This Session …

- Genotype Imputation in Families
- Genotype Imputation and Haplotyping with Unrelated Samples
  - Exercise with Mach and Minimac

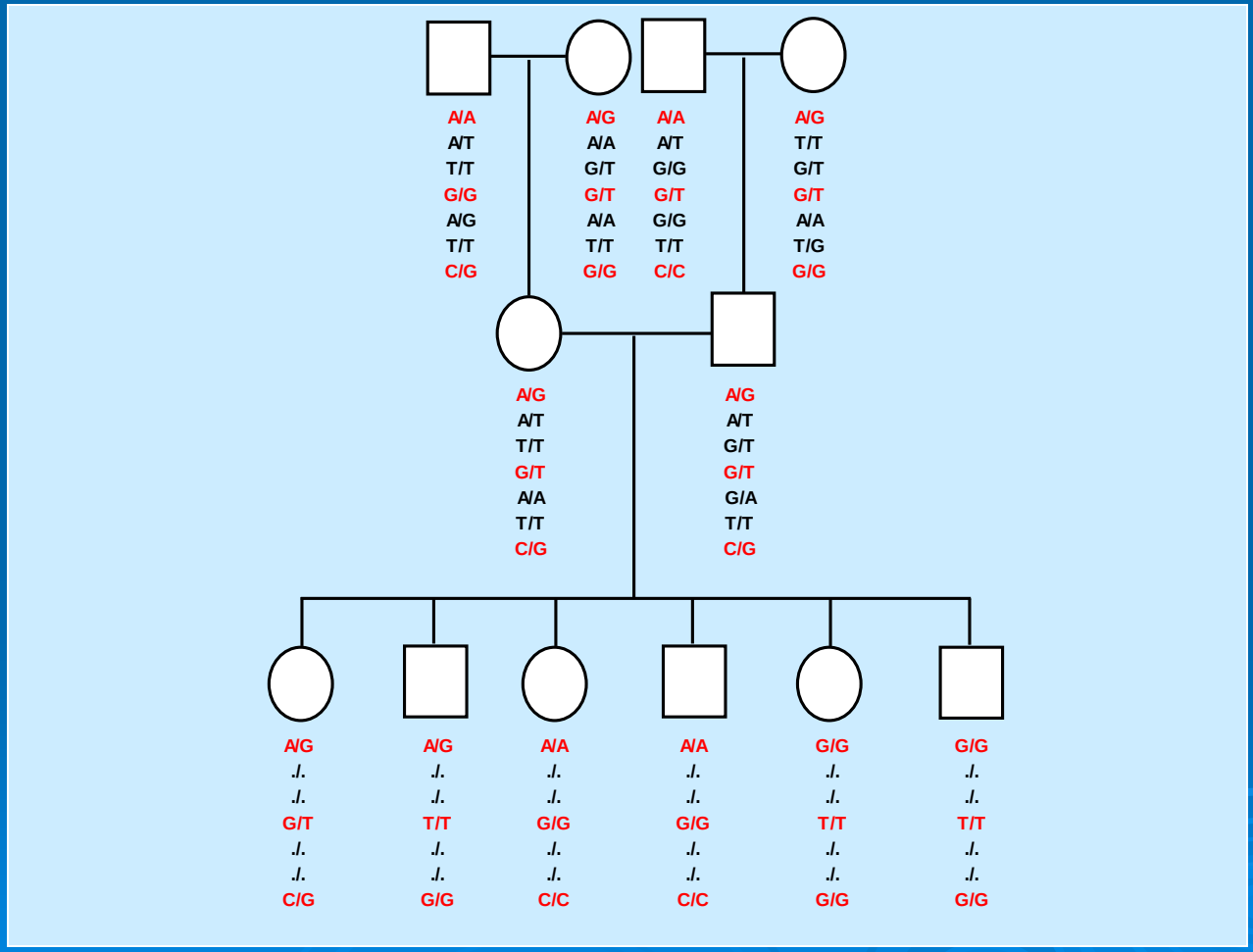
In Silico Genotyping For Genome-Wide Association Studies

Gonçalo Abecasis
Center for Statistical Genetics
University of Michigan

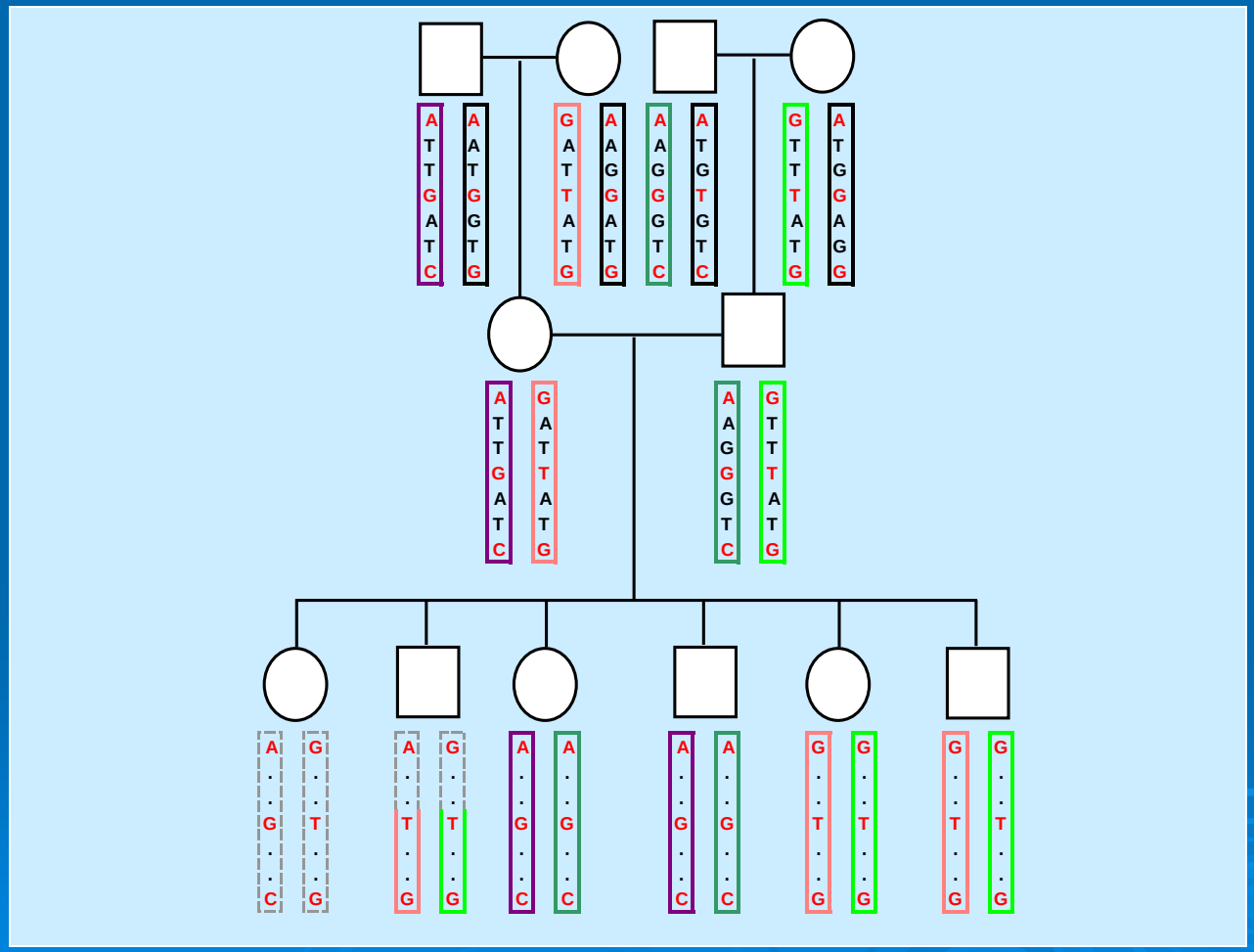
In Silico Genotyping For Family Samples

- Family members will share large segments of chromosomes
- If we genotype many related individuals, we will effectively be genotyping a few chromosomes many times
- In fact, we can:
  - Genotype a few markers on all individuals
  - Find shared haplotype segments
  - Use high-density panel to genotype a few individuals
  - Infer shared segments and then estimate the missing genotypes

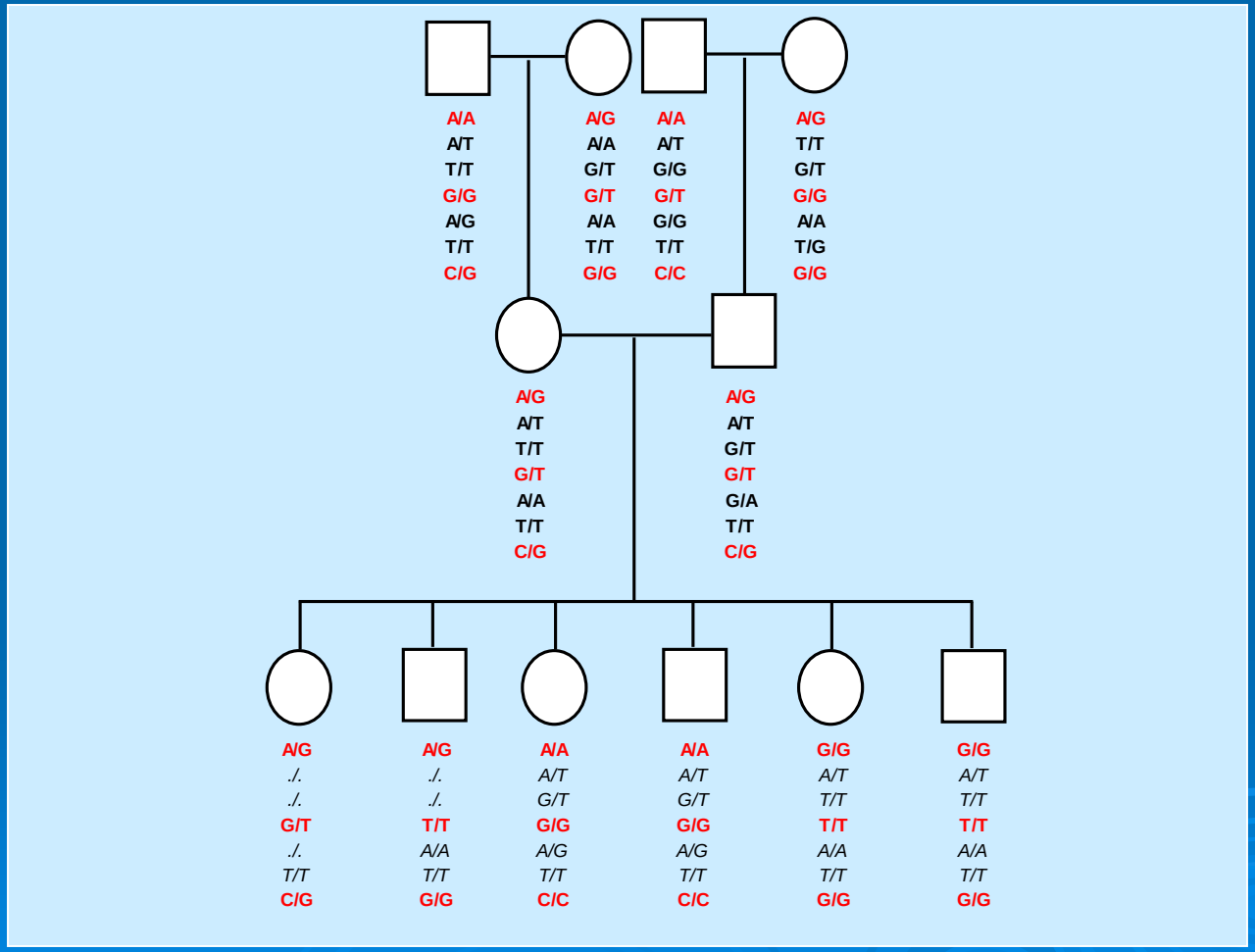
Genotype Inference
Part 1 – Observed Genotype Data



