Haplotype Analysis and Genotype Imputation

Xiaowei Zhan
April 28, 2017
Xiaowei.Zhan@UTSouthwestern.edu

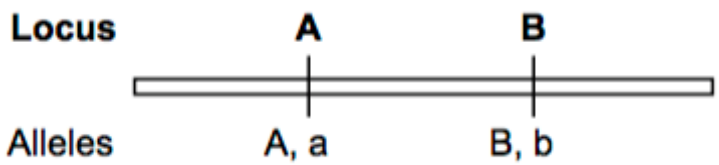
(Slides courtesy of Goncalo Abecasis)

Concepts from yesterday's lecture

- Haplotype and Linkage

Linkage Disequilibrium (LD):

Genotypes at two loci are not independent



Under Linkage Didsequilibrium

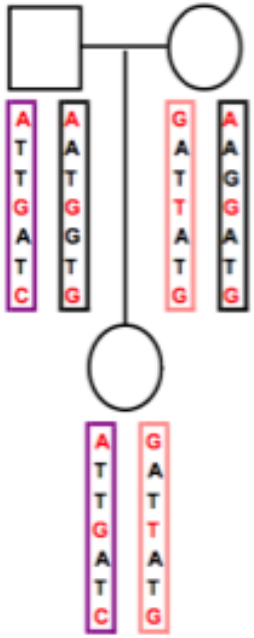
$Pr(AB) \neq Pr(A)\ Pr(B)$

Concepts from yesterday's lecture

- Haplotype

  Related individuals can share long stretches of sequences (haplotype)

  from last lecture under the context of *Identity by Descent*



Today's Outline: Haplotype and Imputation Analysis

- Haplotype analysis
  - Understand the motivation of haplotype analysis
  - Statistical method to infer haplotypes from genotype data
    - Clark's Greedy algorithm
    - E-M algorithm
    - Hidden Markov Model (HMM) ← Lab
  - Haplotype association analysis
- Imputation analysis
  - Understand the concept of imputation
  - How to impute genotypes from familial samples
  - How to imputation genotypes from unrelated individuals
    - Hidden Markov Model (HMM) ← Lab
  - How to use imputed genotypes in association analysis
  - Examples of imputation analyses in GWAS studies

How haplotype analysis can be useful?

Assume we know the haplotype-level genetic data,
how can haplotypes be useful?

- Linkage disequilibrium studies
  (Recall how to calculate D, D' and r^2)
  Genetic variations in the haplotype level.
  e.g. population specific haplotype
- Select markers to genotype
  Select tag SNP based on haplotypes
- Candidate gene studies
  Interpret association results
  Capture the effect of ungenotyped alleles

Haplotype cannot be easily observed

- Biological measurement of haplotypes can be challenging

  X-chromosome in males
  Sperm typing
  Hybrid cell lines
  Other molecular techniques
  …

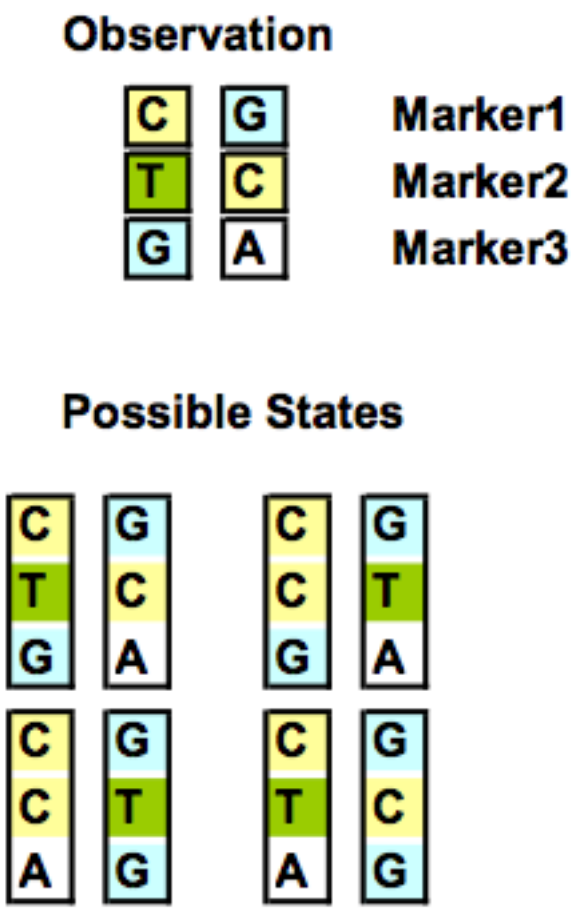
  We only observe genotype data, how to obtain haplotypes?

- *Statistical* approaches to infer haplotypes from genotypes

Observed genotypes ⇔ possible haplotypes

- Two alleles for each individual
  (Genotypes observed)

- Multiple haplotypes are compatible with observed genotype
  (4 haplotype combinations)

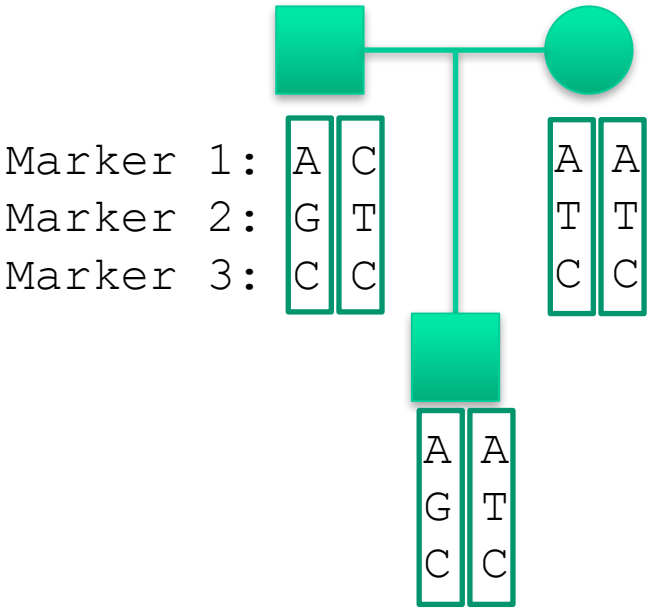


Family information can be helpful

- From pedigree, we can phase many markers

- But still, there can be many ambiguities that cannot be resolved

  Large number of markers => less proportion of known haplotypes



When there are no relatives…

- Rely on linkage disequilibrium
- Assume the number of haplotypes in a population is small
- Haplotypes tend to be similar

Phasing algorithms

Several milestone methods to infer haplotypes

1. Clark's greedy algorithm (1990, Mol Biol Evol, cited by 940, PMID 2108305)
2. E-M algorithm (1995, Mol Biol Evol, cited by 2051, PMID 7476138)
3. Stephen's model (2001, AJHG, cited by 6510, PMC1275651)

Inference of Haplotypes from PCR-amplified Samples of Diploid Populations[1]

Andrew G. Clark
Department of Biology and Genetics Program, Pennsylvania State University

Maximum-Likelihood Estimation of Molecular Haplotype Frequencies in a Diploid Population

Laurent Excoffier and Montgomery Slatkin†*
*Departments of Anthropology and Ecology, University of Geneva and †Department of Integrative Biology, University of California, Berkeley

Am. J. Hum. Genet. 68:978–989, 2001

A New Statistical Method for Haplotype Reconstruction from Population Data

Matthew Stephens,[1,3] Nicholas J. Smith,[2] and Peter Donnelly[1]
Departments of [1]Statistics and [2]Biochemistry, University of Oxford, Oxford; and [3]Department of Statistics, University of Washington, Seattle

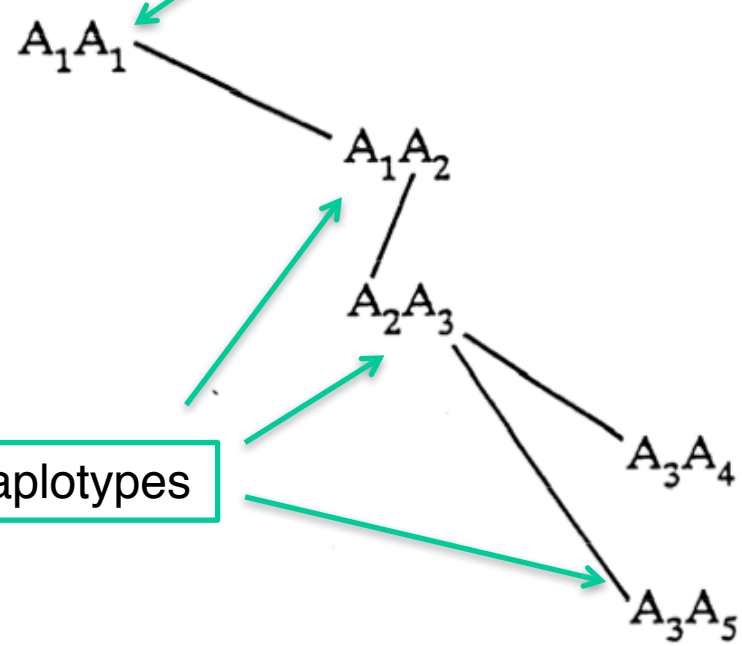
Clark's haplotyping method

Inference of Haplotypes from PCR-amplified Samples of Diploid Populations[1]

Andrew G. Clark
Department of Biology and Genetics Program, Pennsylvania State University

