Combining Data from Different Genotyping Platforms

Gonçalo Abecasis
Center for Statistical Genetics
University of Michigan

The Challenge

- Detecting small effects requires very large sample sizes
- Combined analysis of data from different studies is one way to increase sample size …
- … but these studies may rely on different platforms that have little direct overlap
  - For example, Illumina 317K chip and the Affymetrix 500K chip have only ~51,000 SNPs in common

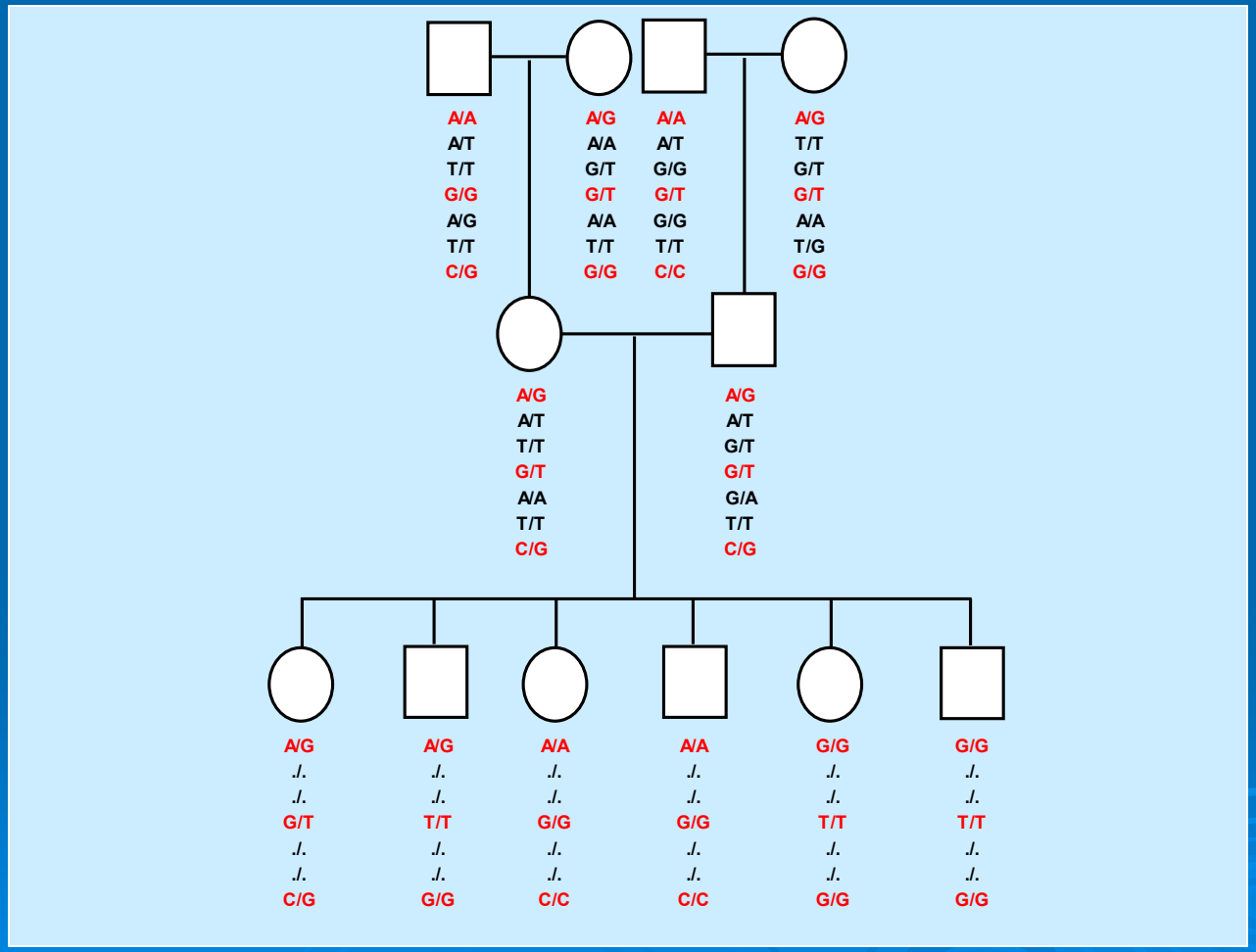
My Talk Today

- *In silico* genotyping
  - Inferring unobserved genotypes
- Estimate genotypes for relatives of individuals in genome-wide association scan
  - Intuition for how in silico genotyping works
- Estimate genotypes for untyped markers, by combining study sample with Hapmap
  - Facilitate comparisons across studies
- Evaluating quality of the inferred genotypes

