

Genetic background and population stratification

Shaun Purcell

Pak Sham

*Social, Genetic & Developmental
Psychiatry Research Centre,
IoP, KCL, London.*

Association & stratification

- Sewall Wright (1951)
 - concepts of population structure & impact on the evolutionary process
- C. C. Li (1972)
 - impact of population structure on disease-gene association studies
 - increase in type I errors
 - decrease in power

Signatures of stratification

- At a single locus
 - non-independence of paternal and maternal alleles
- Across loci
 - non-independence of alleles across loci
 - linkage disequilibrium, LD
 - use LD to map genes
 - spuriously infer indirect association

At a single locus

- Allele frequencies

$$\begin{array}{ll} A_1 & p \\ A_2 & q \end{array}$$

- Genotype frequencies

- expected under “Hardy-Weinberg equilibrium”

$$\begin{array}{ll} A_1A_1 & p^2 \\ A_1A_2 & 2pq \\ A_2A_2 & q^2 \end{array}$$

At a single locus

	Sub-population		
	1	2	<u>1+2</u>
A_1	0.1	0.9	<i>0.5</i>
A_2	0.9	0.1	<i>0.5</i>
A_1A_1	0.01	0.81	<i>0.41 (0.25)</i>
A_1A_2	0.18	0.18	<i>0.18 (0.50)</i>
A_2A_2	0.81	0.01	<i>0.41 (0.25)</i>

Quantifying population structure

- Expected average heterozygosity
 - in random mating subpopulation (H_S)
 - in total population (H_T)
 - from the previous example,
 - $H_S = 0.18$, $H_T = 0.5$
- Wright's fixation index
 - $F_{ST} = (H_T - H_S) / H_T$
 - $F_{ST} = 0.64$
 - 0.01 - 0.05 for European populations
 - 0.1 - 0.3 for most divergent populations

Across loci

- 200 Scandinavians

	B ₁	B ₂
A ₁	160	160
A ₂	40	40

$$\chi^2 = 0$$

- 200 Spaniards

	B ₁	B ₂
A ₁	160	40
A ₂	160	40

$$\chi^2 = 0$$

Across loci

- 400 Scandinavians and Spaniards combined

	B ₁	B ₂
A ₁	320	200
A ₂	200	80

$$\chi^2 = 7.81$$

- Spurious association
 - not reflective of genetic distance
 - *A* and *B* might be on different chromosomes

Solutions

- Family controls
 - related individuals share same sub-population
 - e.g. TDT test
- Index of membership
 - self-reported ethnicity
 - not always accurate / effects may be subtle
 - infer from an individual's genetic background
 - *detection*
 - *look for signatures of population stratification*
 - *correction*
 - *correct tests for inferred substructure*

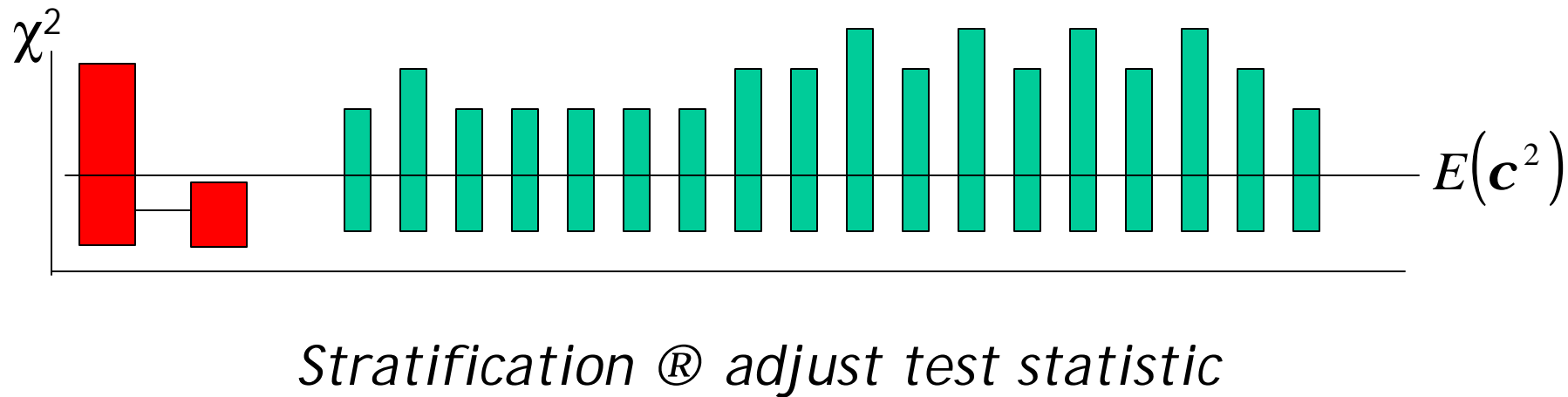
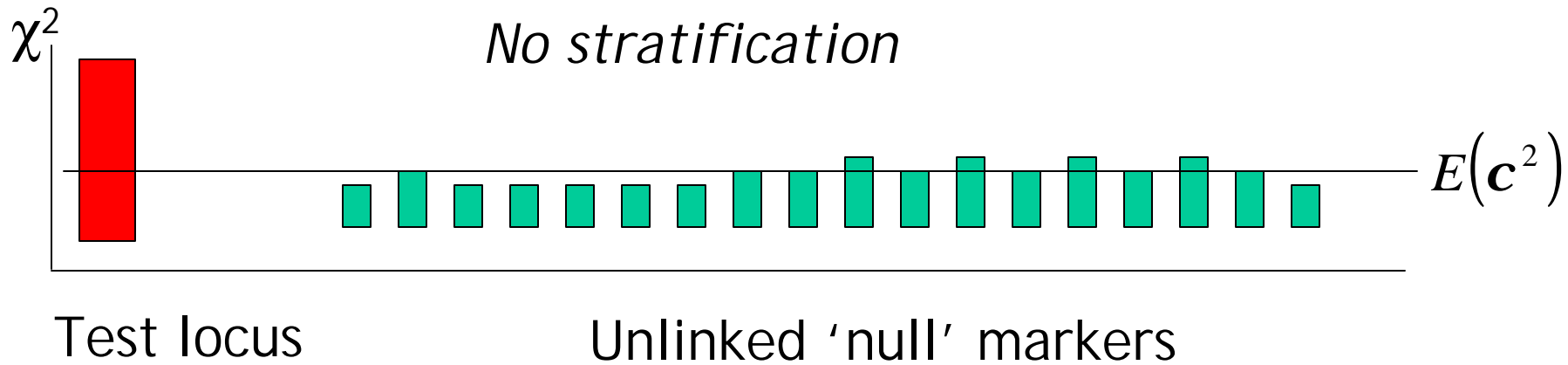
Genetic background approaches

- Genomic Control
- Structured Association
 - Method: multilocus genotype data to detect and correct for stratification
 - Premise: stratification operates globally – on whole genome, whereas LD operates locally at short scales

Genomic control

- χ^2 statistics not distributed as χ^2 under PS
“overdispersion”
 - Pritchard & Rosenberg (1999)
 - assess whether χ^2 statistics for unlinked markers are okay
 - Devlin & Roeder (1999)
 - null locus test statistic T_N distributed χ^2_1
 - in presence of stratification, $T_N / I \sim \chi^2_1$
 - estimate I
 - statistic at test locus $T / I \sim \chi^2_1$

