

# Modelling substructure and pedigree relations in association mapping using GenStat

Marcos Malosetti & Fred van Eeuwijk

European GenStat and ASReml Applied Statistics Conference 2010  
Rothamsted, 14 July 2010



# QTL mapping

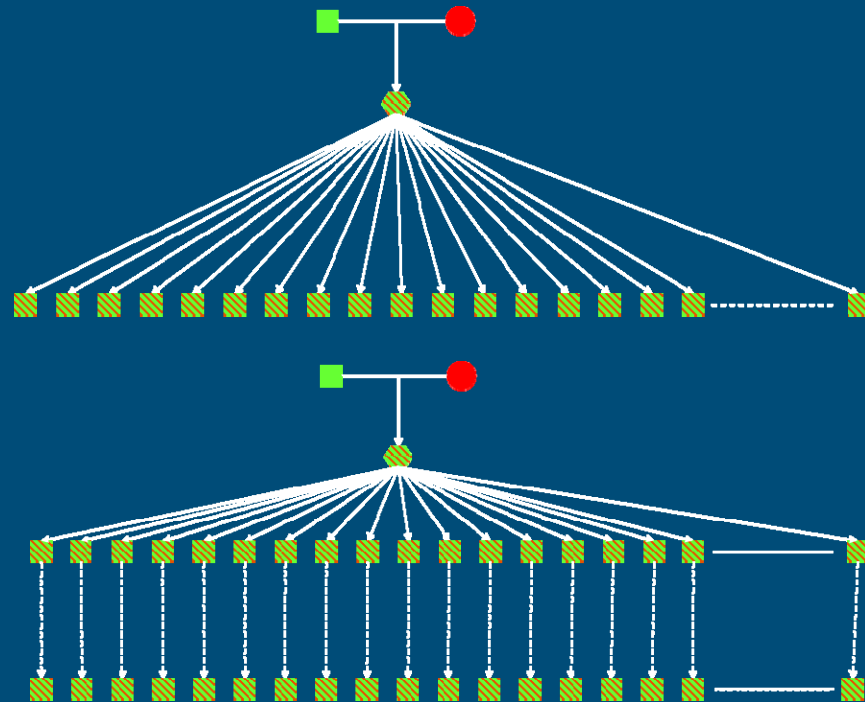
- Objective of QTL mapping studies:
  - describe phenotypes in relation to underlying genetic factors (called **QTL**)
  - QTL = **Q**uantitative **T**rait **L**ocus
- Finding statistical association between information at the DNA level (molecular markers) and phenotypic variation
  - Linkage analysis: common approach, conventional QTL mapping
  - **LD mapping or association mapping** : more recently

# Linkage and linkage disequilibrium

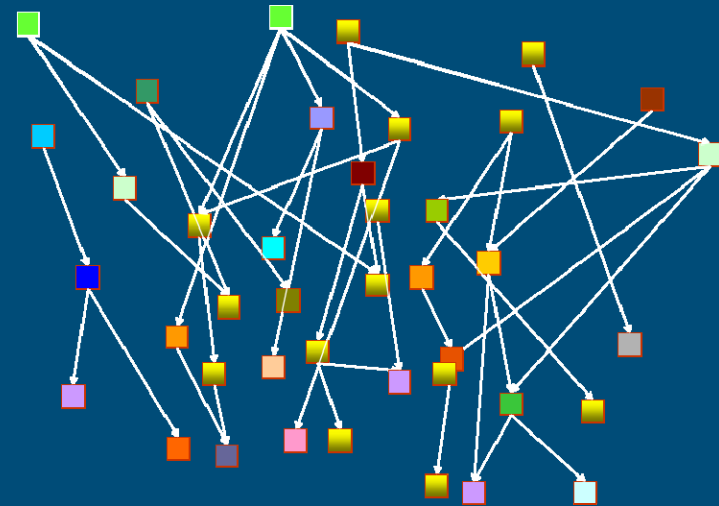
- Statistical association between markers and phenotypes found when there is **linkage (disequilibrium)** between markers and QTLs.
- Linkage disequilibrium (LD): non-random association of alleles at two loci **not necessarily** on the same chromosome.
- Linkage: non-random association of alleles at two loci due **to limited recombination** between the loci.

# Conventional QTL mapping versus LD mapping

Designed crosses



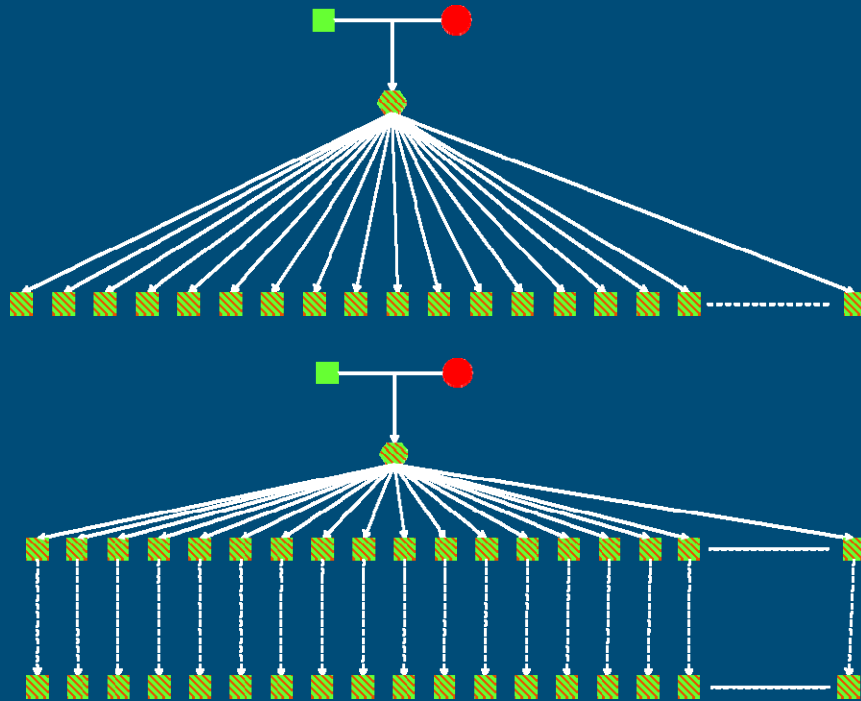
Association panel



Both, linkage analysis and LD mapping, rely on linkage disequilibrium to detect QTLs

