Overview of statistical challenges for whole-genome association studies

Andy Clark
Cornell University
Dec 10, 2006

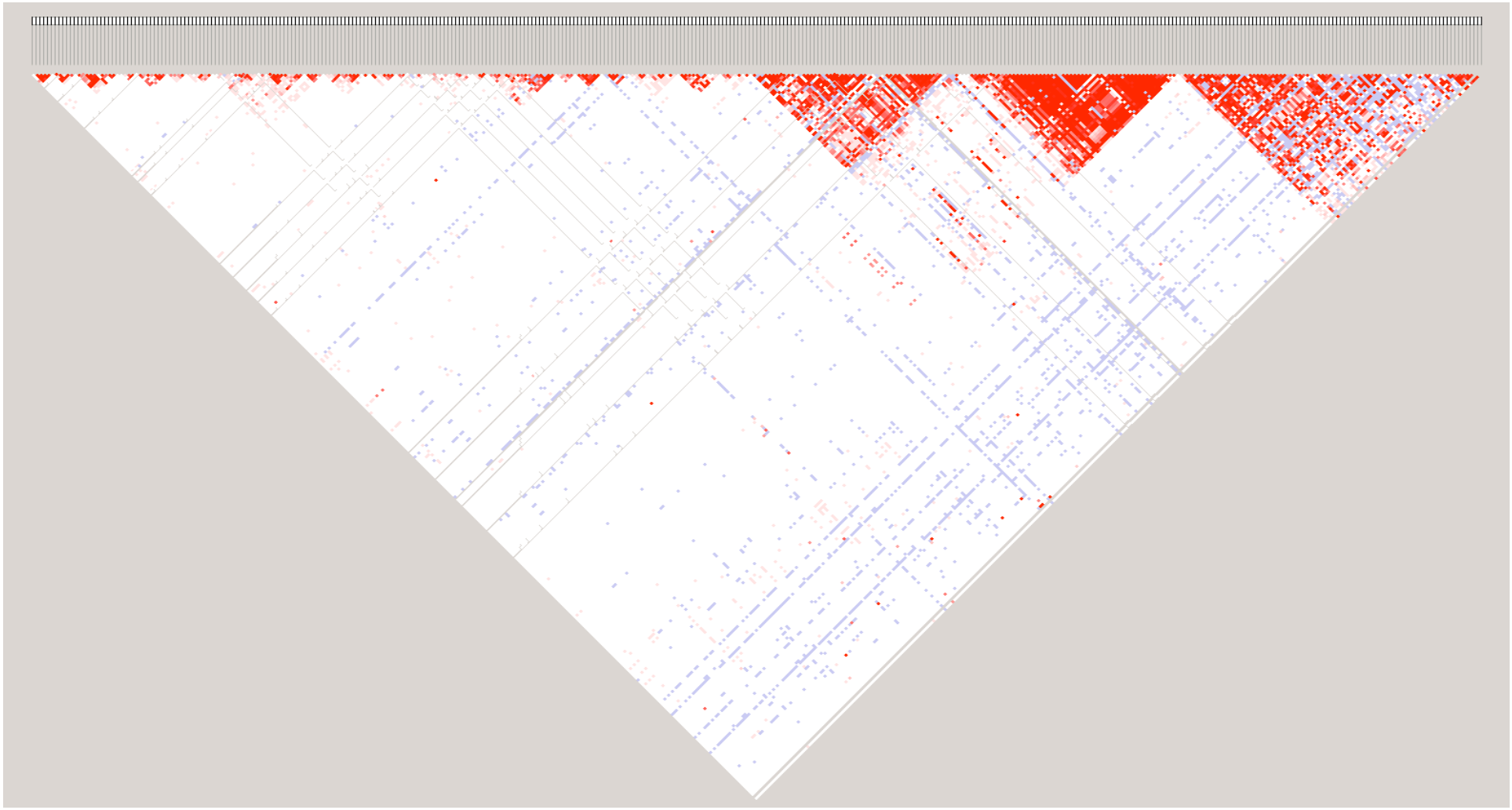
The GWAS Problem

- Complex disorders have resisted identification of causal genes by linkage.
- With sufficient marker density, it should in principle be possible to identify causal loci by whole-genome association testing
- Typical designs include 1000 cases and 1000 controls, and 500,000 markers.

Primary data for association tests

race	sex	city	TRIG	HDL	GT_a10 1280	GT_a10 1564	GT_a10 1616	GT_a10 1787	GT_a10 1899	GT_a10 2110	GT_a10 2954	GT_a10 2957	GT_a10 3132	GT_a10 3253
W	M	Birm	9.9633	5.22222	22	22	22	11	22	11	11	22	22	22
B	M	Birm	10.0317	6.48178	12	22	22	11	22	12	12	12	22	22
B	M	Chic	9.8371	6.76317	12	22	22	11	22	12	12	12	22	22
B	M	Chic	9.1981	3.81049	12	22	22	12	22	11	11	12	22	22
B	M	Minn	10.7382	6.18633	12	22	22	12	22	12	11	11	22	22
B	M	Minn	11.0786	4.68052	22	22	22	11	22	22	11	11	22	22
W	M	Oak	10.1686	3.52323	11	22	22	11	22	11	12	22	22	22
B	M	Oak	10.0122	6.74145	12	22	*	11	22	12	12	12	22	22
W	M	Birm	10.751	4.95946	22	22	22	11	22	11	11	22	22	22
B	M	Birm	10.9515	2.77605	12	22	22	11	22	12	11	12	22	22
B	F	Chic	8.6485	4.46937	12	22	22	12	22	12	11	11	22	22
B	F	Chic	9.538	6.92393	12	22	22	11	22	12	12	12	22	22
W	F	Minn	10.2062	3.7493	22	22	22	11	22	11	11	22	22	22
B	F	Minn	9.3762	5.23078	22	22	22	11	22	12	11	11	22	22
W	F	Oak	10.0905	6.4322	22	22	22	11	22	11	11	22	22	22
B	F	Oak	10.0816	4.05945	22	22	22	11	22	11	11	22	22	22
W	F	Birm	10.123	4.81638	22	22	22	11	22	11	11	22	22	22
B	F	Birm	10.3743	3.81774	12	22	22	11	22	12	12	22	22	22
B	F	Chic	9.5009	3.58655	12	22	22	11	22	11	12	22	-99	22

There is a complex correlation structure to the genotype data



Statistical inference to:

- identify **ETIOLOGY** (what genes are causal and what is the mechanism).

- make **PREDICTION** of phenotype given genotype.