- Very computationally efficient
- Widely use in the 1990's
- Clark's Algorithm

Start from unambiguous haplotypes



Sequentially resolve haplotypes

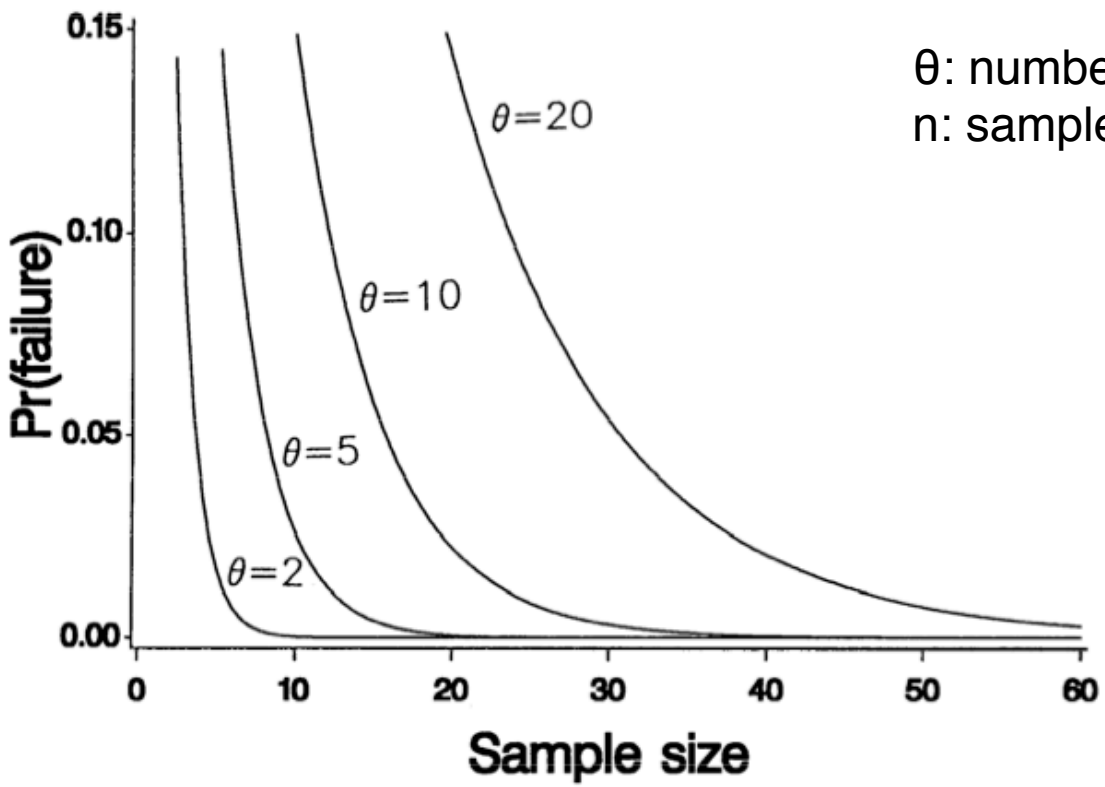
Randomly guess the rest haplotypes → A_6A_7

Limitation: failed to start

- What kind of genotype/haplotype do we need to have to get started?

- What is the probability of failed start?

$$\Pr(\text{failure}) \approx \left[1 - \frac{1}{1+\theta} - \frac{\theta}{(1+\theta)^2}\right]^n$$



θ: number of marker
n: sample size

Pro and Cons

- Andrew Clark's method

  Very fast

  May failed to start with small sample size

  May leave unresolved haplotypes

E-M method

- Excoffier and Slatkin (1995) *Mol Biol Evol* **12:**921-927
- E-M (expectation maximization)

Capable to handle missing genotypes

Consider allele frequencies

When there is m unphased genotypes, there are 2^{m-1} possible haplotypes => computationally expensive (>25 markers)

Maximum-Likelihood Estimation of Molecular Haplotype Frequencies in a Diploid Population

Laurent Excoffier and Montgomery Slatkin†*

*Departments of Anthropology and Ecology, University of Geneva and †Department of Integrative Biology, University of California, Berkeley

Stephen's method

- Stephens et al. (2001) *Am J Hum Genet* **68**:978-89
- Improve the previous EM method by reusing similar haplotypes
- Consider ***genealogical*** information

A New Statistical Method for Haplotype Reconstruction from Population Data

Matthew Stephens,[1,3] Nicholas J. Smith,[2] and Peter Donnelly[1]

Departments of [1]Statistics and [2]Biochemistry, University of Oxford, Oxford; and [3]Department of Statistics, University of Washington, Seattle

Reuse similar haplotypes

- Individual 1: use known haplotypes
- Individual 2: re-use known haplotypes and allow mismatches

Known haplotypes:

22544
22544
22544
22544
33334
33334
23233
14234

Ambiguous individual 1:

Genotype
32344
23534

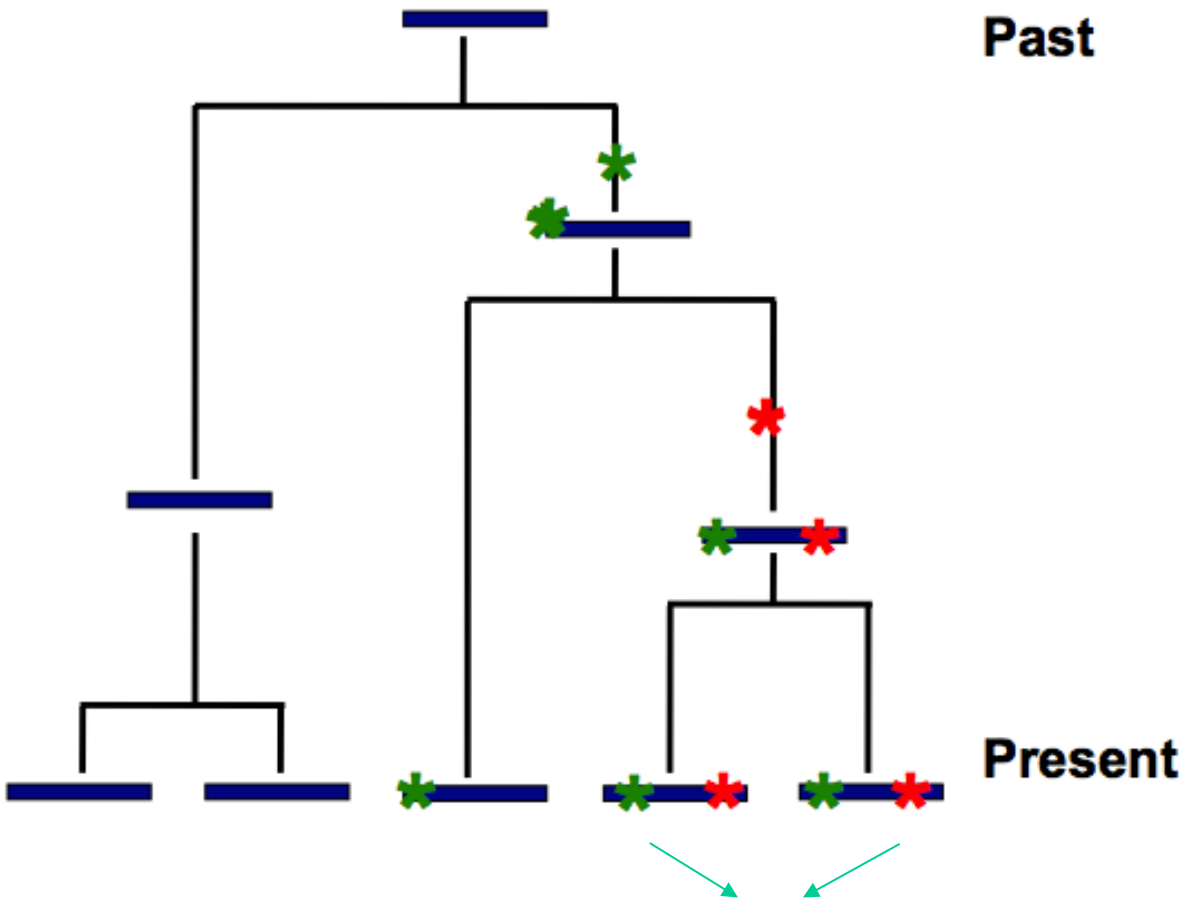
→ 33334
→ 22544

Ambiguous individual 2:

Genotype
32444
23434

→ 33434
→ 22444

Genealogical tree



Similar haplotypes have more recent common ancestor

MCMC method with Gibbs sampler

- MCMC method can iteratively improve solutions

  1. Initialize haplotypes
  2. Sample haplotypes of one individual given other's haplotypes
  3. Update the estimated haplotypes for one individual
  4. Repeat the above millions of times

- MCMC method will converge to an optimal solution
- The result is equivalent to EM algorithm

Stephen's algorithm

- Improve the update step by incorporate genealogical information (coalescent theory)

$$\Pr(h \mid H) = \sum_{\alpha} \sum_{S} \frac{n_\alpha}{n} \left(\frac{\theta}{n+\theta} \right)^S \frac{n}{n+\theta} \left(P^S \right)_{\alpha h}$$

Sum over haplotypes

Sum over number of mutations

S mutations before coalescence

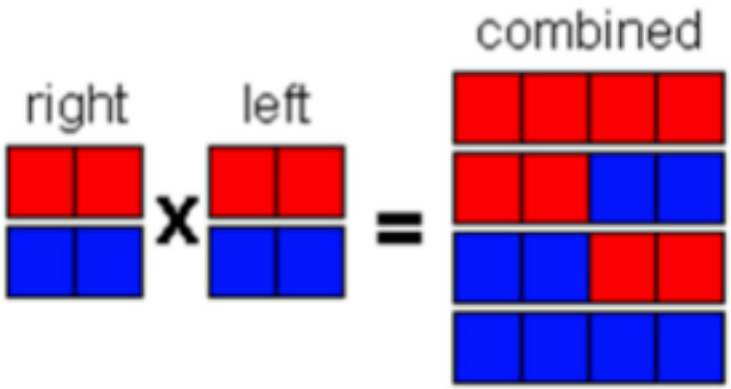
Coalescence

Mutation Matrix

ShapeIT/MaCH software

- Based on Stephen's model, modern phasing software optimizes computational efficiency

