Genome-Wide Association Mapping and Population Stratification

Waseem Hussain
Postdoctoral Research Associate

03/05/2019

Description

- High-throughput phenotyping
- Basic Concepts of Association Mapping
- Work flow for Genome-wide association mapping (GWAS)
- Population stratification
- Methods to account for Population stratification (PS) in GWAS
- Statistical methods for GWAS

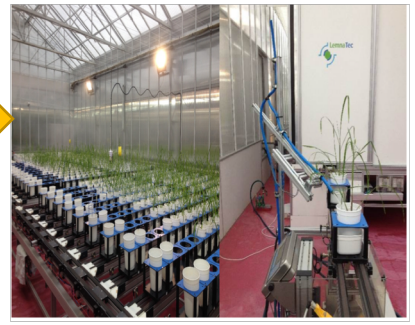
High-throughput Phenotyping

Core collections

Aerial-based Phenotyping

Green-house Phenotyping

Ground-based Phenotyping



Through-out the Growth Stage of Crop

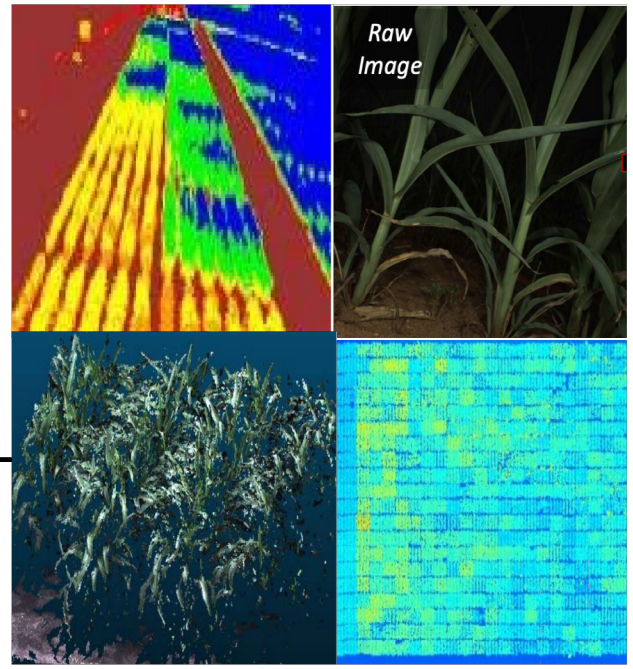


Image acquisition and Processing

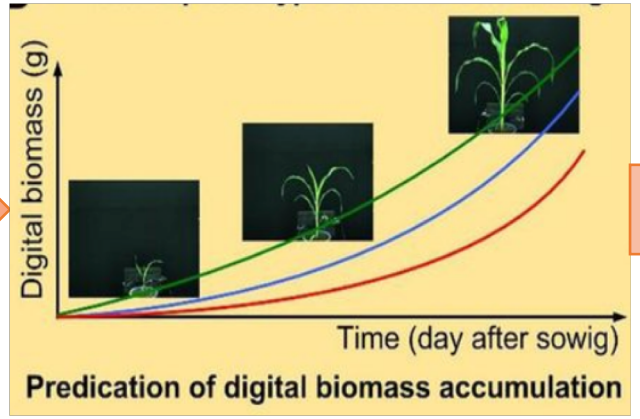
Trait Extraction

Seedling vigor
Green leaf area
NDVI
Leaf senescence
Plant height
Biomass

High-throughput Phenotyping

Seedling vigor
Green leaf area
NDVI
Leaf senescence
Plant height
Biomass

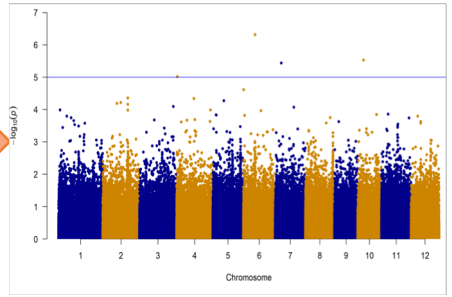
Growth Dynamics and Modelling of Traits



Phenotypes

Trait Dissection and Identification of Loci

Genotypes



Why Mapping genes?

Find markers closely associated with gene for marker assisted gene introgression or predict the breeding value of line.

