IOWA STATE UNIVERSITY

SCHNABLE LAB
Plant Genomics

# A Practical Guide to Genome-wide Association Studies (GWAS)

Zihao Zheng

Schnable Lab

Iowa State University

zhzheng@iastate.edu

@zhzheng92

https://github.com/zhzheng92

AG2PI

Agricultural Genome to Phenome Initiative

# Introduction and objectives

- Part of the AG2PI online workshop series (workshop #4)

- The workshop will be covering
  - Basic theories behind Genome-wide association studies (GWAS)
  - How to conduct a GWAS experiment and what are the resources to do so
  - Hands-on tutorial session using maize and Sorghum (chromosome 9) as examples to demonstrate the analysis of GWAS (https://github.com/zhzheng92/AG2PI_GWAS_workshop_June2021)



Agricultural Genome to Phenome Initiative (AG2PI) is funded by USDA-NIFA award 2020-70412-32615

Join our mailing list to get informed for latest news and upcoming events

Join AG2PI Mailing List →

Get involved with AG2PI to build a stronger G2P community

Get Involved with AG2PI →

AG2PI Twitter    AG2PI YouTube

Copyright © 2020-2021 | Agricultural Genome to Phenome Initiative (AG2PI)

# Genome-wide association studies (GWAS)

- GWAS: an observational study to dissect genetic architecture of complex traits by testing the association of a genome-wide set of genetic variants with variation in phenotype across an assembled population.

- Milestones :
  - Developed in context of human disease genetics in the mid 1990s
  - First GWAS publication in 2002 (Ozaki *et al.*, 2002 )
  - First GWAS publication in plants in 2005 (Aranzana *et al.,* 2005 )
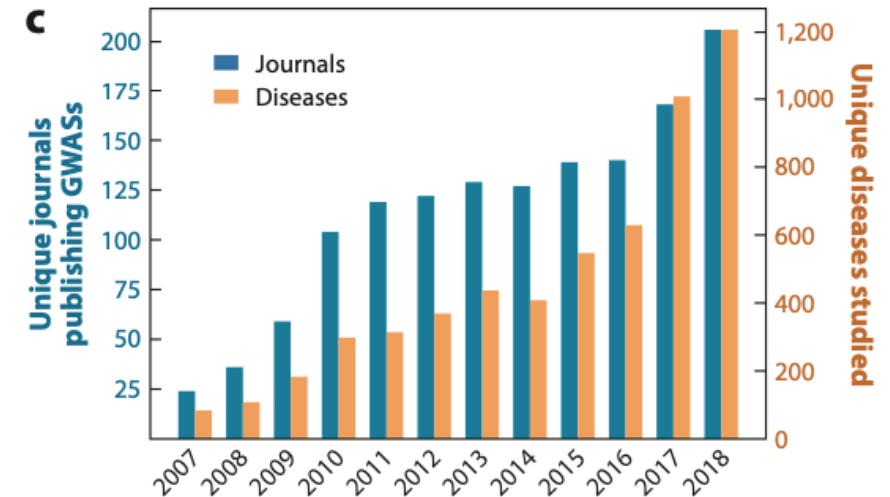


Mills and Tropf (2020)
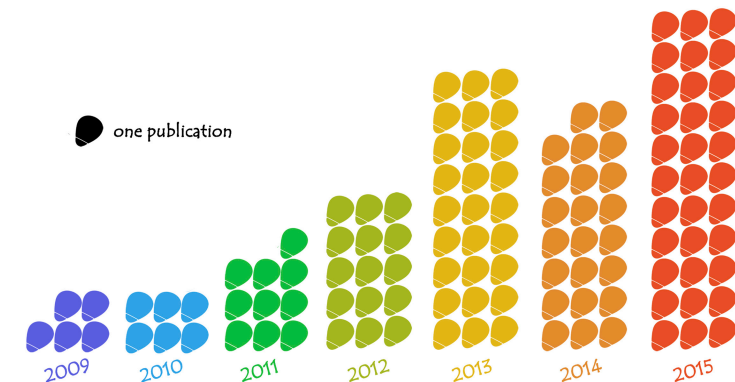


Xiao *et al.* (2017)

# Genome-wide association studies (GWAS)

- GWAS aims to find those genetic markers at which variation in genotype is significantly associated with variation in phenotype.
- Key factors that drive the development of GWAS
  - Germplasm (population)
  - Genetic markers
  - Statistical models
  - Phenotype

germplasm

Variation in genotype → association → Variation in phenotype

model
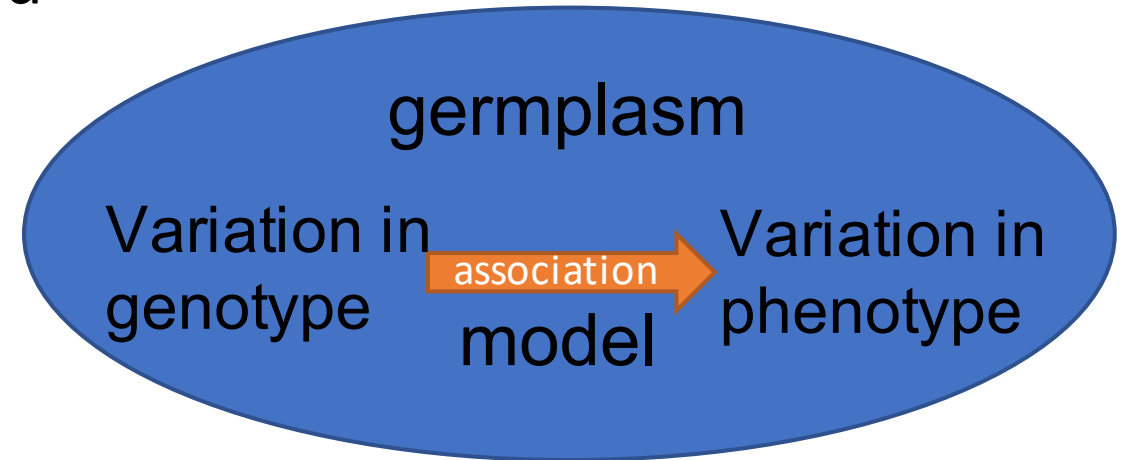
# Linkage analysis

- Linkage analysis for quantitative trait locus (QTL) mapping
  - a statistical method for discovering the locations of loci underlying a trait by testing for **co-segregation** with genetic polymorphisms of known positions in the genome.
  - completely controlled genetic background ($F_2$, Recombinant inbred lines, RILs, multi-parental population etc.)
  - Statistical models: single-marker analysis, interval mapping , composite interval mapping (CIM), multiple-locus CIM
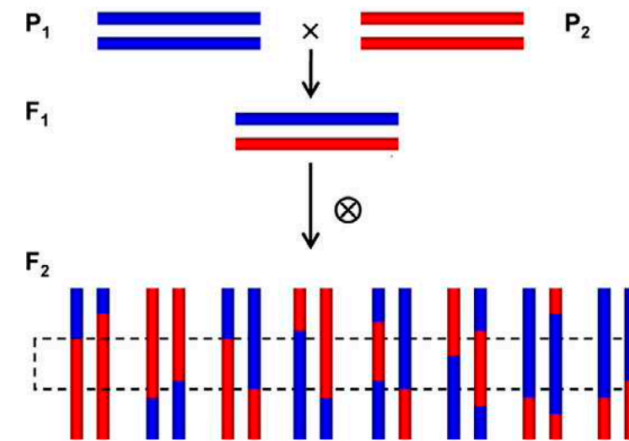
Zhu *et al.* (2008)
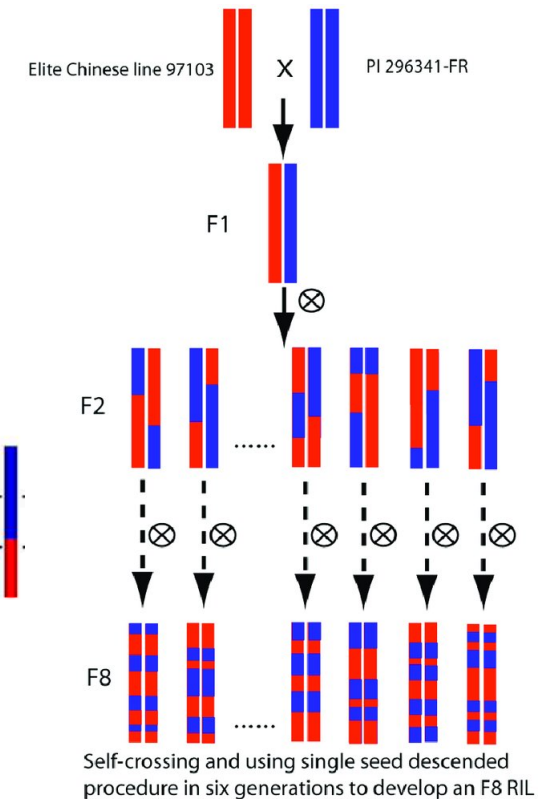
Ren *et al.* (2012)

# Association mapping

- Association mapping
  - A statistical method identifies quantitative trait loci (QTL) by examining the marker-trait associations that can be attributed to the strength of **linkage disequilibrium (LD)** between markers and functional polymorphisms across a set of diverse germplasm.

- Depending on the marker coverage
  - Candidate gene association studies
  - Genome-wide association studies

Natural population



many generations
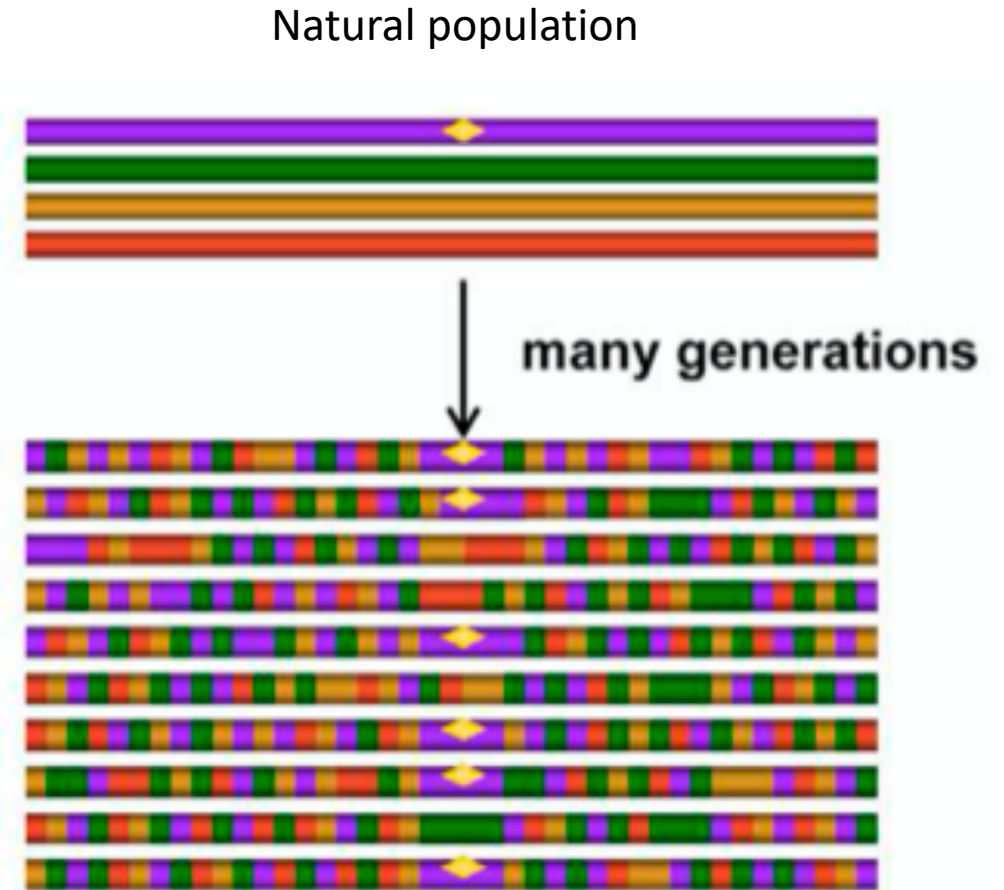
Zhu *et al.* (2008)

# Association mapping

- Linkage disequilibrium (LD)
  - The nonrandom association between two or more loci in a population.
  - In general, the strength of the correlation between two markers is a function of the distance between them: the closer two markers are, the stronger the LD.
  - There's no single best statistic that quantifies the extent of LD



Ümit Seren lecture slides on GWAS



Myles *et al.* 2009

# Association mapping

- Candidate gene association mapping
  - Hypothesis-driven (prior knowledge about the location and the biochemical/regulatory pathway); miss unknown loci associated with the trait
  - Trait-specific (candidate genes)
  - Lower cost (smaller population size and number of SNPs)

