Genome-Wide Association Studies

Caitlin Collins, Thibaut Jombart

MRC Centre for Outbreak Analysis and Modelling
Imperial College London

Genetic data analysis using

30-10-2014

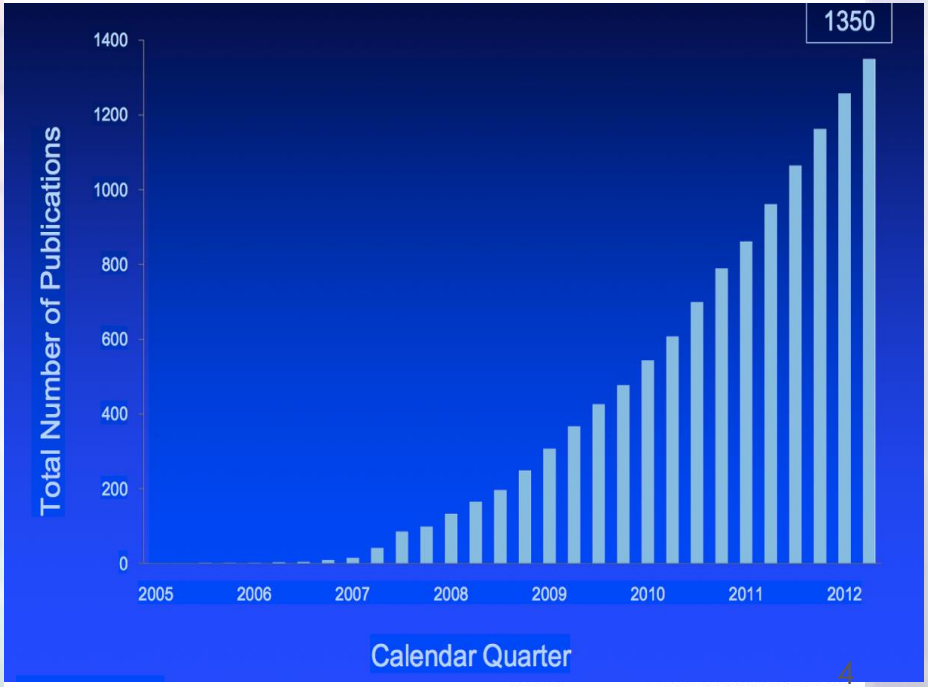
Outline

- Introduction to GWAS
- Study design
  - **GWAS design**
  - **Issues and considerations in GWAS**
- Testing for association
  - **Univariate methods**
  - **Multivariate methods**
    - Penalized regression methods
    - Factorial methods

Genomics & GWAS

The genomics revolution

- **Sequencing technology**
  - **1977** – Sanger
  - **1995** – 1st bacterial genomes
    - **< 10,000** bases per day per machine
  - **2003** – 1st human genome
    - **> 10,000,000,000,000** bases per day per machine
- **GWAS publications**
  - **2005** – 1st GWAS
    - Age-related macular degeneration
  - **2014** – 1,991 publications
    - **14,342 associations**



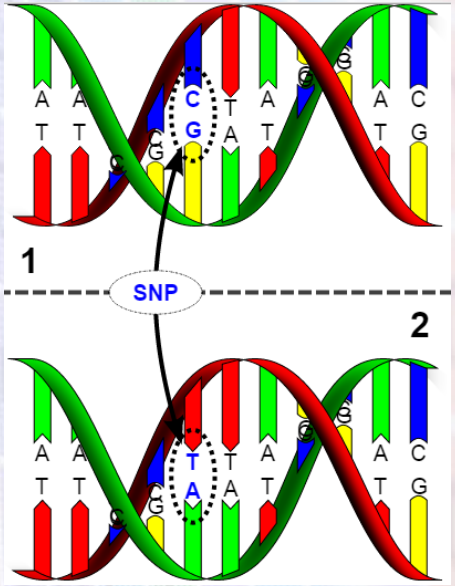
A few GWAS discoveries…

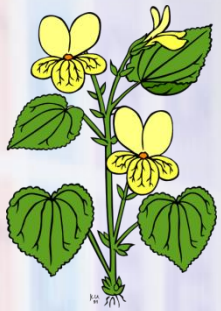
SNP-associated trait categories

- Digestive system disease
- Cardiovascular disease
- Metabolic disease
- Immune system disease
- Nervous system disease
- Liver enzyme measurement
- Lipid or lipoprotein measurement
- Inflammatory marker measurement
- Hematological measurement
- Body weights and measures
- Cardiovascular measurement
- Other measurement
- Response to drug
- Biological process
- Cancer
- Other disease
- Other trait

So what is GWAS?

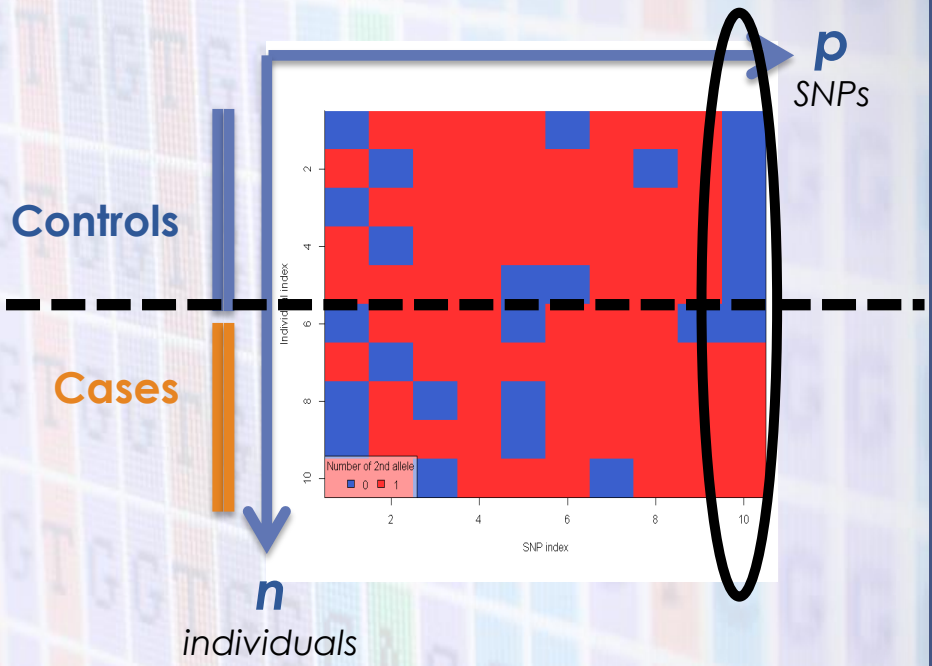
- Genome Wide Association Study
  - ***Looking for SNPs…***
    associated with a phenotype.
- Purpose:
  - **Explain**
    - Understanding
    - Mechanisms
    - Therapeutics
  - **Predict**
    - Intervention
    - Prevention
    - Understanding not required