Genomic control



Genomic control

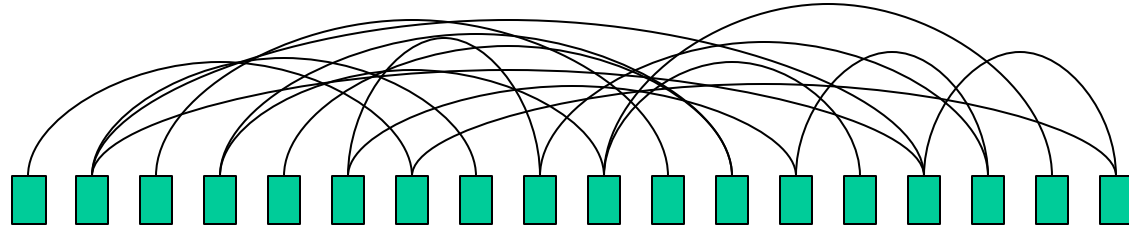
- λ Inflation factor $I \approx 1 + RF \sum_k (f_k - g_k)^2$
 - R number of cases (controls)
 - F Wright's F_{ST} coefficient of inbreeding
 - $g_k (f_k)$ Proportion of cases (controls) from subpopulation k
- Example
 - 2 equiprecurrent subpopulations, $F_{ST} = 0.01$
 - Disease twice as common in one subpopulation
 - $R = 1000$
 - $I \gg 1.5$

Structured association

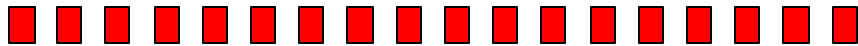
- Assignment of individuals to subpopulations
 - Test for association conditional on subpopulation
- Distance-based approaches
- Model-based approaches
 - Pritchard *et al* (2000)
 - Bayesian framework (STRUCTURE)
 - Satten *et al* (2001)
 - Latent class analysis model
 - Current work
 - Latent class analysis model (L-POP)

Structured association

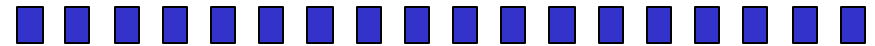
LD observed under stratification



Unlinked 'null' markers



Subpopulation A



Subpopulation B

Advantages of SA

- Structure of intrinsic interest
- Multi-allelic
- Handles allelic heterogeneity between subpopulations
- Does not assume that F_{ST} is constant across the genome

Structured association

- Genotype a number of loci across the genome
- Loci must be *unlinked*
 - *in a non-stratified sample*, would not expect to observe correlations between these loci
 - *in a stratified sample*, would not expect to observe correlations between these loci *within sub-population*

Latent Class Analysis

- K sub-populations, latent classes
 - Sub-populations vary in allele frequencies
 - Random mating within subpopulation
- Within each subpopulation
 - Hardy-Weinberg and linkage **equilibrium**
- For population as a whole
 - Hardy-Weinberg and linkage **disequilibrium**

Latent Class Analysis

- **Goal** : assign each individual to class C of K
- **Key** : conditional independence of genotypes, G within classes

$P(C | G)$ posterior probabilities

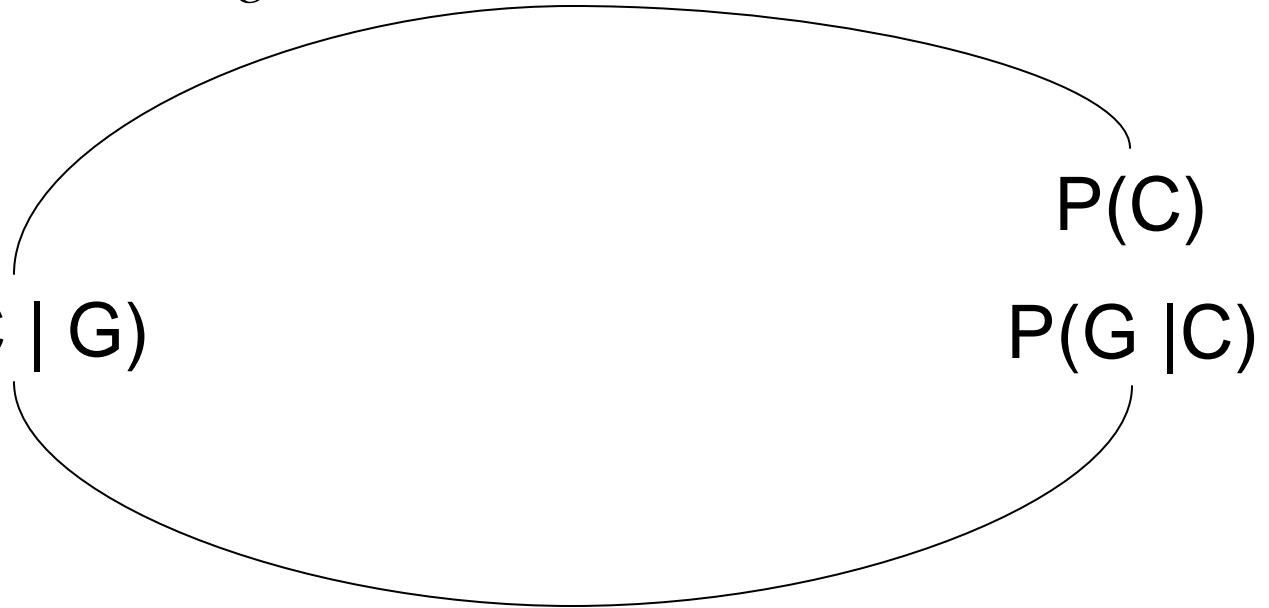
$P(C)$ prior probabilities

$P(G | C)$ class-specific allele frequencies

E-M algorithm

E step:

counting individuals and alleles in classes



M step:

Bayes theorem, assume conditional independence

E-step

- Initially, random values for $P(C | G)$
- Estimate prior class probabilities

$$P(C) = \frac{\sum_i P(C | G)}{N}$$

Sum over $i = 1$ to N individuals

E-step

- Estimate class-specific allele frequencies

$$P(G_l = k | C) = \frac{\sum_i P(C | G)(D_{i1} + D_{i2})}{(2N)P(C)}$$

- For nonmissing data D_{i1} , D_{i2}
 - = 1 if paternally / maternally inherited allele is k ,
 - = 0 otherwise

M-step

- For each individual, posterior probabilities

$$P(C | G) = \frac{P(G | C)P(C)}{\sum_j P(G | C)P(C)}$$

Sum over $j = 1$ to K classes

Assumes conditional independence

$$P(G | C) = \prod_l P(G_l = k_1 | C)P(G_l = k_2 | C)$$

Product over $l = 1$ to L loci

Likelihood

- Likelihood of an individual

$$L_i = \sum_j P(G | C) P(C)$$

- Use AIC to select optimal K solution

$$AIC = -2 \sum_i \ln L_i - 2df$$

Correction

- Pritchard *et al*, Satten *et al*
 - Test of association combined with detection of structure
 - Binary disease traits
- $P(C|G)$ as covariates
 - $K-1$ covariates
 - Alternatively, assign to class with highest $P(C|G)$
 - Applicable to any type of analysis / trait
 - Can allow for interactions (i.e. different effects between subpopulations)