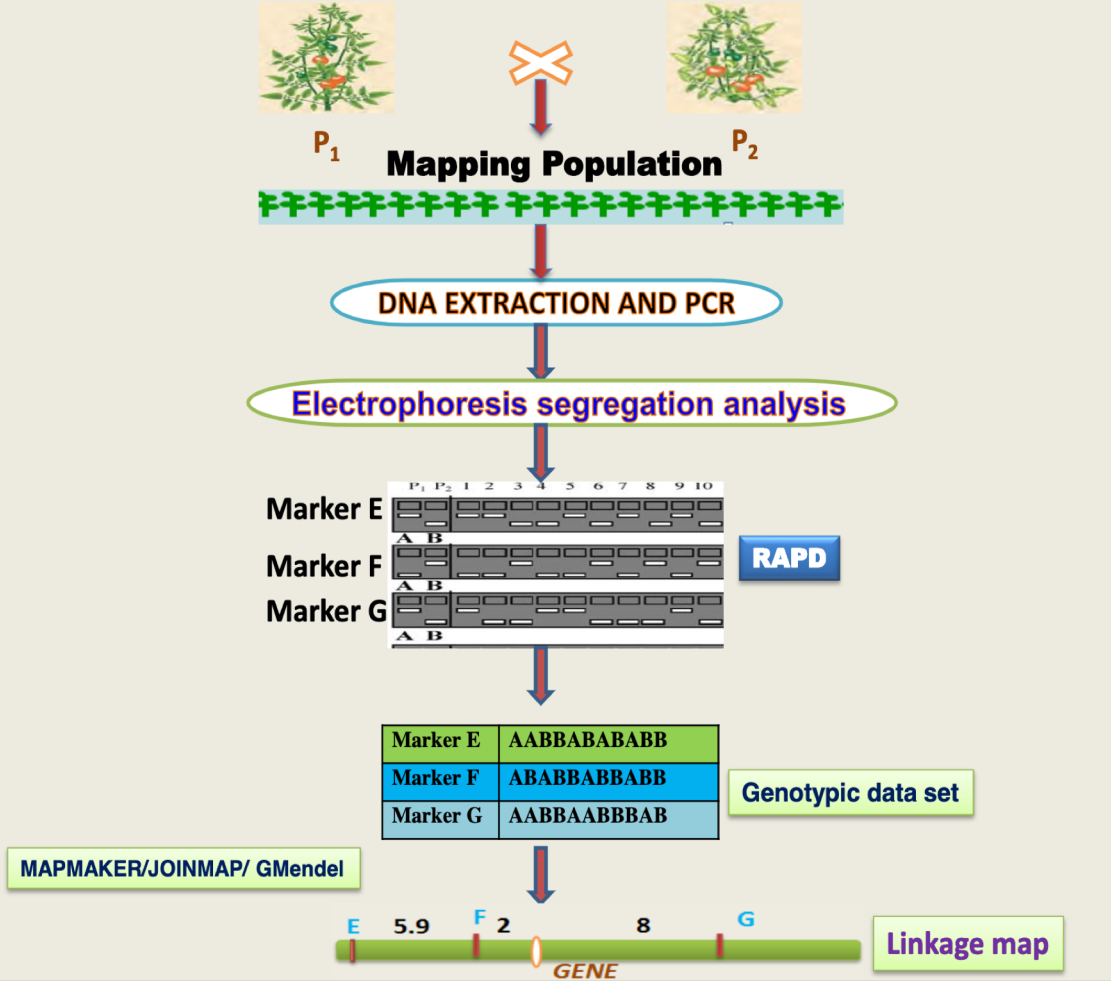
Two Main Approaches

Family-based Linkage Mapping

LD-based Association Mapping

Family Based-Linkage Mapping

Greatly successful for major genes and rare variants

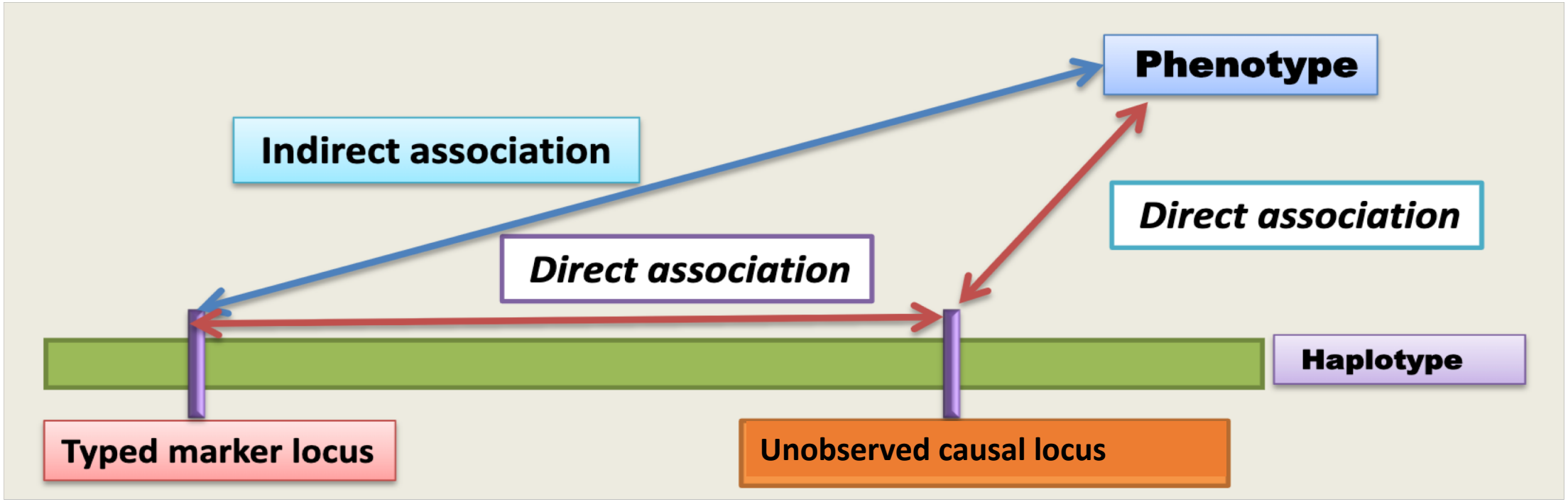


Drawbacks

- Small fraction of variation.
- Only alleles differing between parents.
- Low map genetic resolution-due to limited recombination.
- Inconsistency across mapping populations
- Linked markers not suitable for un-related genotypes.

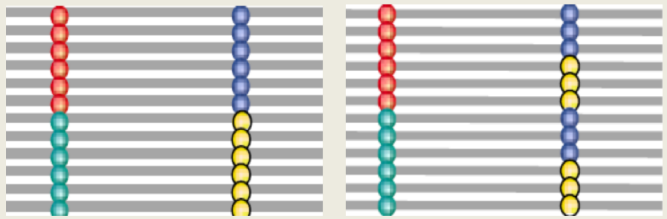
Linkage Disequilibrium -based Association Mapping

- A natural population survey to determine marker trait associations using genome-wide markers.
- Exploits Linkage Disequilibrium (LD) between markers.
- LD is defined as non-random association of alleles.
- Power depends upon degree of LD between marker and functional variant.



