Lecture: GWAS and Population Stratification

Waseem Hussain
Postdoctoral Research Associate

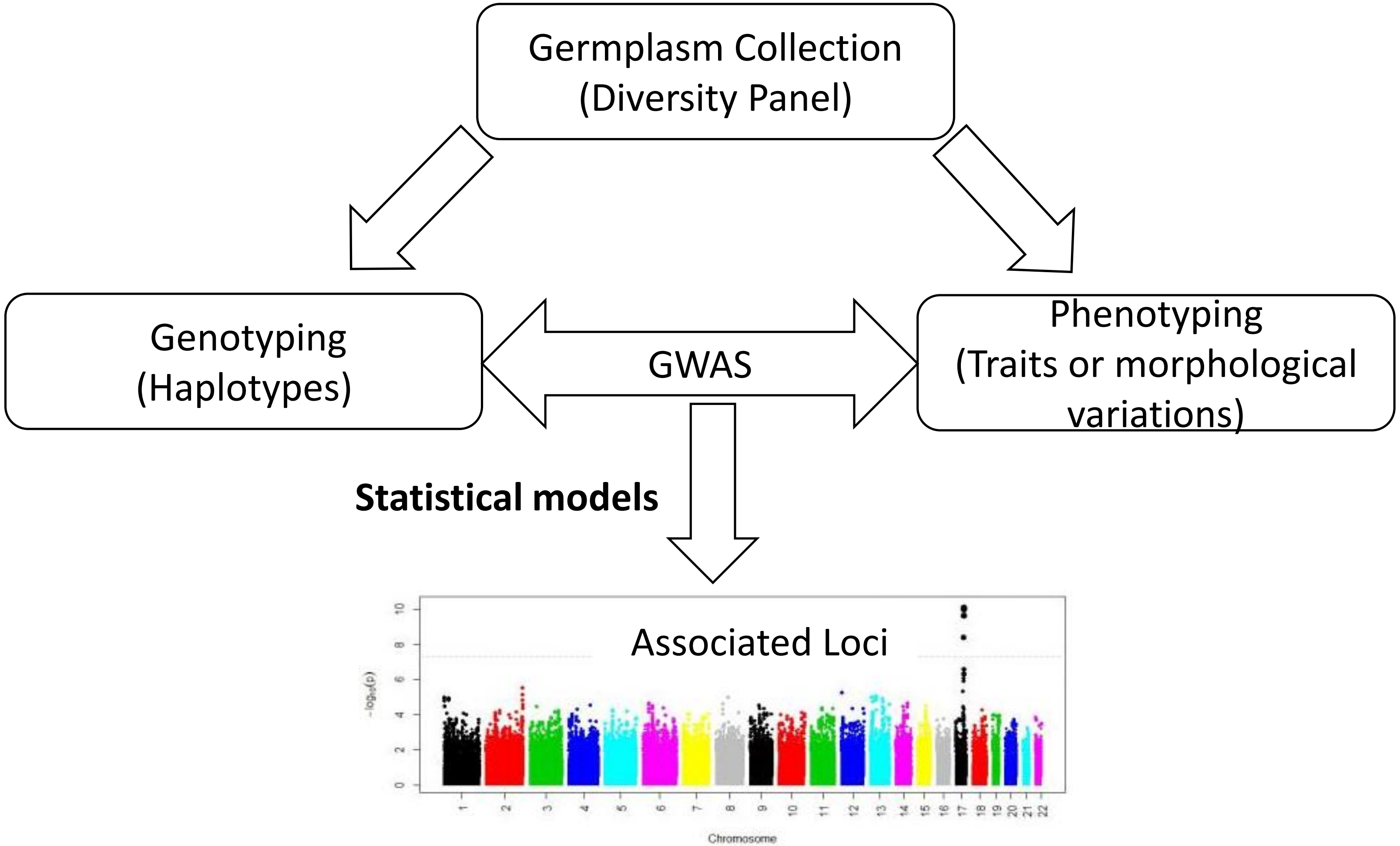
03/29/2018

Description

- What is GWAS and Work flow for GWAS
- Population stratification
- Methods to account for PS in GWAS
- Statistical methods for GWAS

Introduction

- A natural population survey to determine marker trait associations using genome-wide markers.
- Exploits LD between markers



Germplasm Collection (Diversity Panel)

Genotyping (Haplotypes)