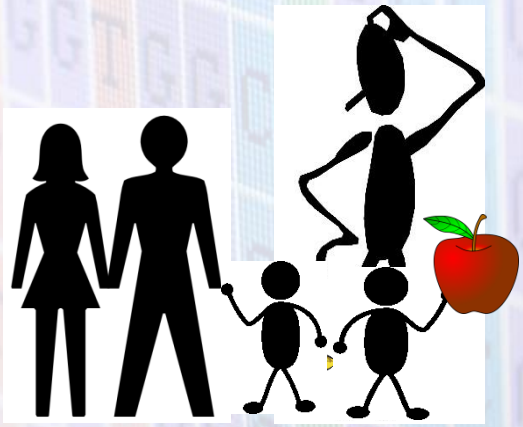


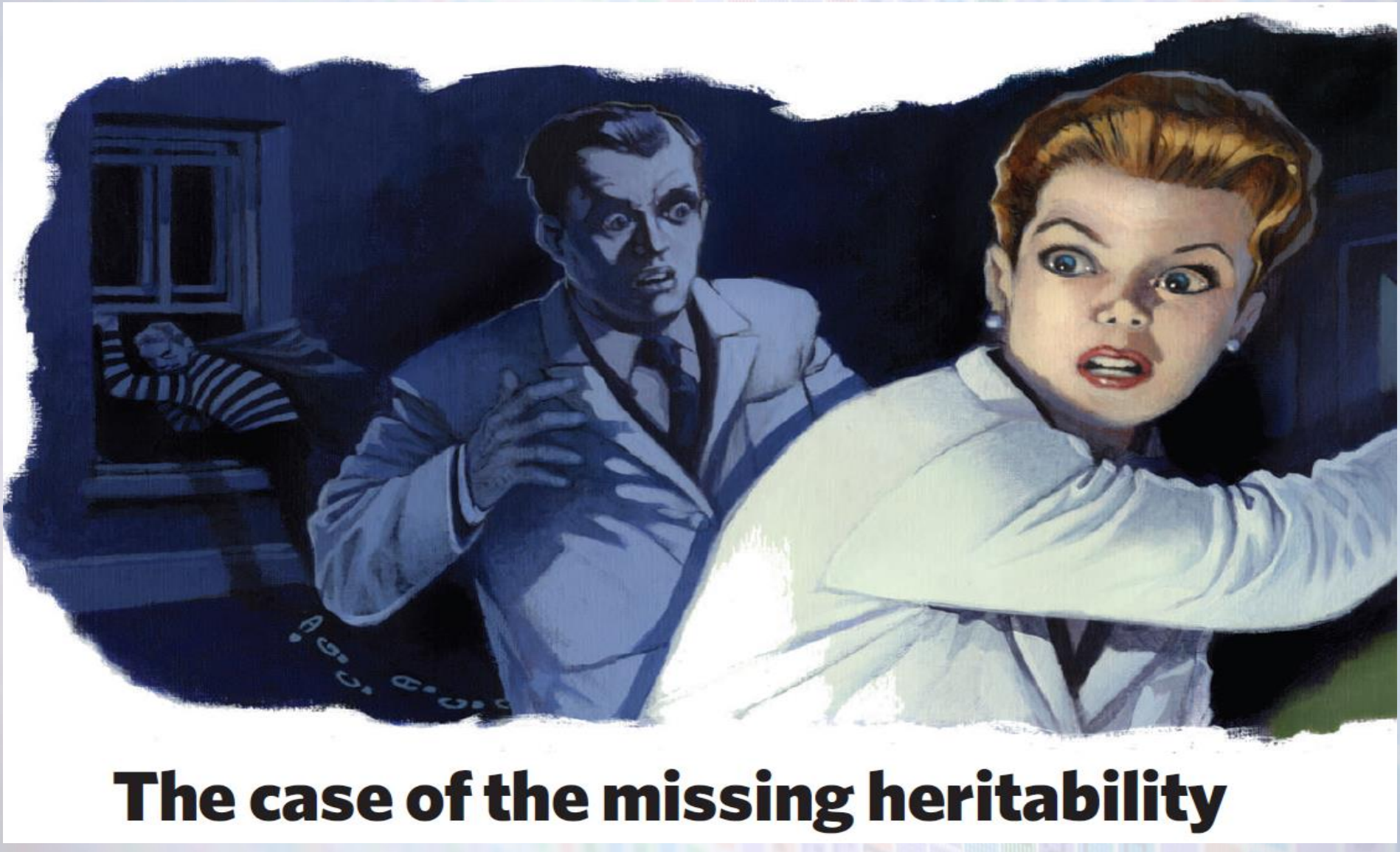
Association

- Definition
  - ***Any relationship between two measured quantities that renders them statistically dependent.***



- Heritability
  - ***The proportion of variance explained by genetics***
  - P = G + E + G*E
    - **Heritability > 0**



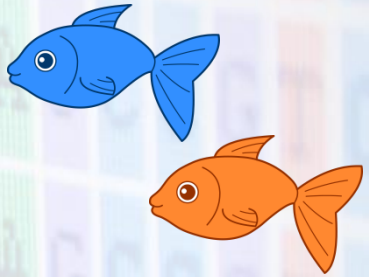

The case of the missing heritability

Why?

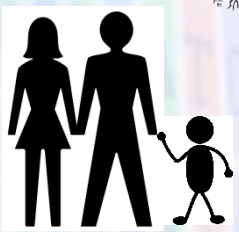
- Environment, Gene-Environment interactions
- Complex traits, small effects, rare variants
- Gene expression levels
- GWAS methodology?