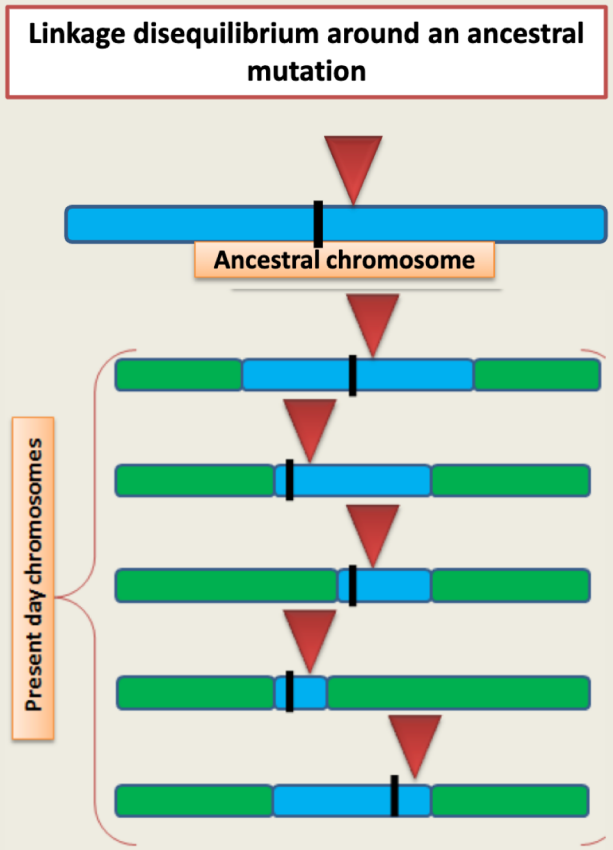
What is Linkage Disequilibrium

➢Non- random association of alleles at adjacent loci….



➢The closer two markers are, the stronger the LD

The resolution with which a QTL can be mapped is a function of how quickly LD decays over distance.



LD measures

Commonly used to quantify LD is r^2

$$r^2 = \frac{D^2}{p_A(1-p_A)p_B(1-p_B)}$$

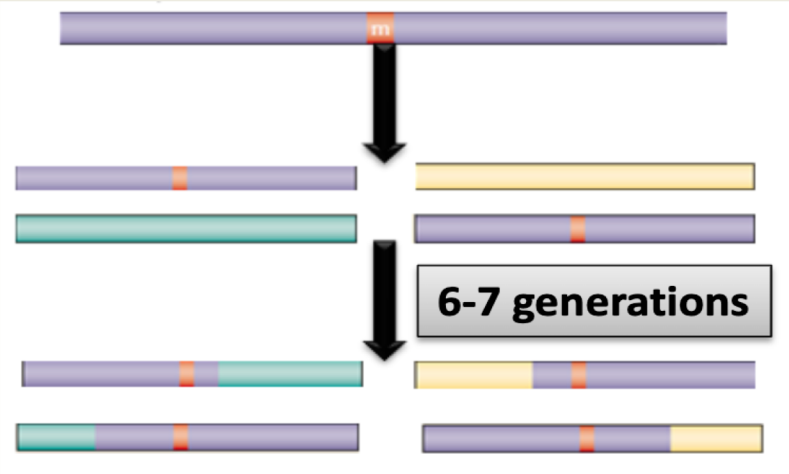
$$D = p_{AB} - p_A p_B$$
$$= p_{AB}p_{ab} - p_{Ab}p_{aB}$$

Advantages of Association mapping

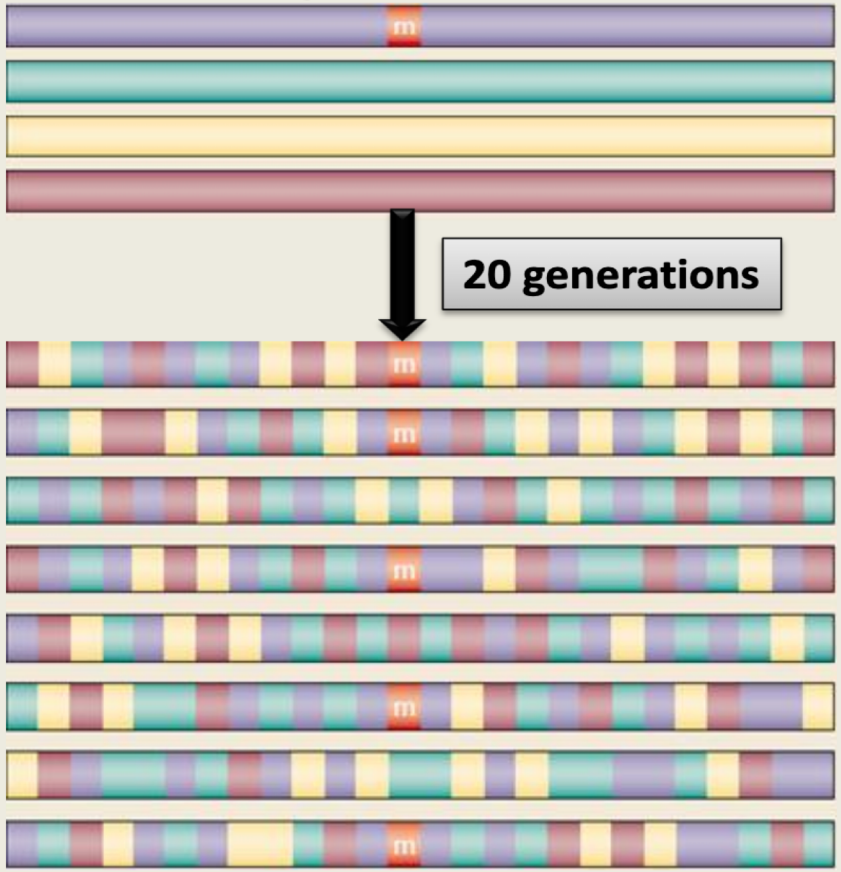
	Conventional	LD mapping
Mapping population	Biparental, structured	Natural/ breeding pool, not structured
Meiosis cycle	Few (6-7)	Several
QTL precision	Less	High –Great resolution
Trait variation	Explains between parents	Natural
LD break up	Less	more
Perennial crops	Not applicable	Effective
Markers	Specific	Diverse genotypes
Cost and ease	More cost and labour	Less cost and reduced time