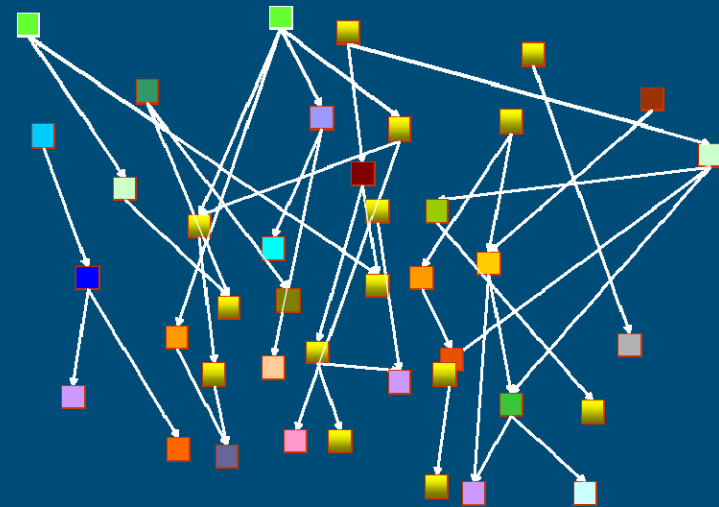
# Conventional QTL mapping versus LD mapping

Designed crosses



LD marker-QTL **is only** consequence of **linkage**.

Association panel



LD marker-QTL **can be** consequence of **linkage** but also other factors can cause LD.

# Genetic relatedness / population substructure

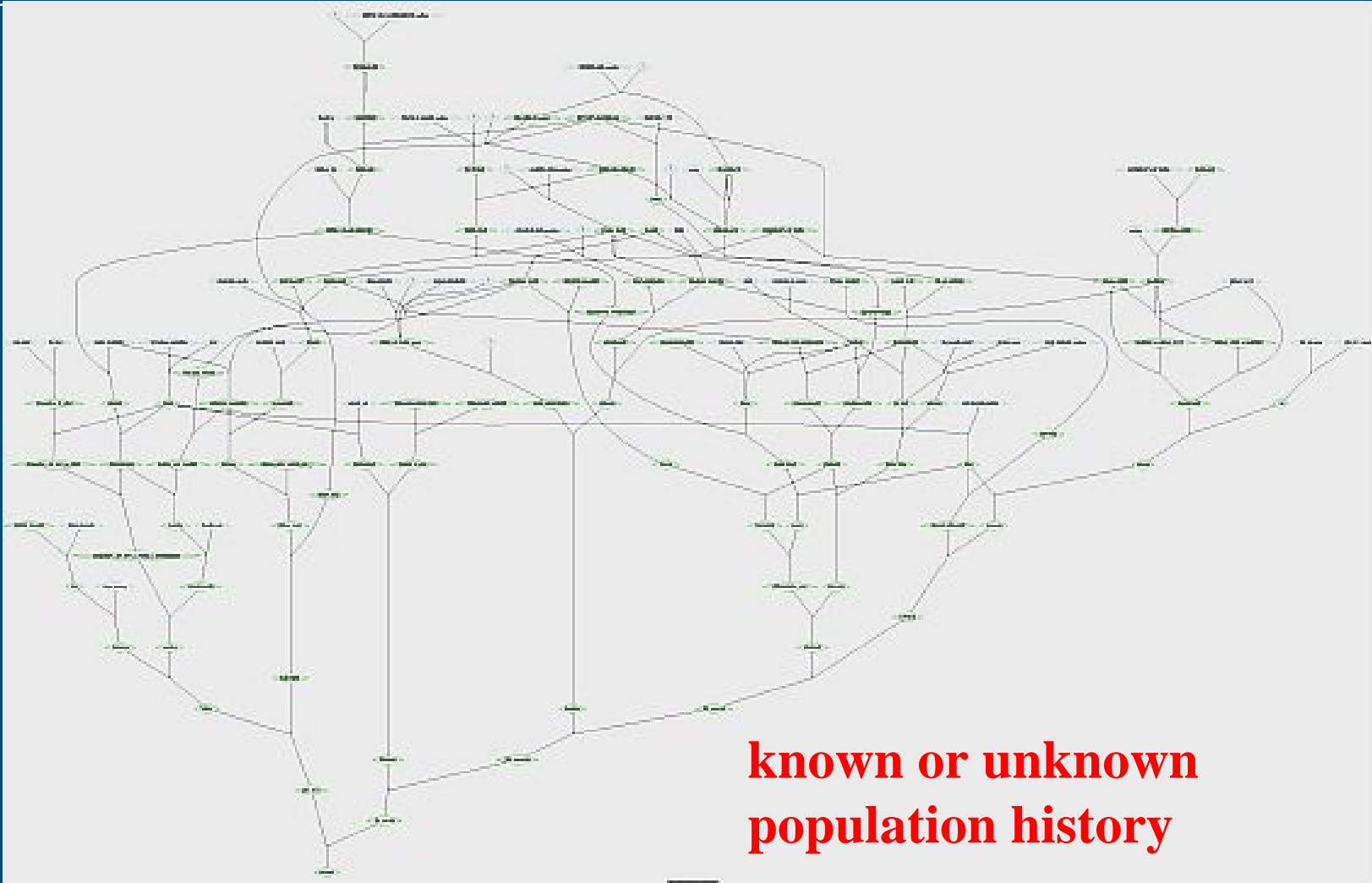
- In linkage QTL analysis, all genotypes of an offspring population have on expectation equal relatedness/correlation
- In LD mapping the genotypes have different degrees of relatedness between them:
  - “Unstructured”: coefficient of coancestry  $\theta_{ij}$
  - “Structured” (population substructure): special form of genetic relatedness, group-wise correlation differences between genotypes (block diagonal VCOV)
- Not accounting for relatedness (coancestry or population substructure) will cause spurious associations



