# Genetic background and population stratification

*Shaun Purcell*[1,2] *& Pak Sham*[1]

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

[2] *Whitehead Institute, MIT, Cambrdige, 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

| | |
|---|---|
| $A_1$ | $p$ |
| $A_2$ | $q$ |

- Genotype frequencies
  - expected under “Hardy-Weinberg equilibrium”

| | |
|---|---|
| $A_1A_1$ | $p^2$ |
| $A_1A_2$ | $2pq$ |
| $A_2A_2$ | $q^2$ |

# At a single locus

|  | Sub-population<br>1 | <br>2 | <br>1+2 |
|---|---|---|---|
| $A_1$ | 0.1 | 0.9 | *0.5* |
| $A_2$ | 0.9 | 0.1 | *0.5* |
|  |  |  |  |
| $A_1A_1$ | 0.01 | 0.81 | *0.41 (0.25)* |
| $A_1A_2$ | 0.18 | 0.18 | *0.18 (0.50)* |
| $A_2A_2$ | 0.81 | 0.01 | *0.41 (0.25)* |

# 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, 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*

# Genetic background approaches

- Genomic Control
- Structured Association

  - Method: multilocus genotype data to detect and correct for stratification

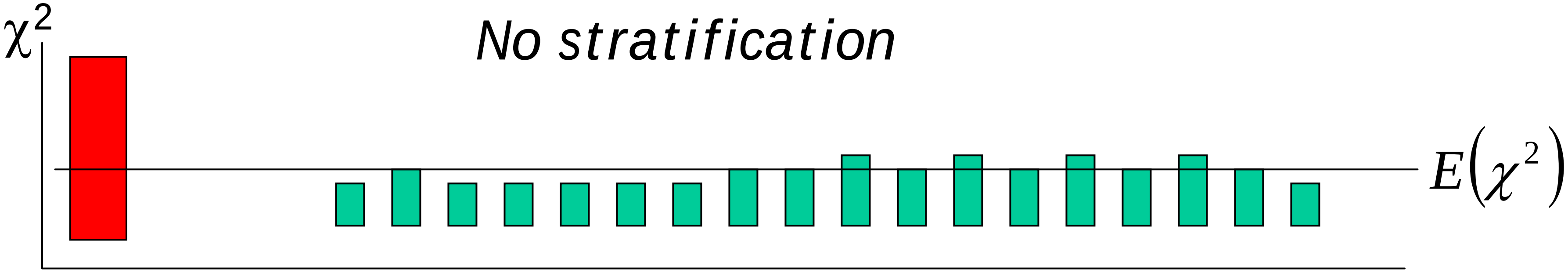
  - Premise: stratification operates globally – on whole genome, whereas LD operates locally at short scales

# Genomic control

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

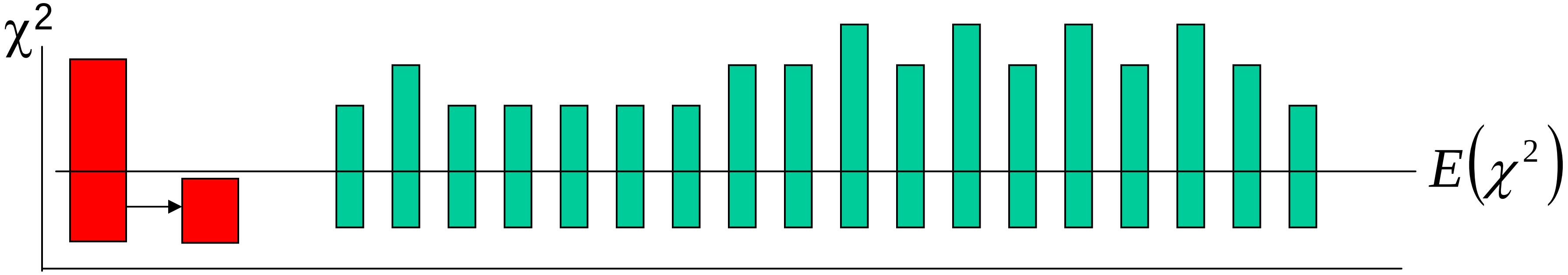
# Genomic control

*No stratification*



Test locus

Unlinked ‘null’ markers



*Stratification → adjust test statistic*

# Genomic control

- Simple estimate of inflation factor

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

  - 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, haplotpes, 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 equifrequent subpopulations, $F_{ST} = 0.01$
  - Disease twice as common in one subpopulation
  - R = 1000
  - **λ ≈ 1.5**

