

Genetic background and population stratification

Shaun Purcell^{1,2} & Pak Sham¹

¹Social, Genetic & Developmental Psychiatry Research Centre, IoP, KCL, London.

²Whitehead Institute, MIT, Cambridge, MA, USA.

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

$$A_1 \quad p$$

$$A_2 \quad q$$

- Genotype frequencies

- expected under “Hardy-Weinberg equilibrium”

$$A_1A_1 \quad p^2$$

$$A_1A_2 \quad 2pq$$

$$A_2A_2 \quad q^2$$

At a single locus

| | <u>Sub-population</u> | | |
|----------|-----------------------|----------|--------------------|
| | <u>1</u> | <u>2</u> | <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 subpopulations (H_S)
 - in combined 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, between-within model
- 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*

Key references

- Pritchard et al (2000) *Genetics*
 - STRUCTURE method
- Thomas & Witte (2002) *Can Epi, Biomarkers & Prev*
- Wacholder *et al* (2002) *Can Epi, Biomarkers & Prev*
 - Point/Counterpoint: should we worry about stratification
- Cardon & Palmer (2003) *The Lancet*
 - review article
- Rosenberg *et al* (2002) *Science* (2003) *AJHG*
 - Diversity panel; informativeness of markers
- Freedman *et al* (2004) *Nat Genet*
 - empirical assessment of stratification in case/control studies

Genetic background approaches

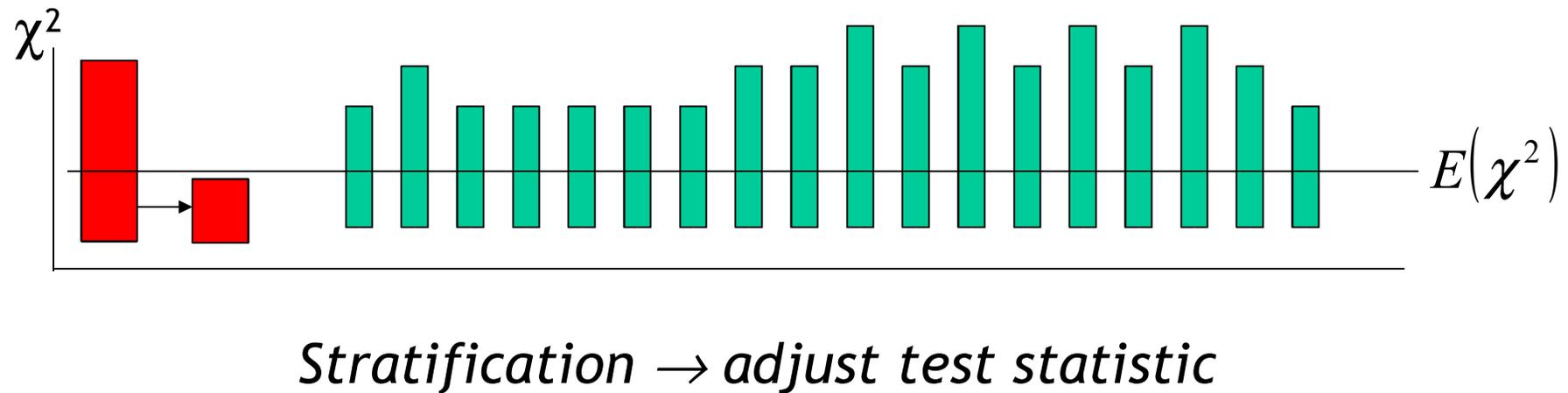
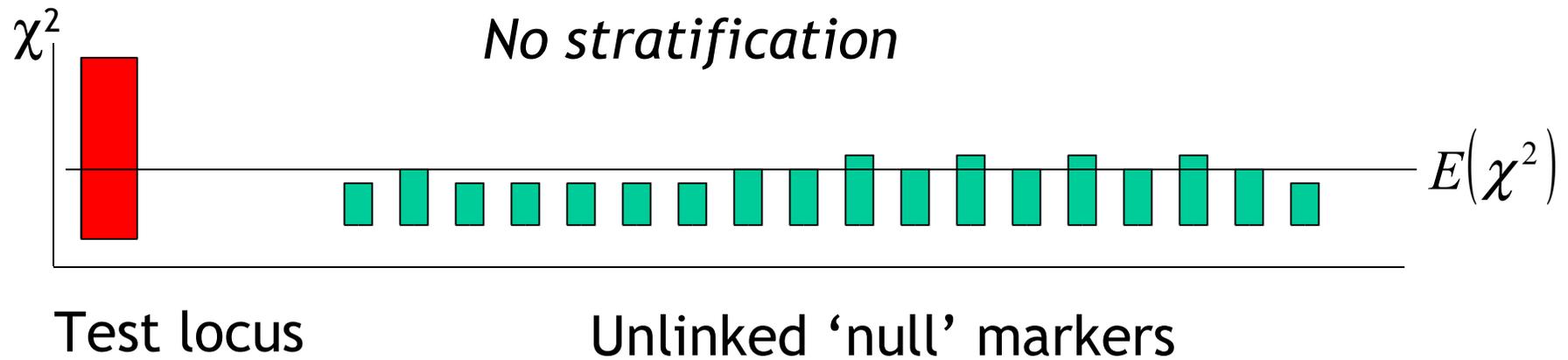
- Genomic Control
- Structured Association
 - Method: multilocus genotype data to detect and correct for stratification
 - Premise: stratification operates globally – on whole genome, whereas LD operates locally at short scales