GWAS

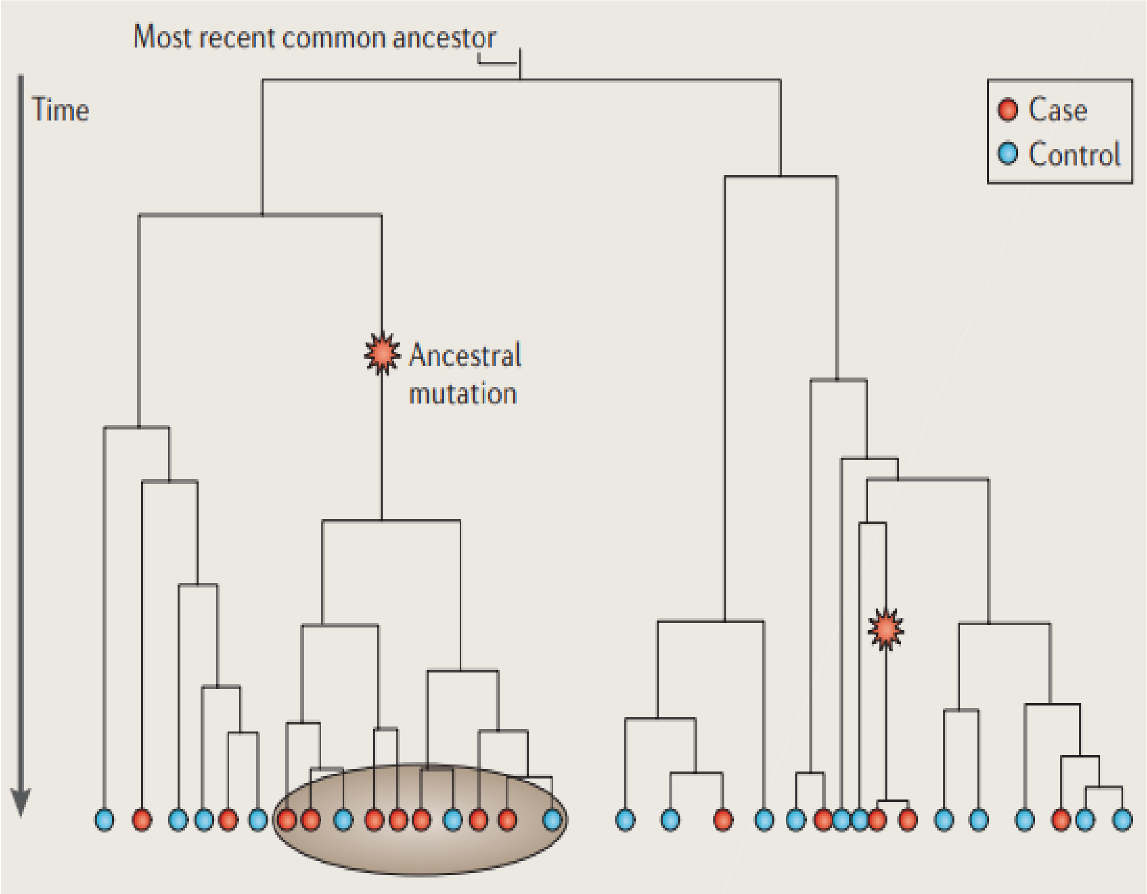
Phenotyping (Traits or morphological variations)

Statistical models

Associated Loci

Rational for Association mapping

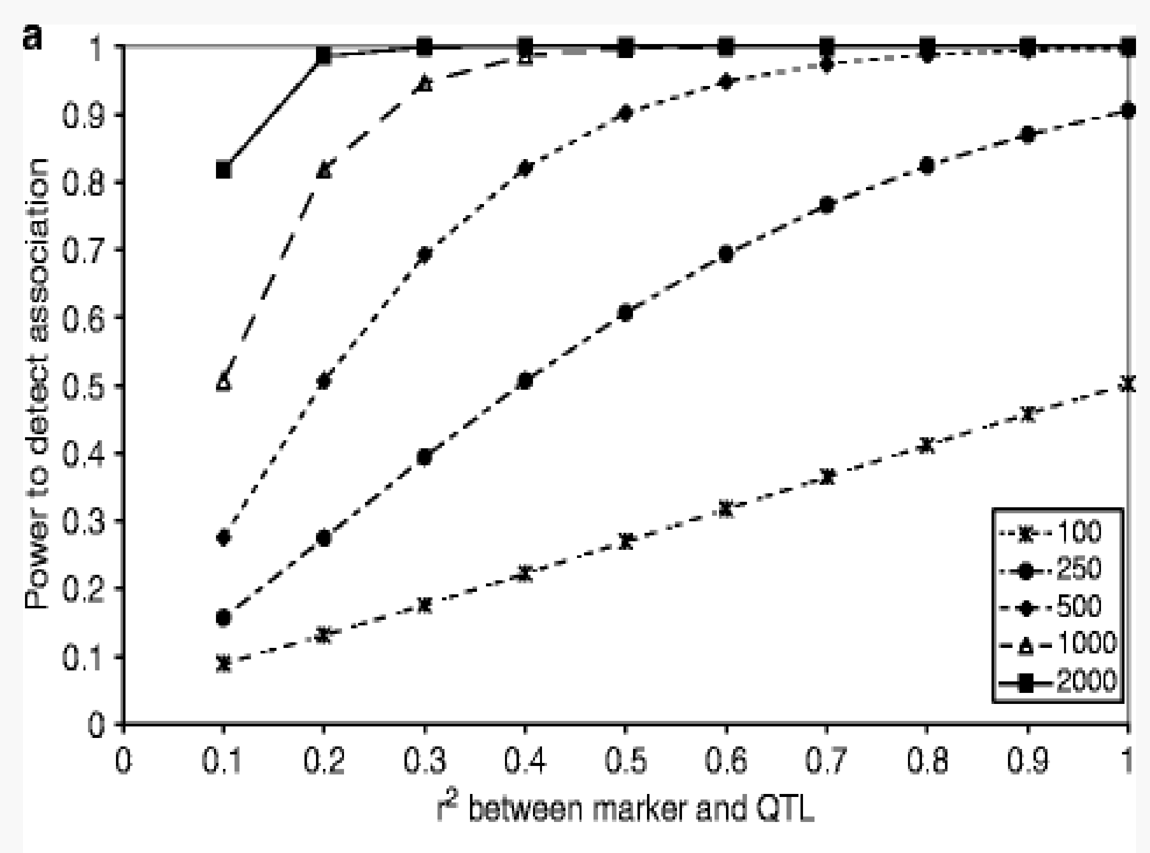
- Individuals should be unrelated, presumed to be distinct.
- Powerful for common variants and Minor allele frequency need to be > 5%