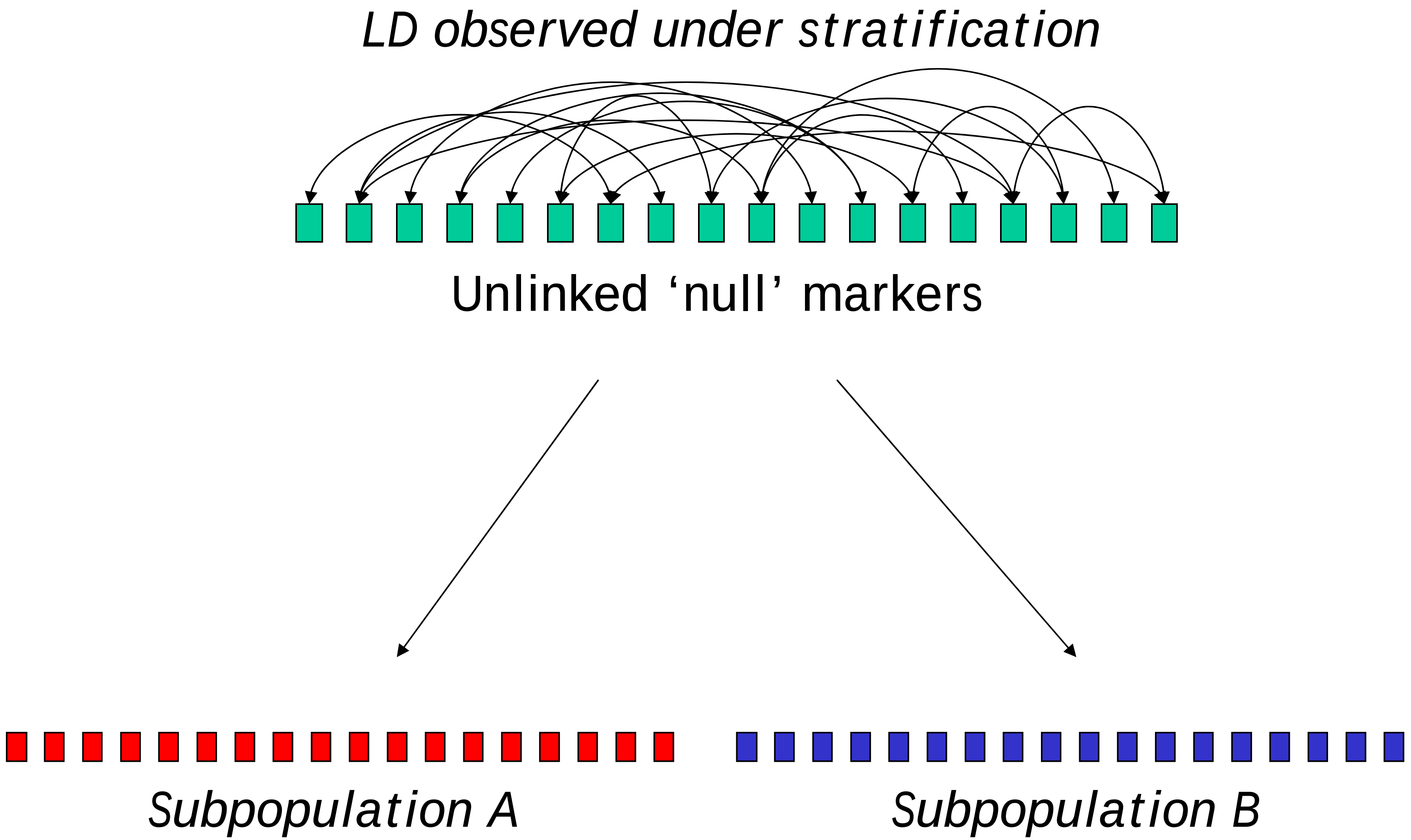
# Structured association

- Assignment of individuals to subpopulations
  - Test for association conditional on subpopulation

- Distance-based approaches

- Model-based approaches
  - Pritchard *et al* (2000)
    - Bayesian framework (STRUCTURE / STRAT)
  - Satten *et al* (2001)
    - Latent class analysis model
  - Purcell & Sham
    - Latent class analysis model (L-POP / L-ASSOC)

# 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**

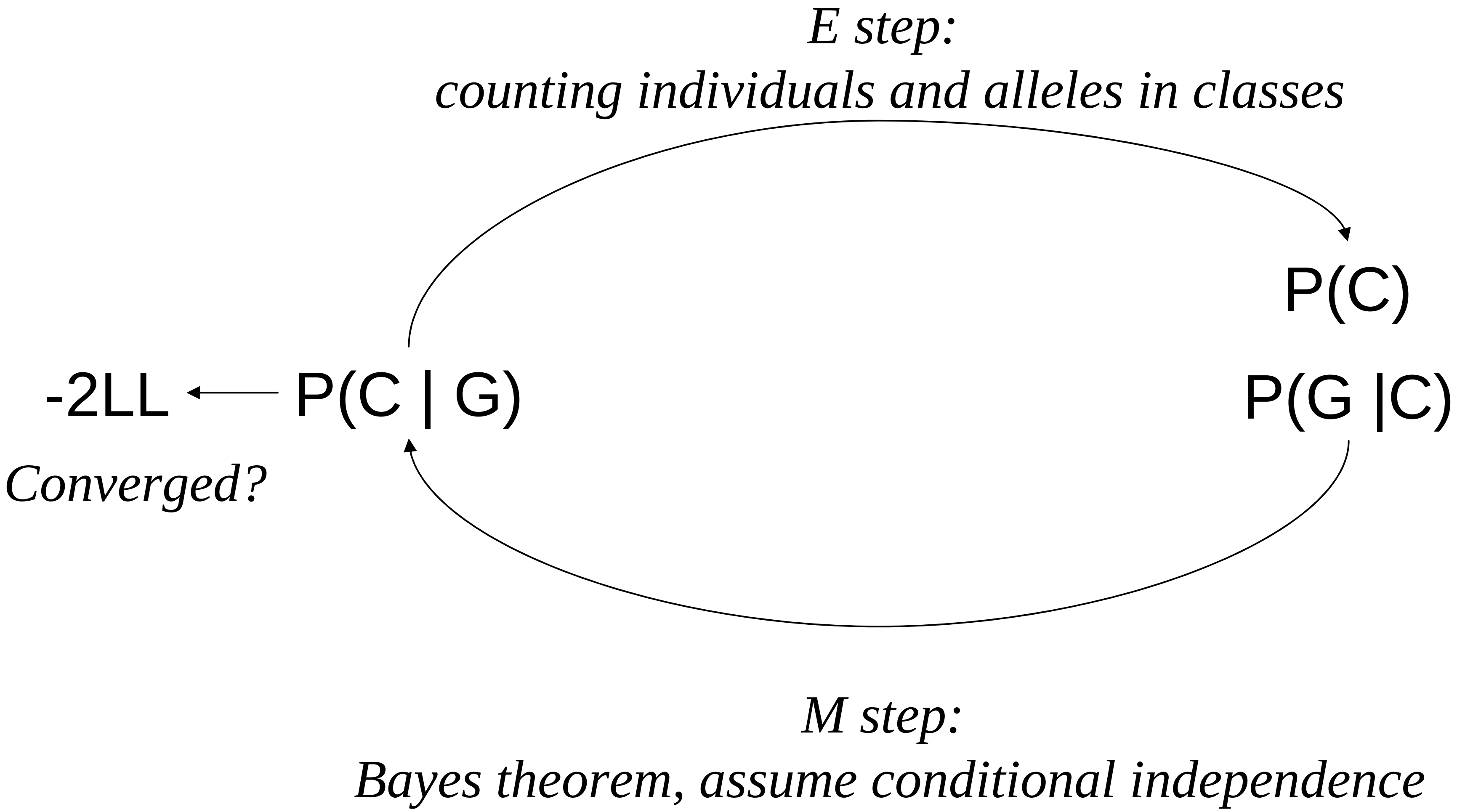
# Latent Class Analysis

- **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*

P(C)

P(G |C)

-2LL ← P(C | G)