Genotype Inference
Part 2 – Inferring Allele Sharing



Genotype Inference
Part 3 – Imputing Missing Genotypes



Our Approach

- Consider full set of observed genotypes G
- Evaluate pedigree likelihood L for each possible value of each missing genotype g_{ij}
- Posterior probability for each missing genotype

$$P(g_{ij} = x \mid G) = \frac{L(G, g_{ij} = x)}{L(G)}$$

- Implemented both using Elston-Stewart (1972) and Lander-Green (1987) algorithms

Standard Linear Model for Genetic Association

- Model association using a model such as:

$$E(y_i) = \mu + \beta_g g + \beta_c c + \ldots$$

- y_i is the phenotype for individual i
- g_i is the genotype for individual i
  - Simplest coding is to set g_i = number of copies of allele *'1'*
- c_i is a covariate for individual i
  - Covariates could be estimated ancestry, environmental factors…

- β coefficients are estimated covariate, genotype effects
- Model is fitted in variance component framework

Model With Inferred Genotypes

- Replace genotype score g with its expected value:

$$E(y_i) = \mu + \beta_g \bar{g} + \beta_c c + \ldots$$

- Where

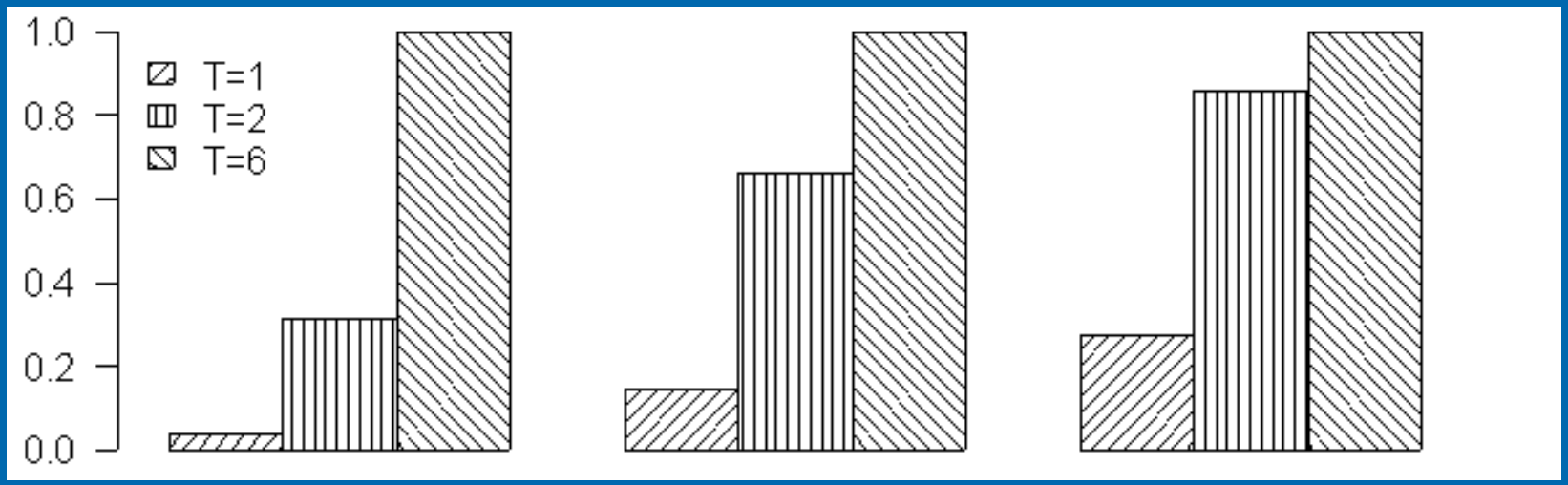
$$\bar{g}_i = 2P(g_i = 2 \mid G) + P(g_i = 1 \mid G)$$

- Association test can then be implemented as a score test or as a likelihood ratio test

- Alternatives would be to
  - (a) impute genotypes with large posterior probabilities; or
  - (b) integrate joint distribution of unobserved genotypes in family

Power in Sibships of Size 6

Without Parental Genotype Data



Analyze Observed Data

Impute when Posterior >.99

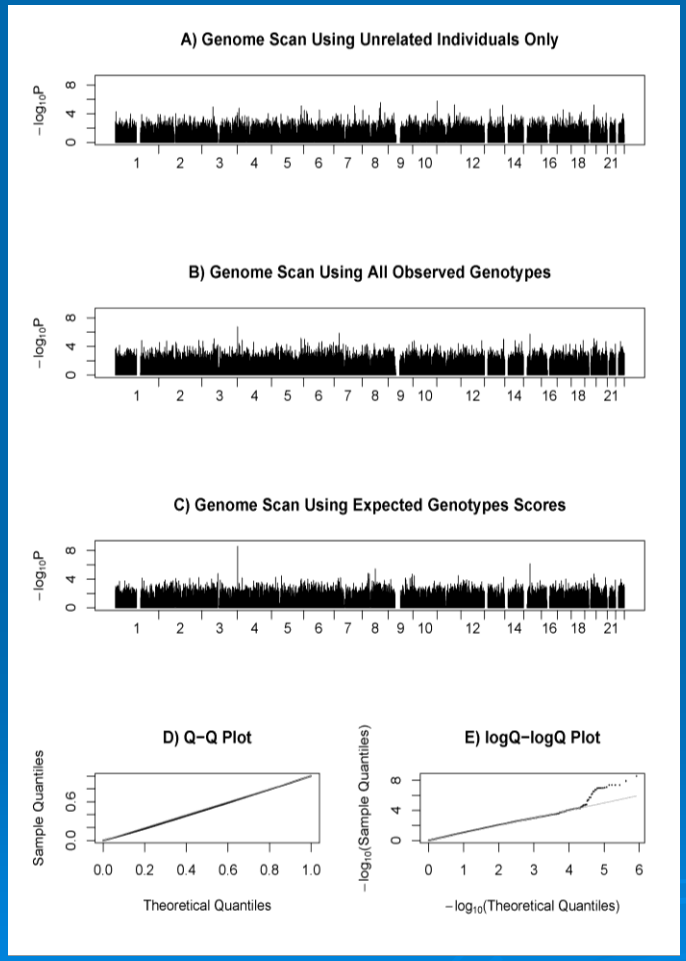
Using Expected Genotype Score

T is the number of genotyped offspring.
QTL explains 5% of variance, polygenes explain 35%,
250 sibships, $\alpha = 0.001$.

Application: Gene Expression Data

- Cheung et al (2005) carried out a genome wide association with 27 expression levels as traits
- Measured in grandparents and parents of CEPH pedigrees and took advantage of HapMap I genotypes
- TSC genotypes also available for ~6000 SNPs in the offspring of each CEPH family