Balding, 2006
https://www.nature.com/articles/nrg1916.pdf

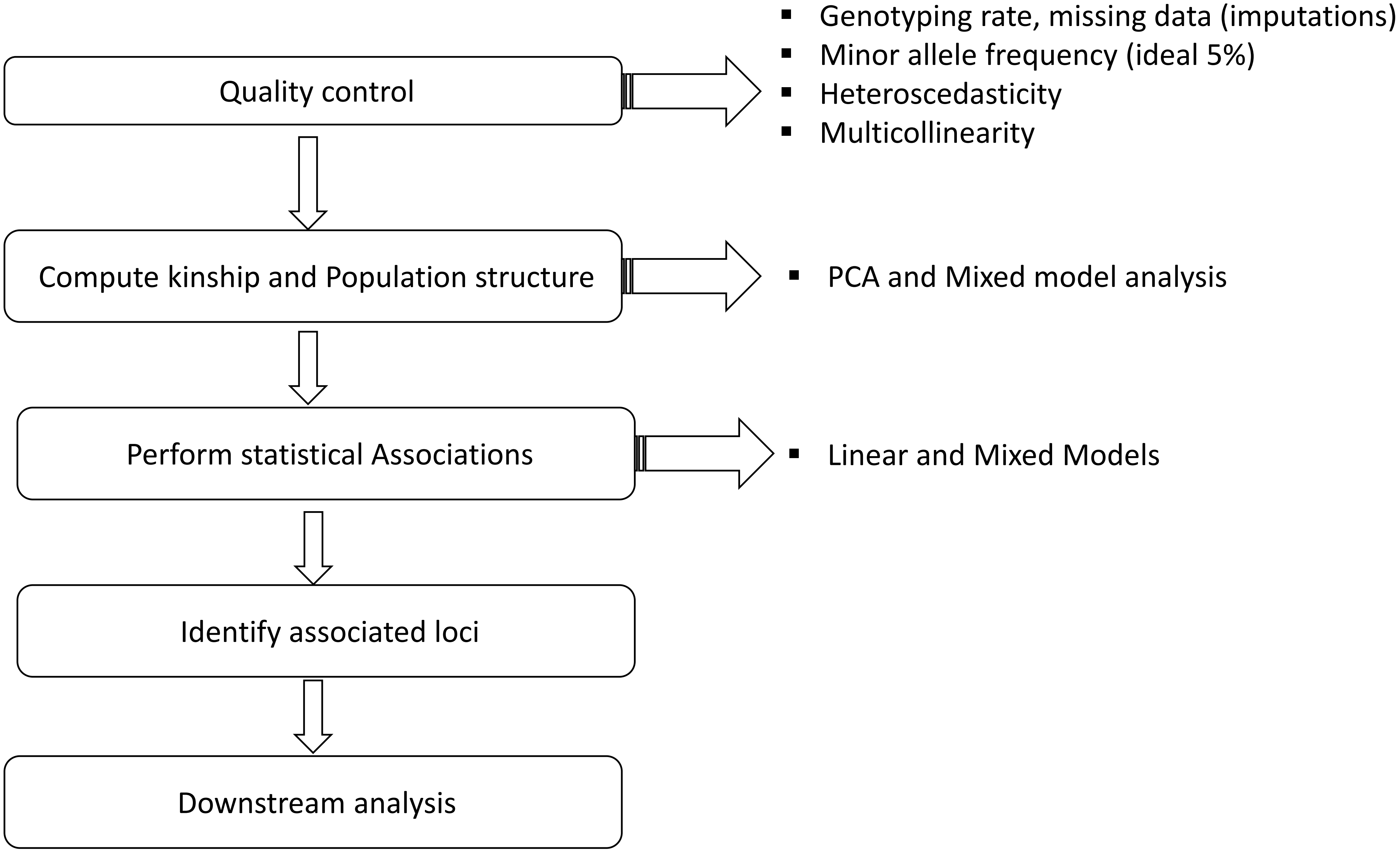
Rational for Association mapping

- Sufficiently large sample
- Polymorphic alleles covering whole genome
- Statistically powerful methods to detect genetic associations
- Individuals should be unrelated, presumed to be distinct.
- Powerful for common variants and Minor allele frequency need to be > 5%



Balding, 2006
https://www.nature.com/articles/nrg1916.pdf

Work flow for GWAS



Quality control

- Genotyping rate, missing data (imputations)
- Minor allele frequency (ideal 5%)
- Heteroscedasticity
- Multicollinearity

Compute kinship and Population structure

- PCA and Mixed model analysis

Perform statistical Associations

- Linear and Mixed Models

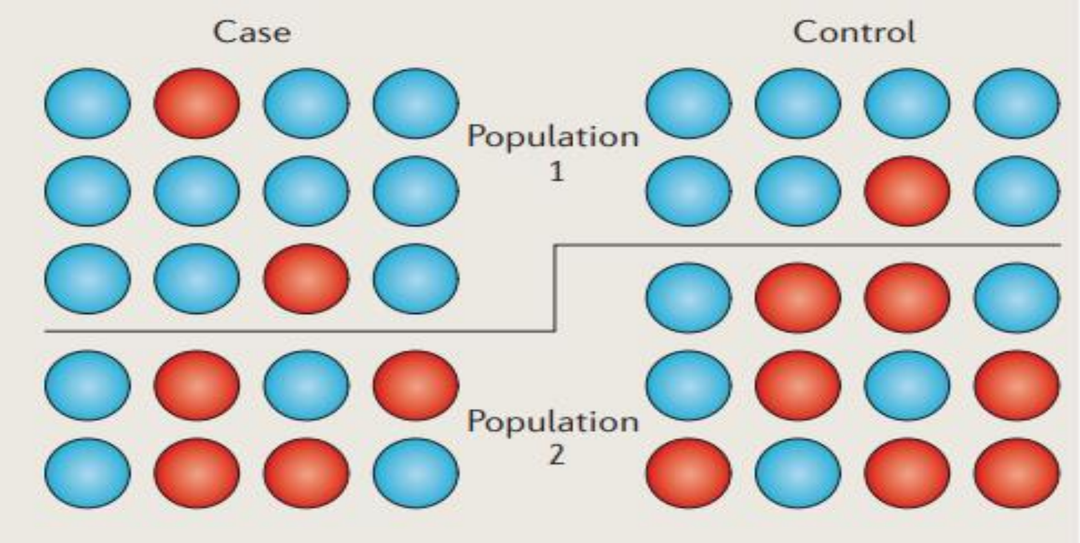
Identify associated loci

Downstream analysis

Population stratification

Difference in allele frequencies between sub-populations due to ancestry

- Can lead to spurious associations if allele frequencies vary between subpopulations..



- Test statistics inflated, high false positive rate
- Inflation of genomic heritability
- Overestimation of prediction accuracy

Balding, 2006
https://www.nature.com/articles/nrg1916.pdf

Methods to control Population stratification

- **Genomic Control**: Estimates inflation factor λ

  λ > 1 indicates stratification

  Limitation: λ same for all markers

- **Structured Association methods**: Assigns individuals to hypothetical subpopulations

  Correct number of subpopulations can never be fully resolved

- **Principle component analysis**: Provides fast and effective way to diagnose the population structure
- **Mixed-Model Approaches**: Involves Kinship and cryptic relatedness

Principle Component Analysis

- Reduce dimensions of data into few components.
- PCA is to find a new set of orthogonal axes (PCs), each of which is made up from a linear combination of the original axes
- Good in detecting major variations in data.
- PCA used in GWAS to generate axes of major genetic variation to account for structure.