# Statistical modelling

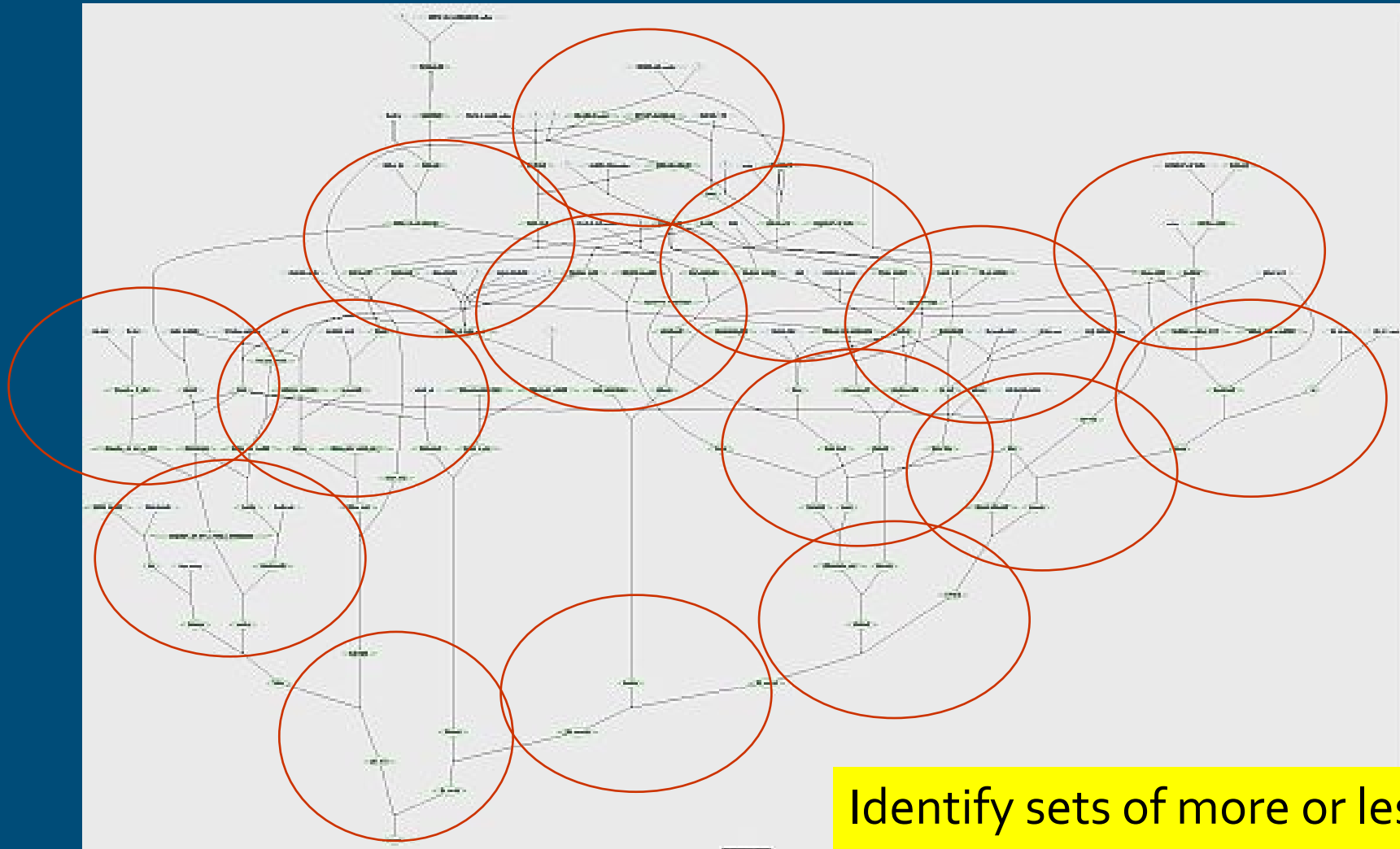
- From the modelling, LD mapping is not much different from linkage QTL mapping.
- Models for linkage QTL mapping are also applicable to LD mapping situations **provided** that LD due to linkage is separated from LD due to other reasons
  - LD mapping requires more elaborated modelling of **genetic relatedness**
  - Appropriate structuring for the VCOV's of the random genotypic effects

# Association panel: a set of interconnected genotypes



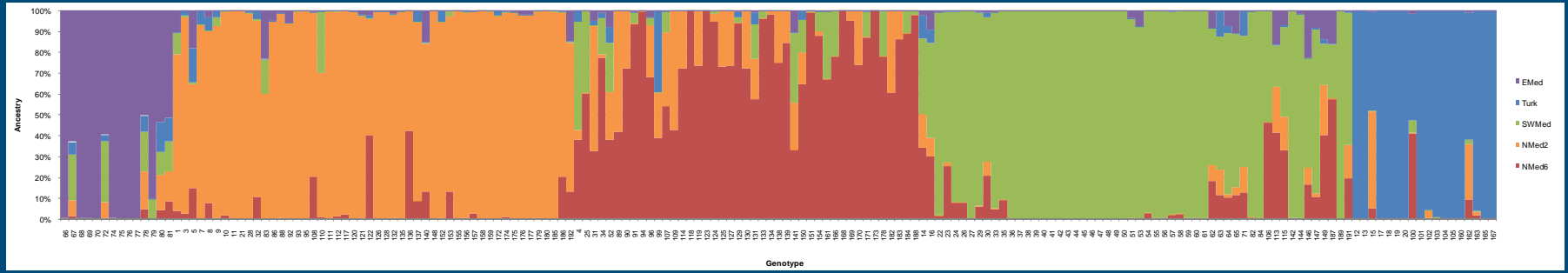


# Genetic correlation between individuals: “structured”



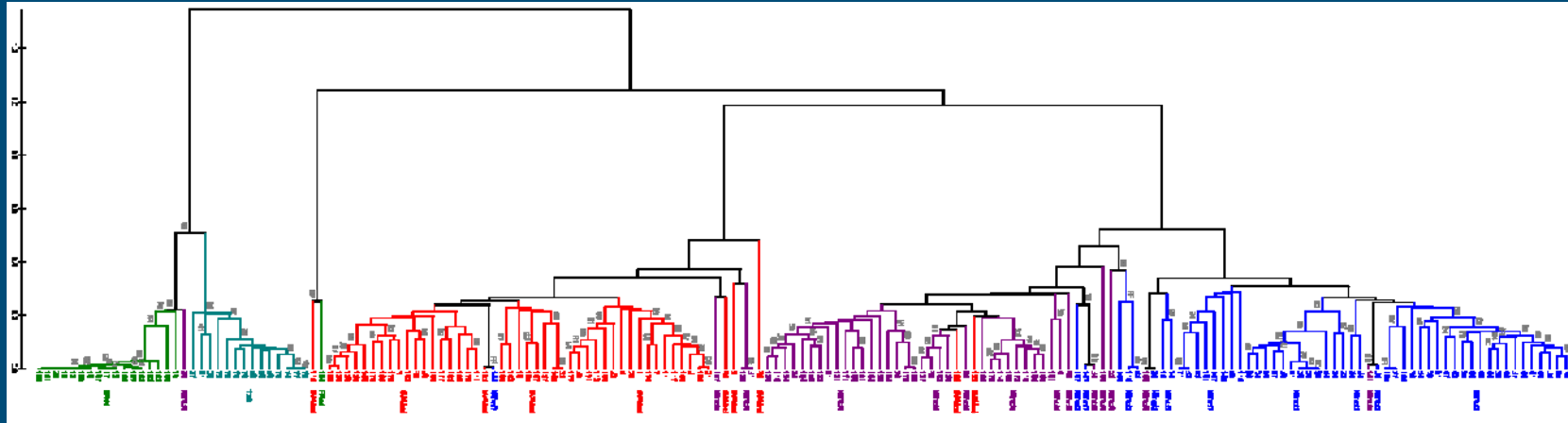
Identify sets of more or less homogenous genotypes