Objectives

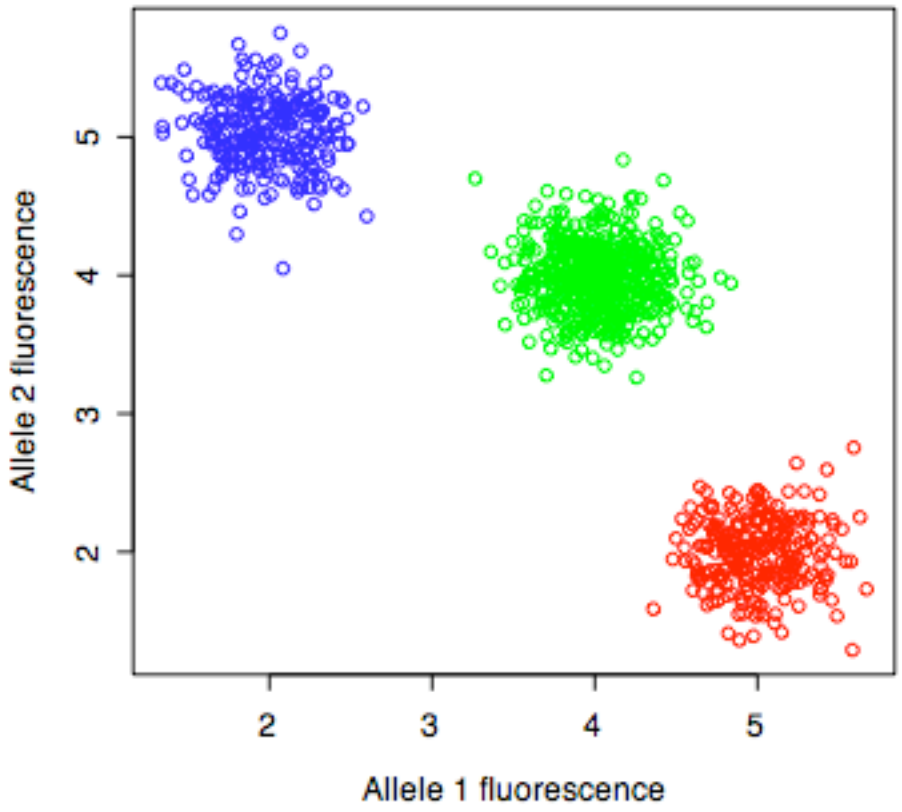
1. Genotype calling
2. Hardy-Weinberg testing
3. One SNP at a time tests
4. Haplotype-based inference
5. Coalescent-based inference
6. Genotype x environment
7. Epistasis
8. Replication

1. Genotype calling

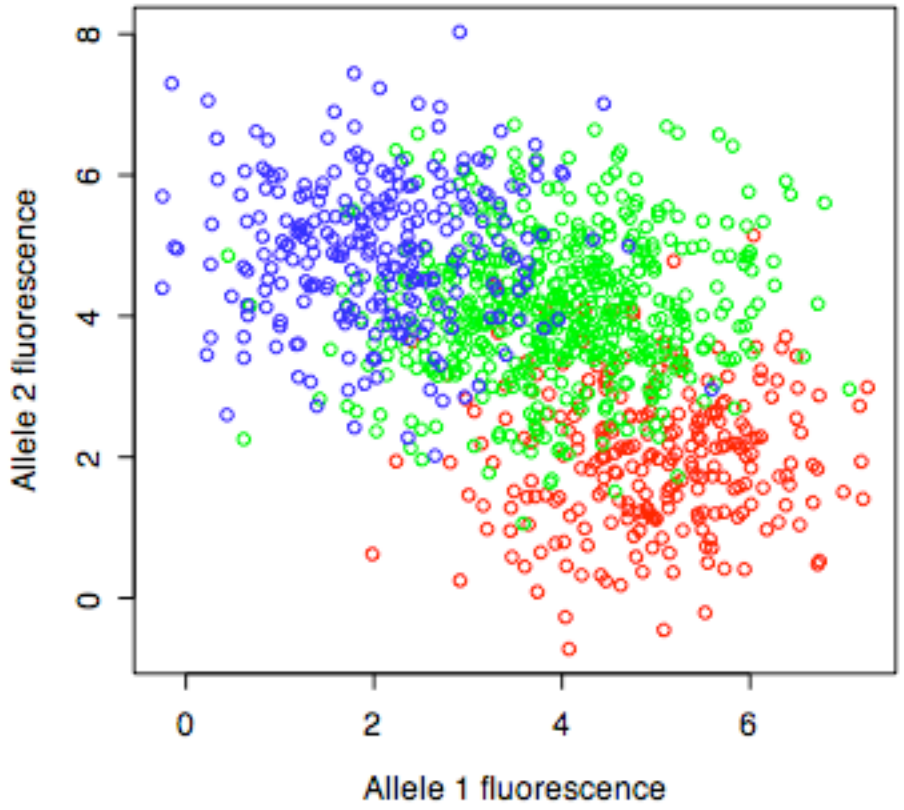
- Microarray SNP calling entails conversion of a continuous signal (hybridization intensity) to a discrete genotype call.

- Combining data from different arrays requires normalization.

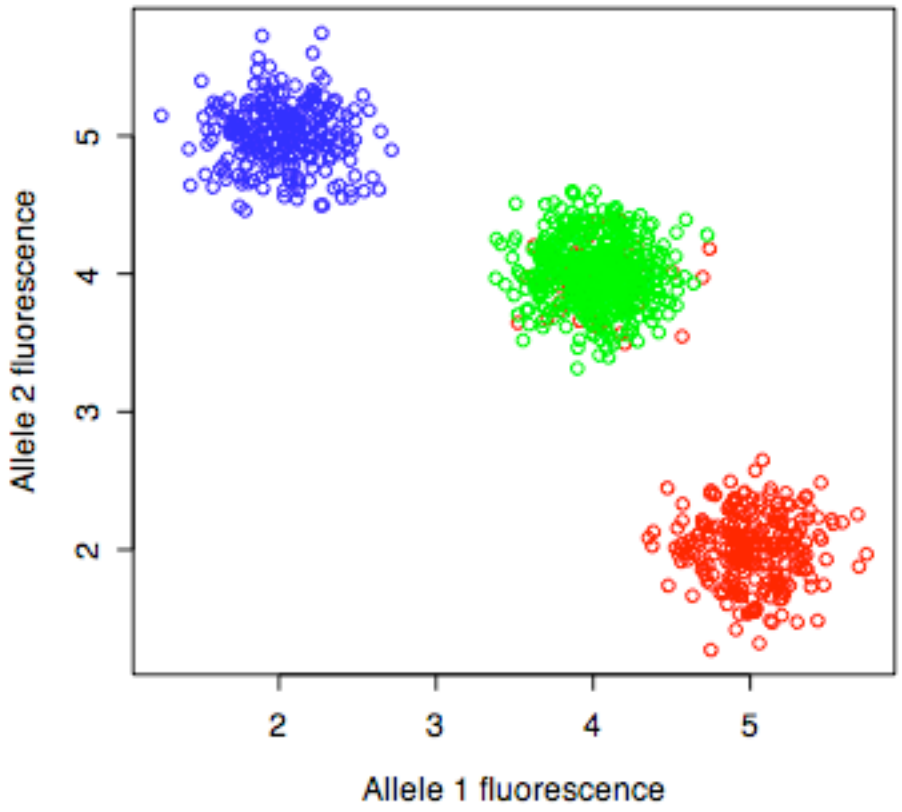
Genotyping platforms give a quantitative score for each allele



Some SNPs have poor separation



And others have obvious miscalls



2. Hardy-Weinberg testing

- Typically investigators select an arbitrary cut-off, and then delete SNPs that fail HW.

