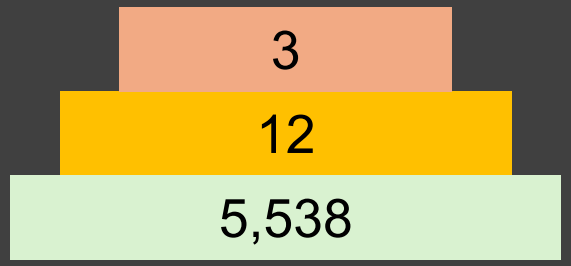
Do you know the number of crops that have been used throughout human history?



Any idea of how many we have lost so far?

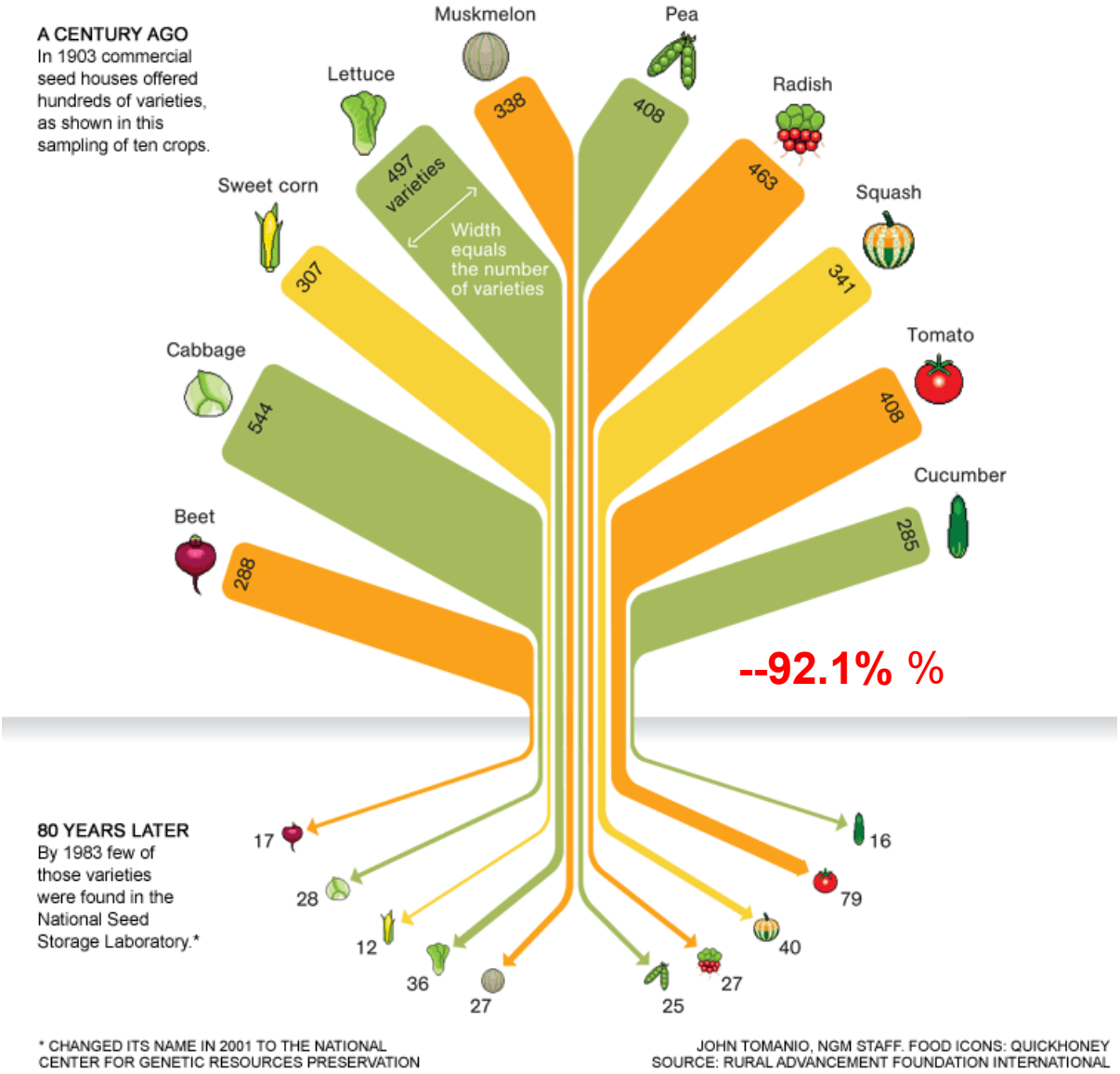
Do you know the number of crops that have been used throughout human history?

3

12

5,538

Any idea of how many we have lost so far?



Do you know the number of crops that have been used throughout human history?

3

12

5,538

Any idea of how many we have lost so far?

Why it's important?

Gros Michel

Cavendish

Cocoa bean prices exceeded $10,000 per ton for the first time in history

World 27.03.2024

Source: iz.ru 2418 EN 中文 DE FR عربى

The exchange value of cocoa beans i
$ 10 thousand per 1 ton, having incr
This was announced on Tuesday, Ma

Climate Change and U.S. Agricultural Exports

How l
Comp

Study links climate change with wheat blast; warns crop yield could drop by 75% in South America, Africa by 2050

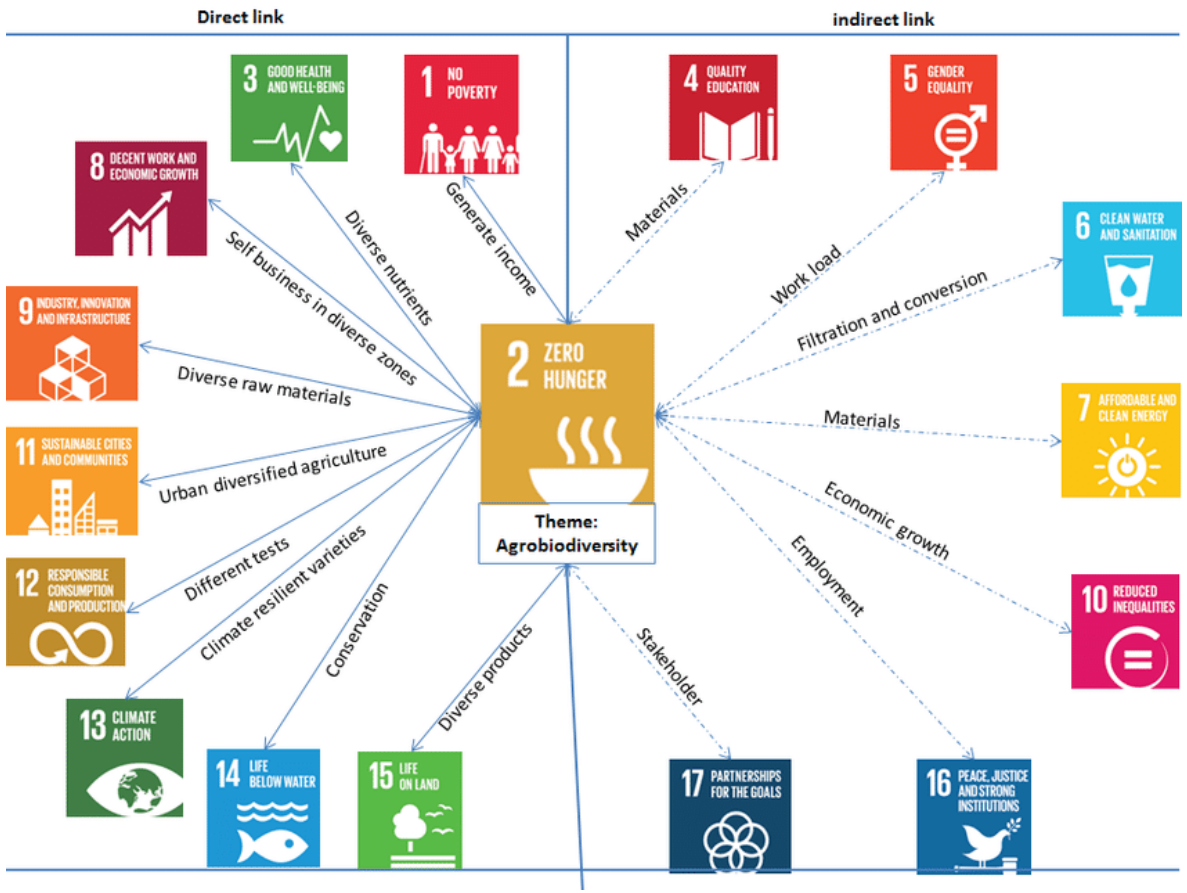
CIMMYT's study warns that climate change could cause wheat blast to reduce global wheat yields by 13% by 2050.

