“In Silico” Genotyping for Genome Wide Association Scans

Turning a Flood of Data into a Deluge

Gonçalo Abecasis

University of Michigan

Lots of Genotypes Are Good… How About Even More Genotypes?

- If millions of genotypes are good, wouldn’t billions be better?
- Spend more dollars, euros, pounds, and …
  - Examine more individuals …
  - Examine more SNPs …
- Inexpensive “in silico” genotyping strategies
- Estimate genotypes for individuals related to those in GWAS sample
  - Intuition for how *in silico* genotyping works
- Estimate additional genotypes for individuals in the GWAS sample
  - Facilitate comparisons across studies
  - Improve coverage of the genome

In Silico Genotyping For Family Samples

- Family members share large segments of chromosomes

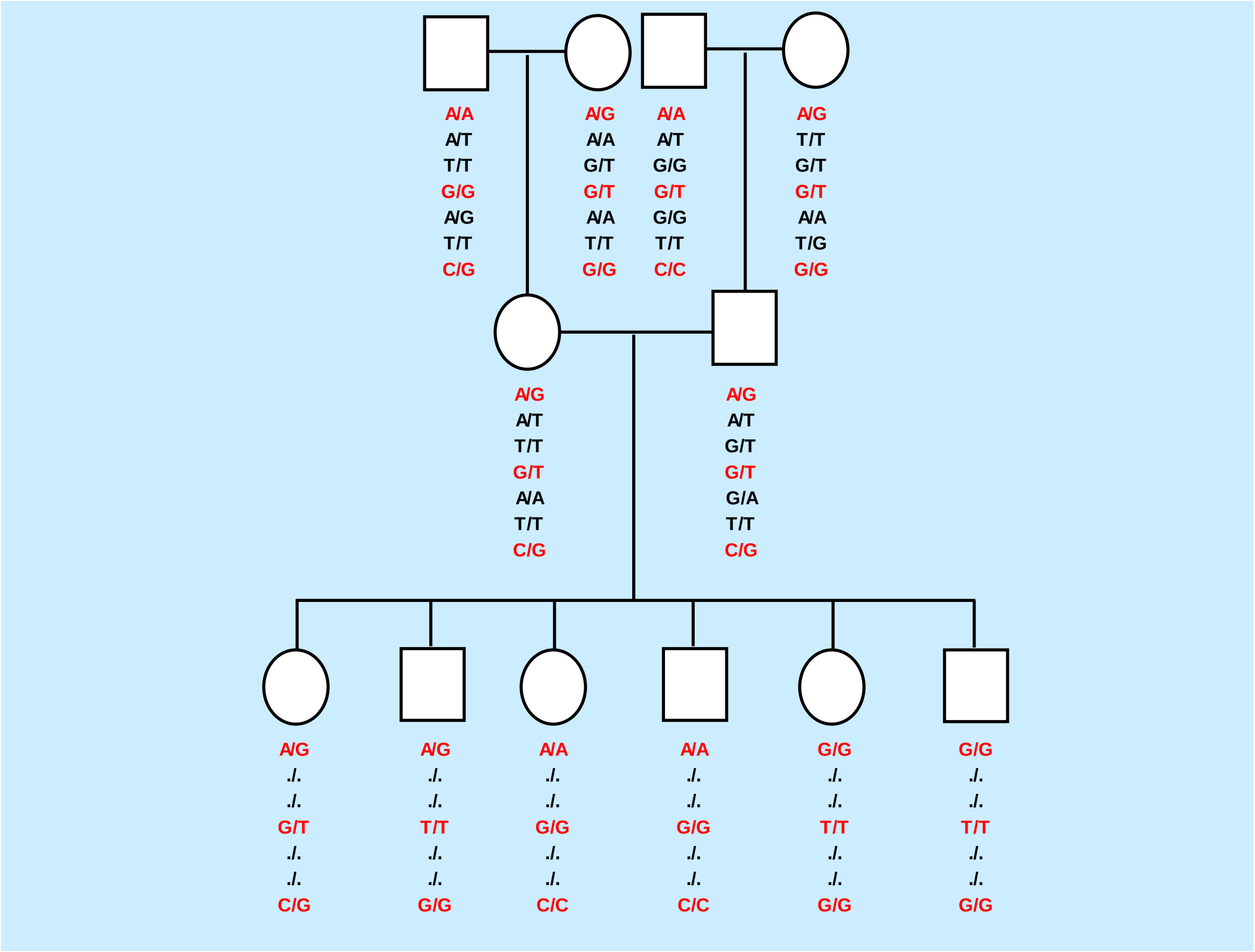
- If we genotype many related individuals, we will effectively be genotyping a few chromosomes many times

- An alternative is to:
  - Genotype a few markers on all samples
  - Identify shared chromosomal segments that segregate in family
  - Use a high-density panel to genotype a few samples per family
  - Estimate missing genotypes in samples without high density data