Study Design

GWAS design



- Case-Control
  - Well-defined "case"
  - Known heritability

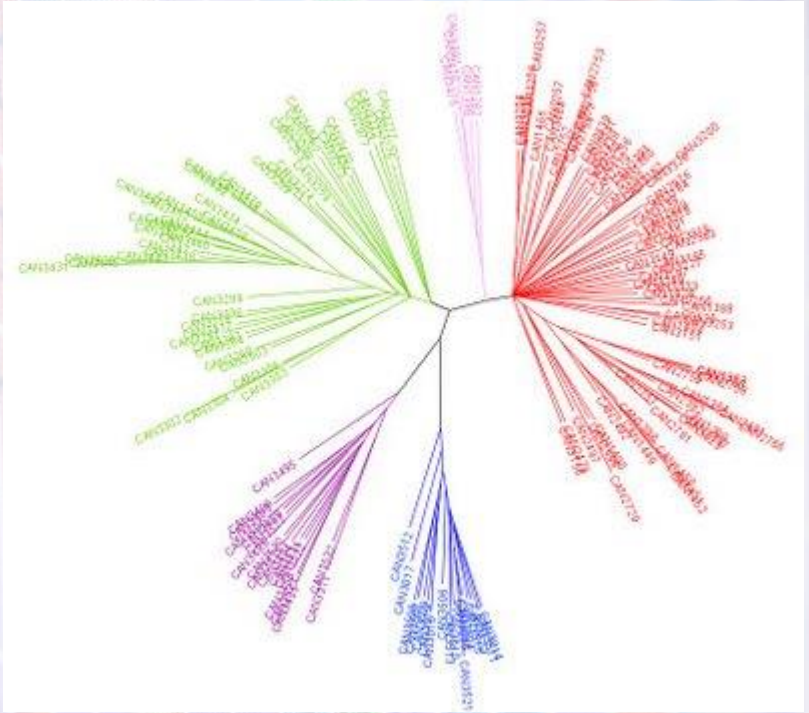


- Variations
  - Quantitative phenotypic data
    - Eg. Height, biomarker concentrations
  - Explicit models
    - Eg. Dominant or recessive

Issues & Considerations

- Data quality
  - 1% rule
- Controlling for confounding
  - Sex, age, health profile
  - Correlation with other variables
- Population stratification*
- Linkage disequilibrium*

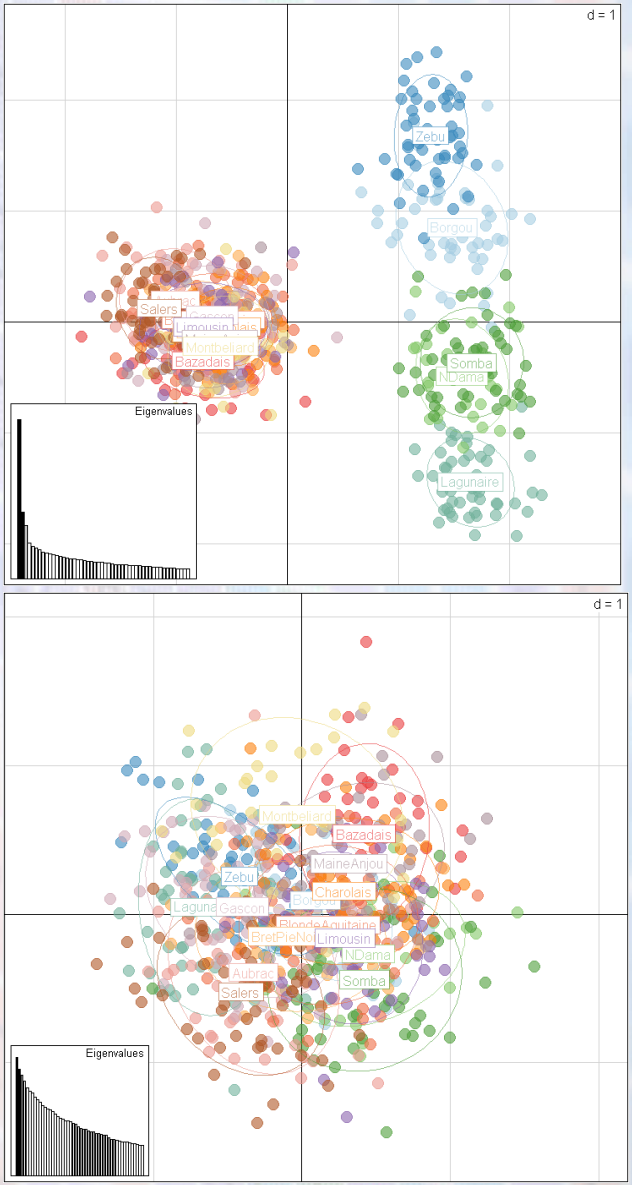
Population stratification



- Definition
  - **"Population stratification" = population structure**
  - ***Systematic difference in allele frequencies btw. sub-populations…***
    - … possibly due to different ancestry
- Problem
  - Violates assumed population homogeneity, independent observations
    - → **Confounding, spurious associations**
  - Case population more likely to be related than Control population
    - → **Over-estimation of significance of associations**