- Genome-wide association mapping
  - A comprehensive approach to systematically search the genome for causal genetic variation; no prior information about candidate genes required
  - Higher cost
    - Community efforts to assemble the diversity panel
    - Advance of NGS made genotyping for large number of SNPs possible

# Linkage mapping and association mapping are complementary

- Linkage mapping
- Pros
  - Relatively small population size
  - Genetic marker with lower density
  - Detected QTL can go through fine-mapping
- Cons
  - Low allelic diversity
  - Experimental crosses take a long time
  - Limited recombination events, lower mapping resolution

- Association mapping
- Pros
  - High allelic diversity
  - No crosses needed, save time
  - Take advantage of historical recombination, higher mapping resolution
- Cons
  - Observational experiment, substantial variation in many phenotypes
  - Large population size to detect QTL
  - High-density genetic marker
  - More validation needed for candidate genes

# The evolution of GWAS

Variation in genotype → association model → Variation in phenotype

- Naïve model: t-test

- False positives accumulate across large number of markers

- **Genetic background of population**
  - **Doesn't address relatedness among individuals within the diversity panel**
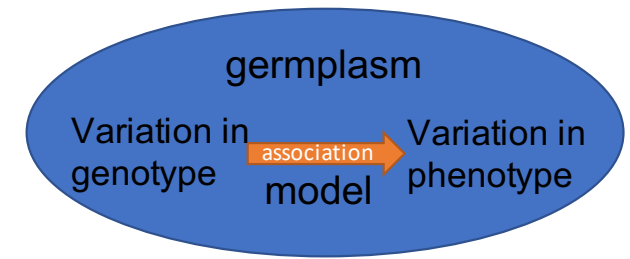
Variation in genetic markers

Variation in phenotype

| | | | | | | |
|---|---|---|---|---|---|---|
| Individual 1 | A | C | G | A | G | 1.3 m |
| Individual 2 | A | C | G | A | T | 1.4 m |
| Individual 3 | A | T | A | A | G | 1.5 m |
| Individual 4 | C | T | A | G | T | 1.8 m |
| Individual 5 | A | C | G | G | T | 2.0 m |
| Individual 6 | A | T | G | G | G | 2.1 m |
| | A/C | T/C | G/A | A/G | G/T | |

Roberto Fritsche-Neto lecture slides on GWAS
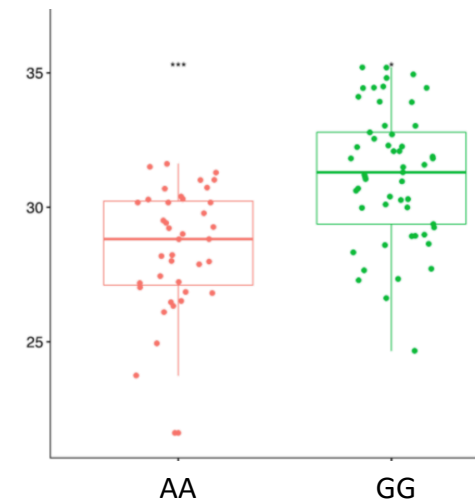
AA          GG

10

# Relatedness among individuals


germplasm
Variation in genotype → Variation in phenotype
association model

- Population structure

- Systematic difference in allele frequencies between subpopulations (caused by geographical and climate distance, familial relationship etc.)

- Can lead to spurious association (false positives/Type I error)



caudatum (101)
caudatum–guinea (46)
bicolor–caudatum (30)
caudatum–durra (35)
durra (81)
bicolor–durra (23)
guinea (64)
bicolor–guinea (4)
durra–guinea (5)
guinea–kafir (3)
kafir (34)
caudatum–kafir (24)
bicolor–kafir (2)
durra–kafir (3)
bicolor (28)
wild (6)
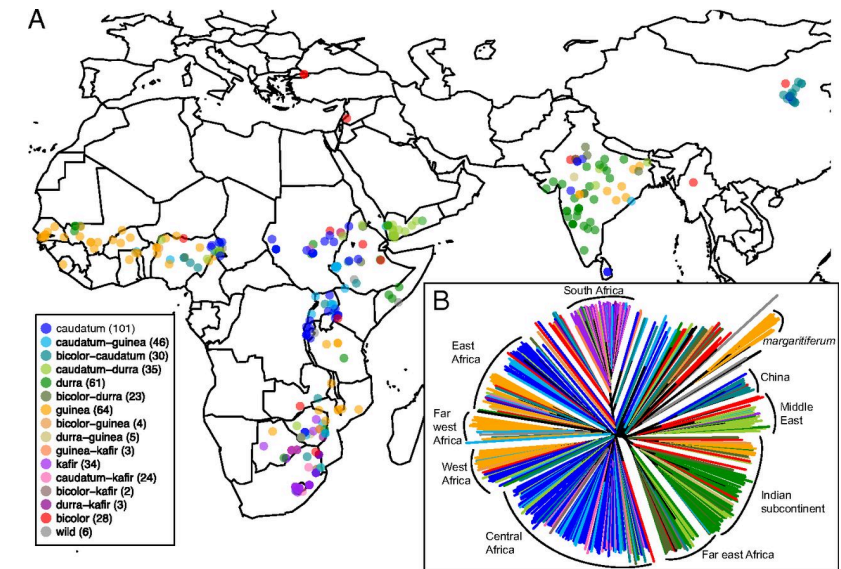
Morris *et al*. 2013

| | West Africa | | | | | | | | | | East Africa | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Plant height (PHT) | 10 | 10 | 12 | 11 | 13 | 9 | 11 | 10 | 13 | 12 | 4 | 6 | 5 | 7 | 6 | 6 | 4 | 5 | 9 | 5 |
| Disease resistant (DS) | S | S | S | S | R | S | S | S | S | R | R | S | R | R | R | R | R | R | S | R |
| SNP1 (A/G) PHT | A | A | A | G | A | A | A | A | G | A | G | G | G | G | G | G | A | G | A | G |
| SNP2 (C/G) DS | C | C | C | C | C | G | C | C | G | C | G | G | G | G | C | G | G | C | G | G |

# Relatedness among individuals
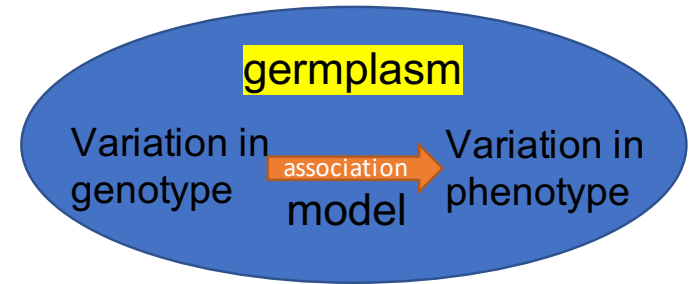


- Unequal familial relationship (kinship)
- Familial relatedness from recent coancestry among individuals usually exists in a collection of diversity panel, which can lead to spurious association.
  - Estimate relationships among individuals using molecular markers due to incomplete pedigree information (kinship matrix)
  - Two individuals sharing a lot of genotypes at SNPs are likely belong to the same family
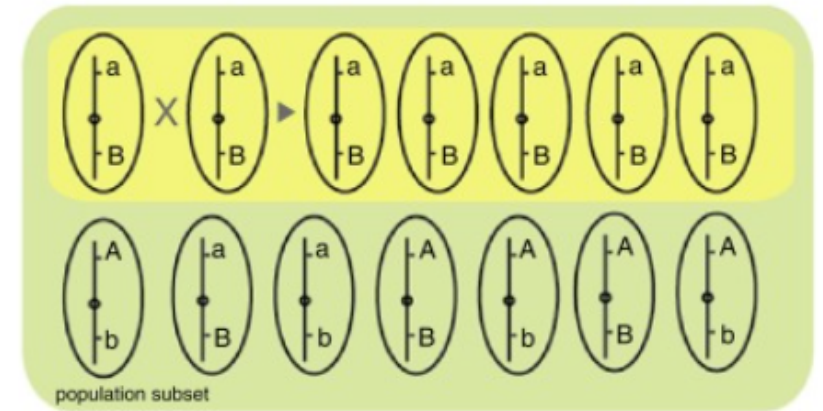


Lipka *et al.* (2018)

# Mixed linear model framework



germplasm

Variation in genotype → association model → Variation in phenotype
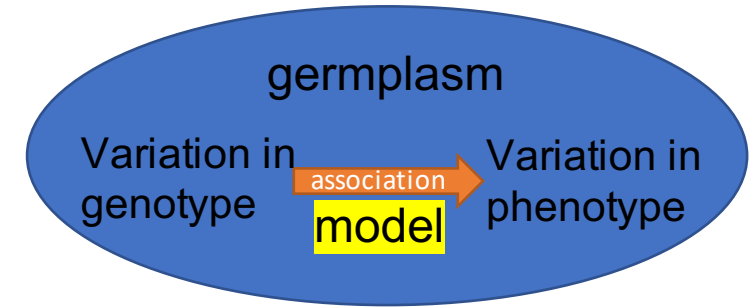
- A unified mixed-model framework for association mapping that accounts for multiple levels of relatedness
  - Control for population structure estimated by the SNP data (STRUCTURE or principle component analysis, PCA)
  - Control for unequal familial relationship using kinship matrix
- A tutorial on mixed linear model



Source of variation

$$Y = \mu + SNP + Q(or\ PCs) + Kinship + e$$

population mean (observation) | (Fixed effect) | population structure (Fixed effect) | Unequal relatedness (Random effect) | (Error)

General linear model (GLM)

Mixed linear model (MLM)

# Mixed linear model framework

germplasm

Variation in genotype → association model → Variation in phenotype
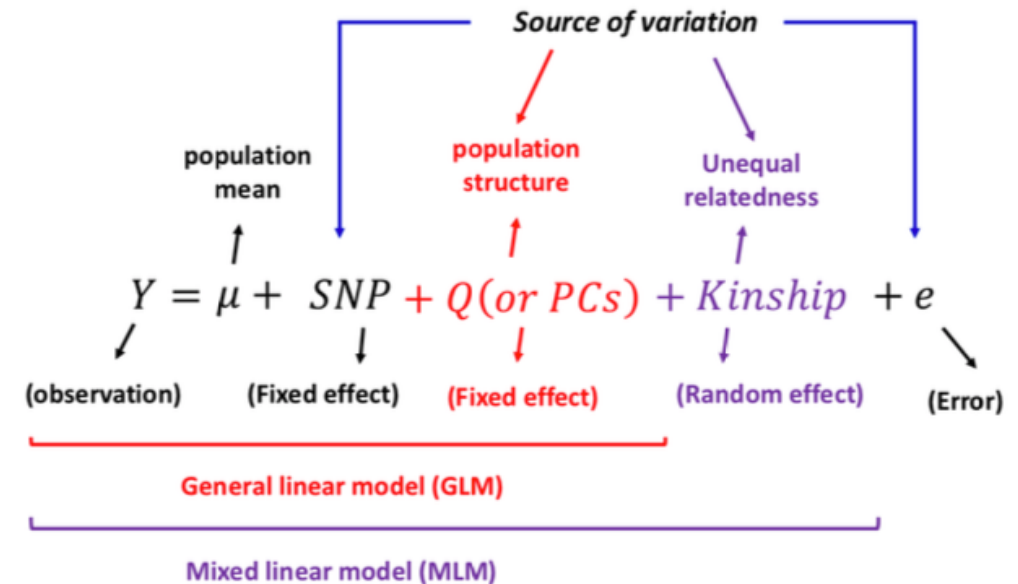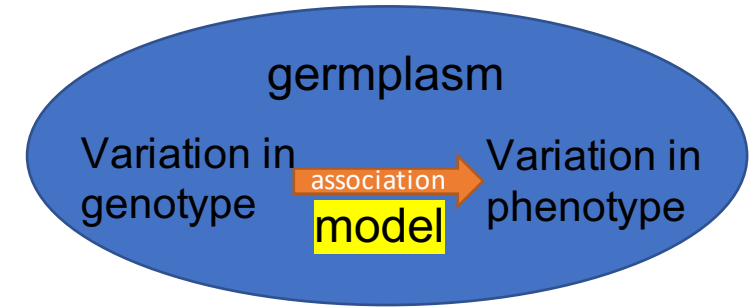
- By controlling population structure and kinship at the same time, mixed linear model is able to reduce Type I and Type II errors in GWAS.