Example: Gene Expression Data

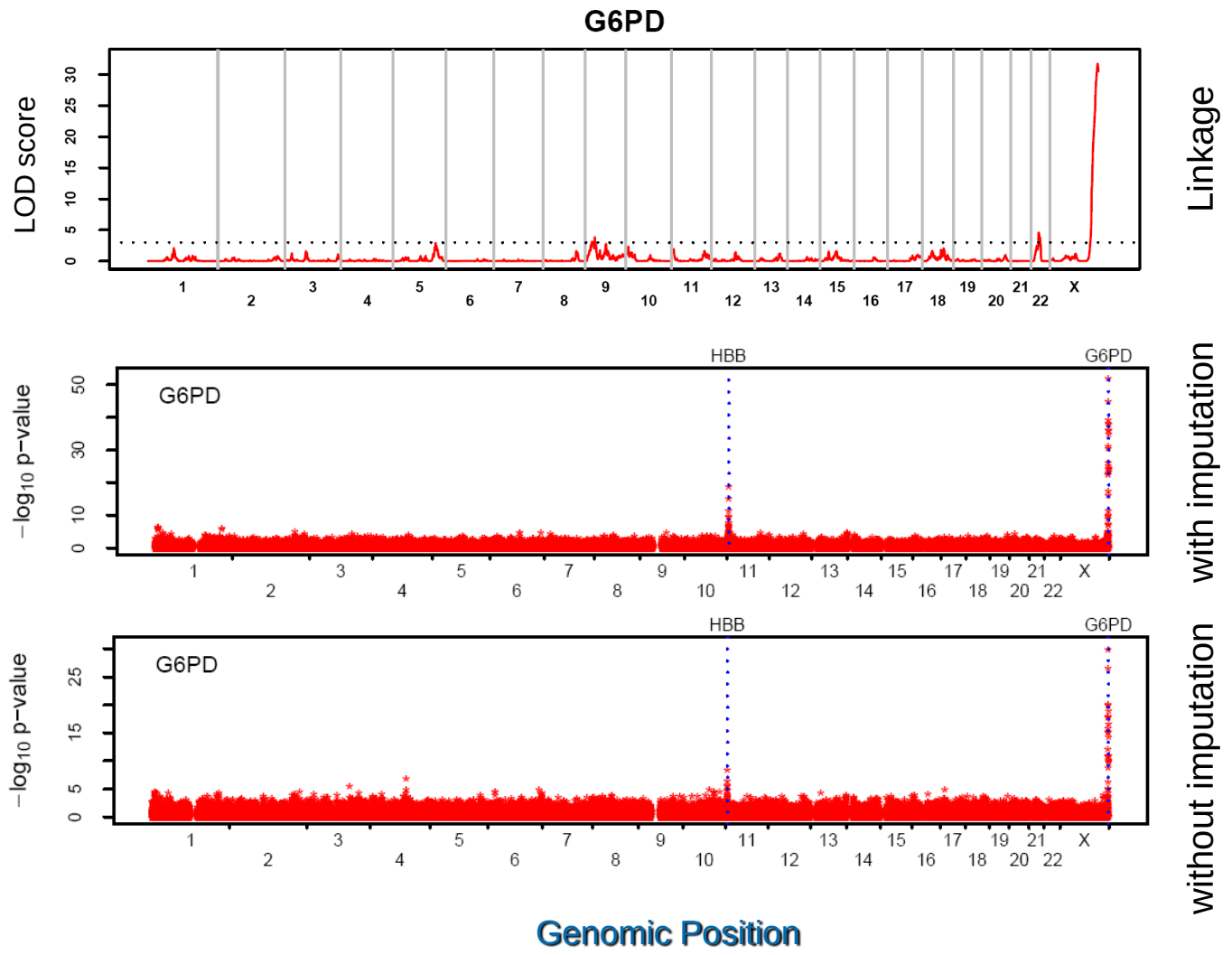


- Panels show GWA scan with CTBP1 expression as outcome
  - Gene is at start of chromosome 4
- Using observed genotypes, most significant association maps in *cis* for 15/27 traits
  - 12 of these reach $p < 5 * 10^{-8}$
- Using inferred genotypes, most significant association maps in *cis* for 19/27 traits
  - 15 of these reach $p < 5 * 10^{-8}$
- Data from Cheung et al. (2005)

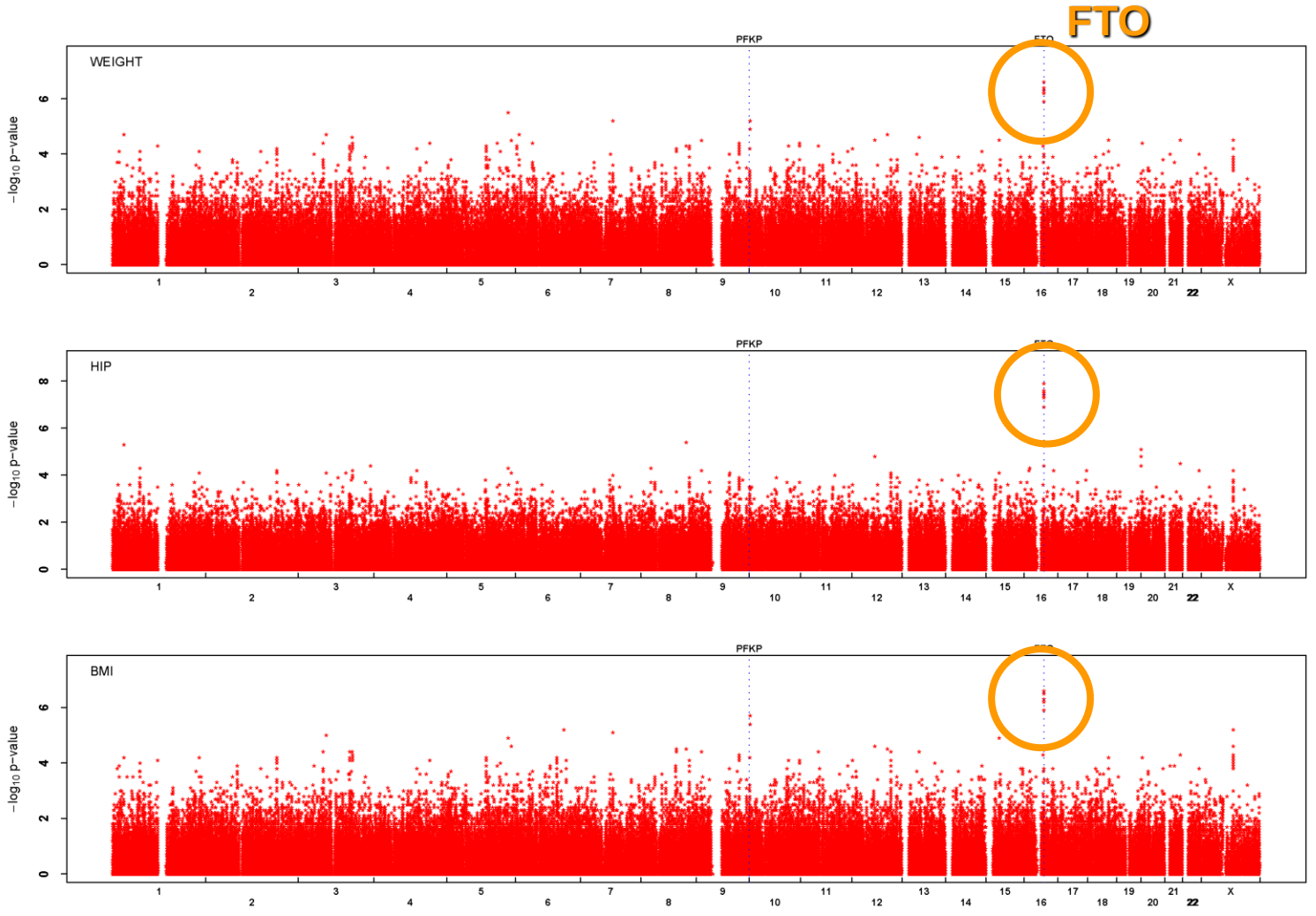
Quantitative Trait GWAS in Sardinia

- 6,148 Sardinians from 4 towns in Ogliastra
- Measured 98 aging related quantitative traits
- Genotyping:
  - 10,000 SNPs measured in ~4,500 individuals
  - 500,000 SNPs measured in ~1,400 individuals

An Example Where We Know The Answer



FTO and Obesity Related Traits



Scuteri et al, PLoS Genetics, 2007

So Far …

- Inferring unobserved genotypes
- Estimate genotypes for relatives of individuals in genome-wide association scan
  - Increase power
- Tests for association in families where only a few individuals are genotyped in detail
  - Limited genotypes may be available for their relatives

Coming Up

- More *in silico* genotyping!
- Estimate genotypes for untyped markers, by combining study sample with Hapmap
  - Facilitate comparisons across studies
- Evaluating quality of the inferred genotypes

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

In Silico Genotyping For Case Control Samples

- In families, we expected relatively long stretches of shared chromosome
- In unrelated individuals, these stretches will typically be much shorter
- The plan is still to identify stretches of shared chromosome between individuals…
- … we then infer intervening genotypes by contrasting study samples with densely typed HapMap samples