By CIMMYT

We study Agrobiodiversity

Agrobiodiversity refers to the diverse animals, plants, and microorganisms used in agriculture and food production.

- Genetic level: Varieties and breeds for food, fodder, fiber, energy, and medicine
- Species level: Soil organisms, predators, pollinators
- Habitat level: Organisms in agroecosystems

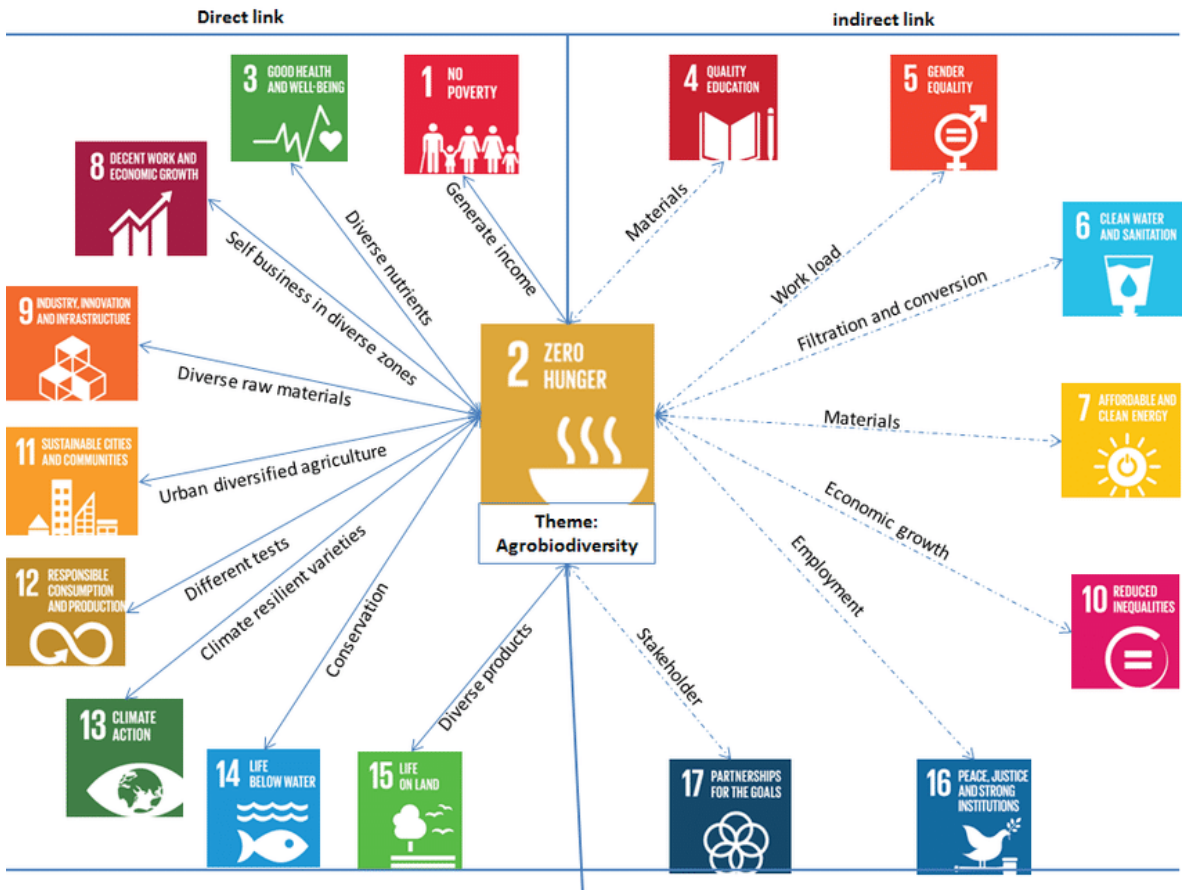


We study Agrobiodiversity

Agrobiodiversity refers to the diverse animals, plants, and microorganisms used in agriculture and food production.

- Genetic level: Varieties and breeds for food, fodder, fiber, energy, and medicine
- Species level: Soil organisms, predators, pollinators
- Habitat level: Organisms in agroecosystems

Tobias Recha and Geon Kang



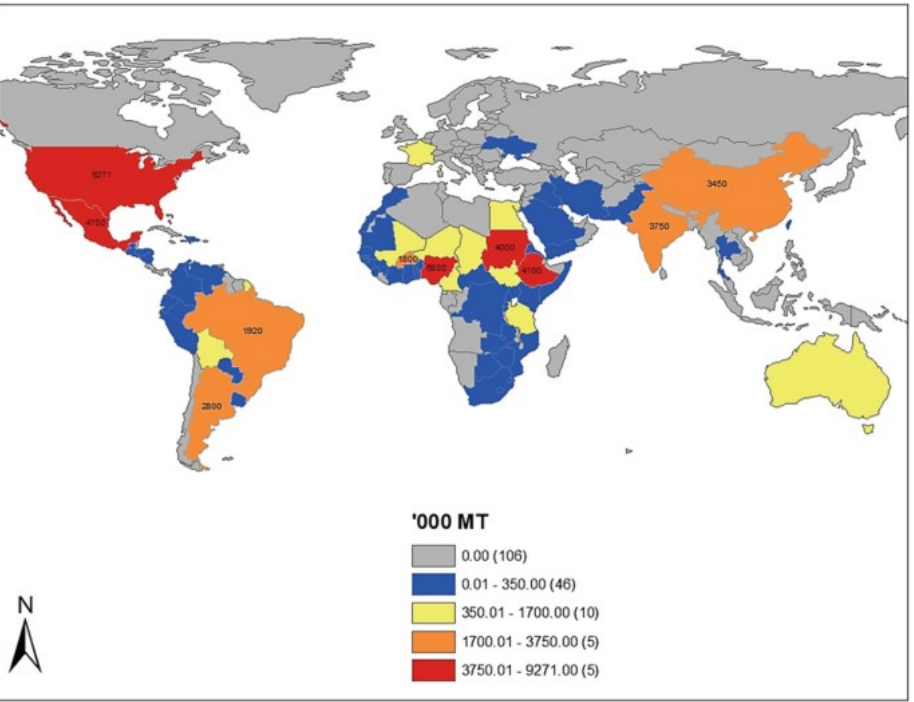
Aim: To solve global food security through crop improvement

Exploring Phenotypic and Genotypic Variations Across Sorghum Varieties

Ph.D. Agrobiodiversity
Tobias Recha
Geon Kang

CENTER OF PLANT SCIENCES
Sant'Anna
Scuola Universitaria Superiore Pisa

Introduction



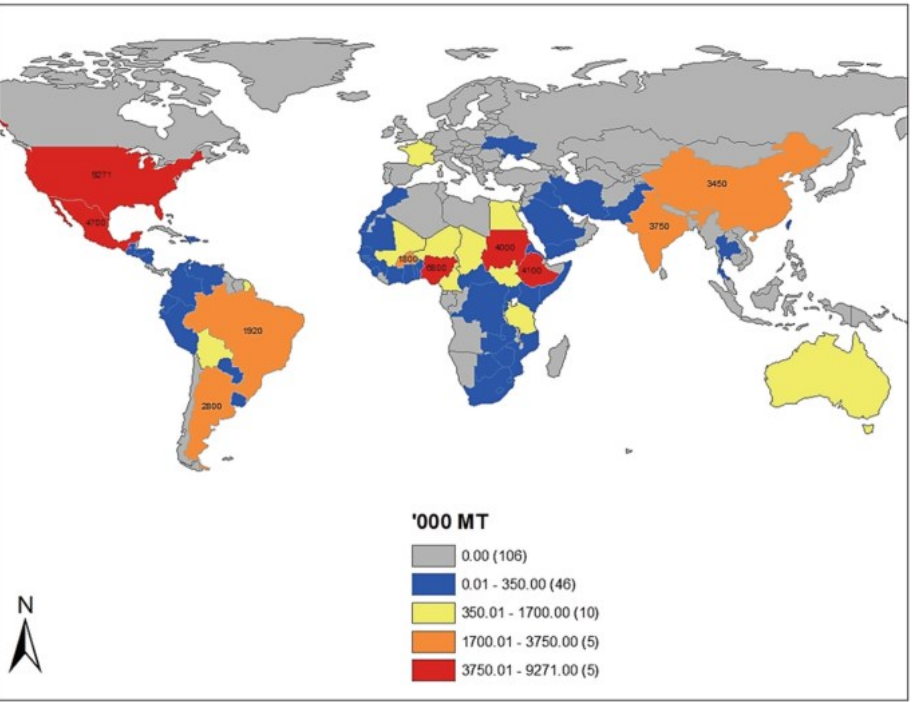
Sorghum (*bicolor. L.), Diploid, 2n = 20 :*
A Resilient Crop for Global Food Security

Resilience: Thrives in drought, heat; crucial in sub-Saharan Africa.