## Type I and Type II Error

| Null hypothesis is... | True | False |
|---|---|---|
| Rejected | **Type I error** False positive Probability = α | **Correct decision** True positive Probability = 1−β |
| Not rejected | **Correct decision** True negative Probability = 1−α | **Type II error** False negative Probability = β |
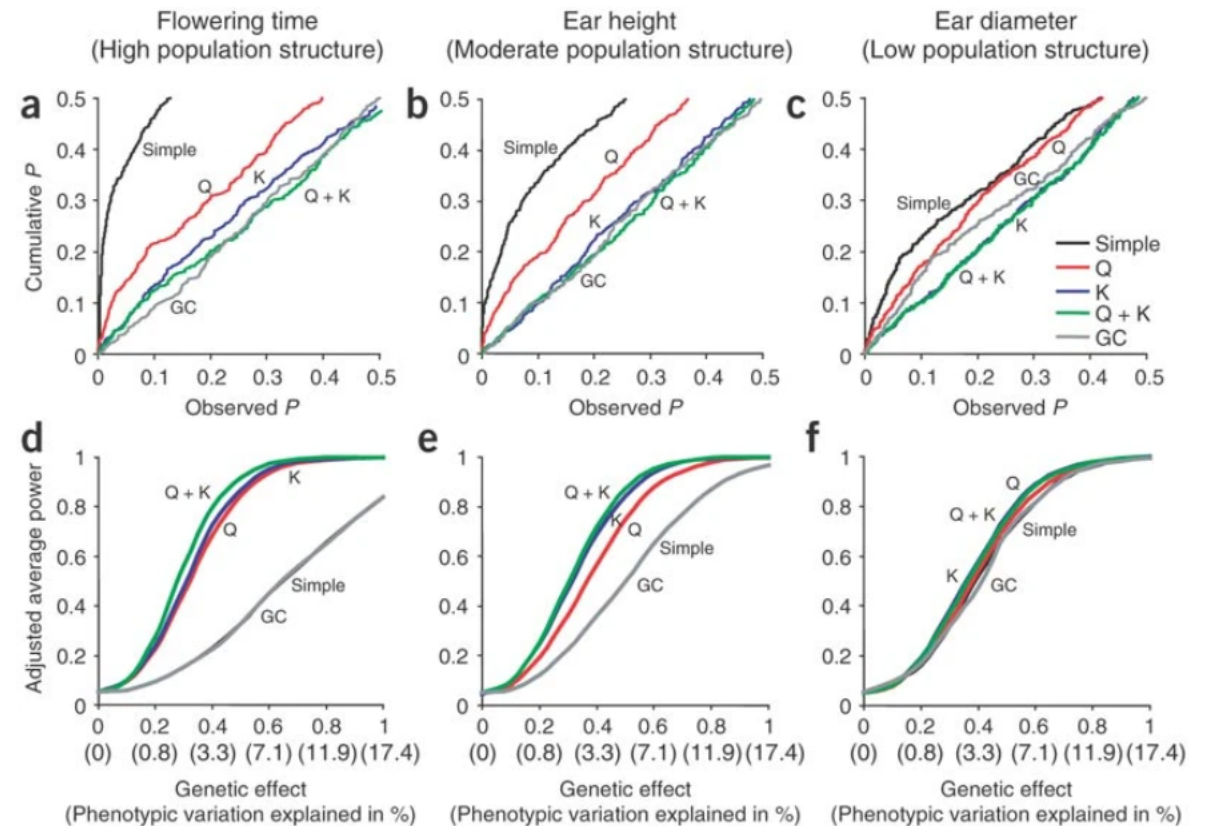
Yu *et al.* (2006)

# Improvement of GWAS models
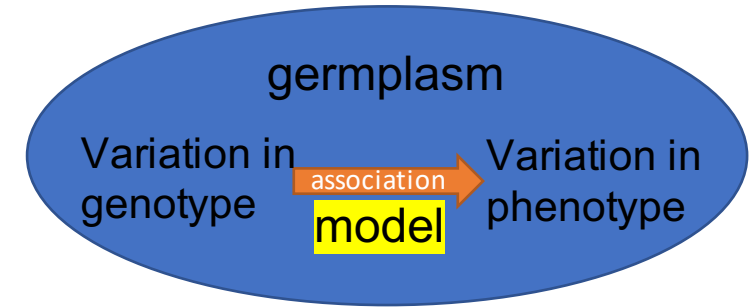


- Improvement have been made based on the mixed linear model framework
  - Increase computational efficiency
  - Increase statistical power

- Alternative framework for GWAS
  - Bayesian methods: Bayesian RR-BLUP, BayesA, BayesB, BayesCπ, etc.
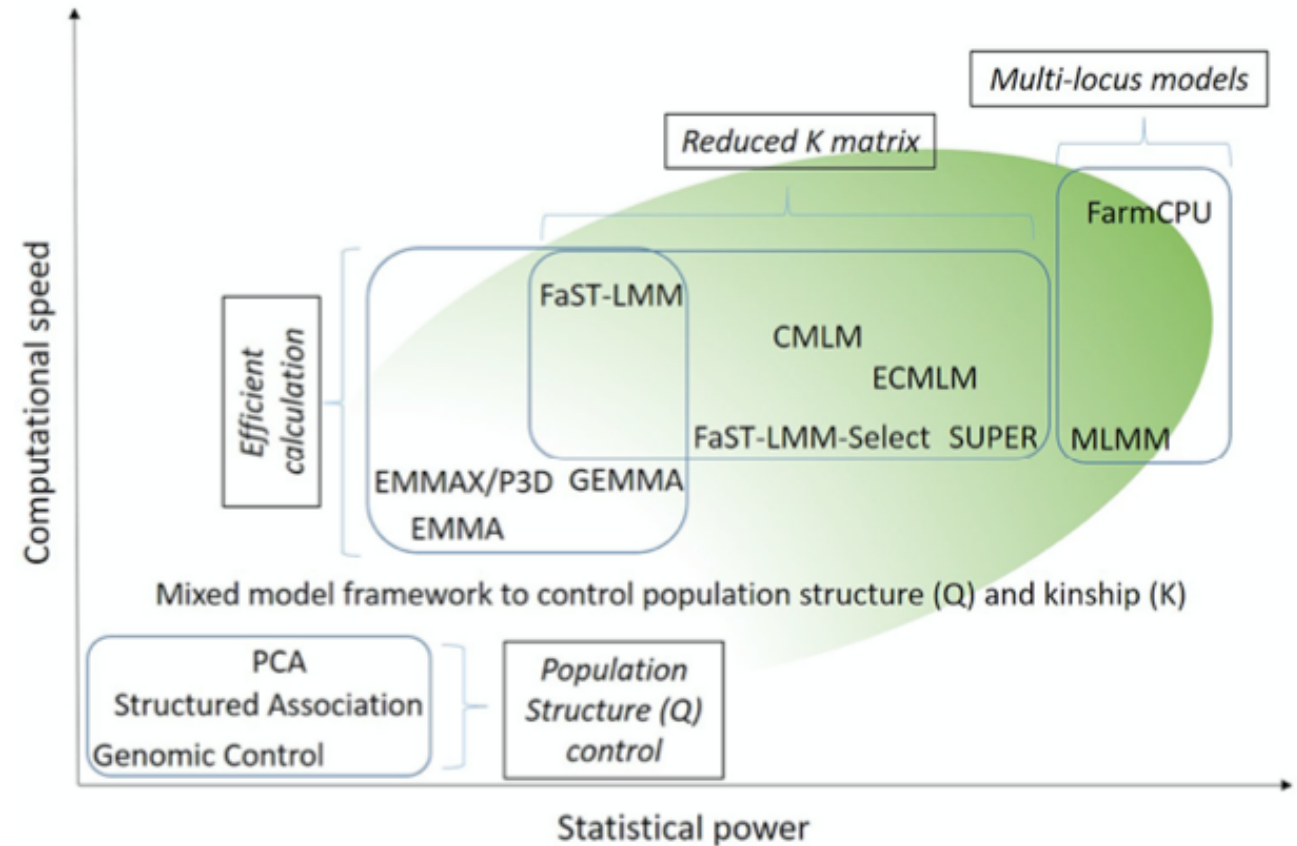


Cortes *et al.* (2020)

# Expanding the list of phenotypes
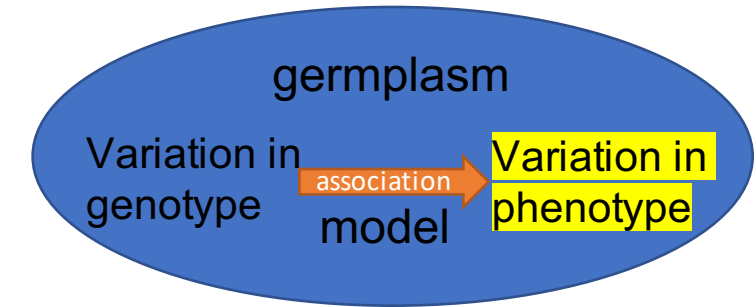


- Phenotype: the set of observable characteristics of an individual resulting from the interaction of its genotype with the environment
  - Morphological: physical form and structure
  - Biochemical: e.g. metabolome
  - Molecular: e.g. transcriptome, proteome

- The expression the phenotype across different environments and developmental stages

Chen *et al.* (2013)

16
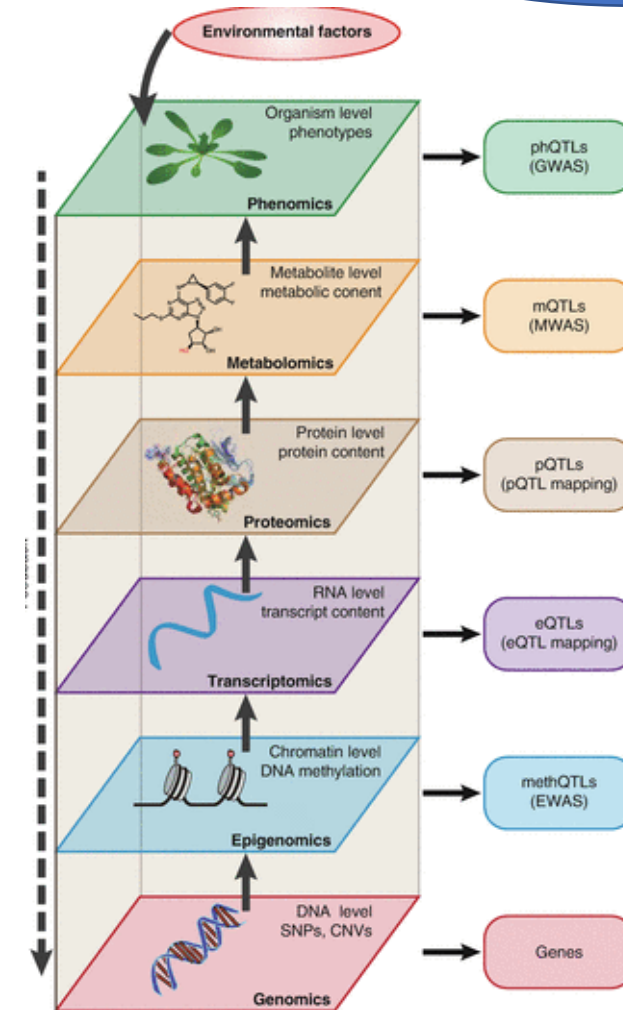
# Summary

- Key factors that drive the development of GWAS
  - Germplasm (population)
  - Genetic markers
  - Statistical models
  - Phenotype
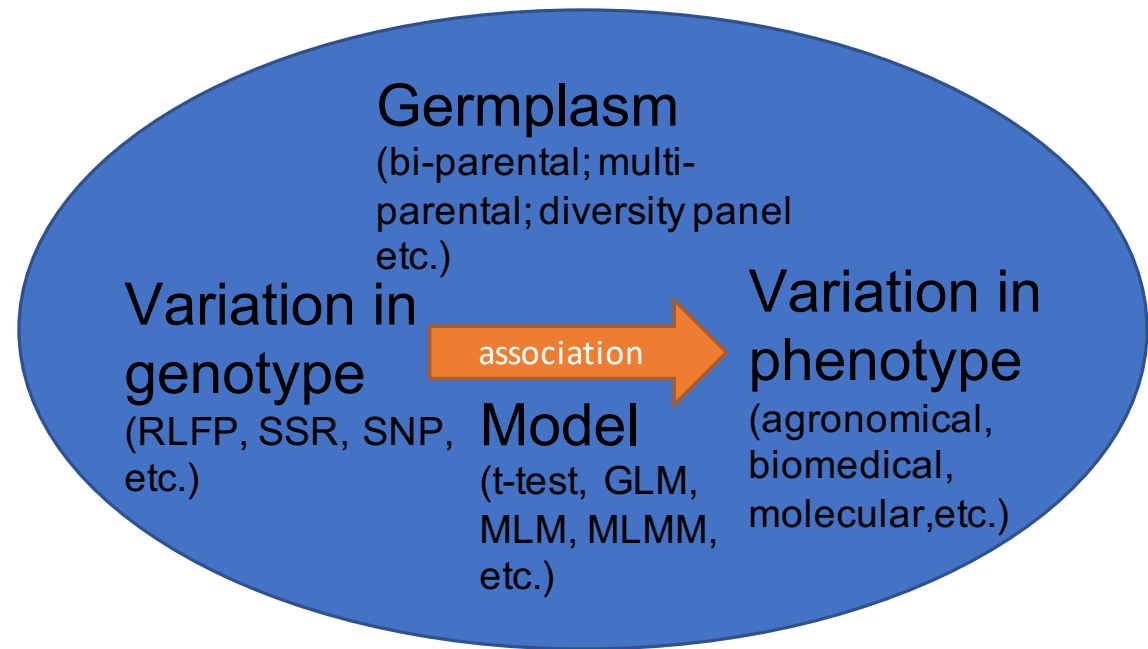
# Some challenges of GWAS

- How to deal with loci with low minor allele frequency (low MAF -> small p-val, more false positives, low confidence on signals with low MAF)

- How to establish an appropriate significance threshold
  - Bonferroni correction: overly stringent
  - False discovery rate (FDR): assumption of test statistics are independent
  - Permutation test

- Synthetic association
  - the causative genes are sometimes located away from the GWAS peaks.