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, $T_N / \lambda \sim \chi^2_1$
 - estimate λ
 - statistic at test locus $T / \lambda \sim \chi^2_1$

Genomic control



Genomic control

- Simple estimate of inflation factor

$$\hat{\lambda} = \text{median}\{\chi_1^2, \chi_2^2, \dots, \chi_N^2\} / 0.456$$

- using the median protects from outliers
 - i.e. if some of the null markers are also QTL
- bounded at minimum of 1
 - i.e. should never increase test statistic
- principle extended to multiple alleles, haplotypes, quantitative traits
 - Must formulate all tests as 1 df tests, however

Genomic control

λ Inflation factor $\lambda \approx 1 + RF \sum_k (f_k - g_k)^2$

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 equipotent subpopulations, $F_{ST} = 0.01$
- Disease twice as common in one subpopulation
- $R = 1000$

$\lambda \approx 1.5$

Structured association

- Assignment of individuals to subpopulations
 - Test for association conditional on subpopulation
- Distance-based approaches
- Model-based approaches
 - Pritchard *et al* (2000) *Genetics*
 - Bayesian framework (STRUCTURE / STRAT)
 - Satten *et al* (2001) *AJHG*
 - Latent class analysis model
 - Purcell & Sham (2005) *Hum Hered*
 - Latent class analysis model (L-POP / L-ASSOC)

IBS distance-based metric

Pair from same population

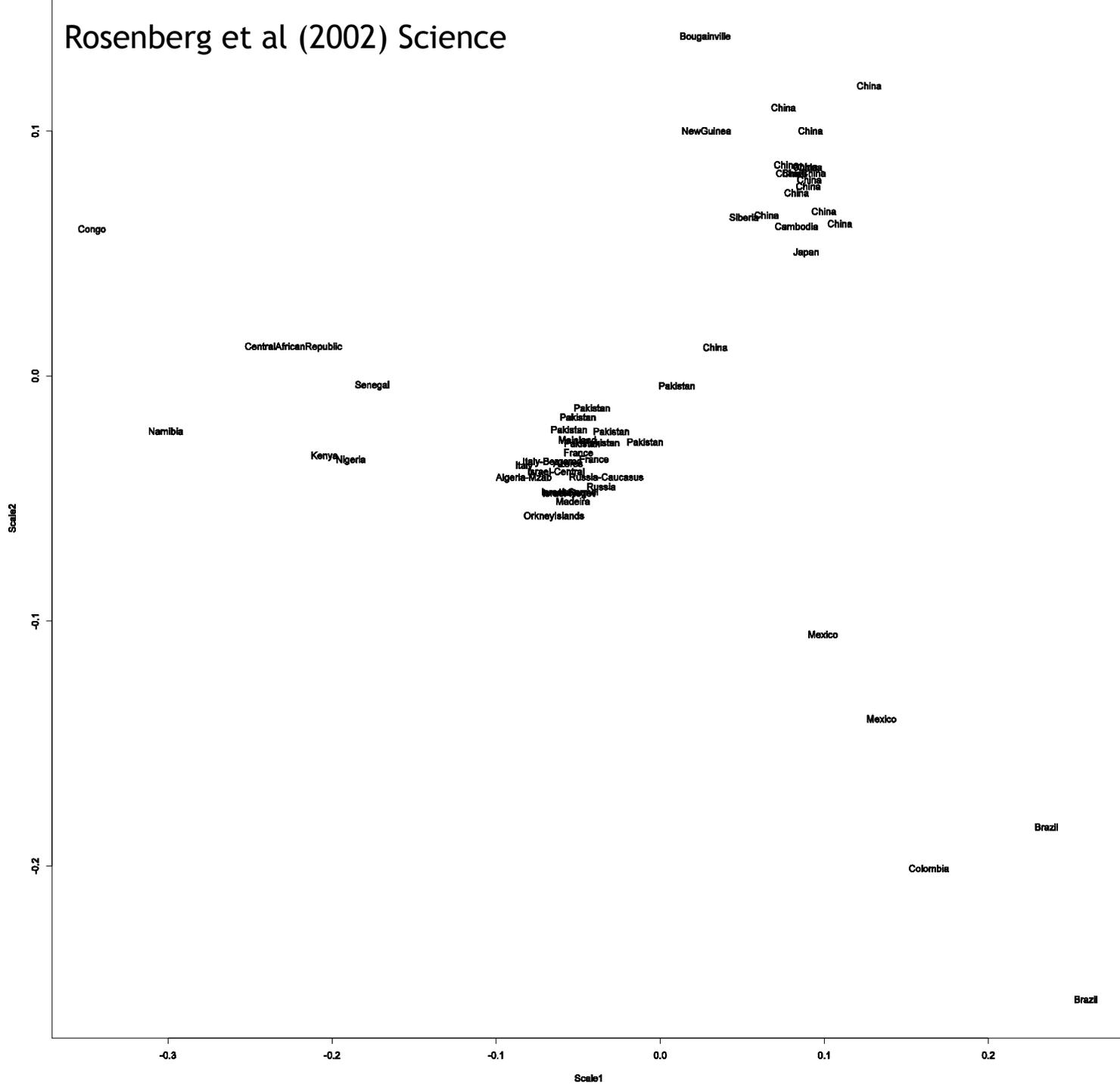
| | | | | | | |
|-----|-----|-----|-----|-----|-----|-----|
| | G/J | K/U | H/H | W/W | S/S | ... |
| | G/G | K/K | H/H | U/B | S/S | ... |
| IBS | 1 | 1 | 2 | 0 | 2 | ... |