How PCA is conducted to account for population structure

m individuals

n SNPs

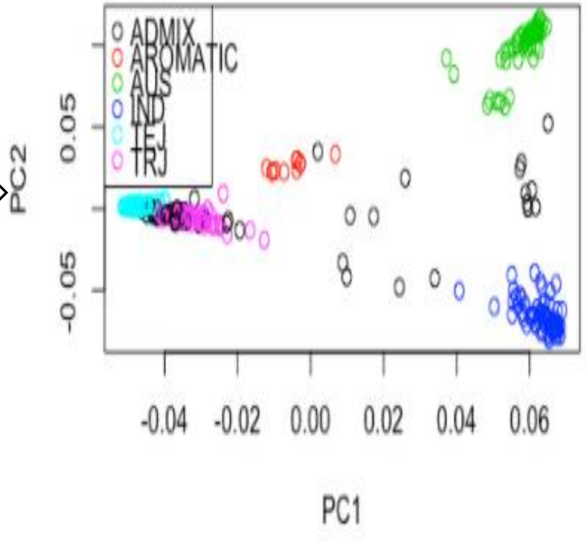
$$\begin{array}{cccccccccccccccc}
1 & 1 & 1 & 1 & 1 & 2 & 1 & 0 & 0 & 2 & 2 & 0 & 2 & 2 & 2 & 0 \\
0 & 0 & 1 & 0 & 2 & 2 & 2 & 0 & 0 & 2 & 2 & 2 & 2 & 2 & 2 & 0 \\
1 & 1 & 1 & 1 & 2 & 2 & 2 & 0 & 0 & 2 & 1 & 1 & 2 & 2 & 2 & 0 \\
1 & 1 & 2 & 0 & 2 & 2 & 1 & 0 & 1 & 2 & 2 & 1 & 2 & 2 & 2 & 0 \\
2 & 2 & 2 & 2 & 2 & 1 & 2 & 0 & 1 & 2 & 1 & 1 & 2 & 2 & 2 & 0 \\
1 & 1 & 1 & 0 & 1 & 2 & 1 & 0 & 1 & 2 & 2 & 1 & 2 & 2 & 2 & 0 \\
1 & 1 & 2 & 1 & 2 & 1 & 2 & 0 & 0 & 2 & 1 & 1 & 1 & 2 & 1 & 1 \\
2 & 2 & 2 & 1 & 1 & 1 & 1 & 0 & 1 & 2 & 1 & 0 & 1 & 2 & 1 & 1 \\
1 & 2 & 2 & 0 & 0 & 2 & 0 & 0 & 2 & 2 & 2 & 1 & 2 & 2 & 2 & 0
\end{array}$$

n x m matrix

Decompose the matrix

Eigen and Single Value Decomposition

PC1	PC2	....	PCm
.	.		.
.	.		.
.	.		.
.	.		.
.	.		.
.	.		.
.	.		.
.	.		.
.	.		.
.	.		.
.	.		.
.	.		.



Algorithm for PCA: Eigen and Single Value Decomposition

Step 1: Compute the variance-covariance as **$G= XX^T/N-1$**

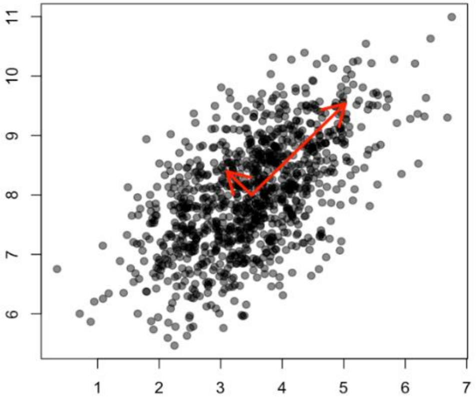
Step 2: Compute the Eigen decomposition of covariance matrix **($G=UDU^T$)**

Singular Value Decomposition **SVD ($X=U\sum V^T$)** (in case of m x n matrix and dense SNP data)