Liu *et al.* (2019)

# Future of GWAS


germplasm
Variation in genotype → association model → Variation in phenotype
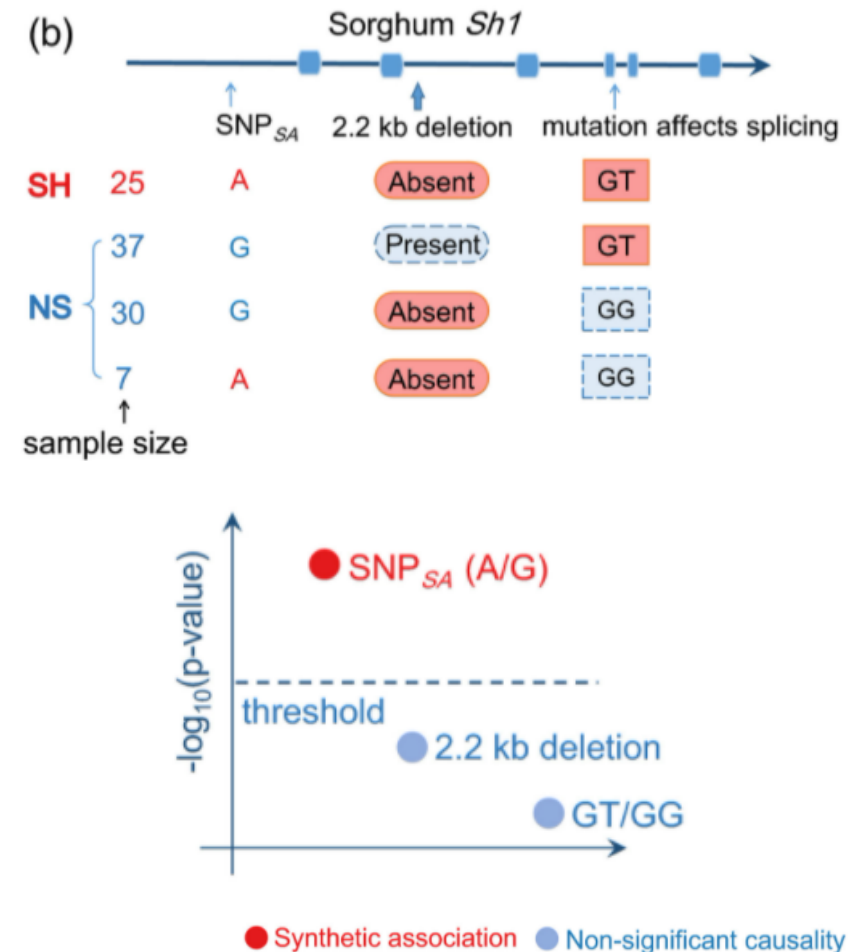
- Using omics data to associate with variation in phenotypes



Present
Future

Genome   Epigenome   Transcriptome   Proteome   Metabolome   •••   Phenome

Xiao *et al.* (2017)



A   GWAS
AtFLC   AtDOG1
$-\log_{10}(FDR)$
Chromosome

B   TWAS
AtFLC
AtSOC1   AtAGL16   AtSPL15   AtAGL24   AtGID1C
$-\log_{10}(FDR)$
Chromosome

Li *et al.* (2021)

# Future of GWAS



germplasm

Variation in genotype → association model → Variation in phenotype
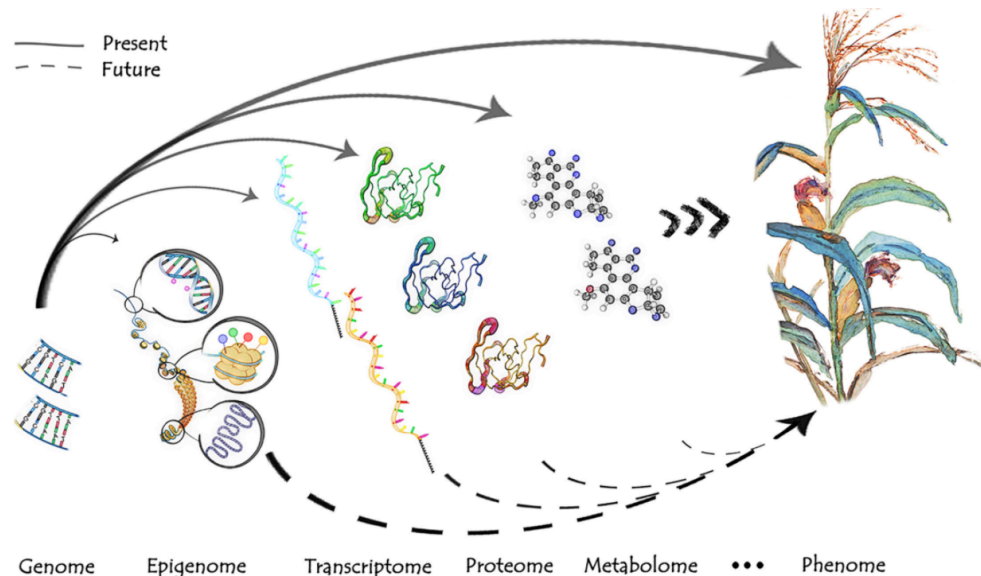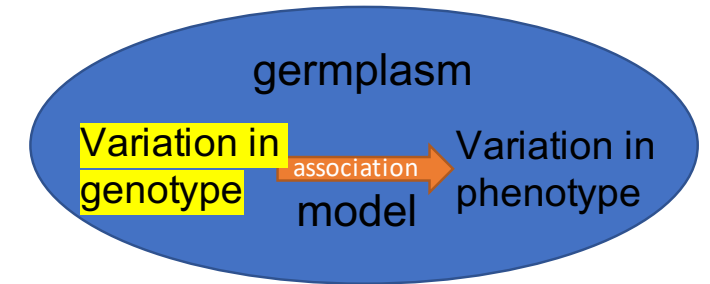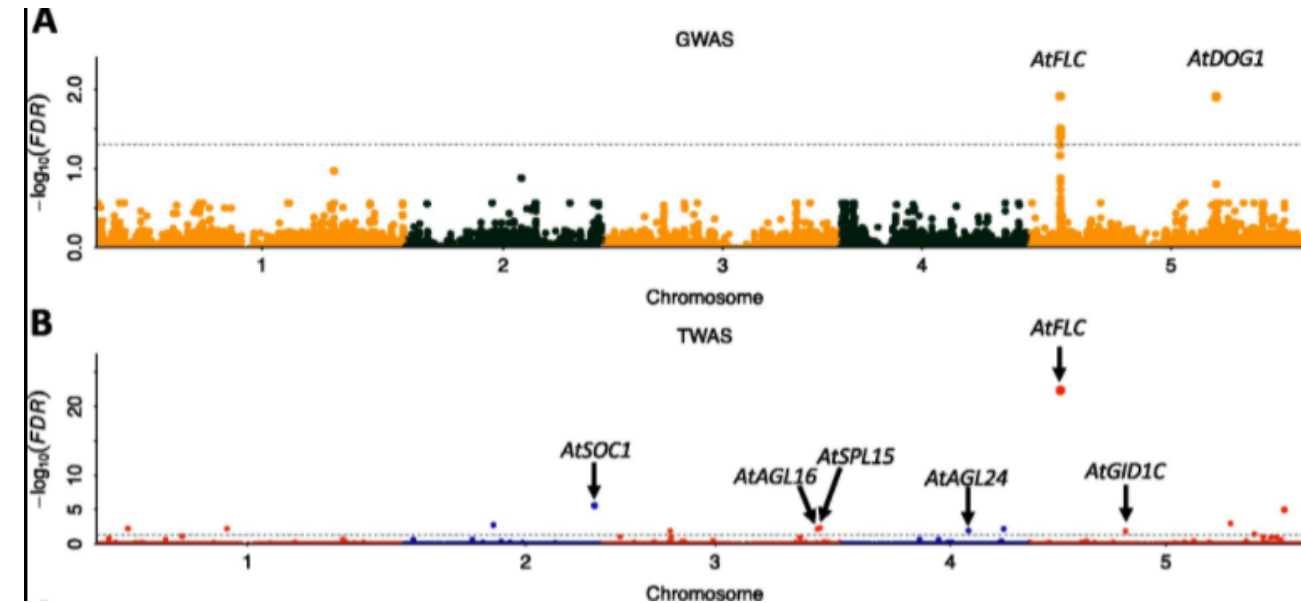
- From single reference genome to pan-genome



Qin *et al.* (2021)

