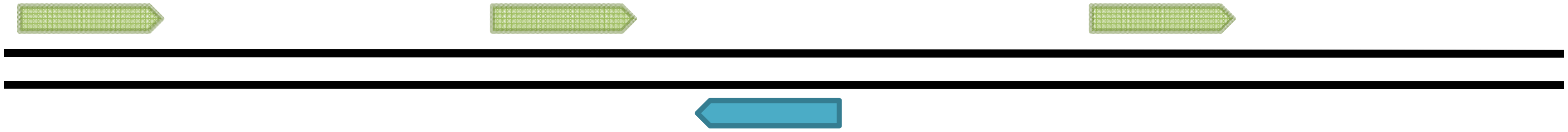
Paired end sequences

Paired End Sequencing

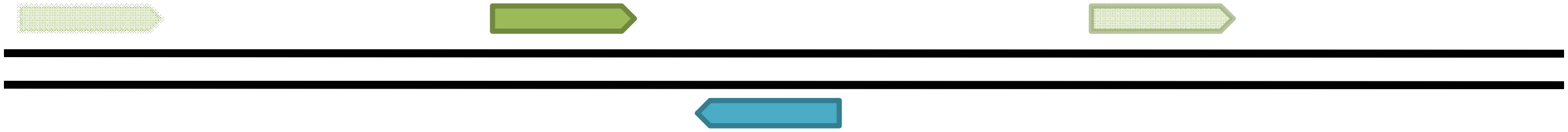
Paired Reads



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

Next Generation Sequencing
and
Imputation

Human Genome Sequencing and Medical Genetics

- Genetic studies of complex diseases, such as cancer and diabetes, require thousands of patients …

- To date, these studies have used on a subset of known variable sites to *“skim”* the genome cost-effectively

- Now, that the human genome has now been sequenced a handful of times …

- … how do we *scale up* sequencing technologies so that we can examine thousands of individuals (or more!)?

Sequence Based Genotype Calls

- Default approach is to use uniform prior
  - 1 difference from reference ~1000 base-pairs or so
  - 66% of these sites are heterozygous

  - Prior that <1/1000 bases differ from reference requires deep sequencing

- If sequencing many individuals, we can use a different prior based on estimates of allele frequency for each site
  - Allele frequency information can dramatically shift prior
  - Low coverage data can be used much more effectively

- Use a model similar to that for imputing HapMap genotypes
  - Increases proportion of called genotypes even further
  - Allows effective use of low depth (even 1-2x) sequence data

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

.	G	.	G	A	.	.	T	.	C	.	T	.	T	.	.	.	.	T	G	.
C	.	A	.	.	.	C	T	C	C	C	.	.	.	C	.	.	.	.	.	.
C	C	A	.	G	.	.	C	T	.	.	.	.	.	.	.	.	.	T	G	.
.	.	.	.	.	.	.	C	T	T	T	.	C	.	.	.	.	.	.	.	.
.	.	.	.	.	.	T	.	.	C	.	.	A	C	C	.	.	A	T	G	.
.	.	.	.	.	.	C	.	C	C	.	G	A	C	C	.	C	A	.	G	G
C	G	A	.	A	.	.	.	.	.	.	G	.	C	.	.	T	.	T	.	.
.	.	.	.	.	.	.	C	.	T	.	T	.	.	.	.	.	.	.	A	.
C	G	.	.	A	.	.	C	T	.	.	.	.	.	C	T	.	G	.	.	.
C	G	A	A	.	.	T	.	.	T	.	T	.	T	.	C	T	.	.	G	C
.	G	A	.	A	T	C	.	.	C	.	T	.	T	T	.	.	.	.	G	.
.	.	A	.	.	.	.	.	.	C	C	.	A	C	.	T	C	A	T	G	.
.	.	A	.	G	.	.	C	.	T	T	.	.	.	T	.	T	G	.	G	C
C	G	A	.	.	.	T	.	.	T	.	.	.	T	T	.	T	.	.	G	C
.	.	.	G	A	C	.	C	.	.	.	.	.	.	.	.	.	.	T	G	.
T	.	.	.	.	T	.	.	C	.	.	.	.	C	C	.	.	.	.	.	.
.	.	.	G	A	T	C	.	C	C	.	G	.	.	C	T	T	.	.	G	C
.	.	.	G	A	.	T	.	T	T	.	T	.	T	T	.	T	.	.	.	.
.	G	A	G	.	.	T	.	T	.	.	G	A	.	.	T	C	G	.	.	C
.	.	A	A	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	G	.