Blockwise computation (ShapeIT)



Hidden Markov Model (MaCH)

Markov haplotying

Same model can be easily adapted for imputation

ShapeIT: Delaneau, Olivier, Jonathan Marchini, and Jean-François Zagury. "A linear complexity phasing method for thousands of genomes." *Nature methods* 9.2 (2012): 179-181.(cited by 467)

MaCH: Li, Yun, et al. "MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes." *Genetic epidemiology* 34.8 (2010): 816-834. (cited by 1373)

Summary on haplotype inference

- Three classic statistical approaches to infer haplotypes from genotype data

  Clark's greedy algorithm
  E-M algorithm
  Stephen's genealogical approach

- Lab: practical workshop

  Phase one sample from HapMap3 project using ShapeIT

Association models for haplotype analysis

- Association tests

  are haplotype frequencies the same in two populations

  e.g. case vs. control, population 1 vs. population 2

- The simplistic approach to compare haplotypes reconstructions

  Calculate haplotype frequency for each group

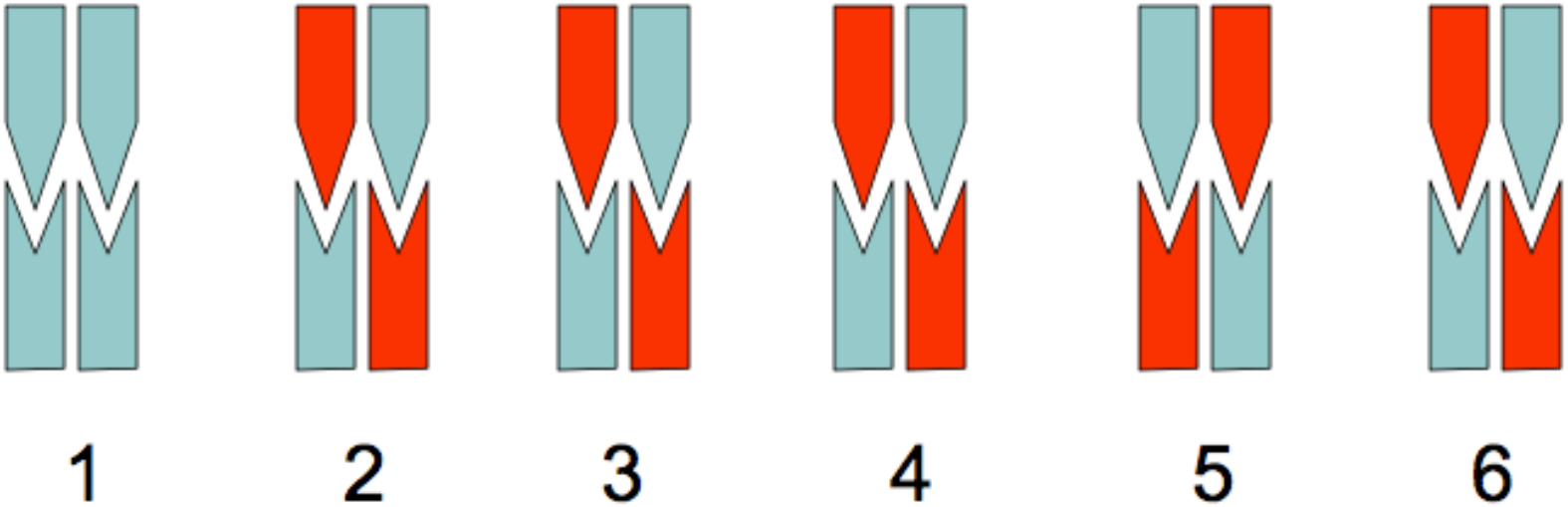
  Find mostly likely haplotype for each individual

  Compare haplotype frequency between the two groups

  → **NOT RECOMMENDED!!!**

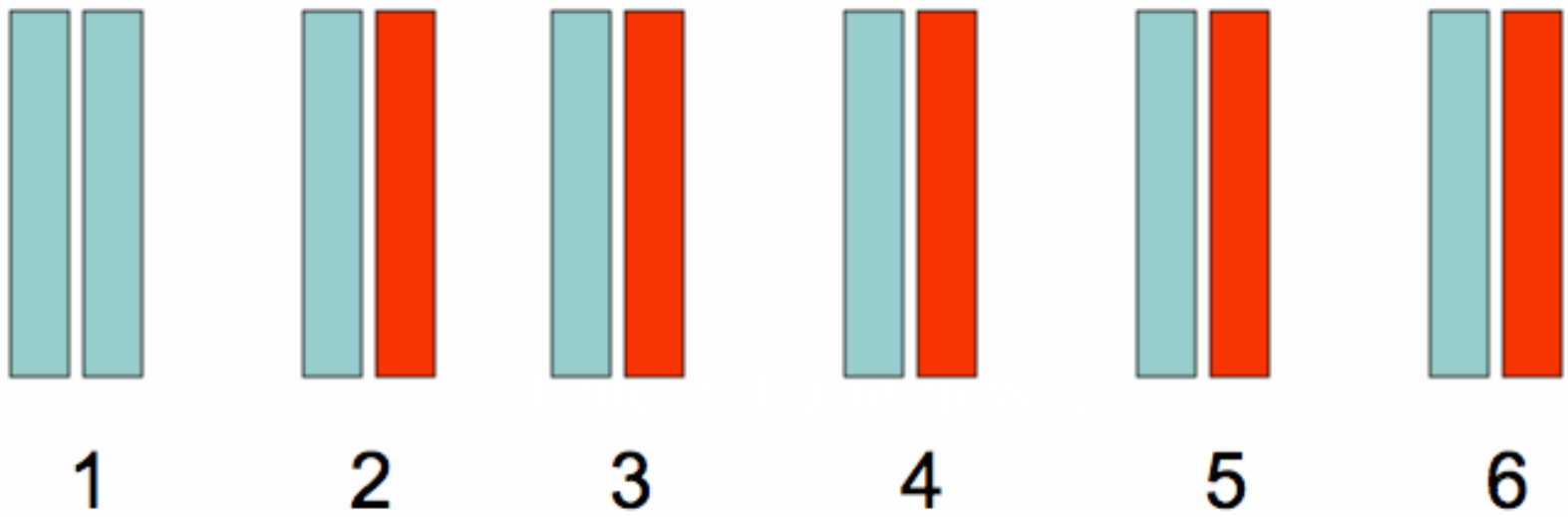
  Question: any caveat in this approach?

Observe genotypes in CASE



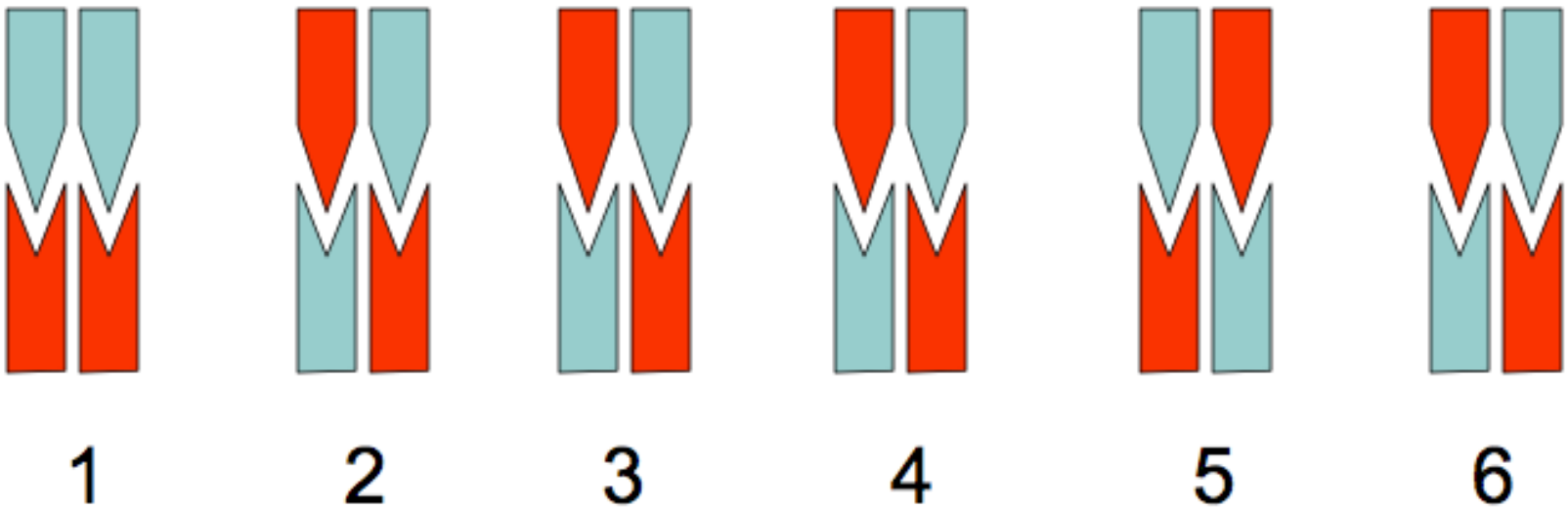
The phase reconstruction in the five ambiguous individuals will be driven by the haplotypes observed in individual 1