Diversity: **Vast phenotypic and genotypic variation.**

Breeding Goals: Improved varieties for diverse climates, higher yield, better nutrition, climate resilience.

Introduction



Sorghum (*bicolor. L.), Diploid, 2n = 20 :*
A Resilient Crop for Global Food Security

Resilience: Thrives in drought, heat; crucial in sub-Saharan Africa.

Diversity: **Vast phenotypic and genotypic variation.**

Breeding Goals: Improved varieties for diverse climates, higher yield, better nutrition, climate resilience.

Many resistance in modern crop is derived from wild relative
What is breeding or crop improvement?

Selection intensity
Selection accuracy
Genetic variance

$$R_t = \frac{ir\sigma_A}{y}$$

Genetic gain

Years per cycle

Understanding phenotypic and genotypic variations is so important!!!

Introduction

Aims:

1. To explore sorghum phenotypic and genetic traits using statistical and computational analysis
2. To promote future multidisciplinary research by conveying information about what we do to non-agronomy major students.

- **Objective:**

1. Investigating relationships between phenotypes and genotypes of sorghums
2. Comparison results of GWAS and Lasso .regression

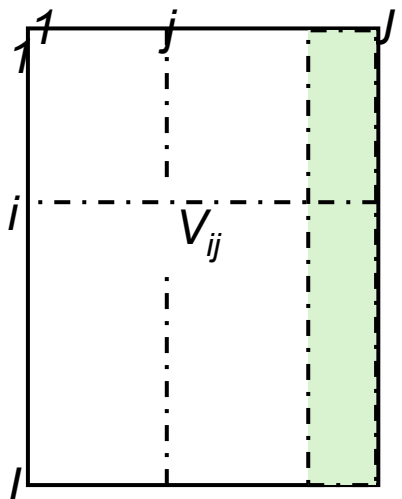
Material and methods

Sorghum Field Trials: Western Kenya

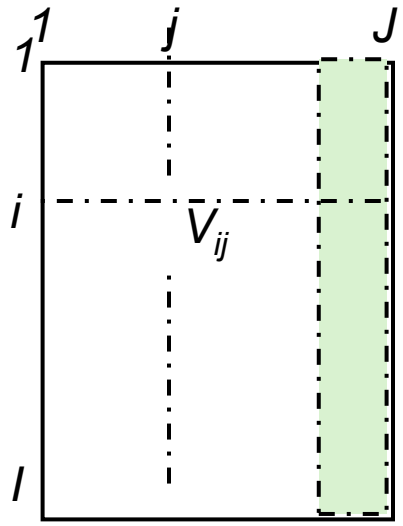
- **Season**: 2023 Long rains
- **Design**: 350 genotypes, randomized blocks, dual replicates
- **Traits Measured**:
  - Growth: Plant Stand, Height, Leaf Size, Tillers
  - Reproduction: Flowering Days
  - Yield: Grain, Biomass, Seed Weight
  - Structure: Lodging, Stem Girth, Panicle Size

Different years phenotype data
350 x traits

Season 1: 2023/03



Season 2: 2023/12



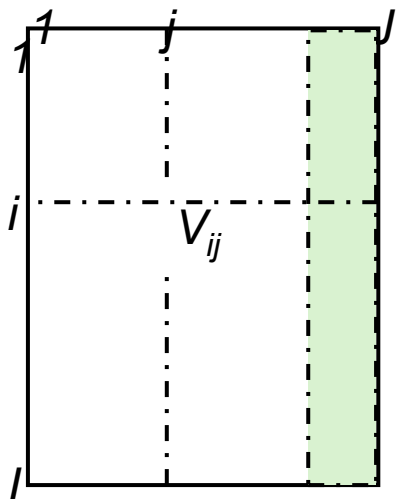
Material and methods

Sorghum Field Trials: Western Kenya

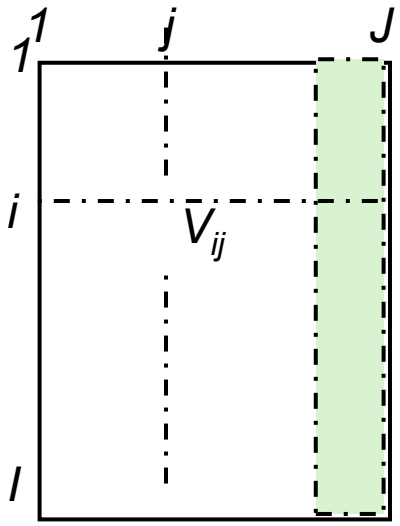
- **Season**: 2023 Long rains
- **Design**: 350 genotypes, randomized blocks, dual replicates
- **Traits Measured**:
  - Growth: Plant Stand, Height, Leaf Size, Tillers
  - Reproduction: Flowering Days
  - Yield: Grain, Biomass, Seed Weight
  - Structure: Lodging, Stem Girth, Panicle Size

Different years phenotype data
350 x traits

Season 1: 2023/03

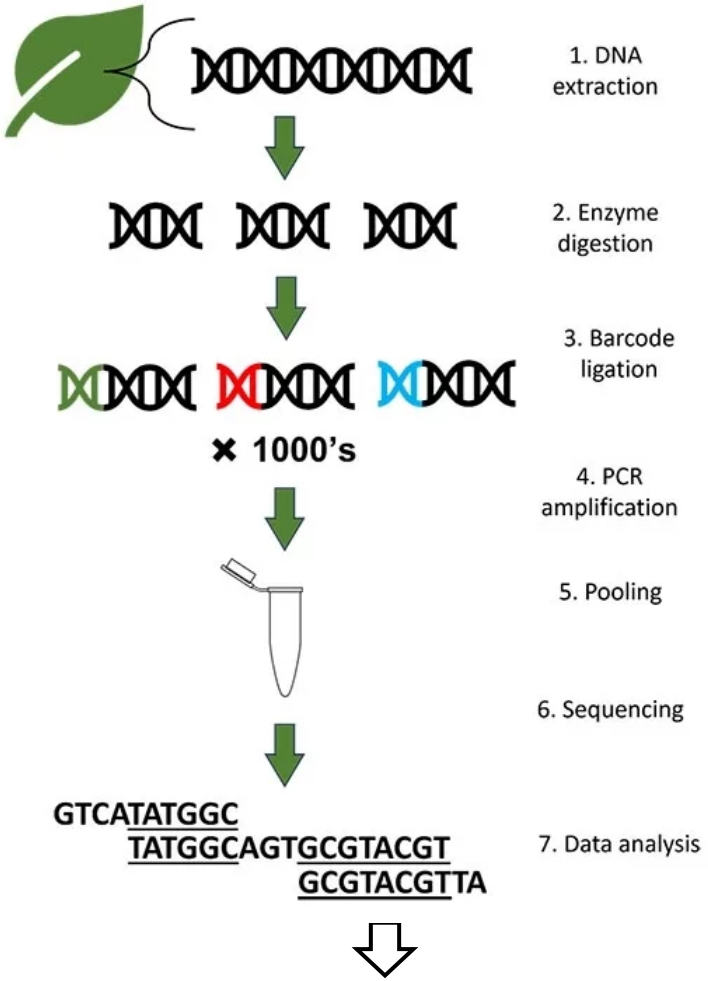


Season 2: 2023/12



+

DNA extraction and sequencing



SNP data frame
350 X 69K (High feature)

Material and methods

- **Unsupervised learning**
  - Hierarchical Clustering
  - Principal Component Analysis (PCA)
  - Locally Weighted Scatterplot Smoothing (LOWESS)
  - Linear Discriminant Analysis (LDA)

1. To identify crop traits linked to the phylogenetic tree.
2. To uncover genetic diversity through linked crop traits.
3. To visualize trait relationships for better understanding.
4. ToExplore how sorghum head type differs from other traits.