*Converged?*

*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

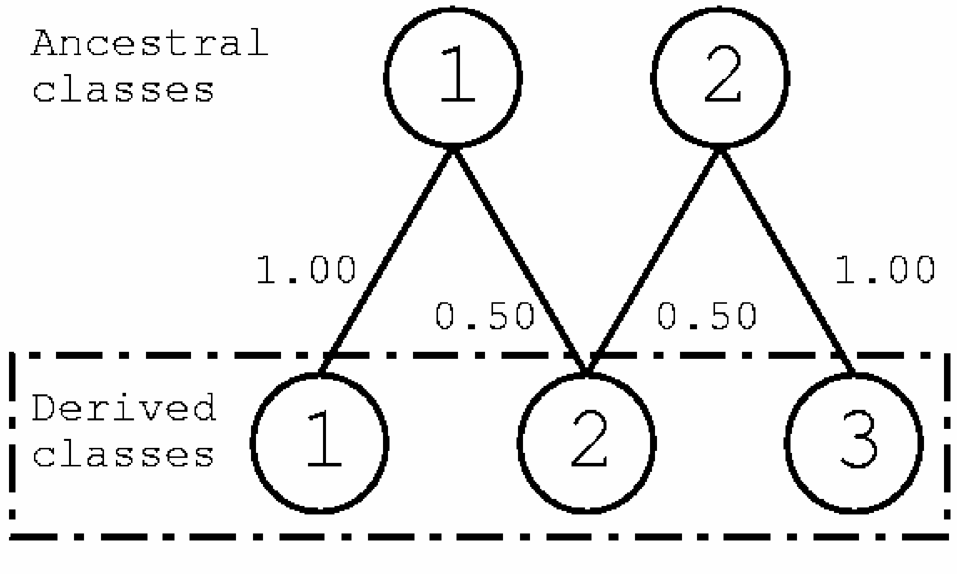
$$L_i = \sum_j P(G \mid C) P(C)$$

- Use AIC to select optimal *K* solution

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

# 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



# 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

$$\sum_{C} L(G \mid X, C) P(C)$$

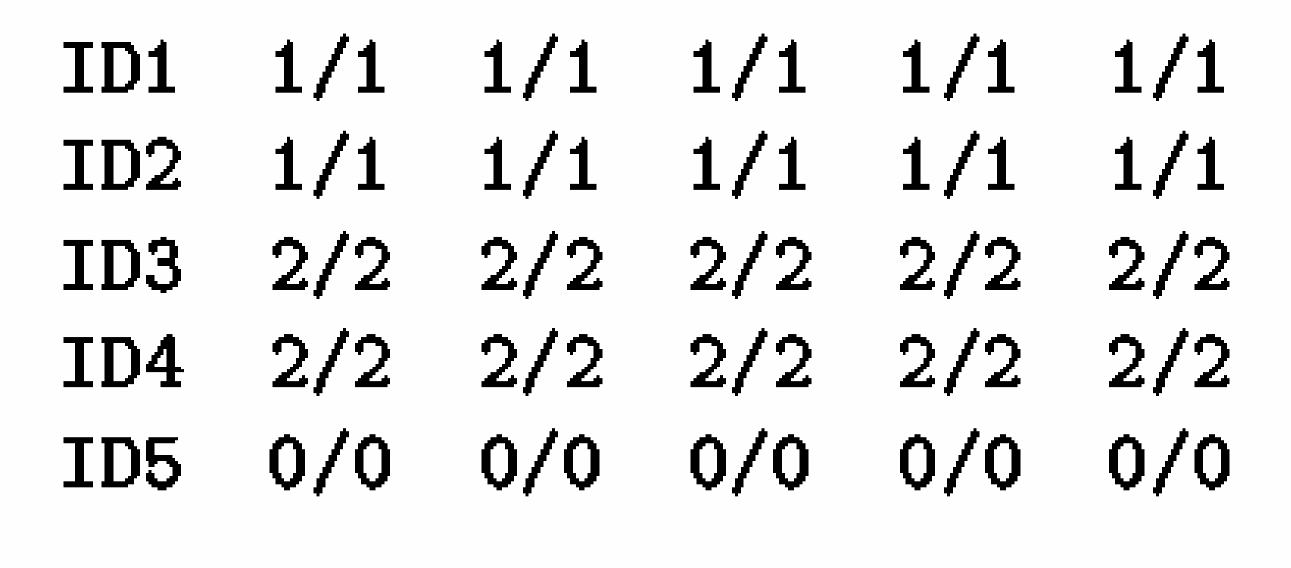
Class-specific likelihood
of genotype conditional
on trait

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

$$L(G \mid X, C) = \frac{L(X \mid G, C) L(G \mid C)}{\sum_{G} L(X \mid G, C) L(G \mid C)}$$

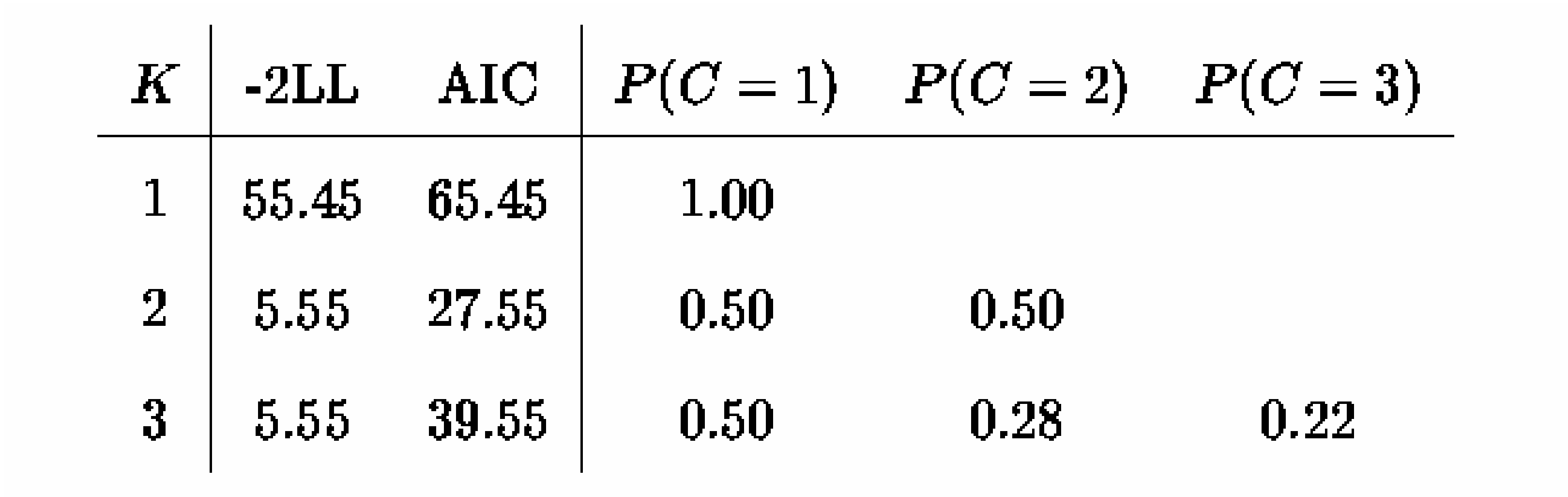
Parameters p, a, d
(potentially class-specific)