- Often $P < 0.001$ is used (tossing 500 perfectly good SNPs)

Expected HW distortion

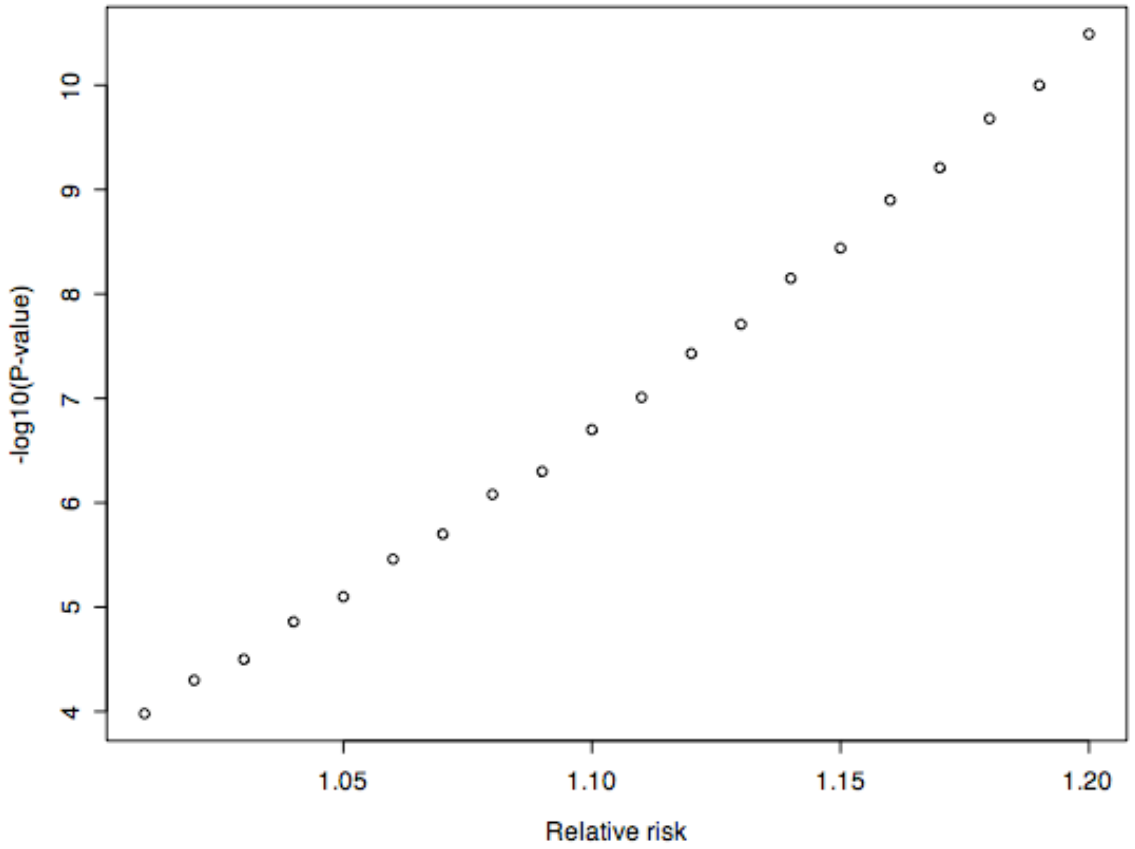
- If genotypes have different risk of disease, we might expect HW departures.

- Testing HW in cases and controls separately.

Affected status and HW distortion

- Assume 1000 cases and 1000 controls
- Allele frequency of 0.3
- Increasing relative risk moves controls into case category

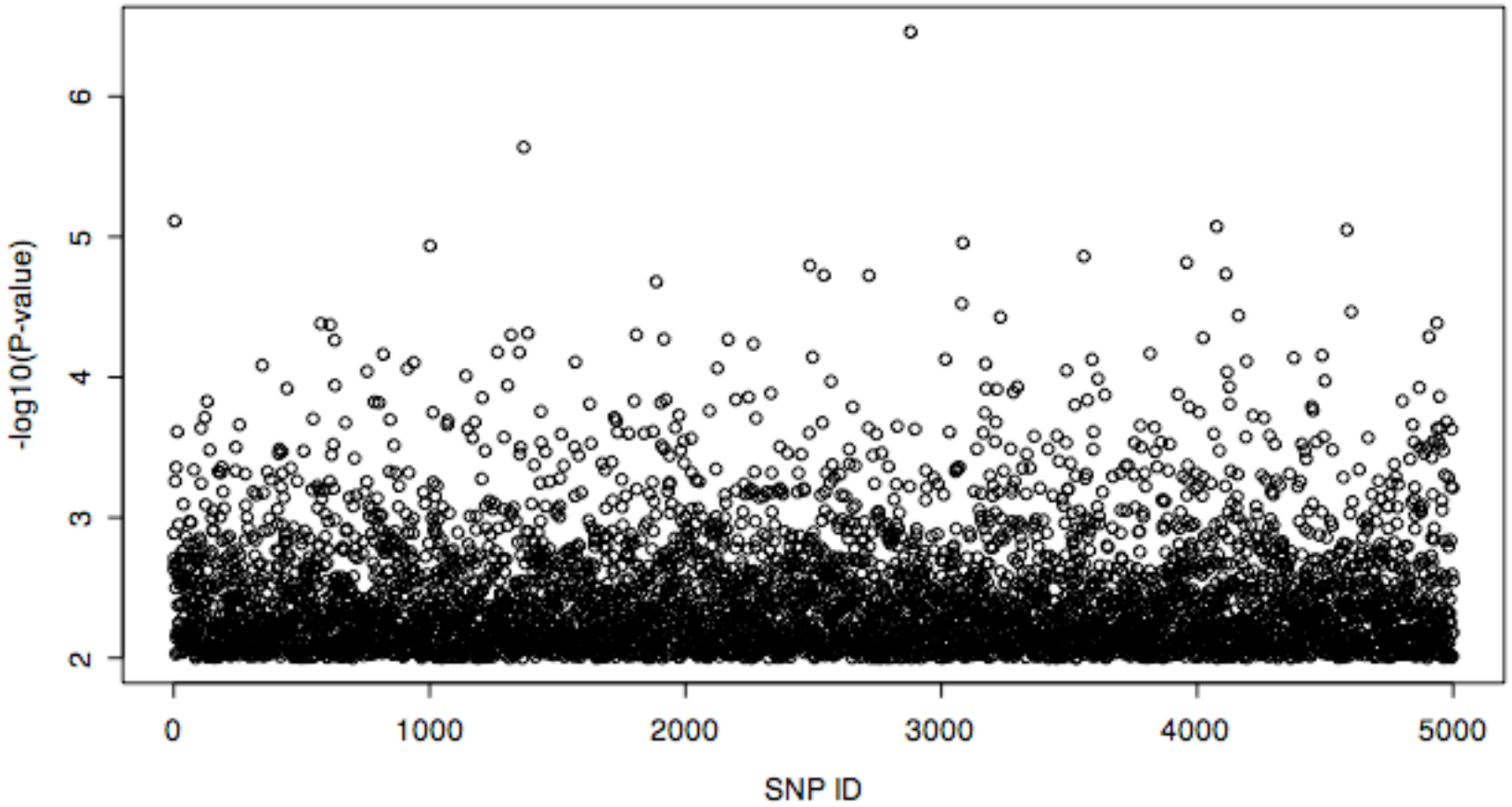
Small relative risk could distort genotype proportions



3. Testing one SNP at a time

- First consider the null hypothesis
  - Each test is a 2 x 3 chi-square test
  - Generate counts by multinomial sampling under the null hypothesis.
  - Calculate nominal P-values (all independent).