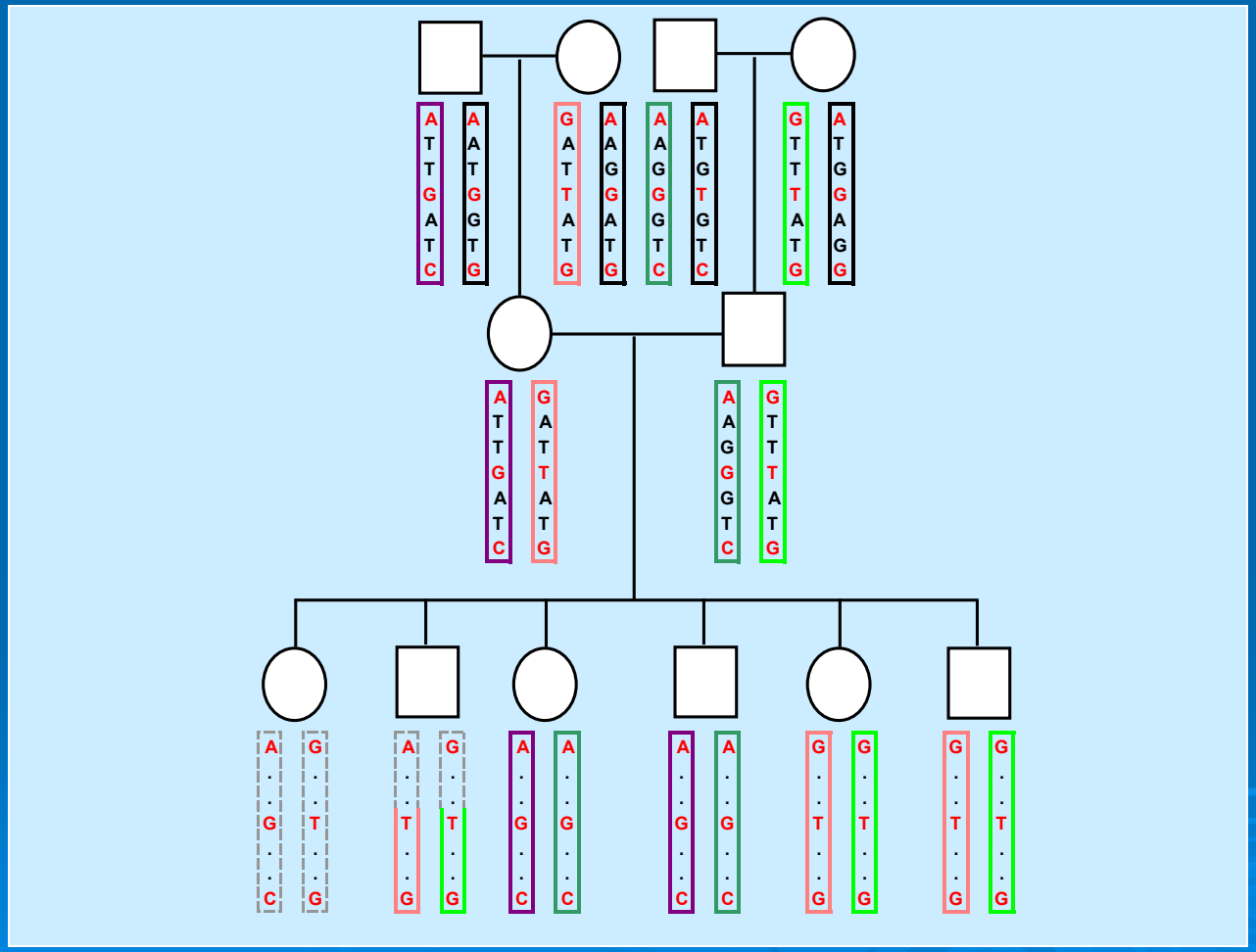
In Silico Genotyping For Family Samples

- 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
  - use high-density panel to genotype a few individuals
  - infer shared segments and then estimate the missing genotypes
  - if relatives have no genotype data, we can still estimate a probability for each of their genotypes

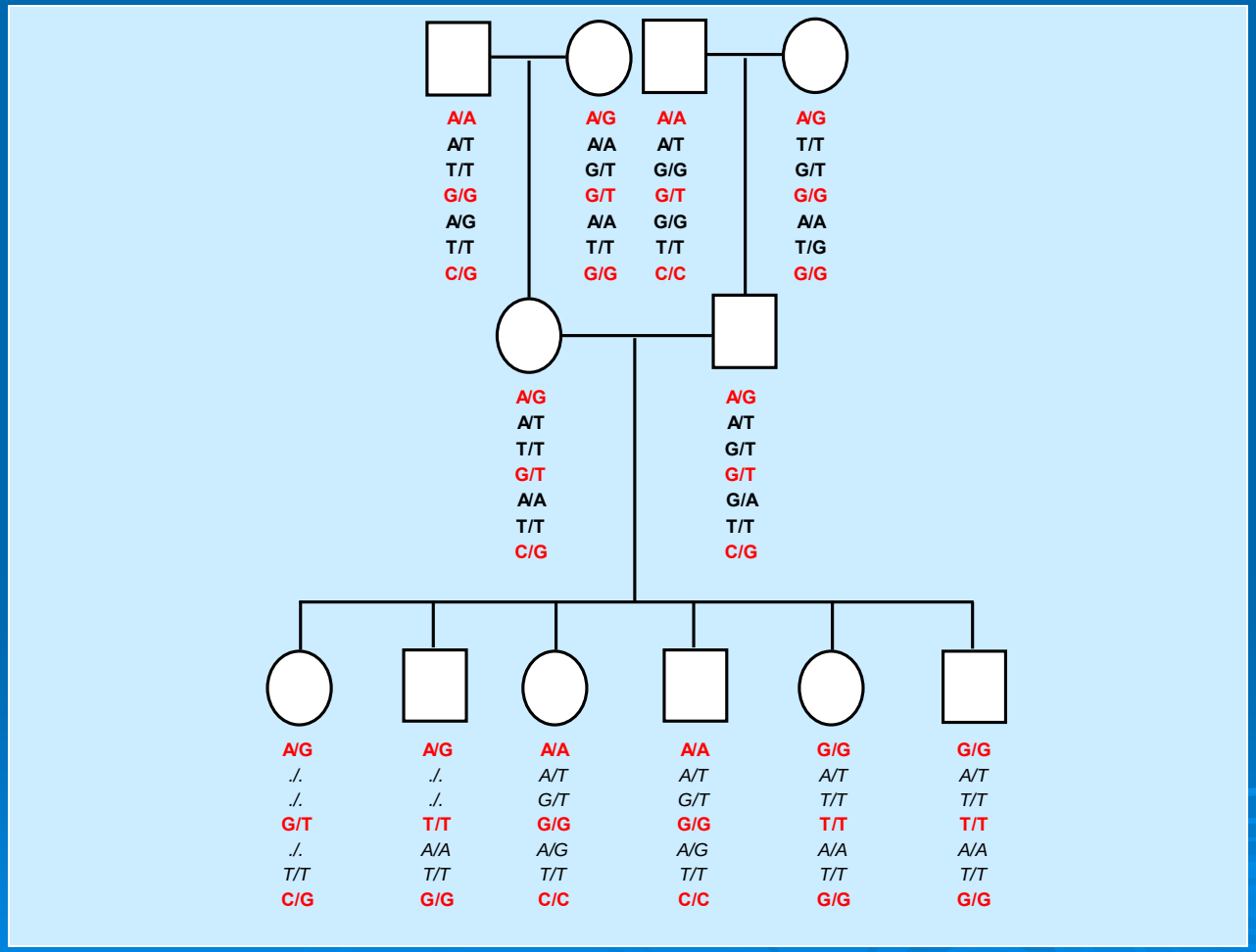
Genotype Inference
Part 1 – Observed Genotype Data



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

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

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

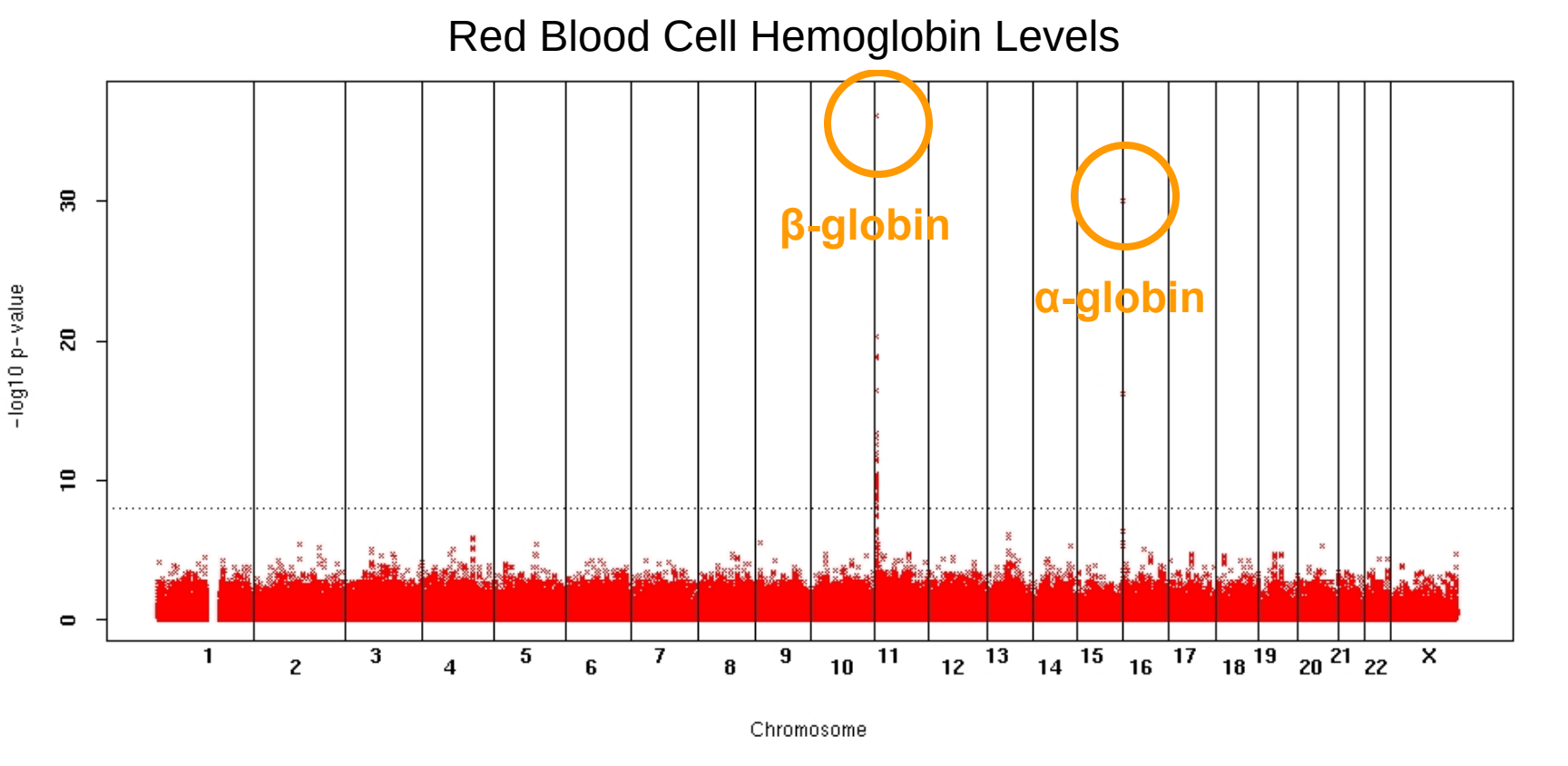
- Association test can then be implemented as a score test or as a likelihood ratio test

