HOS 6236 Molecular Marker Assisted Plant Breeding

Fall 2017

Last Class:

GWAS

Today's Class:

GWAS and DNA imputation

Gene X Environment interaction

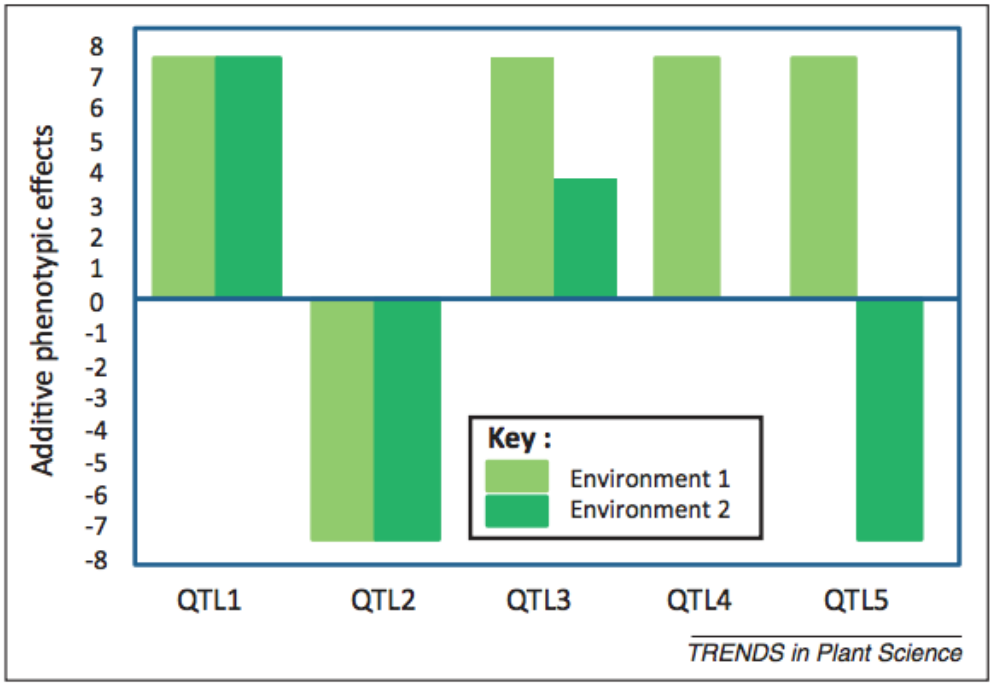


Figure 3. Environment-specific quantitative trait loci (QTL) effects. QTL1 and QTL2 are constitutive QTL with positive or negative additive phenotypic effects, respectively, which do not show a QTL× environment interaction (Q × E). QTL3 shows Q × E because its effect on the phenotype is stronger in one environment than in the other. QTL4 also shows Q × E; however, it is conditionally neutral because it is only detected in environment 1. QTL5 shows the strongest Q × E because it has opposite phenotypic effects when comparing both environments.

El-Soda *et al.* 2014

Leading Edge
Perspective

Cell

An Expanded View of Complex Traits: From Polygenic to Omnigenic

Evan A. Boyle,[1,*] Yang I. Li,[1,*] and Jonathan K. Pritchard[1,2,3,*]
[1]Department of Genetics
[2]Department of Biology
[3]Howard Hughes Medical Institute
Stanford University, Stanford, CA 94305, USA
*Correspondence: eaboyle@stanford.edu (E.A.B.), yangili@stanford.edu (Y.I.L.), pritch@stanford.edu (J.K.P.)
http://dx.doi.org/10.1016/j.cell.2017.05.038

- ”A second surprise was that, in contrast to Mendelian diseases—which are largely caused by protein-coding changes (Botstein and Risch, 2003)—complex traits are mainly driven by noncoding variants that presumably affect gene regulation (Pickrell, 2014; Welter et al., 2014; Li et al., 2016).”

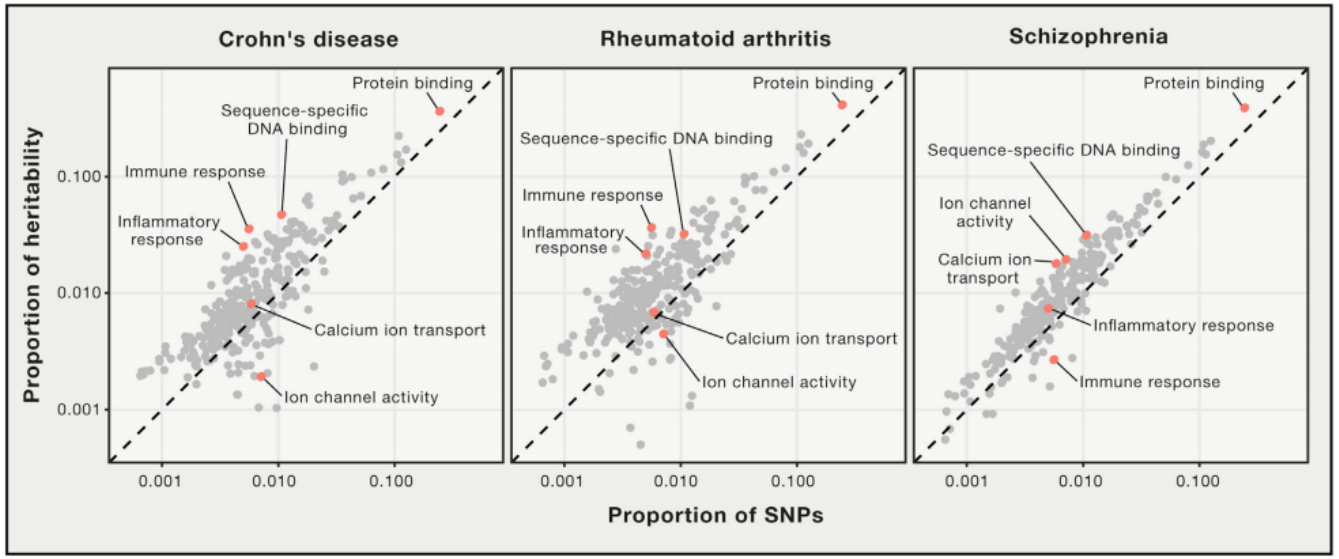


Figure 3. Gene Ontology Enrichments for Three Diseases, with Categories of Particular Interest Labeled
The x axis indicates the fraction of SNPs in each category; the y axis shows the fraction of heritability assigned to each category as a fraction of the heritability assigned to all SNPs. Note that the diagonal indicates the genome-wide average across all SNPs; most GO categories lie above the line due to the general enrichment of signal in and around genes. Analysis by stratified LD score regression (Finucane et al., 2015).