  - The first two steps are optional, but very helpful

Burdick et al, *Nat Genet,* 2006

Genotype Inference
Part 1 – Observed Genotype Data



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

A/G
A/A
G/T
G/T
A/A
T/T
G/G

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

A/G
T/T
G/T
G/T
A/A
T/G
G/G

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

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

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

A/G
./.
./.
T/T
./.
./.
G/G

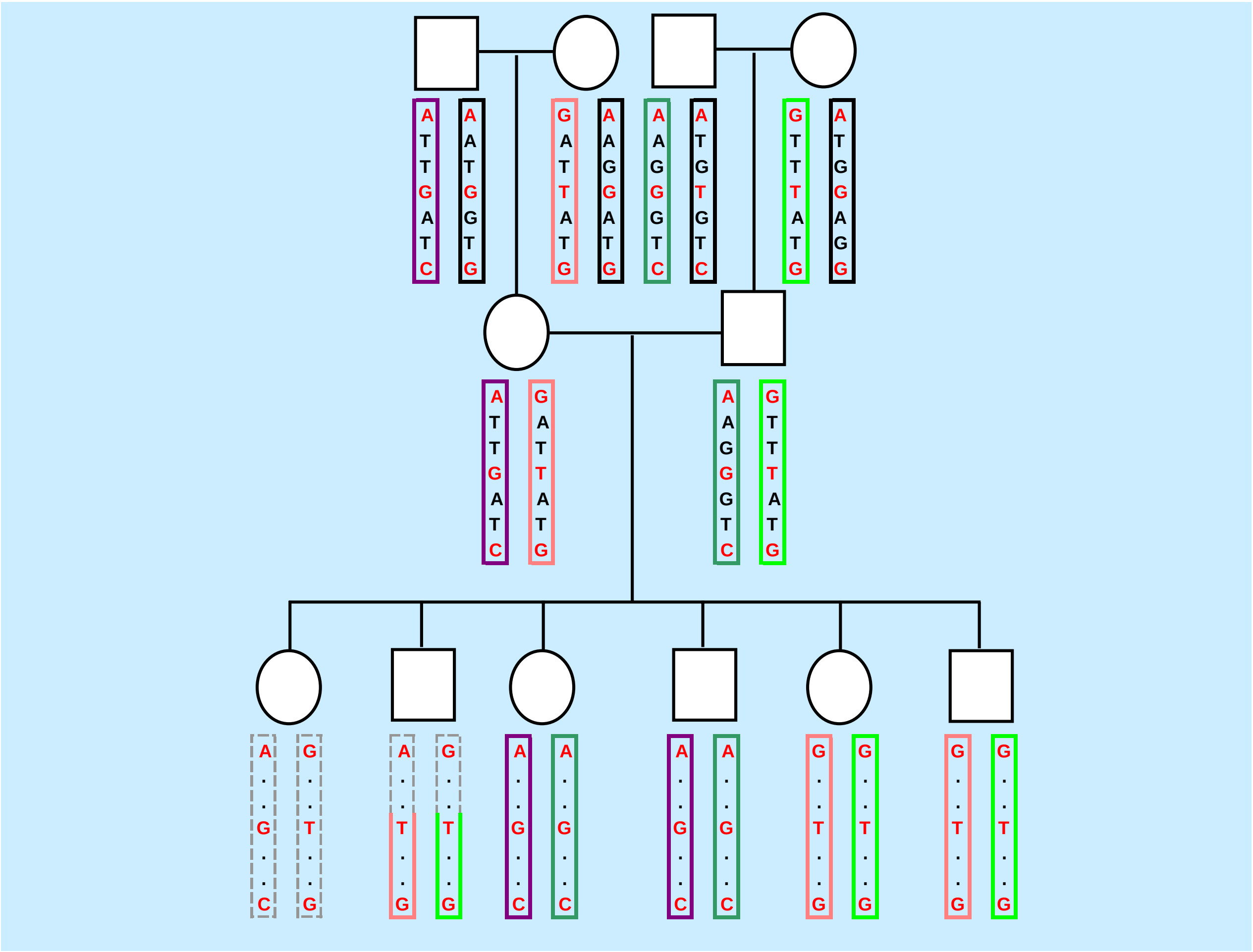
A/A
./.
./.
G/G
./.
./.
C/C

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

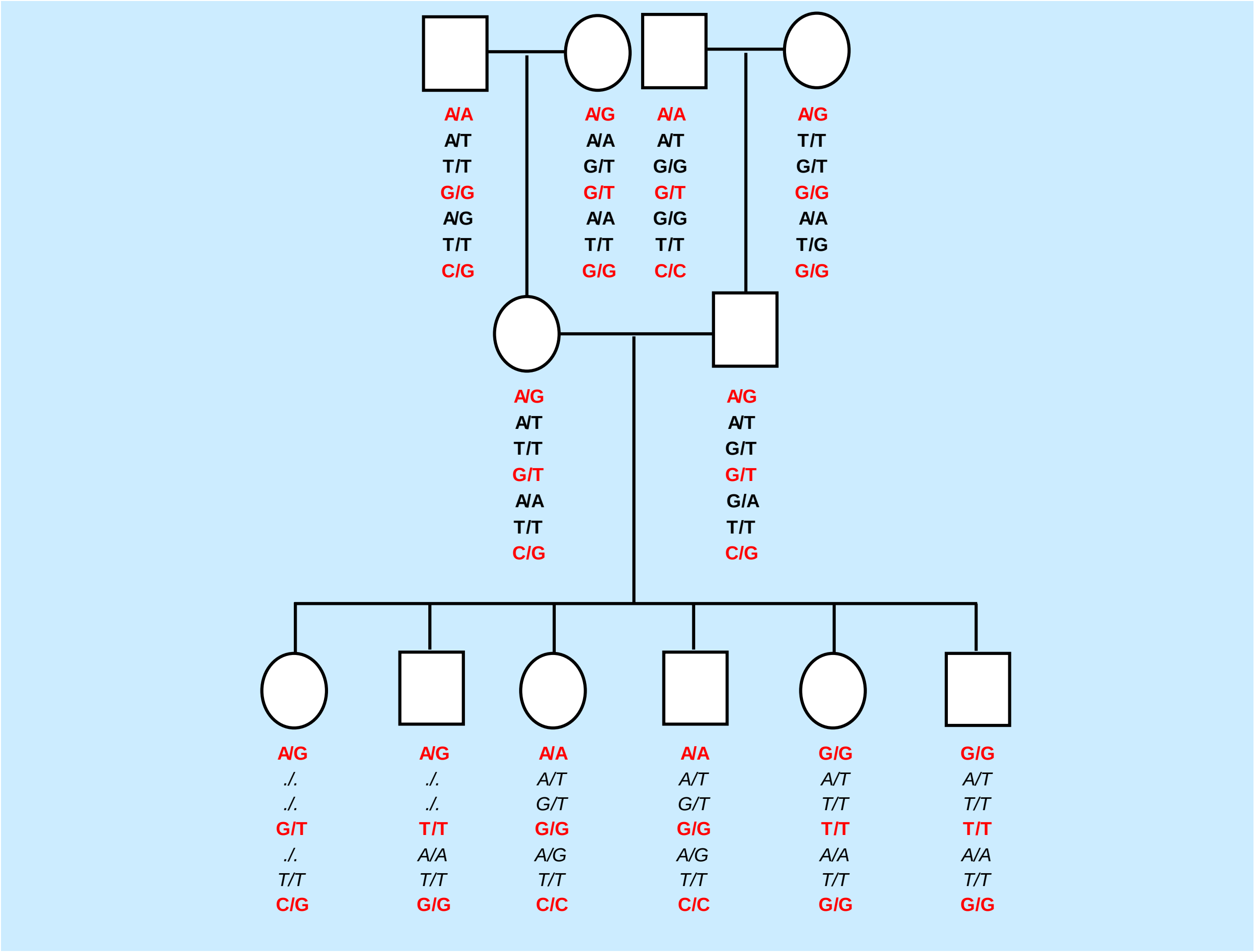
G/G
./.
./.
T/T
./.
./.
G/G

G/G
./.
./.
T/T
./.
./.
G/G

Genotype Inference
Part 2 – Inferring Allele Sharing



Genotype Inference
Part 3 – Imputing Missing Genotypes



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

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

A/G	A/G	A/A	A/A	G/G	G/G
./.	*./.*	*A/T*	*A/T*	*A/T*	*A/T*
./.	*./.*	*G/T*	*G/T*	*T/T*	*T/T*
G/T	T/T	G/G	G/G	T/T	T/T
./.	*A/A*	*A/G*	*A/G*	*A/A*	*A/A*
T/T	*T/T*	*T/T*	*T/T*	*T/T*	*T/T*
C/G	G/G	C/C	C/C	G/G	G/G

Formal 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

Model With Inferred Genotypes

- Replace genotype score g with its expected value:

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

- Where

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

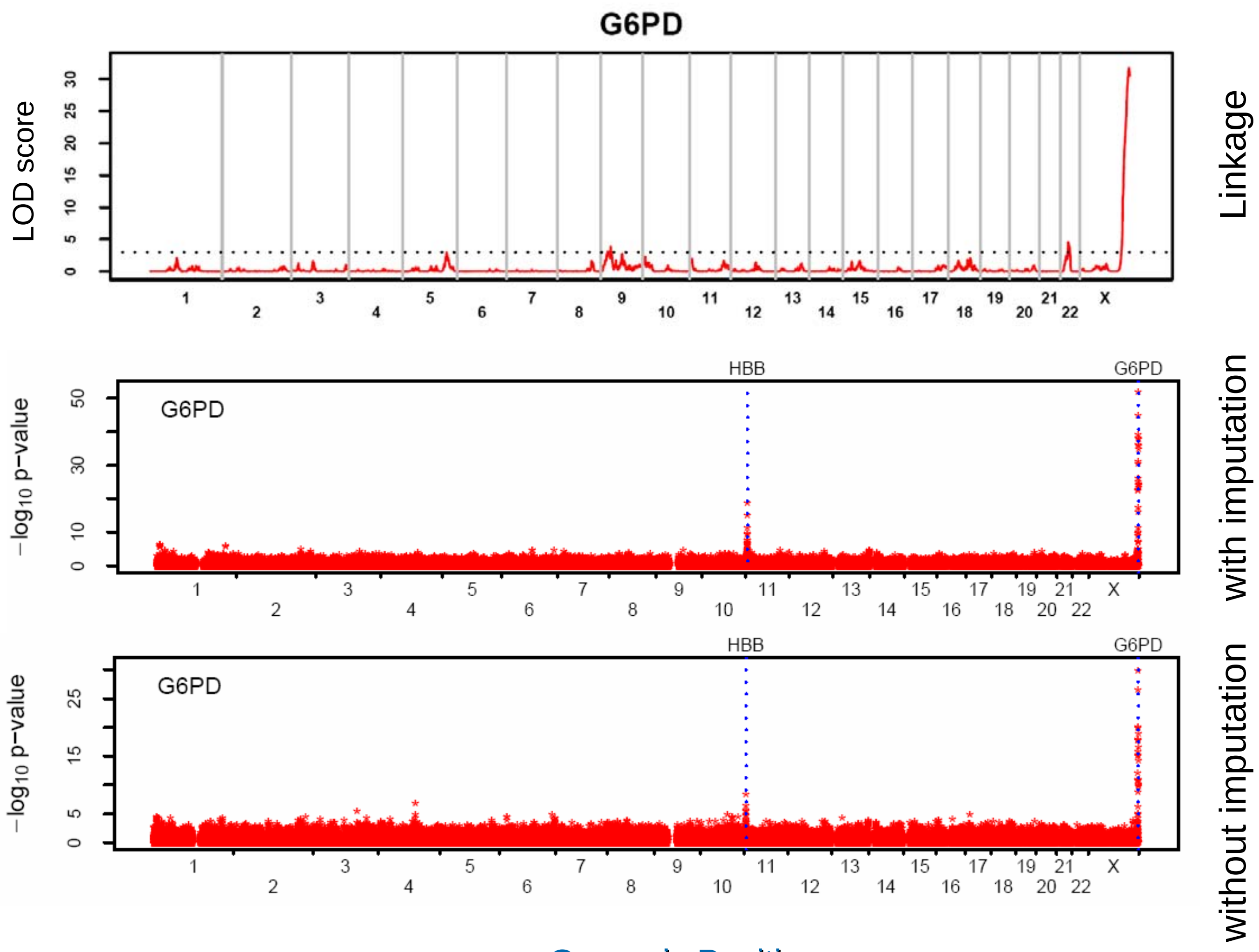
- Association test implemented as score test or as likelihood ratio test
  - Variance component framework to allow for relatedness

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

Quantitative Trait GWAS in Sardinia

- 6,148 Sardinians from 4 towns in Ogliastra
  - Many close relationships among sampled individuals
- 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



Genomic Position

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
- Nevertheless, it may still be possible to identify stretches of shared chromosome …
- … and by comparing shared stretches between densely genotyped individuals and those with sparser data