- **Supervised learning**
  - Ridge regression
  - Lasso regression

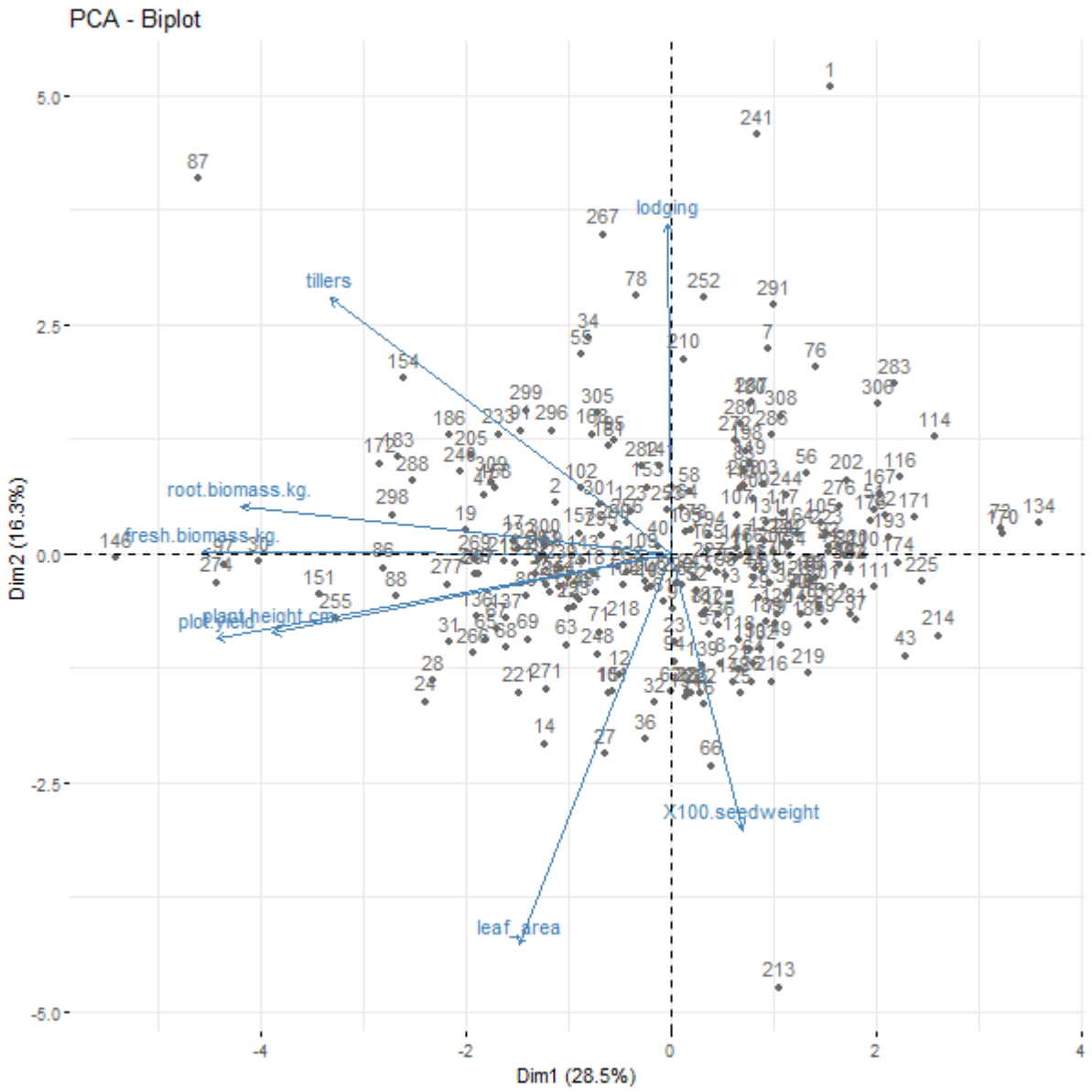
- **Additional statical tool.**
  - Best Linear Unbiased Prediction (BLUP)
  - Ridge regression BLUP
  - Genome-Wide Association Study (GWAS)

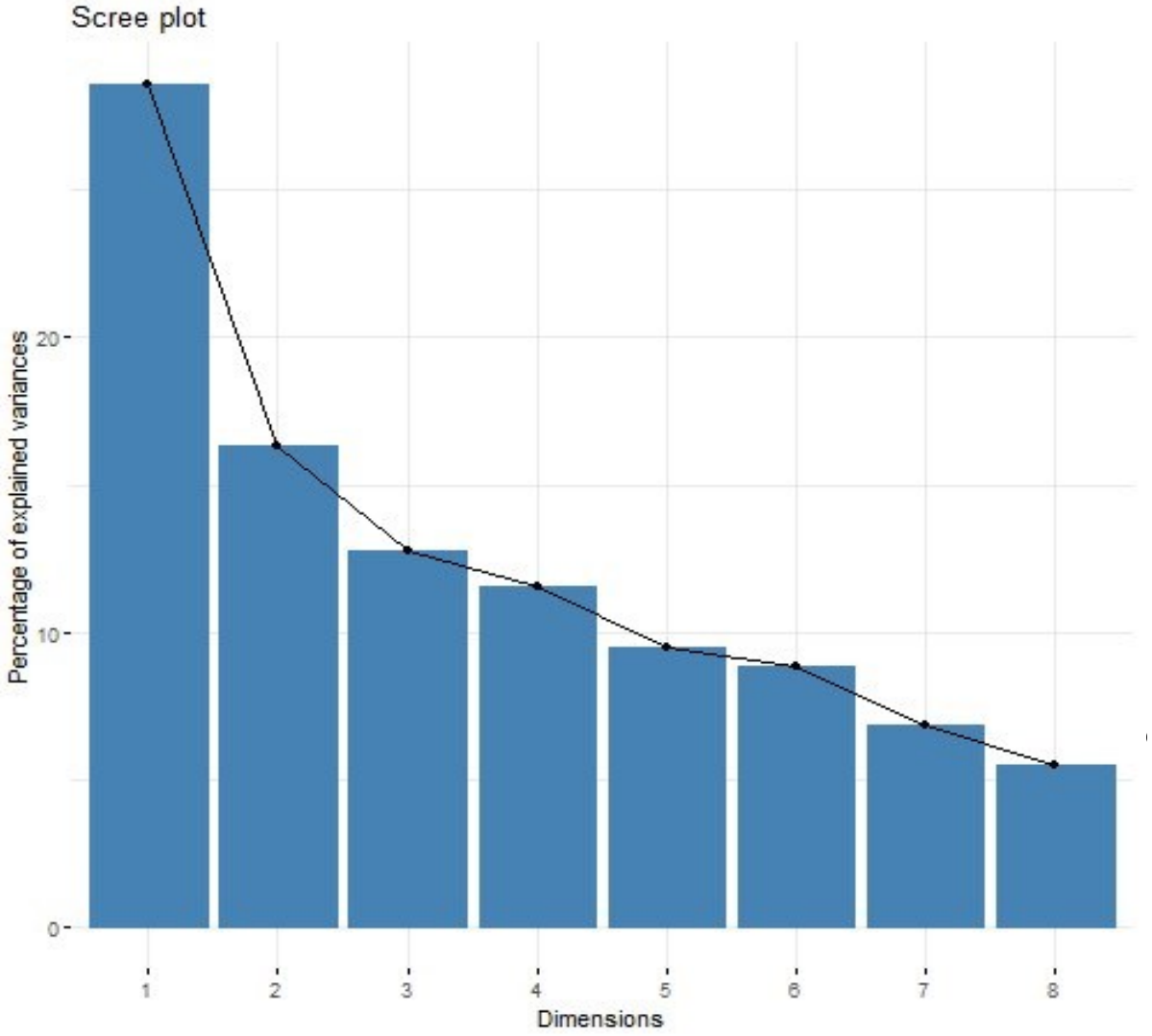
To compare performance between prediction model between phenotype and genotype.