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

Sardinia

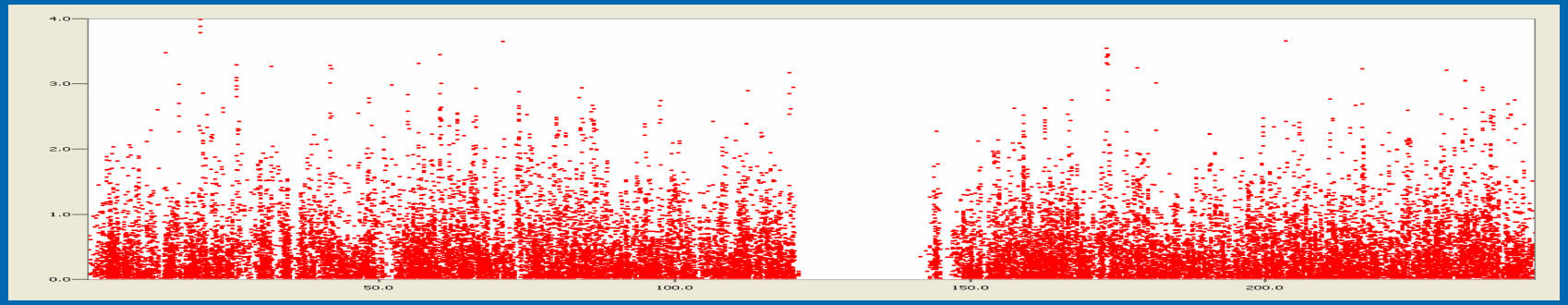
- 6,148 Sardinians from 4 towns in Ogliastra
- Measured 98 aging related quantitative traits
- Genotyping:
  - Affymetrix 10K chip in 4,500 individuals (done)
  - Affymetrix 500K chip in 1,500 individuals (ongoing)
- Large pedigrees, computationally challenging
  - Preliminary results