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

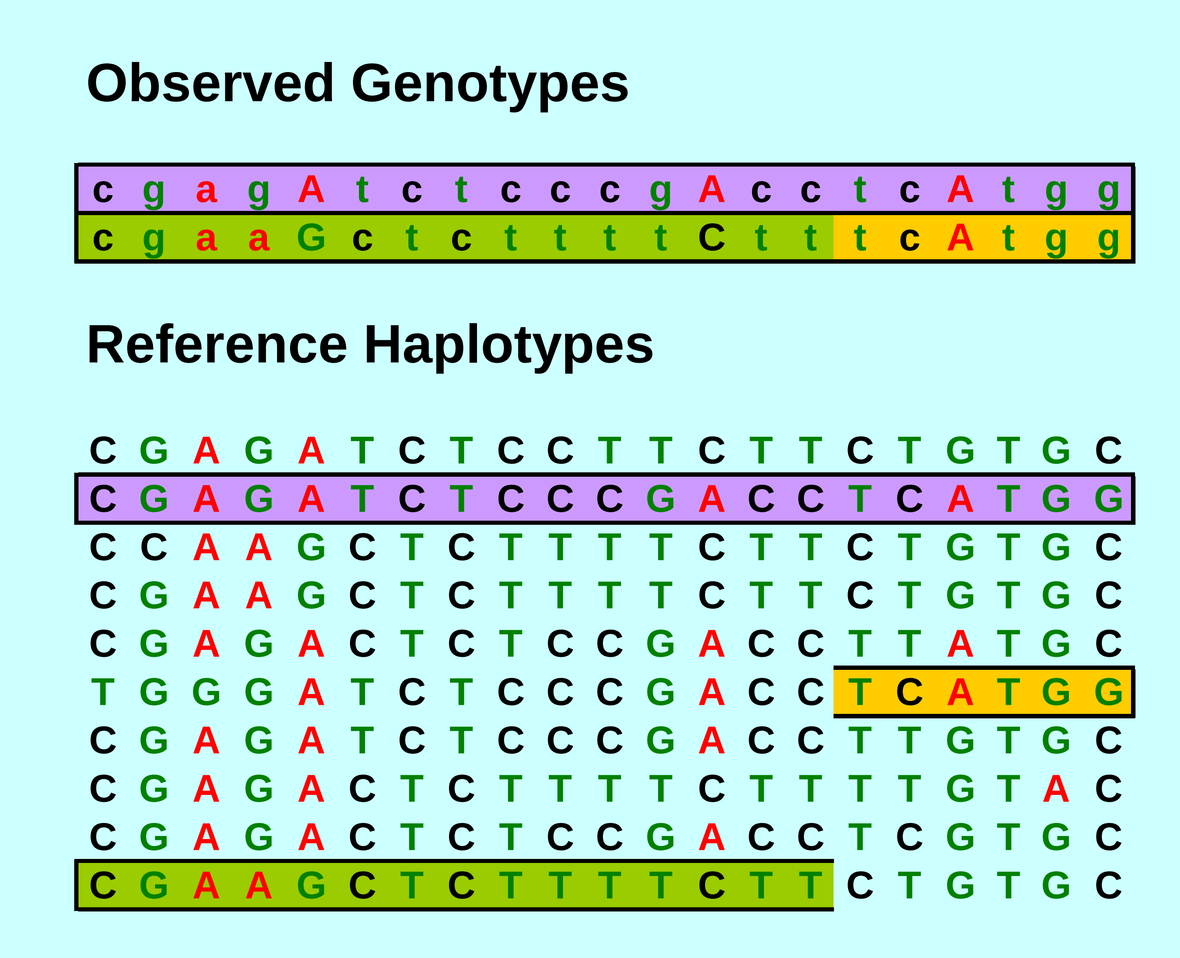
Observed Genotypes

```
. . . . 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

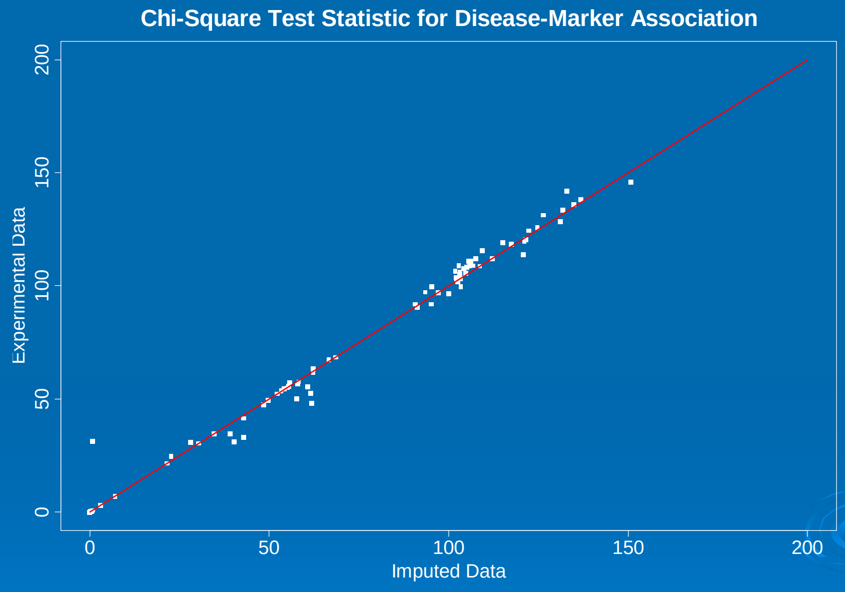
Does This Actually Work? Preliminary Results

- 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



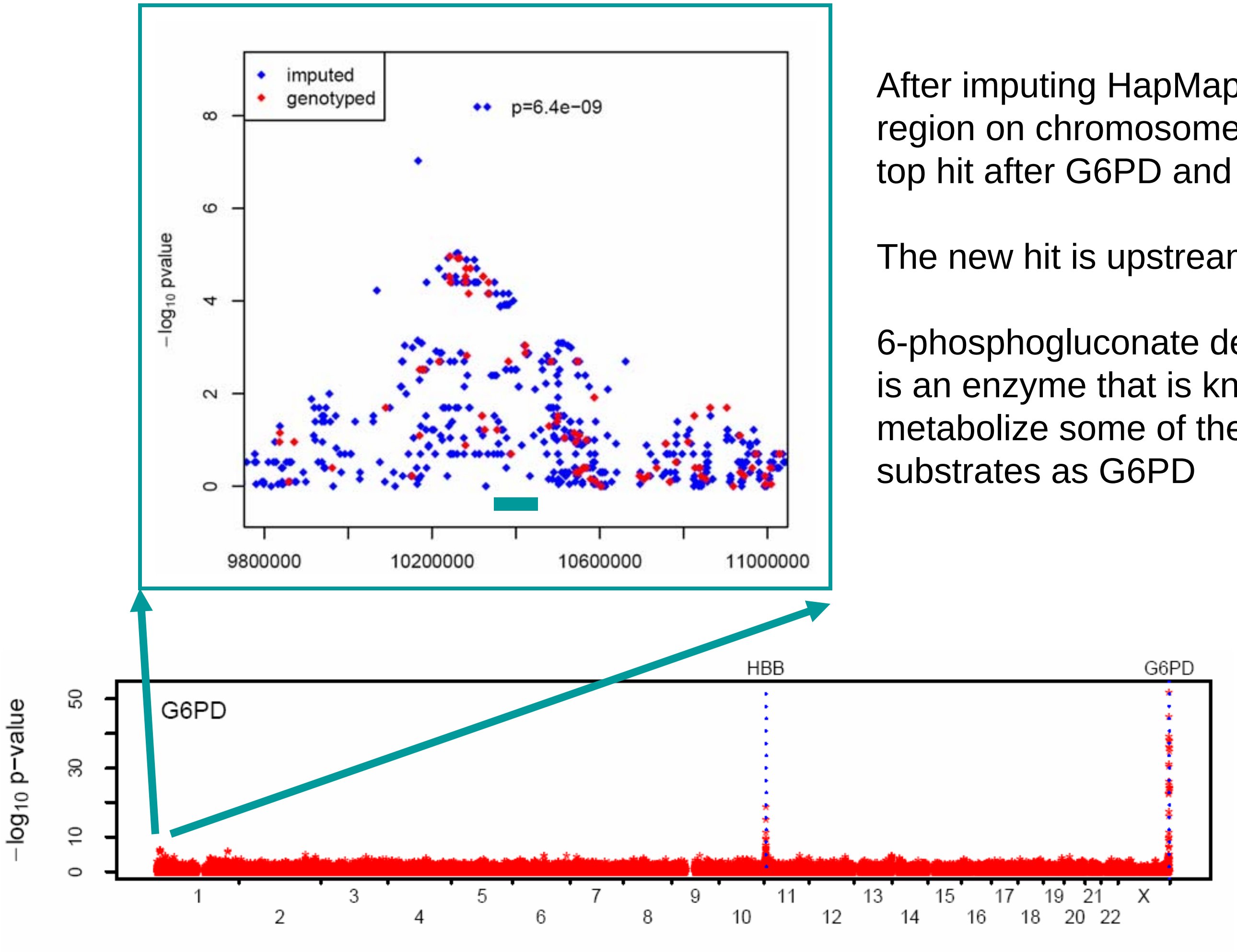
Does This Really Work?

- 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