# Example #1



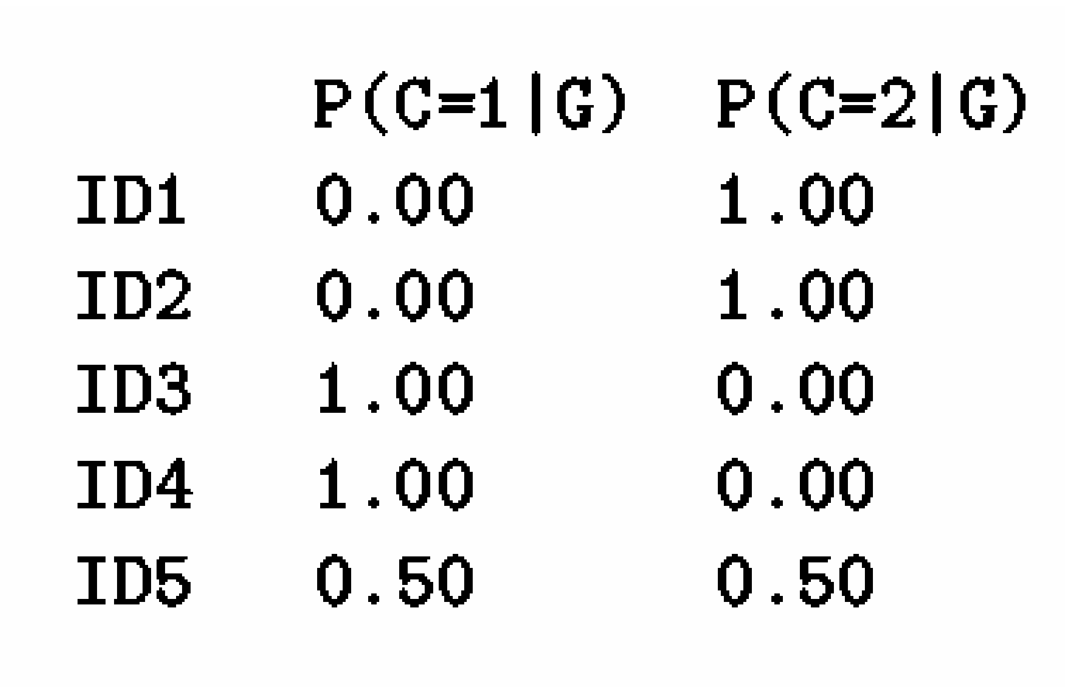
```
ID1  1/1  1/1  1/1  1/1  1/1
ID2  1/1  1/1  1/1  1/1  1/1
ID3  2/2  2/2  2/2  2/2  2/2
ID4  2/2  2/2  2/2  2/2  2/2
ID5  0/0  0/0  0/0  0/0  0/0
```