# Future of GWAS

germplasm

Variation in genotype → association model → Variation in phenotype
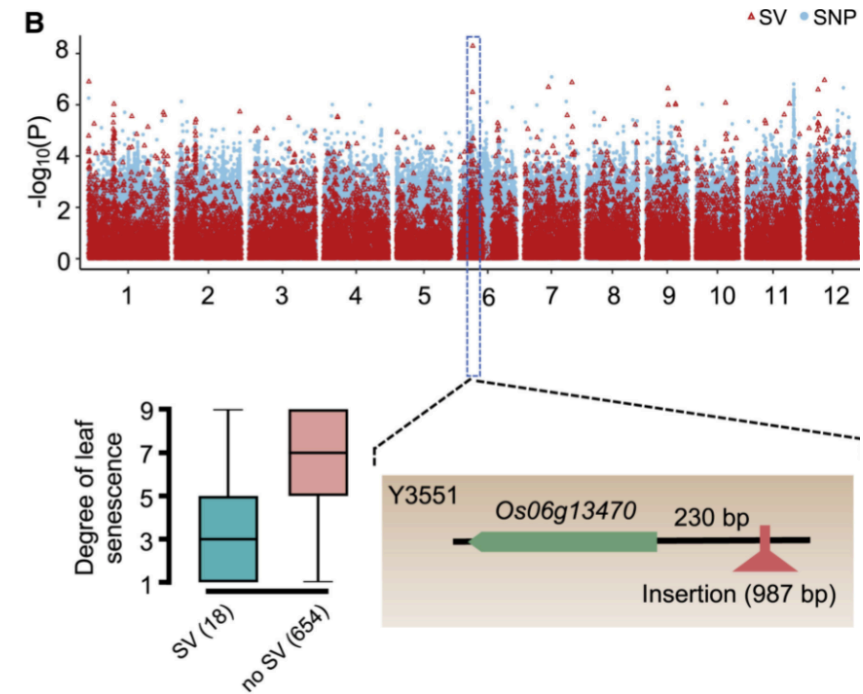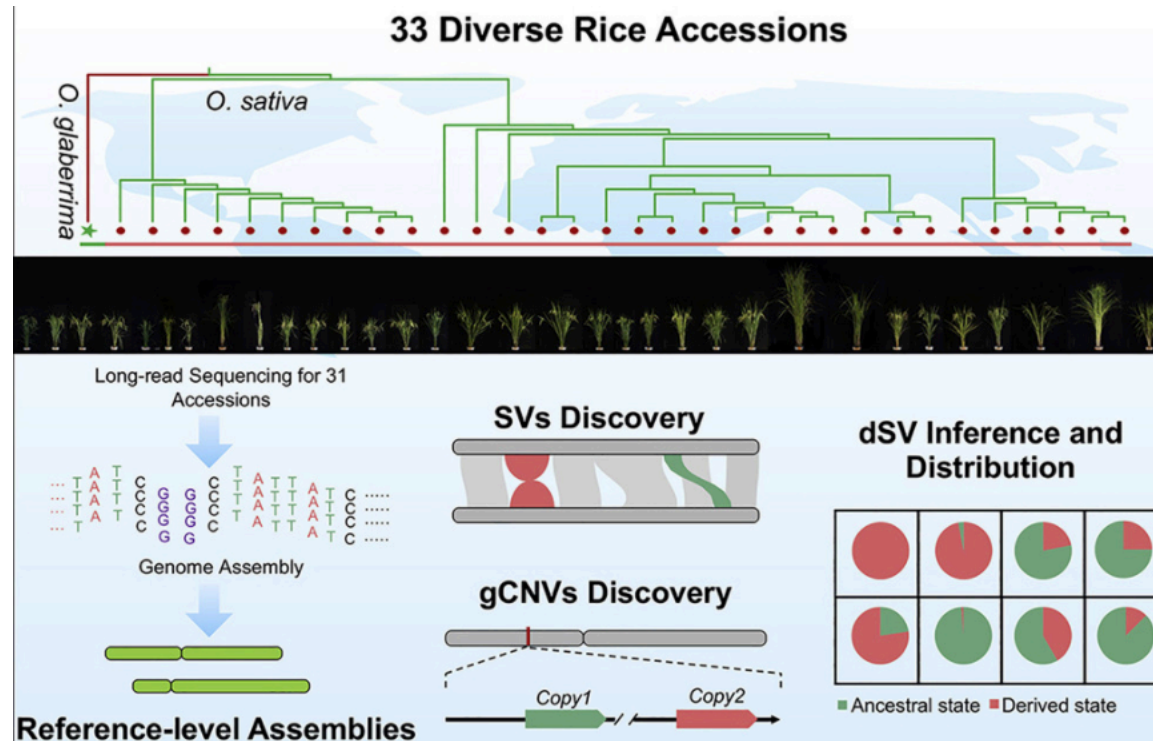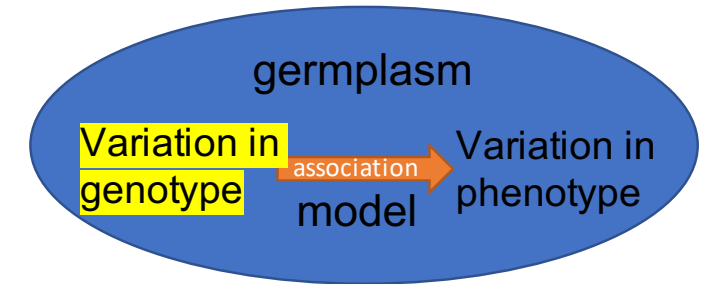
- High-throughput phenotyping/phenomics



Moreita *et al.* (2020)
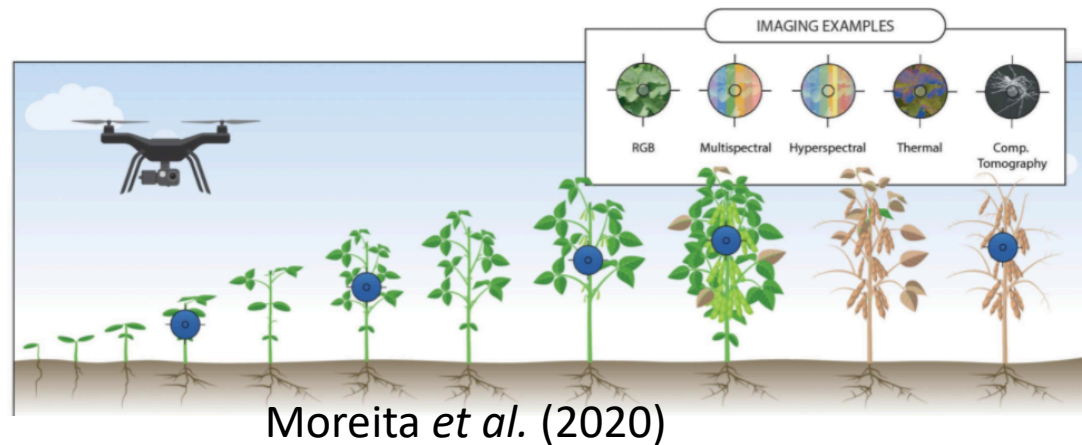
James Schnable Lab at UNL image
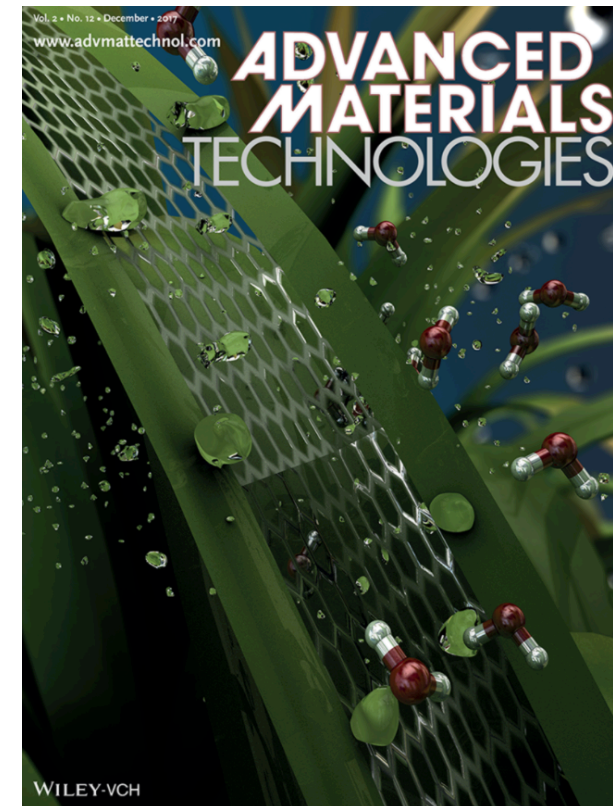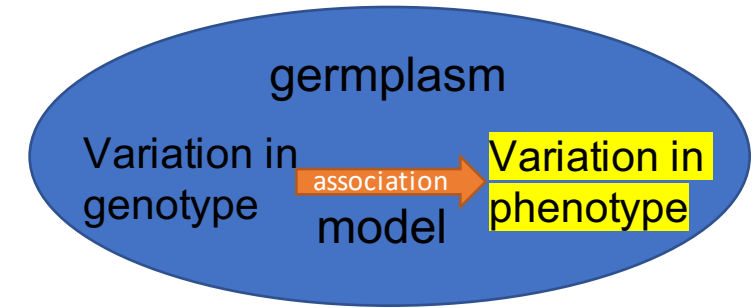


Liang Dong Lab at ISU image

# Species and germplasm

- Carefully consider all genetic aspects and available community resources
    - How was the diversity panel assembled
    - Genetic diversity, genome-wide LD, population structure
    - Genetic marker availability
    - Determine the appropriate sample size and number of markers



**Maize Type**
Flint
Iodent
Mixed
Non Stiff Stalk
Popcorn
Stiff Stalk
Sweet Corn
Tropical
Unknown

Renk *et al.* (2021)

N = 501



g1 - Highland South America
g2 - Lowland South America - Chile
g3 - Lowland South America - Brazil
g4 - West of Sierra Madre Occidental of Mexico
g5 - Mesoamerica, Central America and Caribbean
g6 - Central Valleys and East of Sierra Madre Occidental of Mexico
g7 - Lowland South America - Argentina
g8 - Lowland South America - Paraguay
g9 - Intermediate South America Zone
- Zea luxurians

Silva *et al.* (2020)

N = 575

# Species and germplasm

- Ames diversity panel
  - A world-wide collection of maize inbred lines (>3,000) maintained by the USDA North Central Regional Plant Introduction Station, Ames, IA
  - 2,815 inbred lines were genotyped with GBS (681,257) SNPs
  - LD decayed fast
  - Kernel color, flowering time, plant height were phenotyped



LANDRACES
a. Santo Domingo
b. Assiniboine
c. Longfellow flint
d. Pororo
e. Shoe Peg
f. Hickory King
g. Zapalote chico
h. Araguito
i. Cateto
j. Cravo Riogranense
k. Tuxpeno
l. Costeno
m. Pollo
n. Cuban flint
o. Chapalote
p. Palomero de Jalisco

Romay *et al.* (2013)
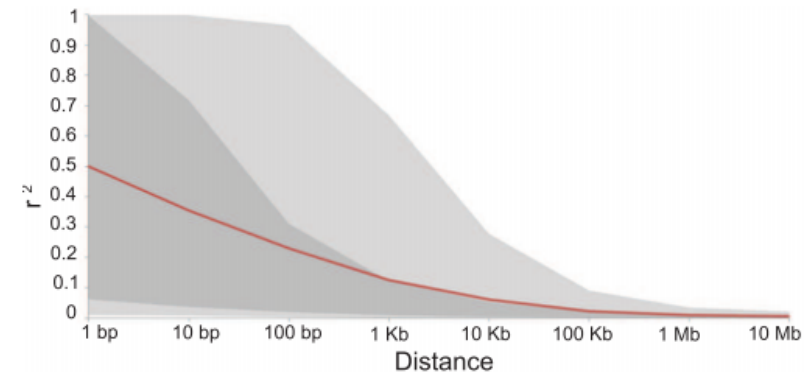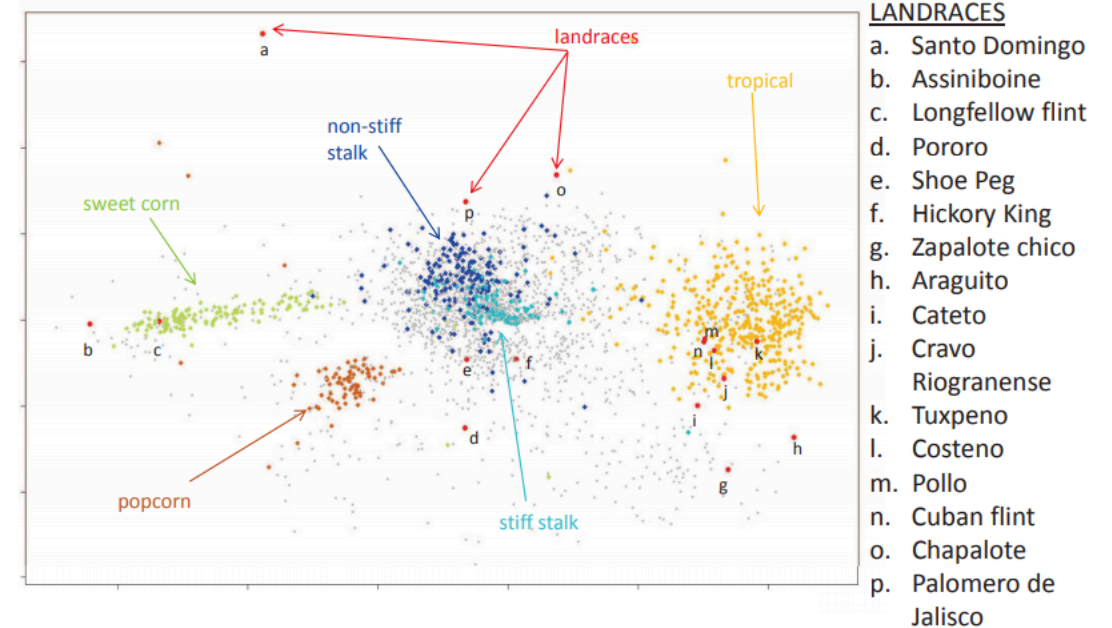
# Species and germplasm

- Maize 282 association panel
  - A collection of 282 inbred lines through the world, biased towards temperate lines
  - Have been used in numerous community mapping studies (57 traits from multiple environments on Panzea.org)
  - Have been genotyped comprehensively
    - WGS (> 83M variant sites)
    - RNA-Seq data from seven tissues



Flint-Garcia *et al.* (2005)



Germinating shoot
Germinating root

Leaf 3 base
Leaf 3 tip

Adult leaf collected during the day
Kernels 350 growing degree days after pollination