ARTICLE

OPEN

doi:10.1038/nature24277

Genetic effects on gene expression across human tissues

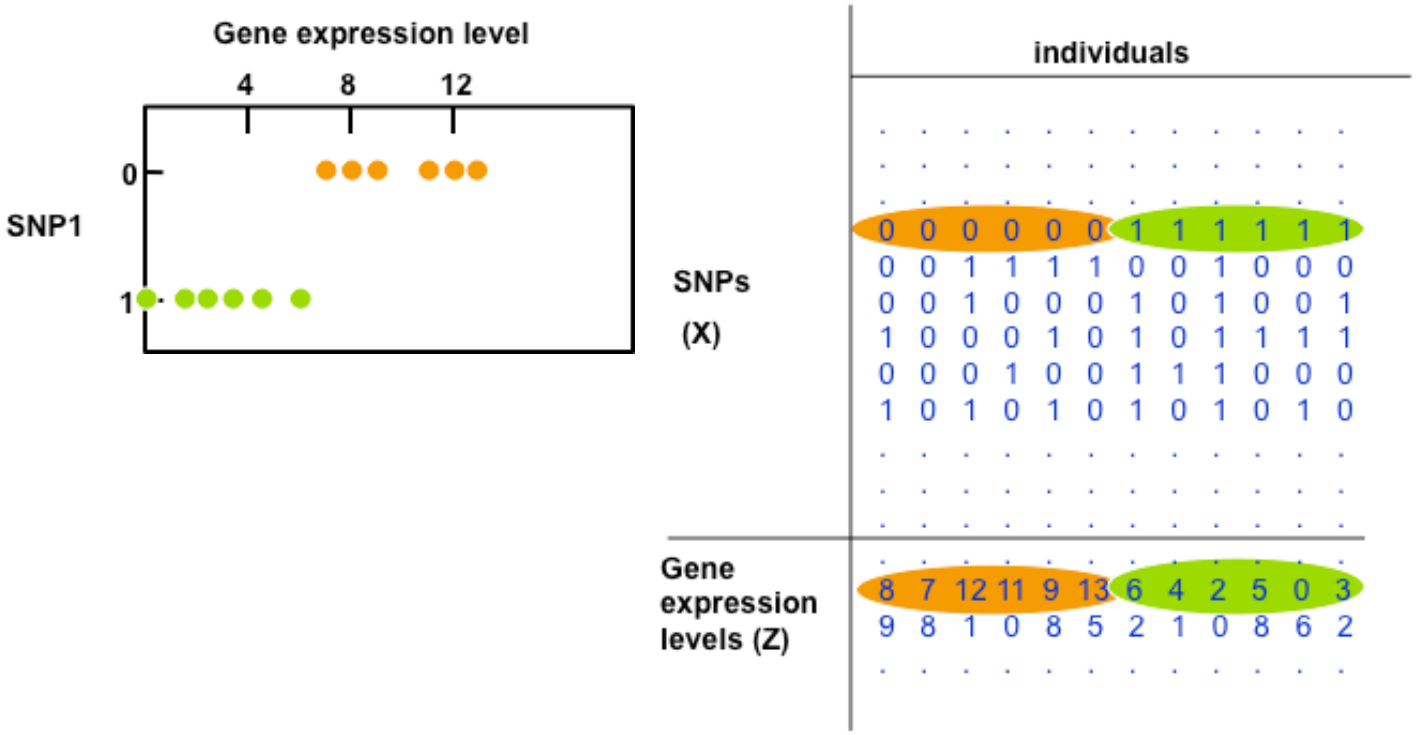
GTEx Consortium*

Characterization of the molecular function of the human genome and its variation across individuals is essential for identifying the cellular mechanisms that underlie human genetic traits and diseases. The Genotype–Tissue Expression (GTEx) project aims to characterize variation in gene expression levels across individuals and diverse tissues of the human body, many of which are not easily accessible. Here we describe genetic effects on gene expression levels across 44 human tissues. We find that local genetic variation affects gene expression levels for the majority of genes, and we further identify inter-chromosomal genetic effects for 93 genes and 112 loci. On the basis of the identified genetic effects, we characterize patterns of tissue specificity, compare local and distal effects, and evaluate the functional properties of the genetic effects. We also demonstrate that multi-tissue, multi-individual data can be used to identify genes and pathways affected by human disease-associated variation, enabling a mechanistic interpretation of gene regulation and the genetic basis of disease.