# Example #1



| $K$ | -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 |

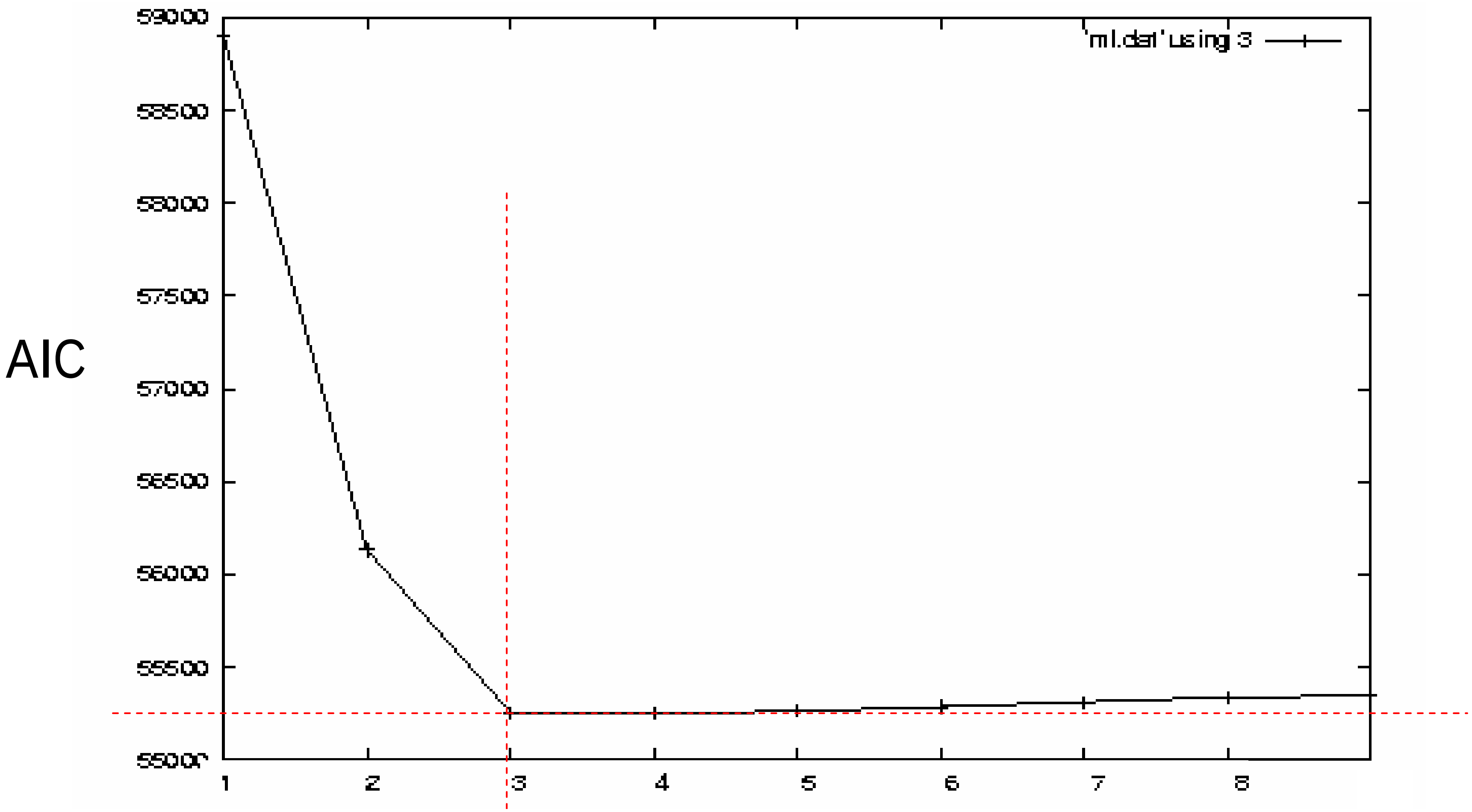
# Example #1



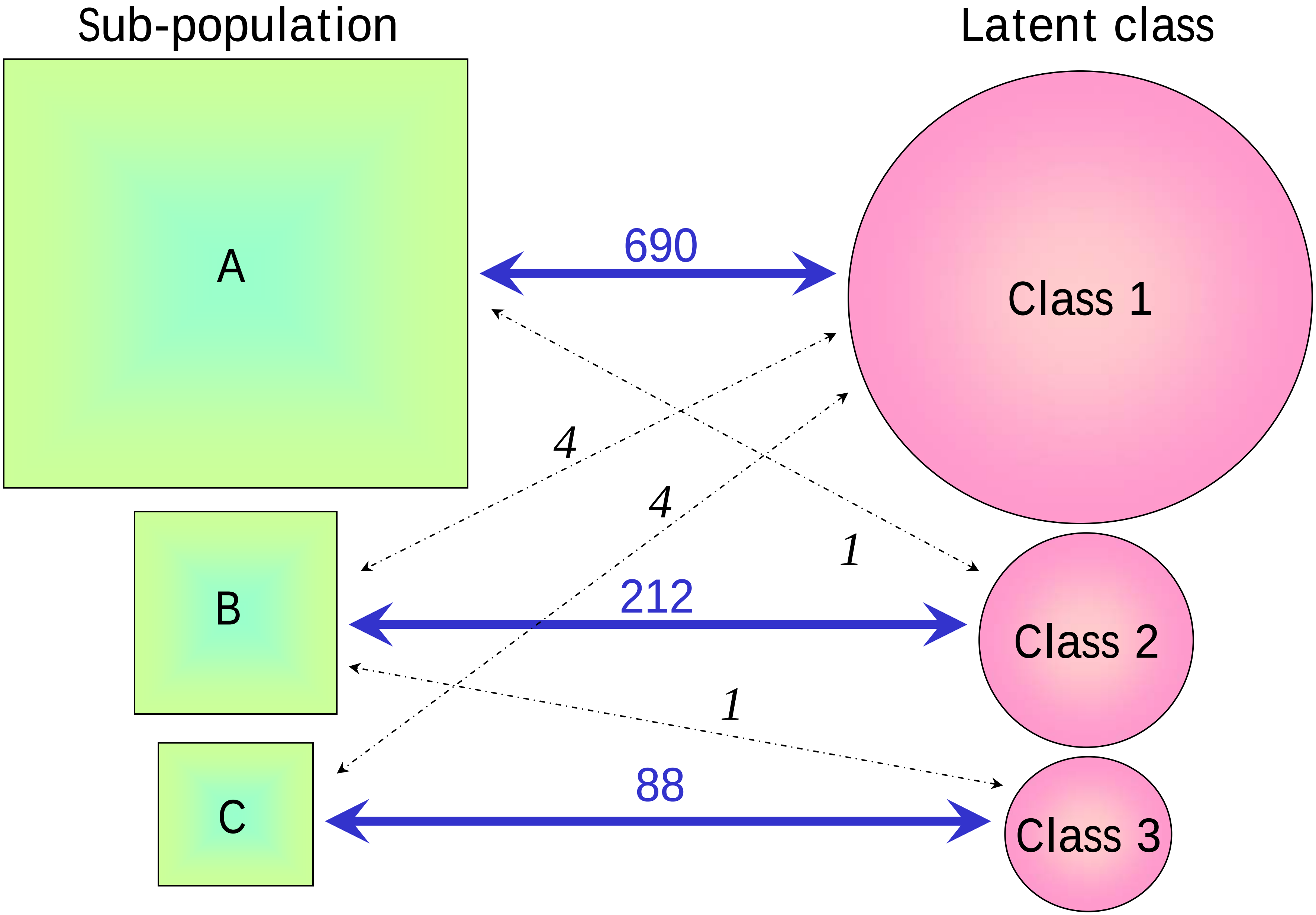
| | P(C=1\|G) | P(C=2\|G) |
|---|---|---|
| ID1 | 0.00 | 1.00 |
| ID2 | 0.00 | 1.00 |
| ID3 | 1.00 | 0.00 |
| ID4 | 1.00 | 0.00 |
| ID5 | 0.50 | 0.50 |