Inferred haplotypes for CASE



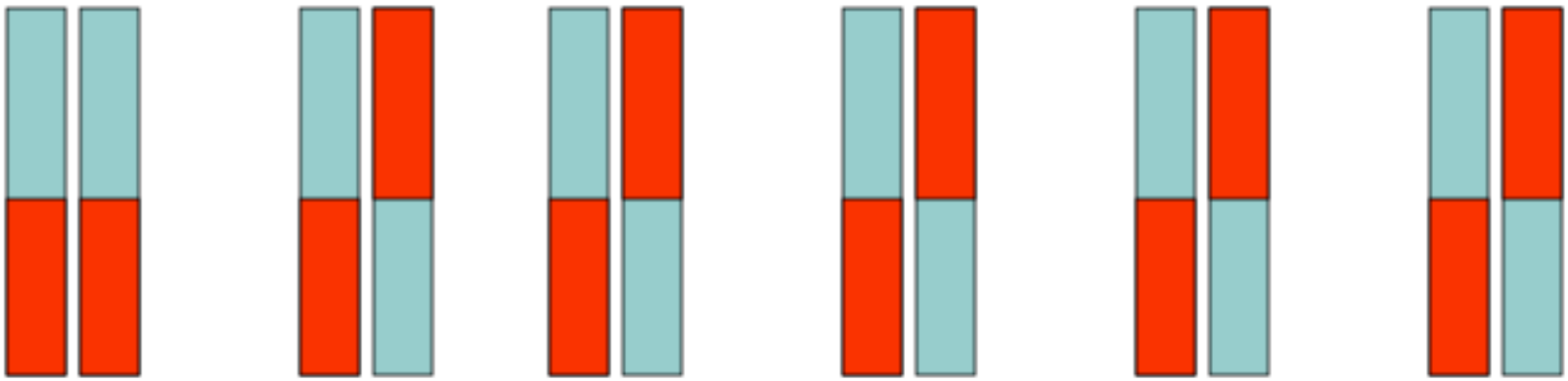
This kind of phenomenon will occur with nearly all population based haplotyping methods!

Observe genotypes for CONTROL



Note these are identical, except for the single homozygous individual

Inferred haplotypes for CONTRL

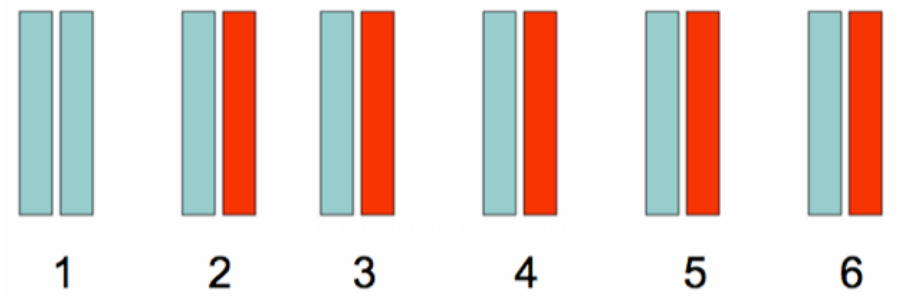


1 2 3 4 5 6

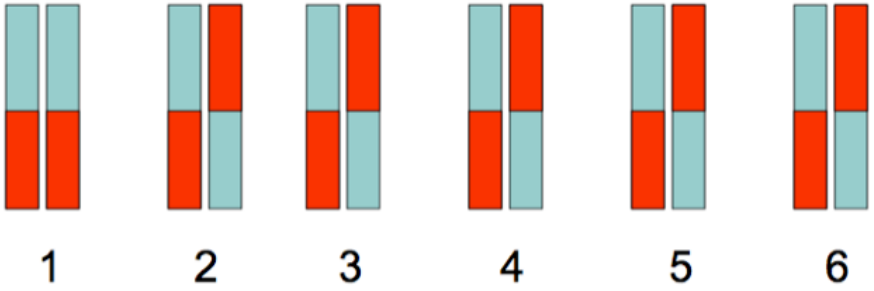
Ooops The difference in a single genotype in the original data has been greatly amplified by estimating haplotypes

Inferred haplotypes for CASE and CONTRL

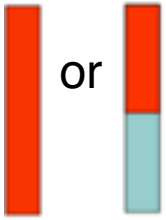
Case



Control



Problematic to conclude haplotype frequencies differs between groups,

although  or

frequencies differ.

Lesson learned

- Do NOT treat case/control in different haplotype inference procedures
- Treat case/control **together**
- Or use likelihood approaches

- Estimated haplotype frequencies, imply a likelihood for the observed genotypes

$$L = \prod_{i} \sum_{H \sim G_i} P(H)$$

individuals

possible haplotype pairs, conditional on genotype

haplotype pair frequency

Likelihood approaches

- Test two sets of models

- Calculate 3 likelihoods:
  - Maximum likelihood for combined sample, L_A
  - Maximum likelihood for control sample, L_B
  - Maximum likelihood for case sample, L_C

$$2\ln\left(\frac{L_B L_C}{L_A}\right) \sim \chi^2_{df}$$

df is hard to obtain
Use permutations

Hypothesis testing

- Previously: test haplotype frequency between two populations
- Often, we want to test

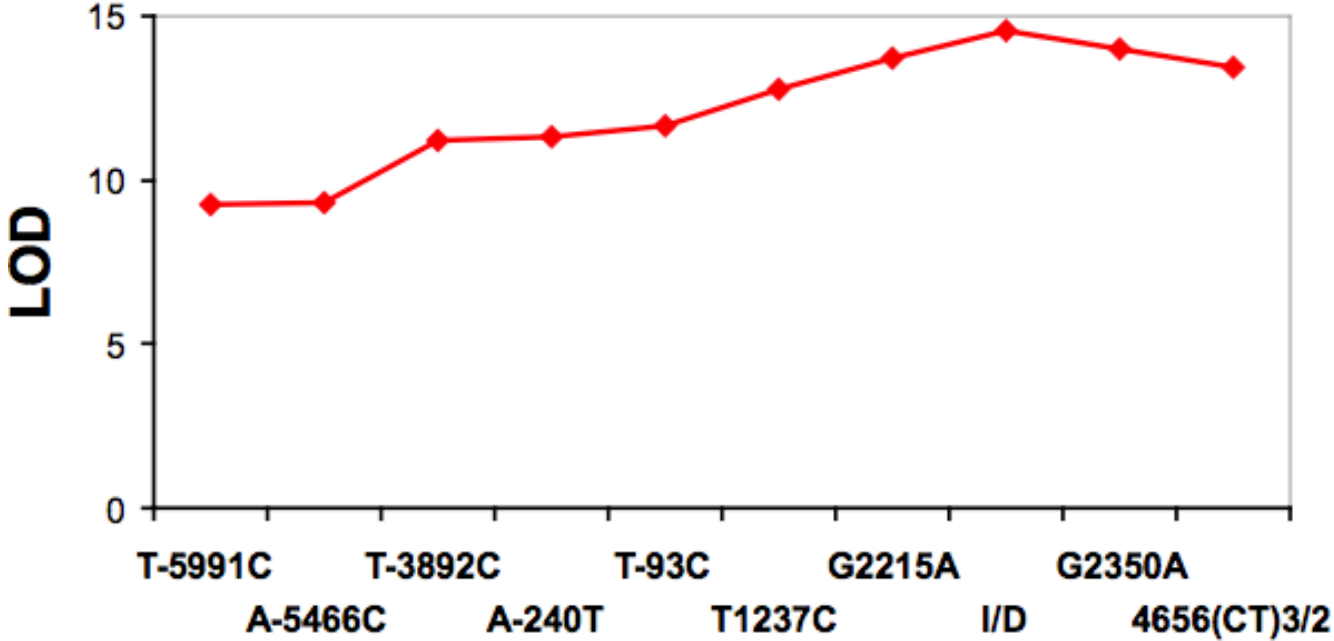
  Are haplotypes different between two populations?

- Note: this is different than single marker test

ACE gene example

- Keavney et al (1998), Hum Mol Genet **7**:1745-1751

- 10 di-allelic polymorphism
  Spanning 26k region
  Common markers



Haplotype analysis

- 3 ACE haplotype clades.
- Clade “B” = Clade “C”
  Similar phenotypic effect
- Interpretations

  Functional variants on the right

  Think: if functional variants on the left,
  which two clades have similar phenotypes?

Regression models

- Predictor

  Haplotype counts

- Regression parameters

  Phenotypic effect of each haplotype

- Response

  Phenotype values

Exemplar design matrix

μ	h_1	h_2	h_3

$$E\begin{Bmatrix} Y_1 \\ Y_2 \\ Y_3 \end{Bmatrix} = \begin{Bmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 1/2 & 1/2 \\ 1 & 1/2 & 0 & 1/2 \end{Bmatrix} \begin{Bmatrix} \mu \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{Bmatrix}$$

Hypothetical set-up when observed haplotypes are:

h_1/h_1 for individual 1
h_2/h_3 for individual 2
h_1/h_3 for individual 3

Zaykin et al, 2002

When haplotype is unknown

- Use Baye's rule to calculate:

$$\Pr(h_2, h_3 \mid G_i) = \frac{\Pr(G_i \mid h_2, h_3) p_{h_2} p_{h_3}}{\sum_{u,v} \Pr(G_i \mid h_u, h_v) p_{h_u} p_{h_v}}$$

- Use partial counts in the design matrix

Is haplotype test more powerful?