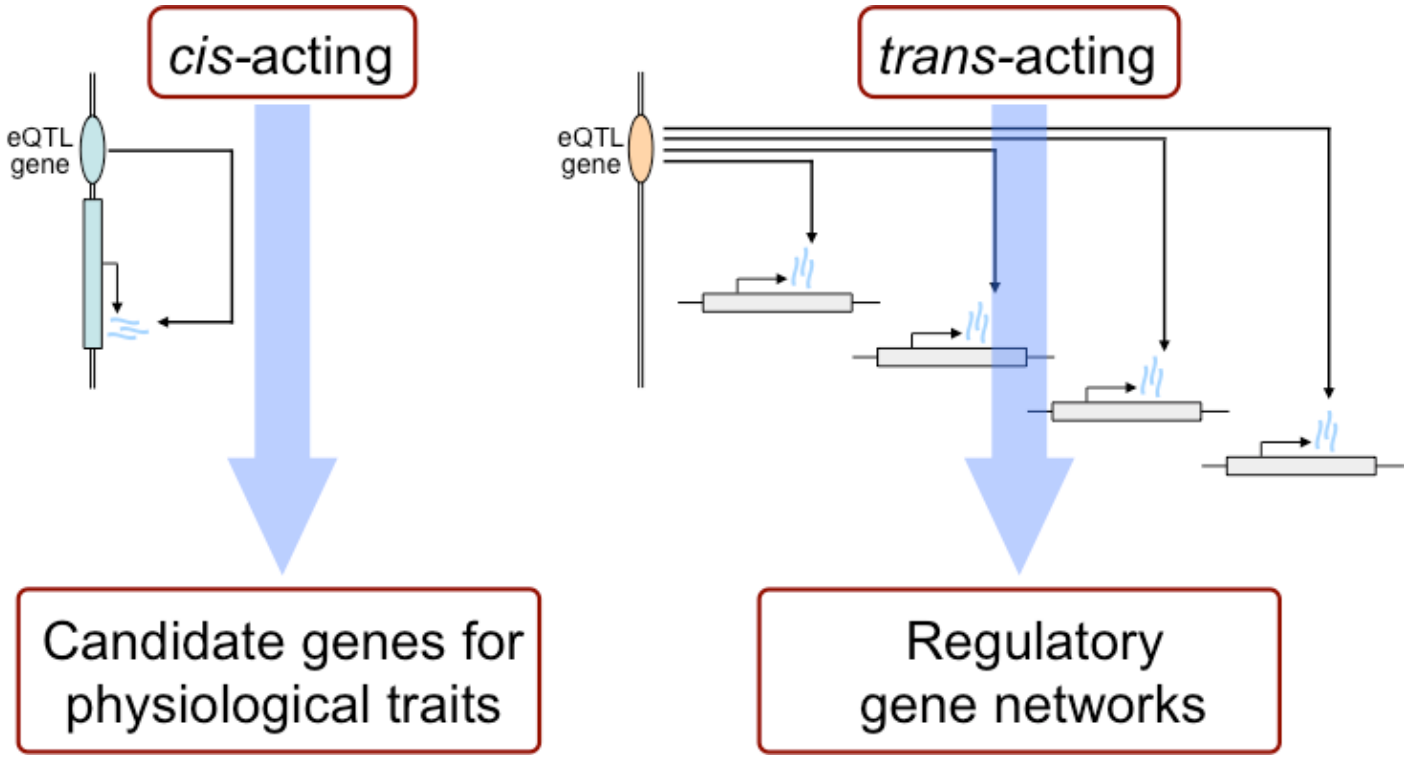
(e)QTL Mapping

- Many disease associated genes have been mapped with QTL
- eQTL mapping:
  - Transcript abundance may act as intermediate phenotype between genetic loci and the phenotype. This abundance may be genetically regulated.
  - Incorporate information of genotype, expression, and phenotypic traits together to construct regulatory networks and to improve understanding of the target phenotype.

(e)QTL Mapping



cis- and trans-acting eQTLs



Genotype Imputation

- Process of predicting genotypes that are not directly assayed by a given genotyping platform.

Why should we impute genotypes?

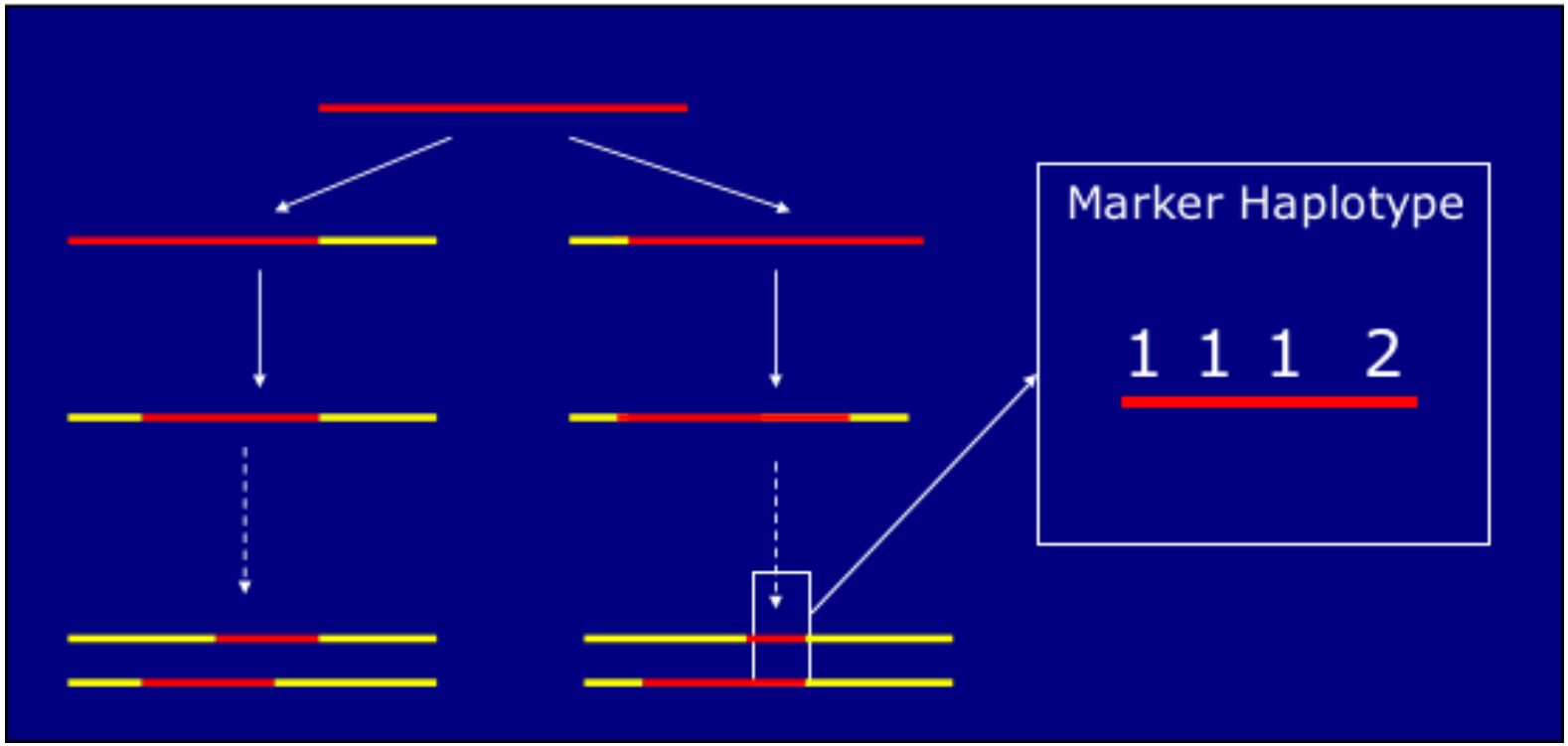
- Fill in missing genotypes from the lab
- Merge data sets with genotypes on different arrays
  - Eg. Affy and Illumina data
- Impute from low density to high density
  - 7K-> 50K (save $$$)
  - 50K->800K
  - 'In silico' genotypes can then be used to boost the number of SNPs that can be tested for association. This increases the power of the study Better persistence of accuracy
- Sequence expensive, can we impute to full sequence data?