# Example #2

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



# *K*= 3



Sub-population

Latent class

A

B

C

Class 1

Class 2

Class 3

690

212

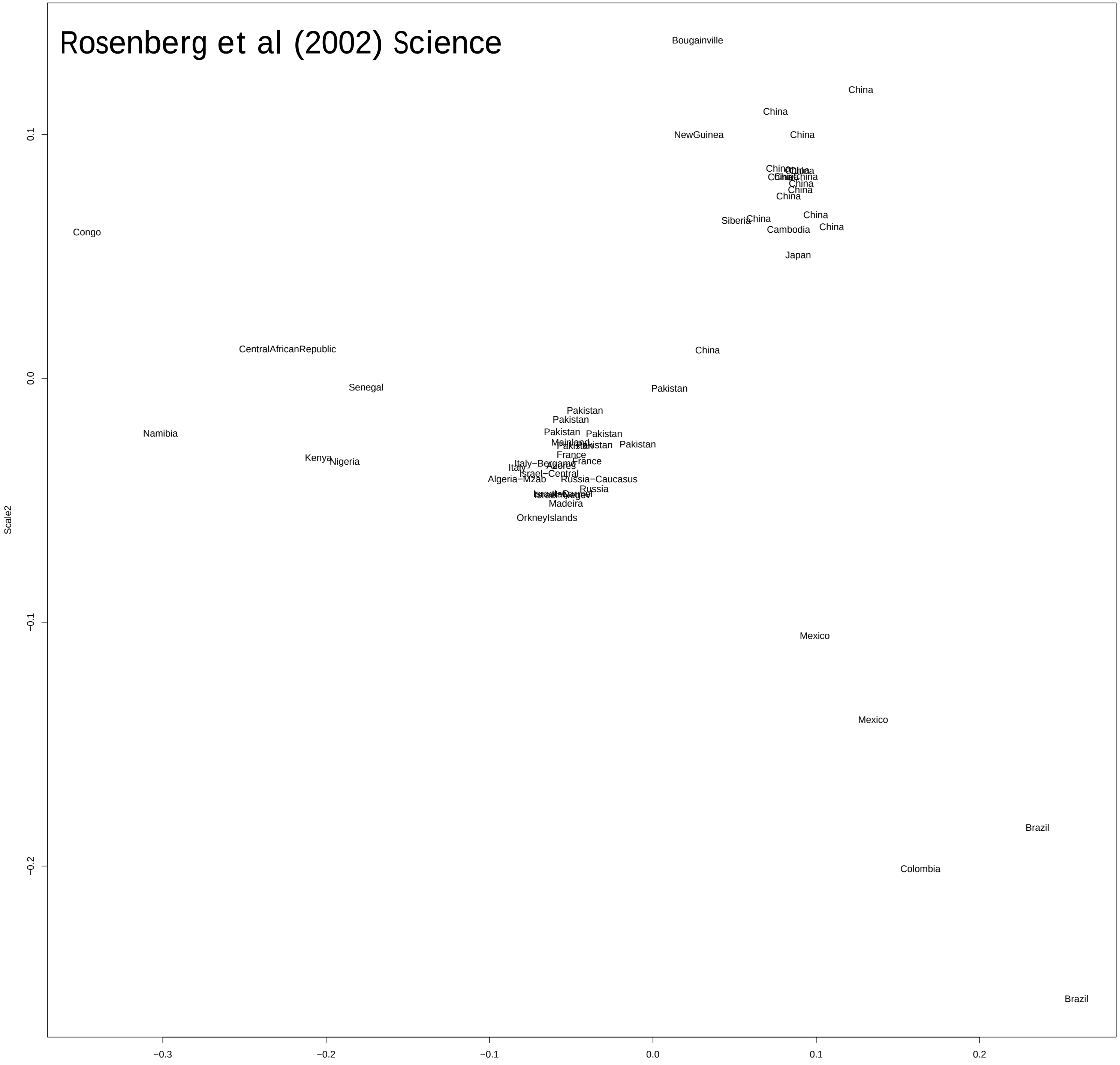
88

4

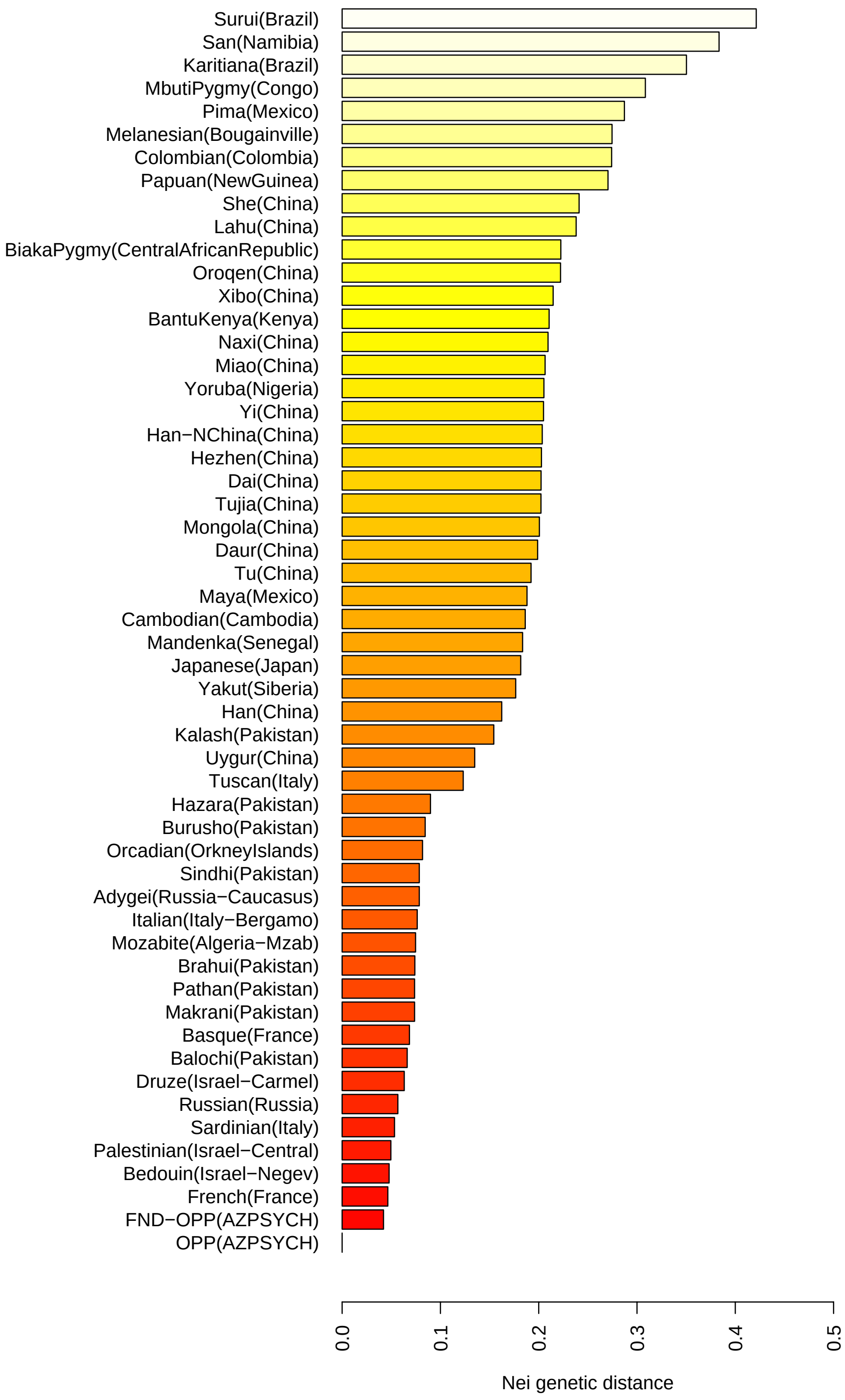
4

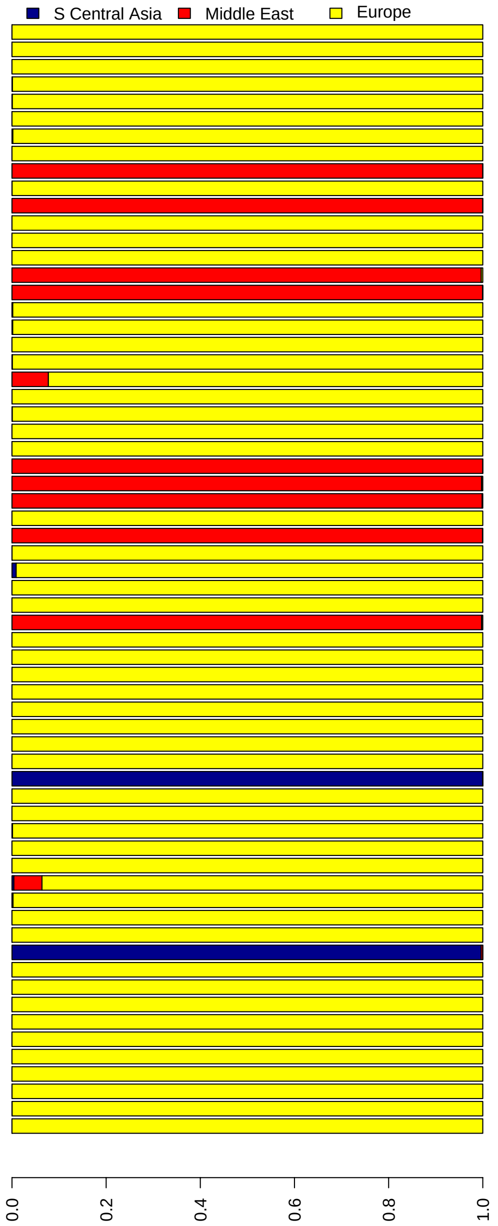
1

1



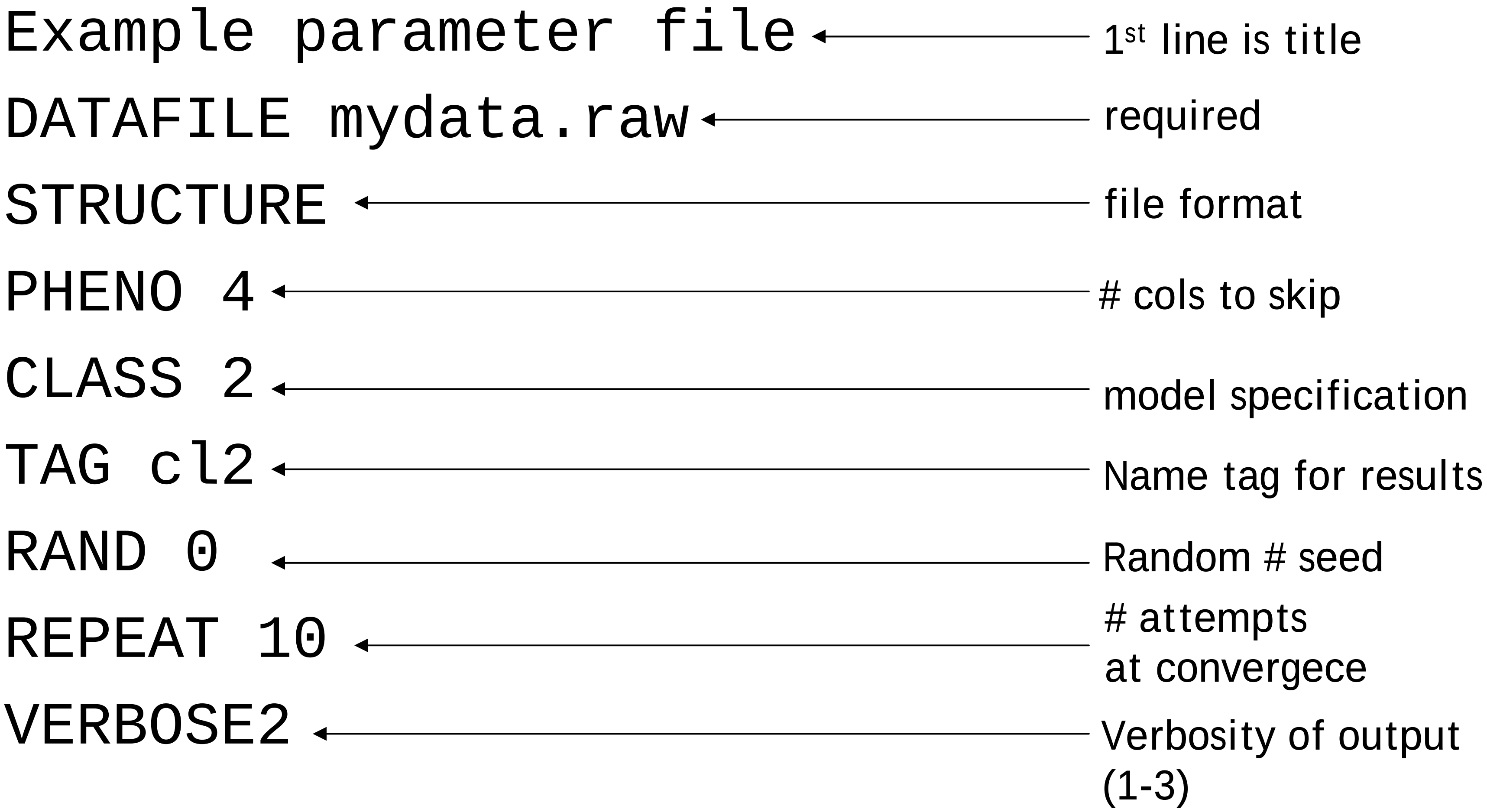
# Rosenberg et al (2002) Science


Surui(Brazil)
San(Namibia)
Karitiana(Brazil)
MbutiPygmy(Congo)
Pima(Mexico)
Melanesian(Bougainville)
Colombian(Colombia)
Papuan(NewGuinea)
She(China)
Lahu(China)
BiakaPygmy(CentralAfricanRepublic)
Oroqen(China)
Xibo(China)
BantuKenya(Kenya)
Naxi(China)
Miao(China)
Yoruba(Nigeria)
Yi(China)
Han−NChina(China)
Hezhen(China)
Dai(China)
Tujia(China)
Mongola(China)
Daur(China)
Tu(China)
Maya(Mexico)
Cambodian(Cambodia)
Mandenka(Senegal)
Japanese(Japan)
Yakut(Siberia)
Han(China)
Kalash(Pakistan)
Uygur(China)
Tuscan(Italy)
Hazara(Pakistan)
Burusho(Pakistan)
Orcadian(OrkneyIslands)