Linkage vs Association mapping: How it leads to high resolution..

LINKAGE

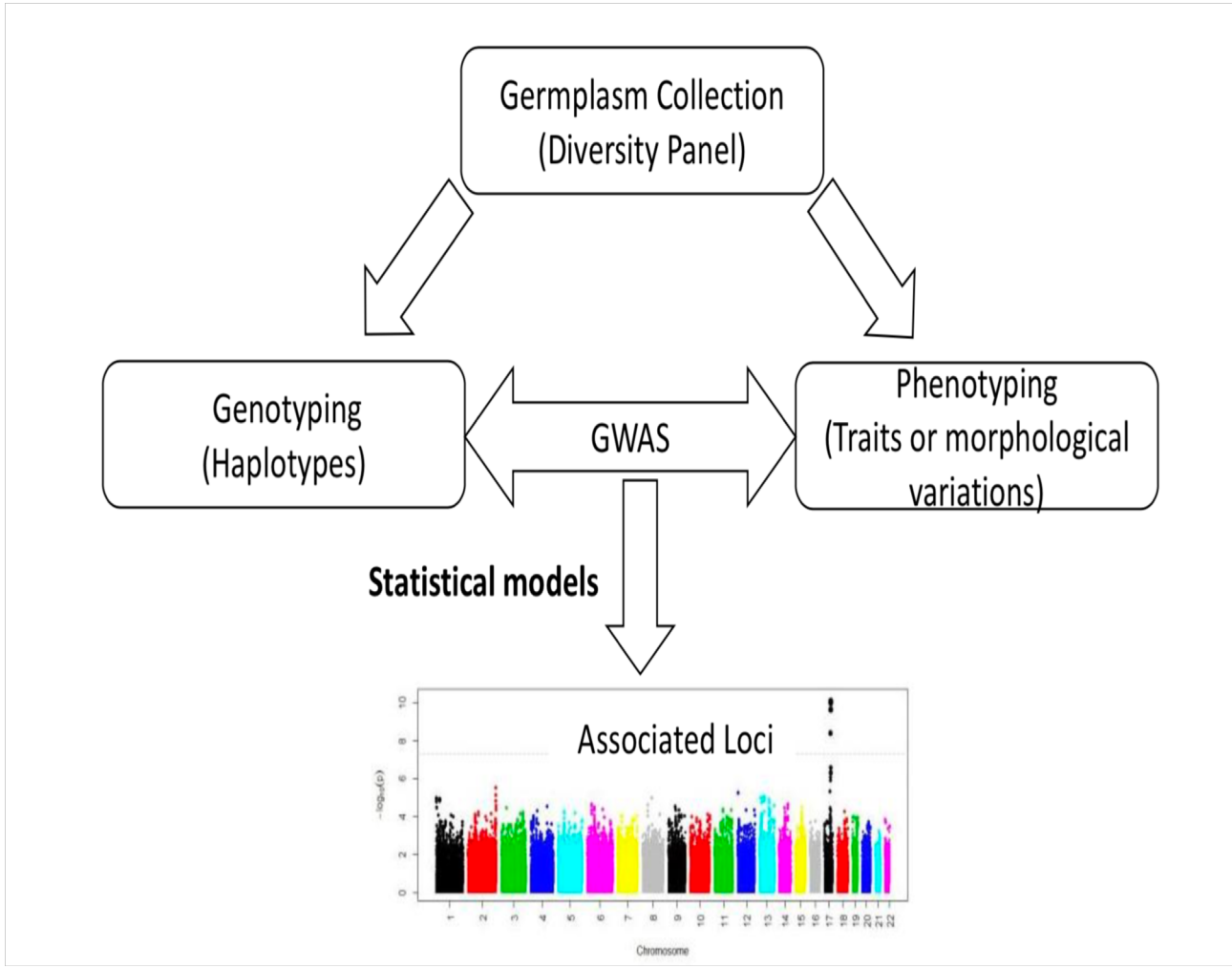


ASSOCIATION MAPPING