Results - Unsupervised learning: PCA – Phenotypes

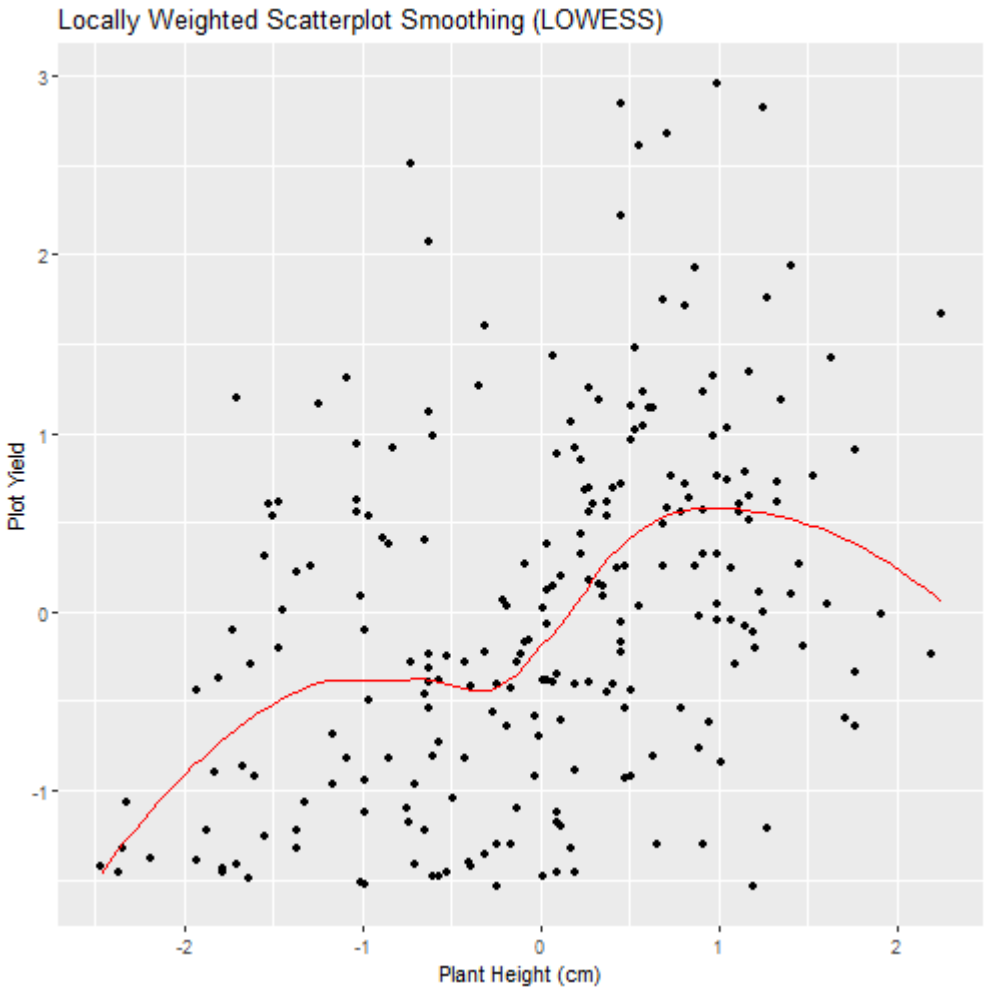
Understanding linked traits in breeding is key

Root biomass, Fresh biomass, plot.yield, and plant height are highly contributed to Dimension 1.