1 Marker

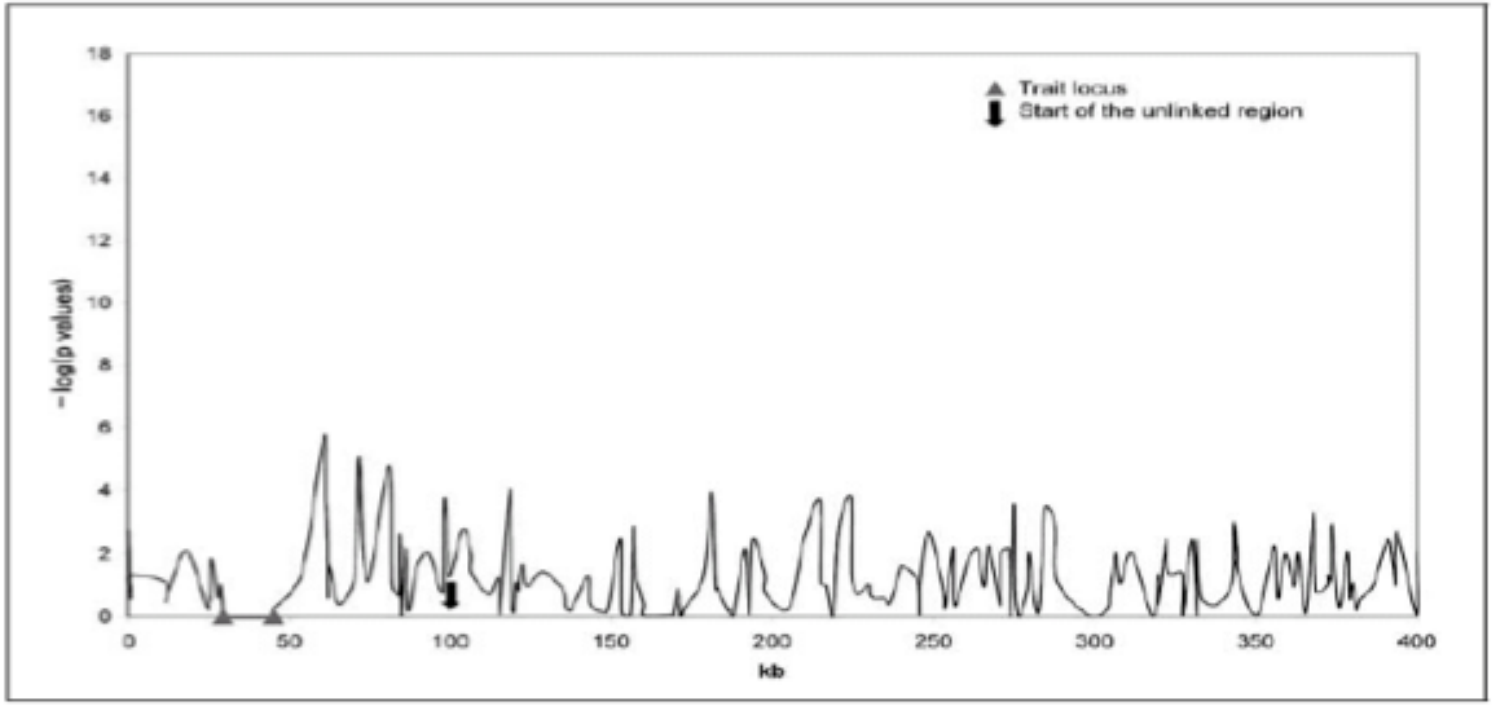


Fig. 1. Sample –log(p values) against the marker map plots for window size of 1 using p values from the asymptotic F test.

Zaykin et al, 2002

Is haplotype test more powerful?

3 Markers

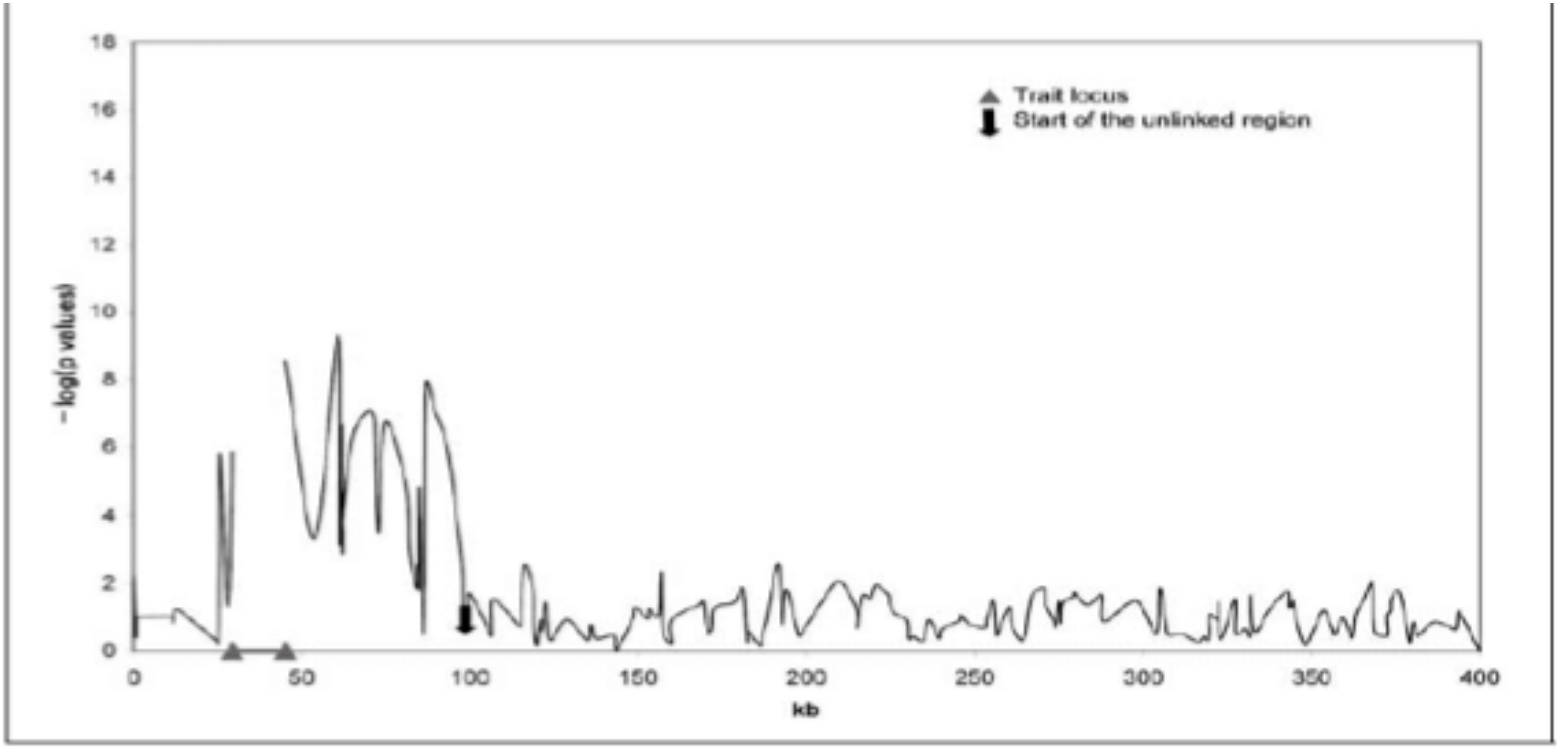


Fig. 2. Sample –log(p values) against the marker map plots for window size of 3 using p values from the asymptotic *F* test.

Zaykin et al. 2002

Is haplotype test more powerful?

5 Markers

Higher than 3-marker and 1-marker test

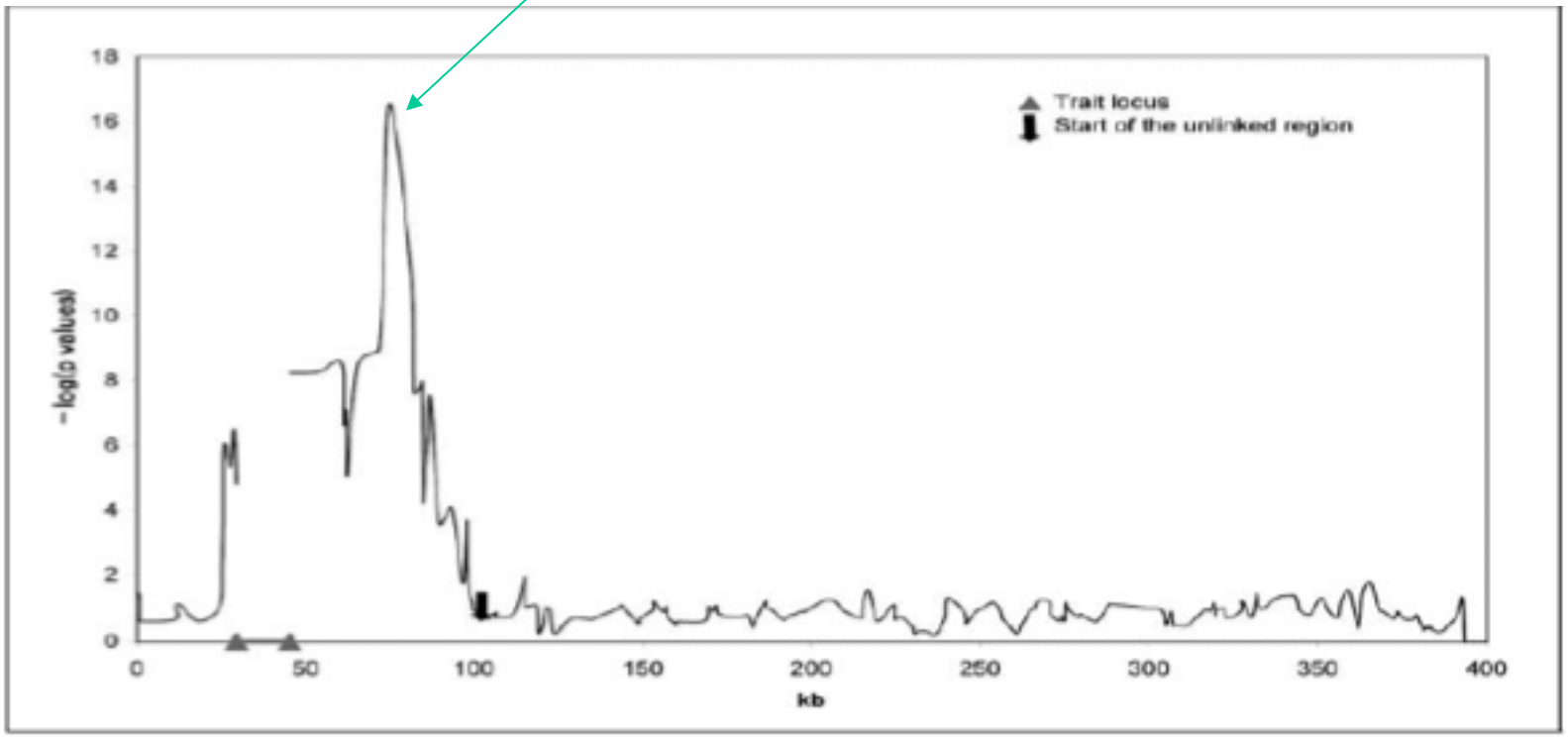


Fig. 3. Sample –log(p values) against the marker map plots for window sizes of 5 using p values from the asymptotic *F* test.