Example #1

ID1	1/1	1/1	1/1	1/1	1/1
ID2	1/1	1/1	1/1	1/1	1/1
ID3	2/2	2/2	2/2	2/2	2/2
ID4	2/2	2/2	2/2	2/2	2/2
ID5	0/0	0/0	0/0	0/0	0/0

Example #1

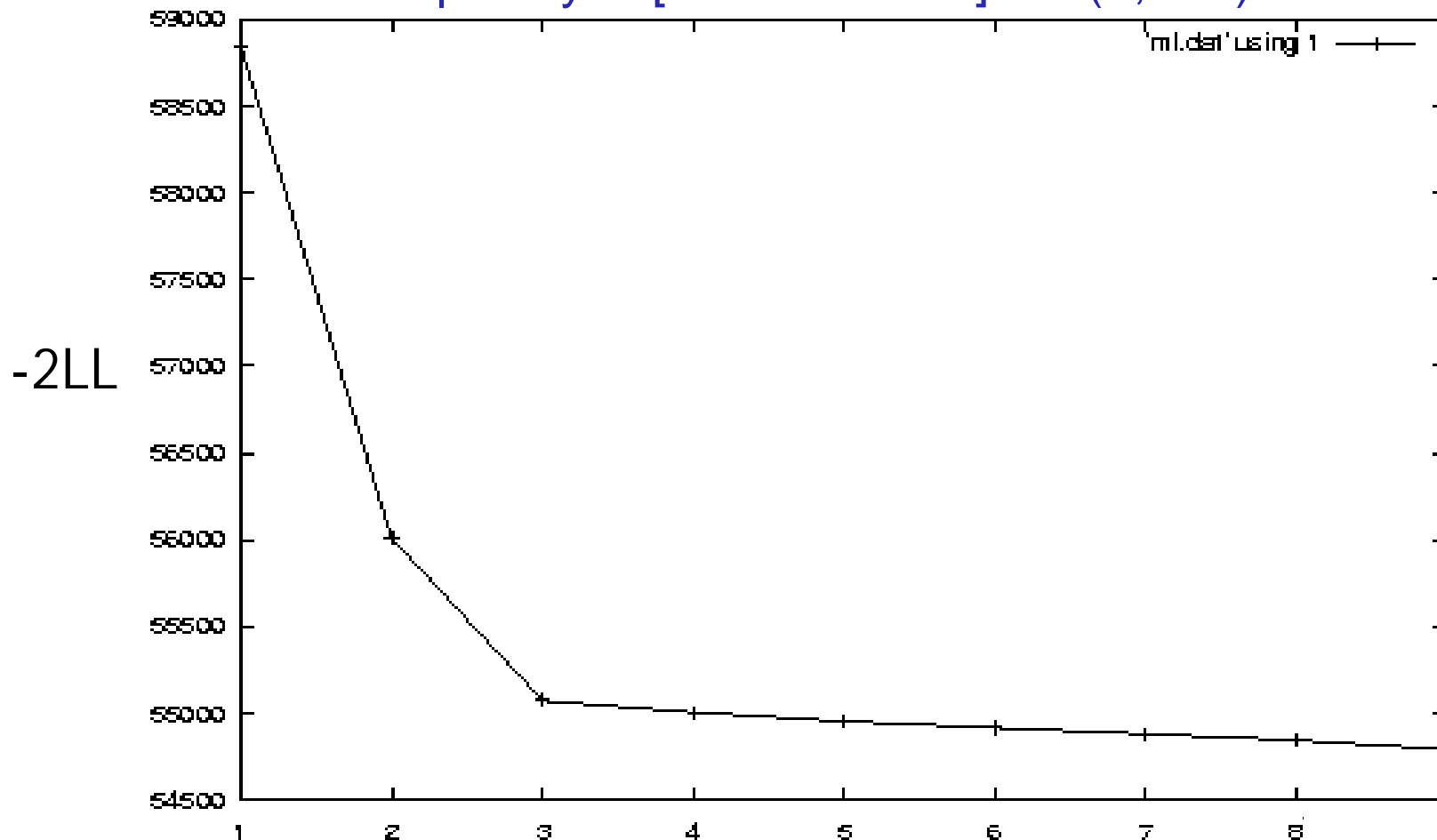
<i>K</i>	-2LL	AIC	$P(C = 1)$	$P(C = 2)$	$P(C = 3)$
1	55.45	65.45	1.00		
2	5.55	27.55	0.50	0.50	
3	5.55	39.55	0.50	0.28	0.22

Example #1

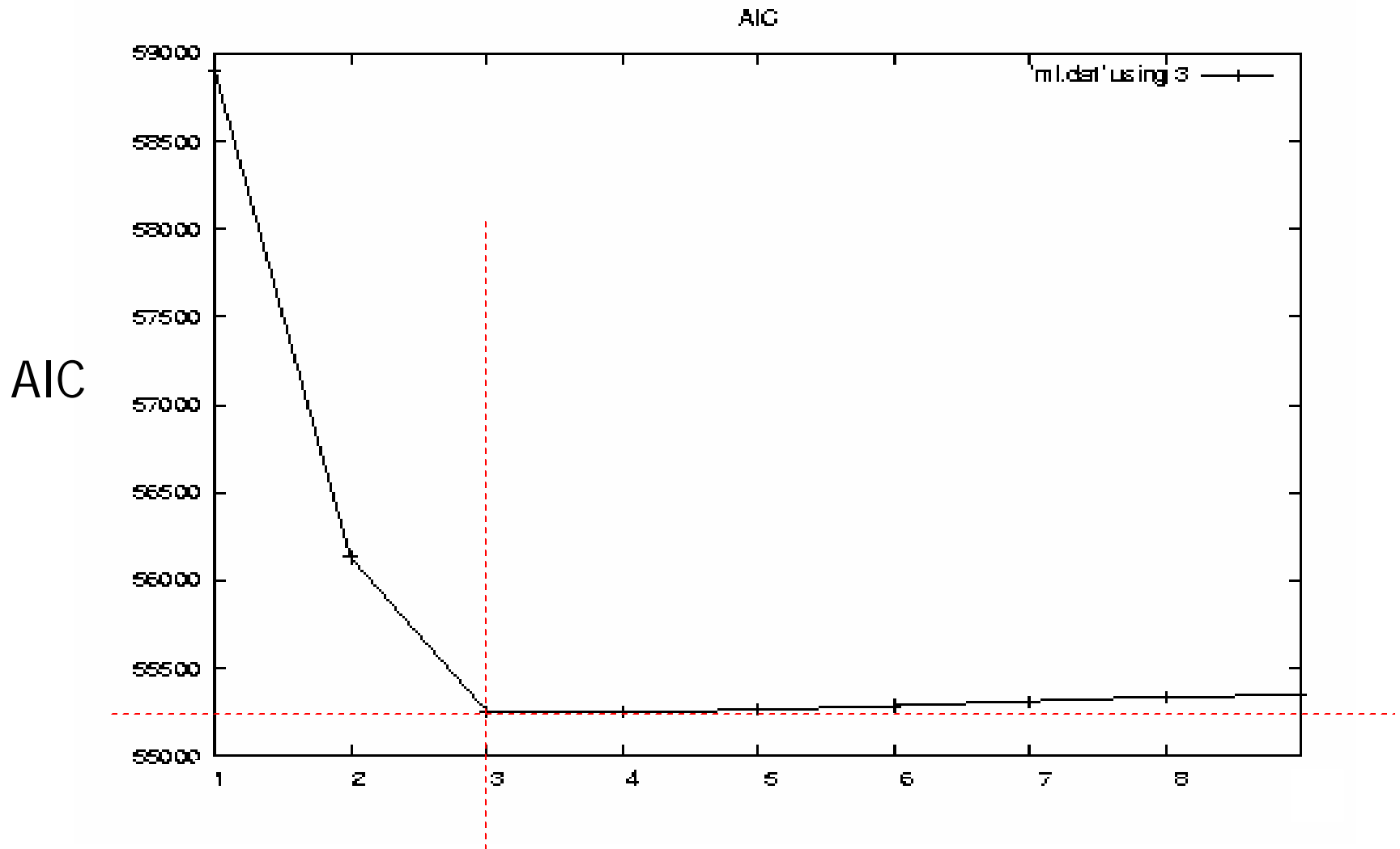
	$P(C=1 G)$	$P(C=2 G)$
ID1	0.00	1.00
ID2	0.00	1.00
ID3	1.00	0.00
ID4	1.00	0.00
ID5	0.50	0.50

Example #2

- 3 subpopulations, 1000 individuals, 30 SNPs
 - 70% : 20% : 10%
 - allele frequency $U[0.001, 0.999] + N(0, 0.2)$