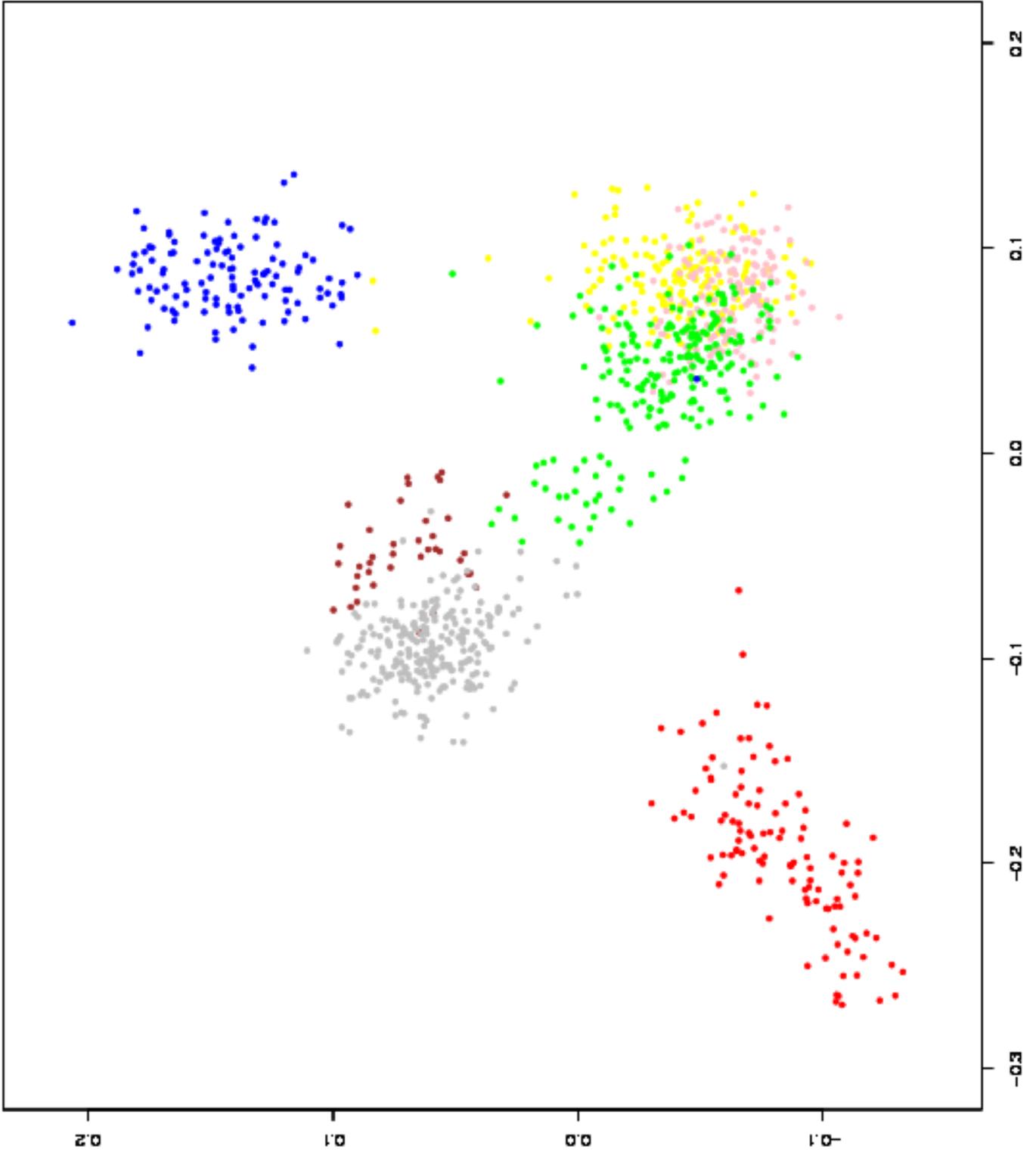
Pair from different population

| | | | | | | |
|-----|-----|-----|-----|-----|-----|-----|
| | G/J | K/U | H/H | W/W | S/S | ... |
| | K/J | L/A | G/D | U/B | S/L | ... |
| IBS | 1 | 0 | 0 | 0 | 1 | ... |

1. Distance matrix (applying clustering or multidimensional scaling)
2. Permutation test: do these two individuals come from different populations?

