Genotype Imputation

Biostatistics 666

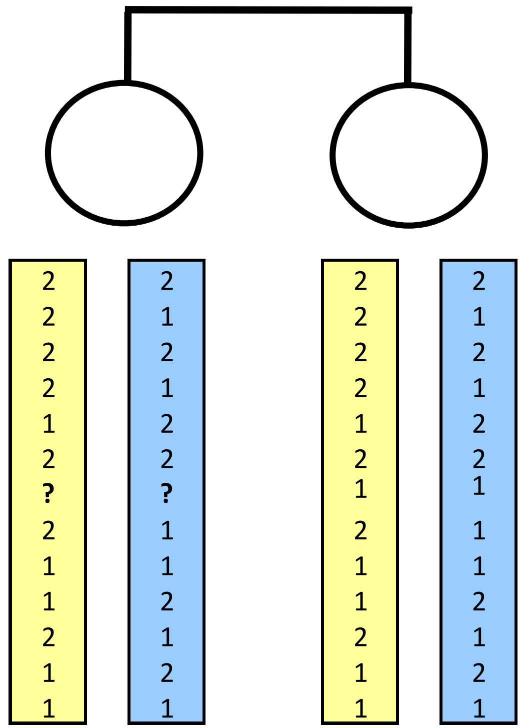
Previously

- Hidden Markov Models for Relative Pairs
  - Linkage analysis using affected sibling pairs
  - Estimation of pairwise relationships
- Identity-by-Descent
  - Relatives share long stretches of chromosome
  - Sharing at some markers can be used as surrogate for sharing at unobserved markers

Today

- Genotype Imputation / “In Silico” Genotyping
  - Use genotypes at a few markers to infer genotypes at other unobserved markers
- Closely related individuals
  - Long segments of identity by descent
- Distantly related individuals
  - Shorter segments of identity by descent

Intuition



Given the above pedigree, what are the likely values of the genotype marked ?/? …?

In Silico Genotyping For Family Samples

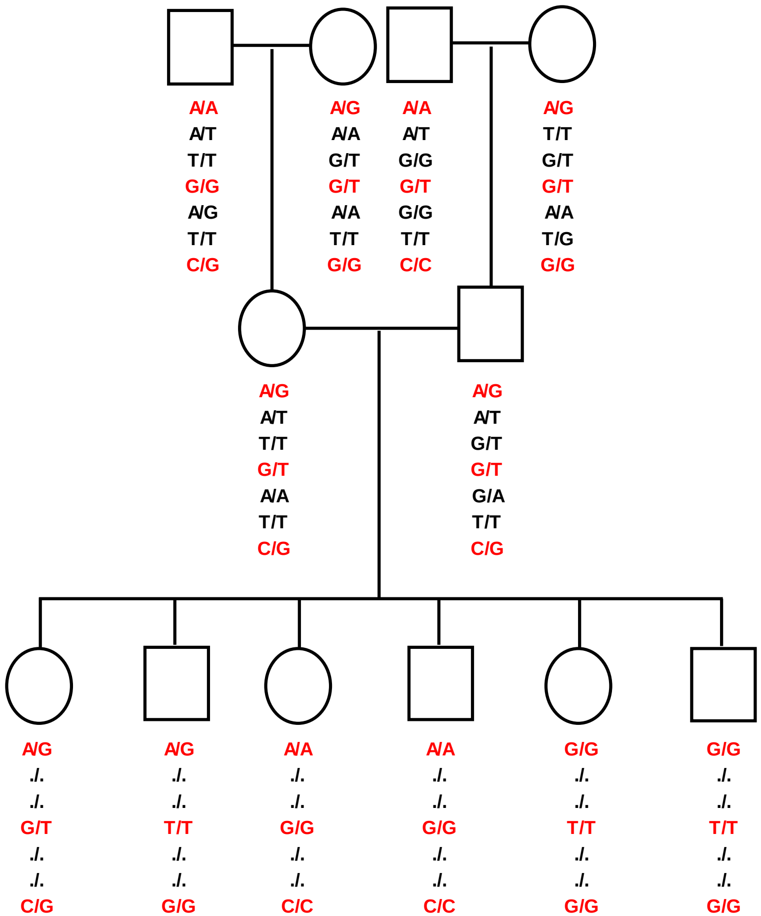
- Family members will share large segments of chromosomes

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

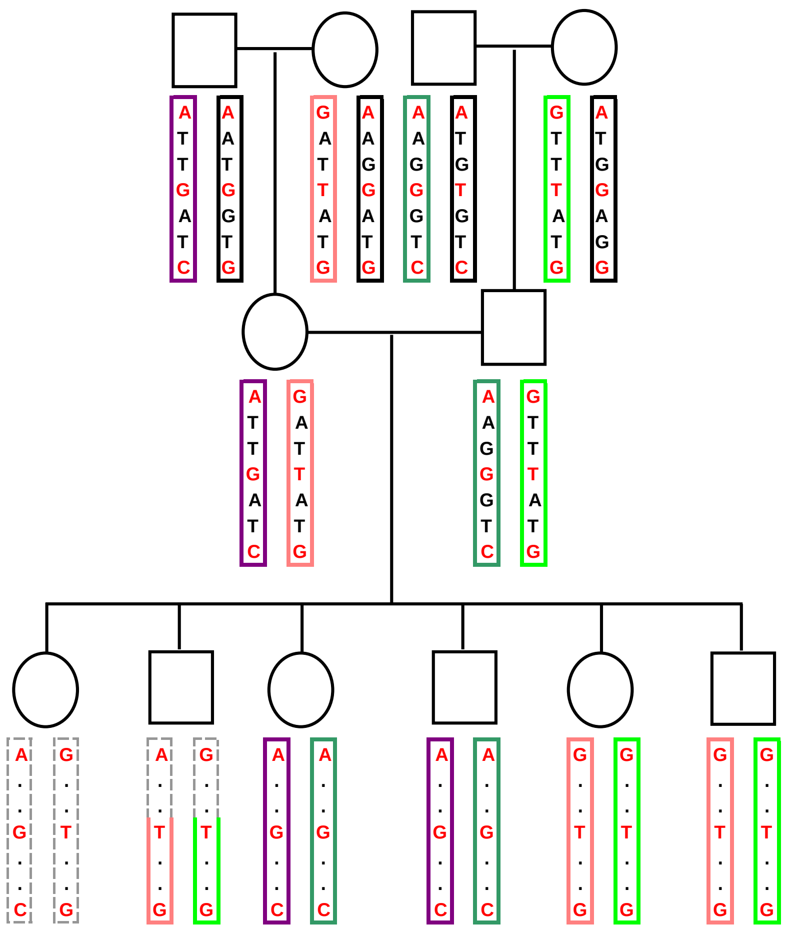
- In fact, we can:
  - Genotype a few markers on all individuals
  - Identify shared segments of haplotypes
  - Genotype additional markers on a subset of individuals
  - Fill in missing genotypes that fall in shared segments