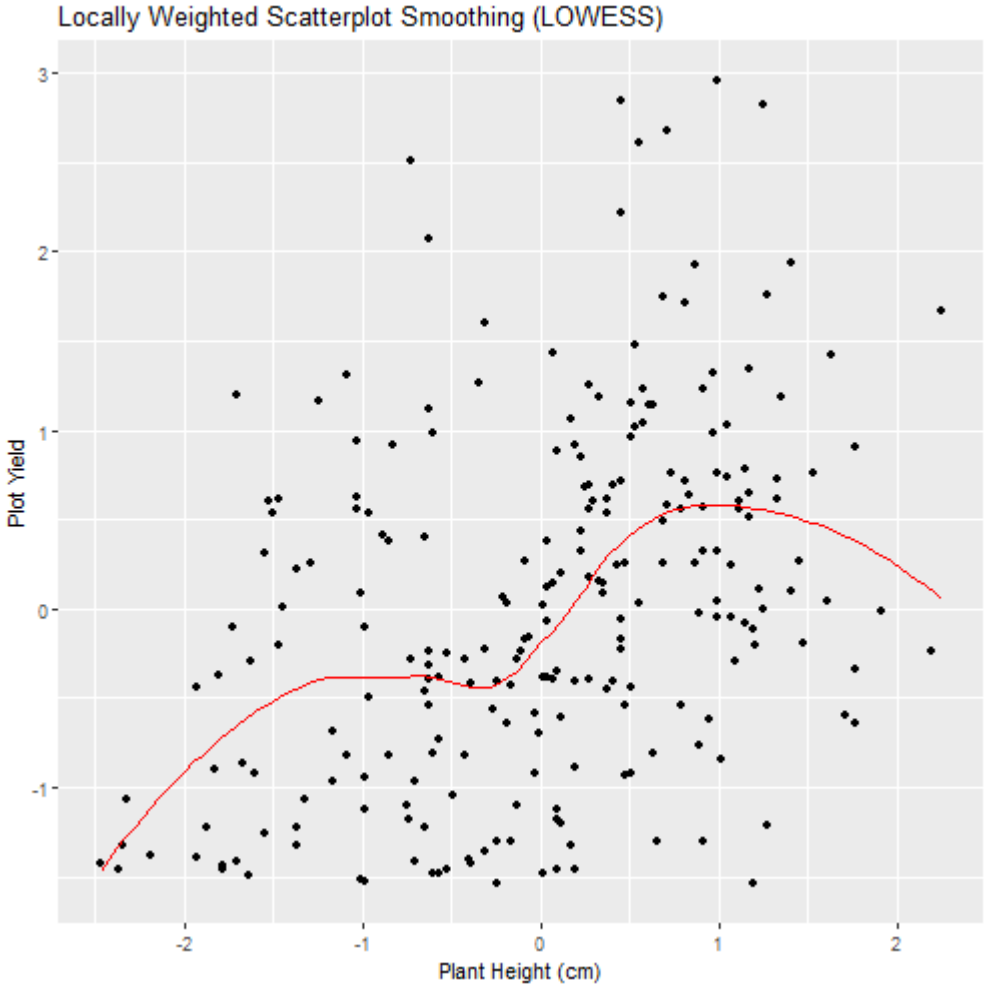




Results - Unsupervised learning: Locally Weighted Scatter Plot Smoothing

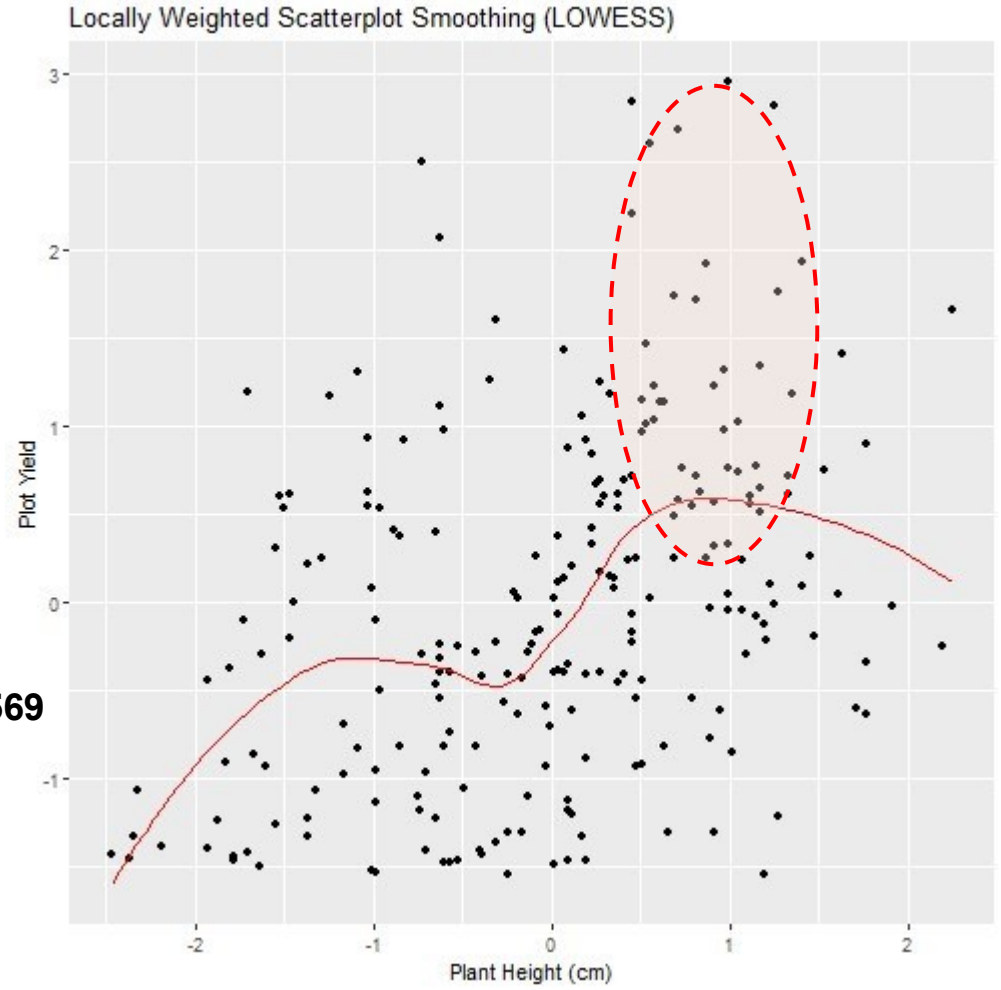


Results - Unsupervised learning: Locally Weighted Scatter Plot Smoothing



LOOCV

spans	degrees	err
0.6833333	2	0.8030569



It seems wise to choose varieties in the orange zone.

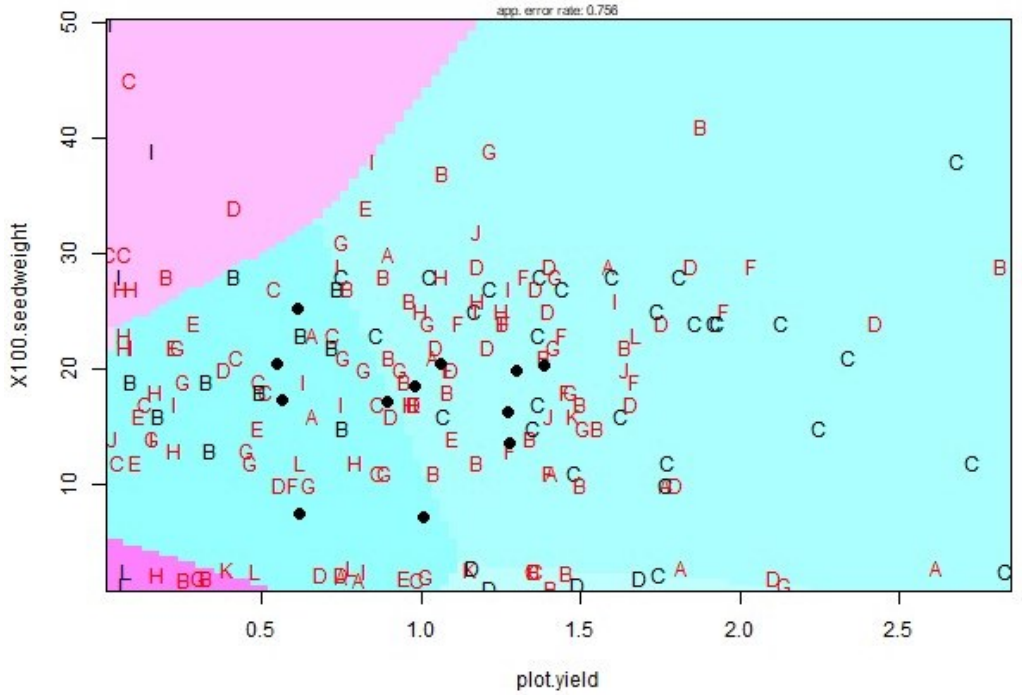
Results - Unsupervised learning: Linear Discriminant Analysis (LDA) - Sorghum head

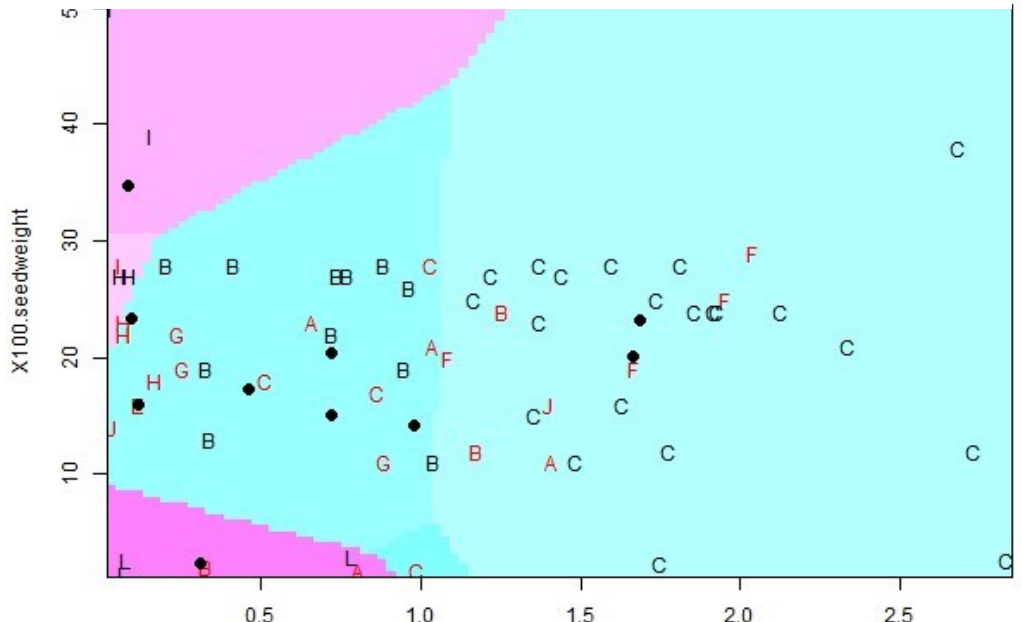
Sorghum head type

Very loose drooping primary branches	A
Compact elliptic	B
Half broomcorn	C
Semi-loose drooping primary branches	D
Loose erect primary branches	E
Very lax panicle typical of wild sorghums	F
Loose drooping primary branches	G
Compact oval	H
Semi-loose erect primary branches	I
Very lax panicle typical of wild sorghums	J
Semi-compact elliptic	K
Very loose erect primary branches	L