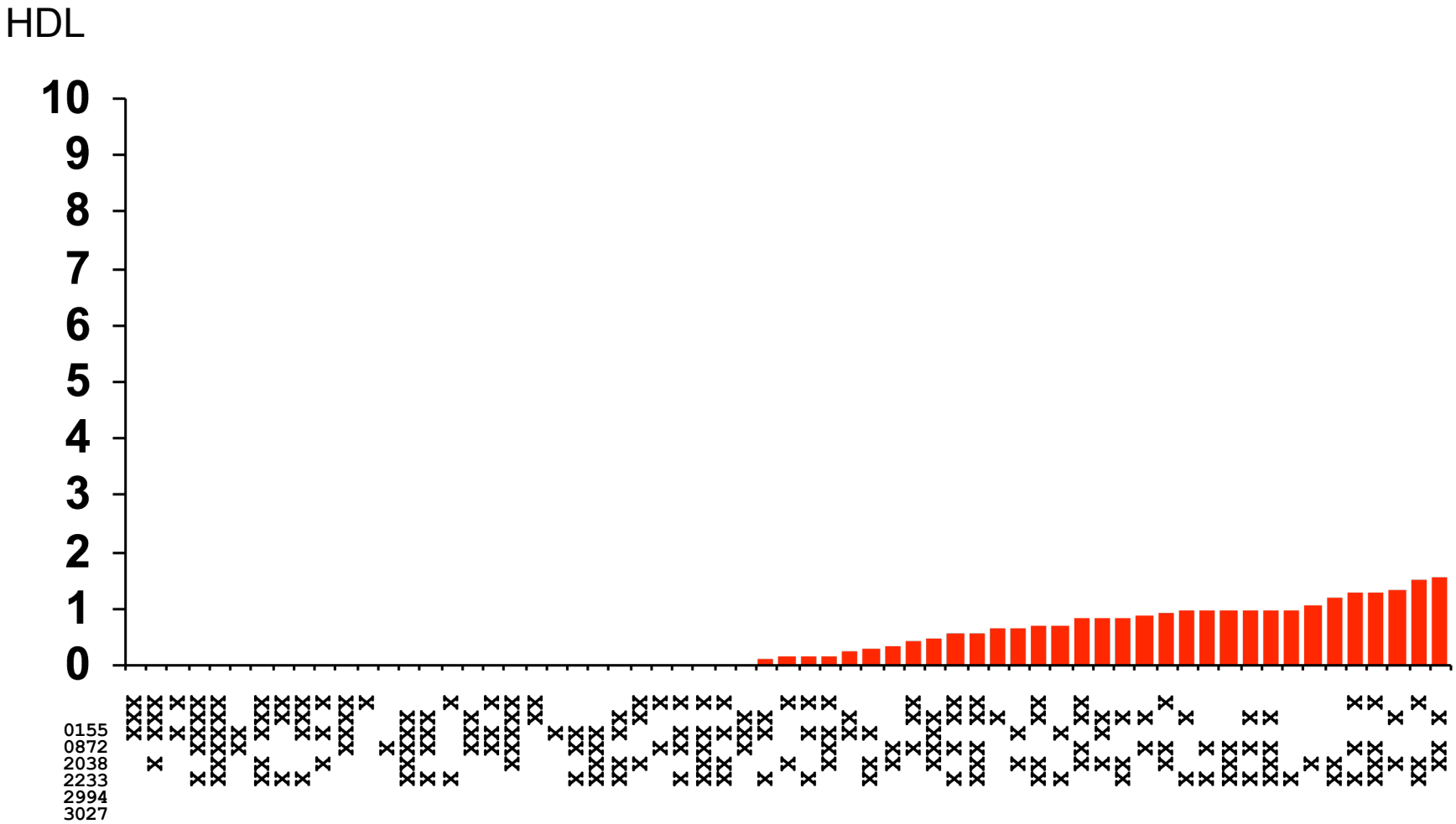
5000 most significant tests



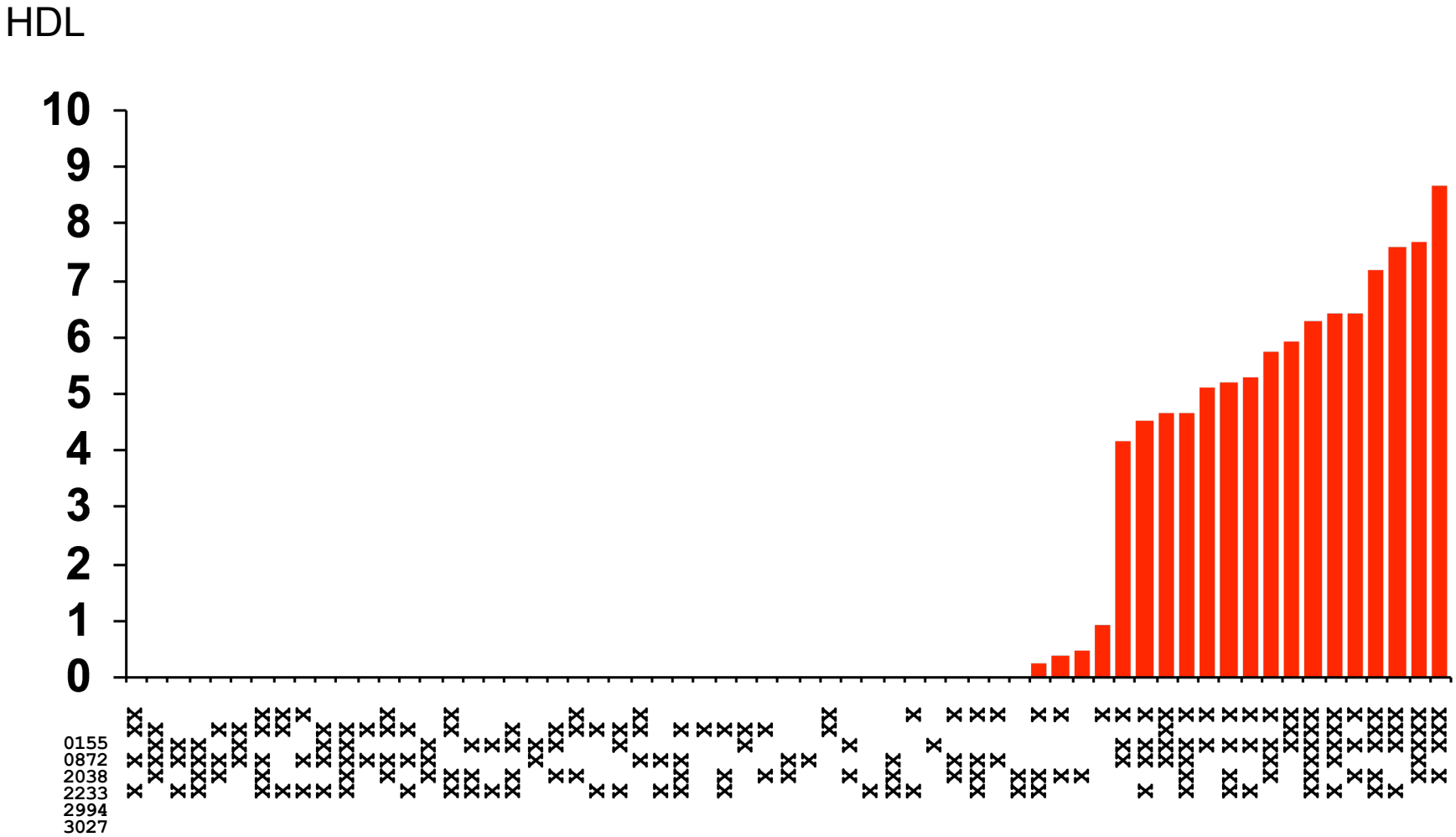
4. Haplotype inference

If risk is a consequence of a particular combination of SNPs within a gene, potentially phasing the SNPs will provide a boost in power.