  - Even without information on shared segments, it may be possible to learn about genotypes of relative members

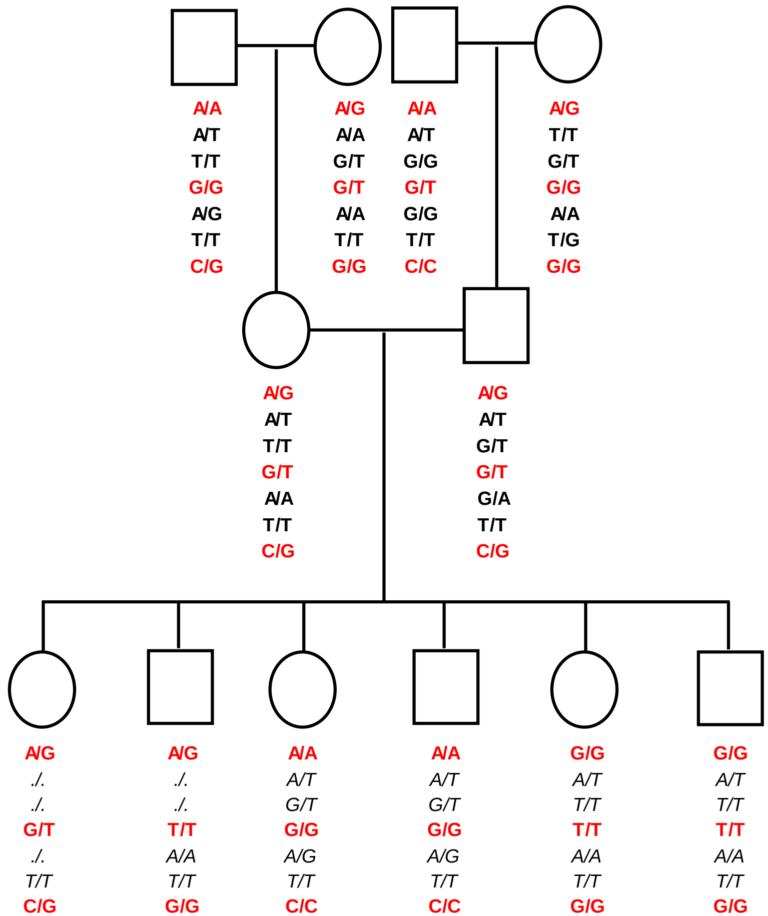
Genotype Inference
Part 1 – Observed Genotype Data



Genotype Inference
Part 2 – Inferring Allele Sharing



Genotype Inference
Part 3 – Imputing Missing Genotypes



Genotype Imputation in Families

- Suppose a particular genotype g_{ij} is missing
  - Genotype for person *i* at marker *j*
- Consider full set of observed genotypes G
- Evaluate pedigree likelihood L for each combination of $\{G, g_{ij} = x\}$
- Posterior probability that $g_{ij} = x$ is

$$P\left(g_{ij} = x \middle| G\right) = \frac{L(G, g_{ij} = x)}{L(G)}$$

- For pairs, same HMM as for linkage analysis or checking relatedness.
- Large pedigrees, Lander-Green (1987) or Elston-Stewart (1972) algorithm.

Standard Linear Model for Genetic Association

- Model association using a model such as:

$$E(y_i) = \mu + \beta_g g_i + \beta_c c_i + \cdots$$

- 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 the first allele
- 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:

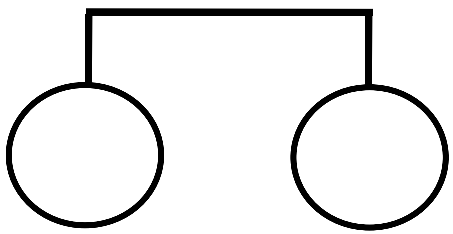
$$E(y_i) = \mu + \beta_g \bar{g} + \beta_c c + \cdots$$

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

- Association test can then be implemented in variance component framework, just as before

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

Example I



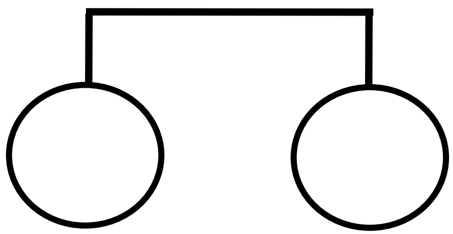
1/1 1/1
1/1 ./.
1/1 1/1

- $L(G) = .0061$
- $L(G, g_{22} = 1/1) = .00494$
- $L(G, g_{22} = 1/2) = .00110$
- $L(G, g_{22} = 2/2) = .00006$

- Assumptions:
  - Two alleles per marker
  - Equal allele frequencies
  - $\Theta = 0$

- $P(g_{22} = 1/1 | G) = 0.81$
- $P(g_{22} = 1/2 | G) = 0.18$
- $P(g_{22} = 2/2 | G) = 0.01$
- $\bar{g} = 1.80$

Example II



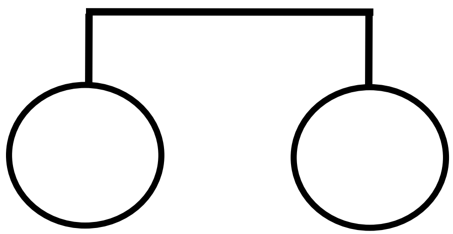
1/1	**2/2**
1/1	**./.**
1/1	**2/2**

- $L(G) = .000244$
- $L(G, g_{22} = 1/1) = .000061$
- $L(G, g_{22} = 1/2) = .000122$
- $L(G, g_{22} = 2/2) = .000061$

- Assumptions:
  - Two alleles per marker
  - Equal allele frequencies
  - $\Theta = 0$

- $P(g_{22} = 1/1|G) = 0.25$
- $P(g_{22} = 1/2|G) = 0.50$
- $P(g_{22} = 2/2|G) = 0.25$
- $\bar{g} = 1.00$