Zaykin et al. 2002

Summary for haplotype association tests

- Testing haplotypes can improve power
- Testing one haplotype is usually not enough
- The significance value need to be empirically evaluated

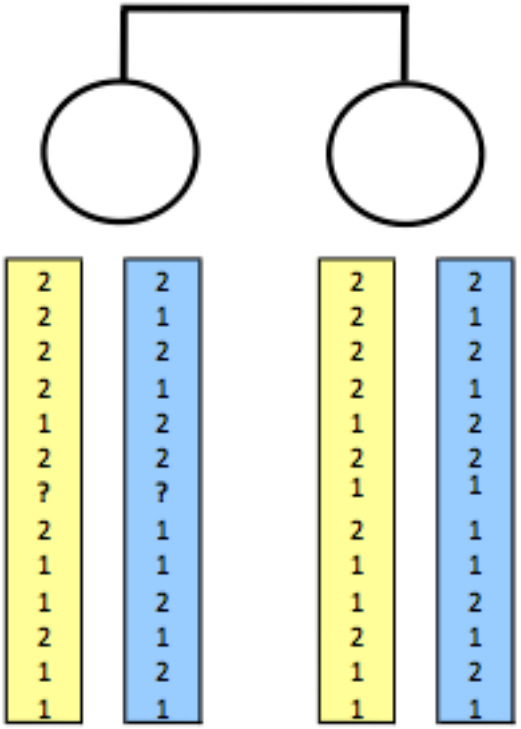
  e.g. permute case/control labels.

Imputation

- Genotype Imputation / “In silico” Genotyping
- Related individuals
  - Share segments of identity by descent

Intuition

- What is genotype of ?/?

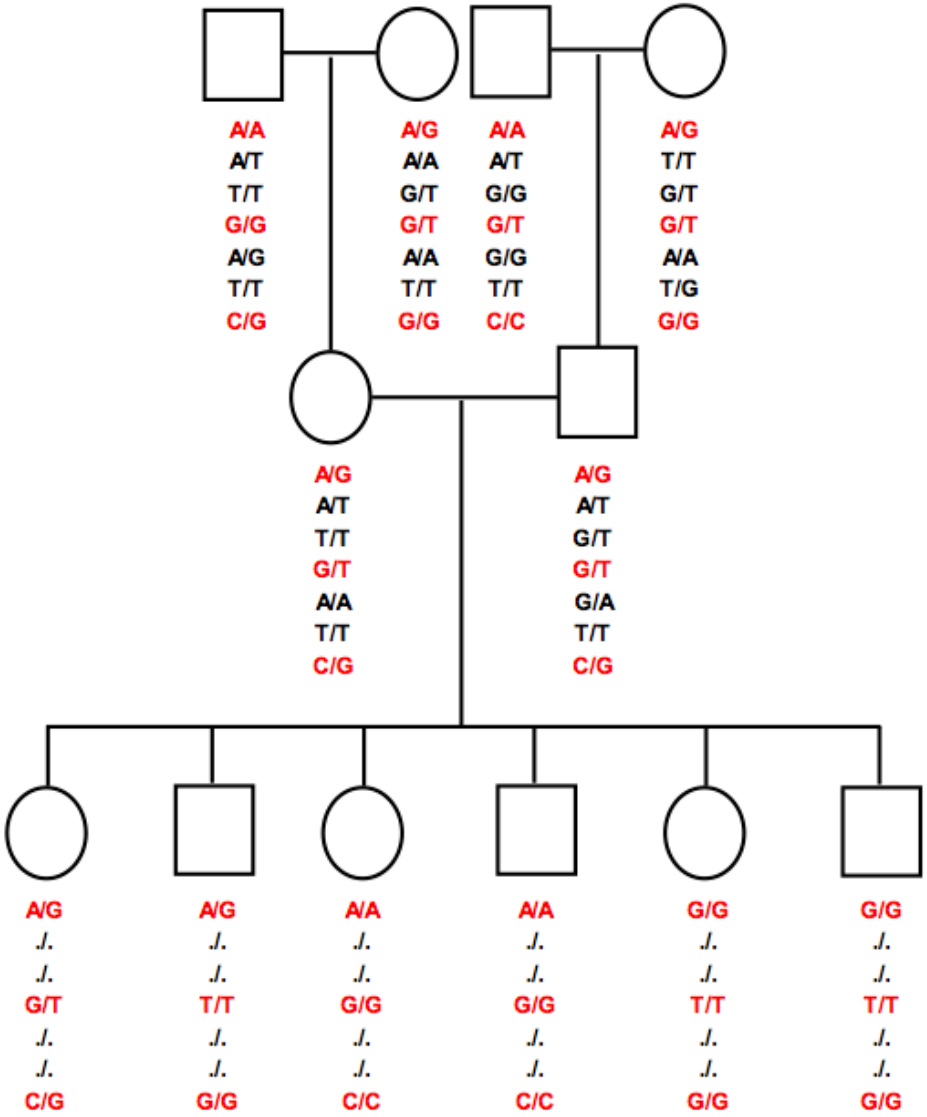


Imputation for family samples

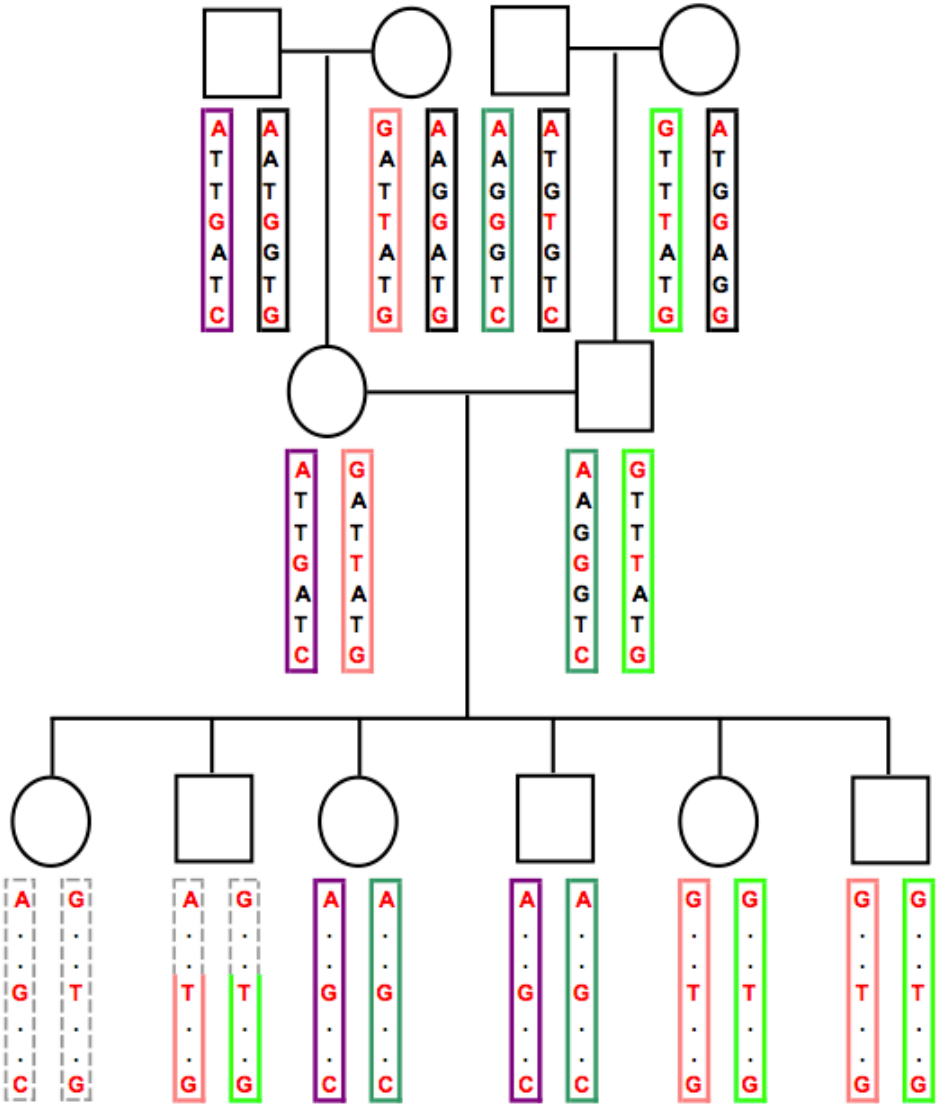
- Family samples share large segments of chromosomes
- Use this information, we can design a cost-effective way for genotyping

  1. Genotype a few markers for all samples
  2. Infer shared segments of haplotypes
  3. Genotype additional markers
  4. Fill in the missing genotypes