Important concepts

- **Identity by state (IBS)**
  - A pair of individuals have the same allele at a locus
- **Identity by descent (IBD)**
  - A pair of individuals have the same alleles at a locus and it traces to a common ancestor
- Imputation methods determine whether a chromosome segment is IBD

Tracing segments IBD

- A chunk of ancestral chromosome is conserved in the current population



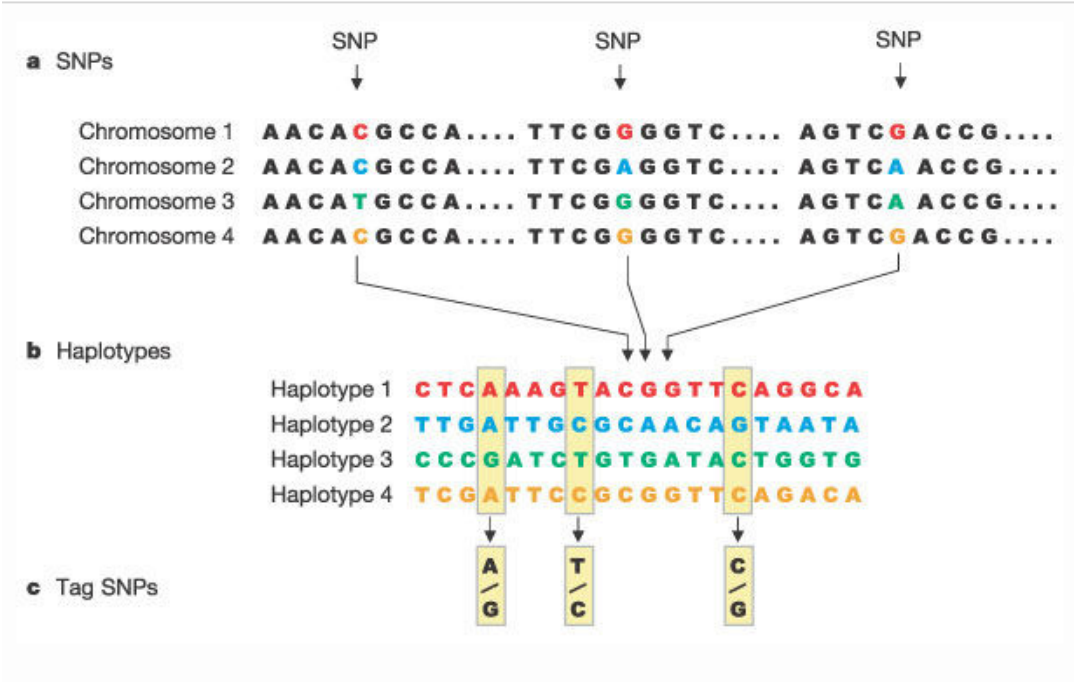
feature

The International HapMap Project

The International HapMap Consortium*

*Lists of participants and affiliations appear at the end of the paper

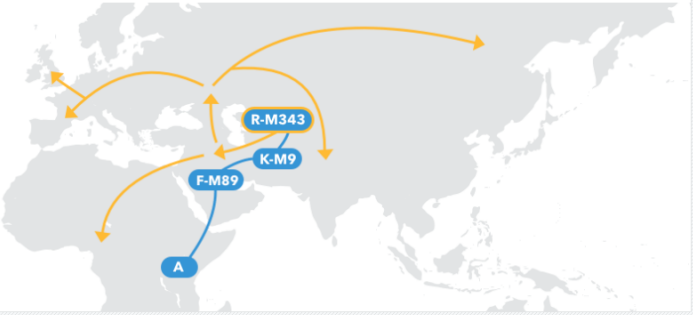
The goal of the International HapMap Project is to determine the common patterns of DNA sequence variation in the human genome and to make this information freely available in the public domain. An international consortium is developing a map of these patterns across the genome by determining the genotypes of one million or more sequence variants, their frequencies and the degree of association between them, in DNA samples from populations with ancestry from parts of Africa, Asia and Europe. The HapMap will allow the discovery of sequence variants that affect common disease, will facilitate development of diagnostic tools, and will enhance our ability to choose targets for therapeutic intervention.



Migrations of Your Paternal Line

Haplogroup R-M343

Your ancestral path forked off again between 20,000 and 25,000 years ago in western Asia, at the beginning of the last great peak of the Ice Age. Massive glaciers covered northern Eurasia, but farther south in the Iranian Plateau your ancestors flourished. When the Ice Age finally gave way to our warmer climate nearly 11,500 years ago, a new era of migrations from the Middle East began and eventually carried haplogroup R-M343 across three continents.



You share a paternal-line ancestor with Niall of the Nine Hostages.

The spread of haplogroup R-M269 in northern Ireland and Scotland was likely aided by men like Niall of the Nine Hostages. Perhaps more myth than man, Niall of the Nine Hostages is said to have been a King of Tara in northwestern Ireland

One of those branches is haplogroup C, which traces back to a woman who lived in Central Asia nearly 24,000 years ago. Today C most common in Siberia and Central Asia, but members are also found in China and as far west as the Ural e branches C1b, C1c, C1d, and C4c migrat

GENOME-WIDE ASSOCIATION STUDIES

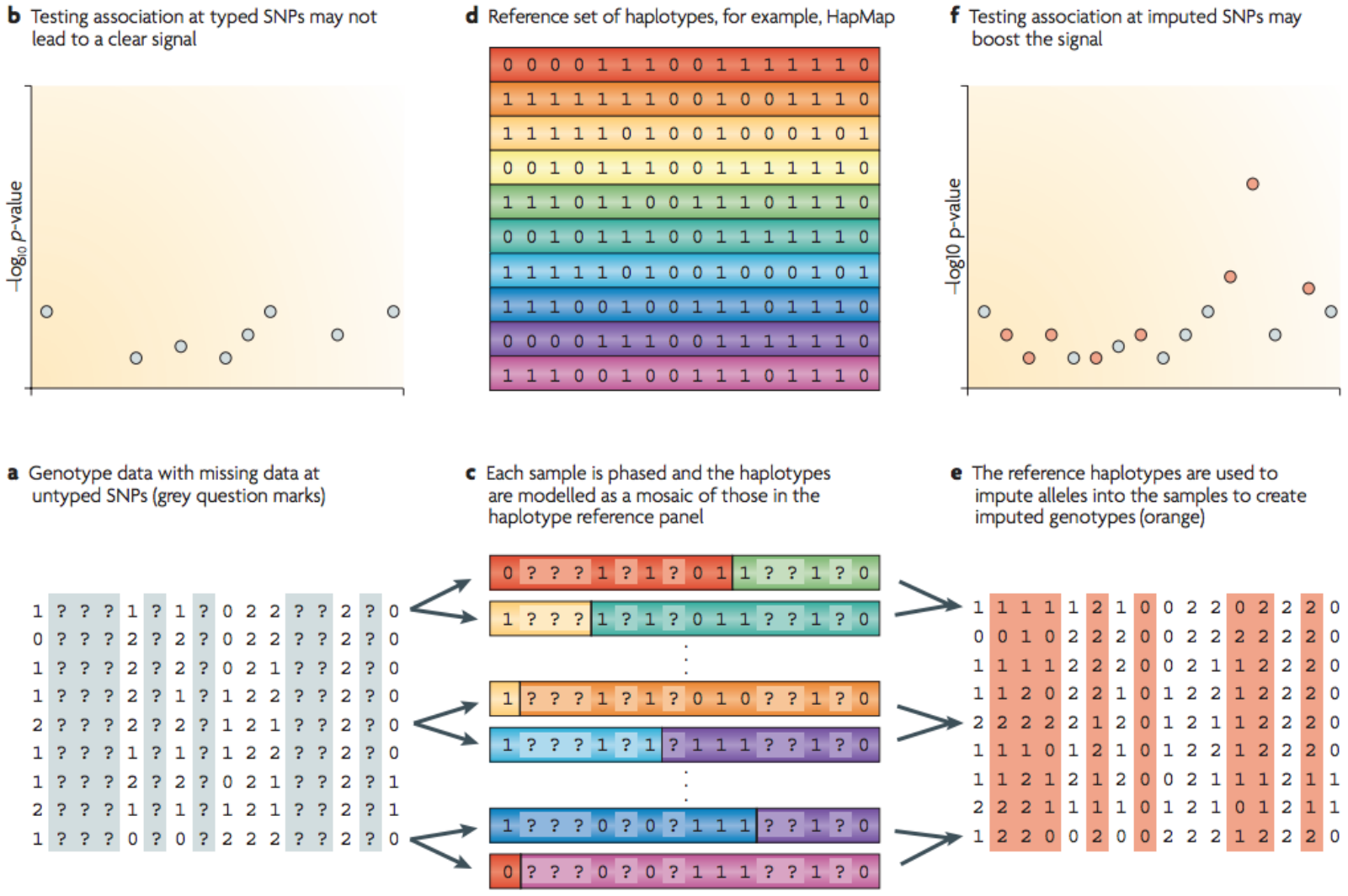
Genotype imputation for genome-wide association studies