Example III

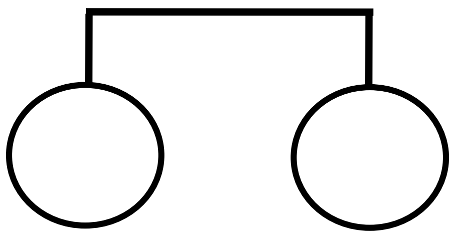


- L(G) = .0054
- L(G, g_{22} = 1/1) = .00392
- L(G, g_{22} = 1/2) = .00136
- L(G, g_{22} = 2/2) = .00012

- Assumptions:
  - Two alleles per marker
  - Equal allele frequencies
  - **Θ = 0.10**

- $P(g_{22} = 1/1|G) = 0.73$
- $P(g_{22} = 1/2|G) = 0.25$
- $P(g_{22} = 2/2|G) = 0.02$
- $\bar{g} = 1.70$

Example IV



1/1	**2/2**
1/1	**./.**
1/1	**2/2**

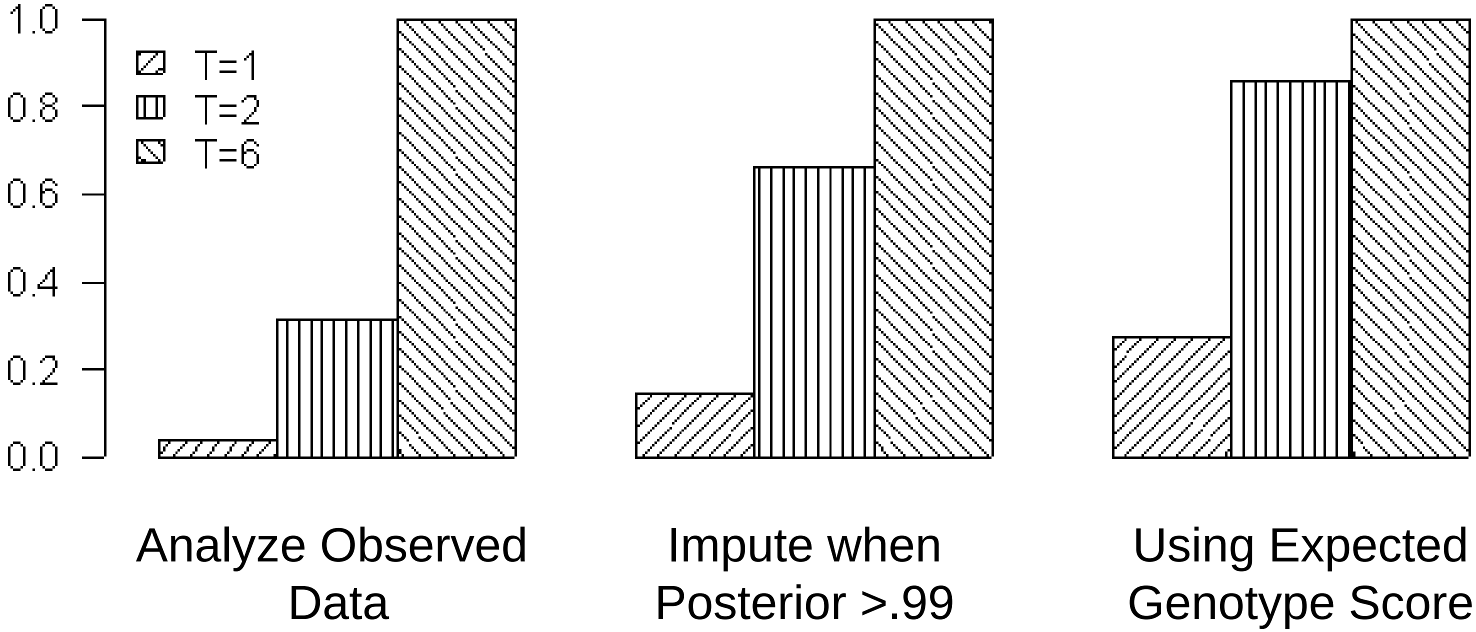
- $L(G) = .000121$
- $L(G, g_{22} = 1/1) = .000033$
- $L(G, g_{22} = 1/2) = .000061$
- $L(G, g_{22} = 2/2) = .000028$

- Assumptions:
  - Two alleles per marker
  - Equal allele frequencies
  - **$\Theta = 0.10$**

- $P(g_{22} = 1/1 | G) = 0.273$
- $P(g_{22} = 1/2 | G) = 0.499$
- $P(g_{22} = 2/2 | G) = 0.227$
- $\bar{g} = 1.05$

Power in Sibships of Size 6

Without Parental Genotype Data

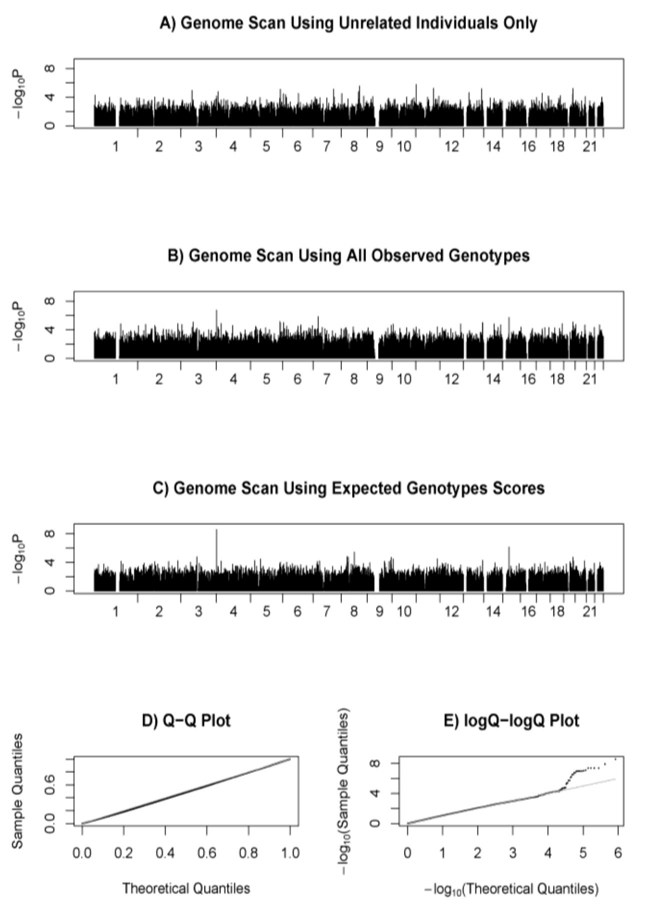


T is the number of genotyped offspring.
QTL explains 5% of variance, polygenes explain 35%,
250 sibships, $\alpha = 0.001$.