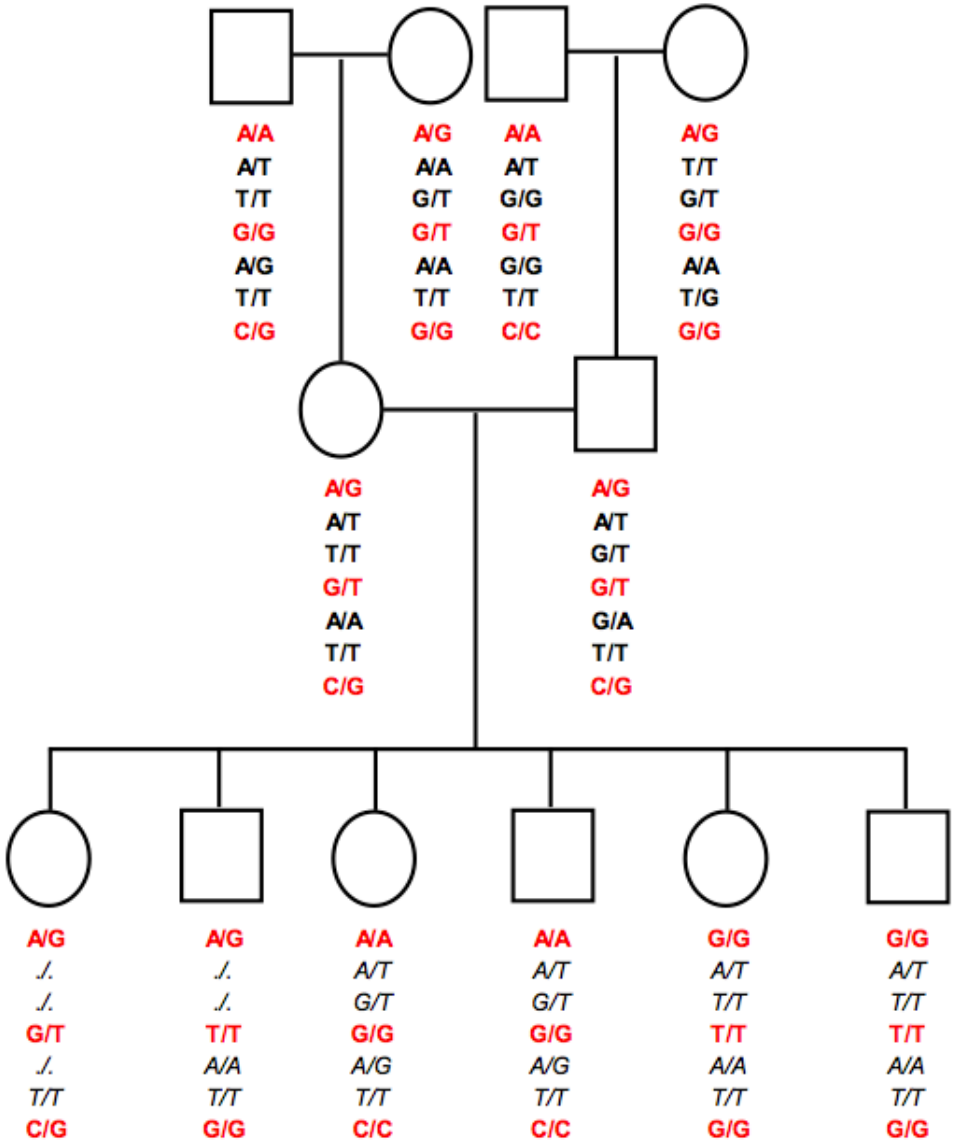
Imputation for family samples



Infer allele sharing



Impute missing genotype



Genotype imputation for family samples

- A particular genotype g_{ij} is missing (true value is 0, 1 or 2).
- Observed genotypes are G
- Using a pedigree likelihood (Lander-Green algorithm or Elston-Stewart algorithm), we can calculate

$$P(g_{ij} = 0, 1 \text{ or } 2 \mid G)$$

and

$$\bar{g}_{ij} = E(P(g_{ij} = 0, 1 \text{ or } 2 \mid G)) = 2*P(g_{ij} = 2 \mid G) + P(g_{ij} = 1 \mid G)$$

Association test of *observed* genotypes

- 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

Association test of *imputed* 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

Imputation for unrelated individuals

- Family samples share longer segments of chromosome
- Unrelated individual share much short segments
- It is still possible to infer stretches of sharing between unrelated individuals
- Then the study design of unrelated individuals can be similar to family samples