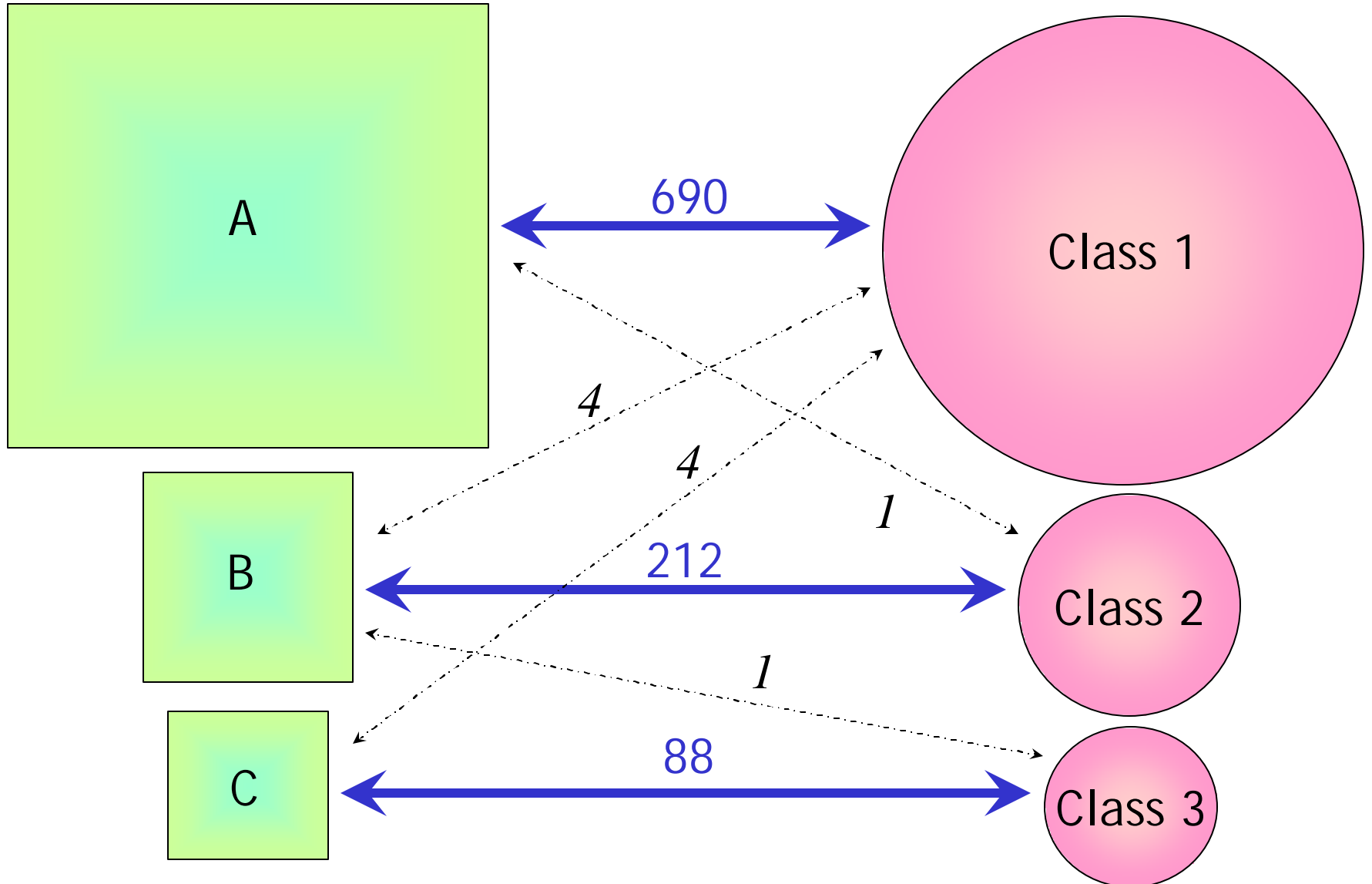
Example #2



$K=3$

Sub-population

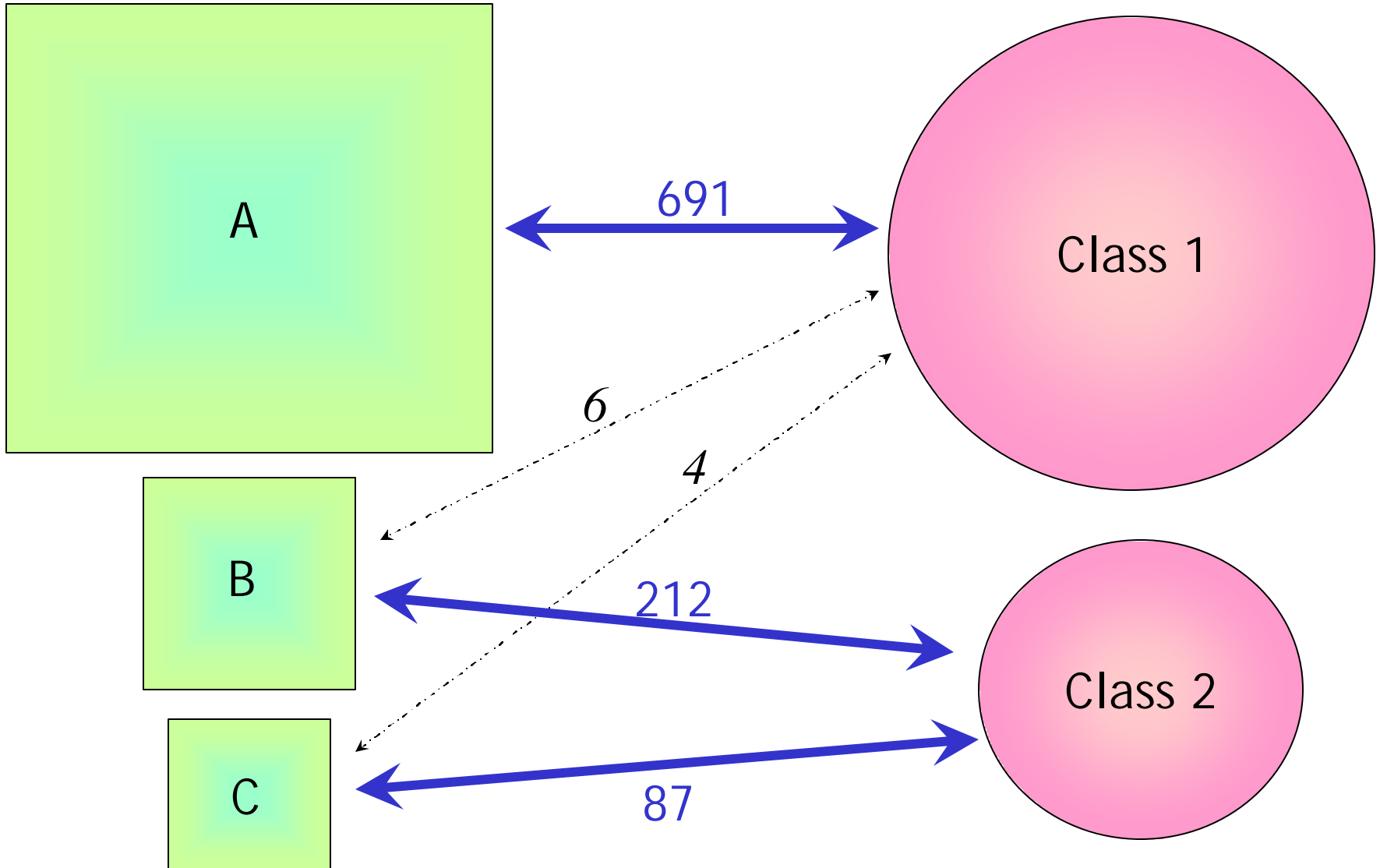
Latent class



$K=2$

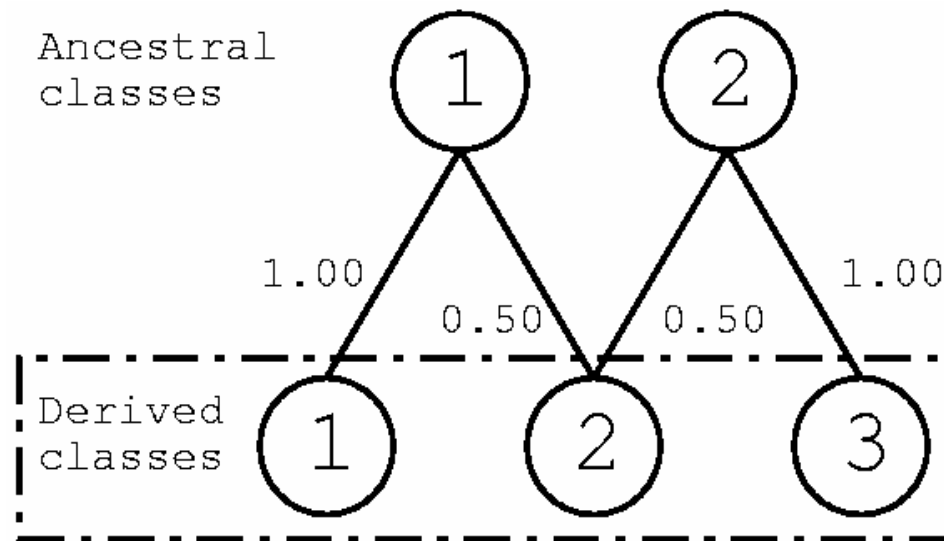
Sub-population

Latent class



Allowing for admixture

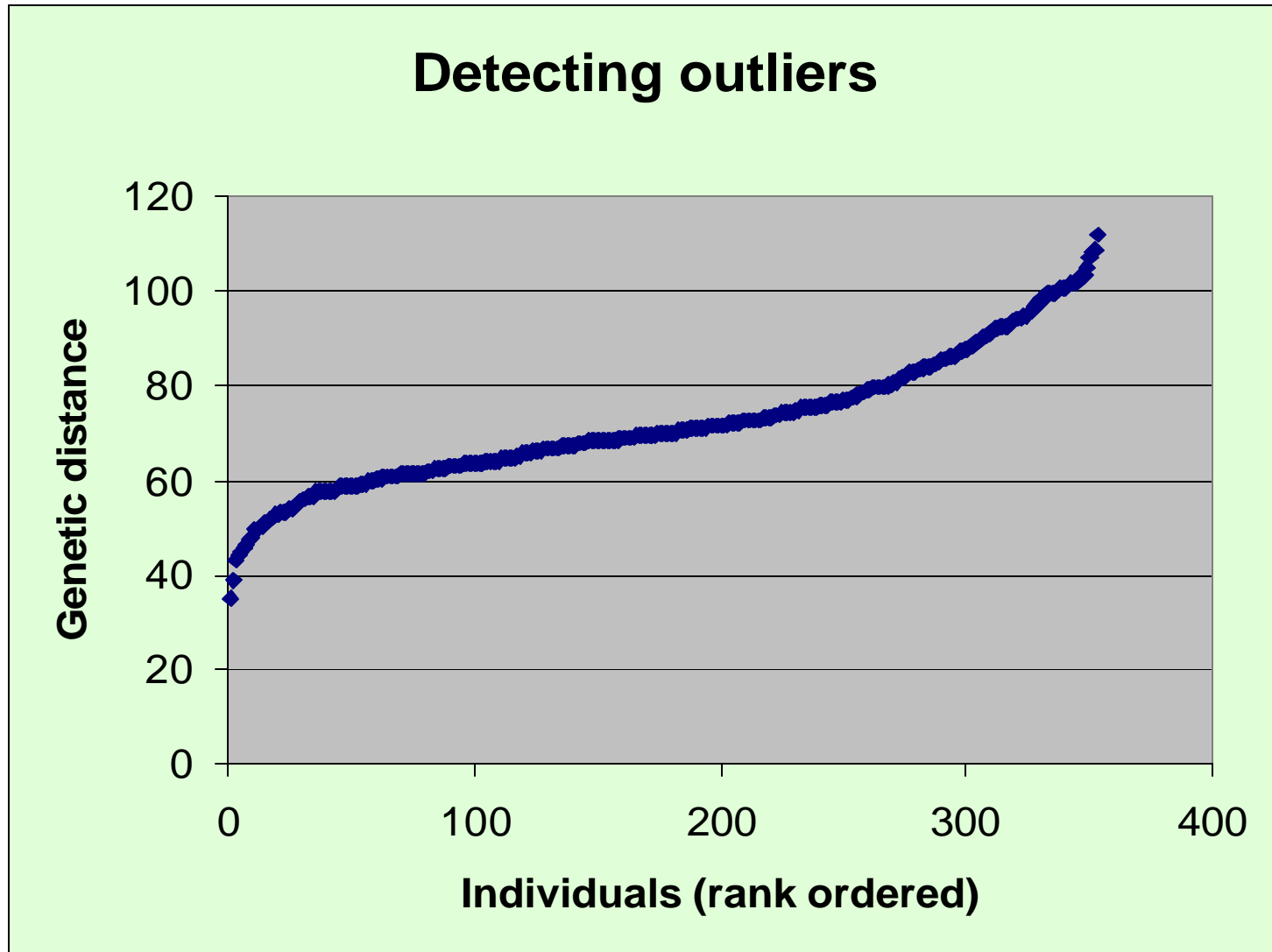
- Stratification within a sample
 - we have assumed sub-populations are distinct
- Admixture within an individual