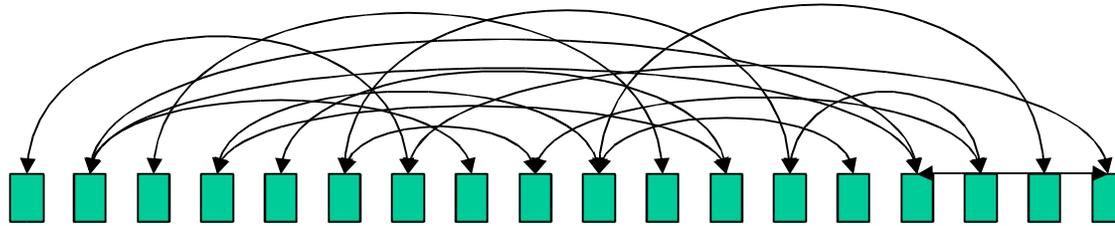
Rosenberg et al (2002) Science



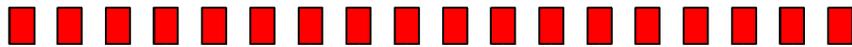
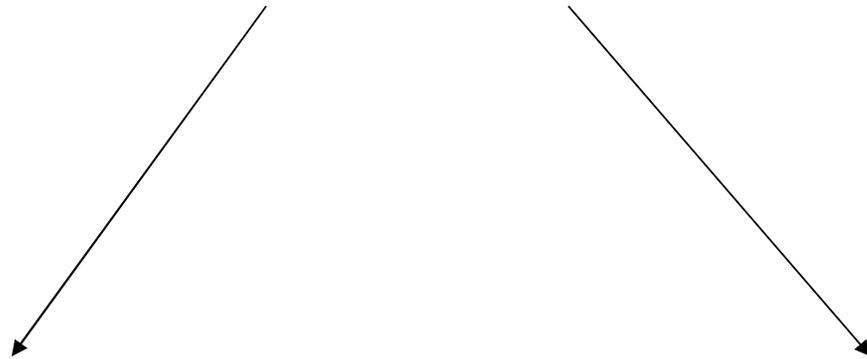


Structured association

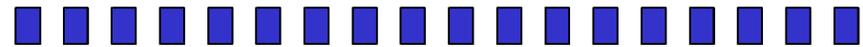
LD observed under stratification



Unlinked 'null' markers



Subpopulation A



Subpopulation B

Advantages of SA

- Structure of intrinsic interest
- Any test of association can be used
- Allows allelic heterogeneity between subpopulations
- Does not assume constant F_{ST} 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

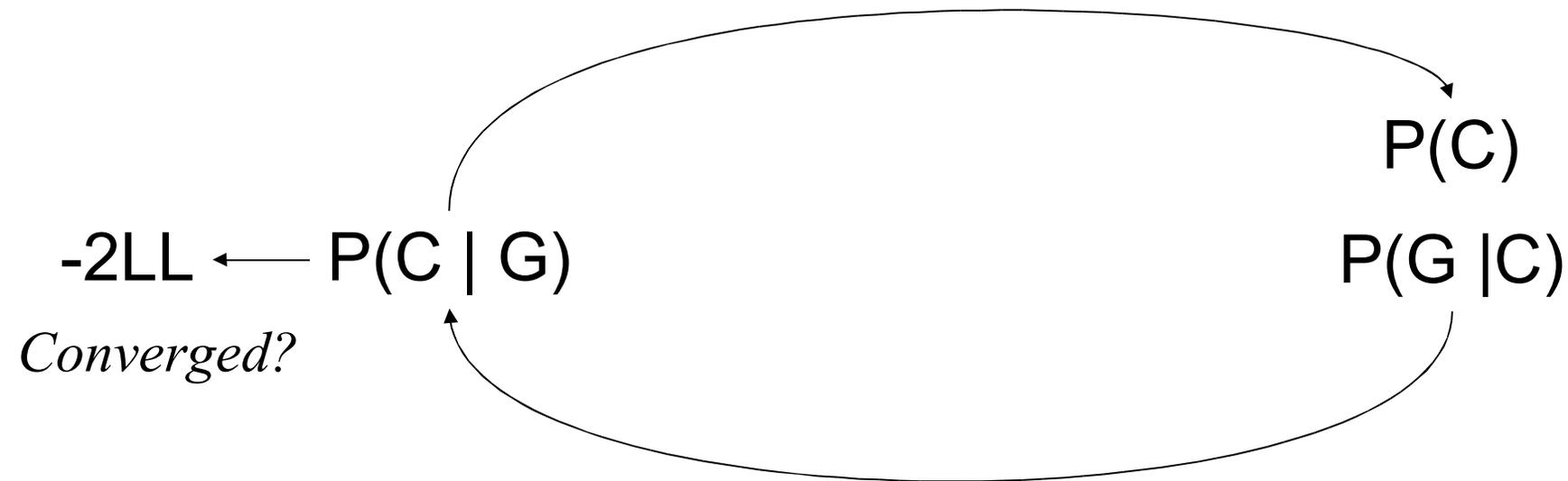
$P(C)$ prior probabilities

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

E-M algorithm

E step:

counting individuals and alleles in classes



M step:

Bayes theorem, assume conditional independence

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

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

Product over $l = 1$ to L loci

Likelihood

- Likelihood of an individual

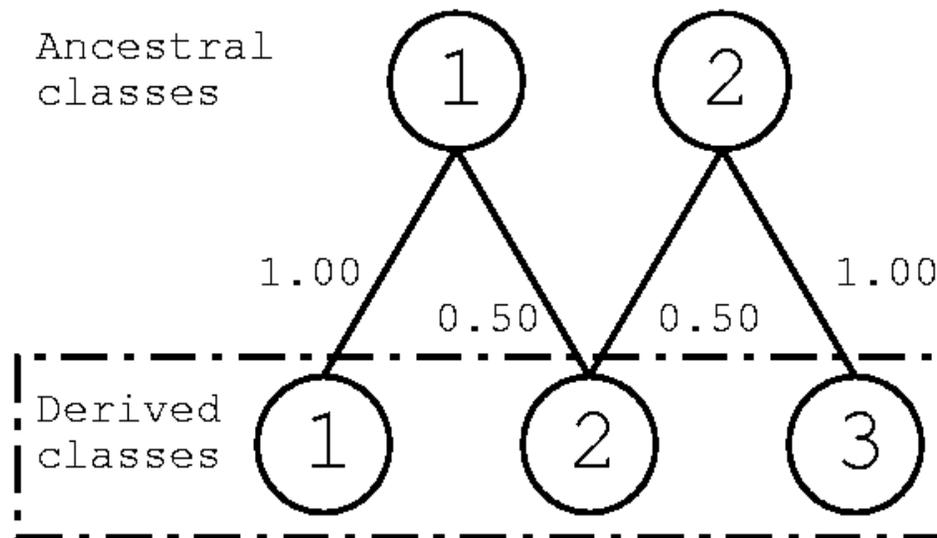
$$L_i = \sum_j P(G|C)P(C)$$

- Use AIC to select optimal K solution

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

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



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 |

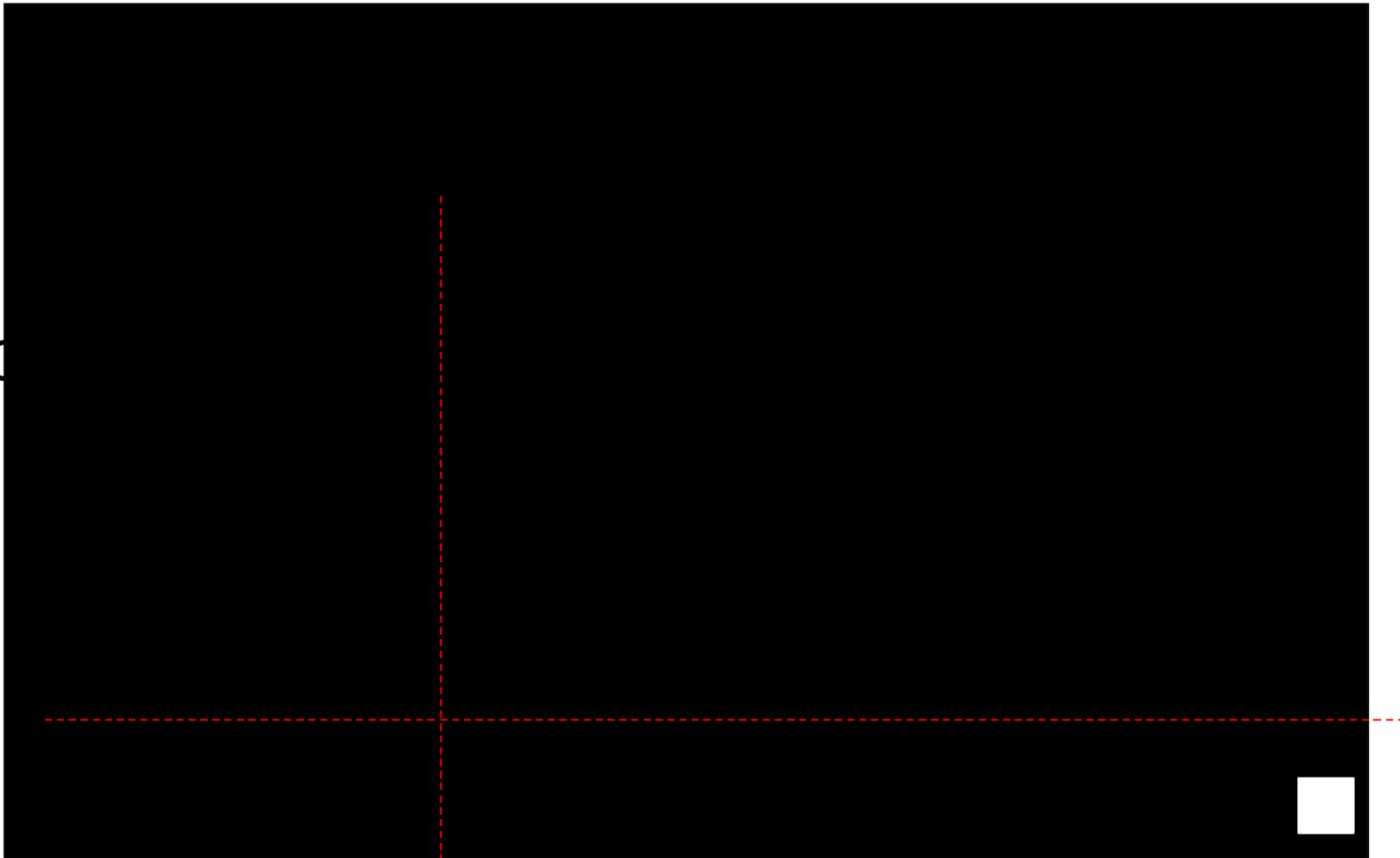
Example #1

| | $P(C=1 G)$ | $P(C=2 G)$ |
|-----|--------------|--------------|
| 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 |