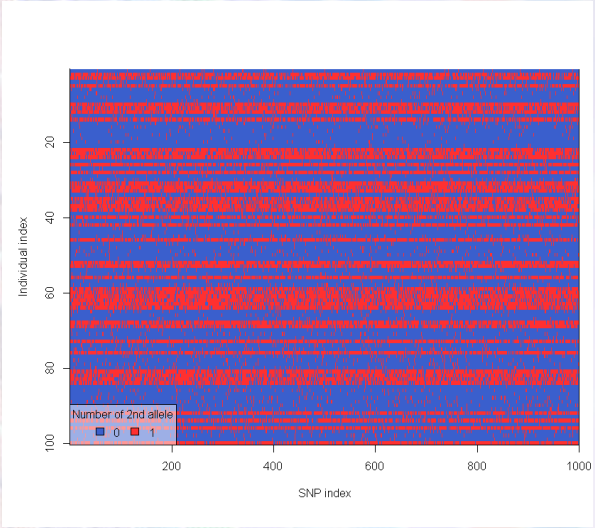
Population stratification II

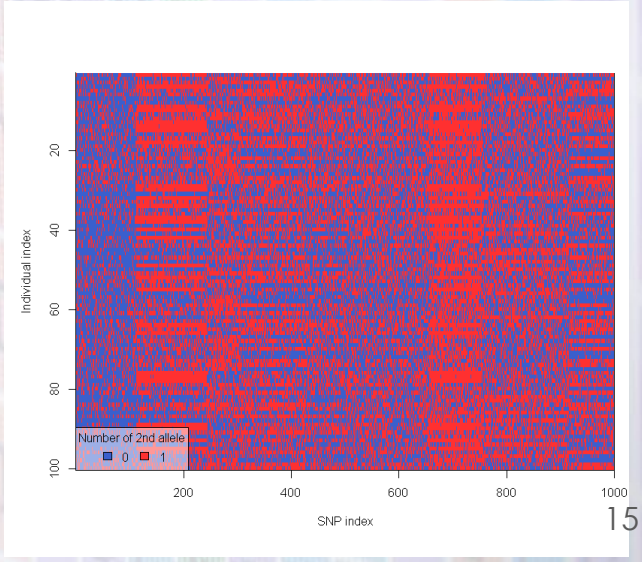
- Solutions
  - **Visualise**
    - Phylogenetics
    - PCA
  - **Correct**
    - Genomic Control
    - Regression on Principal Components of PCA



Linkage disequilibrium (LD)

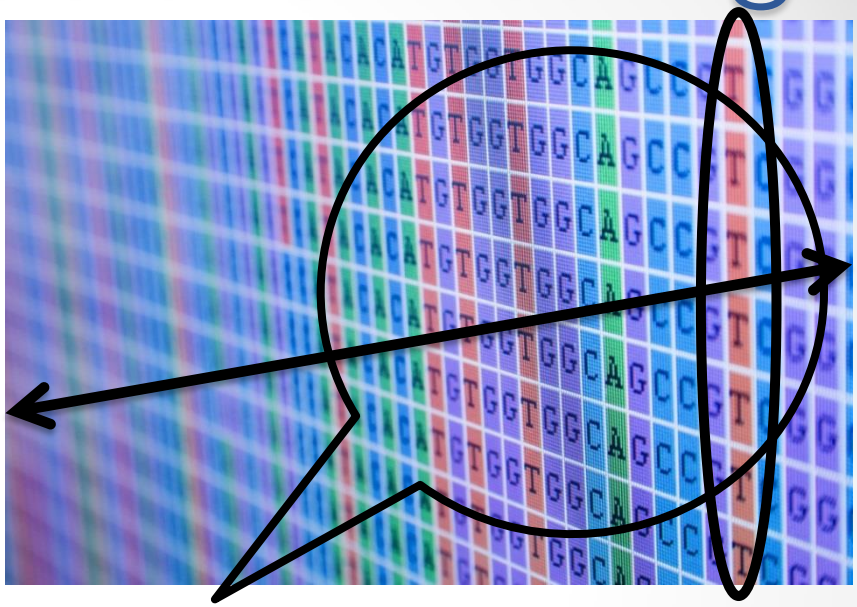
- Definition
  - ***Alleles at separate loci are NOT independent of each other***
- Problem?
  - **Too much LD is a problem**
    - → noise >> signal
  - **Some (predictable) LD can be beneficial**
    - → enables use of "marker" SNPs





Testing for Association …

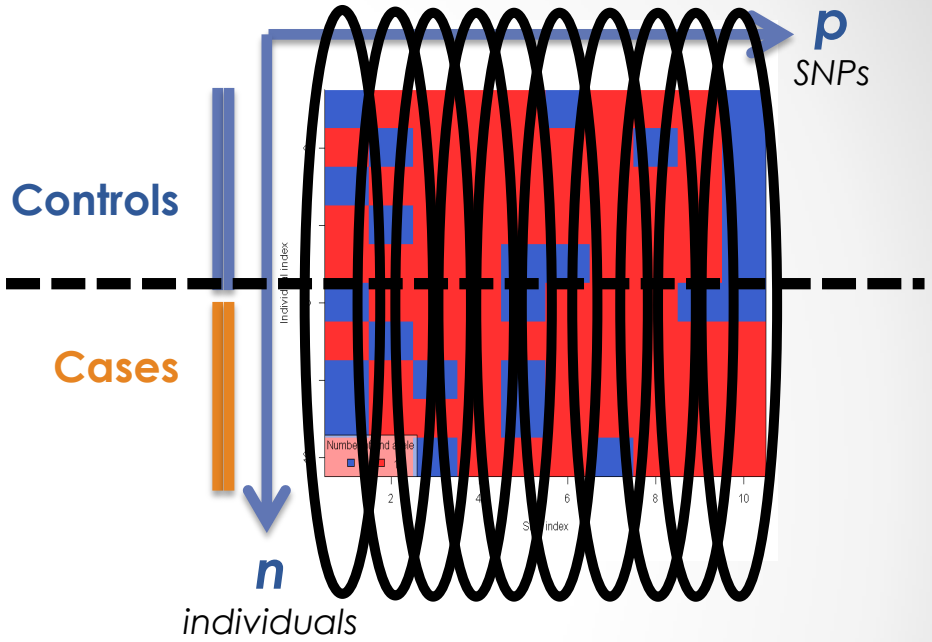
Methods for association testing



- Standard GWAS
  - **Univariate methods**
- Incorporating interactions
  - **Multivariate methods**
    - Penalized regression methods (LASSO)
    - Factorial methods (DAPC-based FS)

Univariate methods

- Approach
  - Individual **test statistics**
  - **Correction** for multiple testing



- Variations
  - **Testing**
    - Fisher's exact test, Cochran-Armitage trend test, Chi-squared test, ANOVA
    - Gold Standard—**Fischer's exact test**
  - **Correcting**
    - Bonferroni
    - Gold Standard—**FDR**