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
```

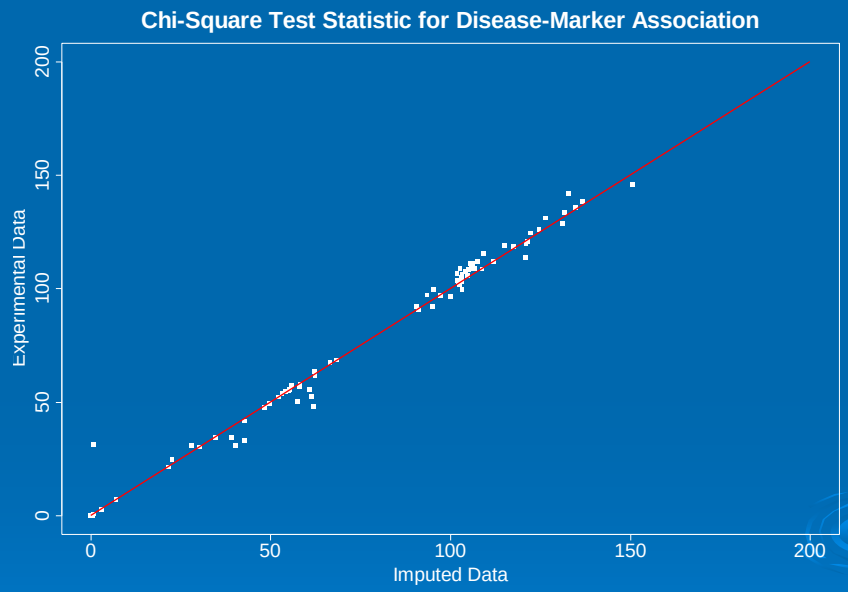
Implementation

- 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 This Really Work? Preliminary Results

- Used 11 tag SNPs to predict 84 SNPs in CFH
- Predicted genotypes differ from original ~1.8% of the time
- Reasonably similar results possible using methods, such as, PHASE and fastPHASE

Comparison of Test Statistics, Truth vs. Imputed



Does This Really Work?

- Used about ~300,000 SNPs from Illumina HumanHap300 to impute 2.1M HapMap SNPs in 2500 individuals from a study of type II diabetes (Scott et al, Science, 2007)

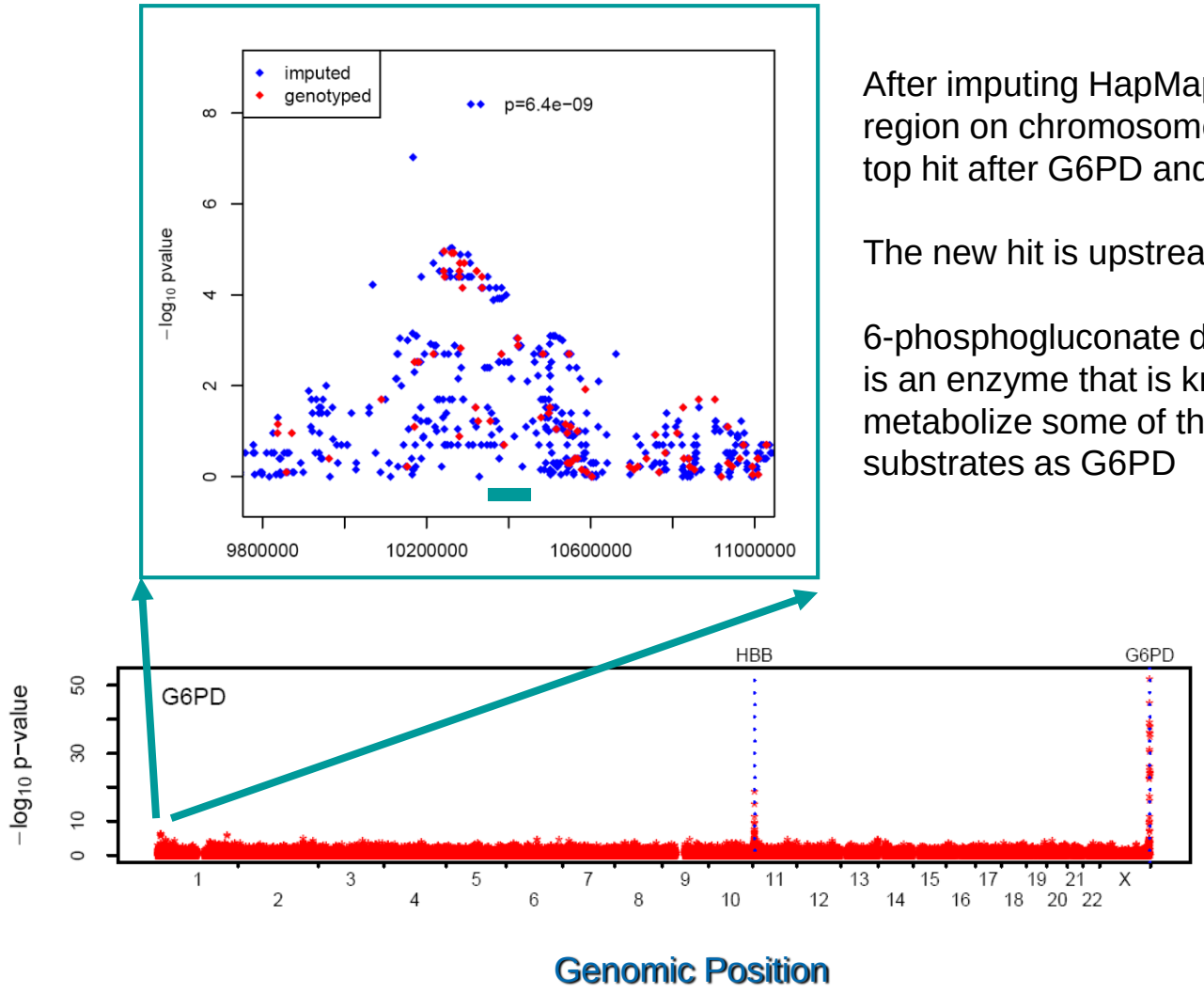
- Compared imputed genotypes with actual experimental genotypes in a candidate region on chromosome 14
  - 1190 individuals, 521 markers not on Illumina chip

- Results of comparison
  - Average r^2 with true genotypes 0.92 (median 0.97)
  - 1.4% of imputed alleles mismatch original
  - 2.8% of imputed genotypes mismatch
  - Most errors concentrated on worst 3% of SNPs

Does this really, really work?

- 90 GAIN psoriasis study samples were re-genotyped for 906,600 SNPs using the Affymetrix 6.0 chip.
- Comparison of 15,844,334 genotypes for 218,039 SNPs that overlap between the Perlegen and Affymetrix chips resulted in discrepancy rate of 0.25% per genotype (0.12% per allele).