ApoA2
North Karelia, Females



ApoA2
North Karelia, Males



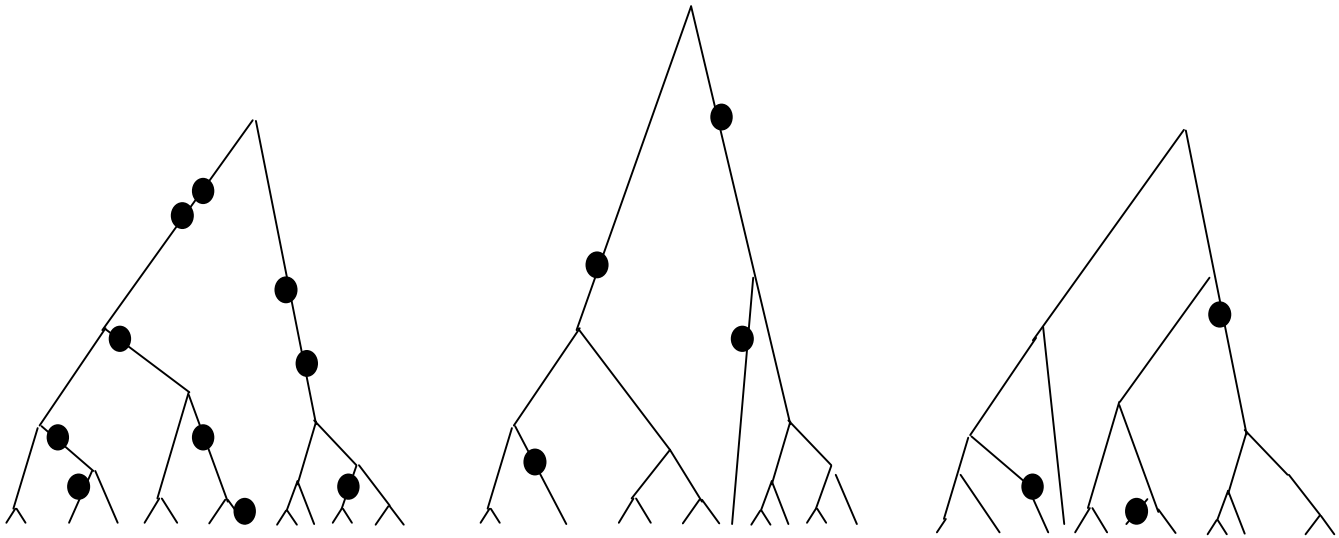
5. Tests that use the neutral coalescent

- Use SNP data to infer the ancestral recombination graph, or an approximation of it.

- Or, identify subsets of sites that form a perfect bifurcating tree ("perfect phylogeny").

- Assess whether the data fit a model where an unobserved mutation on each branch is causal to disease risk.

Perfect phylogenies



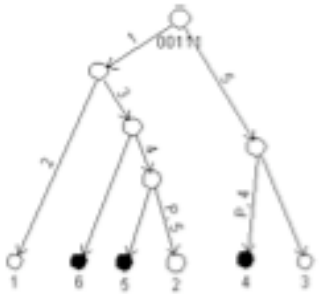
Zollner S. and J. Pritchard. 2005. Coalescent-based association mapping and fine mapping of complex trait loci. Genetics, 169:1071-1092.

Minichiello, M. and R. Durbin. 2006. Mapping trait loci using inferred ancestral recombination graphs. Am J. Hum. Genet.

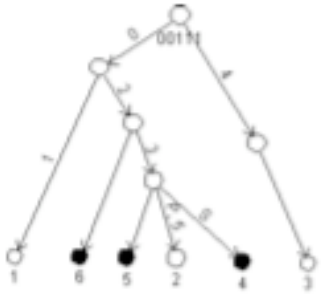
1	1	1	1	1	1
2	1	0	0	0	0
3	0	0	1	1	0
4	0	0	1	0	1
5	1	0	0	0	1
6	1	0	0	1	1

(a) SNP data.

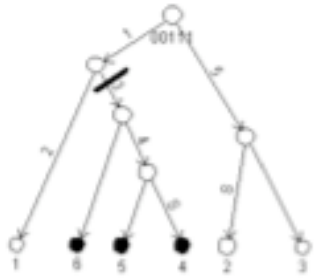
(b) An ARG deriving the data.



(c) Marginal tree for positions between sites 1-3



(d) Marginal tree for positions near site 4



(e) Marginal tree to the right of site 4 fits the phenotypes well

•As appealing as these methods may be, there is huge overhead in inference of the ARG, and the information is latent in the simple SNP genotypes.

•Also, the methods are appropriate for only sets of SNPs that are in linkage disequilibrium (I.e. intragenic sets of SNPs).

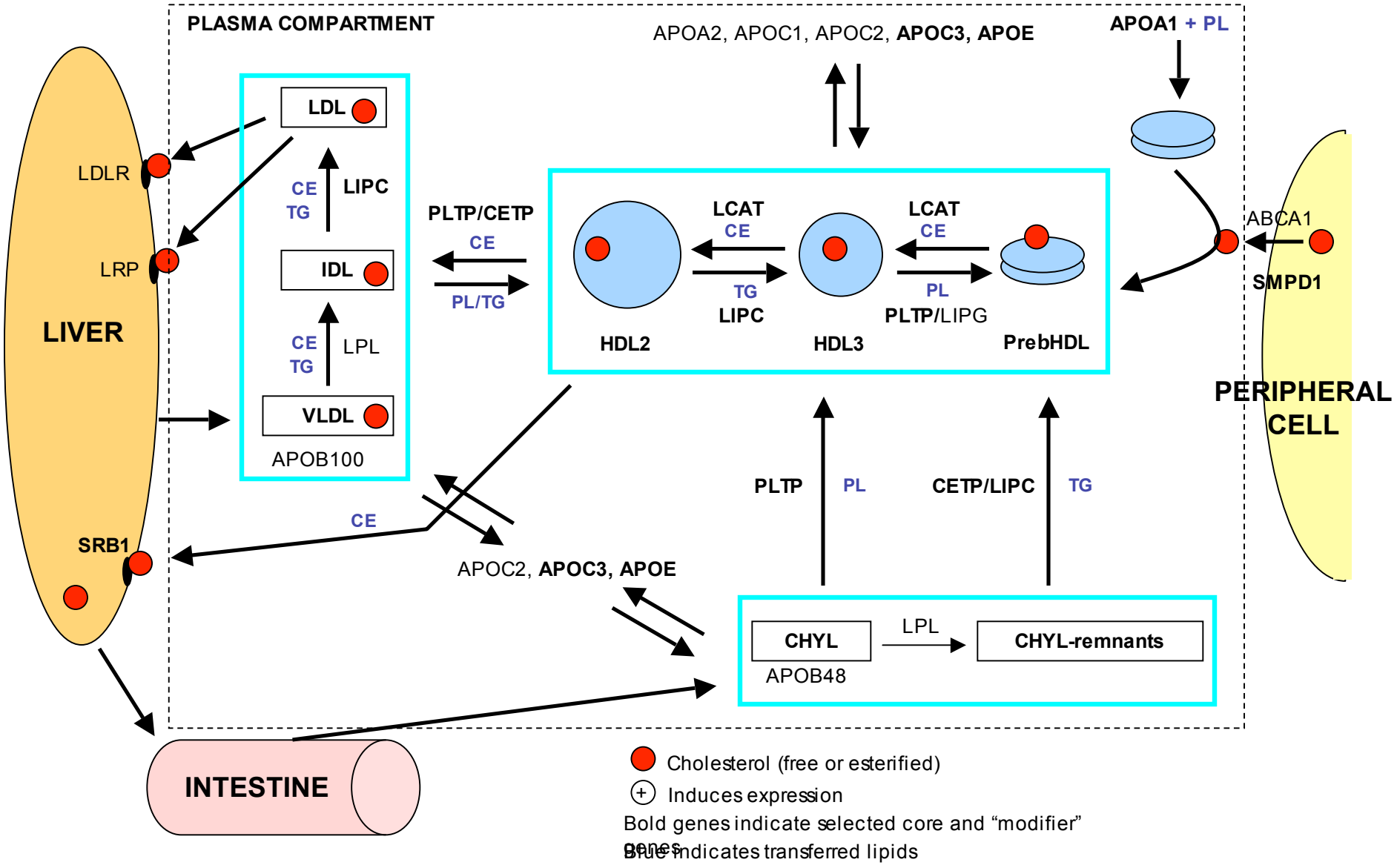
•Methods that explore combinations of SNPs, such as classification trees, are likely to do just as well.