 - an individual's genome has descended from 2 or more pure sub-populations



Genetic outlier detection

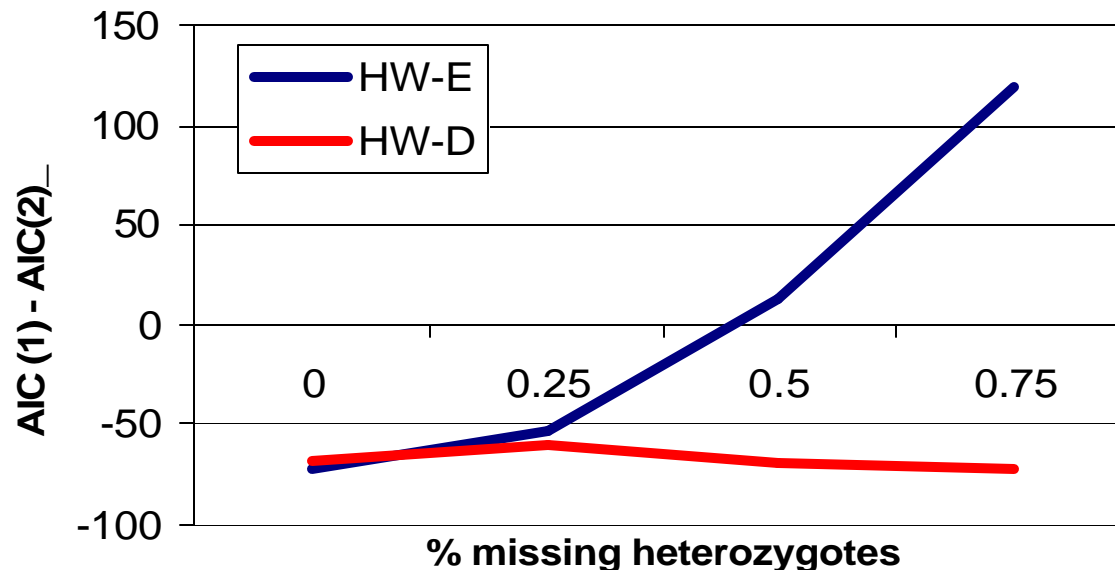
- Identify individuals with unusual genetic makeup
 - sample $\ln L_0$ likelihood for $K = 1$
 - for individual i
 - sample $\ln L_i$ likelihood for $K = 2$
 - with i fixed to class 2
 - all other individuals fixed to class 1
 - $(\ln L_i - \ln L_0)$ is a measure of 'genetic distance'