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
- SNP consortium 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)

Point of Situation…

- When analyzing family samples …

- FOR INDIVIDUALS WITH KNOWN RELATIONSHIPS
  - Impute genotypes in relatives
  - 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 Unrelated Individuals

- In families, long stretches of shared chromosome
- In unrelated individuals, shared stretches are much shorter
- The plan is still to identify stretches of shared chromosome between individuals…
- … we then infer intervening genotypes by contrasting samples typing at a few sites with those with denser genotypes

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

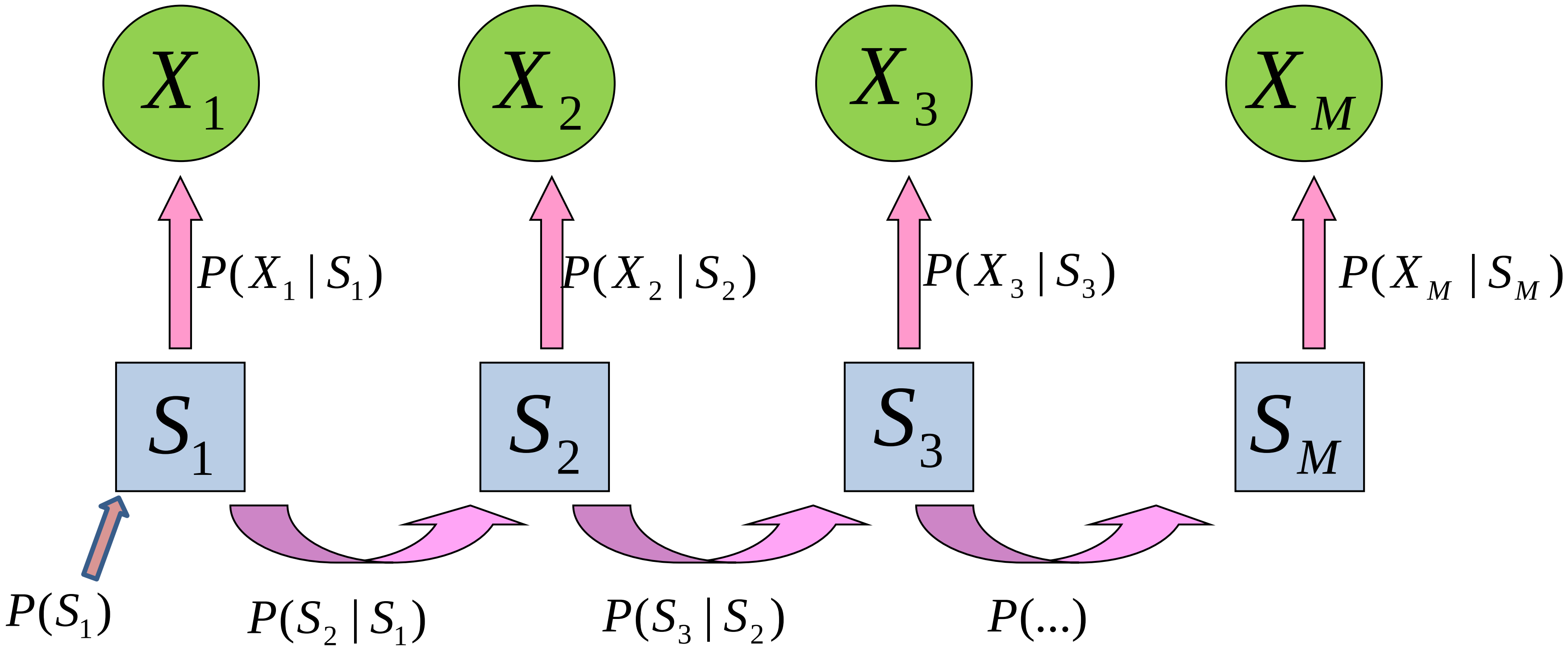
C G A G A C T C T C C G A C C T C G T G C

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
- At each position, we assume that the haplotype being modeled copies a template haplotype
- Each individual has two haplotypes, and therefore copies two template haplotypes

Markov Model



The final ingredient connects template states along the chromosome …

Possible States

- A state *S* selects pair of template haplotypes
  - Consider S_i as vector with two elements ($S_{i,1}$, $S_{i,2}$)

- With *H* possible haplotypes, H^2 possible states
  - $H(H+1)/2$ of these are distinct

- A recombination rate parameter describes probability of switches between states
  - $P((S_{i,1} = a, S_{i,2} = b) \rightarrow (S_{i+1,1} = a, S_{i+1,2} = b))$ $(1-\theta)^2$
  - $P((S_{i,1} = a, S_{i,2} = b) \rightarrow (S_{i+1,1} = a^*, S_{i+1,2} = b))$ $(1-\theta)\theta/H$
  - $P((S_{i,1} = a, S_{i,2} = b) \rightarrow (S_{i+1,1} = a^*, S_{i+1,2} = b^*))$ $(\theta/H)^2$

Emission Probabilities

- Each value of *S* implies expected pair of alleles
- Emission probabilities will be higher when observed genotype matches expected alleles
- Emission probabilities will be lower when alleles mismatch
- Let *T(S)* be a function that provides expected allele pairs for each state *S*