Back to Sardinia G6PD Activity Example …



Genomic Position

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

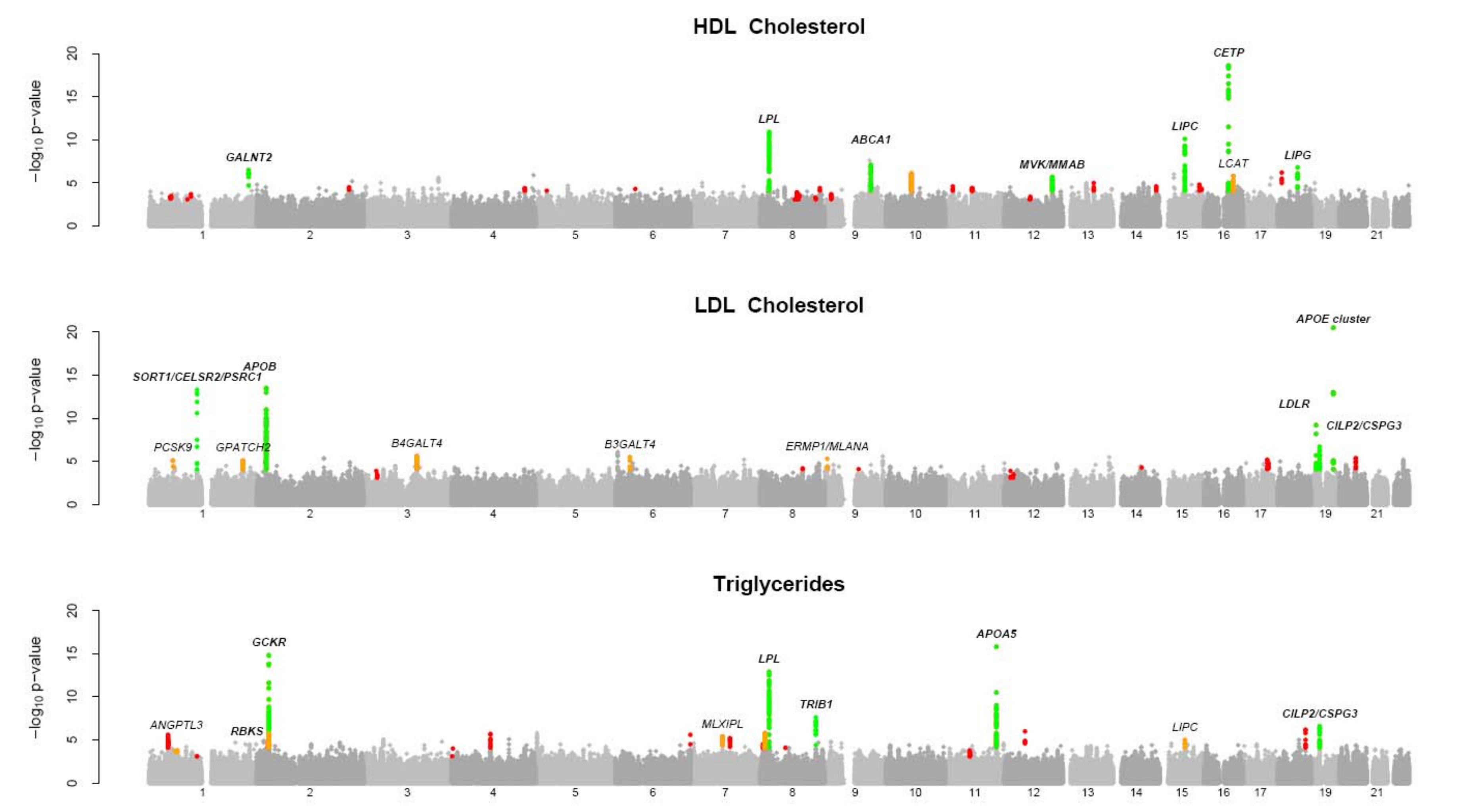
The new hit is upstream of 6PGD

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

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

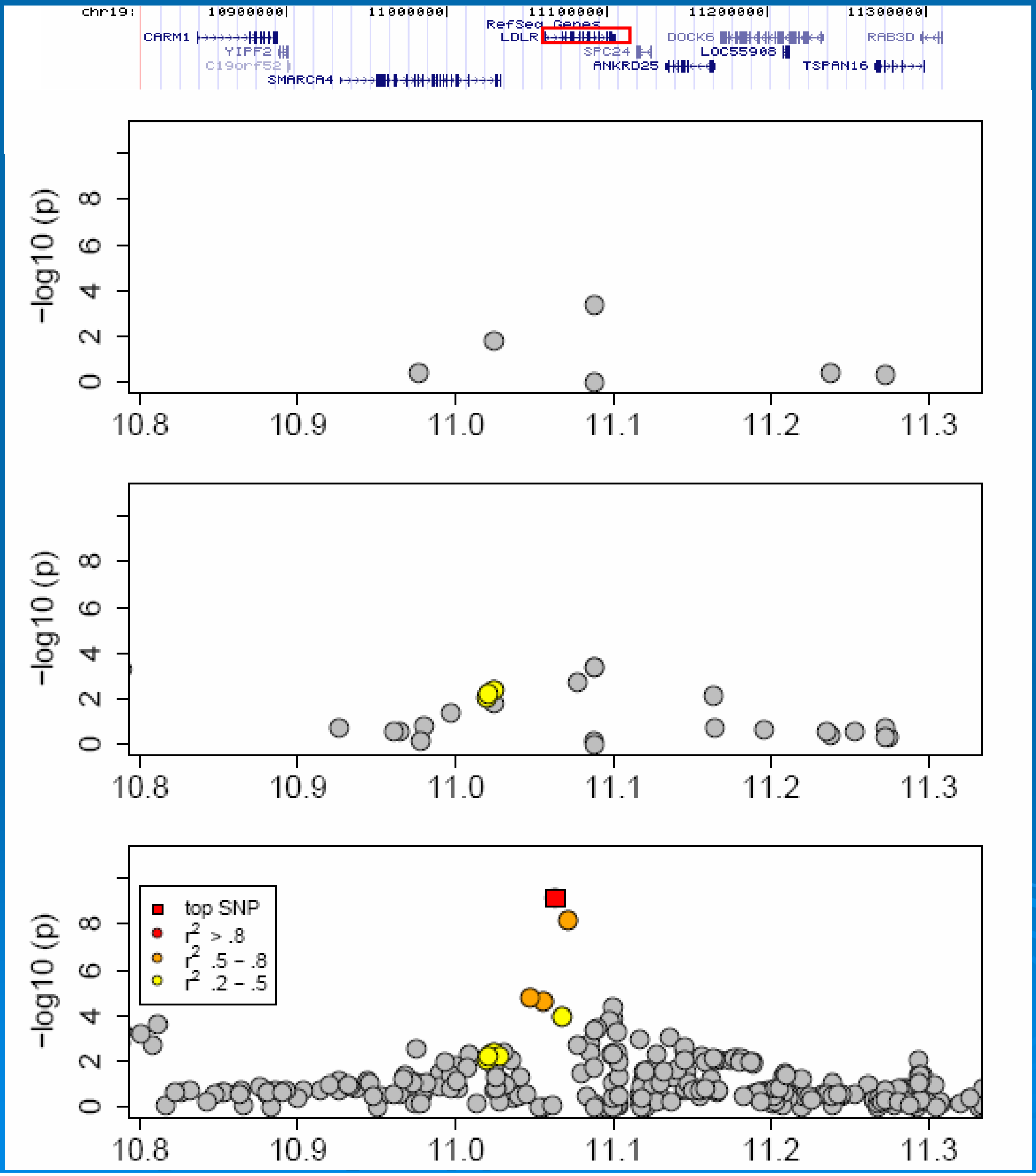


LDL-C association near LDLR

SNPs typed by all 3 groups (44,998)