# Quantifying genetic relatedness / Structuring VCOV(G)



- Bayesian clustering STRUCTURE (Pritchard et al. 2000)
  - It can be computationally intensive
  - Population assumptions not always compatible with plant populations (mainly developed to be used in human population genetics)
  - Not always obvious how to define the model to use

# Quantifying genetic relatedness / Structuring VCOV(G)



- Classical multivariate approaches
  - Simple, fast
  - Similar results with STRUCTURE
  - Where to define boundaries between groups?
- Other criteria (e.g. geographical origin)

# Eigenanalysis

Patterson et al 2006.pdf - Adobe Reader

File Edit View Document Tools Window Help

1 (1 of 20) 134% Find

OPEN ACCESS Freely available online PLoS GENETICS

## Population Structure and Eigenanalysis

Nick Patterson<sup>1\*</sup>, Alkes L. Price<sup>1,2</sup>, David Reich<sup>1,2</sup>

1 Broad Institute of Harvard and MIT, Cambridge, Massachusetts, United States of America, 2 Department of Genetics, Harvard Medical School, Boston, Massachusetts, United States of America

**Current methods for inferring population structure from genetic data do not provide formal significance tests for population differentiation. We discuss an approach to studying population structure (principal components analysis) that was first applied to genetic data by Cavalli-Sforza and colleagues. We place the method on a solid statistical footing, using results from modern statistics to develop formal significance tests. We also uncover a general “phase change” phenomenon about the ability to detect structure in genetic data, which emerges from the statistical theory we use, and has an important implication for the ability to discover structure in genetic data: for a fixed but large dataset size, divergence between two populations (as measured, for example, by a statistic like  $F_{ST}$ ) below a threshold is essentially undetectable, but a little above threshold, detection will be easy. This means that we can predict the dataset size needed to detect structure.**

Citation: Patterson N, Price AL, Reich D (2006) Population structure and eigenanalysis. PLoS Genet 2(12): e190. doi:10.1371/journal.pgen.0020190

### Introduction

A central challenge in analyzing any genetic dataset is to explore whether there is any evidence that the samples in the data are from a population that is structured. Are the individuals from a homogeneous population or from a practical, even on the largest datasets. This is our main aim in this paper.

Using some recent results in theoretical statistics, we introduce a formal test statistic for population structure. We also discuss testing for *additional* structure after some structure has been found. Finally, we are able to estimate the

start papers KINGSTON... LD mappin... LD course ... marker-tra... Patterson ... EN 0:02

# Eigenanalysis

- PCA on genotype x marker scores matrix with a formal test for the number of axes (dimensions)
- No discrete groups, but set of PC's be used as covariates in marker – trait association analysis
- When PC's are introduced in random part of a mixed model, they will approximate the full genetic relationship matrix
- Straightforward, simple, and is easy to program in a conventional statistical package

# Mixed models and LD mapping in GenStat

- LD mapping models should accommodate the complex genetic relationships in the population.
- Mixed models are particularly suitable (GenStat).
- Suite of GenStat procedures developed to run different models for LD mapping.
- Procedures can be run from the GUI.

# A mixed model for LD mapping

$$\underline{P} = \underline{\text{genotype}} + \underline{\text{error}}$$



$$\underline{P} = \underline{\text{marker}} + \underline{\text{genotype}^*} + \underline{\text{error}}$$

$$\underline{P}_i = \mu + \underline{G}_i + \underline{\varepsilon}_i$$

$$\underline{P}_i = \mu + x_i \alpha + \underline{G}_i + \underline{\varepsilon}_i$$

$$\underline{G} \sim N(0, \sigma_{\text{genotype}}^2) \quad \underline{\text{error}} \sim N(0, \sigma^2)$$

$$x_i = -1 \text{ if mm}$$

$$x_i = 0 \text{ if Mm}$$

$$x_i = 1 \text{ if MM}$$

# A naive mixed model for LD mapping

$$\underline{P} = \text{marker} + \underline{\text{genotype}^*} + \underline{\text{error}}$$

$$\underline{P}_i = \mu + x_i\alpha + \underline{G}_i + \underline{\varepsilon}_i$$

Standard assumption:

$$\underline{G} \sim N(0, \sigma_{\text{genotype}}^2) \quad \underline{\text{error}} \sim N(0, \sigma^2)$$

- This model assumes UNRELATED genotypes

Relationship matrix  $K=I$

$$\underline{G} \sim N(0, I\sigma_{\text{genotype}}^2)$$

$$K = \begin{bmatrix} 1 & & & & \\ & 1 & & & \\ & & 1 & & \\ & & & \ddots & \\ & & & & \dots & 1 \end{bmatrix}$$

**This model ignores genetic relatedness/ population structure**

K should be in the model to correct for relatedness

$$\underline{P} = \text{marker} + \underline{\text{genotype}^*} + \underline{\text{error}}$$

$$\underline{P}_i = \mu + x_i\alpha + \underline{G}_i + \underline{\varepsilon}_i$$

Change model assumption:



$$\underline{G} \sim N(0, 2K\sigma_g^2) \quad \underline{\text{error}} \sim N(0, \sigma^2)$$

- Now the relationship matrix (K) is in the model
- K = kinship matrix derived from pedigree/marker information

Relationship matrix  $K \neq I$

$$K = \begin{bmatrix} \theta_{11} & & & & \\ \theta_{12} & \theta_{22} & & & \\ \theta_{13} & \theta_{23} & \theta_{33} & & \\ \vdots & \vdots & \vdots & \ddots & \\ \theta_{1I} & \theta_{2I} & \theta_{3I} & \cdots & \theta_{II} \end{bmatrix}$$

# LD mapping using eigenanalysis

$$\underline{P} = \text{marker} + \underline{\text{PC's}} + \underline{\text{genotype}^*} + \underline{\text{error}}$$

$$\underline{P}_i = \mu + x_i \alpha + \sum_M C_{i,m} + \underline{G}_i + \underline{\varepsilon}_i$$

$$\underline{C} \sim N(0, \sigma_{\text{scores}}^2) \quad \underline{G} \sim N(0, \sigma_{\text{genotype}}^2) \quad \underline{\text{error}} \sim N(0, \sigma^2)$$