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 using Shotgun Reads

- Generate 10 x 1Mb regions
  - Schaffner et al. (2005) coalescent model calibrated on the HapMap
- Estimate “population” allele frequencies by examining 10,000 simulated chromosomes
- Sequence 100 – 400 individuals at varying depths
  - 0.5% per base-pair error rate, no mapping error
- No external reference panel, sequenced individuals serve as a reference for each other
- False positive rates: ~1 false singleton per 10kb

Yun Li

Simulation Results: Common Sites

- Detection and genotyping of Sites with MAF >5% (2116 simulated sites)

  - **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% error rate, 90.6% at hets
  - 200 people 99.6% error rate, 99.4% at hets
  - 400 people 99.8% error rate, 99.7% at hets

Simulation Results: Rarer Sites

- Detection and genotyping of Sites with MAF 1-2% (425 simulated sites)

  – **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% error rate, 92.9% at hets
  – 200 people 99.4% error rate, 95.0% at hets
  – 400 people 99.6% error rate, 95.9% at hets

Accuracy versus Depth Tradeoffs
100 individuals, 2x coverage

Common sites with MAF > 5%

- 2115 simulated sites

  >2088 sites detected in each case

- 98.84% accuracy with no error
- 98.76% accuracy with 0.1% error
- 98.54% accuracy with 0.5% error
- 98.39% accuracy with 1.0% error

- For 98.84% accuracy at 0.5% error:
  - 4.0x coverage gives 99.53% accuracy
  - 2.4x coverage gives 99.0% accuracy

Yun Li

Accuracy versus Depth Tradeoffs
100 individuals, 2x coverage

Rare sites with MAF .5-1%

- 510 simulated sites

- 307 detected with no error
- 118 detected with 0.1% error
- 63 detected with 0.5% error
- 34 detected with 1.0% error

- To detect 307 sites at 0.5% error
  - Need about 12x coverage

Yun Li

1000 Genomes

A Deep Catalog of Human Genetic Variation

Samples and ELSI Group

Leena Peltonen (co-chair) Sanger Institute
Bartha Knoppers (co-chair) University of Montreal
Aravinda Chakravarti (co-chair) Johns Hopkins
Gonçalo Abecasis University of Michigan
Richard Gibbs Baylor College of Medicine
Lynn Jorde University of Utah
Eric Juengst Case Western Reserve University
Jane Kaye Oxford University
Alastair Kent Genetic Interest Group
Rick Kittles University of Chicago
Jim Mullikin National Human Genome Research Institute
Mike Province Washington University in St. Louis
Charles Rotimi Howard University
Yeyang Su Beijing Genomics Institute
Chris Tyler-Smith Sanger Institute
Ling Yang Beijing Genomics Institute

Data Flow Group (being formed)

Paul Flicek (co-chair) European Bioinforma
Stephen Sherry (co-chair) National Center
Ewan Birney European Bioinformatics Institi
Clive Brown Sanger Institute
David Dooling Washington University in St.
Richard Gibbs Baylor College of Medicine
Sol Katzman University of California, Santa
Hoda Khouri National Center for Biotechnology Information
Martin Shumway National Center for Biotechnology Information
Jun Wang Beijing Genomics Institute
George Weinstock Baylor College of Medicine
(Broad representative)

Production Group

Elaine Mardis (co-chair) Washington University in St. Louis
Stacey Gabriel (co-chair) Broad Institute
Richard Durbin Sanger Institute
Richard Gibbs Baylor College of Medicine
David Jaffe Broad Institute
Ruiqiang Li Beijing Genomics Institute
Donna Muzny Baylor College of Medicine
Chad Nusbaum Broad Institute
Aarno Palotie Sanger Institute
Dan Turner Sanger Institute
Jun Wang Beijing Genomics Institute
Wei Wang Beijing Genomics Institute
Rick Wilson Washington University in St Louis