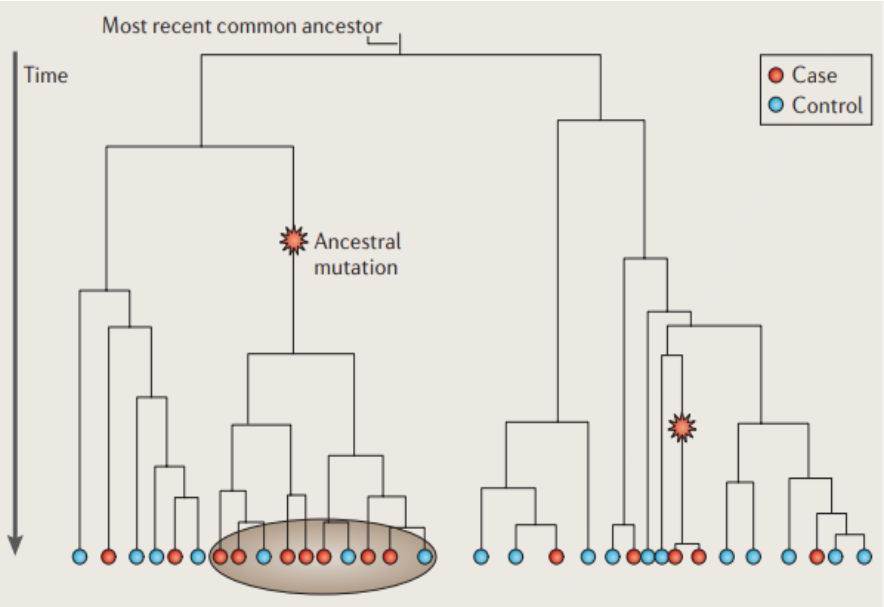
Cardon and John,2001

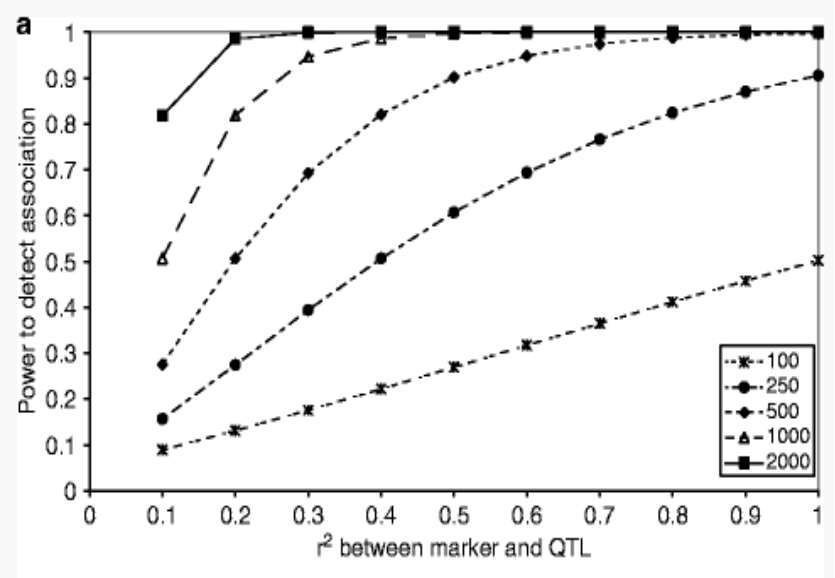
General procedure for Association Mapping