Jonathan Marchini and Bryan Howie‡*

Abstract | In the past few years genome-wide association (GWA) studies have uncovered a large number of convincingly replicated associations for many complex human diseases. Genotype imputation has been used widely in the analysis of GWA studies to boost power, fine-map associations and facilitate the combination of results across studies using meta-analysis. This Review describes the details of several different statistical methods for imputing genotypes, illustrates and discusses the factors that influence imputation performance, and reviews methods that can be used to assess imputation performance and test association at imputed SNPs.

Marchini J, Howie B. Genotype imputation for genome-wide association studies. Nat Rev Genet. 2010 11:499-511.

Box 1 | **How genotype imputation works**



Several methods for genotype imputation

- Two main categories:
  - Family based
  - Population based
  - Or combination of the two

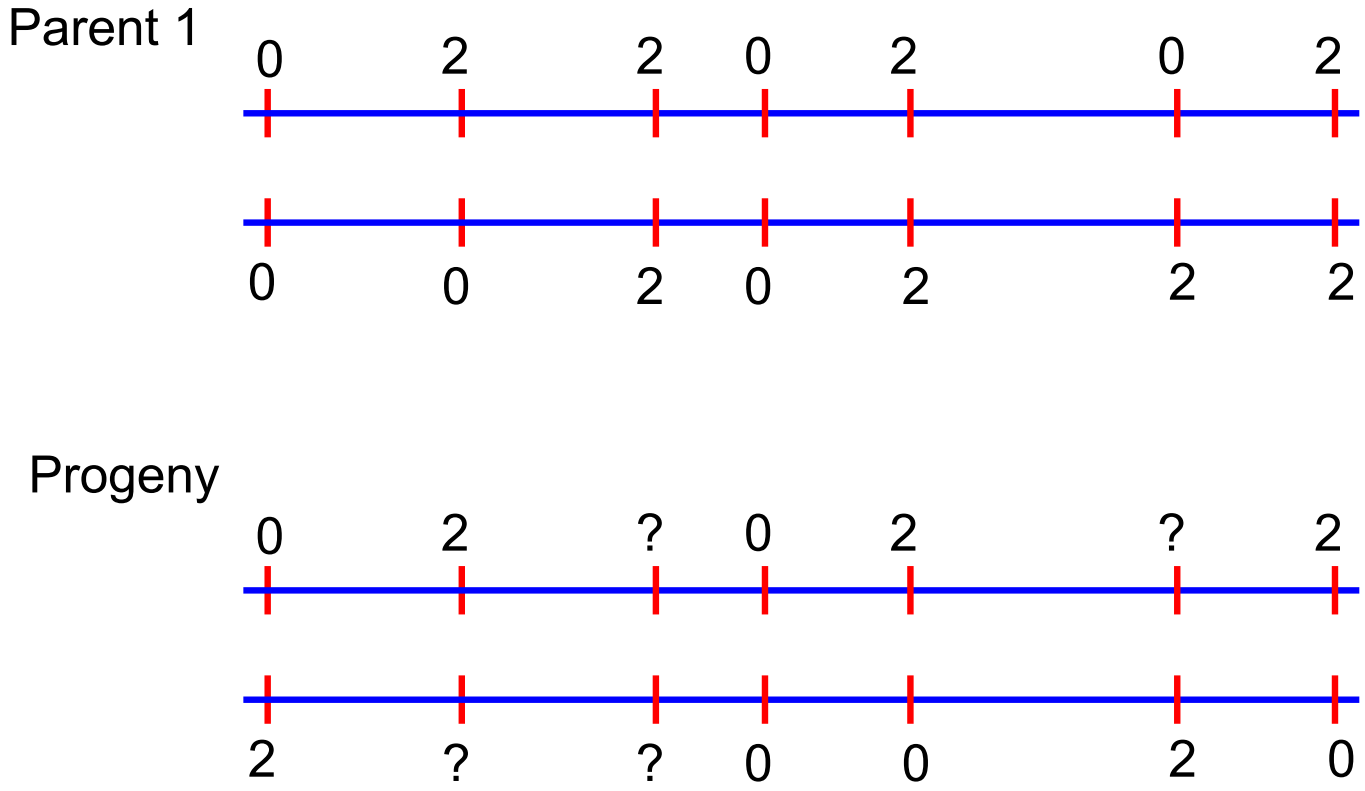
  - Some of the most effective are Beagle (Browning and Browning, 2009), MACH (Li et al., 2010), Impute2 (Howie et al., 2009), AlphaPhase (Hickey et al 2011)