Sindhi(Pakistan)
Adygei(Russia−Caucasus)
Italian(Italy−Bergamo)
Mozabite(Algeria−Mzab)
Brahui(Pakistan)
Pathan(Pakistan)
Makrani(Pakistan)
Basque(France)
Balochi(Pakistan)
Druze(Israel−Carmel)
Russian(Russia)
Sardinian(Italy)
Palestinian(Israel−Central)
Bedouin(Israel−Negev)
French(France)
FND−OPP(AZPSYCH)
OPP(AZPSYCH)
0.0
0.1
0.2
0.3
0.4
0.5
Nei genetic distance


S Central Asia
Middle East
Europe
0.0
0.2
0.4
0.6
0.8
1.0

# Notes on L-POP

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



| | |
|---|---|
| `Example parameter file` | 1st line is title |
| `DATAFILE mydata.raw` | required |
| `STRUCTURE` | file format |
| `PHENO 4` | # cols to skip |
| `CLASS 2` | model specification |
| `TAG cl2` | Name tag for results |
| `RAND 0` | Random # seed |
| `REPEAT 10` | # attempts<br>at convergece |
| `VERBOSE2` | Verbosity of output<br>(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 probabilties** |
| `grep D: results` | get genetic distance matrix |
| `grep I:cl3: results` | get P(C\|G) for solution with `TAG cl3` only |

# 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

HO      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
----------------------------
```

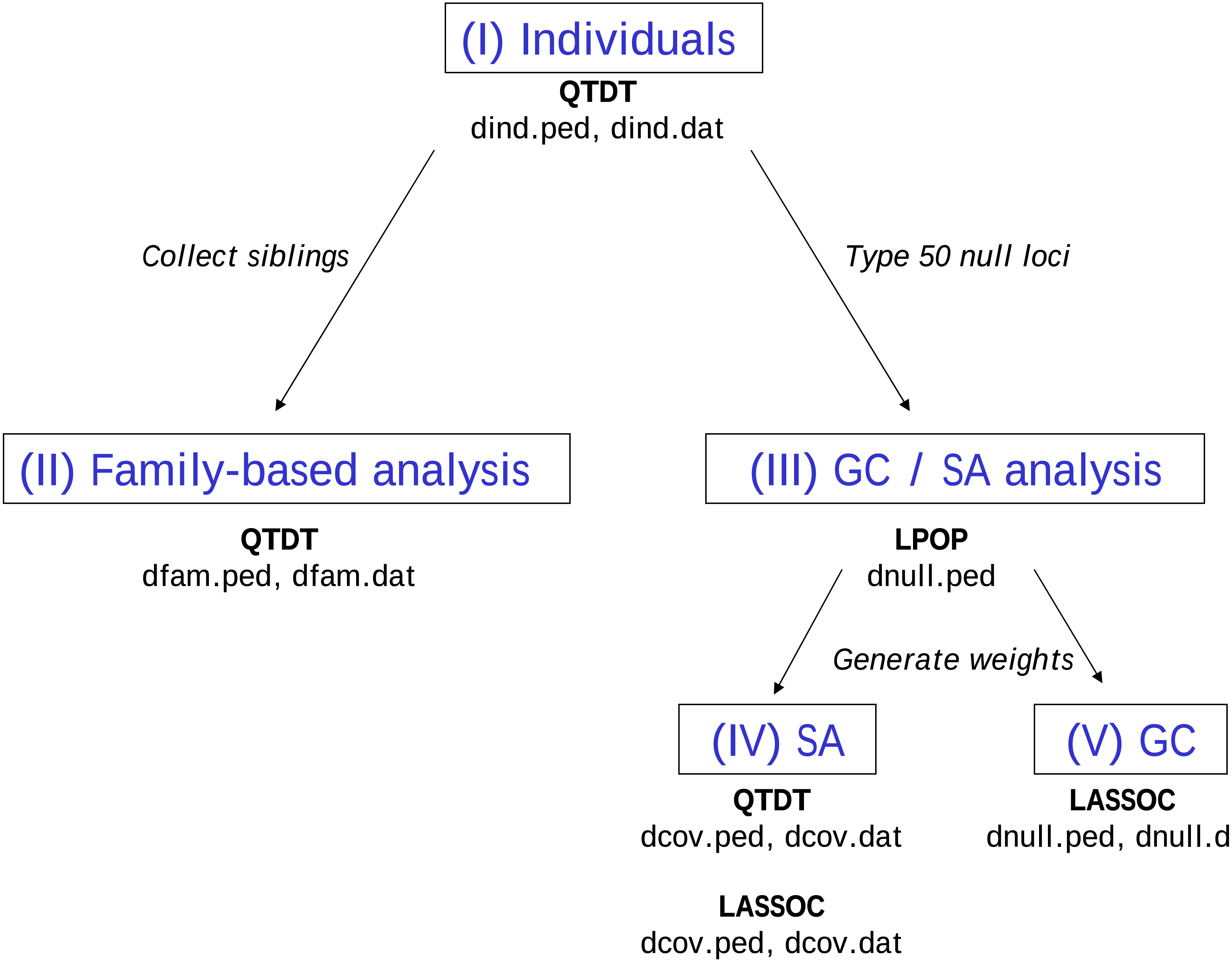
```
lassoc --file data --alt Pa --null P

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

HO      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

**LASSOC**
dcov.ped, dcov.dat

(V) GC

**LASSOC**
dnull.ped, dnull.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 …
3 1 0 0 1   2 1 1 1 -0.423   2 1  1 1  1 2  1 1  2 1  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
(estimated by 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 cl3`)

```
grep P:cl3: results
```

Get posterior probabilities from the 3 class solution

```
grep I:cl3: results
grep I:cl3: 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
```