Steering Committee

Richard Durbin (co-chair) Sanger Institute
David Altshuler (co-chair) Broad / MGH / Harvard
Gonçalo Abecasis University of Michigan
Aravinda Chakravarti Johns Hopkins
Andrew Clark Cornell University
Francis Collins National Human Genome Research Institute
Peter Donnelly Oxford University
Paul Flicek European Bioinformatics Institute
Stacey Gabriel Broad Institute
Richard Gibbs Baylor College of Medicine
Bartha Knoppers University of Montreal
Eric Lander Broad Institute
Elaine Mardis Washington University in St. Louis
Gil McVean Oxford University
Debbie Nickerson University of Washington
Leena Peltonen Sanger Institute
Stephen Sherry National Center for Biotechnology Information
Rick Wilson Washington University in St. Louis
Huanming (Henry) Yang Beijing Genomics Institute

Funders

Alan Schafer Wellcome Trust
Francis Collins National Human Genome Research Institute
Lisa Brooks National Human Genome Research Institute
Audrey Duncanson Wellcome Trust
Adam Felsenfeld National Human Genome Research Institute
Mark Guyer National Human Genome Research Institute
Ruth Jamieson Wellcome Trust
Le Kang Chinese Academy of Sciences
[illegible] Shenzhen Local Municipal Government
Jean McEwen National Human Genome Research Institute
Jane Peterson National Human Genome Research Institute
Anne Pierson National Human Genome Research Institute
Zhiwu Ren National Planning and Development Committee
Jian Wang Beijing Genomics Institute

Analysis Group

Gil McVean (co-chair) Oxford University
Gonçalo Abecasis (co-chair) University of Michigan
David Altshuler Broad / MGH / Harvard
Paul de Bakker Broad / BWH / Harvard
Brian Browning University of Auckland
Sharon Browning University of Auckland
Carlos Bustamante Cornell University
David Carter Sanger Institute
Aravinda Chakravarti Johns Hopkins
Andrew Clark Cornell University
Don Conrad Sanger Institute
Mark Daly Broad / MGH / Harvard
Manolis Dermitzakis Sanger Institute
Peter Donnelly Oxford University
Richard Durbin Sanger Institute
Evan Eichler University of Washington
Paul Flicek European Bioinformatics Institute
Bryan Howie Oxford University
Matt Hurles Sanger Institute
David Jaffe Broad Institute
Lynn Jorde University of Utah
Hoda Khouri National Center for Biotechnology Information
Eric Lander Broad Institute
Charles Lee Brigham and Women's Hospital
Guoqing Li Beijing Genomics Institute
Heng Li Sanger Institute
Ruiqiang Li Beijing Genomics Institute
Yingrui Li Beijing Genomics Institute
Yun Li University of Michigan
Jonathan Marchini Oxford University
Gabor Marth Boston College
Steve McCarroll Broad Institute
Jim Mullikin National Human Genome Research Institute
Simon Myers Oxford University
Rasmus Nielsen University of California, Berkeley
Alkes Price Broad / Harvard
Jonathan Pritchard University of Chicago
Mike Province Washington University in St Louis
Molly Przeworski University of Chicago
Shaun Purcell Broad / MGH / Harvard
Noah Rosenberg University of Michigan
Pardis Sabeti Broad / Harvard
Paul Scheet University of Michigan
Steven Schaffner Broad Institute
Jonathan Sebat Cold Spring Harbor Laboratory
Stephen Sherry National Center for Biotechnology Information
Matthew Stephens University of Chicago
Simon Tavaré University of Southern California
Chris Tyler-Smith Sanger Institute
Jun Wang Beijing Genomics Institute
David Wheeler Baylor College of Medicine
Hongkun Zheng Beijing Genomics Institute

www.1000genomes.org

1000 Genomes Project: Goals

- A public database of essentially all SNPs and detectable CNVs with allele frequency >1% in each of multiple human population samples

- Pioneer and evaluate methods for:
  - Generating data from next-generation sequencing platforms
  - Exchanging and combining data and analytical methods
  - Discovering and genotyping SNPs and CNVs from sequence data
  - Imputation with and from next generation sequencing data

1000 Genomes Project: Plans

- 3 x 400 individuals will be sequenced with:
  - European ancestry
  - East Asian ancestry
  - African ancestry