Several methods for genotype imputation

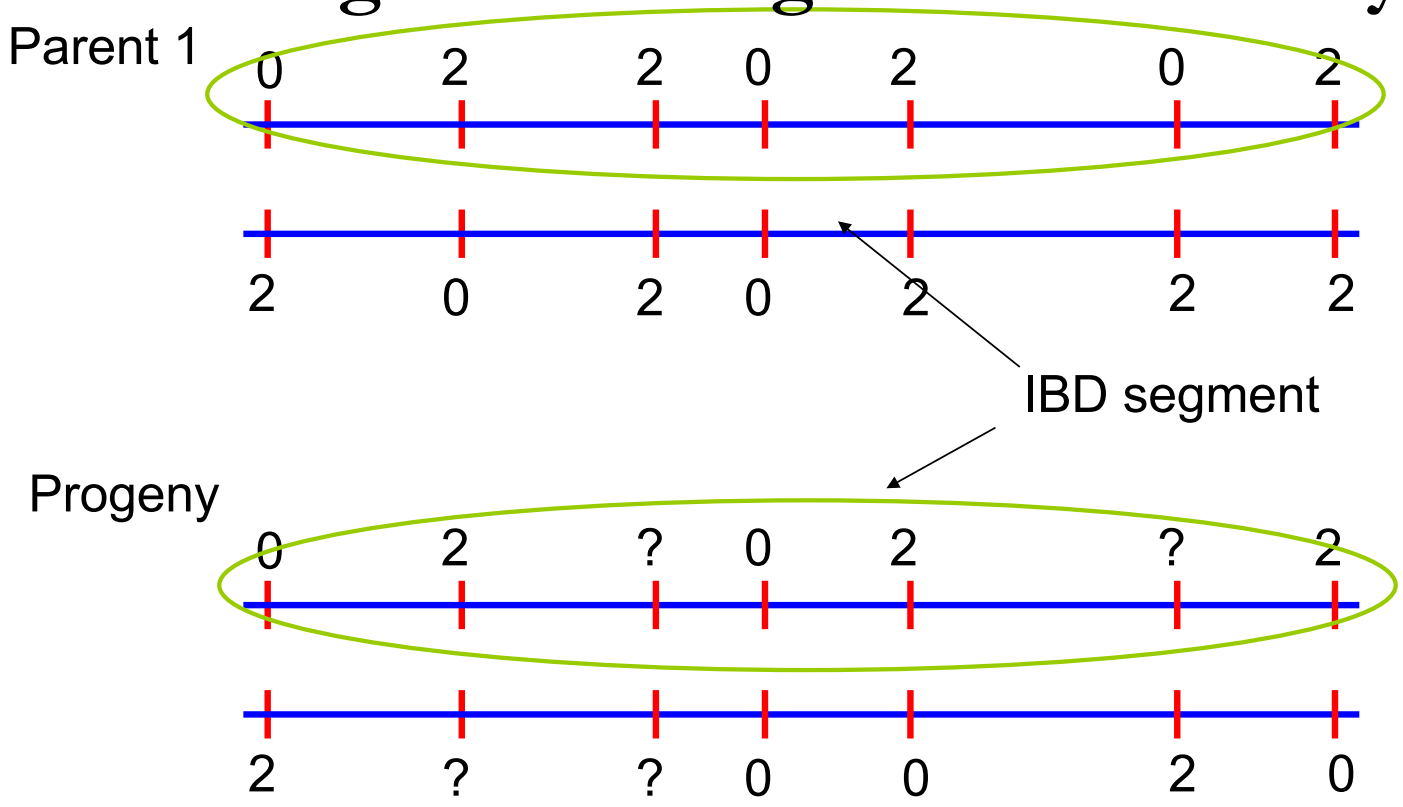
- Two main categories:
  - Family based
  - Population based
  - Or combination of the two

- Some of the most effective are Beagle (Browning and Browning, 2009), MACH (Li et al., 2010), Impute2 (Howie et al., 2009), AlphaPhase (Hickey et al 2011)

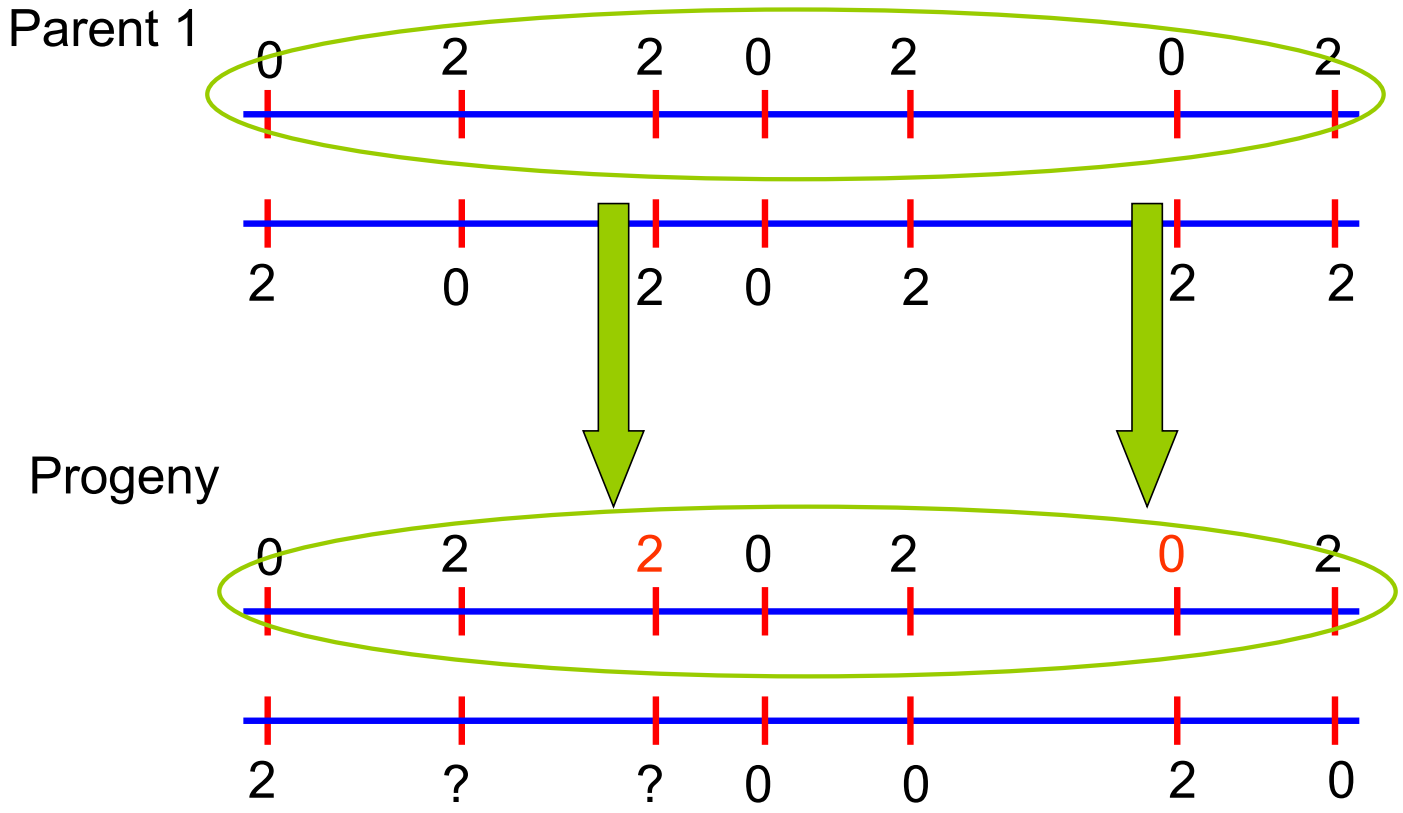
Finding an IBD segment in a family



Finding an IBD segment in a family



Finding an IBD segment in a familyxd



Several methods for genotype imputation

- Two main categories:
  - Family based
  - Population based (exploits LD)
  - Or combination of the two