- Comparison of 57,747,244 imputed and experimentally derived genotypes for 661,881 non-Perlegen SNPs present in the Affymetrix 6.0 array resulted in a discrepancy rate of 1.80% per genotype (0.91% per allele).
- Overall, the average r^2 between imputed genotypes and their experimental counterparts was 0.93. This statistic exceeded 0.80 for >90% of SNPs.

Back to Sardinia G6PD Activity Example …

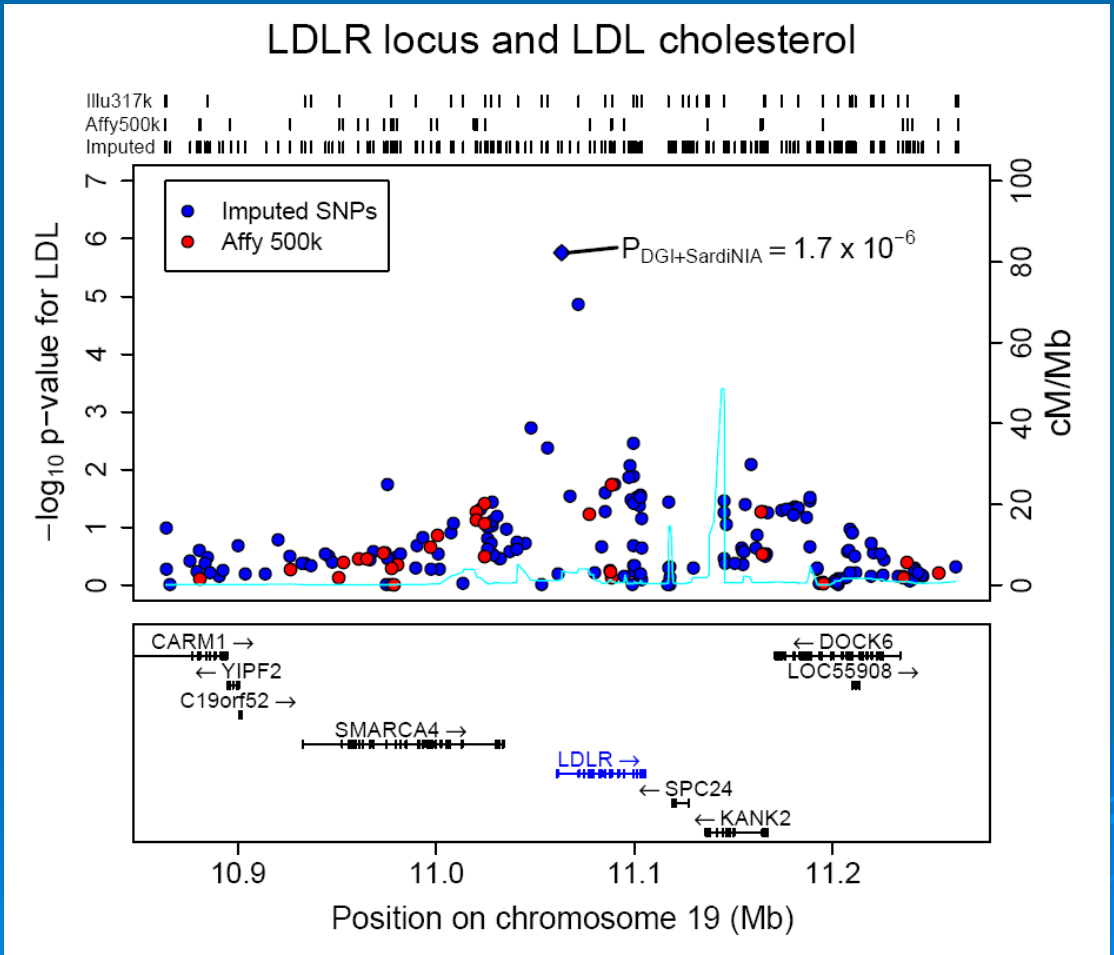


After imputing HapMap SNPs a region on chromosome 1 becomes top hit after G6PD and HBB

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



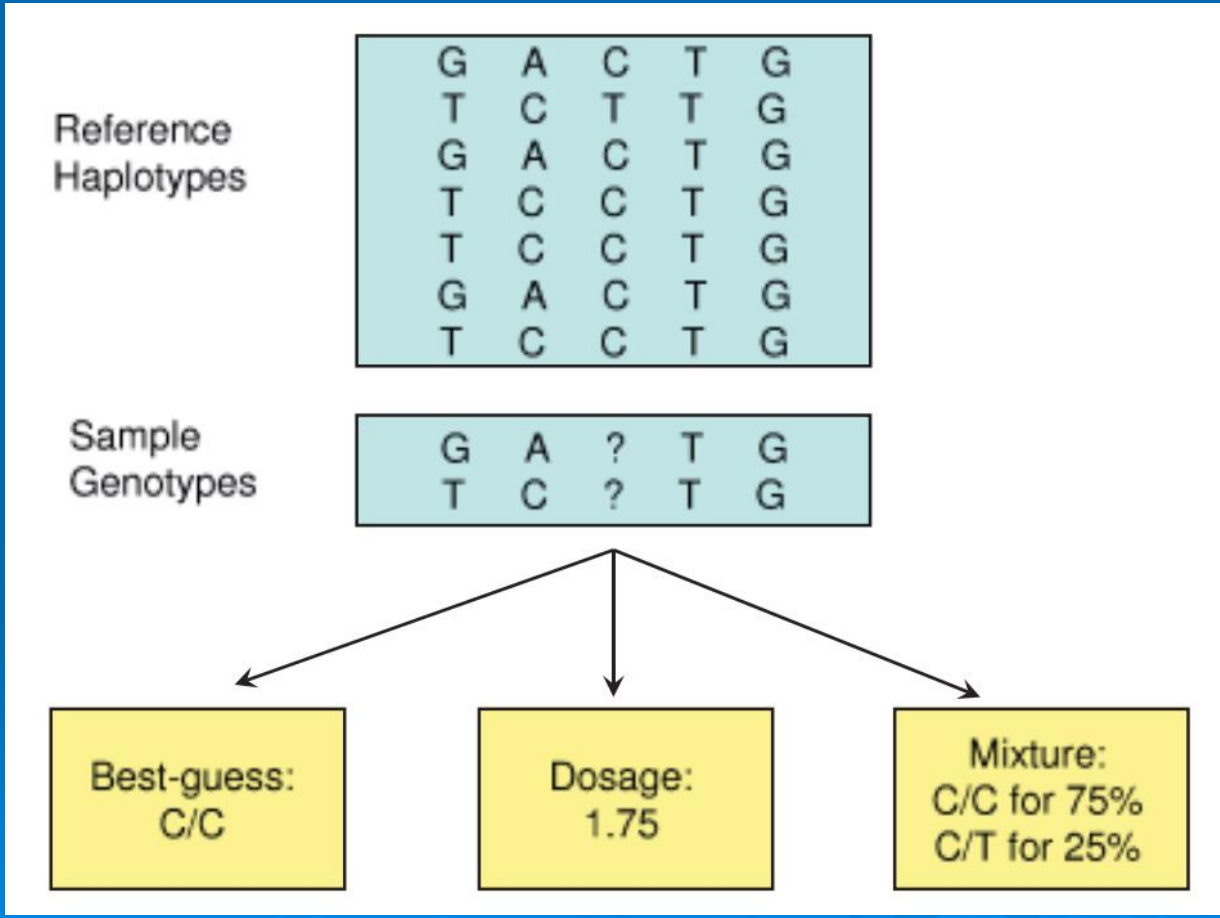
Does Imputation Improve Power?

	Power		
Disease SNP MAF	**tagSNPs**	**Multi-marker tag**	**Imputation**
2.5%	24.4%	25.0%	56.2%
5%	55.8%	56.4%	73.8%
10%	77.4%	78.4%	87.2%
20%	85.6%	86.2%	92.0%
50%	93.0%	93.6%	96.0%

Power for Simulated Case Control Studies

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

Choices for Analysis of Imputed Genotypes



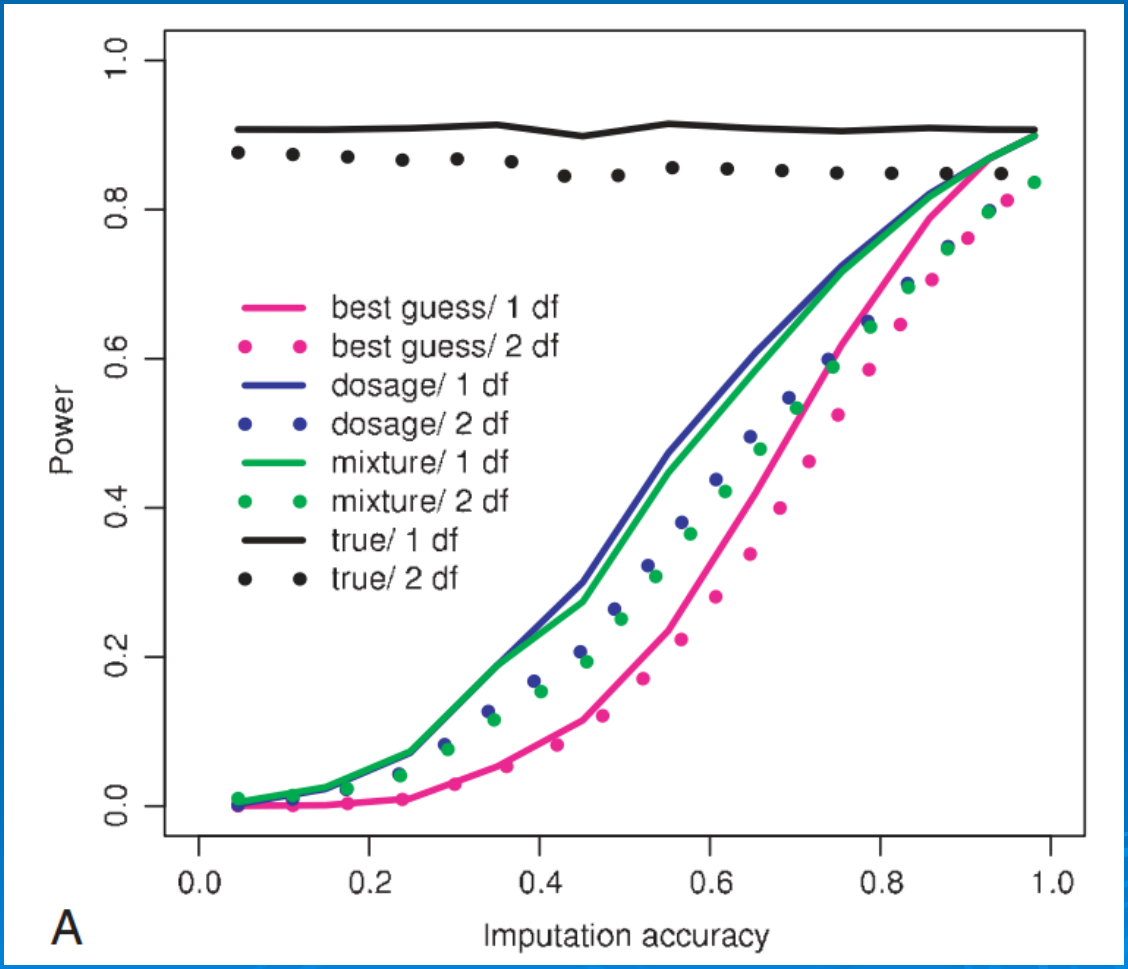
Choices for Analysis

Scenario	N	H^2	Power: Best Guess	Power: Dosage	Power: Mixture
Large sample, small effect					
	1000	3%	63.5%	66.0%	66.8%
Small sample, large effect					
	50	60%	70.1%	75.5%	85.0%

- When effect sizes are small, difference between dosage and mixture models becomes even smaller
- 3% of variance explained would now be considered a large effect for most traits.

Zheng et al, Genetic Epidemiology, 2011

Choices for Analysis

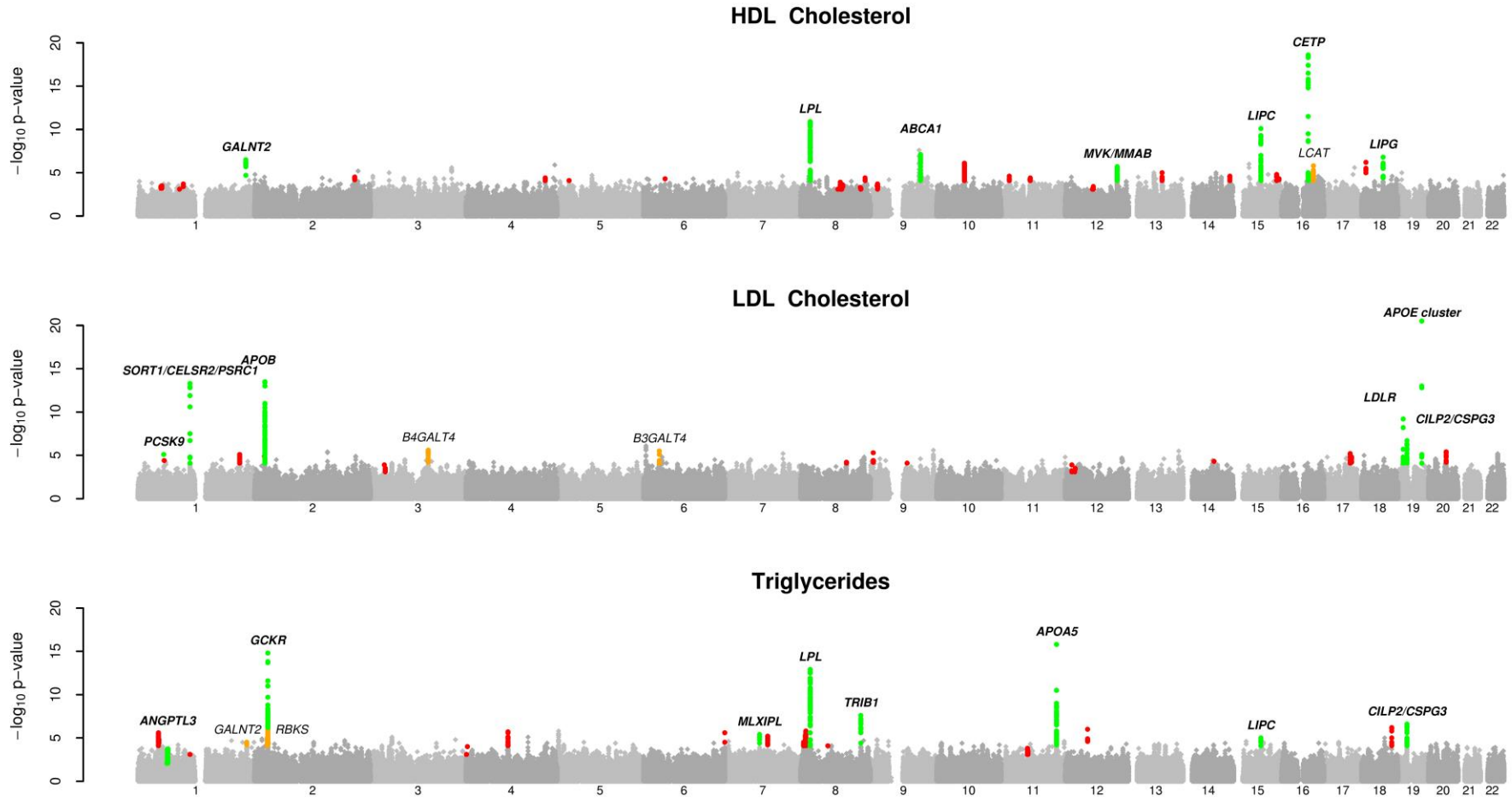