- 4x sequence coverage per individual planned

- Data collection completed by winter 2009

1000 Genomes Project: Pilots

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

- To date, initial data on 105 unrelated individuals and 2 trios available
  - 11,479,146 unique SNPs
    - 5,074,140 are newly discovered
    - 6,405,006 SNPs already in dbSNP 129
  - ftp.1000genomes.ebi.ac.uk

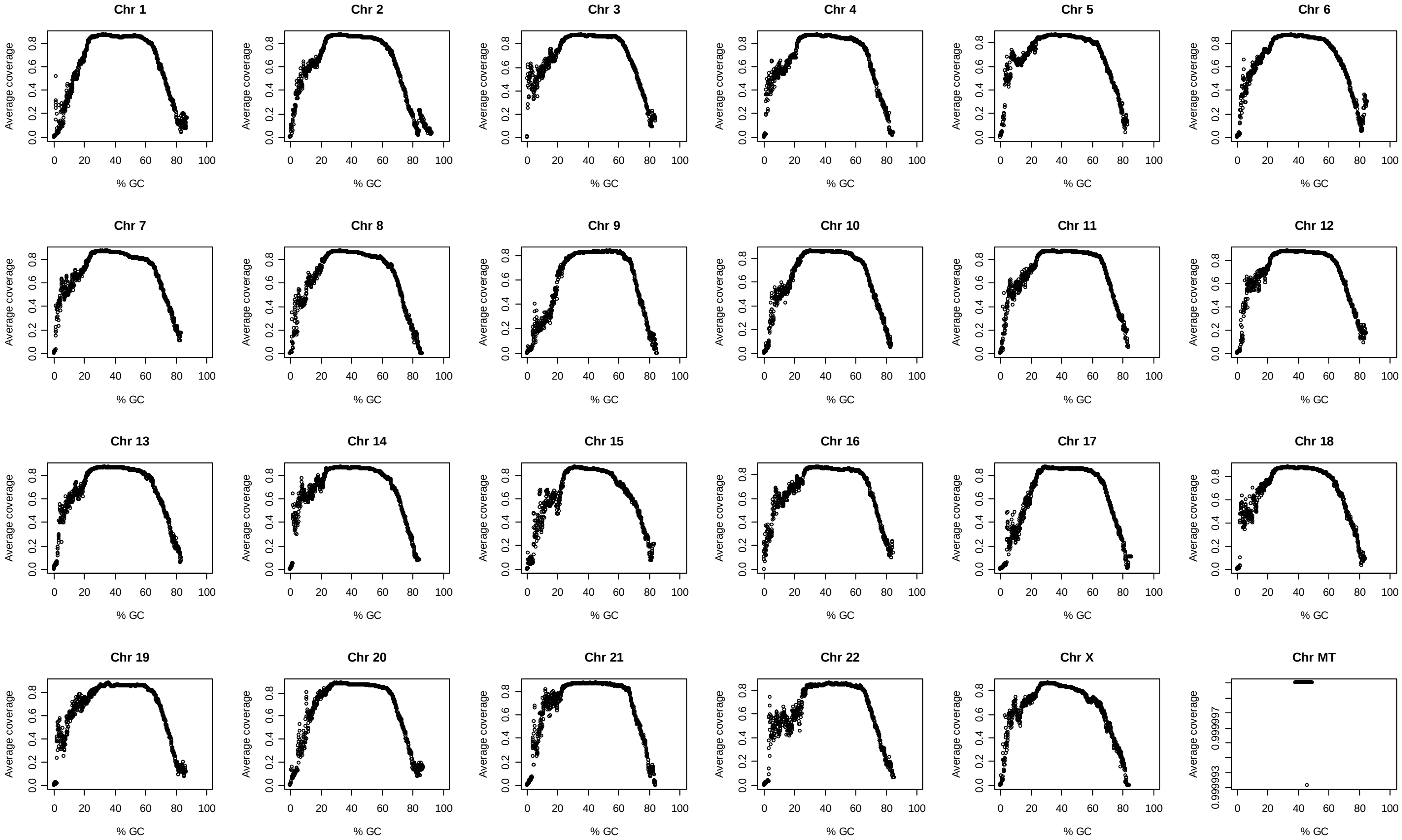
4,047,762 SNPs on CEU trio Comparison with HapMap

- Considered individual genotype calls with Q10

- Compared these calls to HapMap genotypes
  - For sites that match in Phase 2 and Phase 3 HapMap