6. Genotype x environment interaction

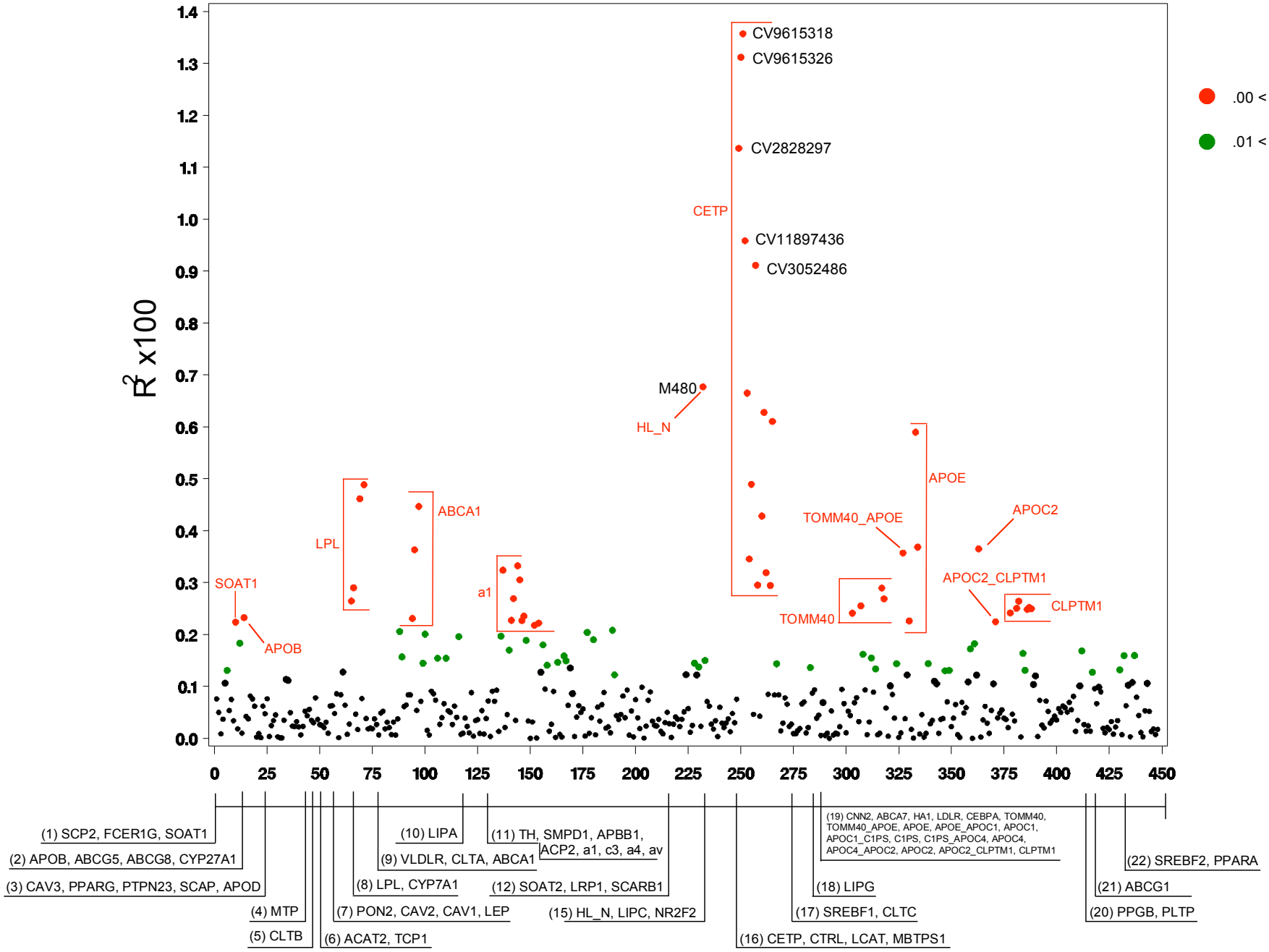
- Many diseases are triggered by environmental insult.

- Consider the genotype as determining susceptibility to adverse response.

Approximately 50% of the risk of cardiovascular disease is environmental in origin.



BASELINE, HDL Adjusted – Cross Classification Model for 448 SNP's

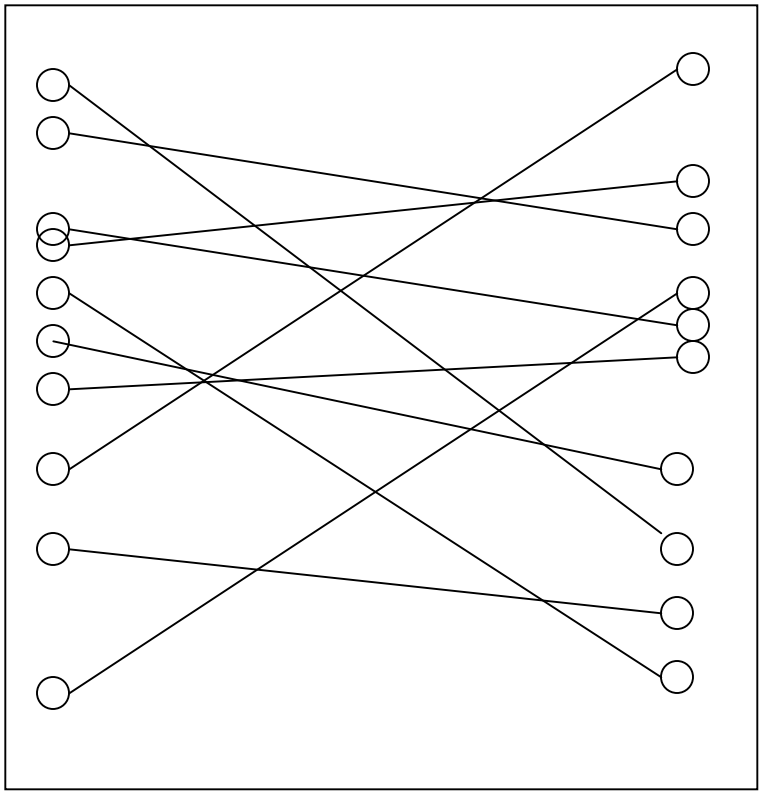


Patrice Milos (Pfizer) and metabolic syndrome.

Torcetrapib and class III trial termination.

Ubiquity of G x E in model organisms

Drosophila
survival %
12 hrs at
6% EtOH



Kristi Montooth

- G x E can only be tested if the environments are measured.