Zheng et al, Genetic Epidemiology, 2011

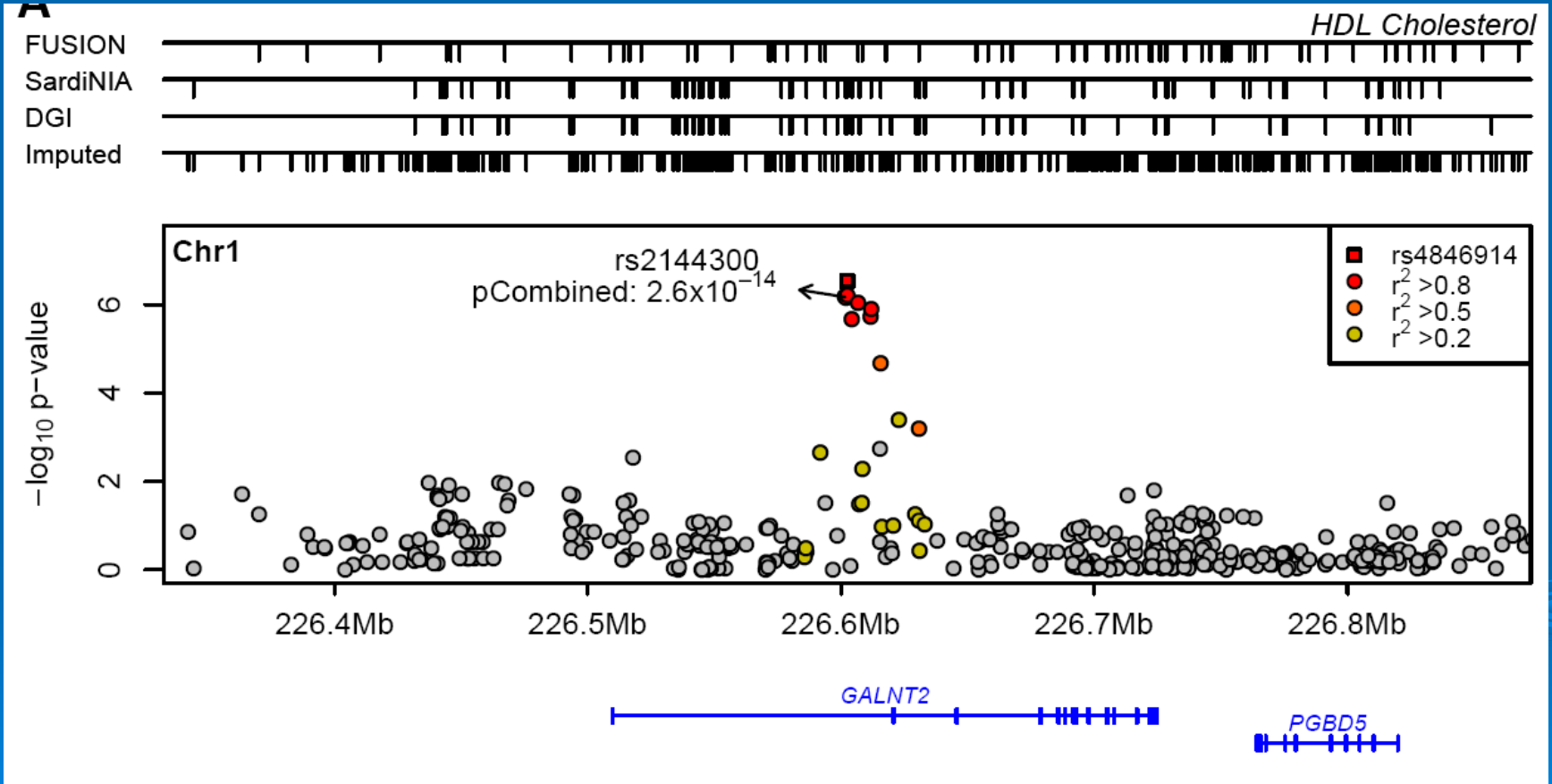
Combined Lipid Scans

- SardiNIA (Schlessinger, Uda, et al.)
  - ~4,300 individuals, cohort
- FUSION (Mohlke, Boehnke, Collins, et al.)
  - ~2,500 individuals
- DGI (Kathiresan, Altshuler, Orho-Mellander, et al.)
  - ~3,000 individuals
- Individually, 1-3 hits/scan, mostly known loci
- Analysis:
  - Impute genotypes so that all scans are analyzed at the same “SNPs”
  - Carry out meta-analysis of results across scans

Combined Lipid Scan Results

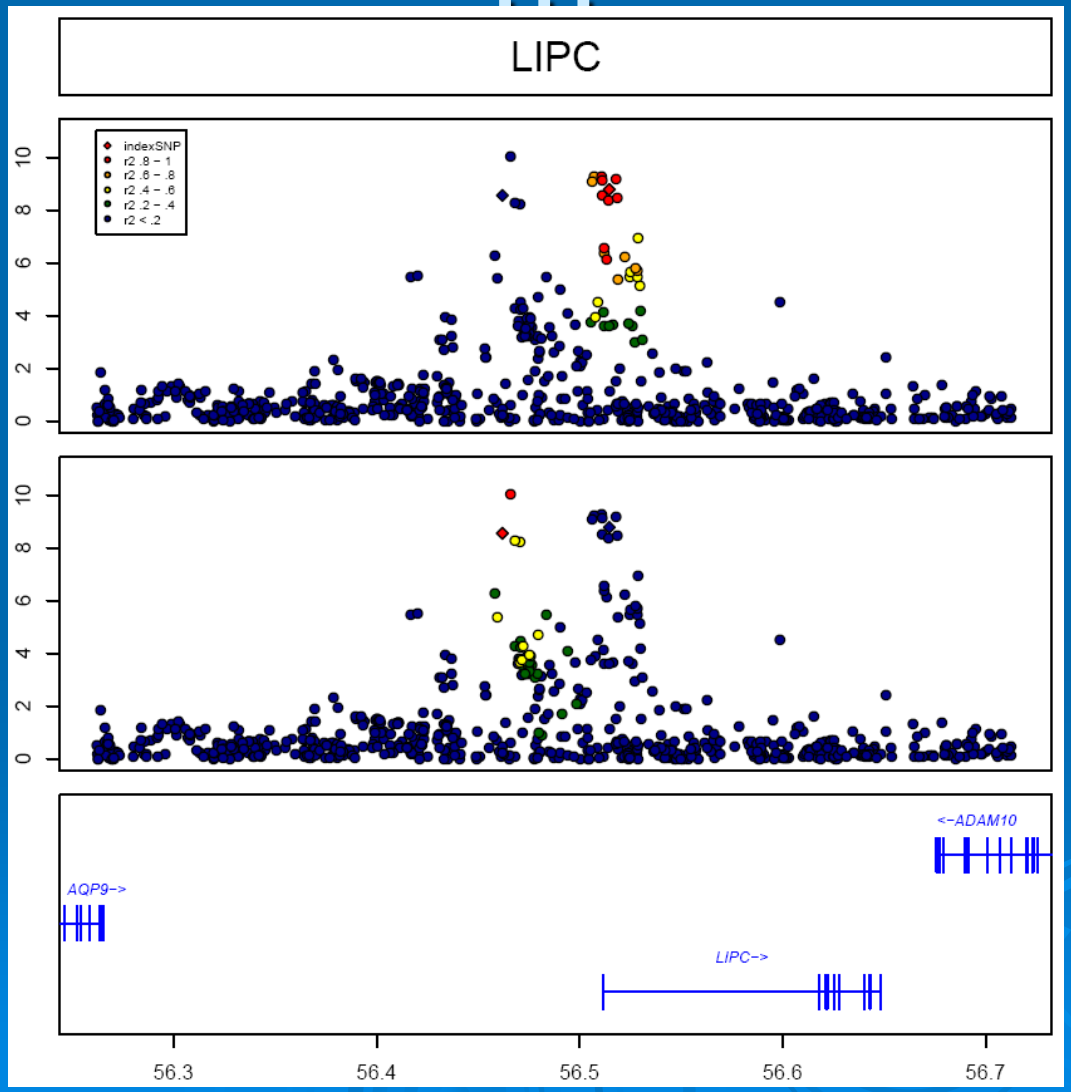


New HDL Locus



Willer et al, *Nat Genet*, 2008

New HDL Signal For An Old Locus

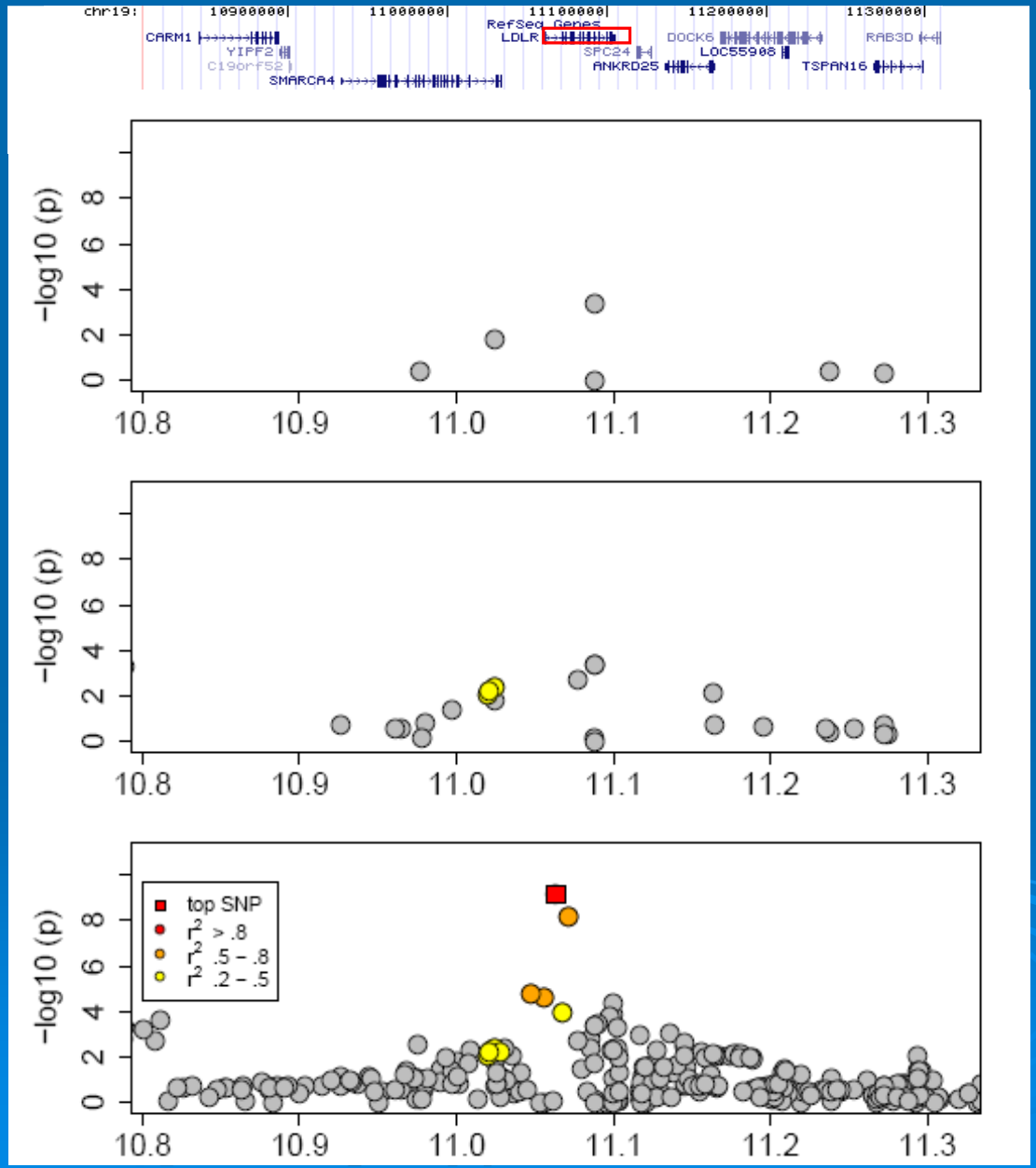


LDL-C association near LDLR

SNPs typed by all 3 groups (44,998)

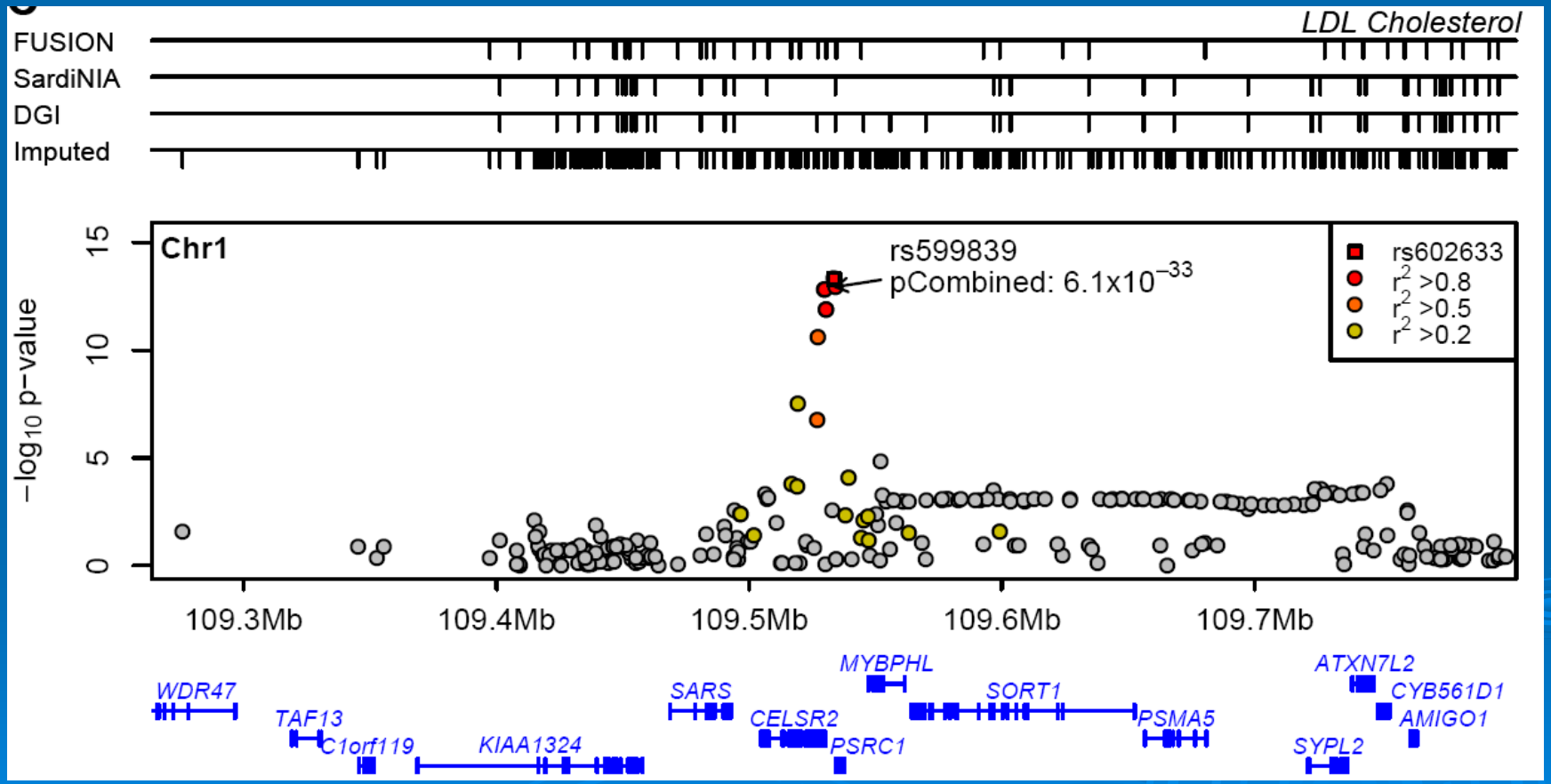
Affy panel SNPs (320,681)

Imputed SNPs (~ 2.25 million)



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

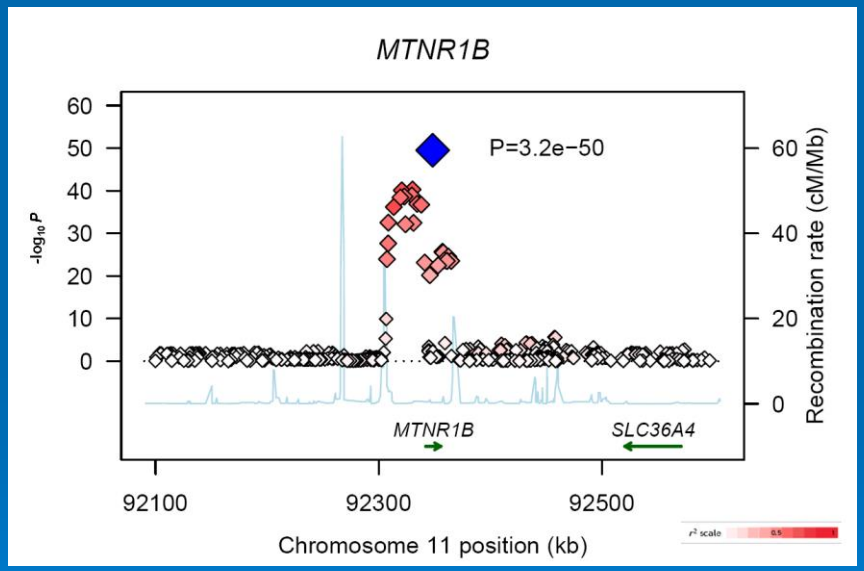
Gene	LDL-C p-value	Frequency CAD cases	Frequency CAD ctrls	CAD p-value	OR
APOE/C1/C4	3.0×10^{-43}	.209	.184	1.0×10^{-4}	1.17 (1.08-1.28)
APOE/C1/C4	1.2×10^{-9}	.339	.319	.0068	1.10 (1.02-1.18)
SORT1	6.1×10^{-33}	.808	.778	1.3×10^{-5}	1.20 (1.10-1.31)
LDLR	4.2×10^{-26}	.902	.890	6.7×10^{-4}	1.29 (1.10-1.52)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

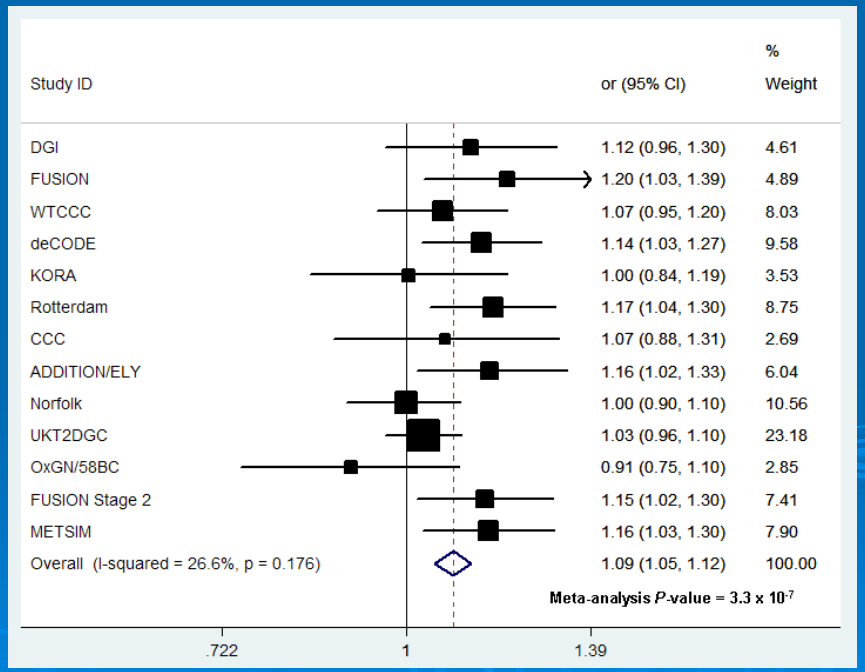
Data from WTCCC

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

Association with glucose,