LDA formula: head.shape~ X100.seedweight+plot.yield

Accuracy test
Training (80%) and test (20%)
Sensitivity: 0.50
Specificity : 0.25



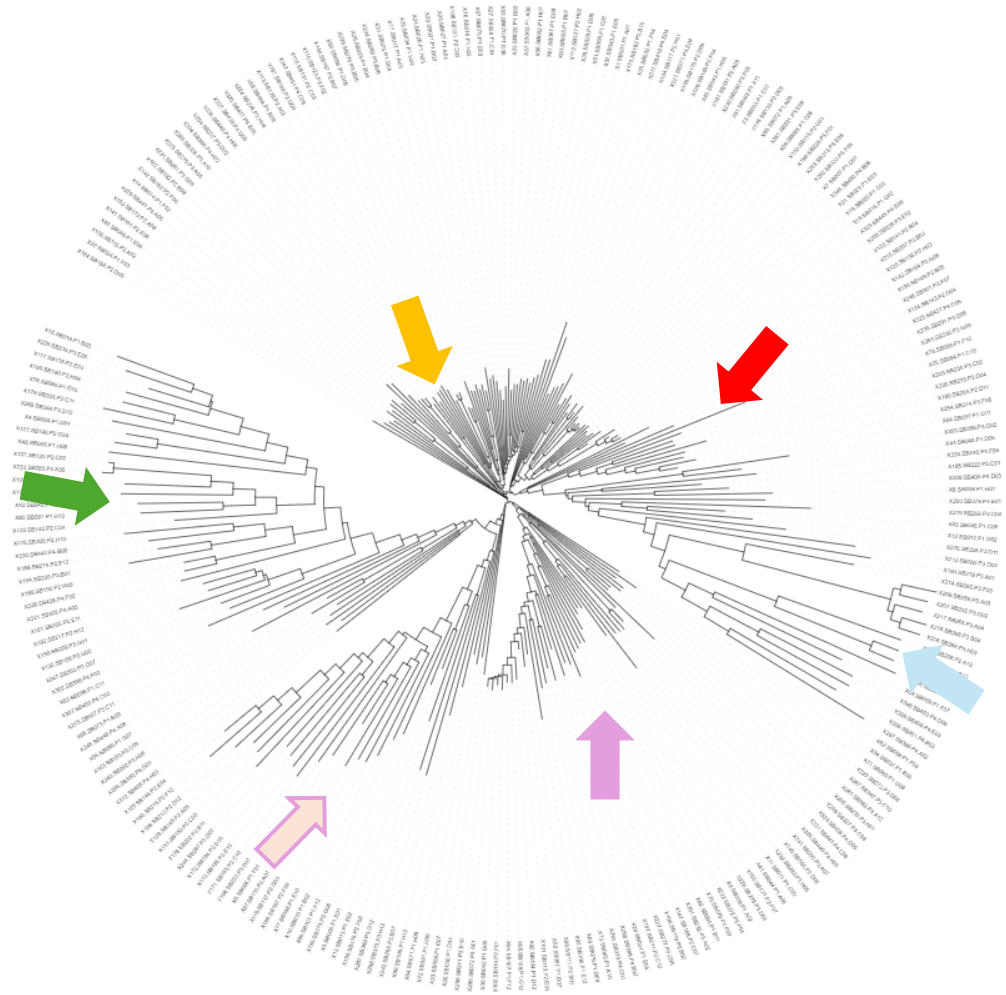


Results - Unsupervised learning: : Hierarchical Clustering and PCA – Genetic level

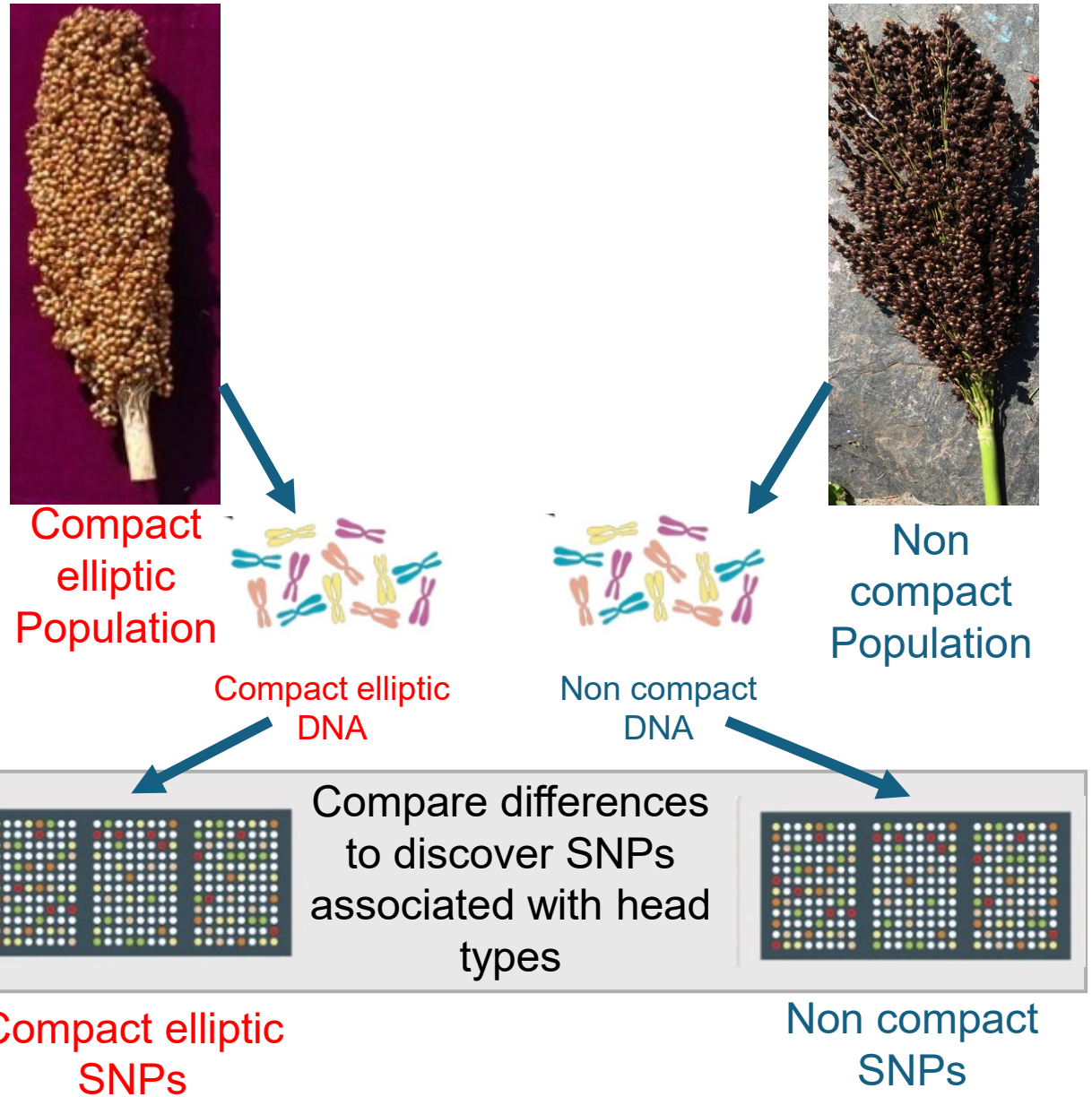
Compact elliptic : "SB020" "SB023" "SB059" "SB060" "SB067" "SB069" “SB072"
"SB141" "SB159" "SB184" "SB234" "SB269" “SB270" "SB398"

Half_broomcorn:"SB005" "SB025" "SB034" "SB071" "SB076" "SB082"
"SB093" "SB107" "SB131" "SB148" "SB160" "SB162" "SB189"
"SB205" "SB292" "SB303" "SB341" "SB342" "SB370" "SB372"
"SB375" "SB382" "SB388" "SB418" "SB442"

Phylogeny of sorghum genotypes



Results – Genome-Wide Association Study (GWAS)



Genome-Wide Association Study (**GWAS**) :
A method that allows scientists to find genes that have an association with observable traits

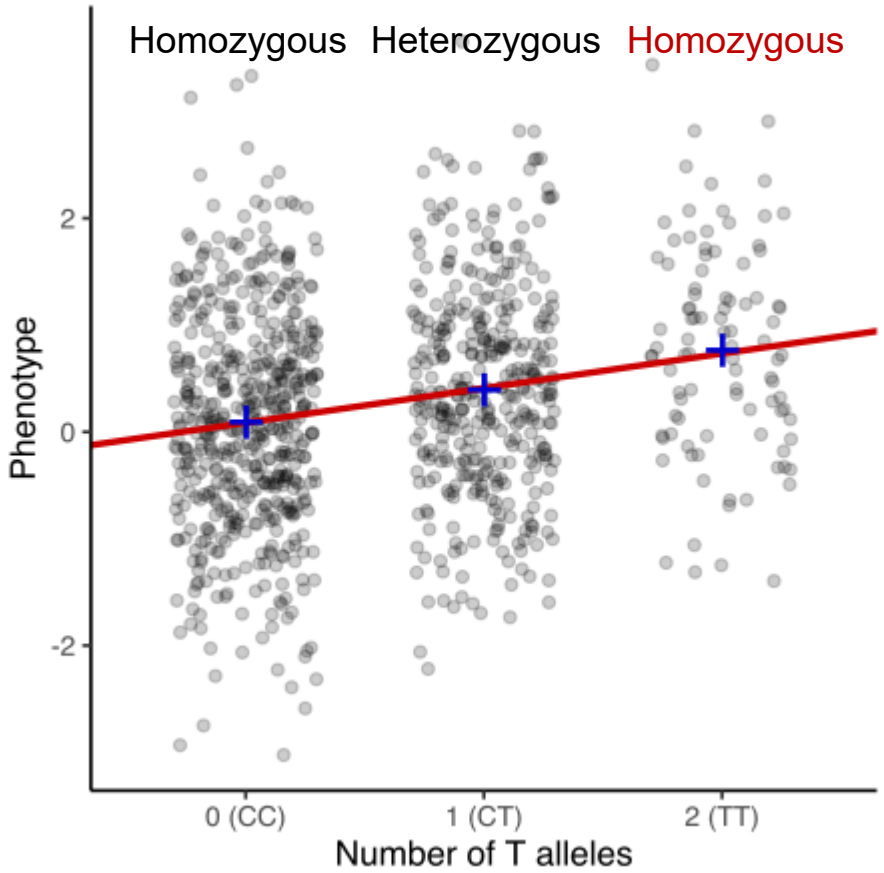
- Logistic regression in GWAS for binary phenotypes.
- It elucidates the logarithm of trait odds based on genotype.