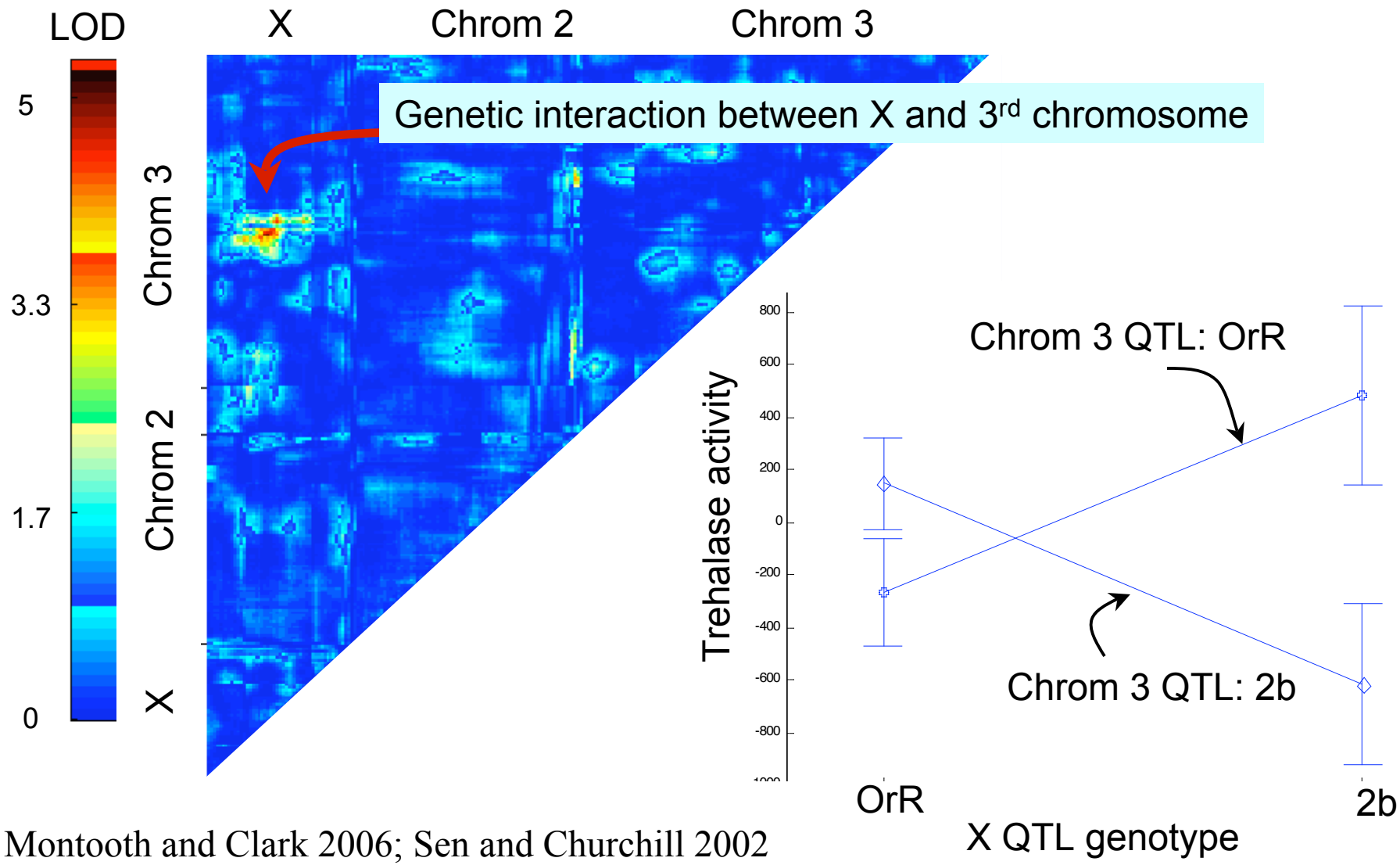
- Many GWAS have made use of previous epidemiological studies.

- Case-control study of schizophrenia seems like wishful thinking.

7. Gene-gene interaction (epistasis)

- We know that complex diseases involve multiple genes.
- Invariably with model organisms, studies designed to detect epistatic effects find them.
- Even though there are few documented cases of strong interaction in humans, these tests have been badly underpowered.

Epistatic interaction is common in model organisms

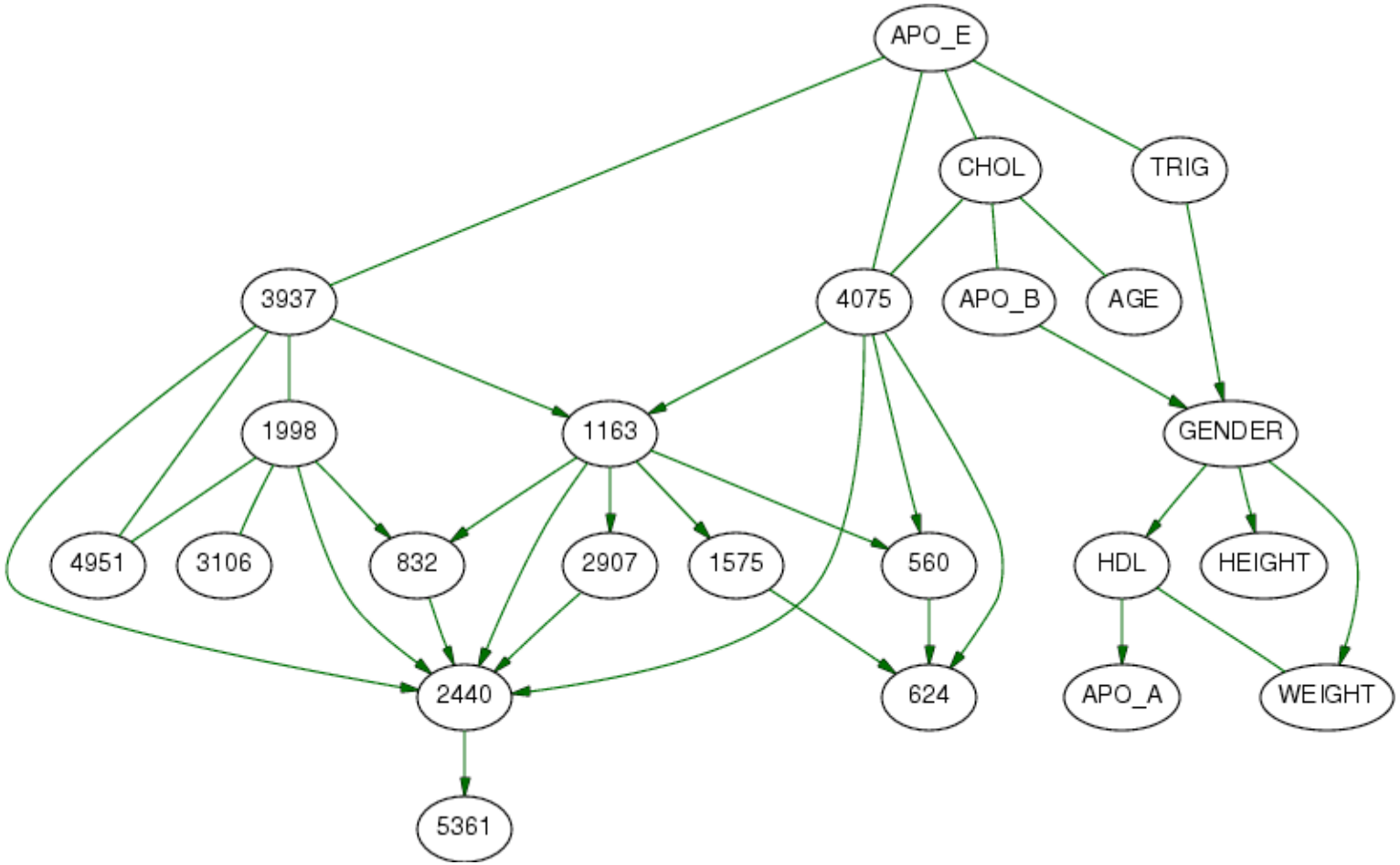


Montooth and Clark 2006; Sen and Churchill 2002

Model selection approaches

- Mixed model – control for hidden structure with linear models
- Bayesian regression (Stochastic variable search)
- Bayesian networks
- Bayesian classification – risk sets (Albrechtsen and Nielsen)