36,000 non-diabetics

Association with diabetes,
18,000 cases vs. 64,000 controls



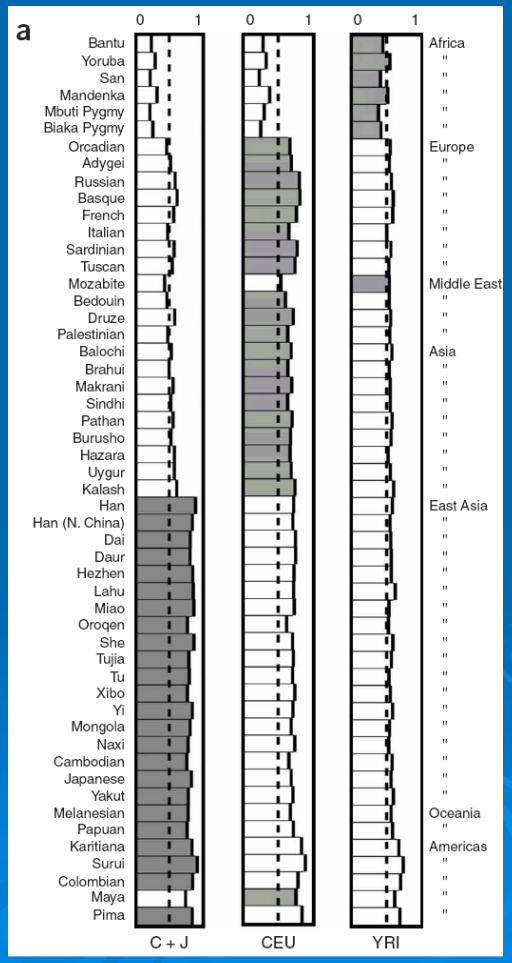


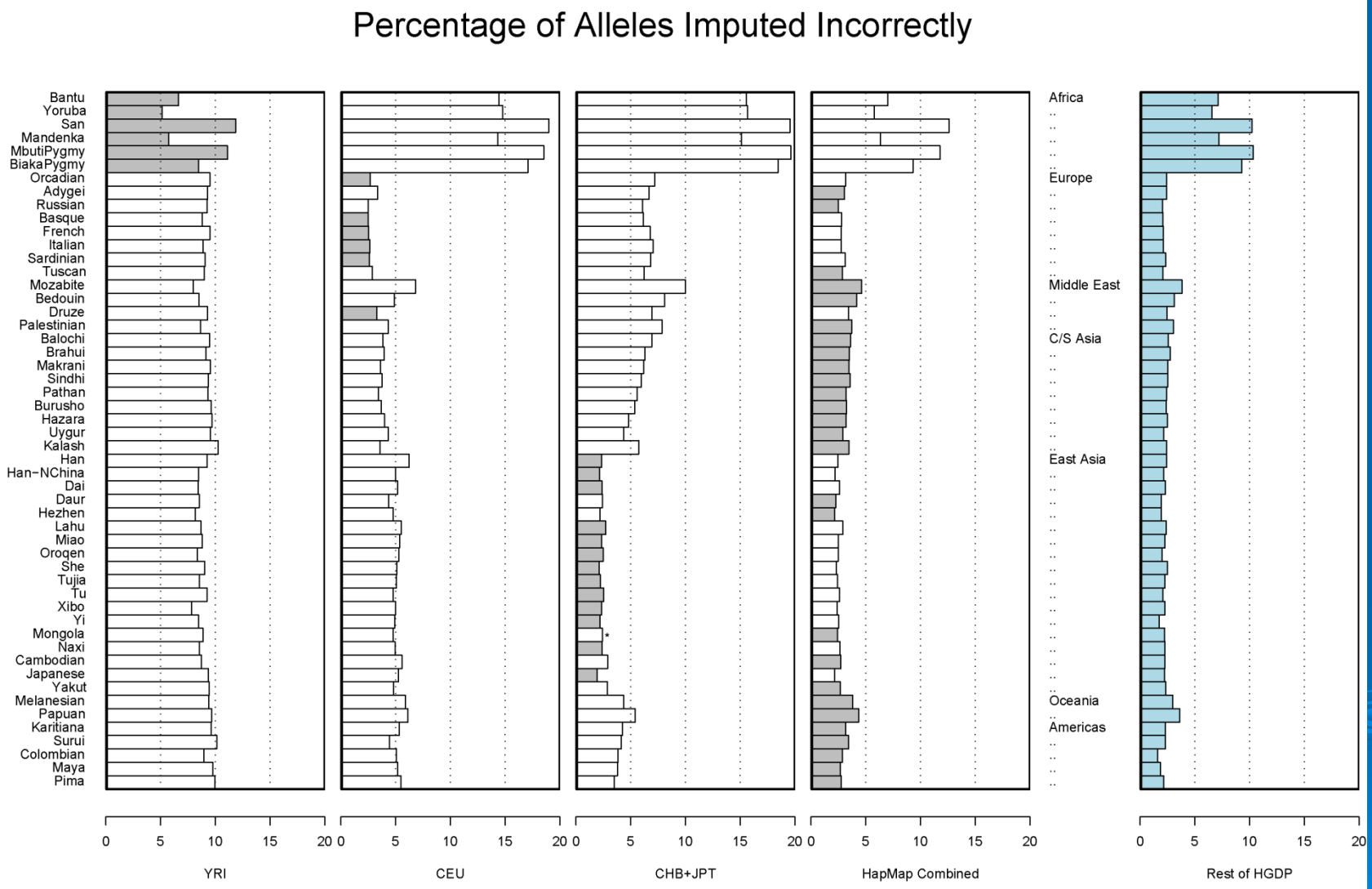
Prokopenko et al, *Nature Genetics*, in press

Does This 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





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

Imputation Improves with Reference Panel Size

		Accuracy By Minor Allele Frequency		
Panel	**# SNPs**	**MAF 1-3%**	**MAF 3-5%**	**MAF >5%**
Pilot (60 EUR)	15M	0.69	0.77	0.91
Interim Freeze (283 EUR)	25M	0.73	0.78	0.92
Phase I Freeze (563 EUR)	39M	0.83	0.85	0.94

- As more individuals are sequenced…
  - Reference panel becomes more complete
  - Imputation quality improves, particularly for rare SNPs

… But Becomes Computationally Challenging

Reference Panel	Samples	Markers	Time per Sample (in minutes)
HapMap 2 CEU	60	2.5 million	14
1000 Genomes Pilot CEU	60	7.3 million	41
1000 Genomes Interim EUR	283	11.6 million	1287
1000 Genomes Phase I EUR	381	18.7 million	3900

- Computational cost for original imputation methods scales …
  - Linearly with number of markers
  - Linearly with number of individuals being imputed
  - Quadratically with reference panel size

… Unless New Methods Used

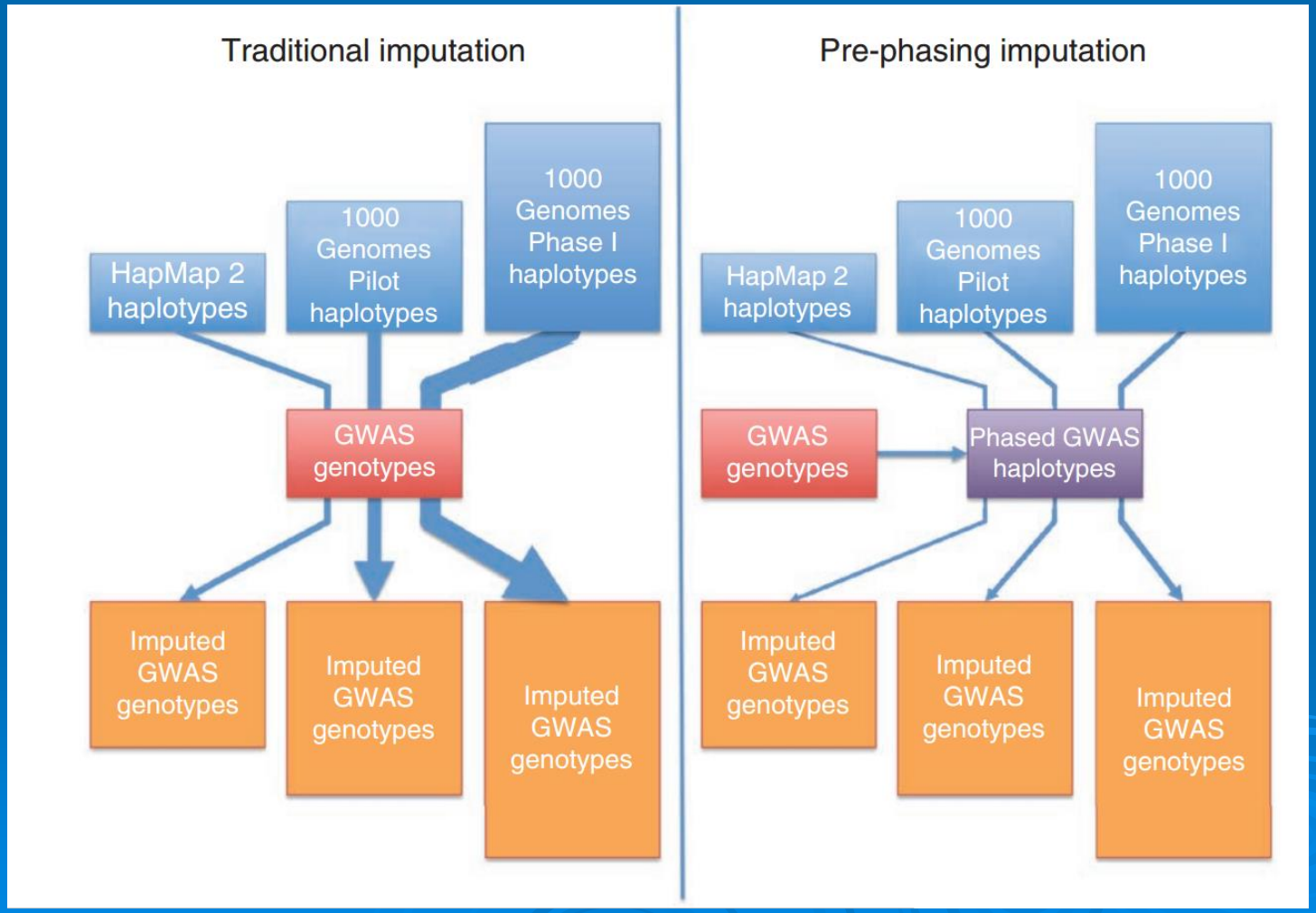
Bryan Howie

Christian Fuchsberger

Reference Panel	Samples	Markers	Time per Sample (in minutes, Standard method)	Time per Sample (in minutes, new method)
HapMap 2 CEU	60	2.5 million	14	1