Preliminary Results from Sardinia

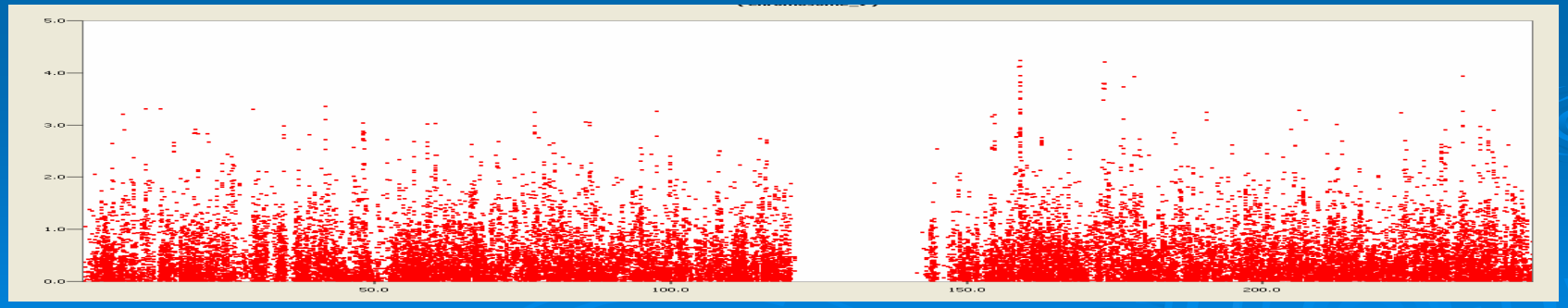


Preliminary Results from Sardinia QT interval, Chromosome 1

Before imputation



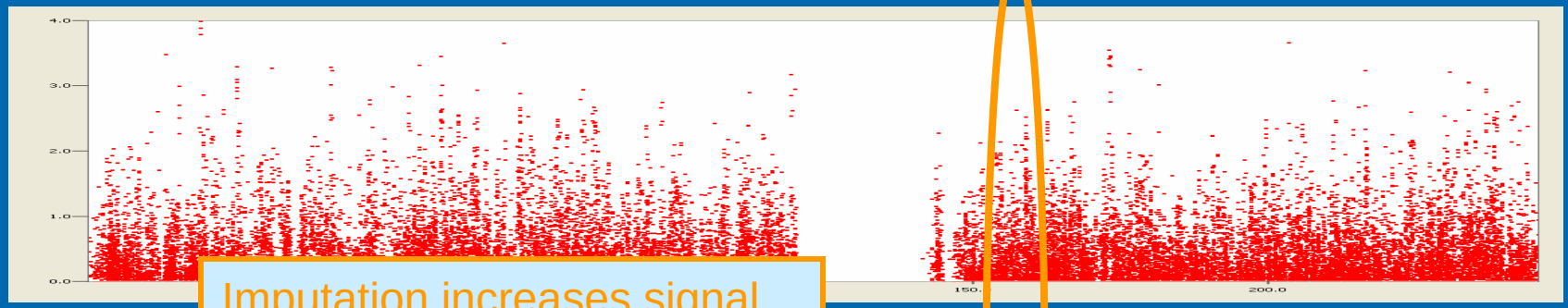
After imputation



Position (in Mb) Along Chromosome 1

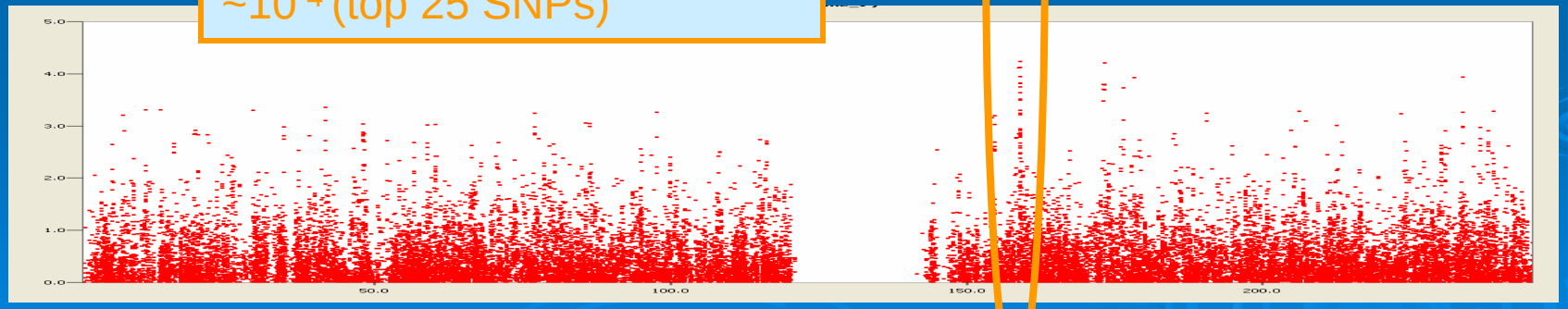
Preliminary Results from Sardinia QT interval, Chromosome 1

Before imputation



Imputation increases signal at NOS1AP increases from ~0.005 (top 3000 SNPs) to ~10^{-4} (top 25 SNPs)