U= is an n x m orthogonal matrix of dimensions n x m

$\sum$= is a diagonal matrix of dimensions n x n

V= orthogonal matrix of n x n

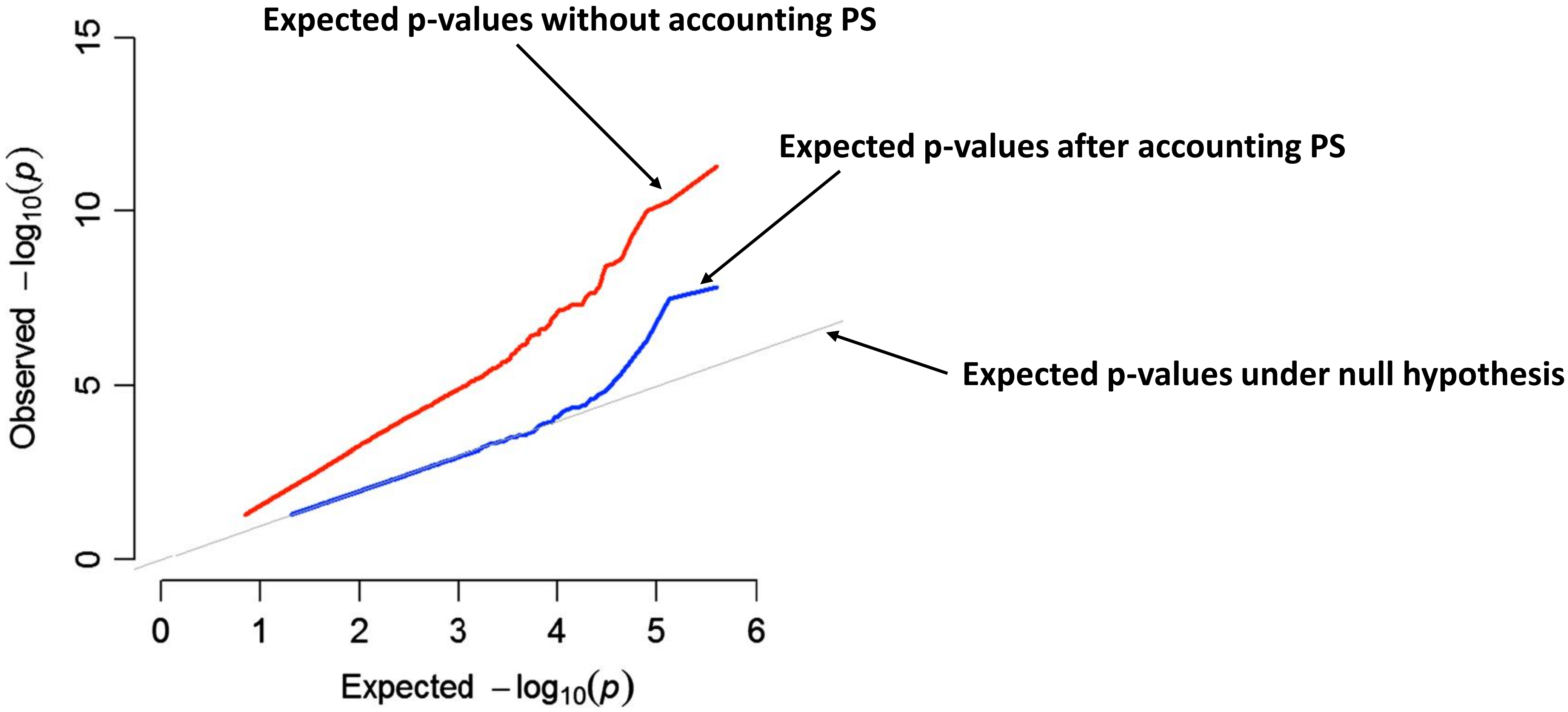


- Singular-decomposition picks out *directions in the data along which the variance is maximised*.
- Singular represent the variance of the data along these directions.

Step 3: Select the top K eigenvalues/PCs that are statistically significant

Step 4: Include the significant eigenvectors in the linear regression model or genotype matrix in mixed model.

Accounting for Population structure



Expected p-values without accounting PS

Expected p-values after accounting PS

Expected p-values under null hypothesis

Q-Q plot of p-values

- PCA only accounts for fixed effects of genetic ancestry
- Does not account for relatedness between individuals.

- Mixed Models
- Use both fixed effects (candidate SNPs and fixed covariates) and random effects (the Genotypic covariance matrix)