1000 Genomes Pilot CEU	60	7.3 million	41	1
1000 Genomes Interim EUR	283	11.6 million	1287	6
1000 Genomes Phase I EUR	381	18.7 million	3900	12

- Improved methods scale linearly with reference panel size
  - This makes computational cost manageable

Speeding Up Imputation: Pre-Phasing



MaCH and Minimac Haplotyping and Imputation

- www.sph.umich.edu/csg/abecasis/Mach
- www.sph.umich.edu/csg/abecasis/Mach/tour
  - We will look at estimating and inferring haplotypes with Mach 1.0

- genome.sph.umich.edu/wiki/minimac
- genome.sph.umich.edu/wiki/minimac:_Tutorial
  - We will look at a simple analysis with minimac

Acknowledgements

- Sardinia Collaborators, led by:
  - David Schlessinger, Antonio Cao, Manuela Uda, Ed Lakatta, Paul Costa
  - Analysis by Serena Sanna, Paul Scheet, Weimin Chen
- FUSION Investigators, led by:
  - Mike Boehnke, Francis Collins, Karen Mohlke, Jaakko Tuomilehto, Richard Bergman
  - Analysis by Cristen Willer and Yun Li
- DGI Investigators:
  - Sekar Kathiresan, Marju Orho-Melander, David Altshuler and colleagues
- MaCH/minimac Development
  - Yun Li, Paul Scheet, Jun Ding, Christian Fuchsberger

References

- References:
  - Chen and Abecasis, *AJHG*, 2007
  - Scuteri et al, *PLoS Genetics,* 2007
  - Willer et al, *Nature Genetics*, 2008
  - Li et al, *Ann Rev of Genomics and Hum Genet*, 2009
  - Li et al, *Genetic Epidemiology*, 2010
  - Howie et al, *Nature Genetics*, 2012

goncalo@umic.edu