- The PC scores represent relatedness / population structure
- PC's impose approximate covariance structure
- Computationally less intensive than full structuring of VCOV(G)

# Population structure

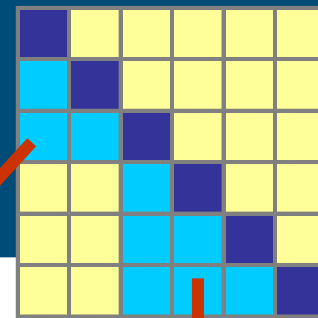
$\underline{P} = \text{marker} + \text{group} + \text{genotype} + \text{error}$

$$\underline{P}_i = \mu + x_i \alpha + \underline{C}_k + \underline{G}_{i(k)} + \underline{\varepsilon}_i$$

$$\underline{C} \sim N(0, \sigma_{\text{group}}^2) \quad \underline{G} \sim N(0, \sigma_{\text{genotype}}^2) \quad \text{error} \sim N(0, \sigma^2)$$

- This model imposes a common covariance between genotypes within a group
- Genotypes from different groups are still assumed unrelated
- Groups from STRUCTURE or clustering

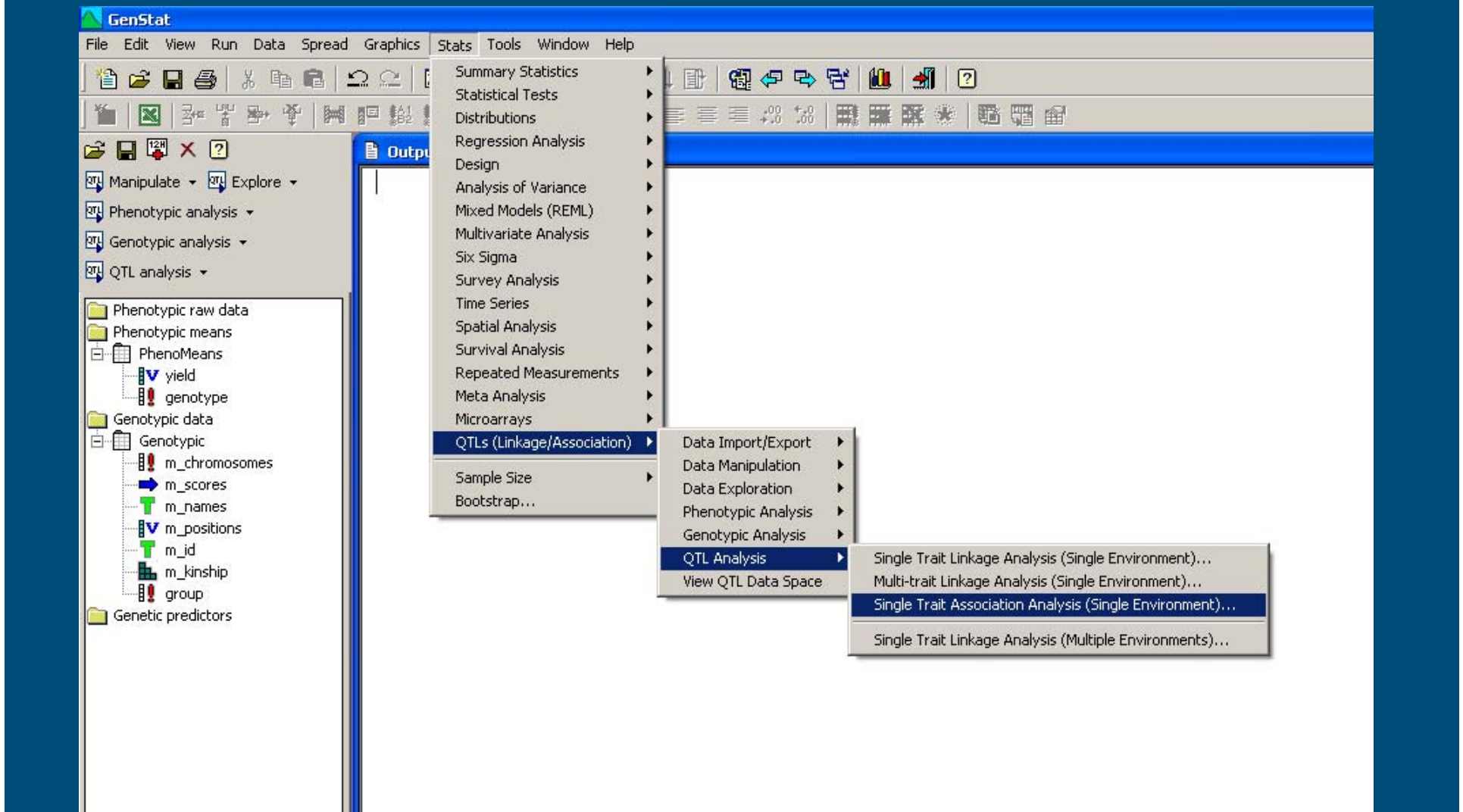
Relationship matrix  $K \neq I$



Group 1

Group 2

# LD mapping in GenStat 13



GenStat

File Edit View Run Data Spread Graphics Stats Tools Window Help

Output

Single Trait Association Analysis

Available Data:

- genotype
- group
- m\_chromosomes

Quantitative trait means: yield

Genotype factor: genotype

Marker genotype scores: m\_scores

Linkage groups: m\_chromosomes

Position within linkage group: m\_positions

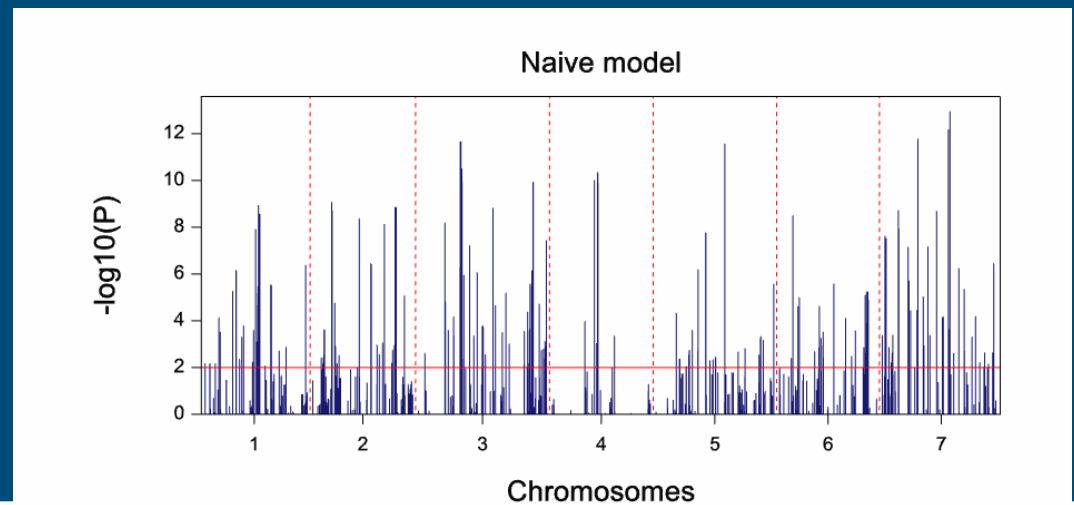
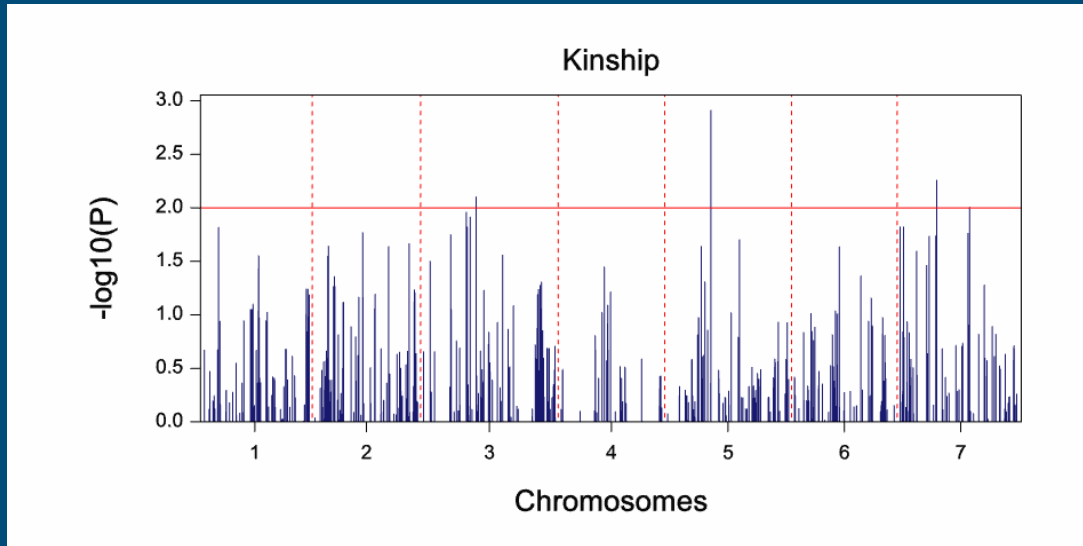
Marker names: m\_names

Relationship model:

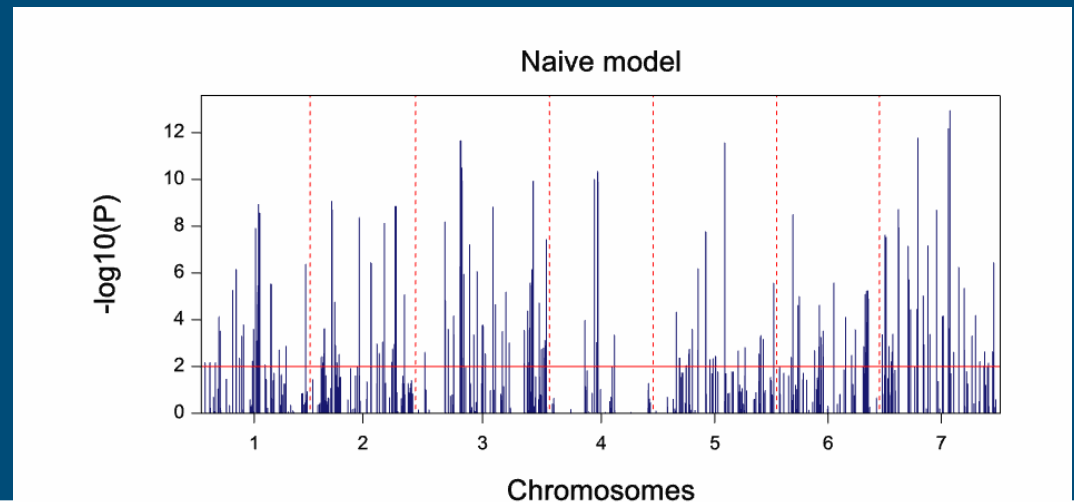
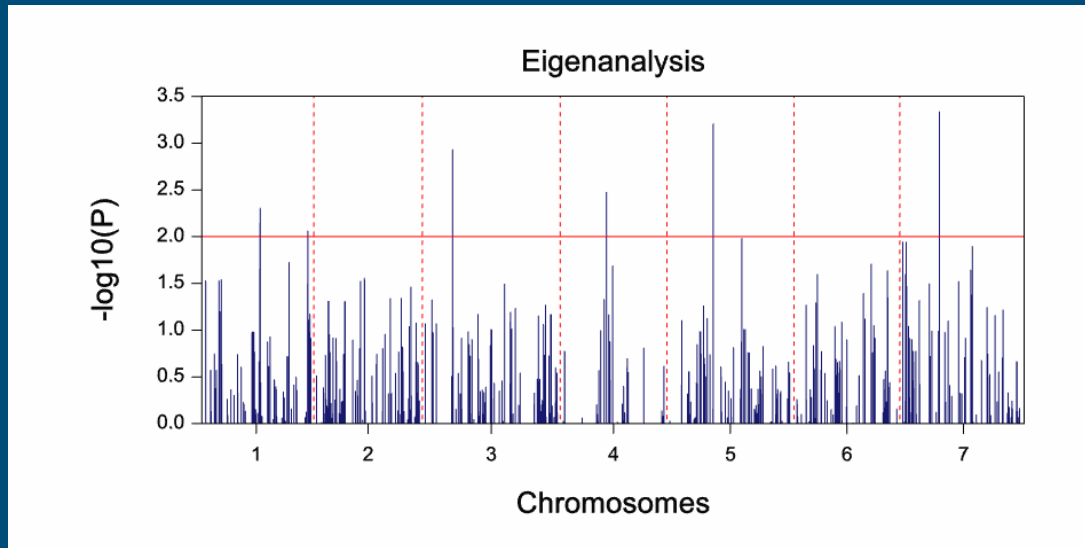
- Eigenanalysis
- Kinship matrix
- Subpopulation groupings
- Null