Bayesian Belief Networks : Whites and ApoE



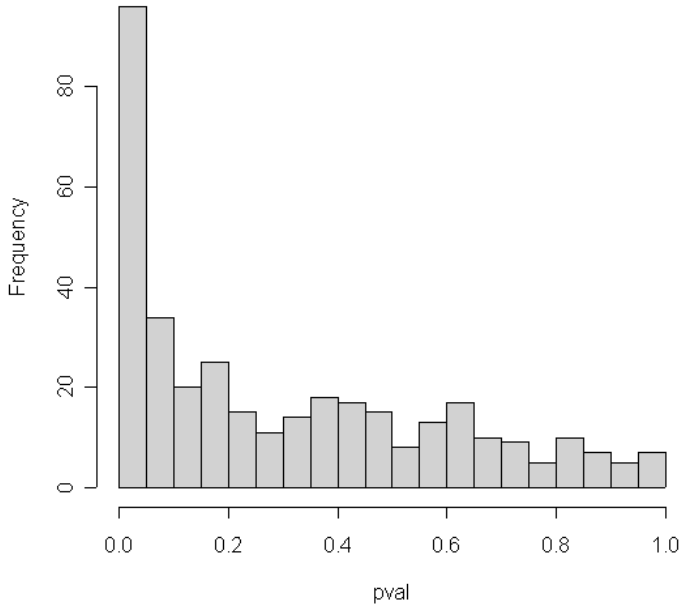
Rodin A, Mosley TH Jr, Clark AG, Sing CF, Boerwinkle E. 2005 J Comput Biol. 12:1-11.

8. Replication

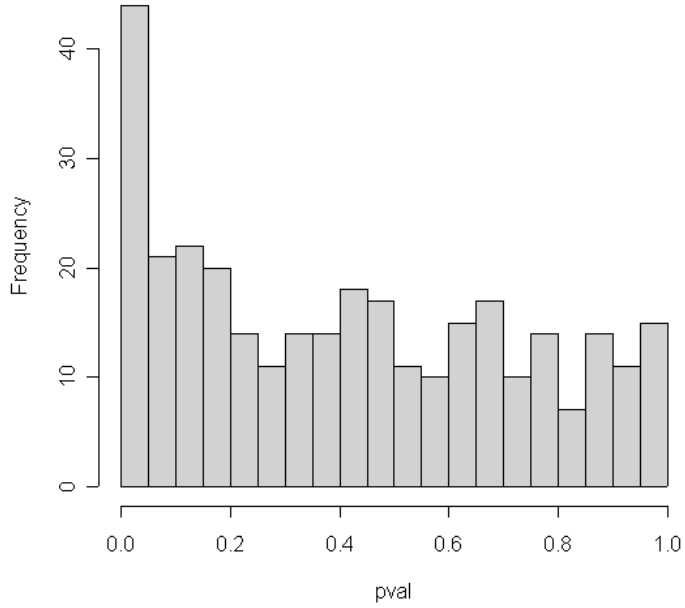
- Given the expected high false positive rate, replication in independent samples is seen as crucial.

- But samples from different cities are different -- there is no such thing as a replicated human population.

Are samples from different US cities like repeated draws from a single multinomial distribution?



African American
89/381 reject at P<0.05



European American
46/381 reject at P<0.05

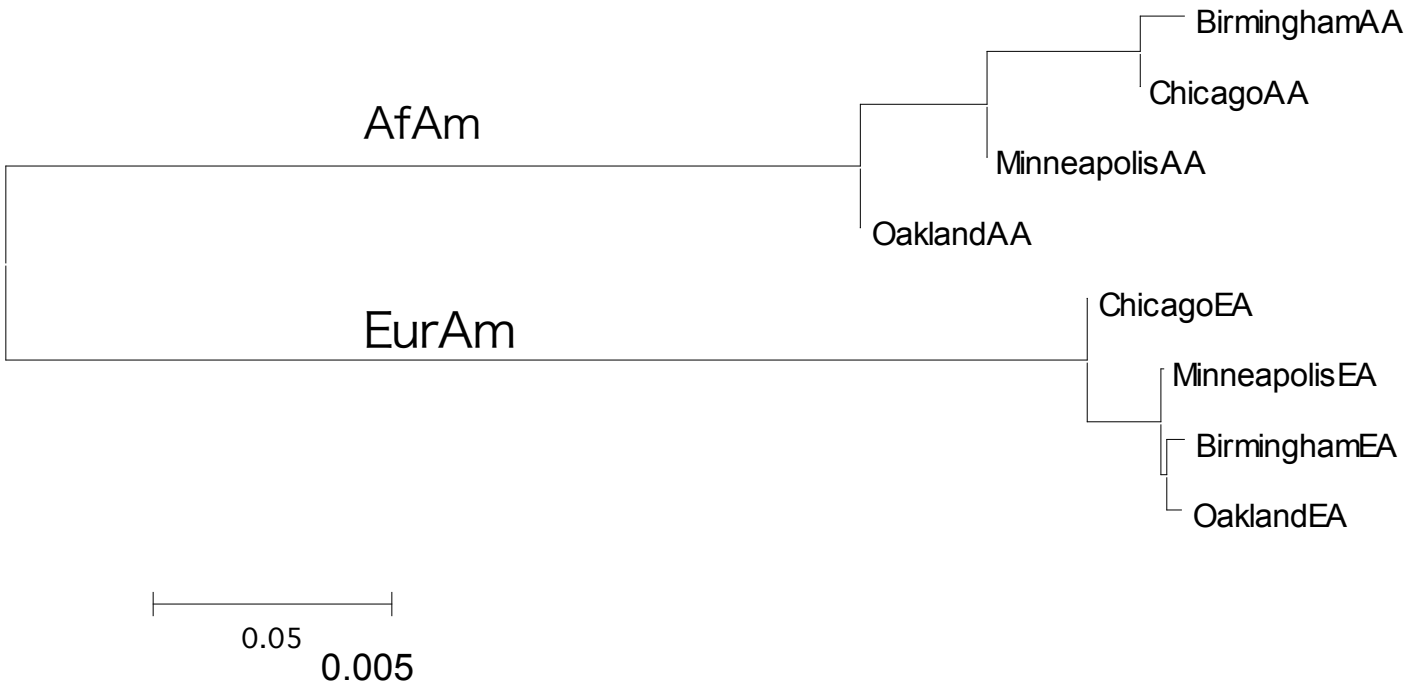
Test of H_0: samples are consistent with four draws from the same population

Consider this as a multinomial sampling problem, with draws taken from either 4 populations with their own parameters, or 4 draws from the same population. Construct a Likelihood Ratio statistic:

$$LR = \frac{\sum_{center(j)} \sum_{SNP(i)} N_{ij} \log(p_{ij})}{\sum_{center(j)} \sum_{SNP(i)} N_{ij} \log(p_{i.})}$$

- For African Americans: P < 0.0013
- For European Americans: P < 0.0084

NJ tree based on pairwise F_{ST}



Sampling design issues

- One-tiered or two (or more)?
- Unrelated or include familial relations?
- Genotype all SNPs at tier 1 and only the positives in subsequent tiers?

When will GWAS fail?

- Disease is genetically heterogeneous.
- Causal alleles fall in gaps of SNP arrays.
- Under-powered - too many small effects.
- High levels of G x E, with environment not adequately measured.
- High levels of epistasis.
- Populations stratified in subtle ways.