Adult leaf collected at night
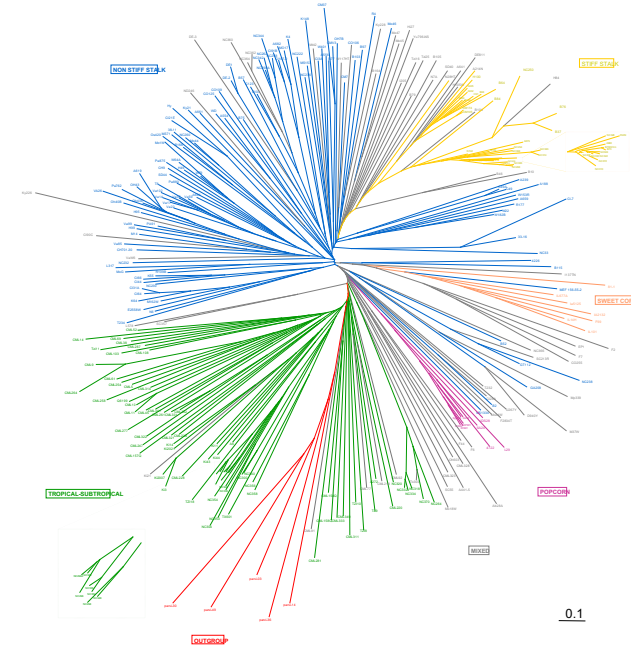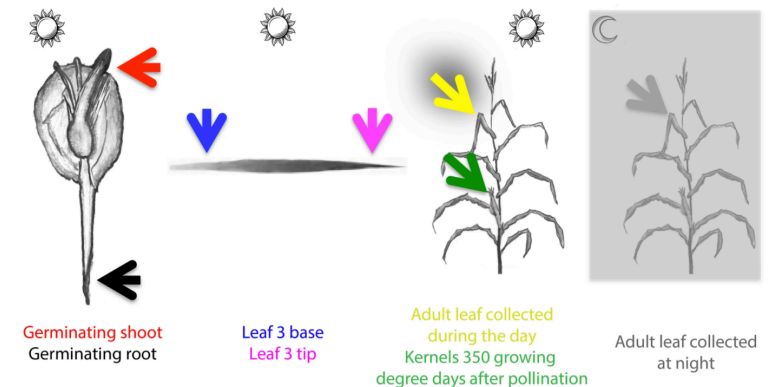
Kremling *et al.* (2018)

# Genetic markers

- SNP discovery enabled by Next Generation Sequencing (NGS)
  - SNP array, GBS, RNA-Seq, WGS etc.
  - Bioinformatic software and pipelines for SNP processing
  - Budget, genome coverage etc.

AG2PI workshop#3 by Dr. Jacob Landis

| Phylogenomics approach | Genomic resources required | Initial bioinformatic investment | Ultimate bioinformatic investment | Initial laboratory cost | Ultimate cost per sample |
|---|---|---|---|---|---|
| Genome skimming | Yes | None | Medium | Low | Medium |
| RAD-Seq | No, but helpful | Medium | High | High | Low |
| RNA-Seq | No, but helpful | Low | High | Low | High |
| Hyb-Seq | Varies[b] | High[b] | Medium | Low[b] | Medium |

Modified from Dodsworth et al., 2019

# Field design and phenotyping

- Randomization and block design (influence of flowering time and correlated traits)

- Biological and technical replications

- Include checks

- Collect environmental data for spatial correction



(a) Field layout    (b) Raw yield data    (c) Fitted values

(d) Residuals' spatial plot    (e) Fitted spatial trend    (f) Genotypic BLUPs
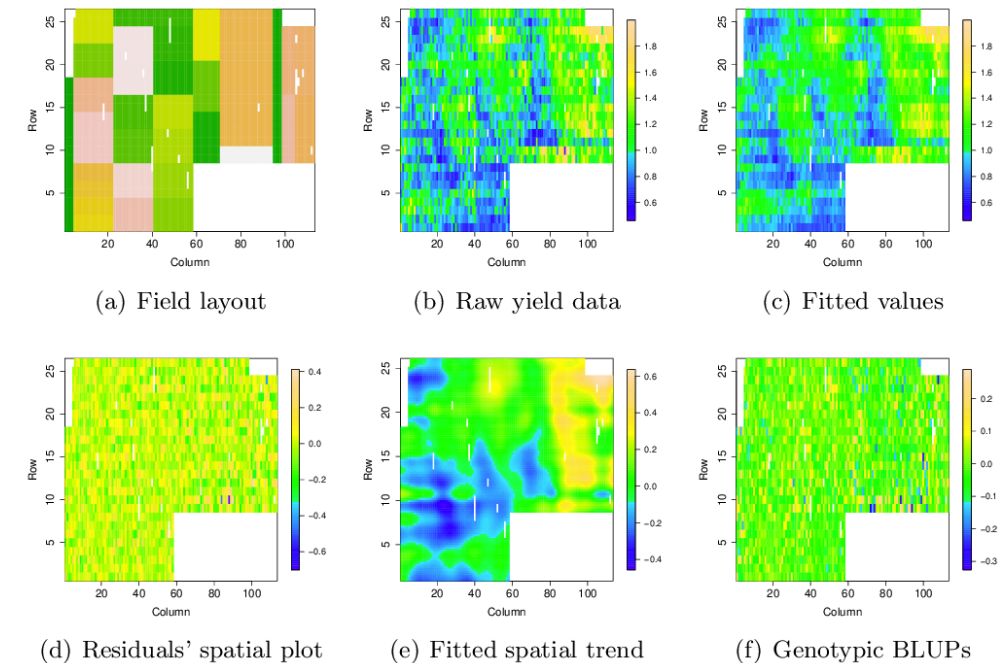
Rodriguez-Alvarez *et al* (2016)

# Field design and phenotyping

- Document how the phenotype was collected clearly
  - The developmental stage of the plant, date of data collection etc.
  - Take photo/video if needed
  - Useful for downstream analysis and reuse of data by other researchers

### Plant Height (PLTHT)

**Description/Procedure:**

Placing measuring stick on ground next to the root crown, "plant height" is measured at the ligule of the flag leaf.
See *Picture 1*

**Timing:** At plant maturity
**n** = 1 representative plant per plot
**Unit:** centimeter [cm]

*Picture 1*

**Notes:** One plant is considered sufficient since these are inbreds and hybrids and are not segregating for traits. Please record date measured.
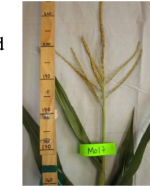
### Pollen Date

**Description/Procedure:**

Taken as [MM/DD/YY] to 50 percent of a plot exhibiting anther exertion on greater than half of main tassel spike. Day of anthesis recording is shown in *Picture 1*, whereas the day after is shown *Picture 2*.

**Timing:** At Flowering
**n** = 1 date per plot
**Unit:** [MM/DD/YY]

*Picture 1*            *Picture 2*

Image Credit: 2004, 2006; Purdue University, RL Nielsen

G2F Initiative SOP, 2017

# Get the phenotype right

- Exploratory analysis
  - Quantitative traits distribution
  - Outlier
- Model the factors in experimental design
  - Replications, treatments, years, locations etc.
  - Best Linear Unbiased Prediction (BLUP)
- Estimate the heritability of traits of interests
- [A tutorial on doing this](#)
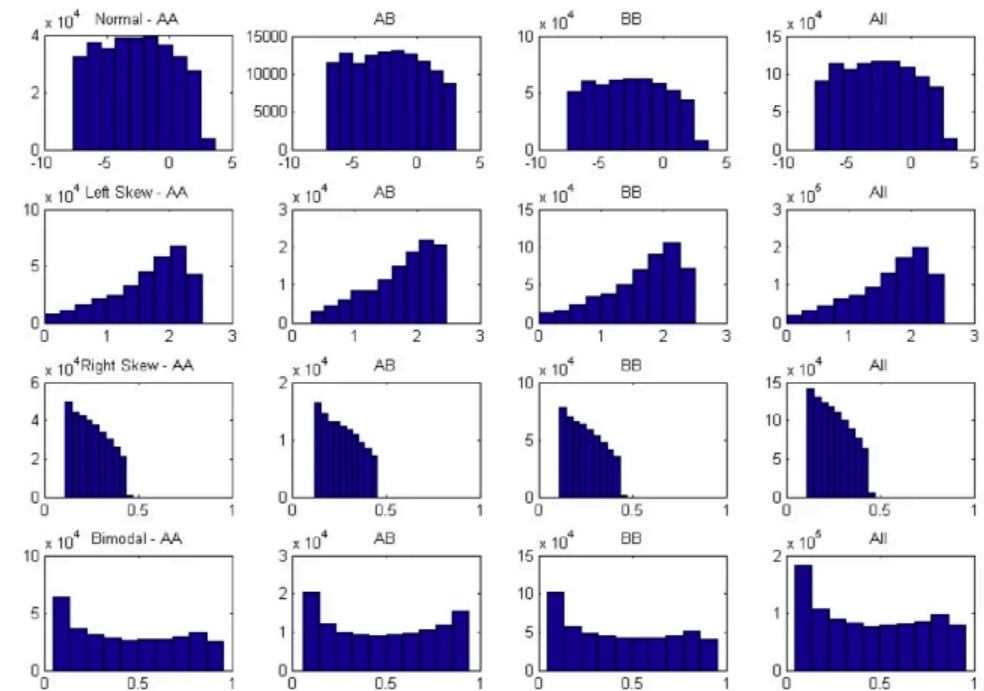


Goh *et al* (2009)

# Get the marker data right

- SNP data format
  - HapMap
  - VCF/BCF
  - ped/bed
  - Numeric
  - …
- Software for format conversion
  - TASSEL
  - VCFtools
  - GTOOL
  - Customized scripts
  - …

**VCF**

```
##fileformat=VCFv4.2
##contig=<ID=2,length=51304566>
##INFO=<ID=AC,Number=A,Type=Integer,Description="Allele count in genotypes">
##INFO=<ID=AN,Number=1,Type=Integer,Description="Total number of alleles in called genotypes">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
#CHROM  POS   ID  REF ALT QUAL FILTER  INFO  FORMAT    SAMPLE1    SAMPLE2    SAMPLE3    SAMPLE4    SAMPLE5    SAMPLE6    SAMPLE7
2  81170  .  C  T   .  .   AC=9;AN=7424  GT:DP:GQ  0/0:4:12   0/0:3:9    0/1:1:3    0/1:9:24   1/0:4:12   0/0:5:15   0/0:4:12
2  81171  .  G  A   .  .   AC=6;AN=7446  GT:DP:GQ  0/1:4:12   0/0:3:9    0/0:1:3    0/0:9:24   0/1:4:12   0/1:5:15   0/0:4:12
2  81182  .  A  G   .  .   AC=5;AN=7506  GT:DP:GQ  0/0:5:15   0/0:4:12   0/0:5:15   0/0:9:24   0/0:4:12   0/0:4:12   0/0:4:12
2  81204  .  T  G   .  .   AC=2;AN=7542  GT:DP:GQ  1/0:5:15   0/0:9:27   0/0:10:30  0/0:15:39  0/0:9:27   1/0:13:39  0/1:14:42
```

**BCF**

```
2  81170  .  C  T   .  .   AC=9;AN=7424  GT:0/0:0/0:0/1:0/1:1/0:0/0:0/0  DP:4:3:1:9:4:5:4        GQ:12: 9: 3:24:12:15:12
2  81171  .  G  A   .  .   AC=6;AN=7446  GT:0/1:0/0:0/0:0/0:0/1:0/1:0/0  DP:4:3:1:9:4:5:4        GQ:12: 9: 3:24:12:15:12
2  81182  .  A  G   .  .   AC=5;AN=7506  GT:0/0:0/0:0/0:0/0:0/0:0/0:0/0  DP:5:4:5:9:4:4:4        GQ:15:12:15:24:12:12:12
2  81204  .  T  G   .  .   AC=2;AN=7542  GT:1/0:0/0:0/0:0/0:0/0:1/0:0/1  DP:5:9:10:15:9:13:14    GQ:15:27:30:39:27:39:42
```

https://en.wikipedia.org/wiki/Variant_Call_Format

# Get the marker data right

- Filter SNP marker before GWAS
- Minor allele frequency (MAF)
  - the frequency at which the *second most common* allele occurs in a given population
  - MAF < 0.05: rare variants
- Missing rate
- Heterozygosity rate



**Effect size**

50.0 — High

High-effect from interacting low risk alleles

State-dependent risk alleles

Rare alleles causing Mendelian disease

Few single high-effect Common variants

A

B

3.0 — Intermediate

Low-frequency variants with intermediate effect

1.5 — Modest

1.1

Rare variants small effect - hard to identify

Combinations of Rare and Common variants

Common variants implicated in common disease by GWAS

Low

0.001    0.005    0.05

Very rare    Rare    Low frequency    Common

**Allele frequency**

Whitcomb *et al.* (2015)

# Address relatedness among individuals

- Control for population structure
- Calculate population structure with the STRUCTURE software (Pritchard Lab, Stanford)
  - Can be applied to microsatellites, RFLPs, AFLPs and SNPs
  - fastSTRUCTURE for large SNP datasets
- Principal Component Analysis (PCA)
  - Include the first few PCs that explain a large portion of variation