- Overlapping calls agree >99.9%
  - Genotypes calls made at 98.3% of HapMap sites
  - Variants called at 0.2% of sites where HapMap genotypes for trio are homozygous for the reference

Coverage vs GC for NA12878



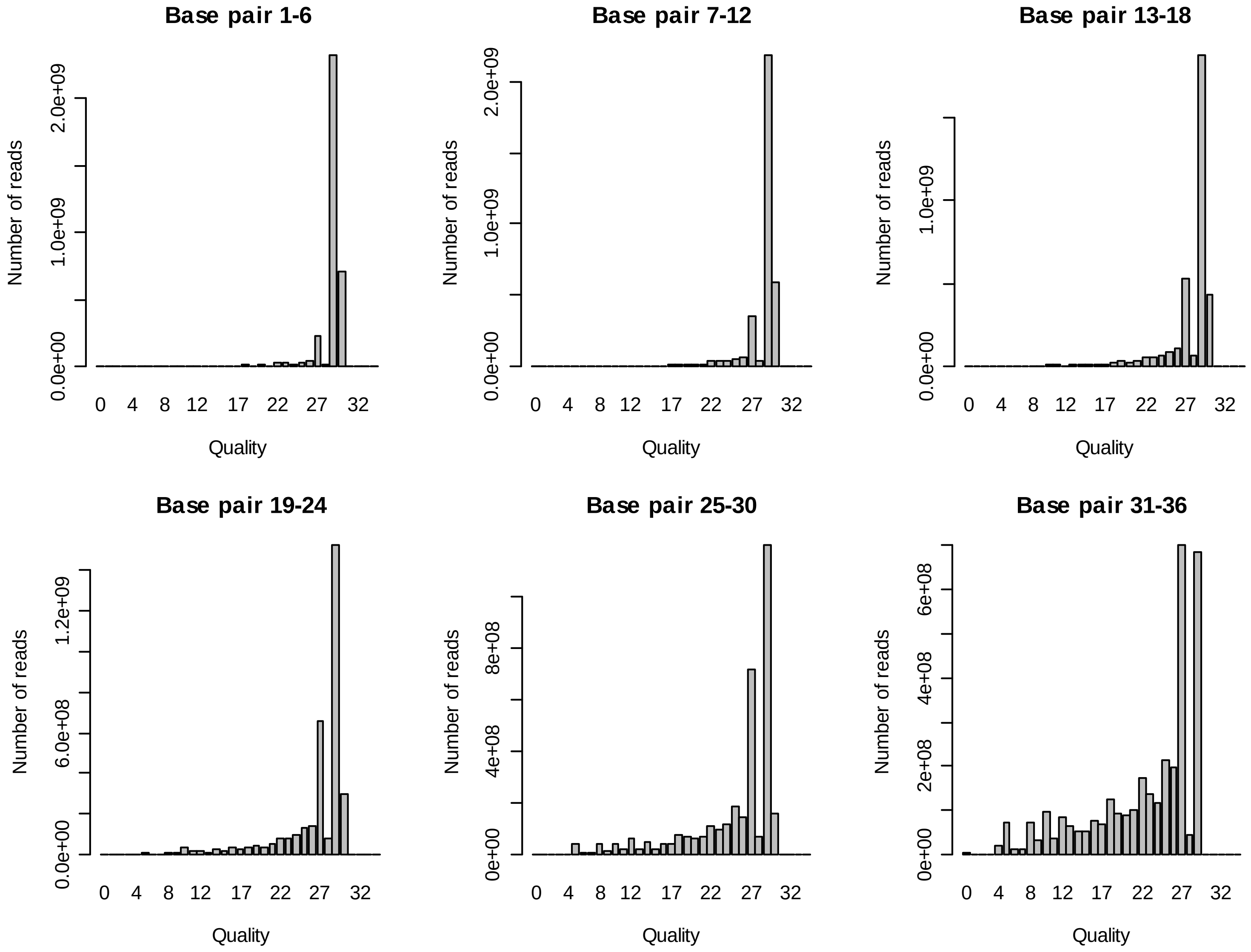
7,809 Mb of 454 sequence (2.7x depth)

2,425 Mb covered sequence (84.9% coverage, 3.4x per covered base)