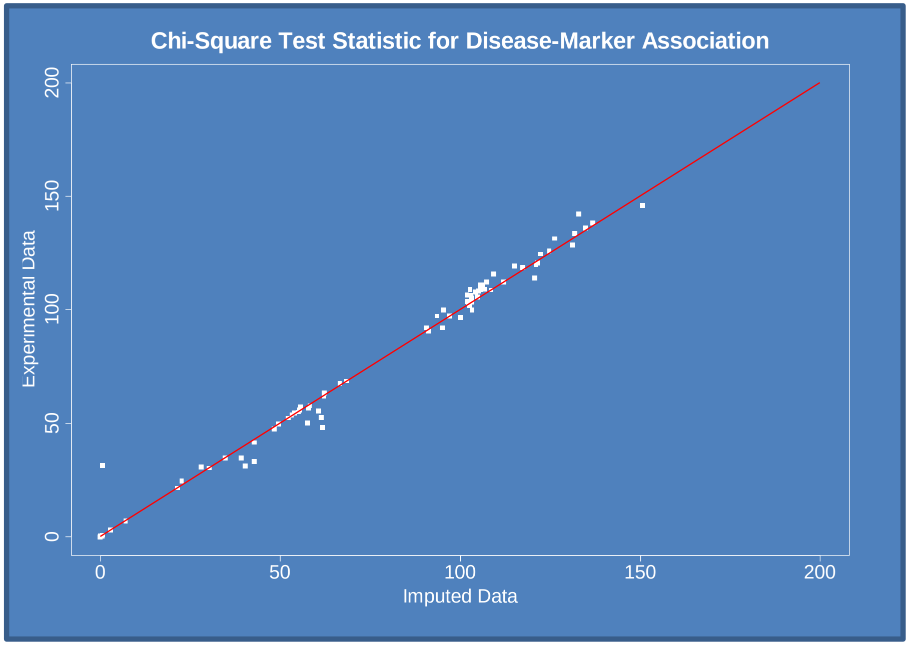
Emission Probabilities

$$P(G_j|S_j) = \begin{cases} (1-\varepsilon_j)^2+\varepsilon_j^2, & T(S_j)=G_j \text{ and } G_j \text{ is heterozygote,} \\ 2(1-\varepsilon_j)\varepsilon_j, & T(S_j)\neq G_j \text{ and } G_j \text{ is heterozygote,} \\ (1-\varepsilon_j)^2, & T(S_j)=G_j \text{ and } G_j \text{ is homozygote,} \\ (1-\varepsilon_j)\varepsilon, & T(S_j) \text{ is heterozygote and} \\ & \quad G_j \text{ homozygote,} \\ \varepsilon_j^2, & T(S_j) \text{ and } G_j \text{ are opposite} \\ & \quad \text{homozygotes.} \end{cases}$$

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 various haplotyping methods

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

- 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

Scott et al, *Science*, 2007

Does this really, really work?

- 90 GAIN psoriasis study samples were re-genotyped for 906,600 SNPs using the Affymetrix 6.0 chip.

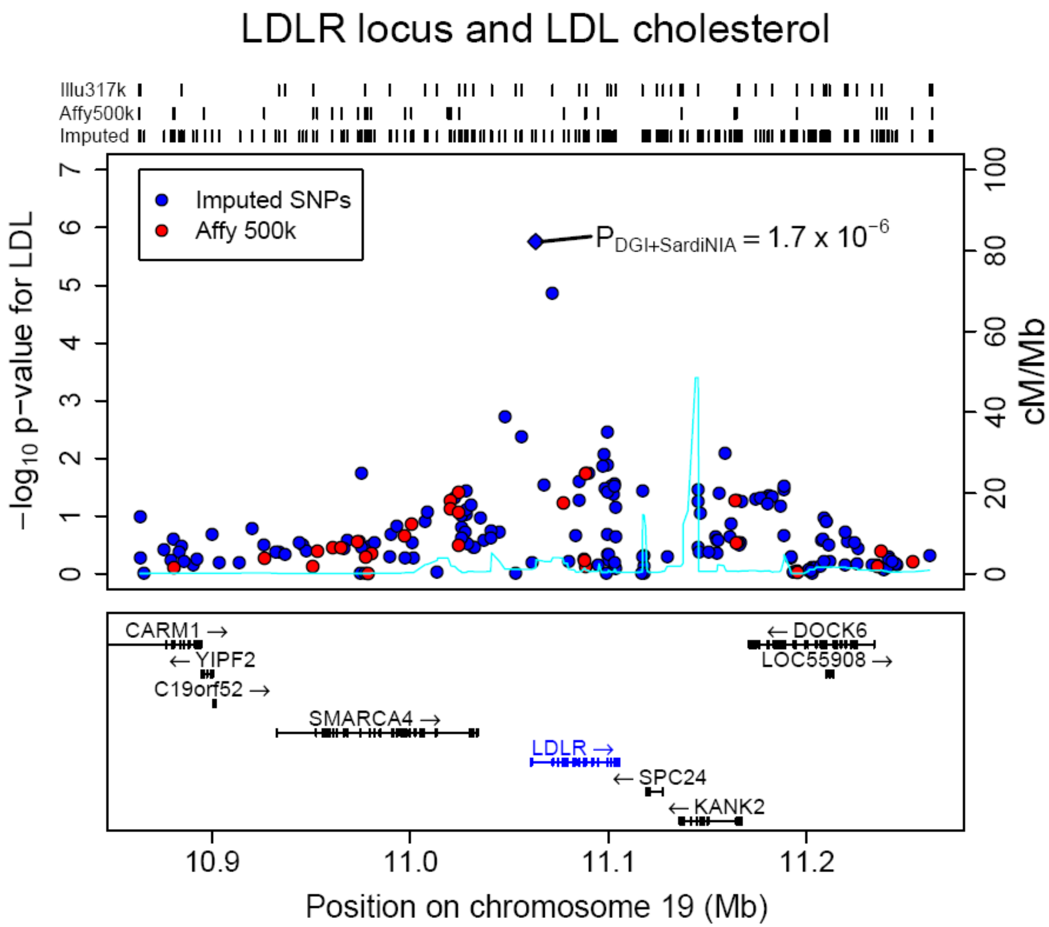
- Comparison of 15,844,334 genotypes for 218,039 SNPs that overlap between the Perlegen and Affymetrix chips resulted in discrepancy rate of 0.25% per genotype (0.12% per allele).

- 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.

Nair et al, *Nature Genetics*, 2009

LDLR and LDL example



Willer et al, *Nature Genetics*, 2008
Li et al, Annual Review of Genomics and Human Genetics, 2009

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 Genotyped SNPs.