Run Options... Store... Cancel Defaults

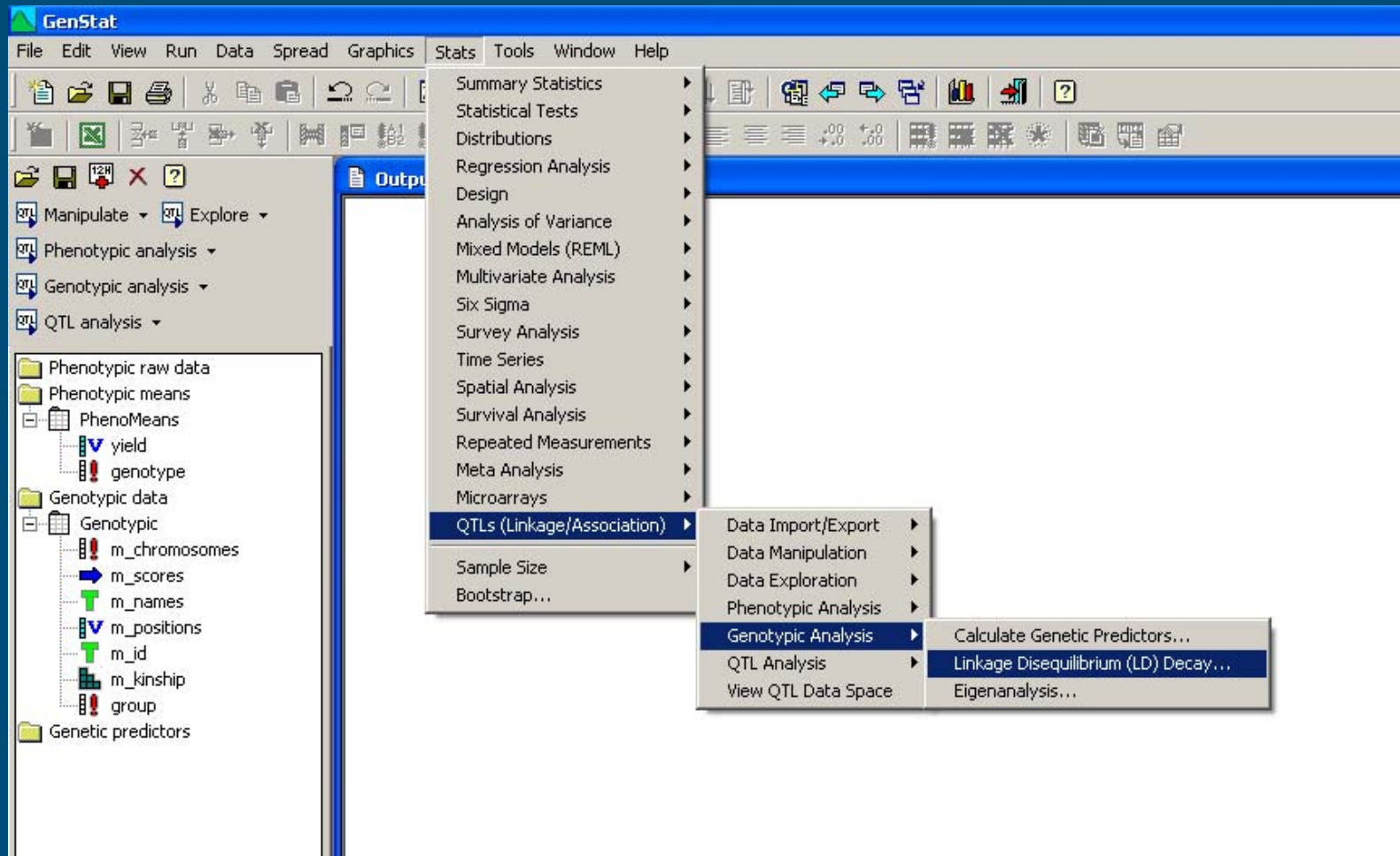
# Correcting for genetic relatedness: kinship vs null