After imp



Position (in Mb) Along Chromosome 1

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$_{Ha}$ plotypes

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$_{Ha}$ plotypes

```
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 (2002) model

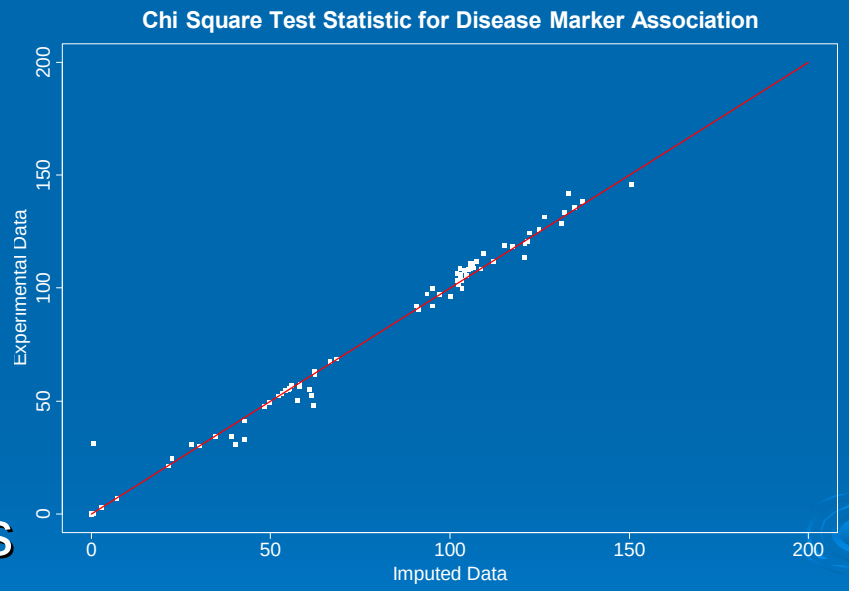
Output of Imputation Runs…

g	A	c	c	t	c	A	t	g	g	Iteration 1
t	C	t	t	t	c	A	t	g	g	
g	A	c	c	t	c	A	t	g	g	Iteration 2
t	C	c	c	t	c	A	t	g	c	
g	A	c	c	t	c	A	t	g	g	Iteration 3
t	C	c	c	t	c	A	t	g	c	
g	A	c	c	t	c	A	t	g	g	Iteration 4
t	C	c	c	t	c	A	t	g	c	
g	A	c	c	t	c	A	t	g	g	"Best Call"
t	C	c	c	t	c	A	t	g	c	
1	1	3/4	3/4	1	1	1	1	1	3/4	Quality Score
g	A	c	c	t	c	A	t	g	g	Reference Allele
1	1	7/4	7/4	2	2	2	2	2	5/4	Dosage

Assessing the Approach: AMD Case Control Study

- Used 11 tag SNPs to predict 84 SNPs in CFH
- Predicted genotypes differ from original ~1.8% of the time
  - ~2.5% for PHASE
  - ~3.2% for fastPHASE
- Calculation took ~3 minutes
  - ~21min for fastPHASE
  - ~1 day for PHASE

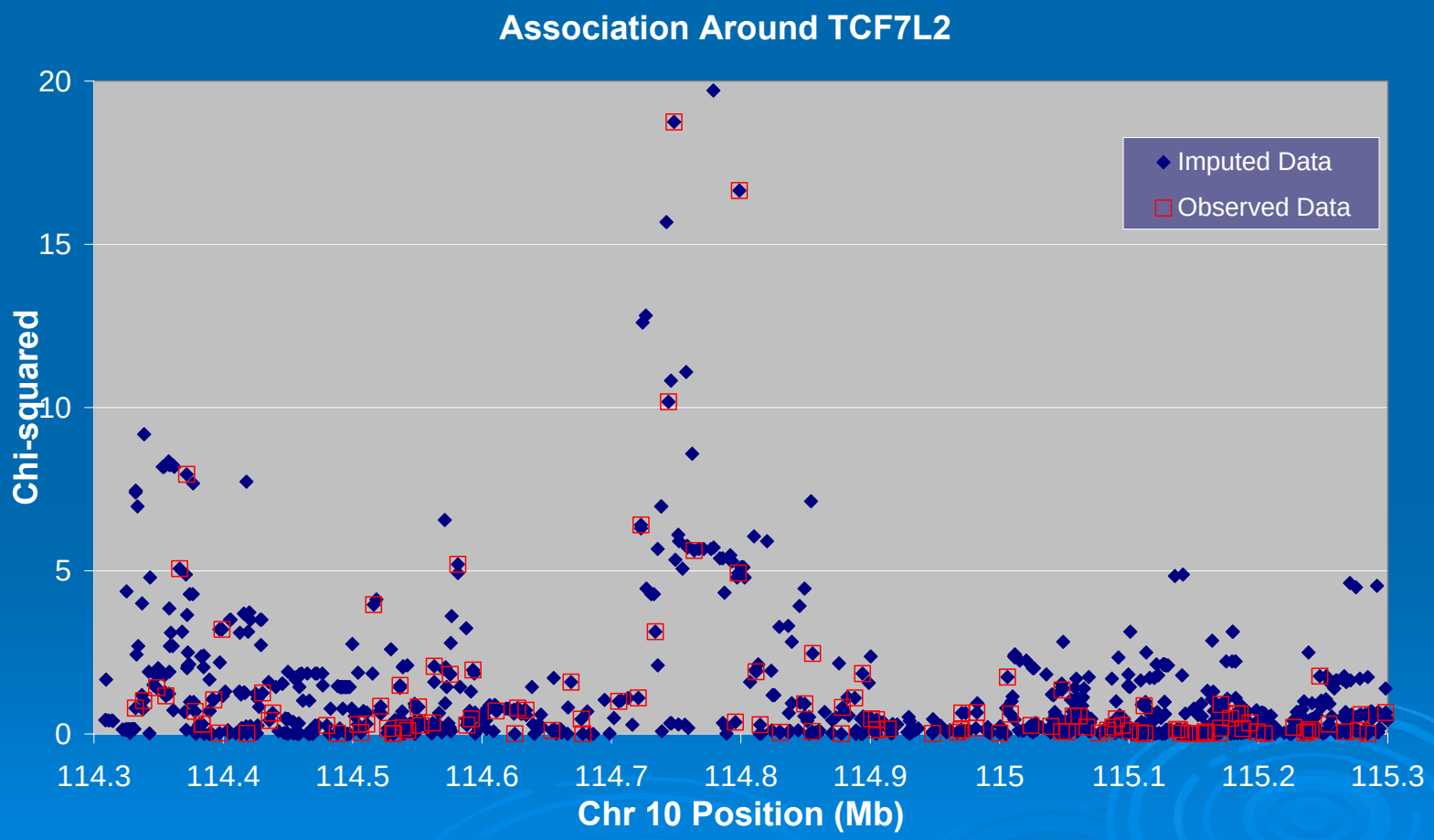
Comparison of Test Statistics, Truth vs. Imputed



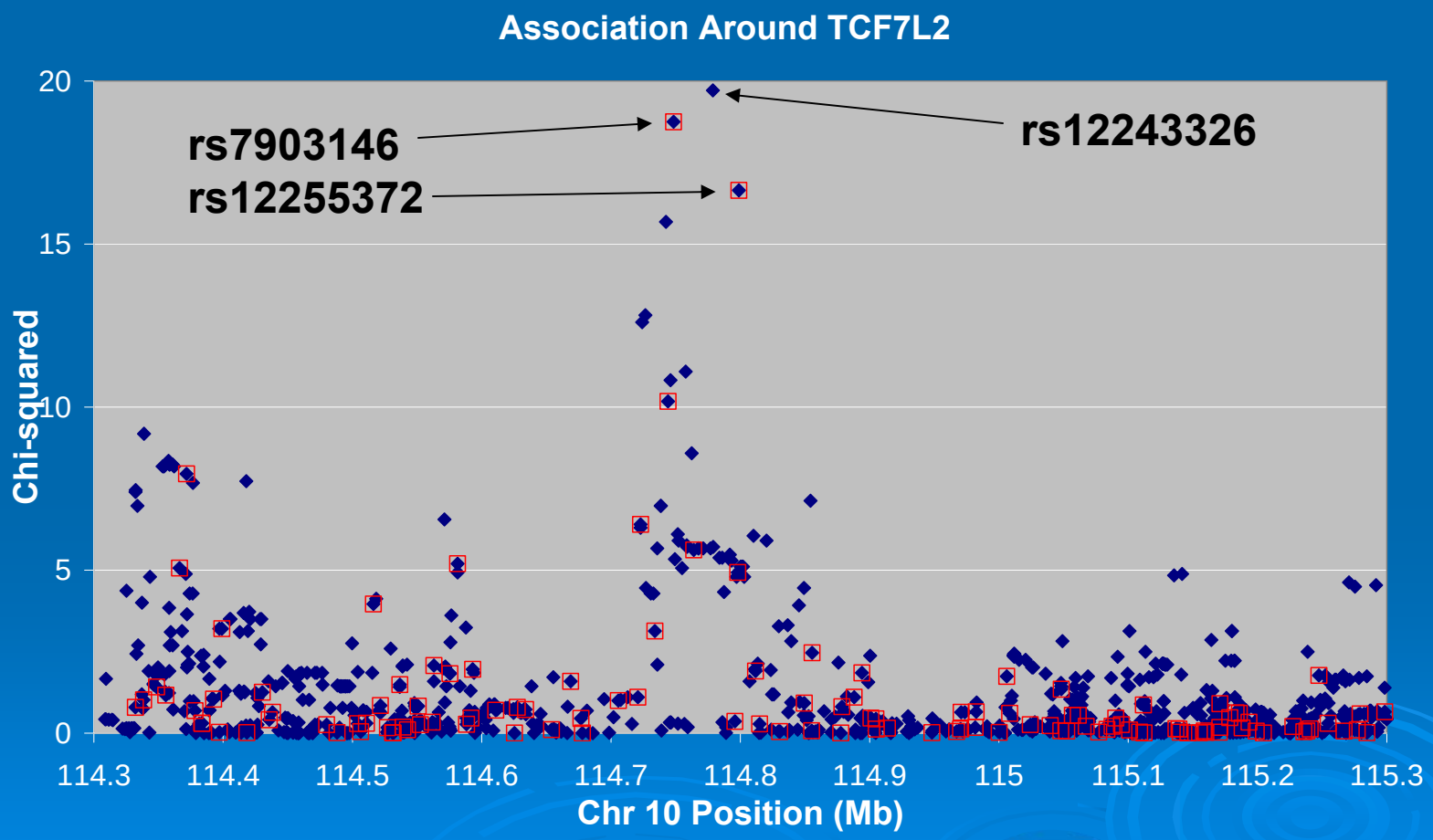
FUSION Example

- Finland United States Investigation of NIDDM Genetics
- Genome-wide association scan in 1200 type II diabetes cases and 1200 controls
  - Imputed 2.5M SNPs for all individuals
  - ~1 week, 50 CPUs
- Genotyping carried out using the Illumina 317K chip
  - To start, I will focus on 127 SNPs around TCF7L2
  - There are 984 Hapmap SNPs in the same interval