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

Combined Lipid Scans

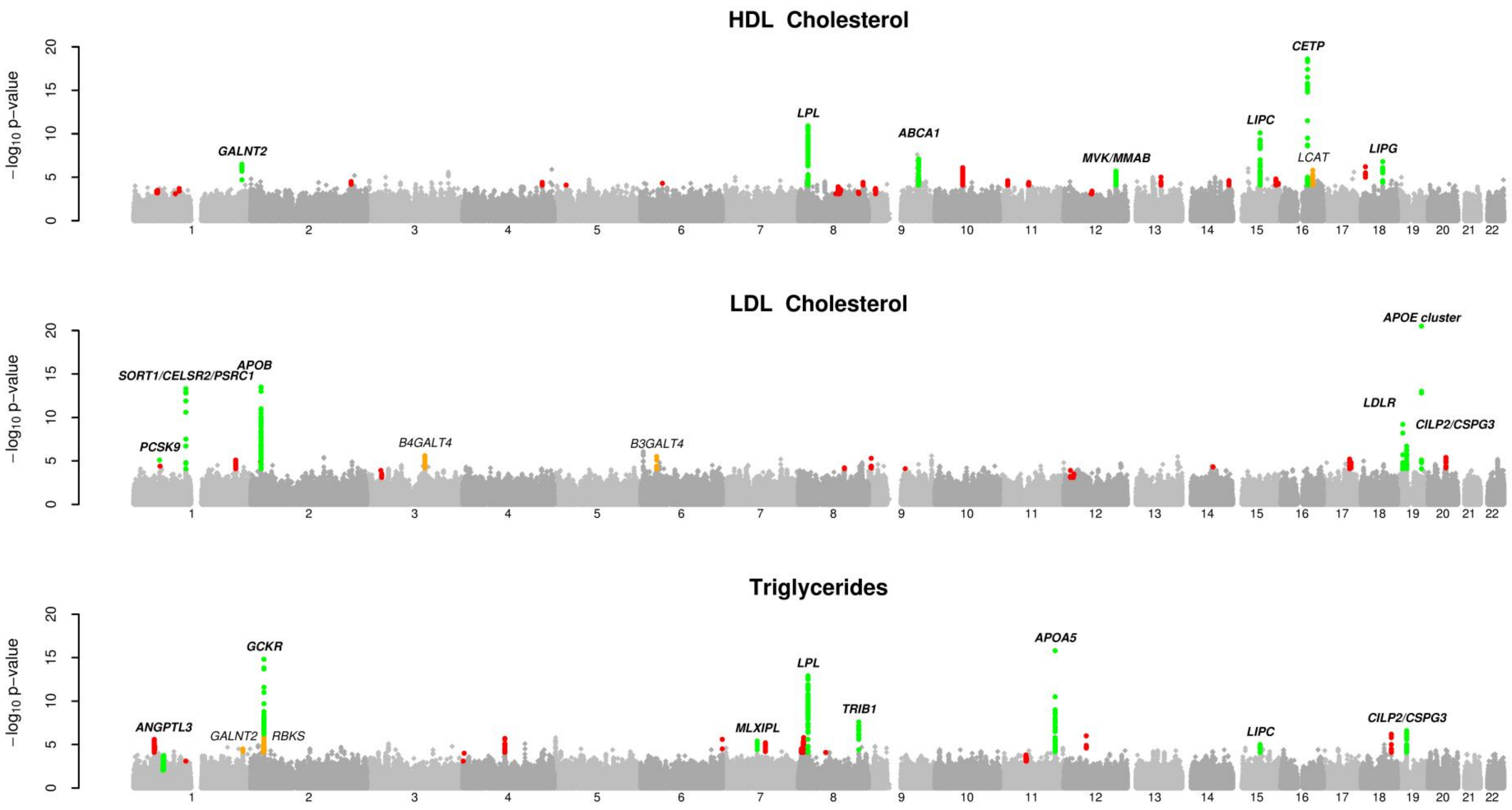
- SardiNIA (Schlessinger, Uda, et al.)
  - ~4,300 individuals, cohort study
- FUSION (Mohlke, Boehnke, Collins, et al.)
  - ~2,500 individuals, case-control study of type 2 diabetes
- DGI (Kathiresan, Altshuler, Orho-Mellander, et al.)
  - ~3,000 individuals, case-control study of type 2 diabetes

- 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

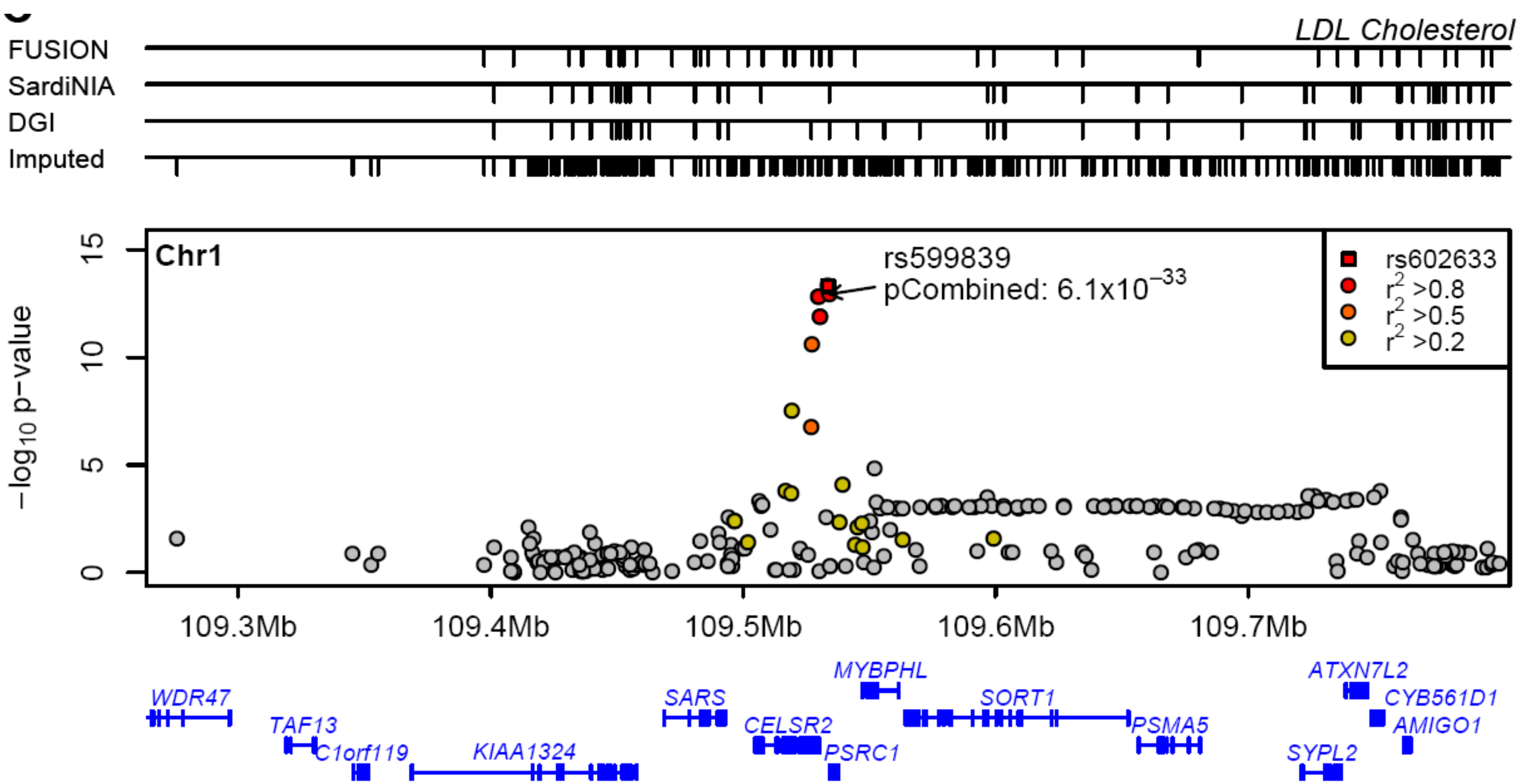
Willer et al, *Nature Genetics*, 2008

Combined Lipid Scan Results
18 clear loci!