Study design

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 stretches

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
```

Impute missing genotypes

Observed Genotypes

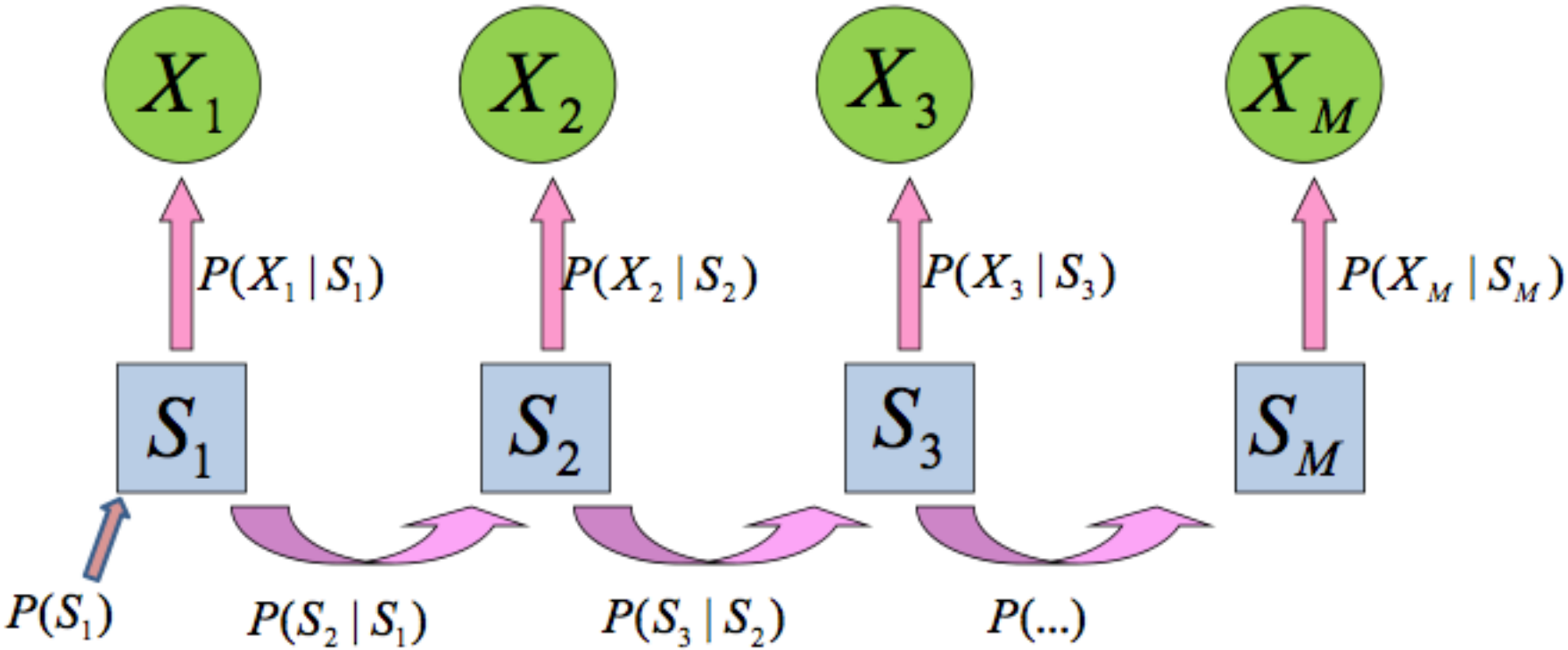
```
c g a g A t c t c c c g A c c t c A t g g
c g a a G c t c t t t t C t t t c A t g g
```

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

- Hidden Markov Model – a very useful but complex model
- Observe X (observed genotypes)
- Goal is to infer the hidden state, S (reference haplotypes)



Commonly used imputation software

1. MaCH – classic, highly accurate

   MaCH: Li, Yun, et al. "MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes." Genetic epidemiology 34.8 (2010): 816-834.

   http://csg.sph.umich.edu/abecasis/mach/download/

2. Minimac – efficient, good for sequence data

   Lab → Das, Sayantan, et al. "Next-generation genotype imputation service and methods." *Nature Genetics* 48.10 (2016): 1284-1287.

   http://genome.sph.umich.edu/wiki/Minimac3

3. Michigan Imputation Server - cloud-based imputation service, large panel of reference haplotypes

   McCarthy, Shane, et al. "A reference panel of 64,976 haplotypes for genotype imputation." *Nature genetics* (2016).

   https://imputationserver.sph.umich.edu/index.html

4. IMPUTE/IMPUTE2 – classic, similar to MaCH, Minimac

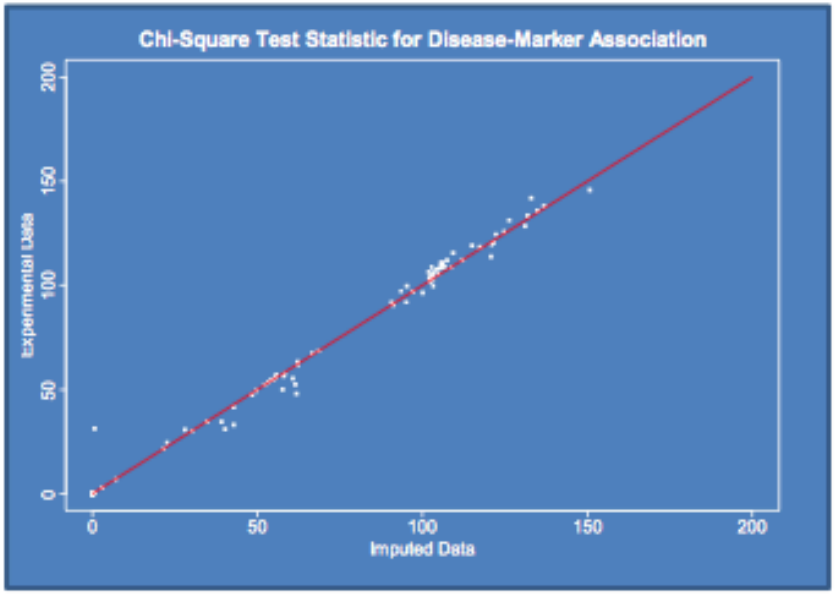
   Howie, Bryan, Jonathan Marchini, and Matthew Stephens. "Genotype imputation with thousands of genomes." *G3: Genes, Genomes, Genetics* 1.6 (2011): 457-470.

   https://mathgen.stats.ox.ac.uk/impute/impute_v2.html

Will imputation work

- 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



Imputation improve study power

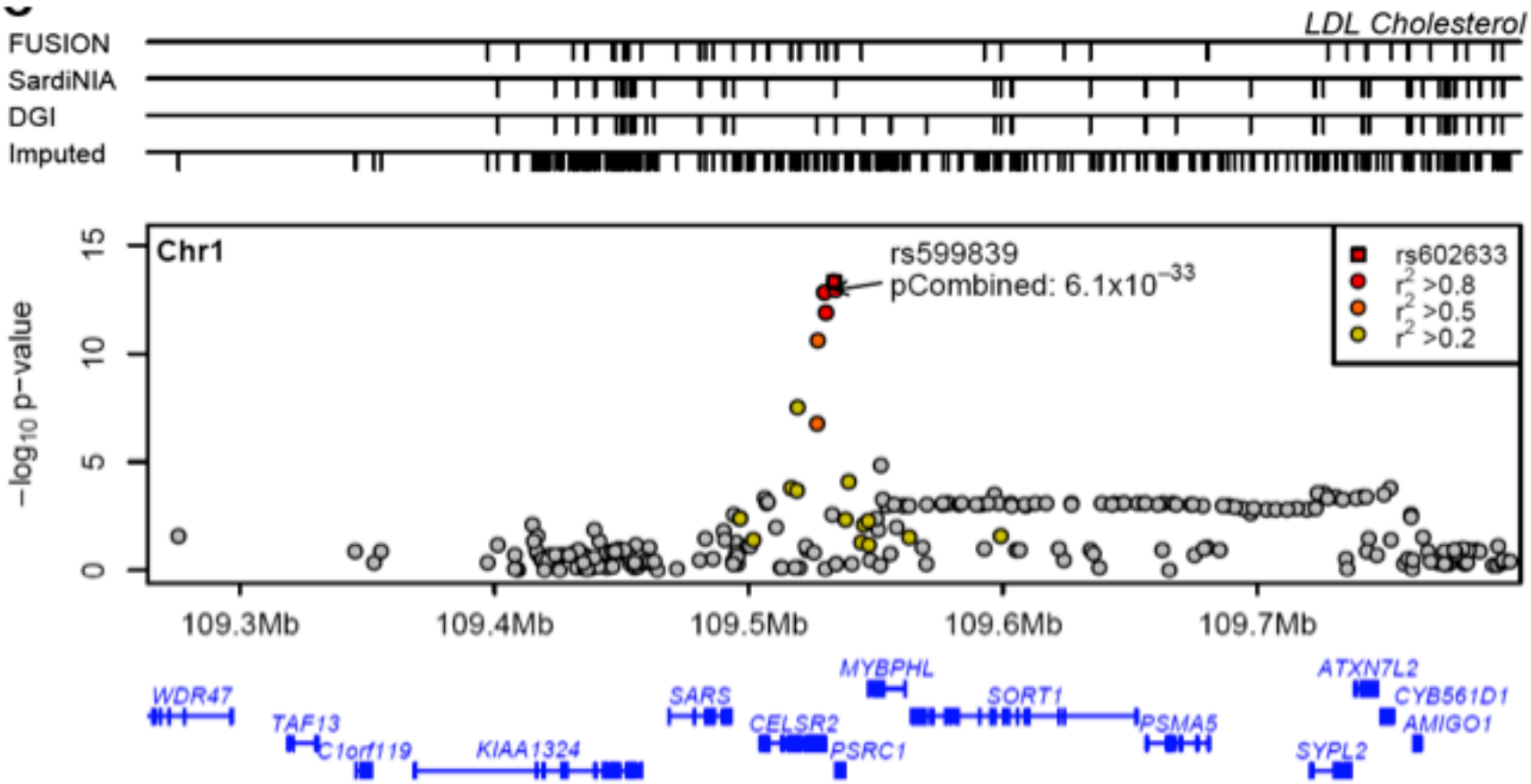
	Power	
Disease SNP MAF	tagSNPs	Imputation
2.5%	24.4%	56.2%
5%	55.8%	73.8%
10%	77.4%	87.2%
20%	85.6%	92.0%
50%	93.0%	96.0%

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 $r^2 > 0.80$.

Enable combination of studies

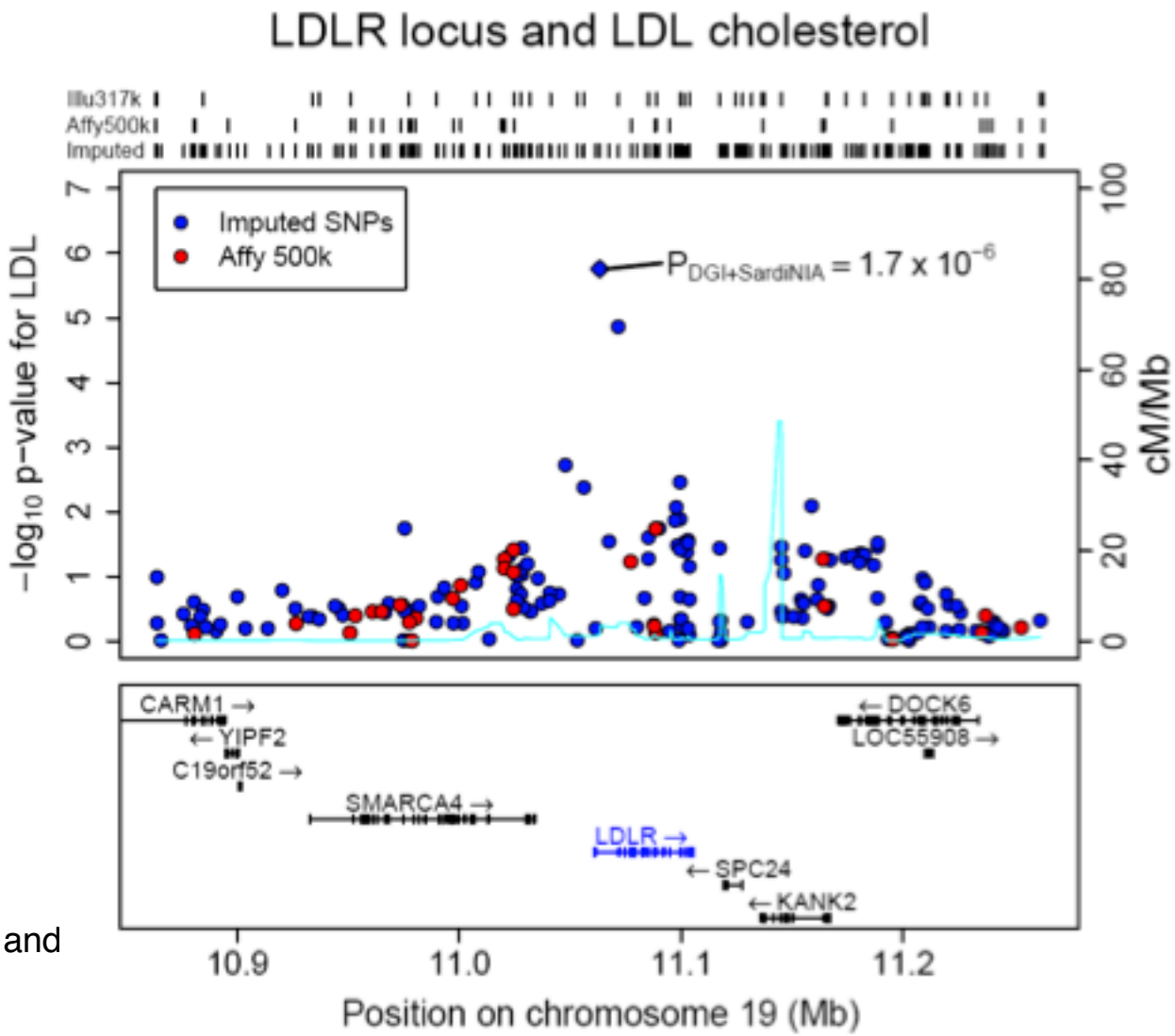
- New LDL loci, previously associated with CAD



NOTE: Imputed SNP is denser than FUSION, SardiNIA, DGI alone.

Boost GWAS signal

- LDLR example



Willer et al, *Nature Genetics*, 2008

Li et al, Annual Review of Genomics and Human Genetics, 2009

Summary

- Genotype imputation (in silico genotyping) can estimate missing genotypes accurately
- Genotype imputation are implemented in Hidden Markov Model
- Benefits of imputation includes:

  Increase power of GWAS study

  Facility combination on studies (different platform, QC et al)

  Better interpretation of GWAS results
- Lab
  - Use MiniMac to impute artificially masked genotypes of one sample in the HapMap3 project