# Correcting for genetic relatedness: PC scores vs null

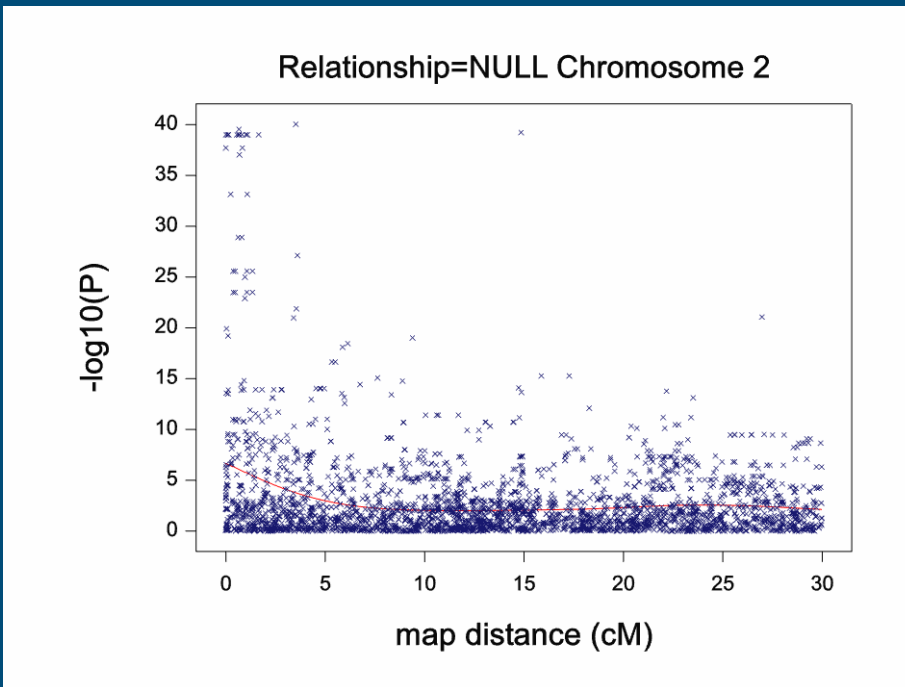


# LD decay plots

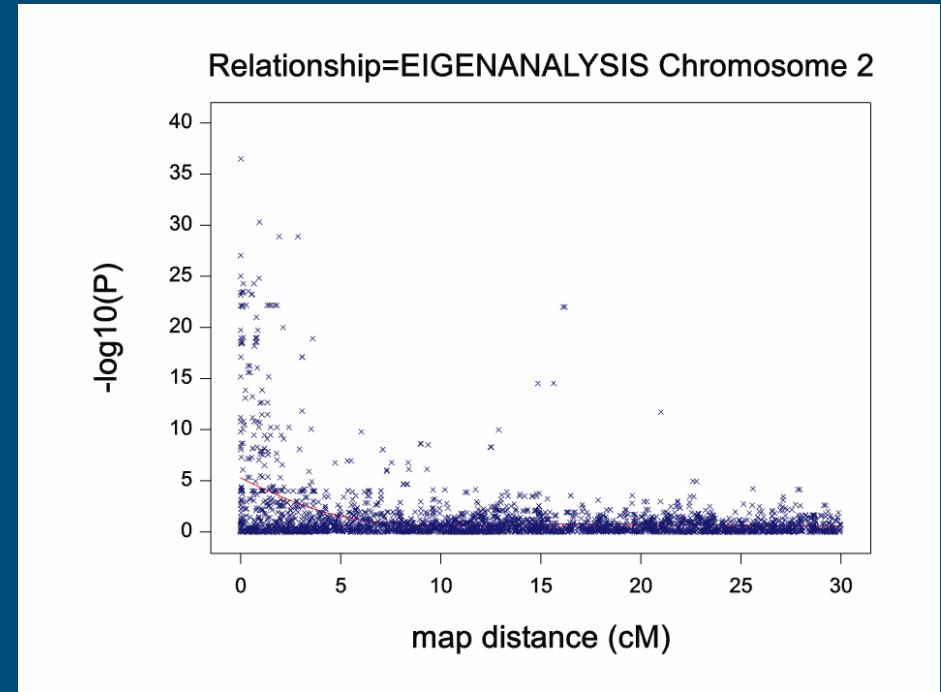


# LD decay plots

No correction



Correction for population structure

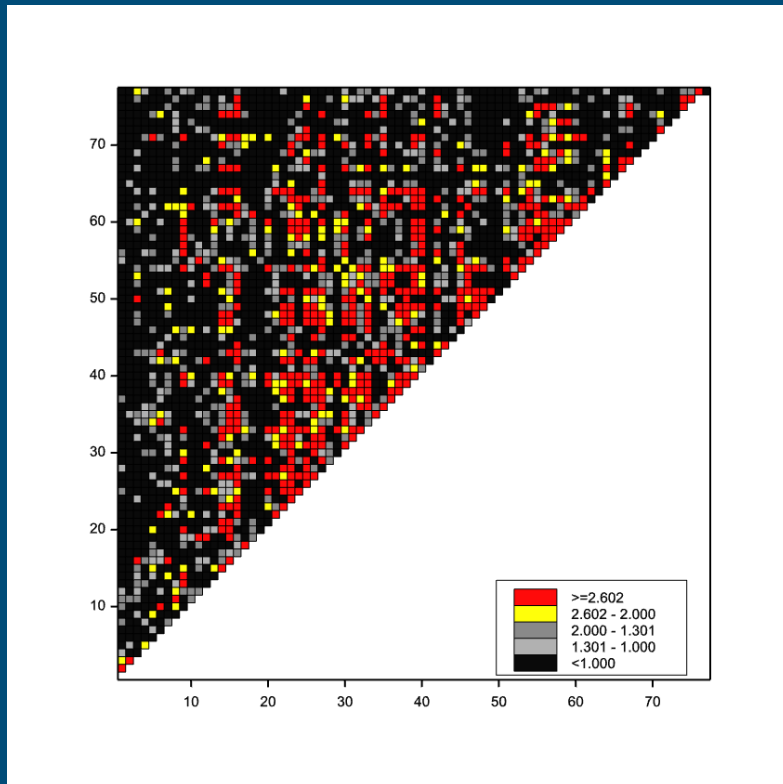


Response marker =  
**predictor marker** +  
error

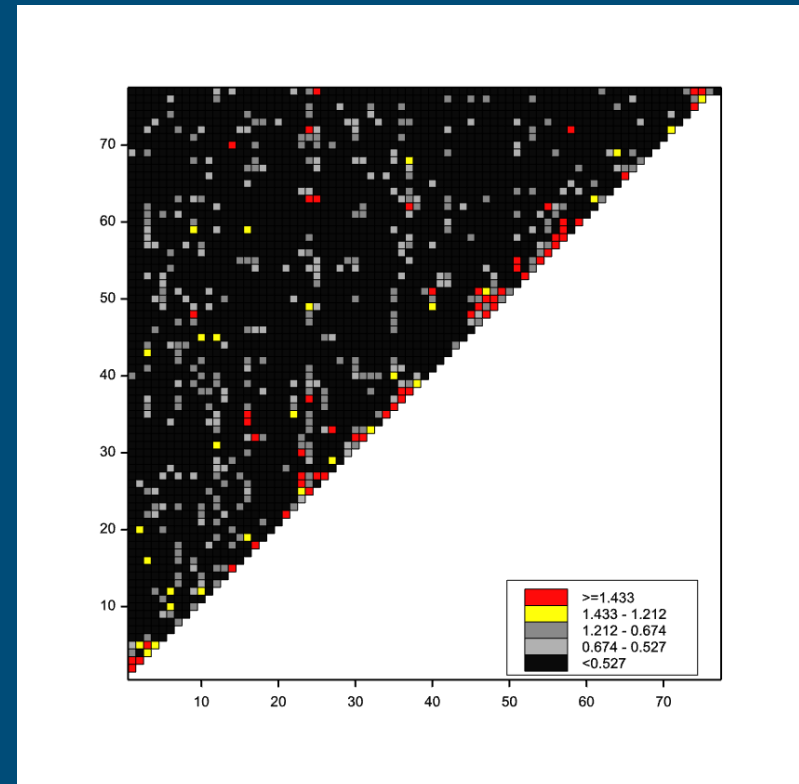
Response marker =  
**PC scores / groups** +  
**predictor marker** +  
error

# LD image plots

No correction



Correction for population structure



# Conclusions

- Study of genetic relatedness crucial in LD mapping
  - Kinship
  - Eigenanalysis
  - Clustering methods (including STRUCTURE)
- Need to control for genetic relatedness when assessing:
  - Marker – marker association (LD decay)
  - Marker – trait association (LD mapping)
- GenStat procedures / GUI can be used to run all these types of analyses

# Thanks for your attention