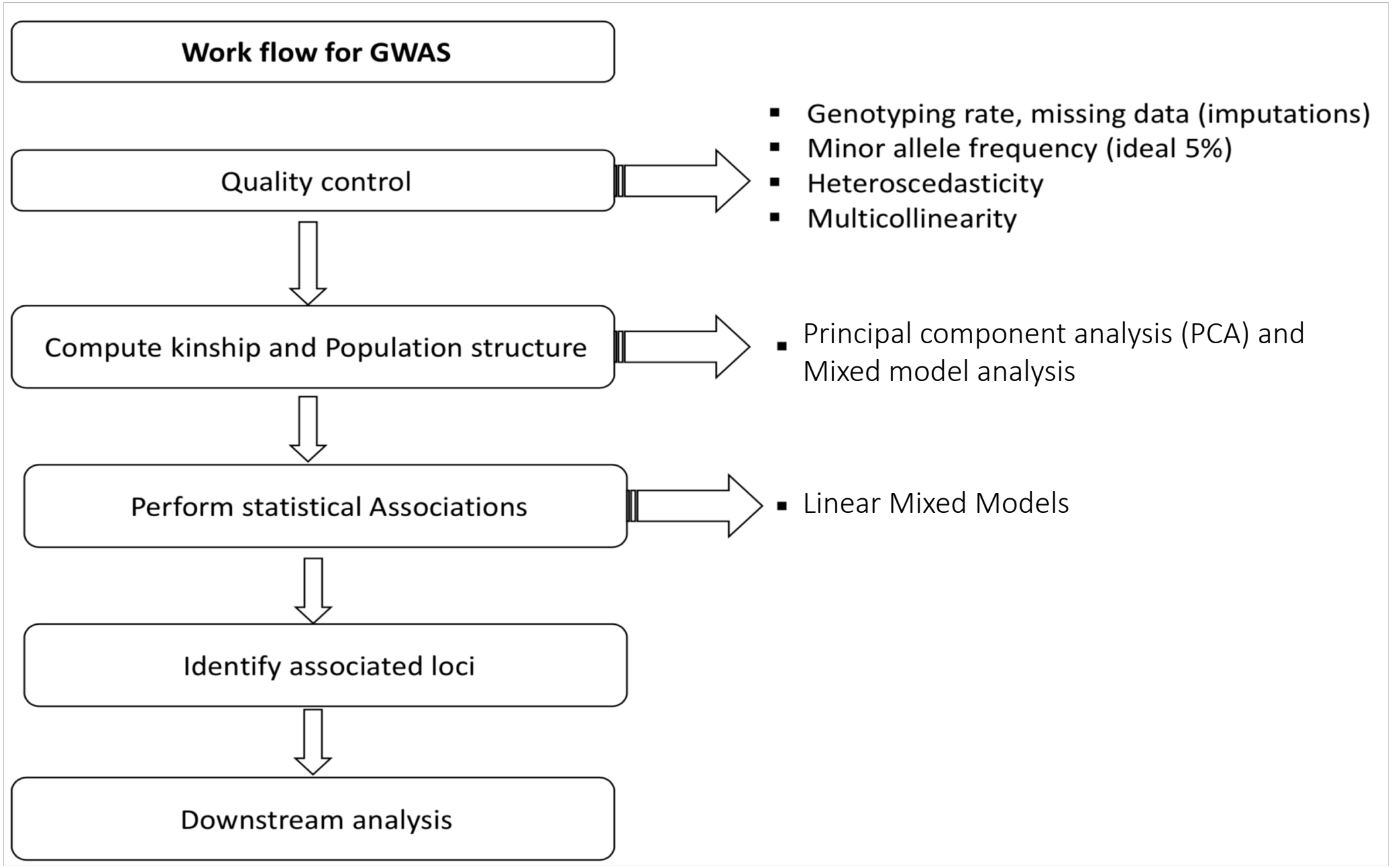
Rational for Association mapping

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



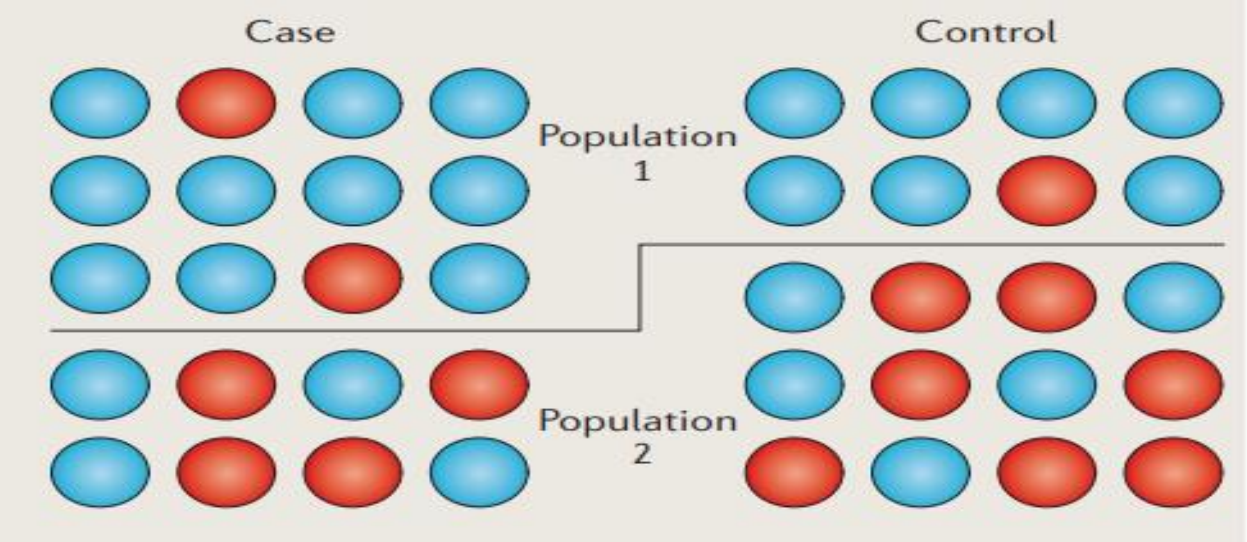


Balding, 2006 https://www.nature.com/articles/nrg1916.pdf


Work flow for GWAS
Quality control
Genotyping rate, missing data (imputations)
Minor allele frequency (ideal 5%)
Heteroscedasticity
Multicollinearity
Compute kinship and Population structure
Principal component analysis (PCA) and Mixed model analysis
Perform statistical Associations
Linear 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