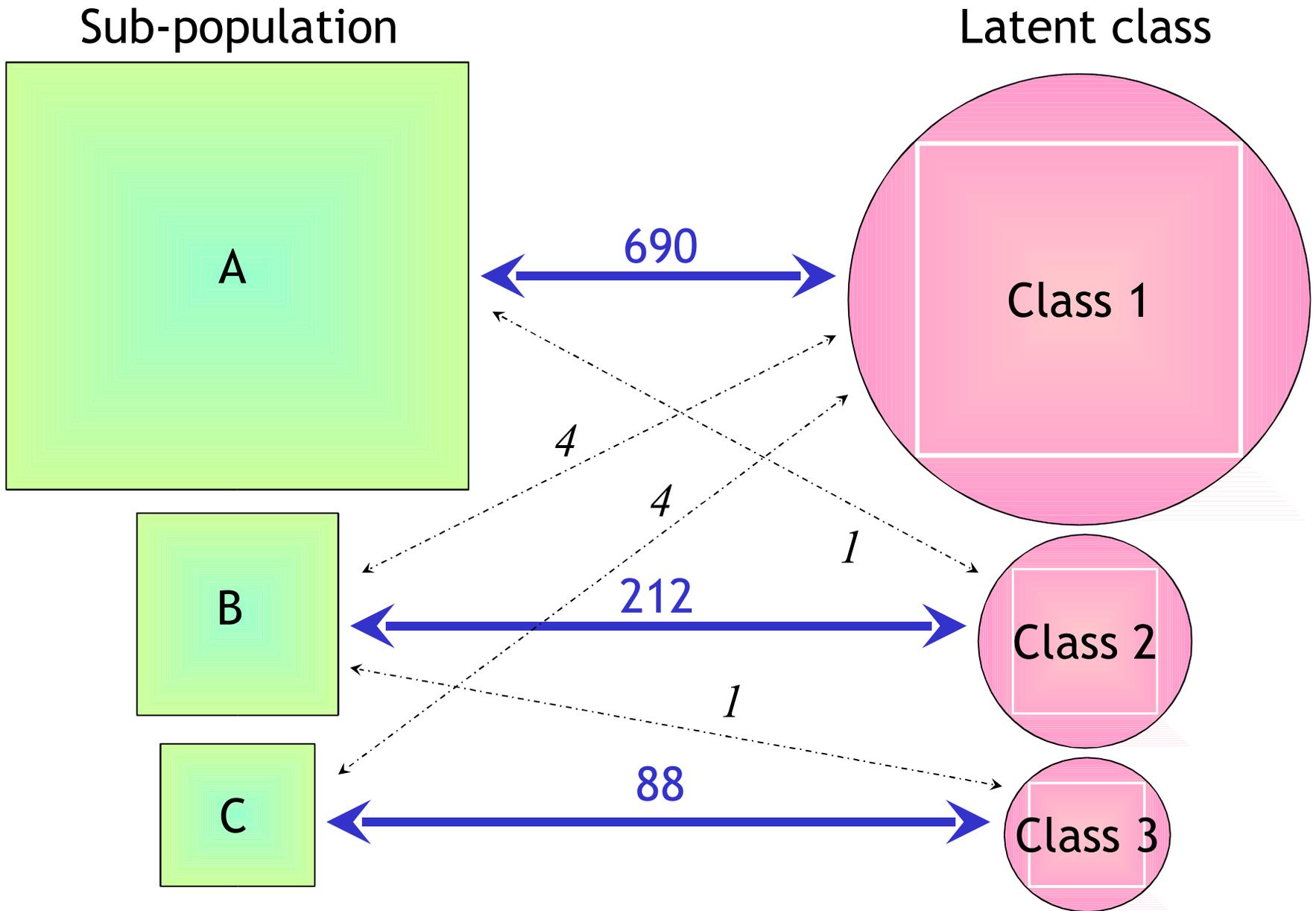
Example #2

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

AIC



$K=3$



Factors affecting power

- Number of markers
- Selection of most informative markers vs random
 - ancestry informative markers (AIMs)

Big impact on
type I error

Small impact on
type I error rates

**Extreme
substructure**

**Very mild
substructure**

Easy to detect

Hard to detect

Notes on L-POP

- Example parameter file (<http://statgen.iop.kcl.ac.uk/lpop/>)

| | | |
|------------------------|---|-------------------------------|
| Example parameter file | ← | 1 st line is title |
| DATAFILE mydata.raw | ← | required |
| STRUCTURE | ← | file format |
| PHENO 4 | ← | # cols to skip |
| CLASS 2 | ← | model specification |
| TAG c12 | ← | Name tag for results |
| RAND 0 | ← | Random # seed |
| REPEAT 10 | ← | # attempts at convergence |
| VERBOSE2 | ← | Verbosity of output (1-3) |

Results format for L-POP

| | |
|---------------------|---|
| grep P: results | get prior class probabilities |
| grep K: results | get likelihood, AIC |
| grep k: results | get likelihood, AIC from all E-M convergences |
| grep I: results | get posterior class probabilities |
| grep D: results | get genetic distance matrix |
| grep I:c13: results | get $P(C G)$ for solution with TAG c13 only |

Correction

- 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)

Testing for association

- Weighted likelihood
- Model probability of genotype conditional on trait

$$\sum_C L(G|X, C)P(C)$$

Class-specific likelihood
of genotype conditional
on trait

Individual's class probabilities
(estimated using L-POP)

$$L(G|X, C) = \frac{L(X|G, C)L(G|C)}{\sum_G L(X|G, C)L(G|C)}$$

Parameters p, a, d
(potentially class-specific)

Notes on L-ASSOC

Data :

Individuals only, quantitative trait
.ped file and .dat file
weights as covariates (C in .dat file)

Parameters :

used to build alt and null models

| | Universal | Class-specific |
|-------------------------|-----------|----------------|
| Allele frequency: | p | P |
| Additive genetic value: | a | A |
| Dominance deviation: | d | D |