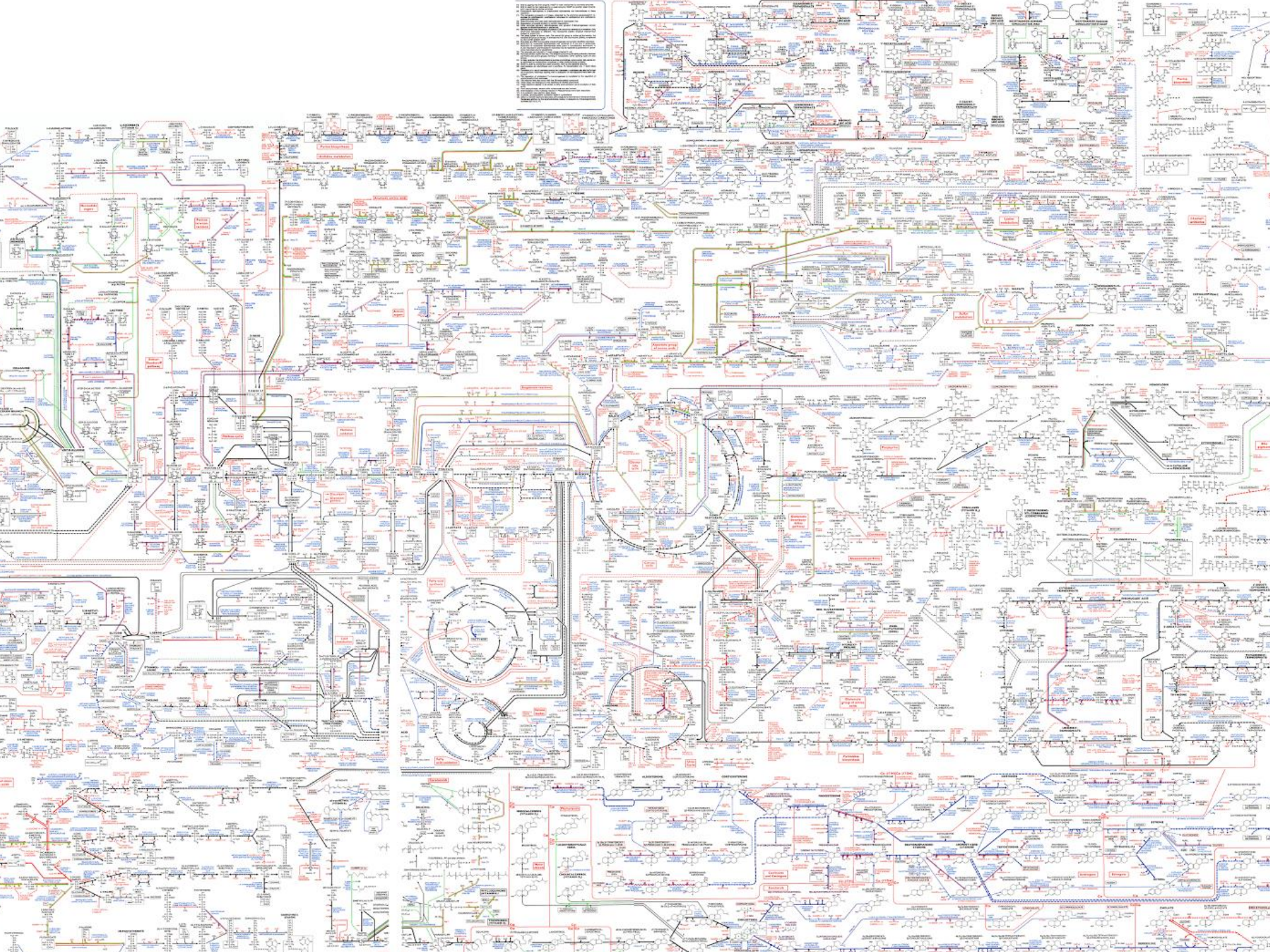
Univariate – Strengths & weaknesses

Strengths

- Straightforward
- Computationally fast
- Conservative
- Easy to interpret

Weaknesses

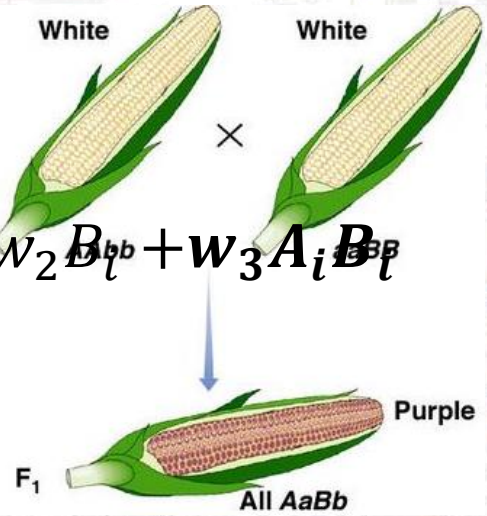
- Multivariate system, univariate framework
- Effect size of individual SNPs may be too small
- Marginal effects of individual SNPs ≠ combined effects

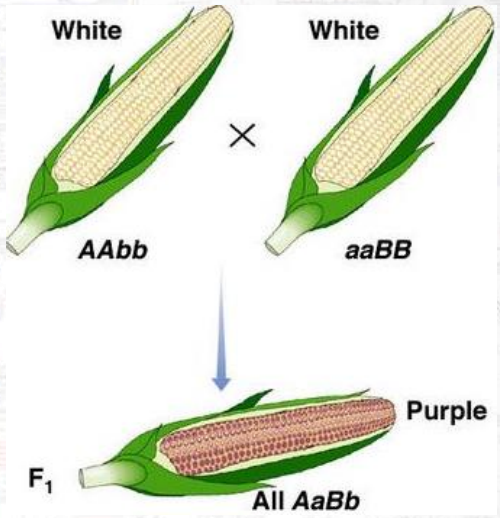


Interactions

- Epistasis
  - *"Deviation from li*
    general linear m

$$Y_i = w_0 + w_1 A_i + w_2 B_i + w_3 A_i B_i$$





- With *p* predictors, t
  - $\binom{p}{k} = \frac{p^k}{k!}$ k-way interactions
    - *p* = 10,000,000 → 5 x 10^{11}
      - That's **500 BILLION** possible pair-wise interactions!