Affy panel SNPs (320,681)

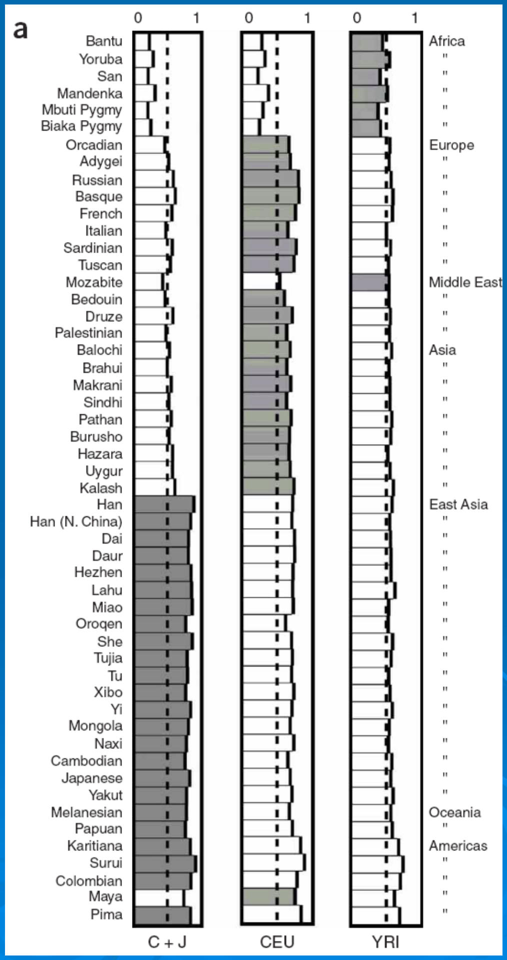
Imputed SNPs (~ 2.25 million)



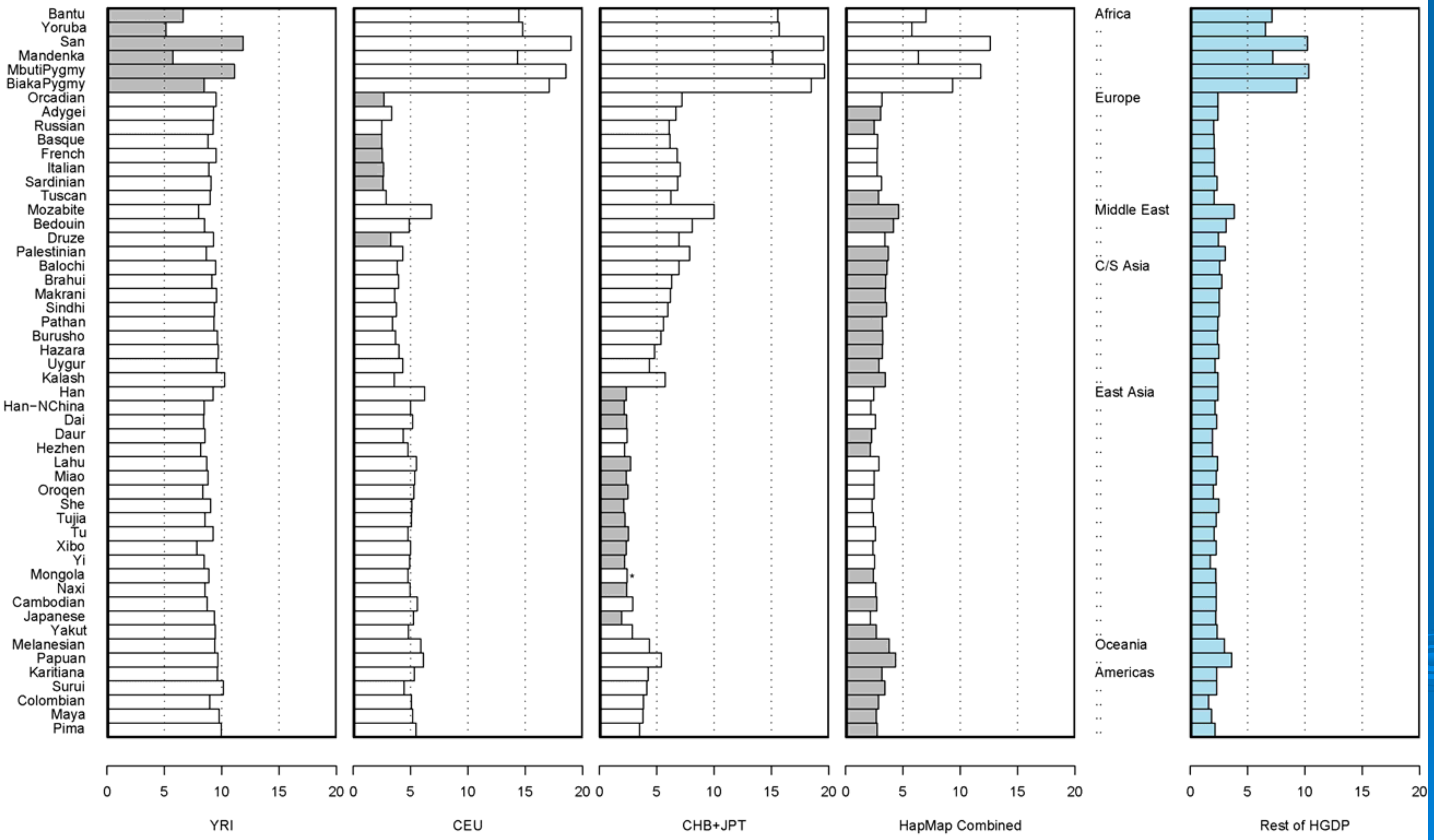
Does This Work Across Populations?

- 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



Percentage of Alleles Imputed Incorrectly



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

Comparison With Impute

- We compared our results with IMPUTE across all the HGDP populations
- We found that:
- Genotypes imputed by MACH were more concordant with original genotypes in 29/52 populations
- Genotypes imputed by IMPUTE were more concordant with original genotypes in 7/52 populations
- Overall, the two methods are more concordant with each other than with the real data

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, David Altshuler and colleagues
- MaCH Development
  - Yun Li, Paul Scheet, Jun Ding

goncalo@umich.edu