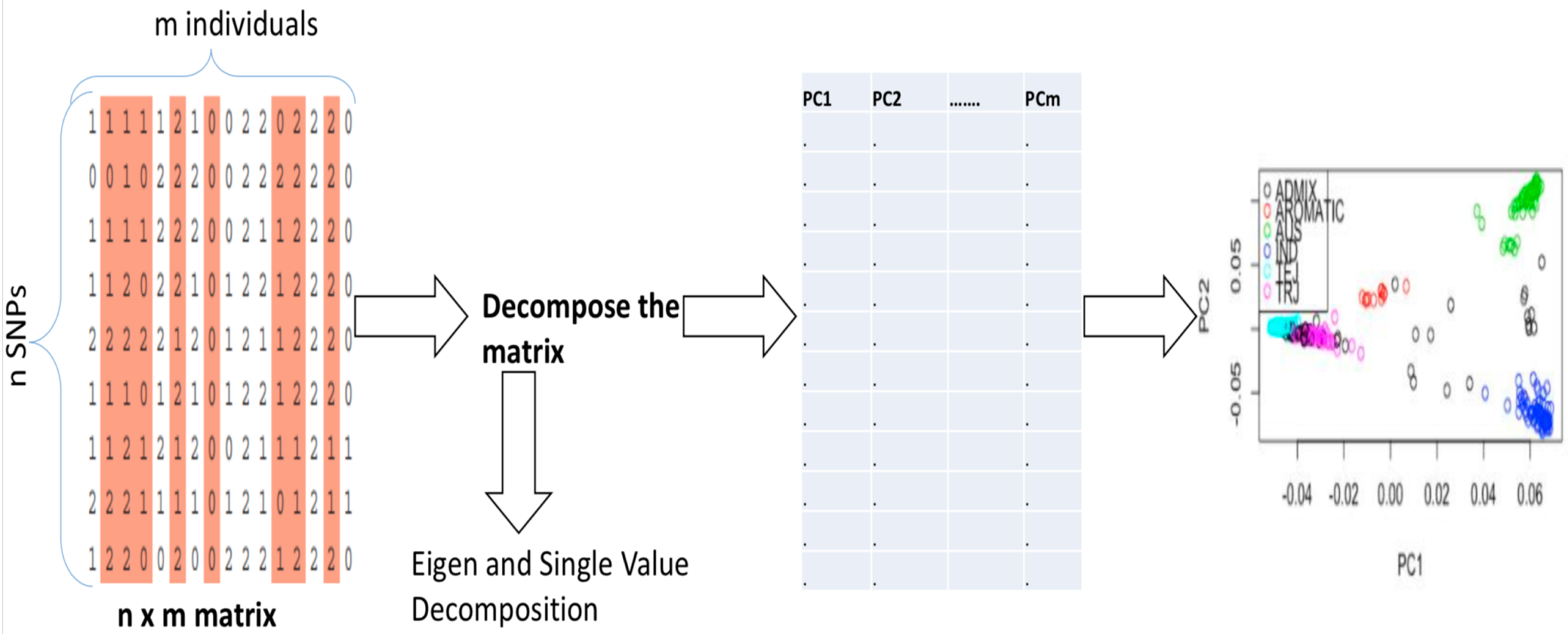
- **Principle component analysis: Provides fast and effective way to diagnose the population structure**

- **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



Algorithm for PCA: Eigen and Single Value Decomposition

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

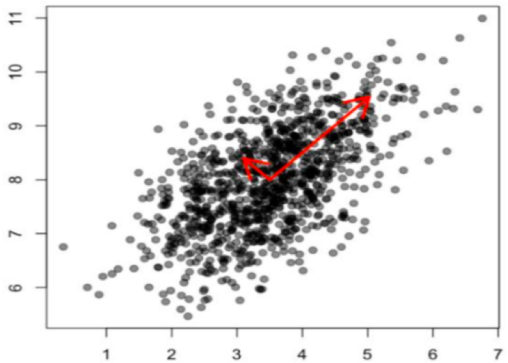
Step 2: Compute the Eigen decomposition of covariance matrix **($G=UDU^T$)**

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

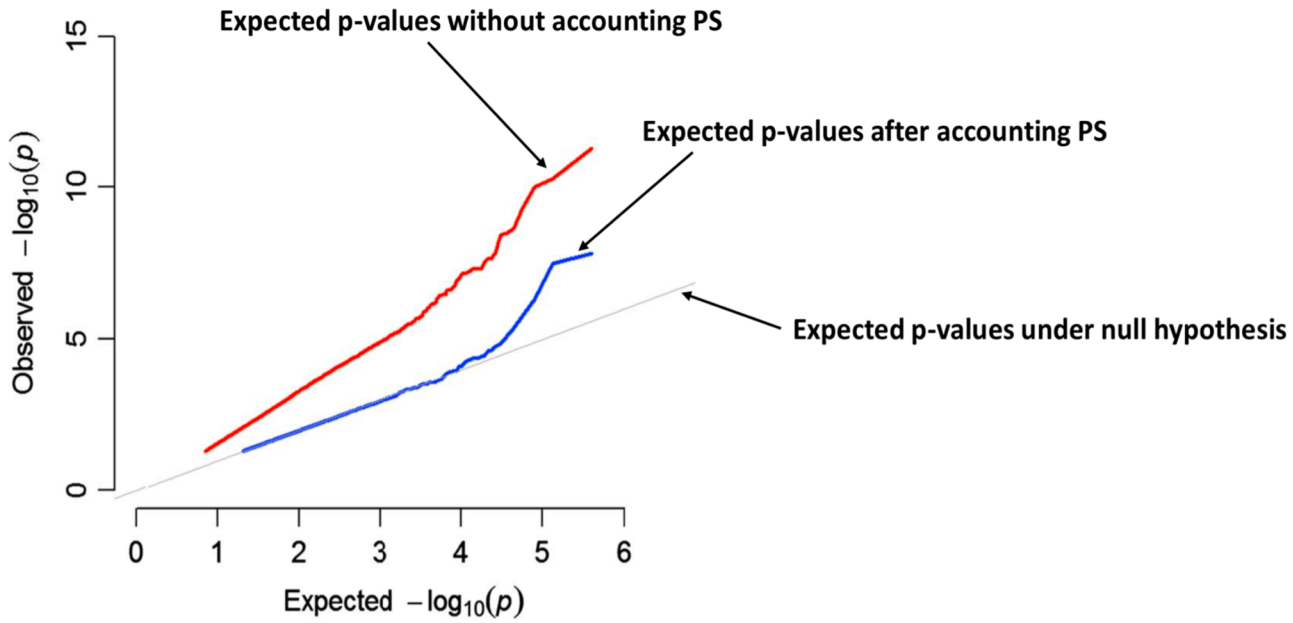


- Singular-decomposition picks out *directions in the data along which the variance is maximised*.
- Singular represent the variance of the data along these directions.