- **Need some way to limit the number of pairwise interactions considered…**

Multivariate methods

Penalized Regression

LASSO penalized regression

The elastic net

Ridge regression

Neural Networks

Genetic programming optimized neural networks

Parametric decreasing metho

Logic Trees

Logic feature selection

Monte Carlo

Logic regression

Logic Regression

Modified Logic Regression-Gene Expression Programming

Genetic Programming for Association Studies

Set association approach

Bayesian Approaches

Bayesian partitioning

Bayesian Logistic Regression with Stochastic Search Variable Selection

Bayesian Epistasis Association Mapping

Factorial Methods

Sparse-PCA

Multi-factor dimensionality reduction method

Supervised-PCA

DAPC-based FS (snpzip)

Odds-ratio-based MDR

Non-parametric Methods

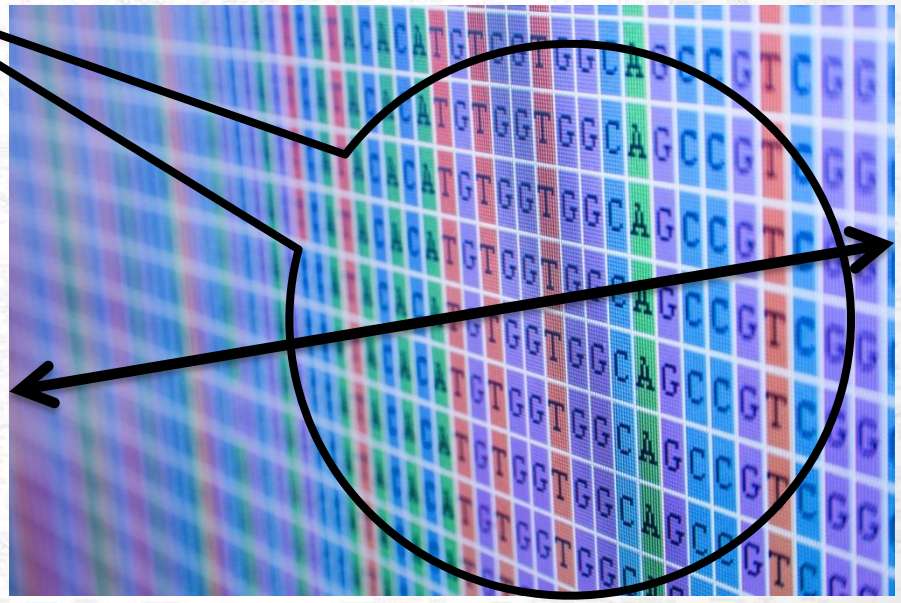
Random forests

Restricted partitioning method

Combinatorial partitioning method

Multivariate methods (ii)

- Penalized regression methods
  - **LASSO penalized regression**
- Factorial methods
  - **DAPC-based feature selection**

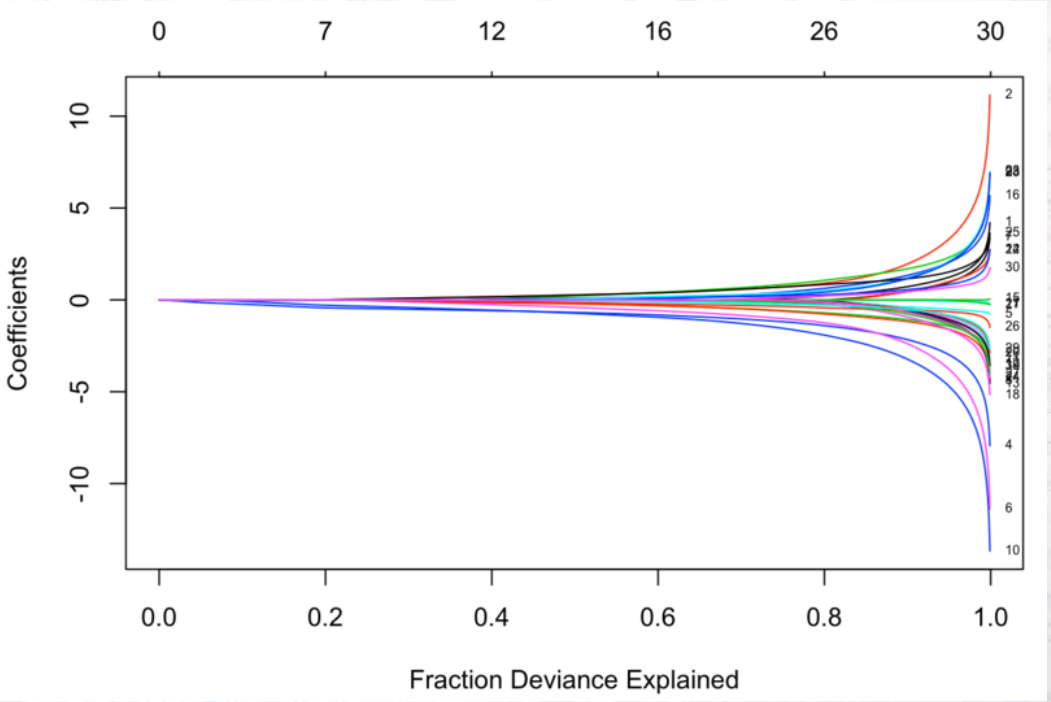


Penalized regression methods

- Approach
  - **Regression** models multivariate association
  - **Shrinkage estimation** → feature selection
- Variations
  - LASSO, Ridge, Elastic net, Logic regression
    - Gold Standard—**LASSO penalized regression**

LASSO penalized regression

- Regression
  - Generalized linear model ("**glm**")
- Penalization
  - **L1 norm**
  - Coefficients → 0
  - Feature selection!



LASSO – Strengths & weaknesses

Strengths

- Stability
- Interpretability
- Likely to accurately select the most influential predictors
- **Sparsity**