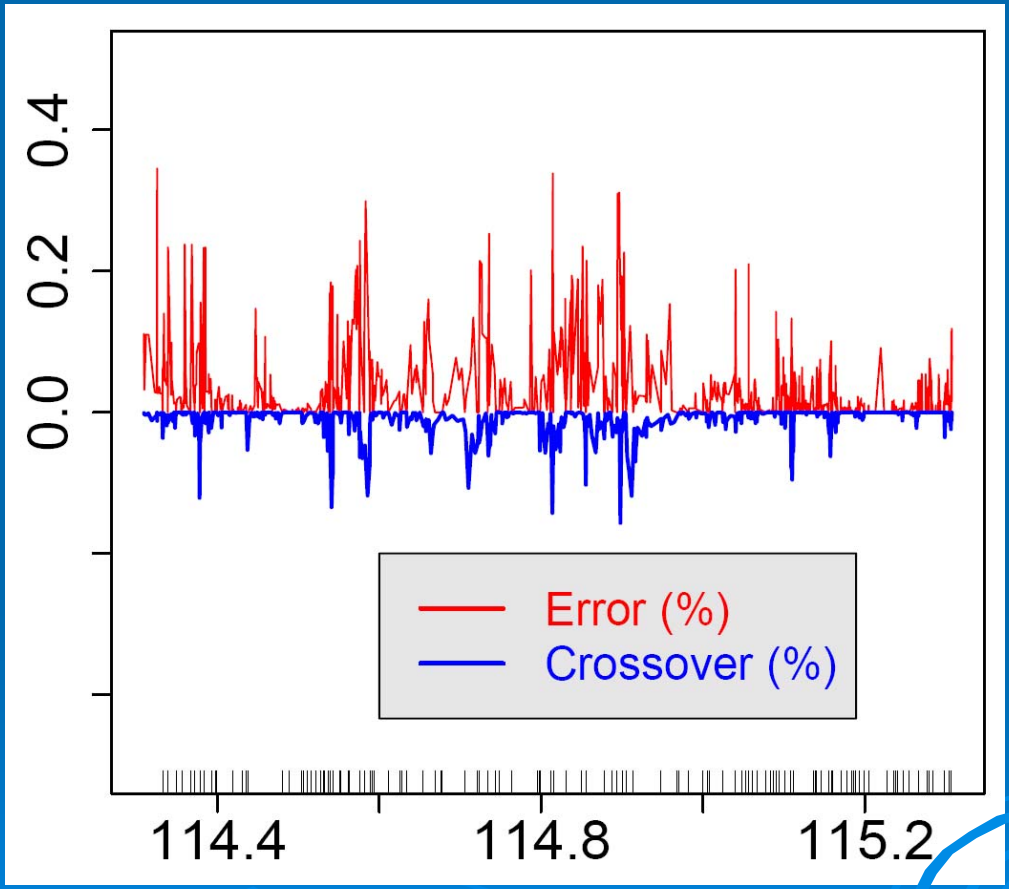
FUSION: TCF7L2



FUSION: TCF7L2



Imputed Data Includes Quality Estimates



FUSION TCF7L2 region. Estimated error rate, at each marker, based on similarity between haplotypes estimated at each iteration. Overall average is just under 3.0%. "Crossover" rates are averaged over Gibbs sampler iterations.

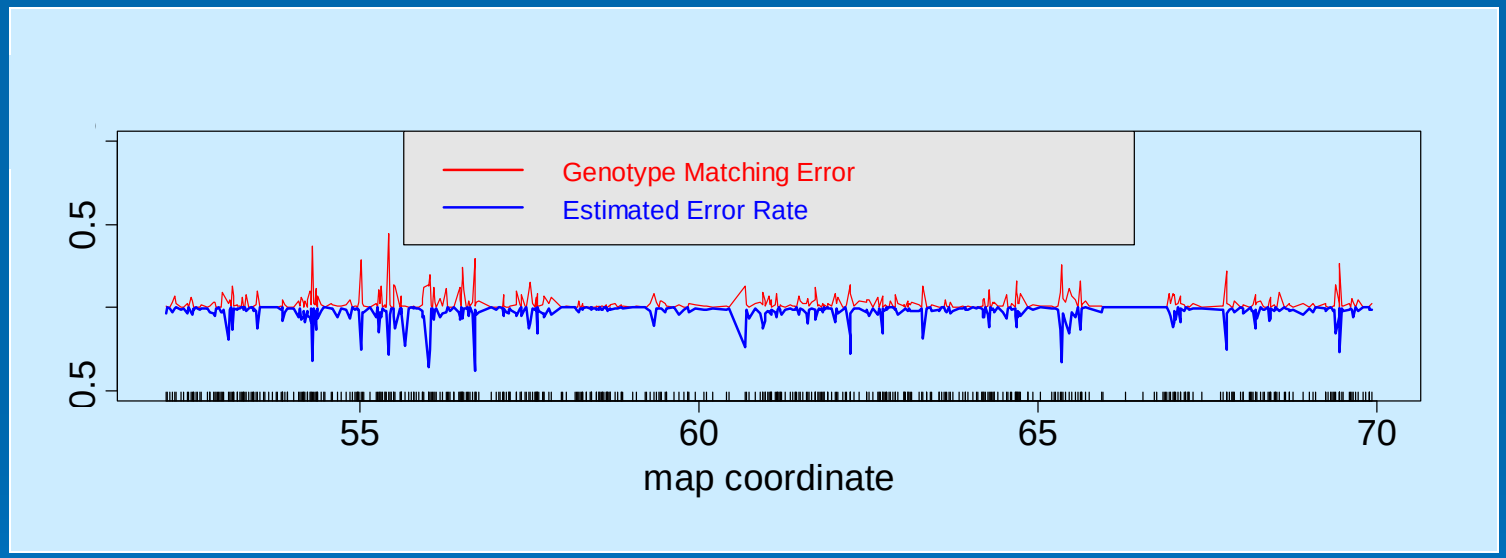
More Thorough Assessment

- Prior to genome-wide association scan
  - FUSION examined 20Mb region on chromosome 14
  - A candidate region that shows evidence for linkage
- The original genotype data
  - 1190 individuals
  - 521 markers not on Illumina HumanHap300 chip
- The imputed genotyped data
  - ~17,000 genotypes using ~2,000 GWA markers
  - ~1.5 days in a one CPU

Do the imputed alleles match?

- 1.5% of alleles mismatch original
  - 3.0% of genotypes mismatch original
- Errors are concentrated on a few markers
  - **14.82%** error for 1% of SNPs with lowest quality scores
  - **11.09%** error for next 1% of SNPs (1st – 2nd percentile)
  - **5.86%** error for next 1% of SNPs (2nd – 3rd percentile)
  - **1.11%** error for top 95% of SNPs

Predicted and Actual Error Rates



Top panel shows actual error rate (imputed vs. actual genotypes)
Bottom panel shows estimated error rate

Does Coverage Improve?

1. R^2 in FUSION with Best Tag in HapMap
2. R^2 in HapMap with Best Tag in HapMap
3. R^2 with (Best-Guess) Imputed Genotypes
4. Squared Allele Dosage Correlation

e.g., Imputed genotypes over 5 rounds:

1/1 1/1 1/2 1/1 1/1

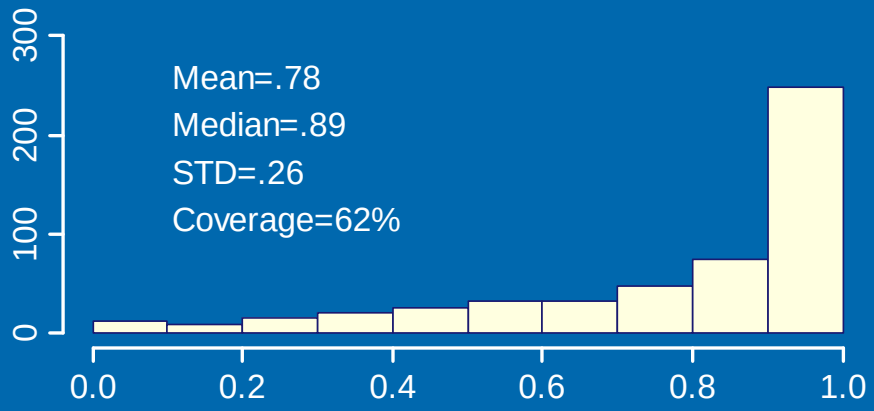
⇒ (Best-Guess) imputed genotype: **1/1**

⇒ Dosage for allele 1: **1.8**

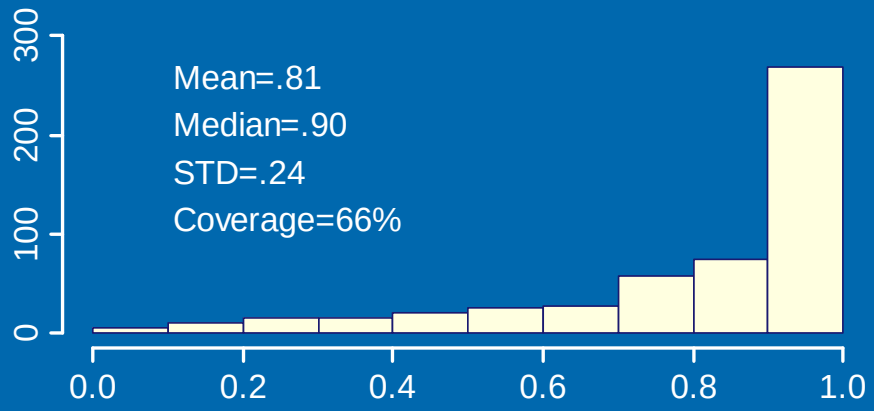
Coverage Comparison (r^2)

521 chromosome 14 SNPs

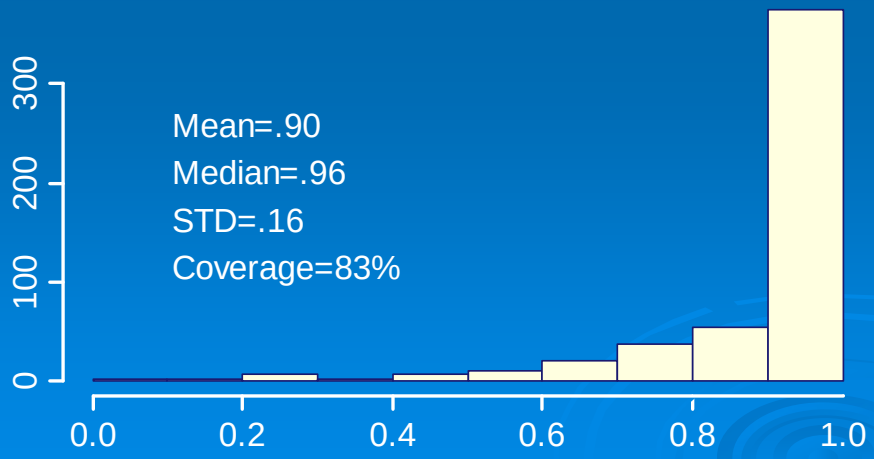
R^2 in FUSION with Best Tag in HapMap



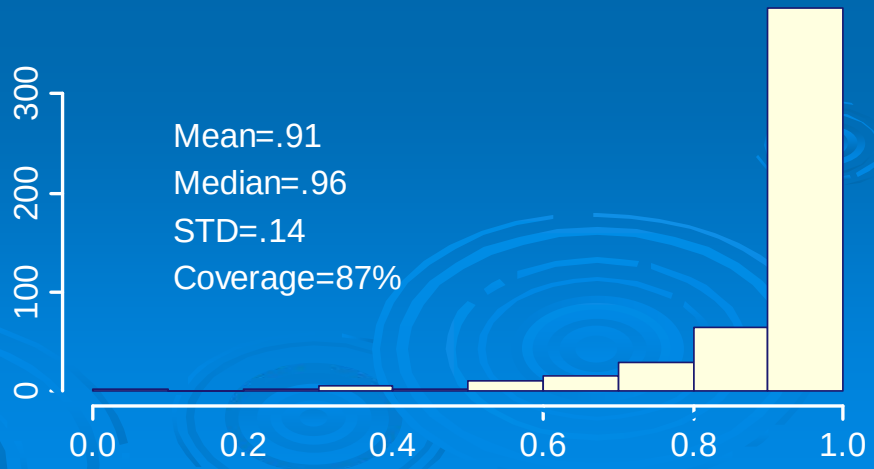
R^2 in HapMap with Best Tag in HapMap



R^2 in FUSION with Imputed Genotype



R^2 in FUSION with Imputed Allele Dose



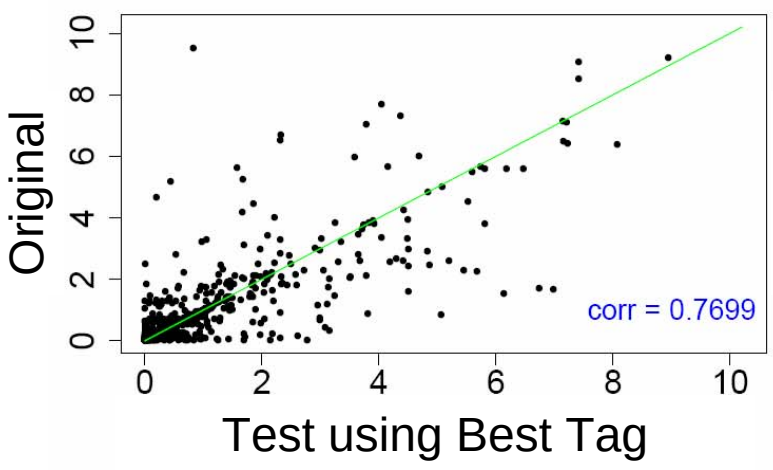
Can we recover original test statistics?

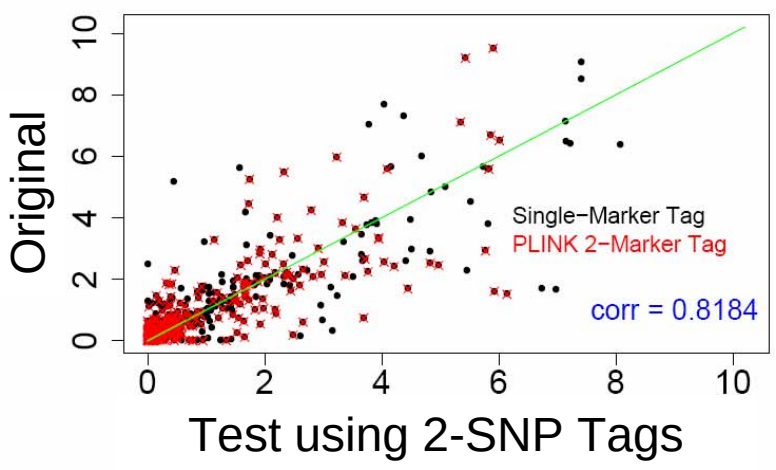
- Chi-squared test statistic in original data

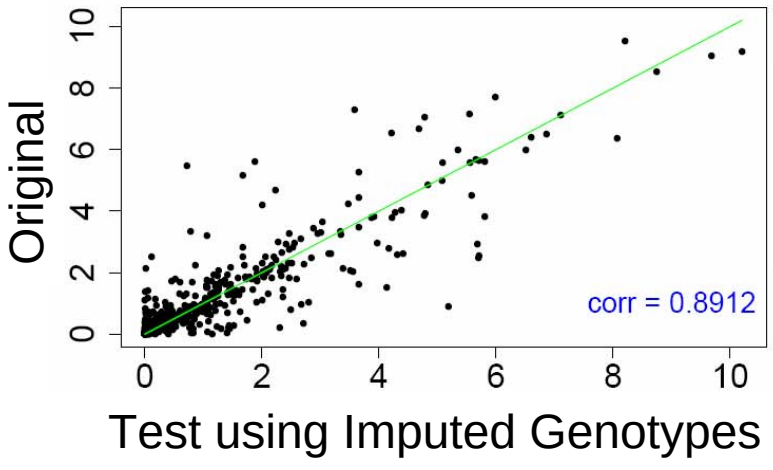
- Chi-squared test statistic for best tag
- Chi-squared test statistic for best 2-SNP tag
- Chi-squared test statistic for imputed alleles
- Chi-squared test statistics for allele doses

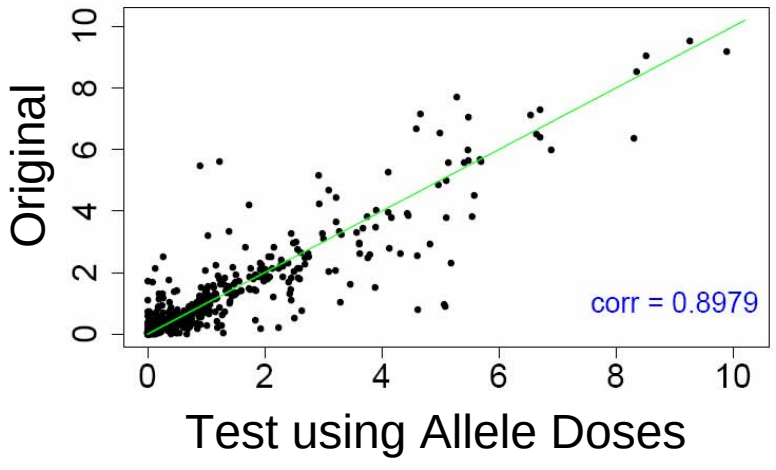
- Compare each of these 4 to original statistic

Test Statistic Comparison









Can we do even better?

- Ask a better statistician?
  - Jonathan Marchini / Peter Donnelly
  - Matthew Stephens
  - Mark Daly / Paul de Bakker
  - Many more?

Can we do even better?

- Ask a better statistician?
- Collect more data?
  - Genotype study samples on two platforms
  - 60 individuals in overlap, 1.78% error rate per allele
  - 100 individuals in overlap, 1.03% error rate
  - 200 individuals in overlap, 0.78% error rate
  - 500 individuals in overlap, 0.41% error rate