Willer et al, *Nature Genetics*, 2008

New LDL Locus,
Previously Associated with CAD



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

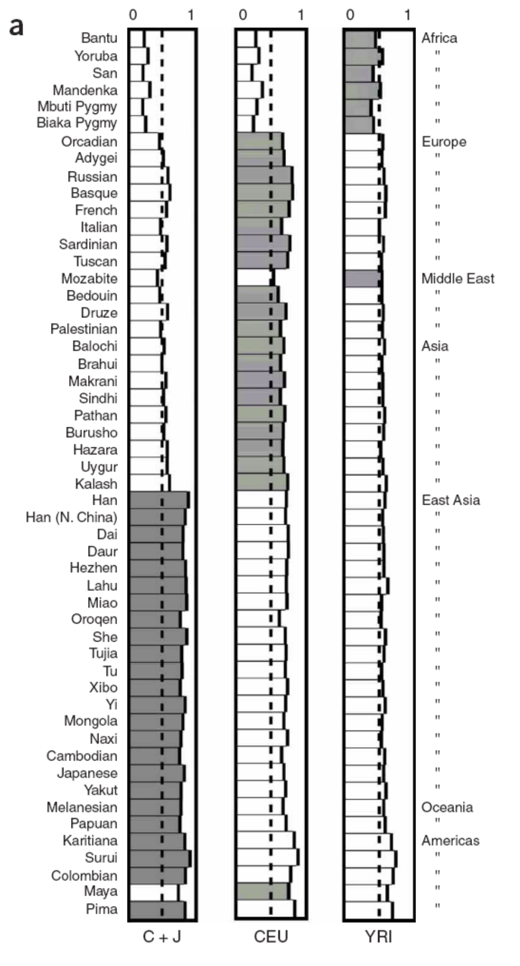
Gene	LDL-C p-value	Frequency CAD cases	Frequency CAD ctrls	CAD p-value	OR
APOE/C1/C4	**3.0×10^{-43}**	**.209**	**.184**	**1.0×10^{-4}**	**1.17 (1.08-1.28)**
APOE/C1/C4	**1.2×10^{-9}**	**.339**	**.319**	**.0068**	**1.10 (1.02-1.18)**
SORT1	**6.1×10^{-33}**	**.808**	**.778**	**1.3×10^{-5}**	**1.20 (1.10-1.31)**
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)**

Comparison to data from WTCCC (Nature, 2007) was made possible by imputation.

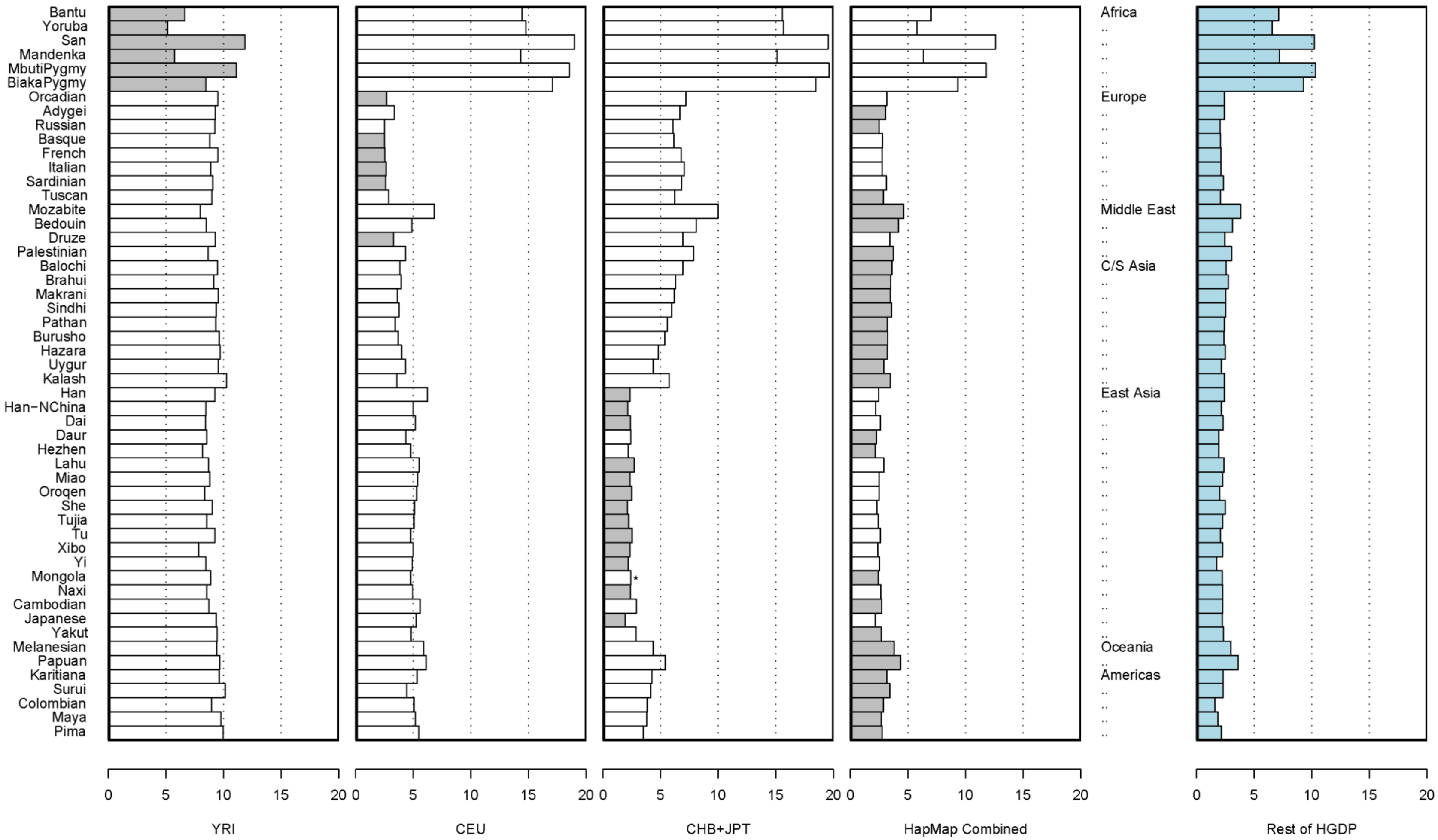
Does This Work Across Populations?