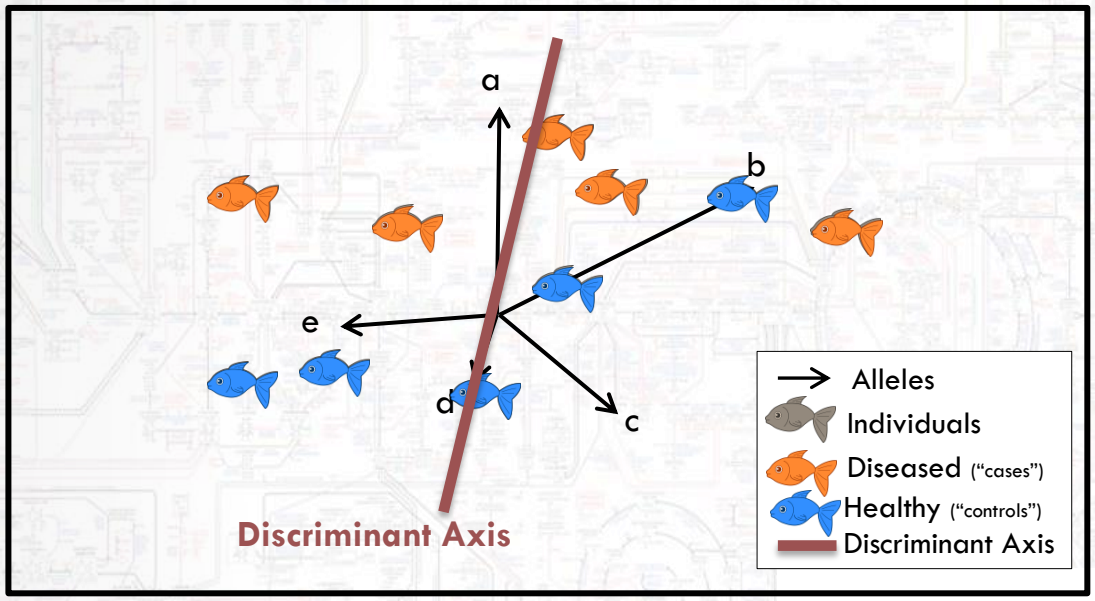
Weaknesses

- Multicollinearity
- Not designed for high-p
- Computationally intensive
- Calibration of penalty parameters
  - User-defined → variability
- **Sparsity**
- **NO p-values!**

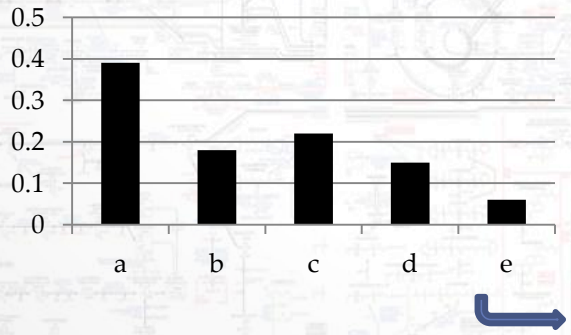
Factorial methods

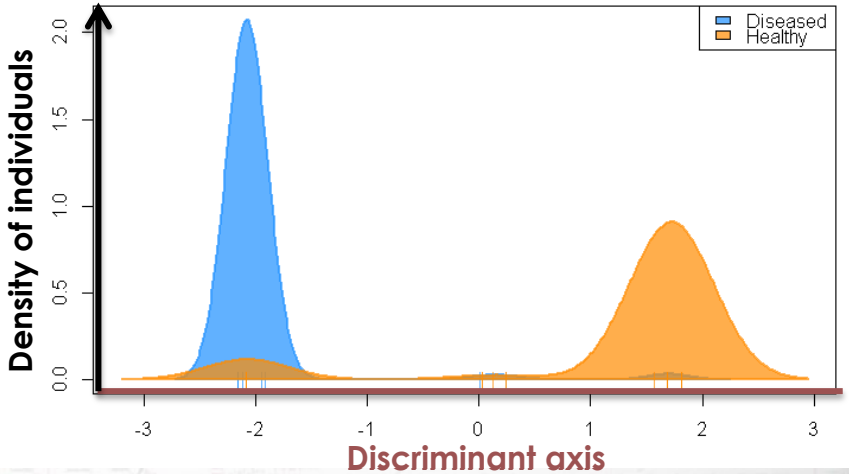
- Approach
  - Place all variables (SNPs) in a **multivariate space**
  - **Identify discriminant axis** → best separation
  - **Select variables** with the highest contributions to that axis
- Variations
  - Supervised-PCA, Sparse-PCA, DA, DAPC-based FS
  - Our focus—**DAPC with feature selection (snpzip)**