$$y= Wa + u+\varepsilon$$

$$var(u)= б^2K$$

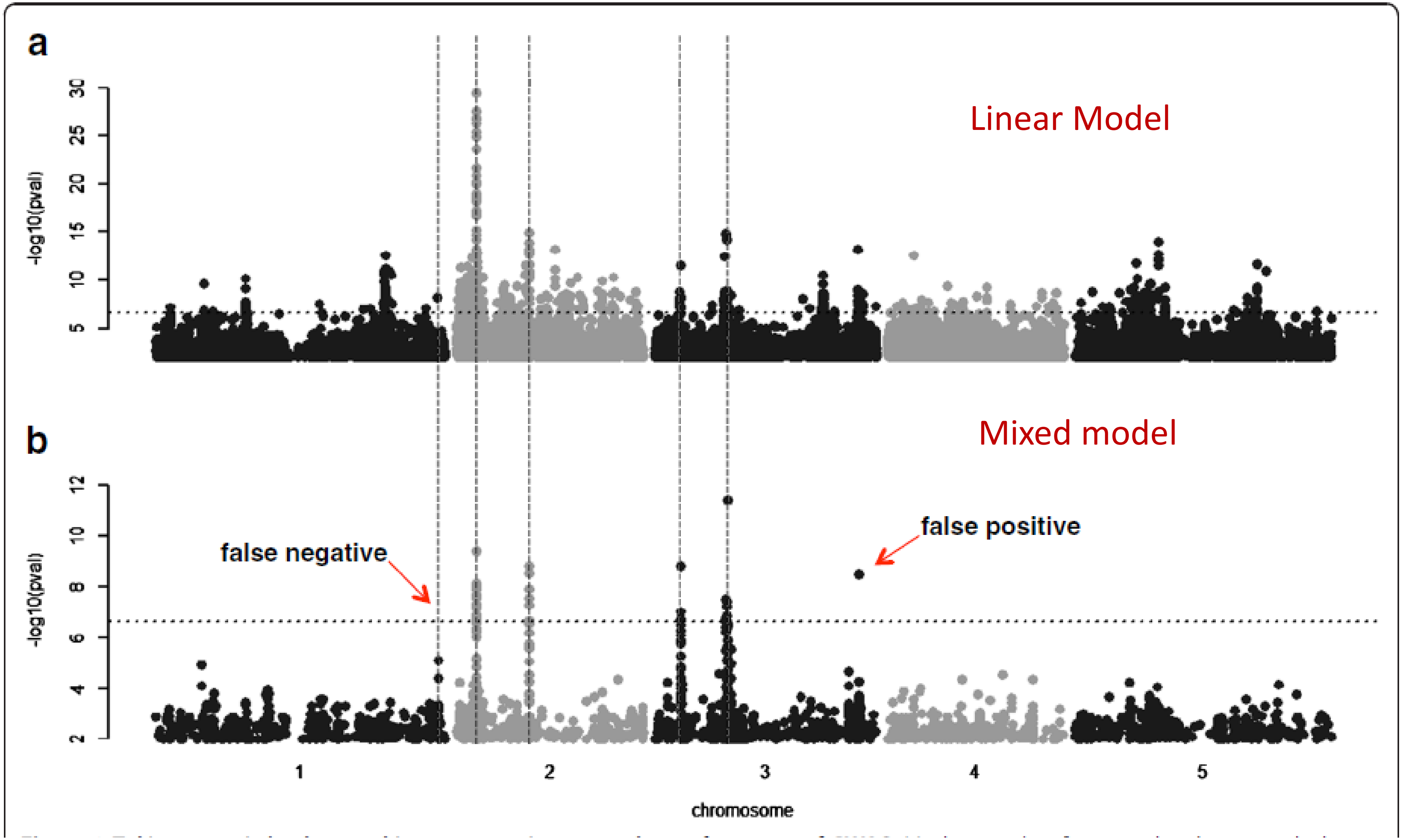
- K is Kinship matrix (pairwise genomic similarity of Individuals)
- Structure of Kinship matrix reflects:

Population structure

Family structure

and Cryptic Relatedness

Peterson et al. 2010
doi: 10.1038/nrg2813



Linear Model

Mixed model

false negative

false positive

GWAS using linear model and Mixed model

Korte and Farlow Plant Methods 2013, 9:29
http://www.plantmethods.com/content/9/1/29

Statistical methods for GWAS

Ordinary least squares

Model: y= Wa + e

To find “a”, effective size of SNP, we minimize the residual sum of squares.
And least square estimator of “a” is given as

$$\hat{\mathbf{a}} = (\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}'\mathbf{y}$$

â is the vector of regression coefficient for markers, i.e., effect size of SNPs
if the Gauss-Markov theorem is met, E[â]=a → BLUE

$$E[\epsilon] = 0, Var[\epsilon] = \mathbf{I}\sigma_\epsilon^2$$

No. of SNPs (n) is greater than individuals (m) n>>>m

(W`W)$^{-1}$ Does not exist, matrix is singular

Assumptions for Guass-Markov to hold true

- Population parameter linear
- No collinearity
- Homoskesdactic errors

Single marker regression

- One marker at a time tested for significance
- Problem: Marker effect may be exaggerated

The expectation of â is

$$E(\hat{\mathbf{a}}|\mathbf{W}) = (\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}'E(\mathbf{y}) = (\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}'\mathbf{W}\mathbf{a} = \mathbf{a}$$

OLS estimate for single SNP model

$$\hat{a}_1 = (\mathbf{w}_1'\mathbf{w}_1)^{-1}\mathbf{w}_1'\mathbf{y}$$

$$\begin{aligned} E(\hat{a}_1|\mathbf{w}_1) &= (\mathbf{w}_1'\mathbf{w}_1)^{-1}\mathbf{w}_1'E(\mathbf{y}) \\ &= (\mathbf{w}_1'\mathbf{w}_1)^{-1}\mathbf{w}_1'[\mathbf{w}_1\mathbf{a}_1 + \mathbf{w}_2\mathbf{a}_2] \\ &= (\mathbf{w}_1'\mathbf{w}_1)^{-1}\mathbf{w}_1'\mathbf{w}_1 a_1 + (\mathbf{w}_1'\mathbf{w}_1)^{-1}\mathbf{w}_1'\mathbf{w}_2 a_2 \\ &= a_1 + (\mathbf{w}_1'\mathbf{w}_1)^{-1}\mathbf{w}_1'\mathbf{w}_2 a_2 \end{aligned}$$