- 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



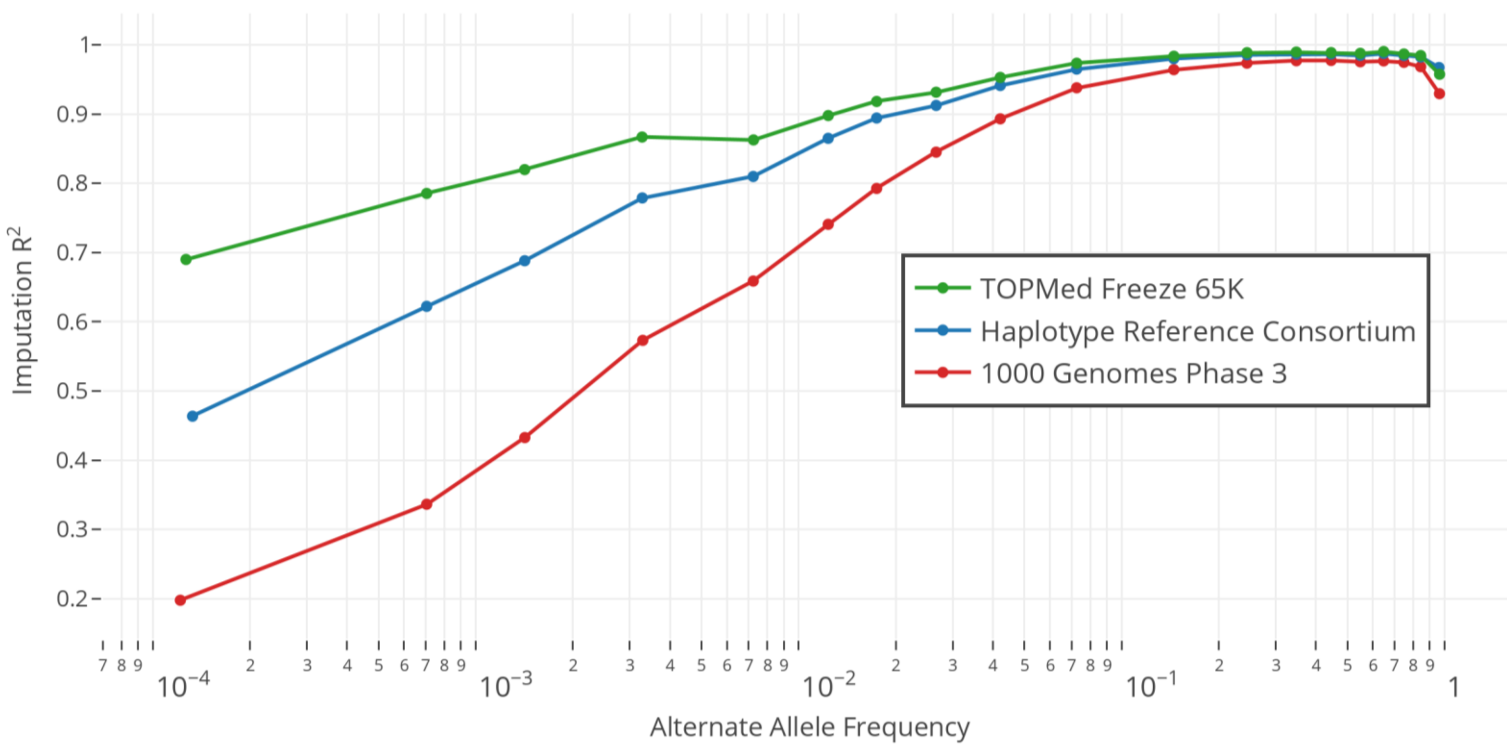
(Evaluation Using ~1 SNP per 10kb in 52 x 300kb regions For Imputation)

Summary

- Genotype imputation can be used to accurately estimate missing genotypes
- Genotype imputation is usually implemented through using a Hidden Markov Model
- Benefits of genotype imputation
  - Increases power of genetic association studies
  - Facilitates analyses that combine data across studies
  - Facilitates interpretation of results

2017 Imputation Accuracy: Europeans

(Complete Genomics as Truth)



Imputatior

https://imputationserver.sph.umich.edu

Michigan Imputation Server

This server provides a free genotype imputation service. You can upload GWAS genotypes (VCF or 23andMe format) and receive phased and imputed genomes in return. Our server offers imputation from HapMap, 1000 Genomes (Phase 1 and 3), CAAPA and the updated Haplotype Reference Consortium (HRC version r1.1) panel. Learn more or follow us on Twitter.

13.4M Genomes

2,623 Users

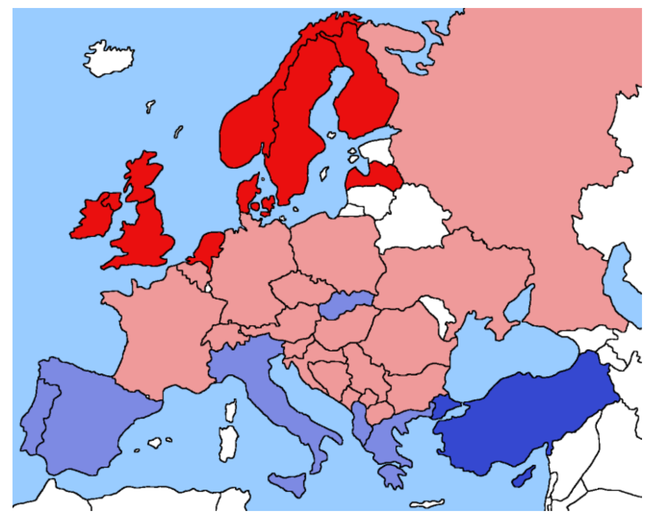
Sign up now | Login

The easiest way to impute genotypes

Upload your genotypes to our server located in Michigan. All interactions with the server are **secured**.

Choose a reference panel. We will take care of pre-phasing and imputation.

Download the results. All results are encrypted with a one-time password. After 7 days, all results are deleted from our server.



Recommended Reading

- Chen and Abecasis (2007) Family based association tests for genome wide association scans. *Am J Hum Genet* **81:**913-926

- Li et al (2010) Using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genetic Epidemiology* **34**:816-834