Beckett *et al* (2017)
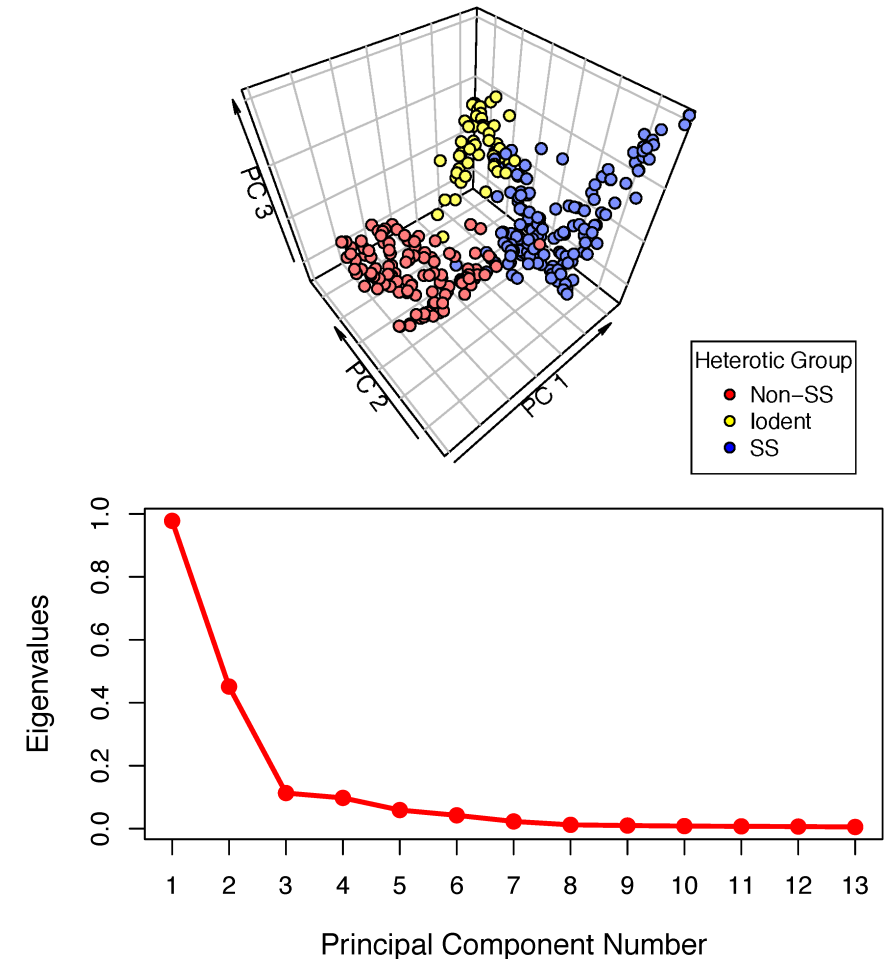
# Address relatedness among individuals

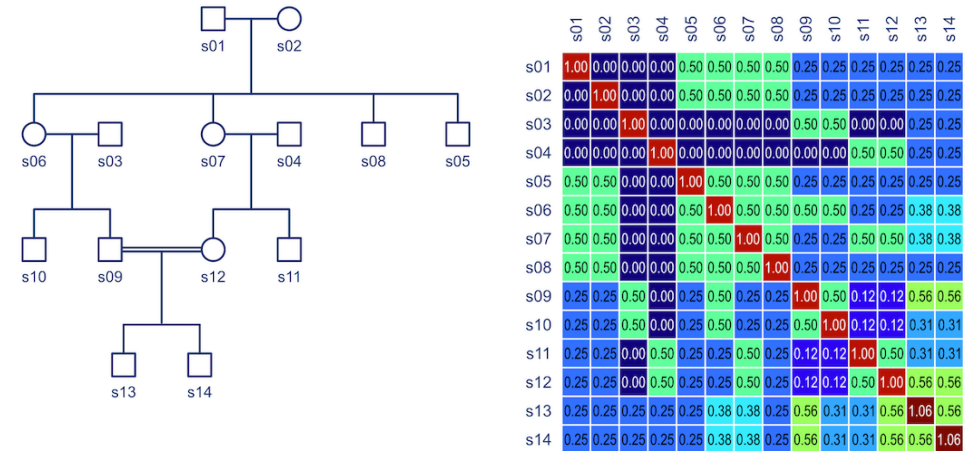- Control for kinship

- Identity by descent (IBD)

- Identity by state (IBS)
  - When two individuals possess the same alleles at a locus

- Algorithm for estimate kinship
  - Loiselle *et al* (1995)
  - VanRaden (2008)

- Explaination



https://brainder.org/2015/07/29/understanding-the-kinship-matrix/



Benjamin Neale lecture notes on population straatification

# GWAS models and software

129 Free Genome-wide Association Study (GWAS) Tools – Software and Resources

https://bioinformaticshome.com/tools/gwas/gwas.html

- Graphic User Interface (GUI)
  - Local
    - TASSEL (Trait Analysis by aSSociation, Evolution and Linkage)
  - Online
    - Cyverse DE, *easyGWAS*, GWAPP etc.
- Command line
  - TASSEL, PLINK, GEMMA etc.
  - GAPIT, MVP etc.

| Year | Method | Positive semidefinite matrix requirement[a] | Strategy for increasing computational speed | | | Computational speed | Statistical speed |
|------|--------|---------|---------|---------|---------|---------|---------|
| | | | Approximate/Two-step approach[b] | Matrix optimization[c] | Low-rank matrix | | |
| 2006 | Standard MLM | | | | | Low | High |
| 2007 | GRAMMAR | | + | | | Very fast | Intermed |
| 2008 | EMMA | + | | + | | Intermediate | High |
| 2010 | EMMAX | + | + | + | | Fast | High/Int |
| 2010 | P3D and CMLM | | + | | + | Fast | High/Int |
| 2011 | FaST-LMM | + | | + | | Fast | High |
| 2012 | GEMMA | + | | + | | Fast | High |
| 2012 | FaST-LMM-Select | + | | + | + | Very fast | High |
| 2014 | ECMLM | | + | | + | Intermediate | High/Int |
| 2014 | SUPER | + | | + | + | Fast | High |

Xiao *et al.* (2017)

# GWAS models and software

- Choice of GWAS model
  - Species and population
  - Statistical power
  - Computational resources
  - Prior knowledge of the traits of interests (genetic architecture, heritability etc.)



Miao *et al.* (2019)
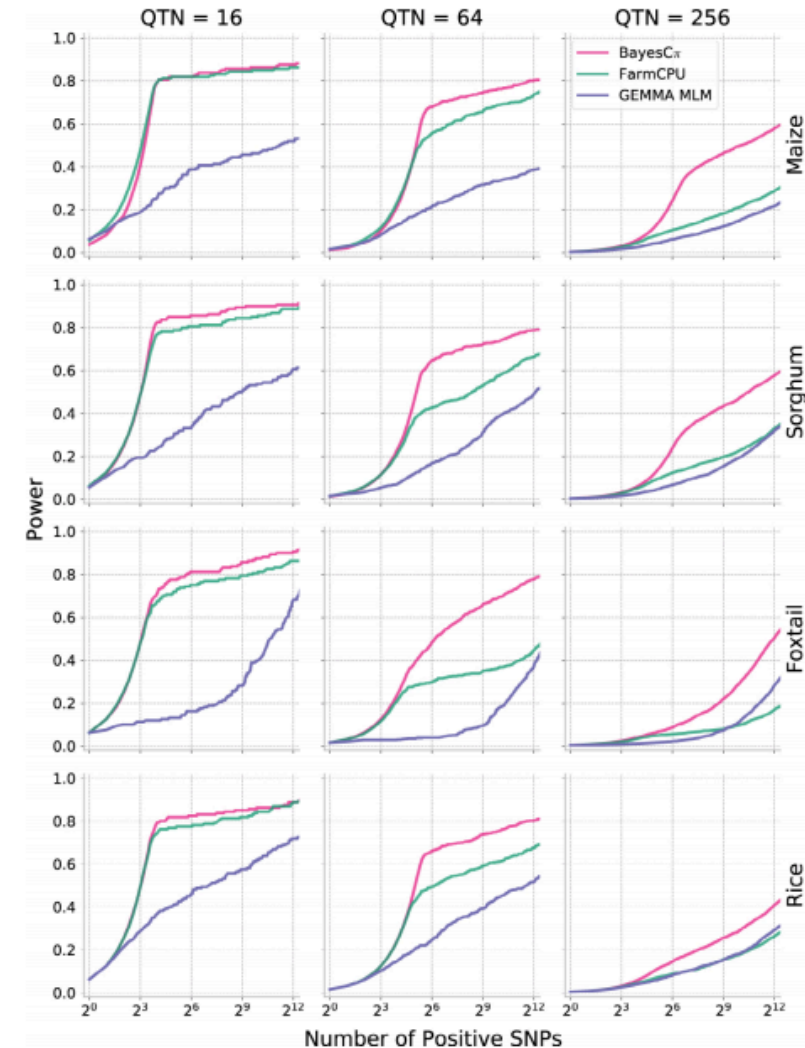
# Interpret GWAS results

Typical outputs from GWAS

Association table

| SNP | Chromosome | Position | P.value | maf | nobs | Rsquare.of.Model.without.SNP | Rsquare.of.Model.with.SNP | FDR_Adjusted_P-values |
|---|---|---|---|---|---|---|---|---|
| Fea2.4 | 4 | 132736424 | 3.13E-07 | 0.290035587 | 281 | 0.079004463 | 0.170450593 | 0.000966617 |
| PZB01223.3 | 3 | 192865132 | 3.88E-05 | 0.346975089 | 281 | 0.079004463 | 0.137165003 | 0.059980668 |
| PZA03748.1 | 2 | 7481079 | 0.000196769 | 0.195729537 | 281 | 0.079004463 | 0.126357611 | 0.193907122 |
| PZA00394.11 | 2 | 56930271 | 0.00025959 | 0.145907473 | 281 | 0.079004463 | 0.124537393 | 0.193907122 |
| PZA002219.6 | 3 | 219309117 | 0.000313461 | 0.379003559 | 281 | 0.079004463 | 0.123302809 | 0.193907122 |

Allele effects table

| SNP | Chromosome | Position | DF | t Value | std Error | effect |
|---|---|---|---|---|---|---|
| PZB00859.1 | 1 | 157104 | 276 | 0.255638527 | 0.501027645 | 0.128081969 |
| PZA01271.1 | 1 | 1947984 | 276 | -0.205390736 | 0.451124818 | -0.092656858 |
| PZA03613.2 | 1 | 2914066 | 276 | -1.143780776 | 0.4887135 | -0.558981106 |
| PZA03613.1 | 1 | 2914171 | 276 | 1.194922452 | 0.540019499 | 0.645281424 |
| PZA03614.2 | 1 | 2915078 | 276 | 0.277681398 | 0.489962989 | 0.136053608 |