Genetic outlier detection



Hardy-Weinberg Equilibrium

- Selective genotyping errors for heterozygotes
 - a cause of Hardy-Weinberg disequilibrium
 - spurious stratification?
- Relax within-class HW-E assumption
 - N=400, 40 SNPs, no stratification



Diagnostic output

- Inter-class genetic distance matrix
 - multidimensional scaling to view results
- Entropy measure for individual class assignment

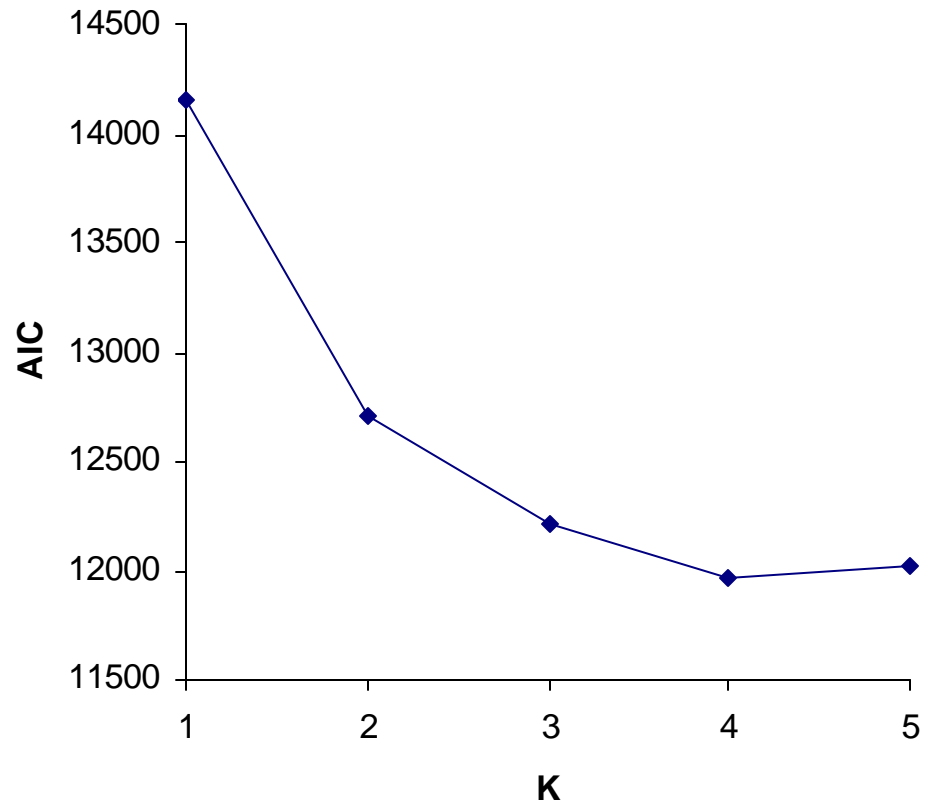
ID	P(C=1)	P(C=2)	P(C=3)	P(C=4)
1	0.01	0	0.99	0
2	0.3	0.1	0	0.6

- Locus-specific inter-class genetic distance
 - handy for identifying linked loci

Satten *et al* data

- Argentinean & native American samples simulated based on known group allele frequencies

- 250 individuals
- 4 subpops (1:1:1:7)
- 12 multi-allelic loci
- “Perfect” solution
 - $K = 4$
 - $P(C) 1:1:1:7$
 - $P(C|G)$ all 1 or 0

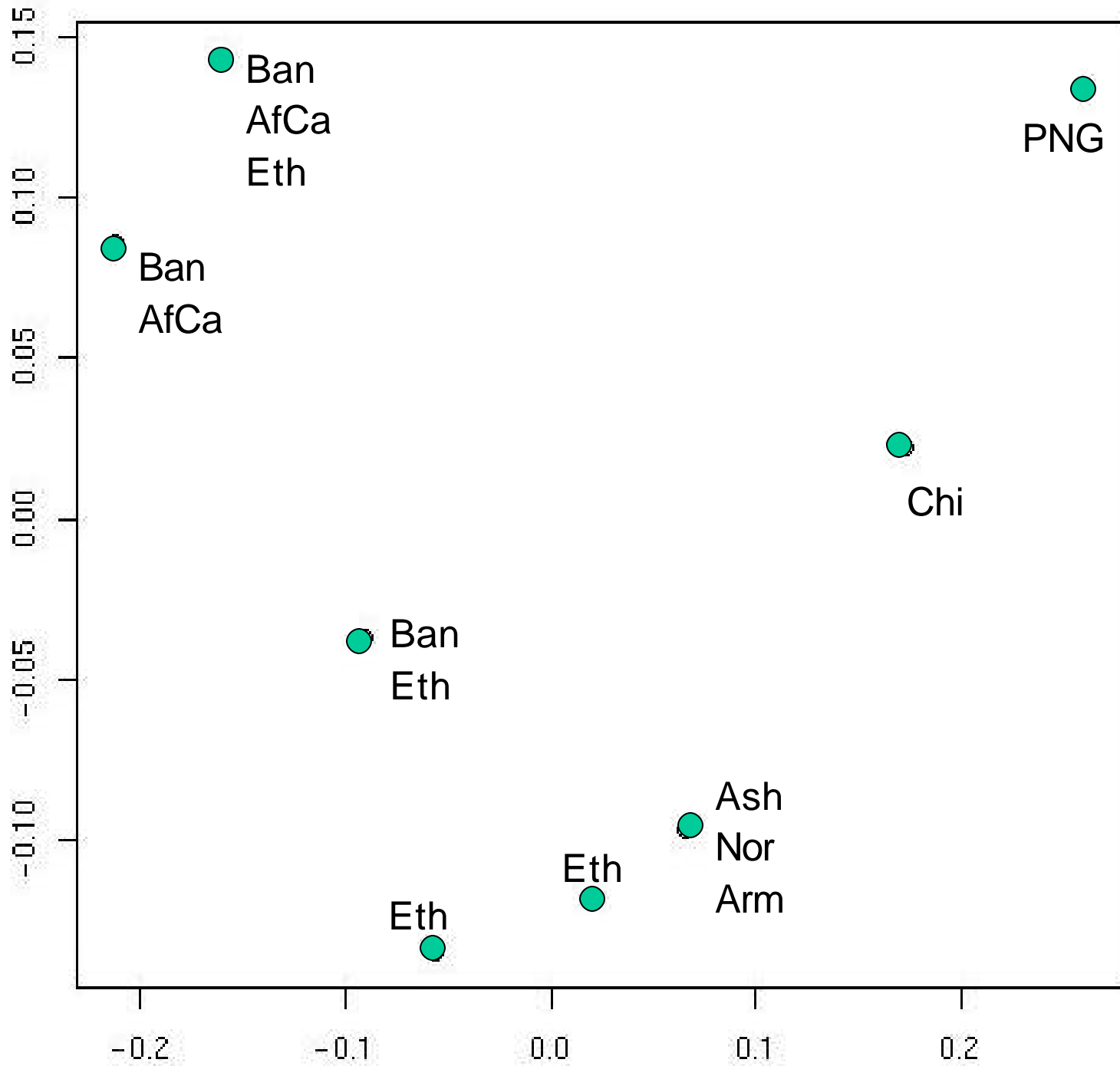


Data application

- Wilson *et al* (2001) Nat Genet
 - Population genetic structure of variable drug response
- 354 individuals; 8 ethnic labels
 - Bantu, Norwegian, Ethiopian, Chinese, Papuan New Guinea, Armenian , Ashkenazi Afro-Caribbean
- Typed on 38 microsatellite marker loci
 - 16 autosomal
 - 22 X chromosome