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



Q-Q plot of p-values

Mixed Models

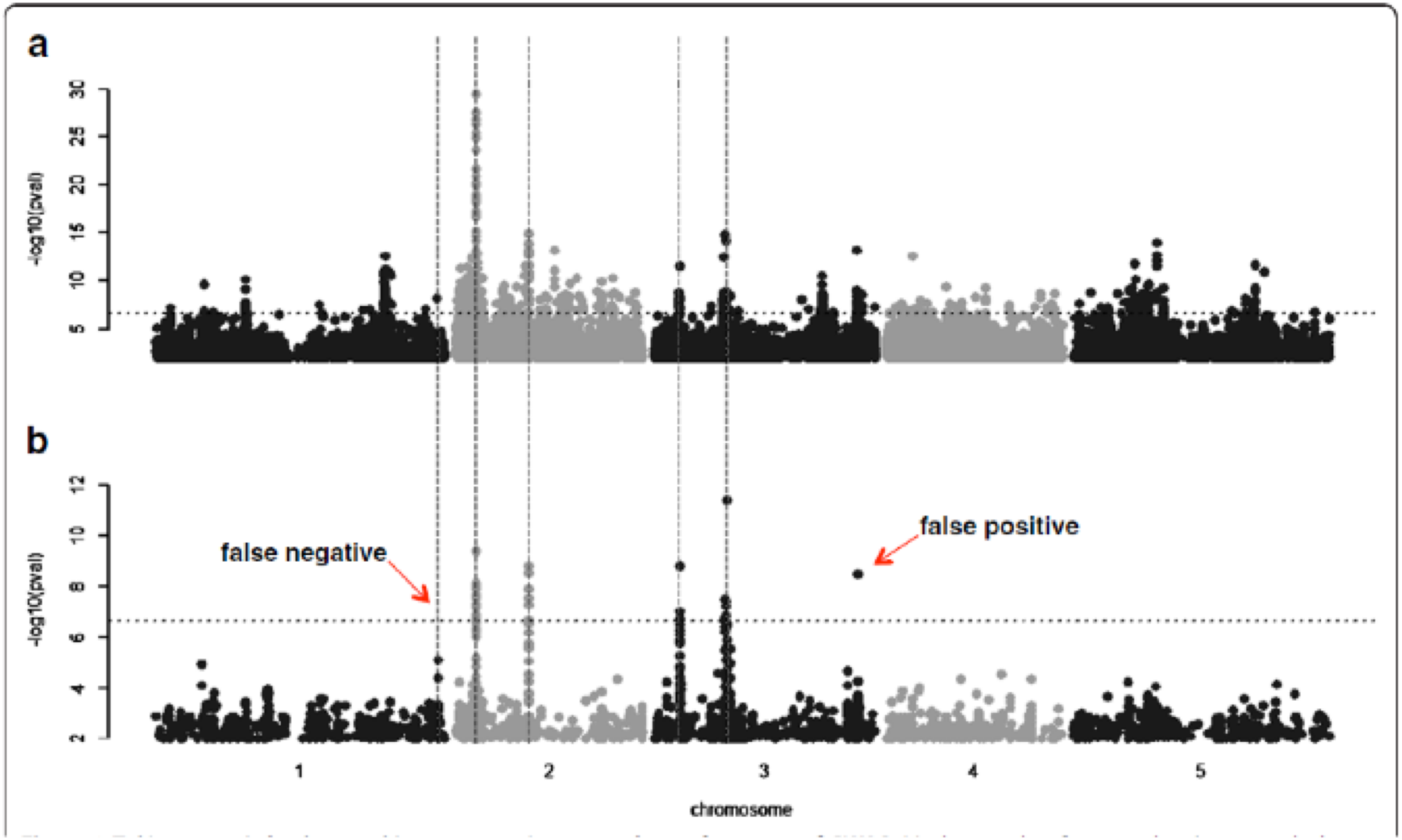
- Use both fixed effects (candidate SNPs and fixed covariates) and random effects (the Genotypic covariance matrix)
  - $y = Wa + u + \varepsilon$

$var(u) = \sigma^2 K$

- 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


a
-log10(pval)
30
25
20
15
10
5
b
-log10(pval)
12
10
8
6
4
2
false negative
false positive
1
2
3
4
5
chromosome

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

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

- 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

$$y_i = \mu + \beta_j \chi_{ij} + \varepsilon_i$$

Phenotype

*j*th marker effect

- 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

Single marker regression

Considering Population Structure

$$y_i = \mu + \beta_j \chi_{ij} + \sum_{k=1}^{K} v_k q_{ik} + \varepsilon_i$$

Principle Components based on marker Data

- PCA only accounts for differences in sub-groups among sub-populations
- Does not account for family relatedness or kinship between individuals

Yu et al. (2006) Nat. Genet. 38: 203

Linear Mixed Models

Accounting for population structure and family relatedness
Single marker based mixed model association

$$y_i = \mu + \beta_j \chi_{ij} + z_g + \varepsilon_i$$

Realized relationship matrix G or A
Captures population structure and polygenic effects

$$g \sim N(0, G\sigma_g^2)$$

Yu et al. (2006) Nat. Genet. 38: 203

- **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

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

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)

Trait Class	Genetic Architecture	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]
		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

- Fits all SNPs simultaneously as random effects

$$y_i = \mu + \sum_{j=1}^{n\_snp} b_j x_{ij} + \varepsilon_i$$

- Distribution assumption for markers varies from model to model

➢ **SNP BLUP**- same variance

➢ **Bayes A**: assumes t-distribution

➢ **BayesB**: only fraction of SNPs has effect on variance

➢ **BayesC**: assumes t-distribution one with large variance for SNP fraction and other with small variance

Schmid and Bennewitz, 2017

GWAS Demonstration in R