May 2008

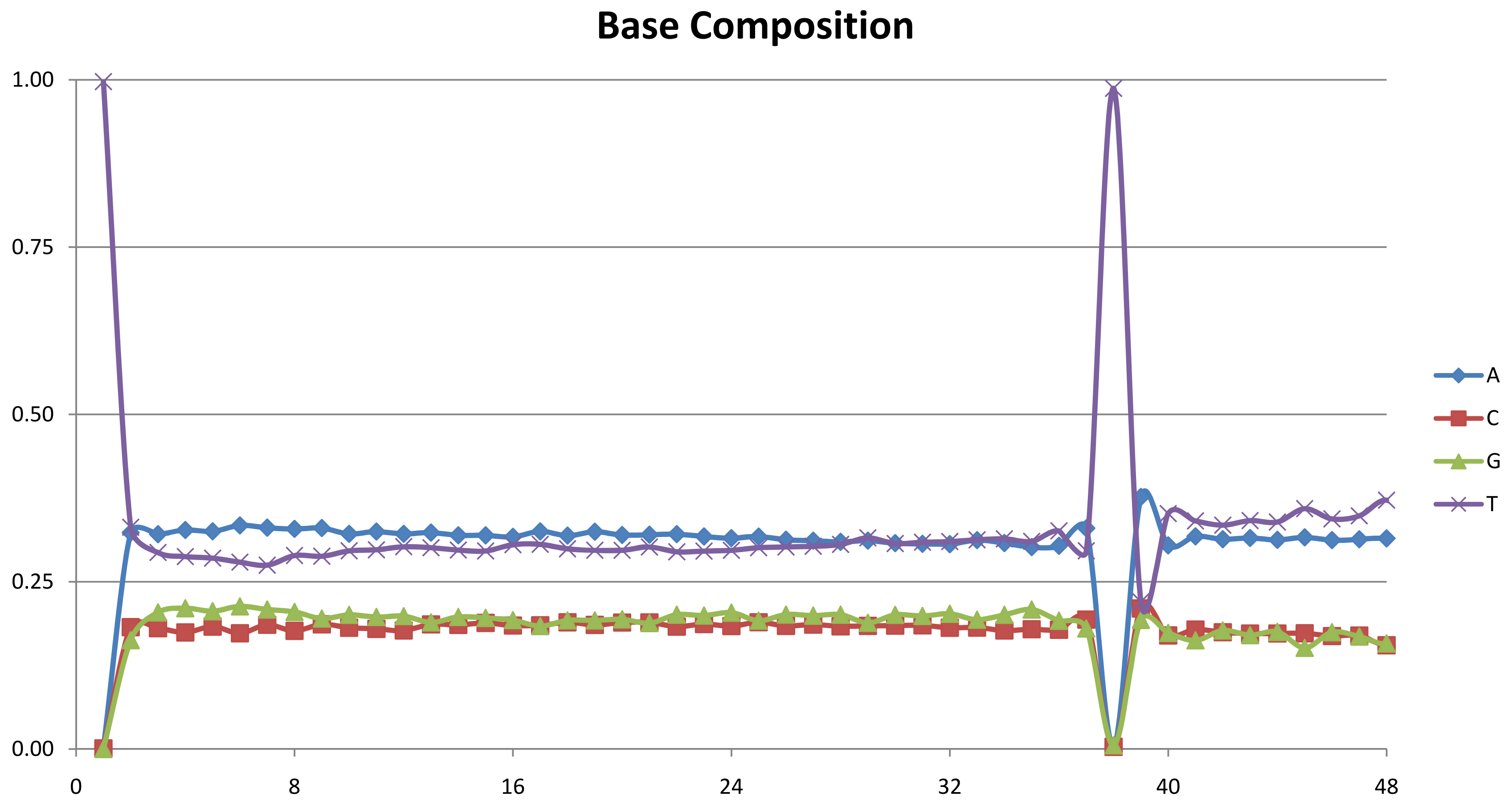
Liming Liang

Solexa Base Quality vs. Read Position

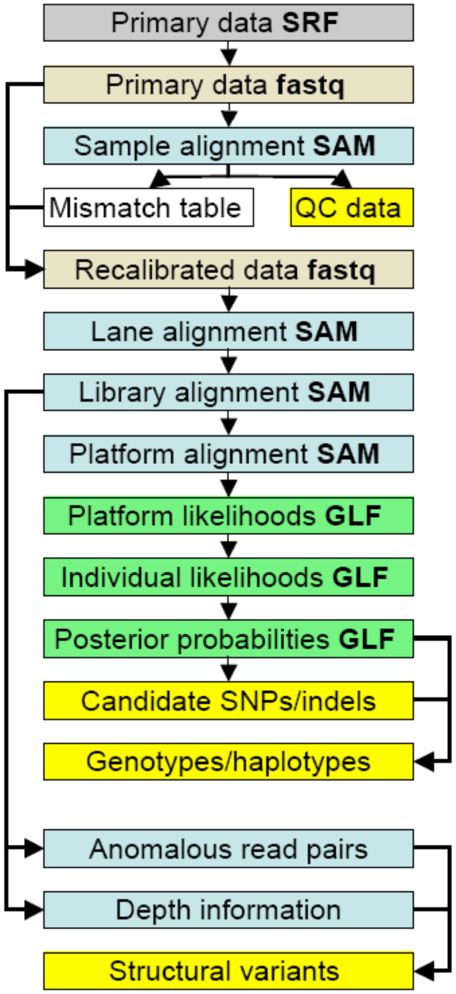


Liming Liang

However, many oddities lurking…



1000 Genome Projects: Data Processing



		Unit (one file per …)	Who? (italics if not done yet)
1. Submit	Primary data **SRF**	*lane*	production centres
2. Extract	Primary data **fastq**	*lane*	DCC
3. Map sample	Sample alignment **SAM**	*lane*	Sanger *to DCC*
4. Recalibrate	Mismatch table QC data	*lane*	Sanger *to DCC*
	Recalibrated data **fastq**	*lane*	Sanger *to DCC*
5. Map	Lane alignment **SAM**	*lane*	Data Processing
6. Merge and remove dups	Library alignment **SAM**	*library*	Data Processing
7. Merge	Platform alignment **SAM**	*platform/individual*	Data Processing
8. Calc likelihoods	Platform likelihoods **GLF**	*platform/individual*	Data Processing
9. Combine l'hoods	Individual likelihoods **GLF**	*individual*	Data Processing
10. Apply priors	Posterior probabilities **GLF**	*individual*	Data Processing
11. Call SNPs/indels	Candidate SNPs/indels	*experiment/population*	Data Processing