DAPC-based feature selection



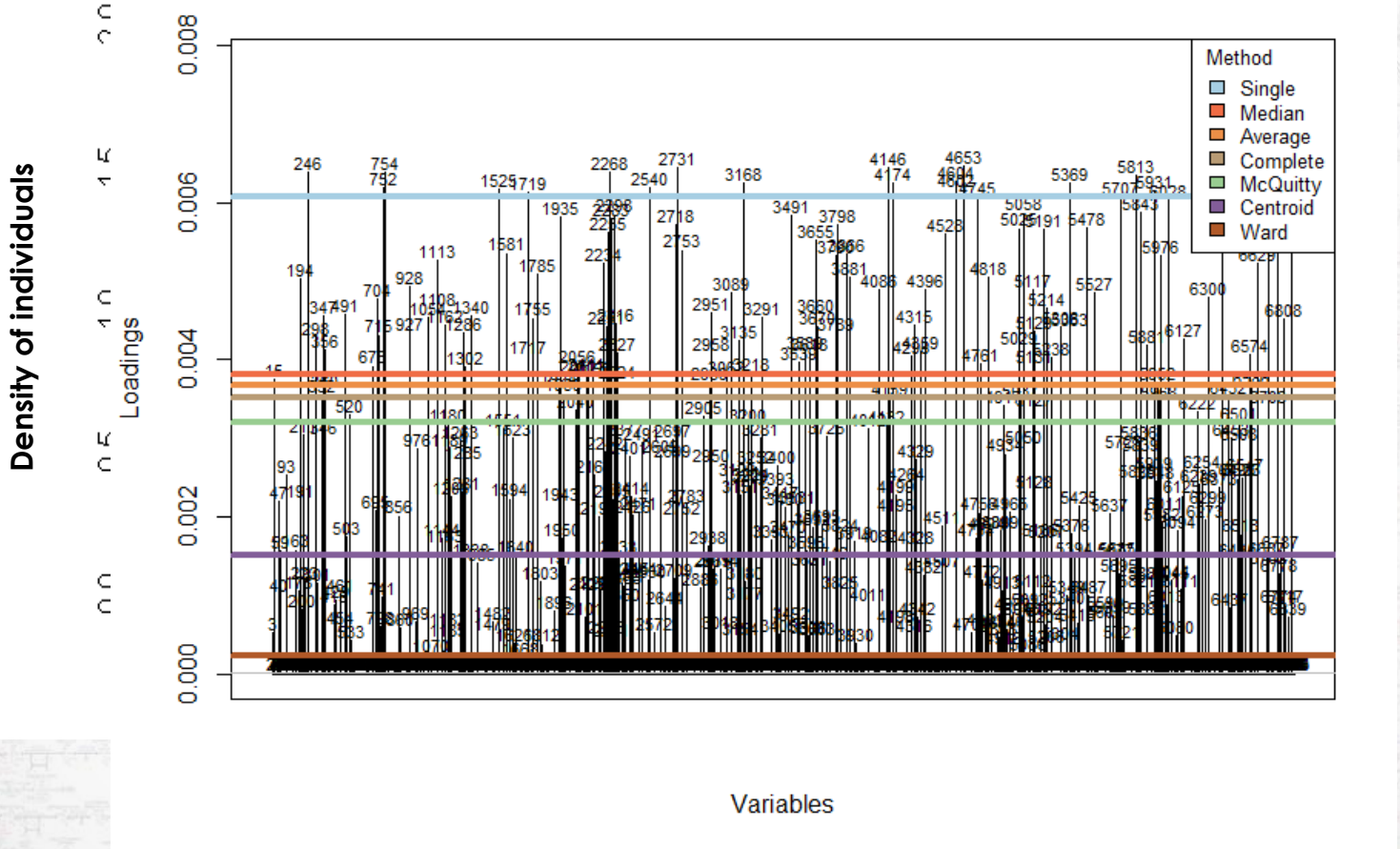
■ Contribution to Discriminant Axis





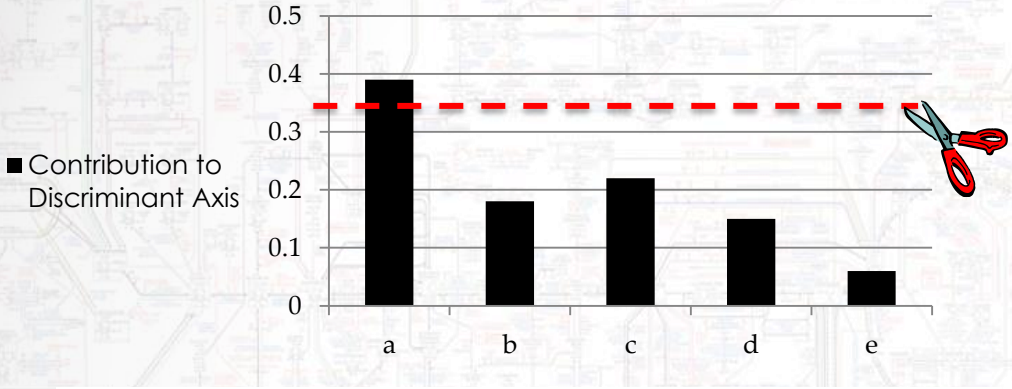
DAPC-based feature selection

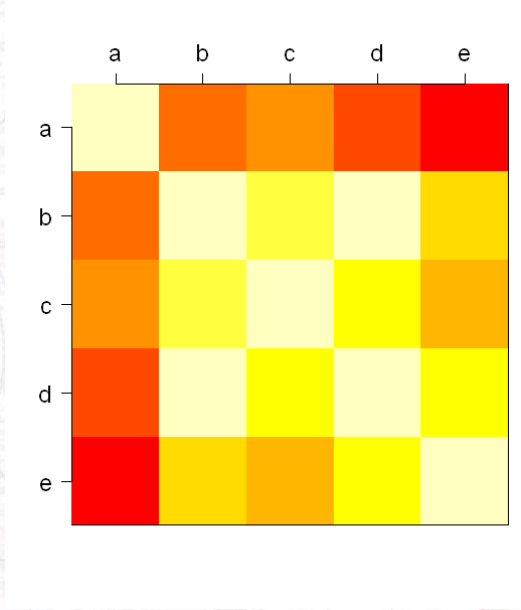
- Where should we draw the line?
  - → **Hierarchical clustering**

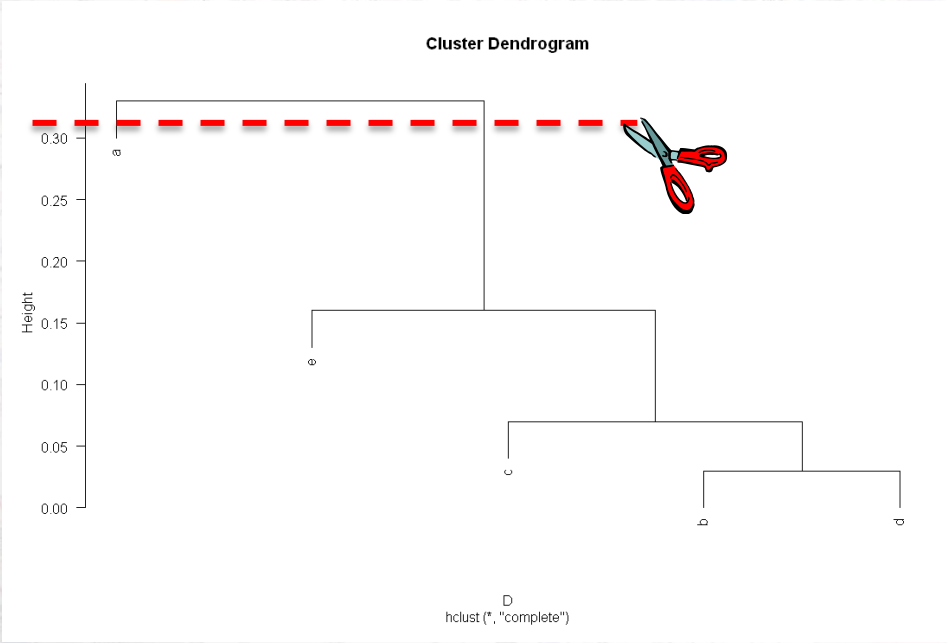


■ Contribution to

Hierarchical clustering (FS)








```
$FS
$FS$`Number of selected vs. unselected alleles`
[1]    1   4

$FS$`List of selected alleles`
[1] 1

$FS$`Names of selected alleles`
[1] "a"

$FS$`Contributions of selected alleles to discriminant axis`
1
0.39
```

DAPC – Strengths & weaknesses

Strengths

- More likely to catch **all** relevant SNPs (signal)
- Computationally quick
- Good ***exploratory*** tool
- **Redundancy > sparsity**

Weaknesses

- Sensitive to n.pca
- N.snps.selected varies
- No "p-value"
- **Redundancy > sparsity**

Conclusions

- Study design
  - **GWAS design**
  - **Issues and considerations in GWAS**
- Testing for association
  - **Univariate methods**
  - **Multivariate methods**
    - Penalized regression methods
    - Factorial methods

Thanks for listening!

Questions?