- OLS is biased if full model holds but fit a mis-specified model
- the same applies when there are more than two SNPs

Linear mixed models for GWAS

- Single marker-based mixed model association (MMA)
- Fit one marker at a time (Yang et al. 2014)

$$\mathbf{y} = \mu + \mathbf{w_j}\mathbf{a_j} + \mathbf{Z}\mathbf{g} + \epsilon$$
$$\mathbf{g} \sim N(0, \mathbf{G}\sigma_g^2)$$

- G (genomic relation matrix) captures population structure and polygenic effects

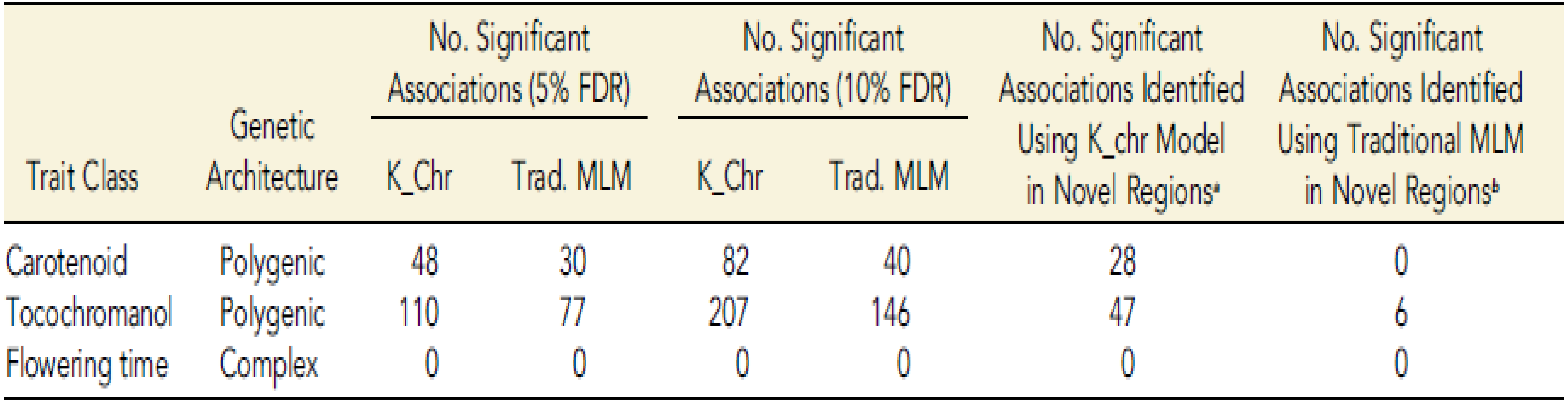
- **Double counting/fitting**
  SNP appears twice in model (once fixed and other time random)
  Candidate/tested markers used to calculate structure and family relatedness

- Alternatively,
- Exclude candidate markers from G, using model one chromosome out

$$\mathbf{y} = \mu + \mathbf{w_j}\mathbf{a_j} + \mathbf{Z}\mathbf{g} + \epsilon$$
$$\mathbf{g} \sim N(0, \mathbf{G}_{-k}\sigma_{g_{-k}}^2)$$

where −k denotes the kth chromosome removed

Comparison of K_Chr model and traditional Unified Mixed Linear Model in the Goodman diversity panel (Maize diversity panel of 281 lines)



		No. Significant Associations (5% FDR)		No. Significant Associations (10% FDR)		No. Significant Associations Identified Using K_chr Model in Novel Regions[a]	No. Significant Associations Identified Using Traditional MLM in Novel Regions[b]
Trait Class	Genetic Architecture	K_Chr	Trad. MLM	K_Chr	Trad. MLM		
Carotenoid	Polygenic	48	30	82	40	28	0
Tocochromanol	Polygenic	110	77	207	146	47	6
Flowering time	Complex	0	0	0	0	0	0

Chen and Lipka, 2016
doi:10.1534/g3.116.029090/-/DC1

Multiple marker models

- Single marker fitting cannot capture the effect of allele due to imperfect LD lead to inflation of type 1 errors particularly using dense SNP set.
- Multiple testing problems.

Multiple Marker models can overcome these:

- Fits all SNPs simultaneously as random effects

$$y_i = \mu + \sum_{j=1}^{n\_\mathrm{SNP}} b_j x_{ij} + e_i .$$

- Distribution assumption for markers varies from model to model

Schmid and Bennewitz, 2017

Demonstration in R