  - Maybe we could use a larger HapMap?

Summary

- It is possible to combine data across studies that rely on different platforms
  - Will add value to genome wide scans
- My (currently) favorite way is to impute missing genotypes
- A lot of interesting statistical and computational problems

Acknowledgements

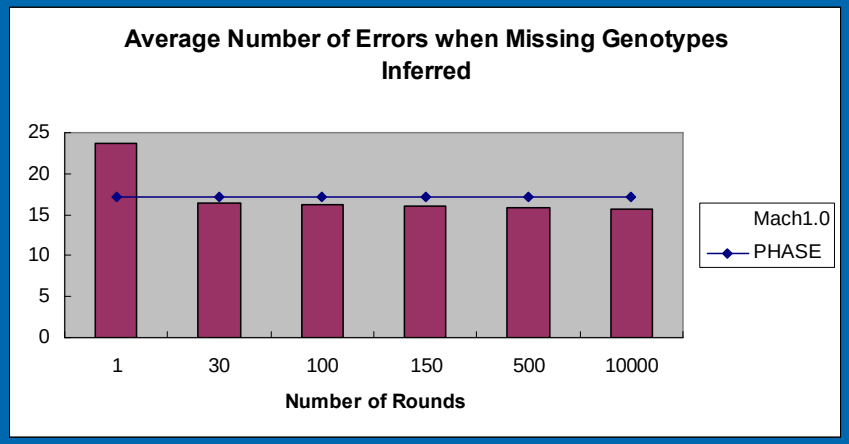
- Yun Li, Paul Scheet, Jun Ding, Weimin Chen, Serena Sanna

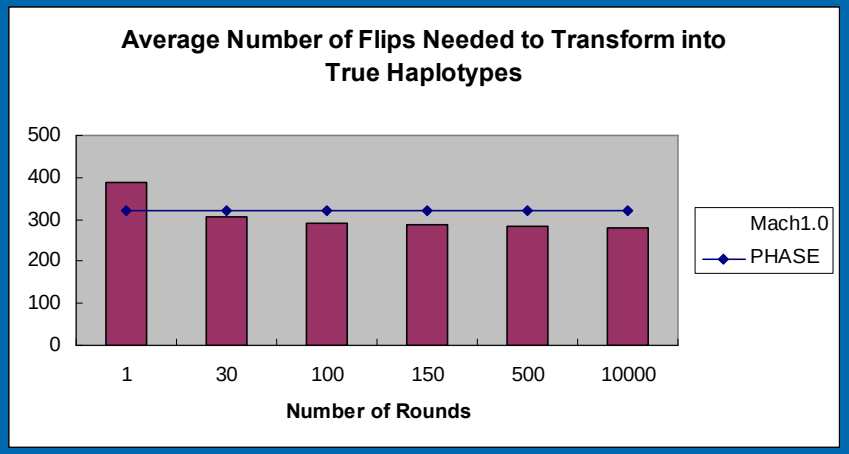
- FUSION Investigators, led by:
  - Karen Mohlke, Mike Boehnke, Francis Collins, Jaakko Tuomilehto, Richard Bergman

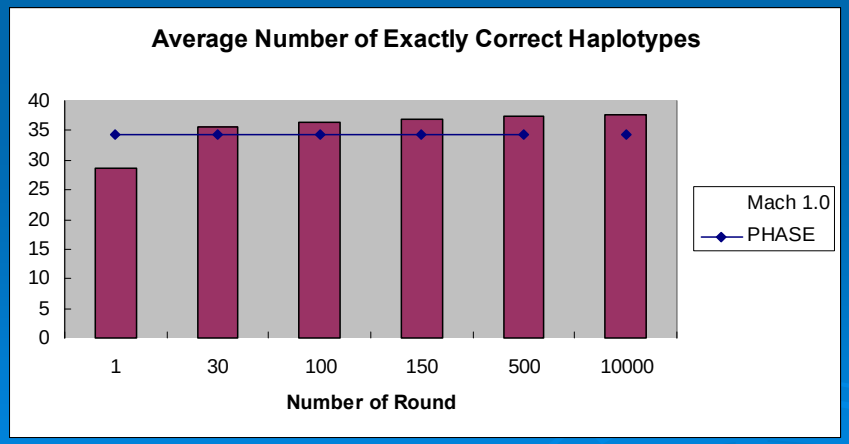
- Sardinia Investigators, led by:
  - David Schlessinger, Manuela Uda, Antonio Cao, Edward Lakatta, Paul Costa

goncalo@umich.edu

Comparison With Phase







Computation Time for this Dataset:

- **Mach 1.0**: ~3 sec per round.
- **PHASE**: ~10h in total.

Simulations follow model of Schaffner et al (2005), Marchini et al (2006).

Mathematical Model

- Markov model, where each haplotype is a mosaic of other “known” haplotypes
- The probability of a particular arrangement depends on number of change-over points

$$\Pr(\mathbf{S}=\mathbf{s}) = \Pr(S_1 = s_1,,..., S_L = s_L) = \Pr(S_1 = s_1)\prod_{j}^{L-1}\Pr(S_{j+1} = s_{j+1} \mid S_j = s_j)$$

- For a specific arrangement of the mosaic, calculate probability of observed alleles

$$\Pr(\mathbf{A}=\mathbf{a} \mid \mathbf{S}=\mathbf{s}) = \Pr(A_1 = a_1,..., A_L = a_L \mid S_1 = s_1,..., S_L = s_L) = \prod_{j}^{L}\Pr(A_j = a_j \mid E_j(s_j))$$