Notes on L-ASSOC

Standard test of association

```
lassoc --file data --alt pa --null p
```

Test of association allowing for stratification

```
lassoc --file data --alt Pa --null P
```

Test of allele frequency differences between strata

```
lassoc --file data --alt P --null p
```

Test of QTL by strata interaction

```
lassoc --file data --alt PA --null Pa
```

Test of all effects

```
lassoc --file data --alt PAD --null P
```

lassoc --file data --alt pa --null p

| Model | SP | p | a | d | va | vd |
|----------|---------|-------|-------|---|-------|----|
| ----- | | | | | | |
| H1 | 1 | 0.498 | 0.020 | | 0.005 | |
| | 2 | 0.498 | 0.020 | | 0.005 | |
| | 3 | 0.498 | 0.020 | | 0.005 | |
| | | | | | | |
| H0 | 1 | 0.498 | | | | |
| | 2 | 0.498 | | | | |
| | 3 | 0.498 | | | | |
| ----- | | | | | | |
| -2LL(H1) | 209.839 | | | | | |
| -2LL(H0) | 216.029 | | | | | |
| LRT | 6.190 | | | | | |
| df | 1 | | | | | |
| p-value | 0.013 | | | | | |
| ----- | | | | | | |

lassoc --file data --alt Pa --null P

| Model | SP | p | a | d | va | vd |
|----------|---------|-------|-------|---|-------|----|
| ----- | | | | | | |
| H1 | 1 | 0.624 | 0.017 | | 0.004 | |
| | 2 | 0.443 | 0.017 | | 0.004 | |
| | 3 | 0.502 | 0.017 | | 0.004 | |
| | | | | | | |
| H0 | 1 | 0.622 | | | | |
| | 2 | 0.446 | | | | |
| | 3 | 0.508 | | | | |
| ----- | | | | | | |
| -2LL(H1) | 209.839 | | | | | |
| -2LL(H0) | 216.029 | | | | | |
| LRT | 1.190 | | | | | |
| df | 1 | | | | | |
| p-value | 0.734 | | | | | |
| ----- | | | | | | |

Practical session

- Goal
 - using QTDT, LPOP and LASSOC, analyse the data under the pshaun/strat/ directory
 - 1. For the two SNP test markers, what does standard association analysis reveal?
 - 2. Is there evidence for population substructure?
 - 3. What is the effect of testing for association conditional on any substructure, using family-based tests?

(I) Individuals

QTDT

dind.ped, dind.dat

Collect siblings

Type 50 null loci

(II) Family-based analysis

QTDT

dfam.ped, dfam.dat

(III) GC / SA analysis

LPOP

dnull.ped

Generate weights

(IV) SA

QTDT

dcov.ped, dcov.dat

(V) GC

LASSOC

dnull.ped, dnull.dat

LASSOC

dcov.ped, dcov.dat

dind.ped

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|--------|
| 1 | 1 | 0 | 0 | 1 | 1 | 1 | 1 | 2 | 1.576 |
| 2 | 1 | 0 | 0 | 1 | 1 | 2 | 1 | 1 | 0.368 |
| 3 | 1 | 0 | 0 | 1 | 2 | 1 | 1 | 1 | -0.423 |

PED details

QTL

Trait

dfam.ped

| | | | | | | | | | |
|---|---|---|---|---|----|----|----|----|-------|
| 1 | 3 | 0 | 0 | 1 | -9 | -9 | -9 | -9 | -9 |
| 1 | 4 | 0 | 0 | 1 | -9 | -9 | -9 | -9 | -9 |
| 1 | 1 | 3 | 4 | 1 | 1 | 1 | 1 | 2 | 1.576 |
| 1 | 2 | 3 | 4 | 1 | 1 | 2 | 1 | 2 | 1.576 |

“Parents”

Siblings

dnull.ped

| | | | | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|--------|---|---|---|---|---|---|---|---|---|---|---|---|-----|---|-----|
| 1 | 1 | 0 | 0 | 1 | 1 | 1 | 1 | 2 | 1.576 | 1 | 1 | 1 | 2 | 2 | 1 | 2 | 2 | 1 | 2 | 2 | 1 | ... | | |
| 2 | 1 | 0 | 0 | 1 | 1 | 2 | 1 | 1 | 0.368 | 1 | 2 | 1 | 1 | 2 | 1 | 2 | 2 | 2 | 1 | 2 | 1 | 2 | 1 | ... |
| 3 | 1 | 0 | 0 | 1 | 2 | 1 | 1 | 1 | -0.423 | 2 | 1 | 1 | 1 | 1 | 2 | 1 | 1 | 2 | 1 | 1 | 2 | 1 | 1 | ... |