12. Call genotypes	Genotypes/haplotypes	*individual*	Data Processing
13. Collect read pair info	Anomalous read pairs	*library*	Structural Variation
13. Collect depth info	Depth information	*library*	Structural Variation
14. Call SVs	Structural variants	*experiment and individual*	Structural Variation

Slide courtesy Richard Durbin

Impact of HapMap Imputation on Power

	Power	
Disease SNP MAF	tagSNPs	Imputation
2.5%	24.4%	56.2%
5%	55.8%	73.8%
10%	77.4%	87.2%
20%	85.6%	92.0%
50%	93.0%	96.0%

Power for Simulated Case Control Studies.
Simulations Ensure Equal Power for Directly Genotype SNPs.

Simulated studies used a tag SNP panel that captures
80% of common variants with pairwise $r^2 > 0.80$.

Impact of HapMap Imputation on Power

	Power	
Disease SNP MAF	**tagSNPs**	**Imputation**
2.5%	24.4%	56.2%
5%	55.8%	73.8%
10%	77.4%	87.2%
20%	85.6%	92.0%
50%	93.0%	96.0%

Power for Simulated Case Control Studies.
Simulations Ensure Equal Power for Directly Genotype SNPs.

Simulated studies used a tag SNP panel that captures
80% of common variants with pairwise $r^2 > 0.80$.