Wilson *et al* example

- STRUCTURE supports of 4 class solution
 - A: Norwegian, Armenian , Ashkenazi, Ethiopian
 - B: Bantu, Afro-Caribbean, Ethiopian
 - C: Chinese
 - D: Papuan New Guinea
- Ethnic labels such as 'Asian' or 'Black' inadequate



Wilson *et al* example

K = 4 solution

STRUCTURE versus L-POP solutions

	1	4	3	2
4	173	0	1	2
1	5	35	1	6
2	3	3	2	42
3	6	1	74	0

Adjusted RAND index = 0.82

Distribution of P(C|G)

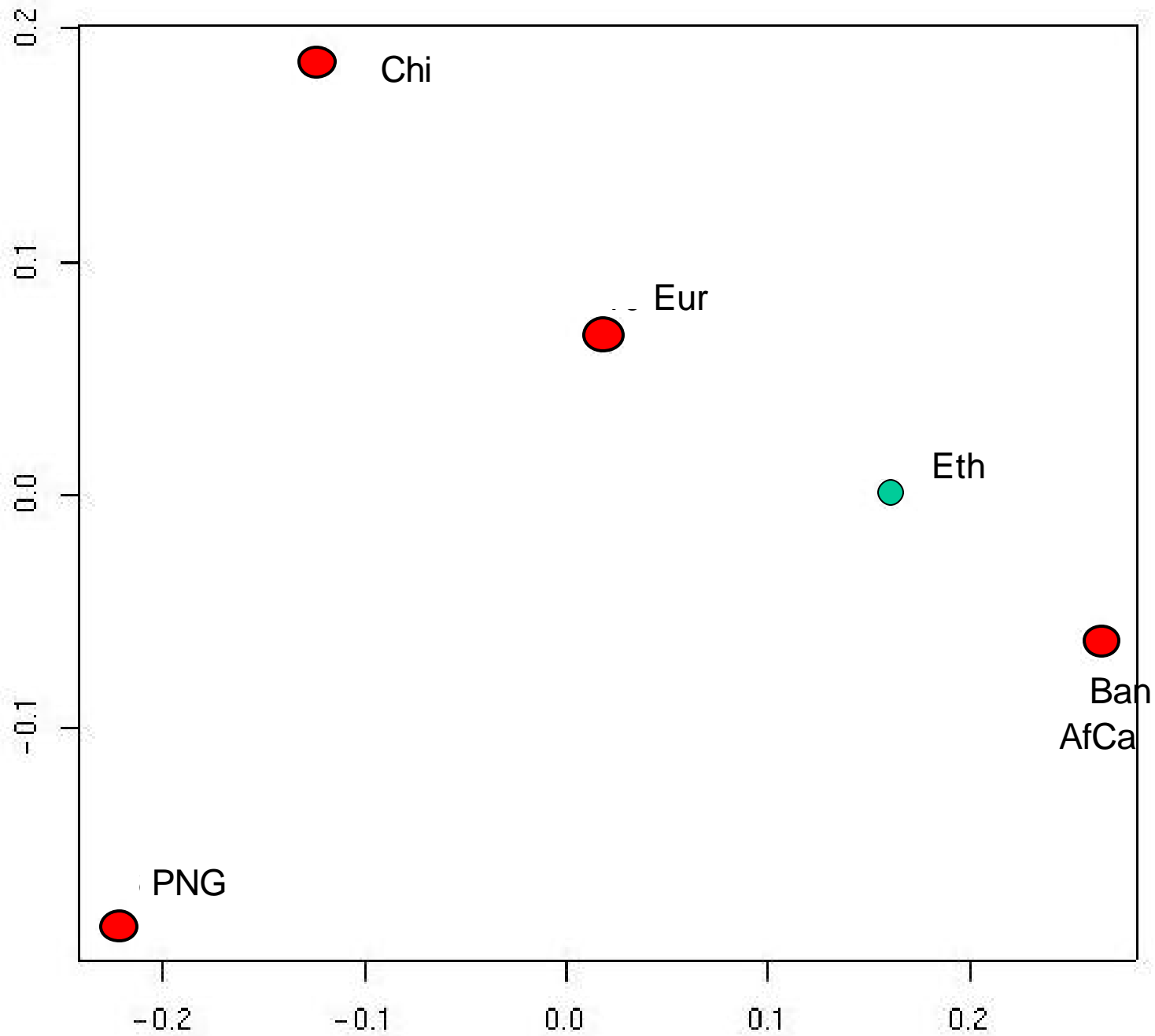


4 class solution

Wilson *et al* example

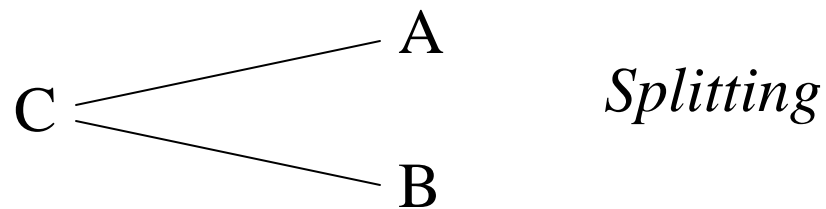
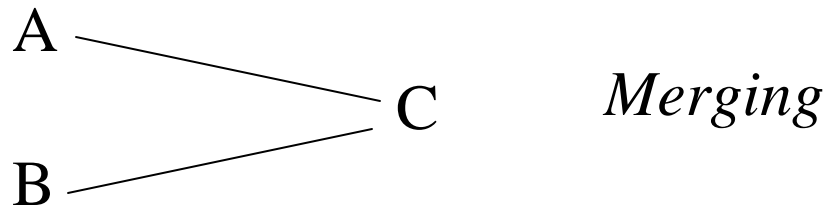
- L-POP best solution
 - 4 classes, allowing for admixture

	Ban	Ash	Eth	Nor	Arm	Chi	PNG	AfCa
A	44		2					22
B					1	36		
C							45	
D		46	15	46	42	1		3
A+D	2		31		1			4