  - Some of the most effective are Beagle (Browning and Browning, 2009), MACH (Li et al., 2010), Impute2 (Howie et al., 2009), AlphaPhase (Hickey et al 2011)

Population-based imputation

- Hidden Markov Models
  - Has “hidden states”
  - For target individuals these are “maps” of reference haplotypes that have been inherited
  - Imputation problem is to derive genotype probabilities given hidden states, sparse genotypes, recombination rates, other population parameters

$$P(Gi|H,\theta,\rho) = \sum_{s} P(Gi|S,\theta)P(S|H,\rho)$$

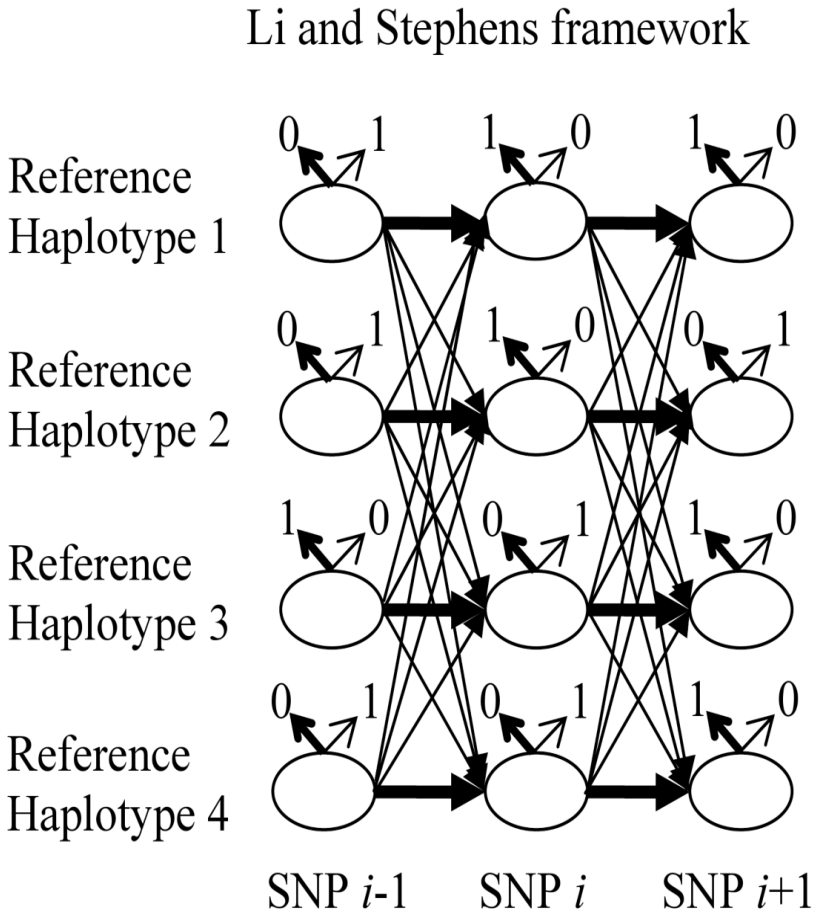
Population-based imputationx

- Consider three markers, 4 reference haplotypes

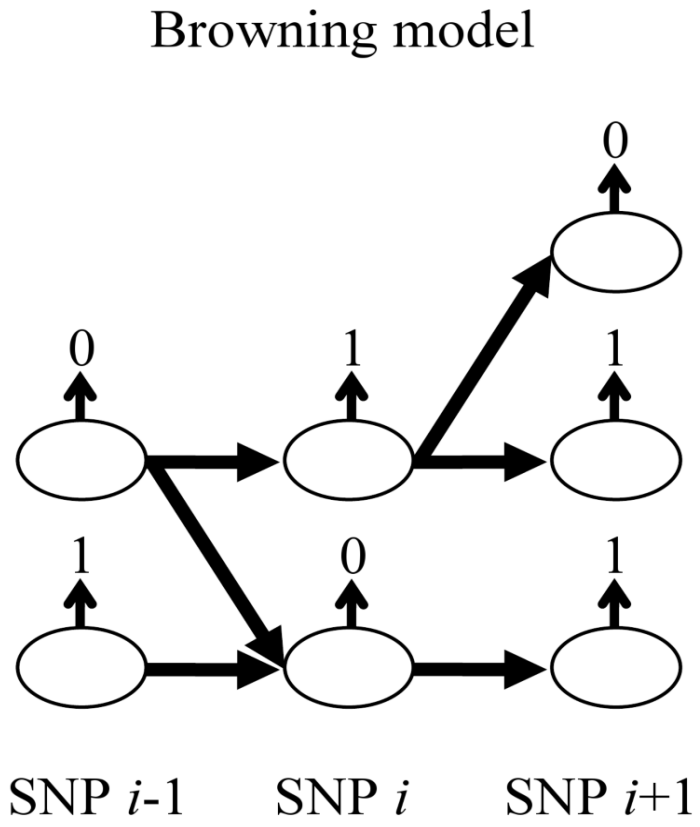
- 0 1 1
- 0 1 0
- 1 0 1
- 0 0 1

- Imputation?

Li and Stephens (MACH)



Browning (BEAGLE)

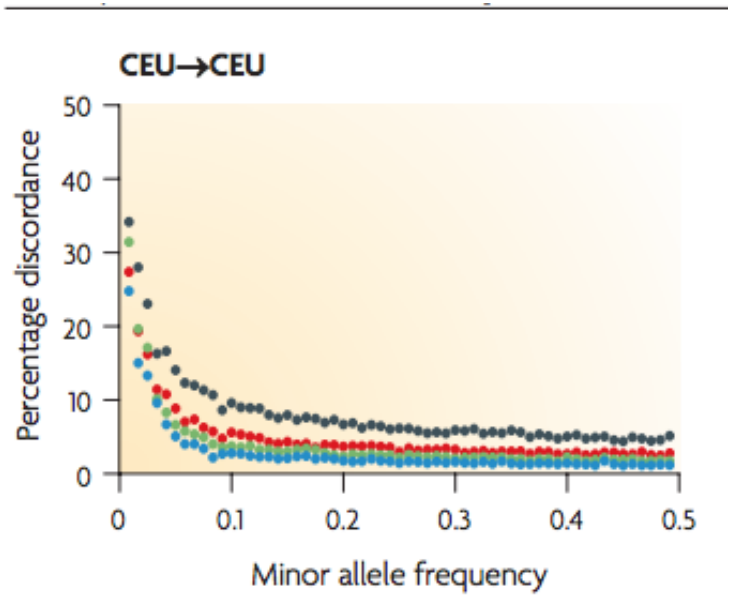


Imputation Accuracy

- Accuracy = correlation of real and imputed genotypes
- Concordance = percentage (%) of genotypes called correctly