35

GAPIT user manual

# Interpret GWAS results

- ## Manhattan Plot
  - A scatter plot used to display large number of data points (SNPs) and their significance in GWAS
  - Each data point is a SNP (not gene)
  - X-axis: genomic position
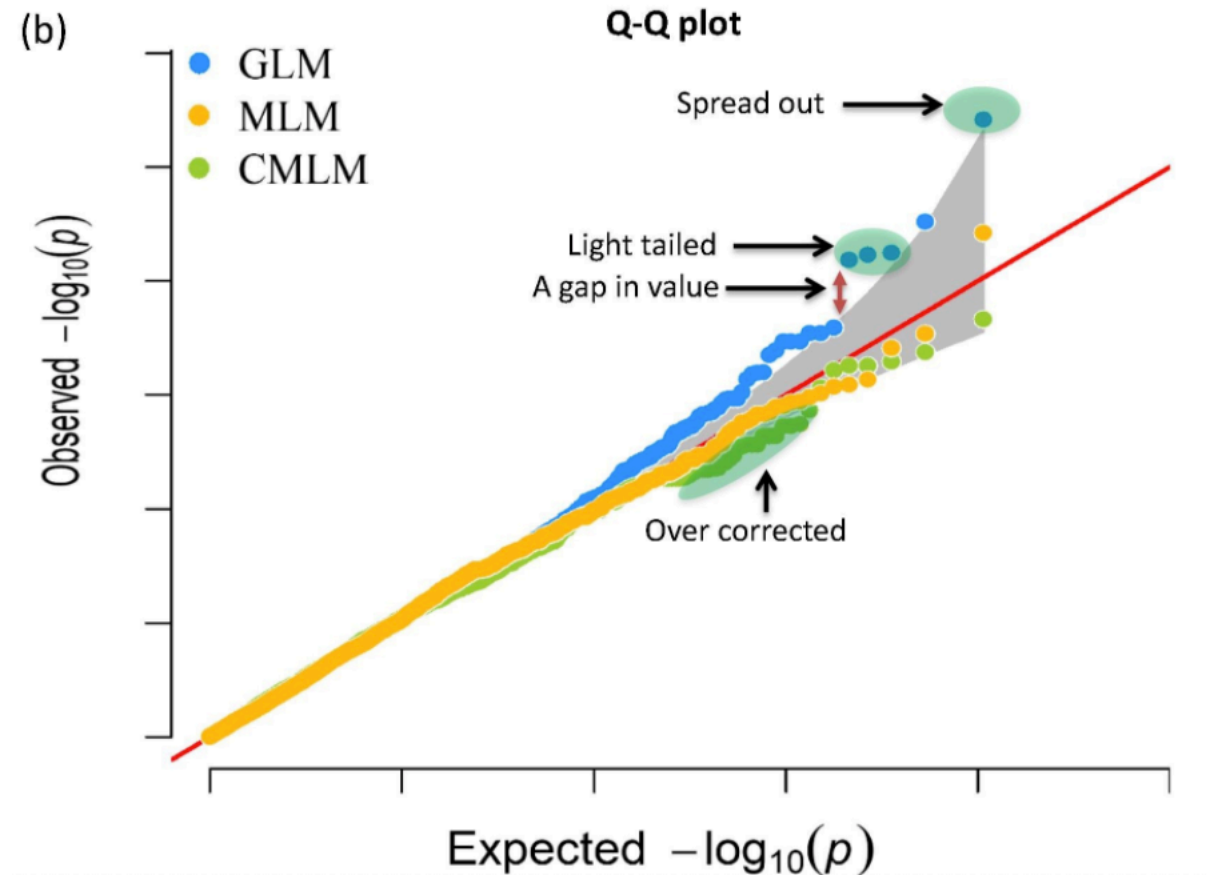  - Y-axis: negative logarithm of the association p-value



MLM

MMLM

FarmCPU

Cortes *et al.* (2020)
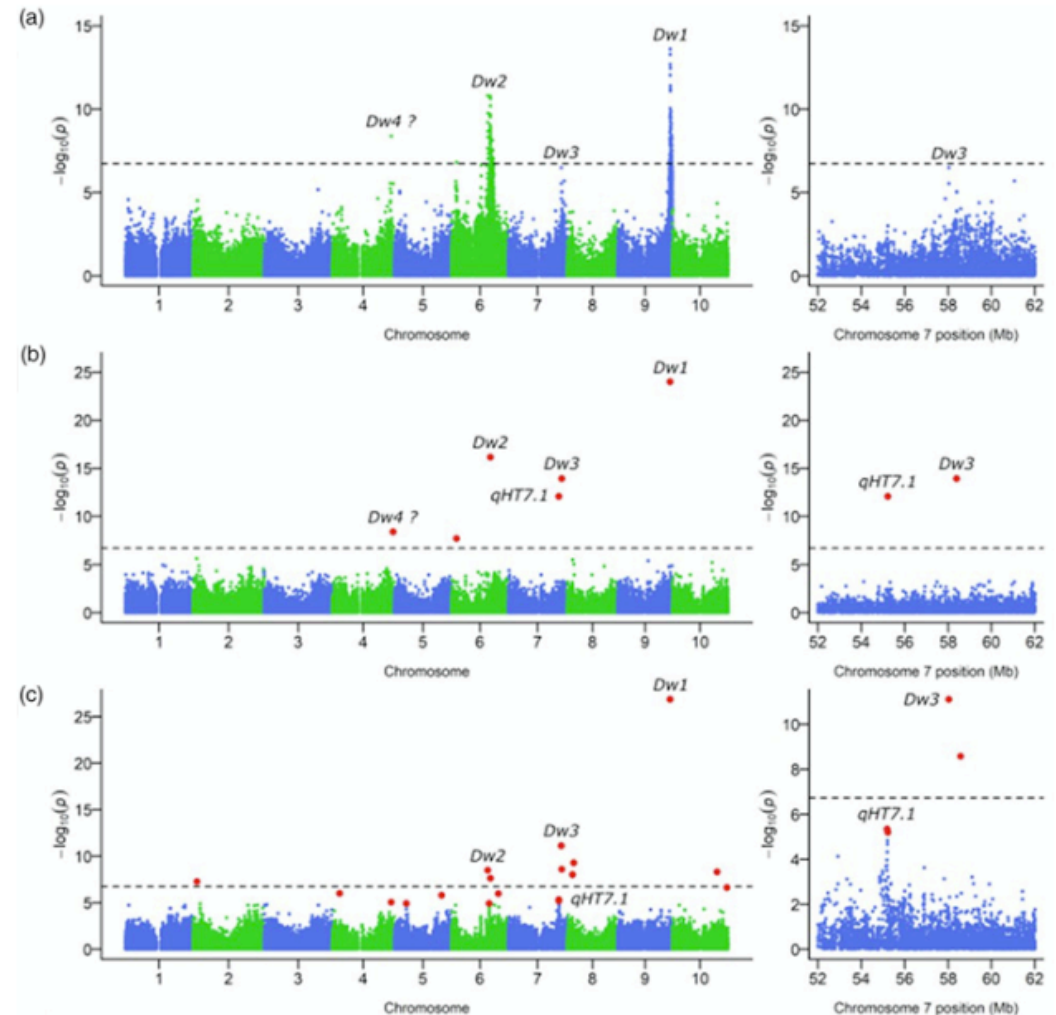
# Interpret GWAS results

- Q-Q plot (quantile-quantile plot)
  - A scatterplot created by plotting two sets of quantiles against each other
  - An essential tool for detecting problems in a GWAS



Alqudah *et al.* (2020)

# Interpret GWAS results

- How to decide on a cutoff for determining which p-values are significant?

- Multiple testing problem
  - E.g. N hypothesis tests were performed
  - If Type I error were set to a level (e.g. 0.05): the probability of incorrectly rejecting the null hypothesis
  - If the N tests were independent, and N is large (e.g. 1 million SNPs), we expect to incorrectly reject the null hypothesis ~50,000 times
  - Multiple testing problem: the more tests performed, the greater the probability of making  Type I errors



**Type I and Type II Error**

| Null hypothesis is... | True | False |
|---|---|---|
| Rejected | **Type I error** False positive Probability = α | **Correct decision** True positive Probability = 1−β |
| Not rejected | **Correct decision** True negative Probability = 1−α | **Type II error** False negative Probability = β |

# Interpret GWAS results

- Control for multiple testing
- Bonferroni correction
  - For a desired type I error α, set the Bonferroni type I error $\alpha_B$ to be

$$\alpha_B = \frac{\alpha}{N}$$

  - E.g. if we have N = 100,000 and we want an overall type I error to be 0.05, we require a test to have have p-value less than $5 \times 10^{-7}$
- False discovery rate (FDR)-based approaches
  - Uses the expected number of false positives to control for type I error
  - FDR = 0.05, 0.1 etc.
  - Benjamini-Hochberg (BH) procedure, Benjamini–Yekutieli (BY) procedure etc.

# Interpret GWAS results

- Search for candidate genes using local LD vs genome-wide LD

- Multiple candidate genes in the genomics region found by GWAS
  - Fine-mapping
  - Expression patterns of genes in related tissues
    - Gene expression datasets
    - Co-expression network
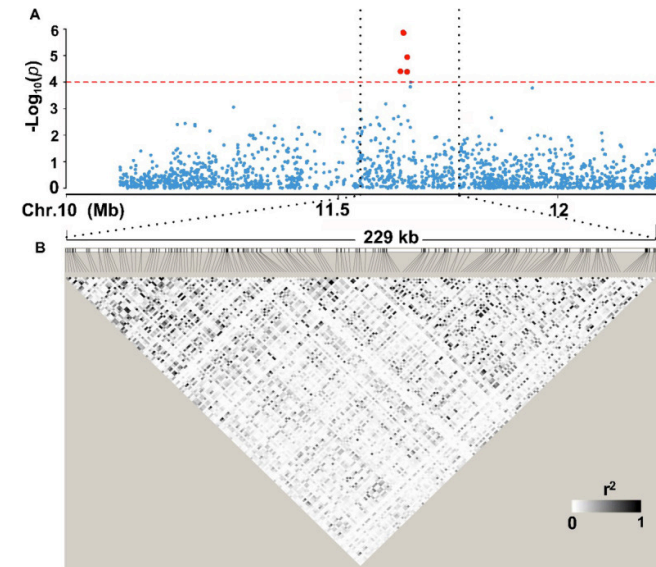


Yang *et al Rice* (2020)
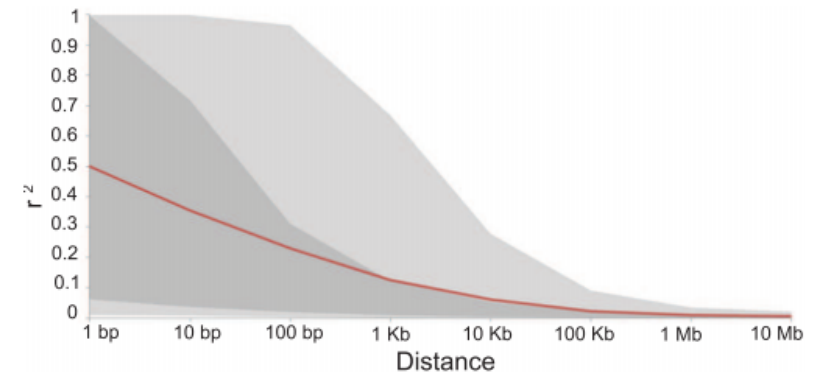


Romay *et al.* (2013)

# Summary

- Brief history and evolution of GWAS
  - Linkage mapping and association studies
  - Four key factors in GWAS: germplasm, genetic markers, statistical models, phenotype
  - Control for population structure and kinship (MLM)
- How to initiate a GWAS experiment
  - Choice of germplasm (genetic diversity, population size etc.), type and number of genetic markers, field experimental design and phenotyping, choice of GWAS model
  - How to interpret the GWAS results

# Acknowledgement

- Dr. Pat Schnable
- Dr. Eric Lyons
- Dr. Ryan Bartelme
- Dr. Nicole Scott
- Dr. Tingting Guo
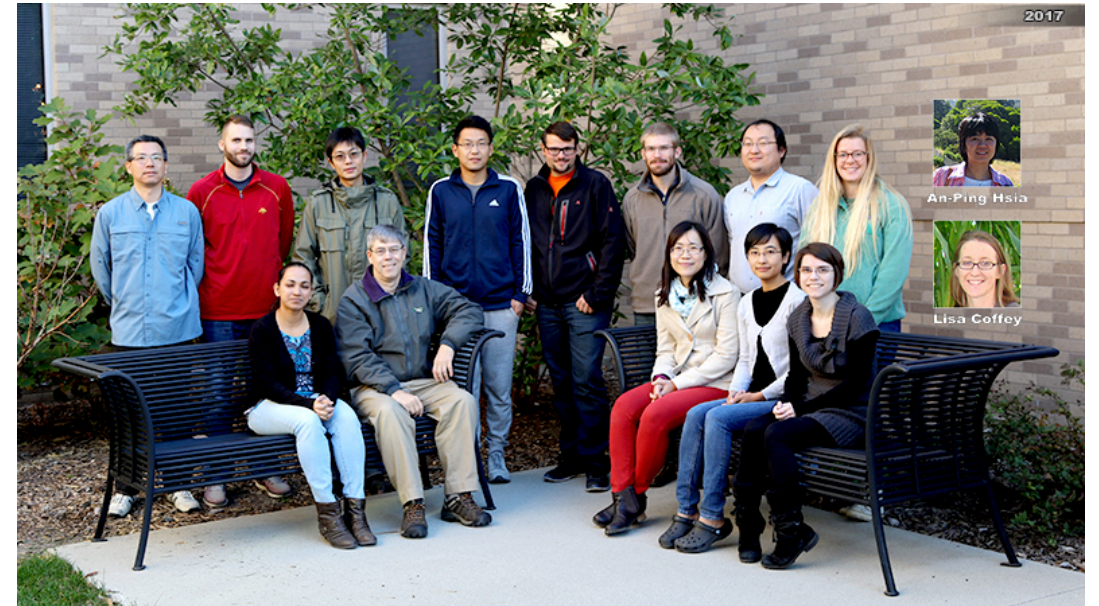


Schnable lab @ ISU



Schnable lab @ CAU

42