Admixture

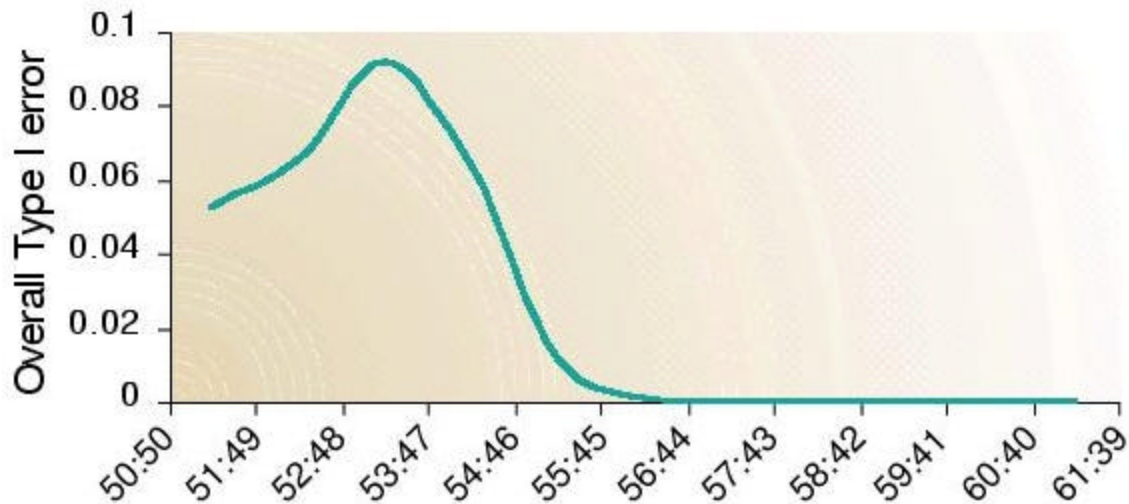
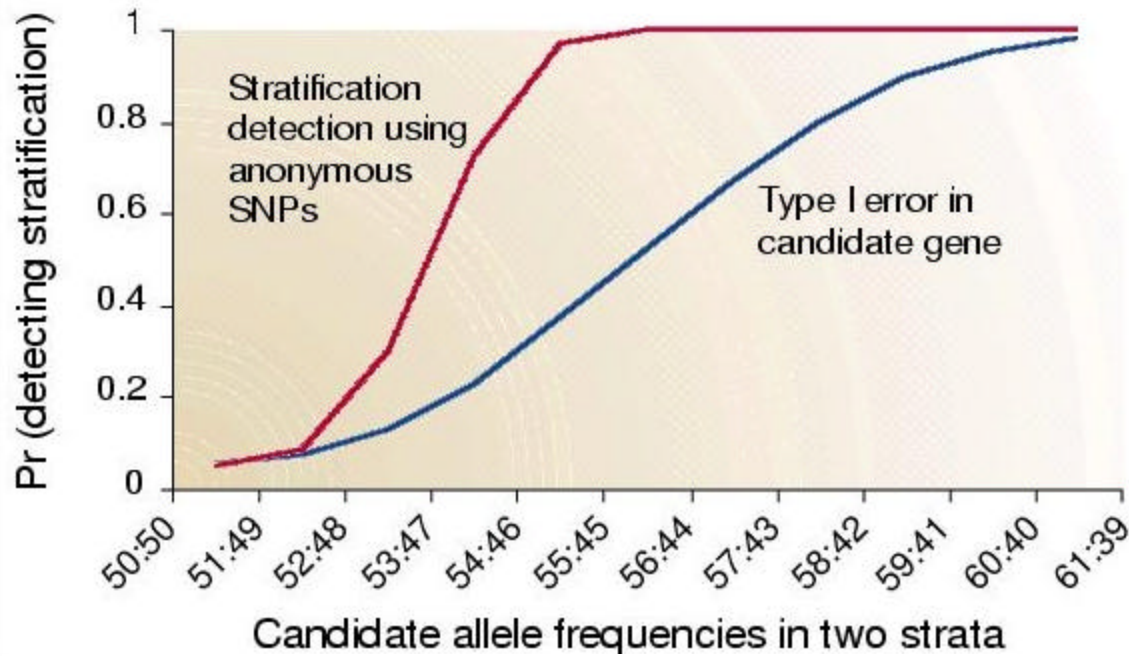
- Only implies that a class has intermediate allele frequencies compared to 2 or more other classes



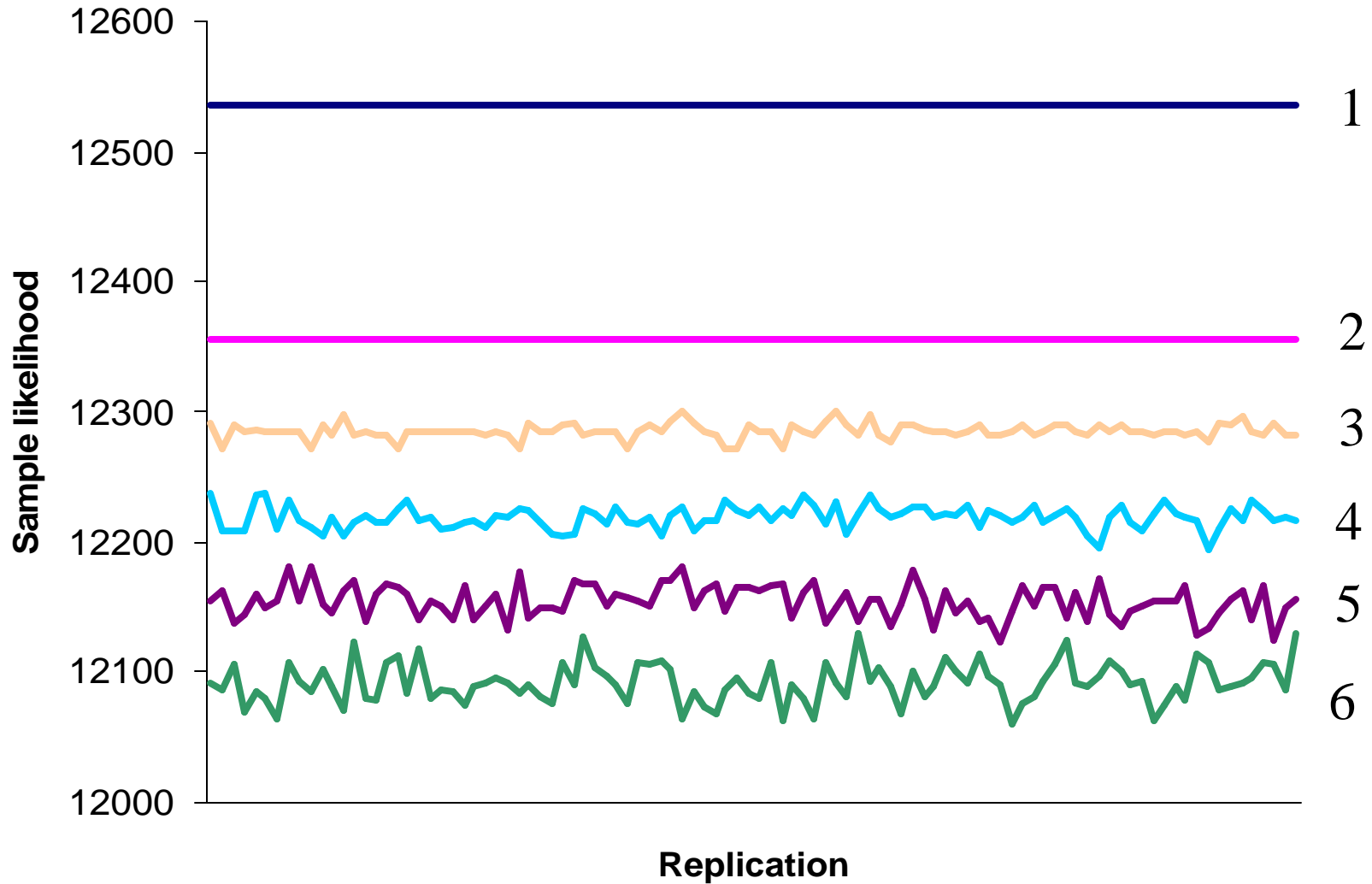
Power issues

- Bacanu *et al* (2000) AJHG
- GC versus TDT
 - in absence of stratification
 - GC more powerful (esp. with common disease)
 - in presence of stratification
 - more complex results: GC works better with low levels of stratification
- Statistical versus economic efficiency
- GC seems to work with as few as 20 loci

Power issues



3 classes



Research Questions

- Detection
 - power issues
 - the nature of population substructure
 - minimum # of SNPs required to detect sub-pops
 - selection of optimal marker sets
 - extend to sibship data
- Correction
 - optimal use of $P(C|G)$ as a covariate
 - effect of “incorrect” K
- **L-POP** is available at
<http://statgen.iop.kcl.ac.uk/lpop/>