Impact of HapMap Imputation on Power

Disease SNP MAF	Power tagSNPs	 Imputation
2.5%	24.4%	56.2%
5%	55.0%	73.8%
10%	77.4%	87.2%
20%	85.6%	92.0%
50%	93.0%	96.0%

Power Drops for Rare Variants

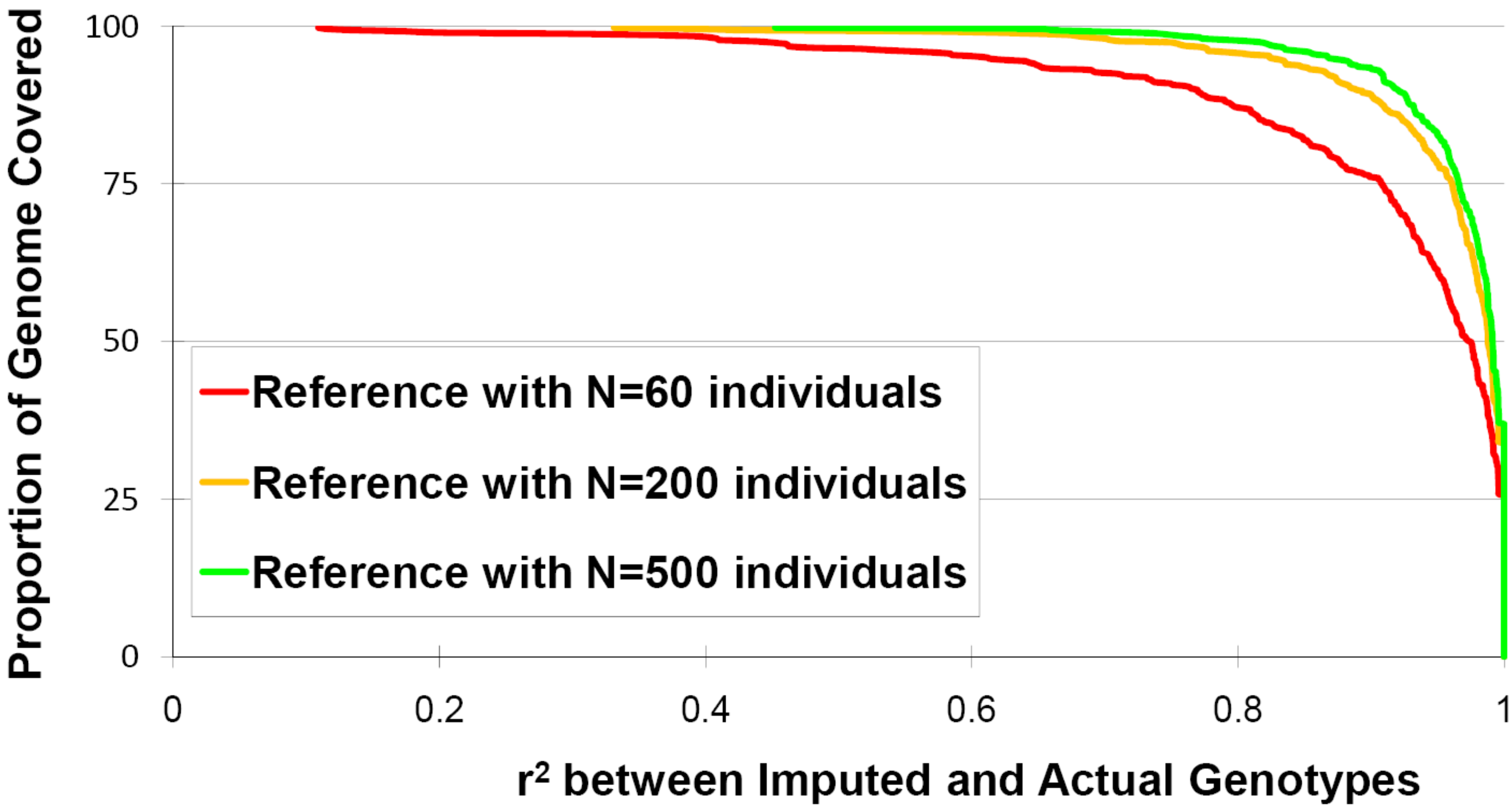
1) Harder to Impute
Should do better using full project data

2) Not Tagged By HapMap
Pilot will address this in part

Power for Simulated Case Control Studies.
Simulations Ensure Equal Power for Directly Genotype SNPs.

Simulated studies used a tag SNP panel that captures
80% of common variants with pairwise $r^2 > 0.80$.

How Might We Use the 1000 Genome Data? Improve Imputation and Power in all GWAS



Increasing reference panels from 60 (HapMap) to 500 individuals (1000 genomes?) should decrease imputation error in GWAS from ~1.4% to ~0.4%.