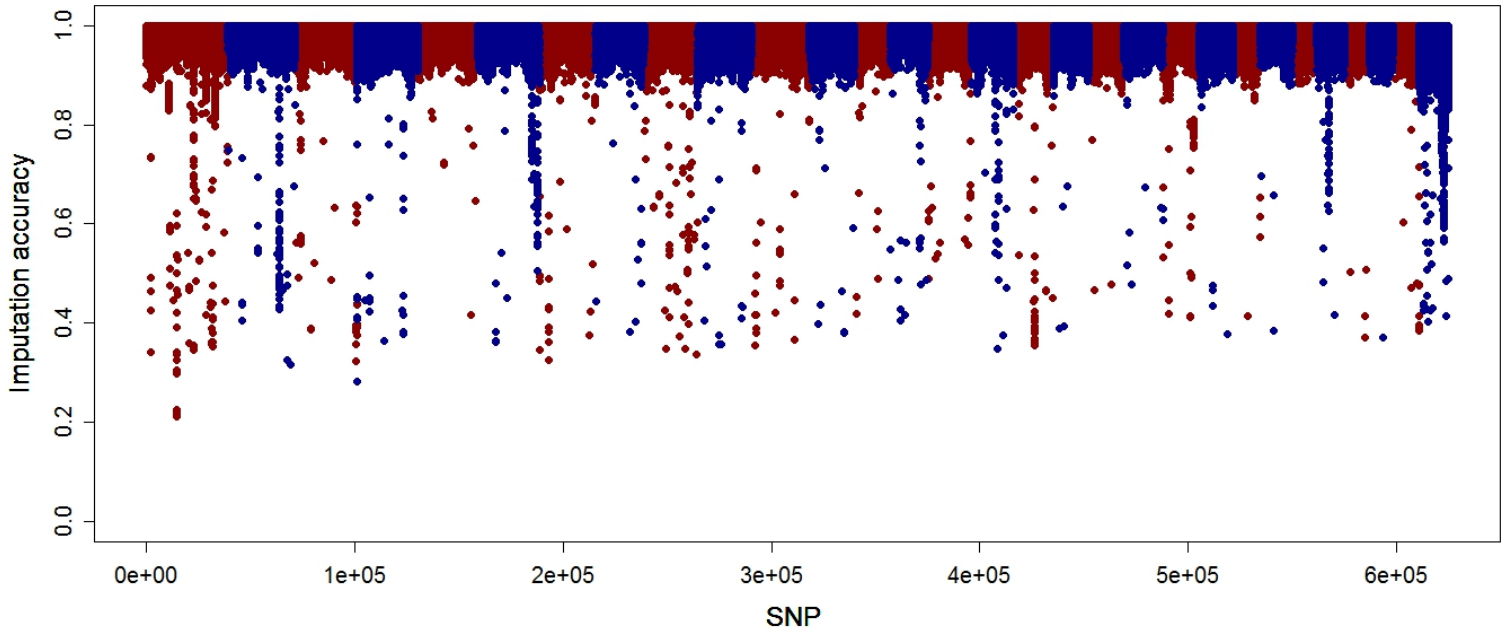
Imputation Accuracy

- Depends on
  - Size of reference set
    - bigger the better!
  - Density of markers
    - extent of LD, effective population si
  - Frequency of SNP alleles
  - Genetic relationship to reference



Imputation Accuracy

- Effect of map errors?



Imputation Accuracy

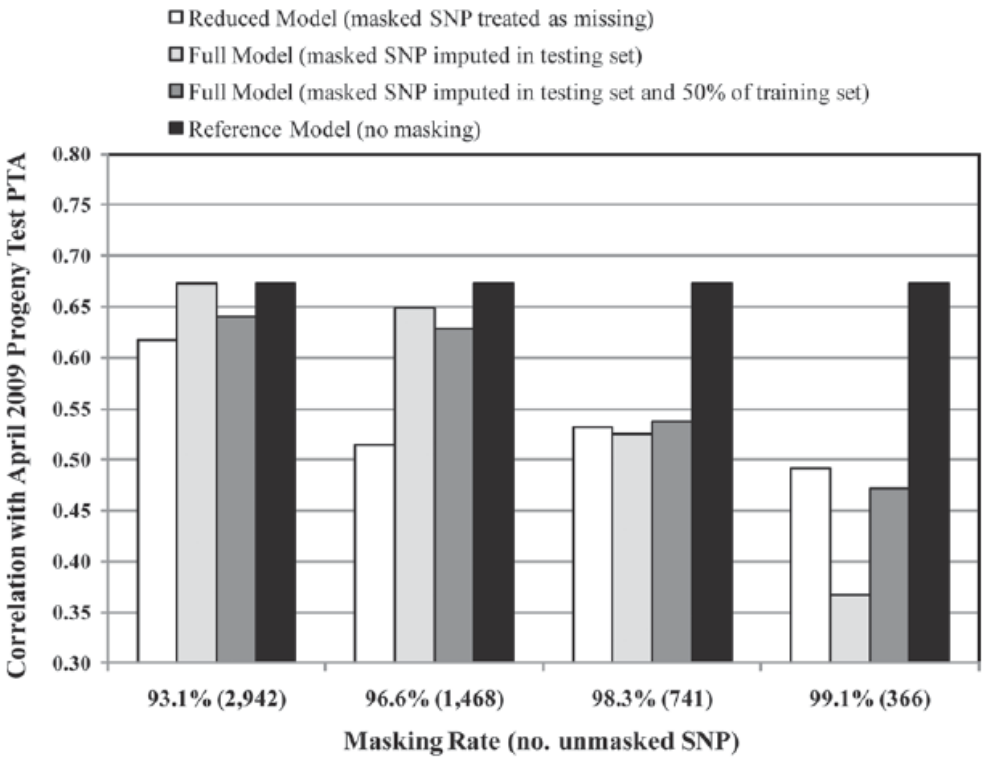


Figure 2. Correlations between predicted direct genomic values for milk yield and corresponding April 2009 progeny-test PTA using full or reduced models with 42,552 or 366, 741, 1,468, or 2,942 single SNP covariates, respectively, with or without imputation of masked genotypes for bulls in the testing set or bulls in the testing set and a randomly chosen 50% of bulls in the training set. The bars denoted as "reference" correspond to correlations from a full model in which all 42,552 SNP genotypes were left as unmasked in both the training and testing sets.