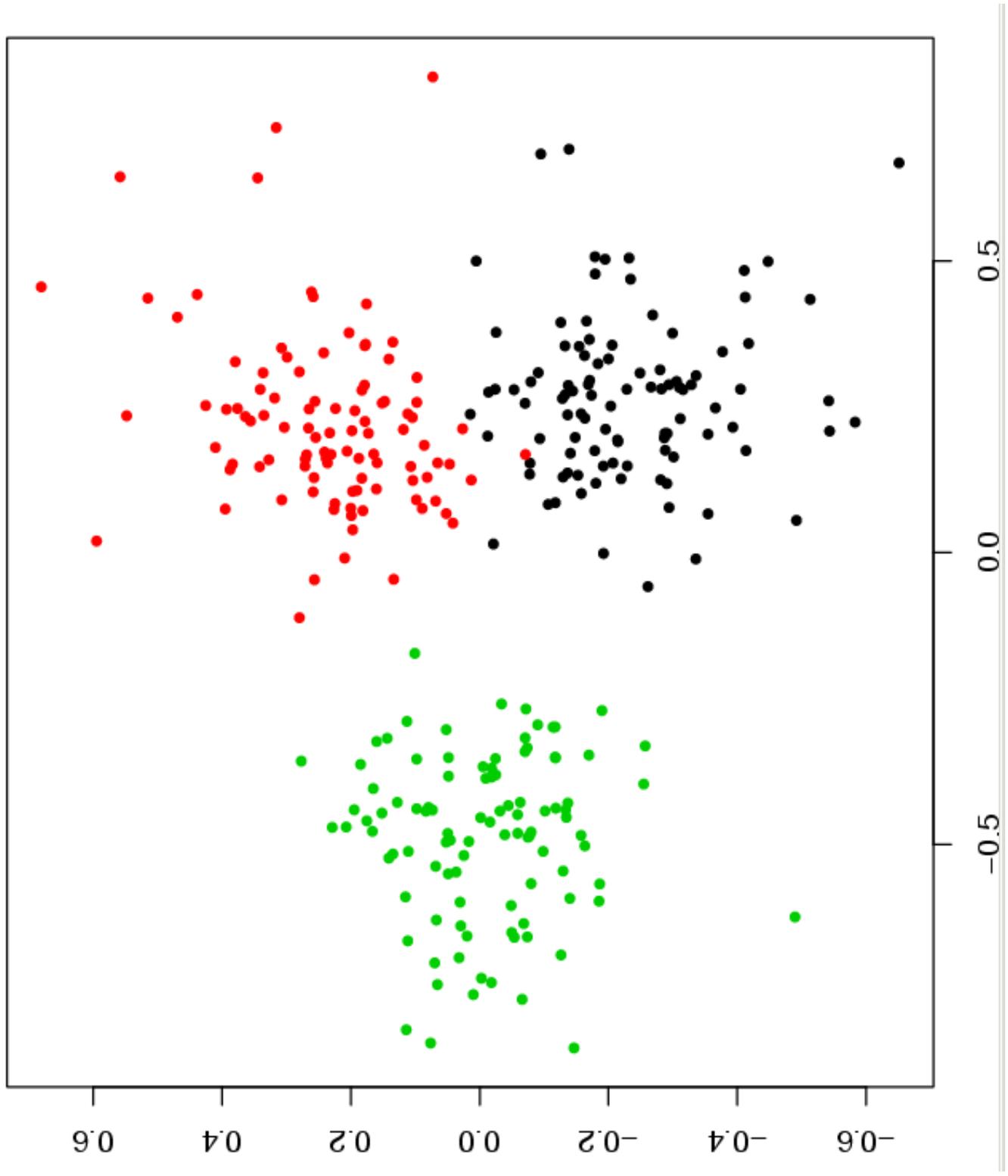
PED details, QTL & trait

Null markers

dcov.ped

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|--------|-------|-------|-------|
| 1 | 1 | 0 | 0 | 1 | 1 | 1 | 1 | 2 | 1.576 | 0.000 | 0.000 | 1.000 |
| 2 | 1 | 0 | 0 | 1 | 1 | 2 | 1 | 1 | 0.368 | 0.000 | 0.150 | 0.850 |
| 3 | 1 | 0 | 0 | 1 | 2 | 1 | 1 | 1 | -0.423 | 0.998 | 0.001 | 0.001 |

Posterior probabilities (LPOP)



Standard QTDT analysis (not controlling for stratification)

```
qtdt -p dind.ped -d dind.dat -at -weg
```

Family-based QTDT analysis (not controlling for stratification)

```
qtdt -p dfam.ped -d dfam.dat -at -weg
```

Family-based QTDT analysis (within test, controlling for stratification)

```
qtdt -p dfam.ped -d dfam.dat -ao -weg
```

Family-based QTDT analysis (test of stratification)

```
qtdt -p dfam.ped -d dfam.dat -ap -weg
```

L-POP stratification analysis

```
lpop < param1 > results
```

```
lpop < param2 >> results
```

```
lpop < param3 >> results
```

```
lpop < param4 >> results
```

Get lowest AIC

```
grep AIC results
```

Get prior class probabilities for 3 class solution (TAG c13)

```
grep P:c13: results
```

Get posterior probabilities from the 3 class solution

```
grep I:c13: results
```

```
grep I:c13: results | gawk '{print $4,$5,$6}' > postprob
```

QTDT analysis, using covariates

```
qtdt -p dcov.ped -d dcov.dat -at -weg
```

LASSOC analysis, not controlling

```
lassoc --file dcov --alt pa --null p
```

LASSOC analysis, controlling stratification

```
lassoc --file dcov --alt Pa --null P
```

LASSOC analysis, testing for stratification

```
lassoc --file dcov --alt P --null p
```

LASSOC analysis, allowing for QTL x strata interaction

```
lassoc --file dcov --alt PA --null P
```

LASSOC analysis of all null loci

```
lassoc --file dnull --alt pa --null p
```

Get median test statistic, divide by 0.456, use to correct QTL tests
e.g. using grep to extract test statistics efficiently

```
lassoc --file dnull --alt pa --null p > gcresults  
grep LRT gcresults
```