$$\log\left(\frac{\Pr(Y=1\,|\,X=x)}{\Pr(Y=0\,|\,X=x)}\right)=\mu+x\beta.$$

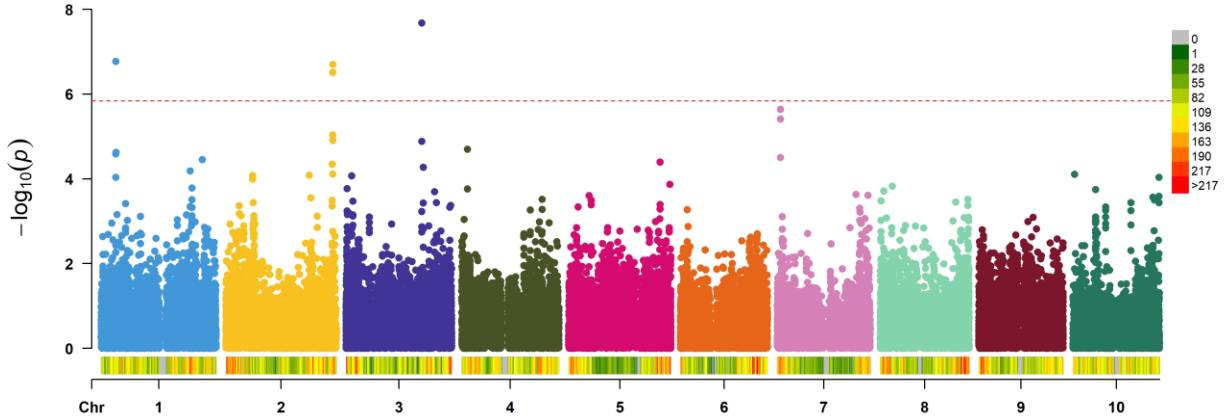
μ is the logarithm of odds ('log-odds') for genotype 0 and β is the log of odds) ratio between genotype 1 and 0 (and exp(β) is the corresponding odds ratio).

Results – Genome-Wide Association Study (GWAS)

$$\log\left(\frac{\Pr(Y=1\,|\,X=x)}{\Pr(Y=0\,|\,X=x)}\right)=\mu+x\beta.$$

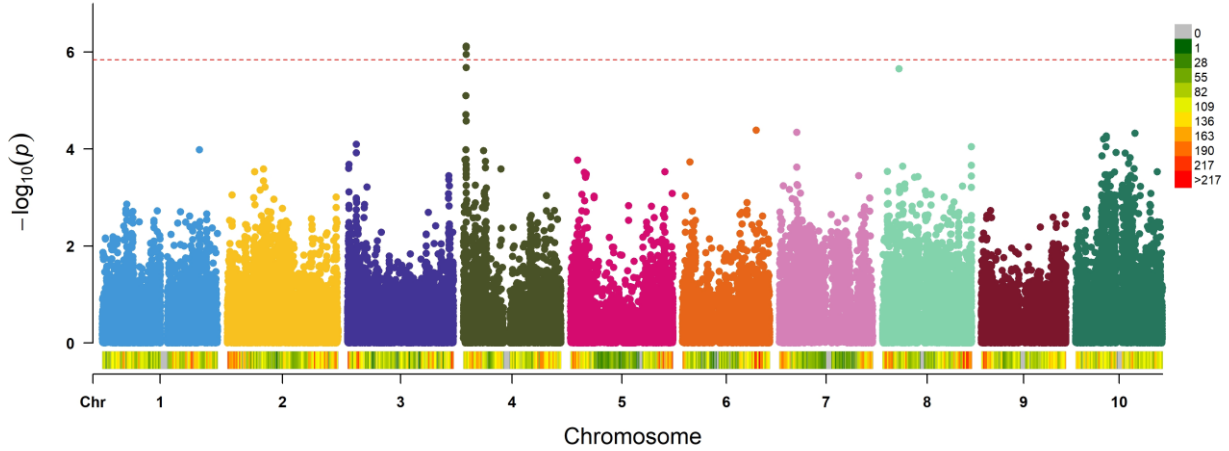


Compact elliptic



SNP 9933:5:+, CHROM1, position 10319147
SNP 124195:45:+, CHROM2, position 75185962
SNP 124229:353:-, CHROM2, position 75221846
SNP 168618:68:+, CHROM3, position 53254930

Half broomcorn

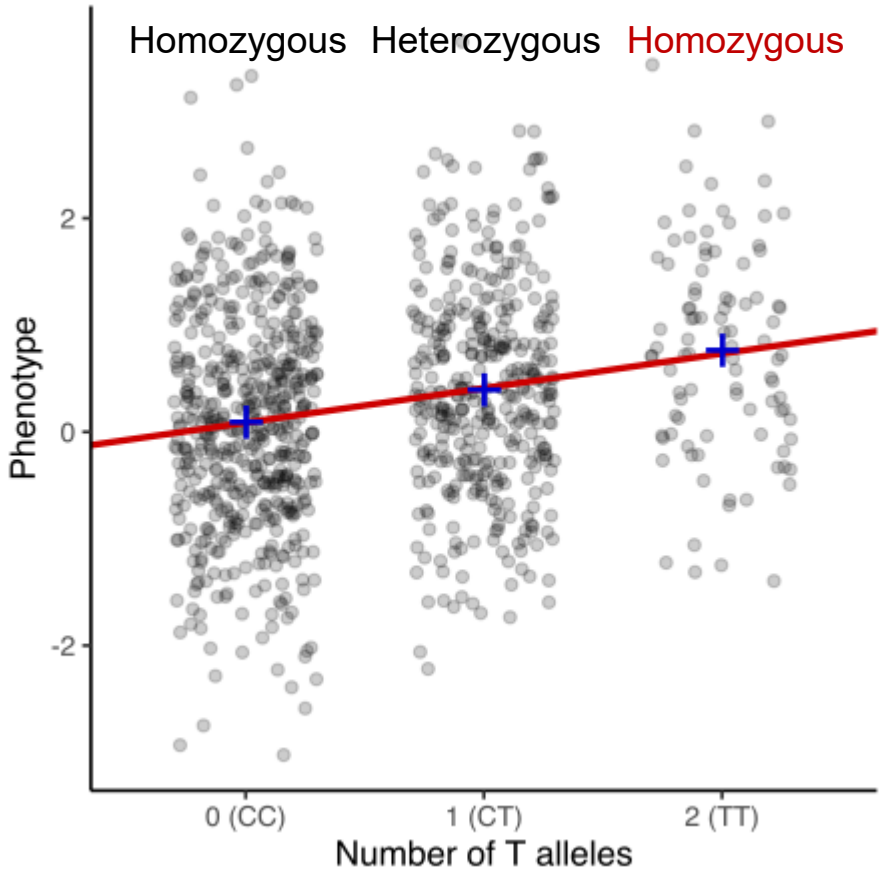


SNP 191353:132:-, CHROM4 position 2076595
SNP 191353:117:-, CHROM4, position 2076610
SNP 191353:94:-, CHROM4, position 2076633

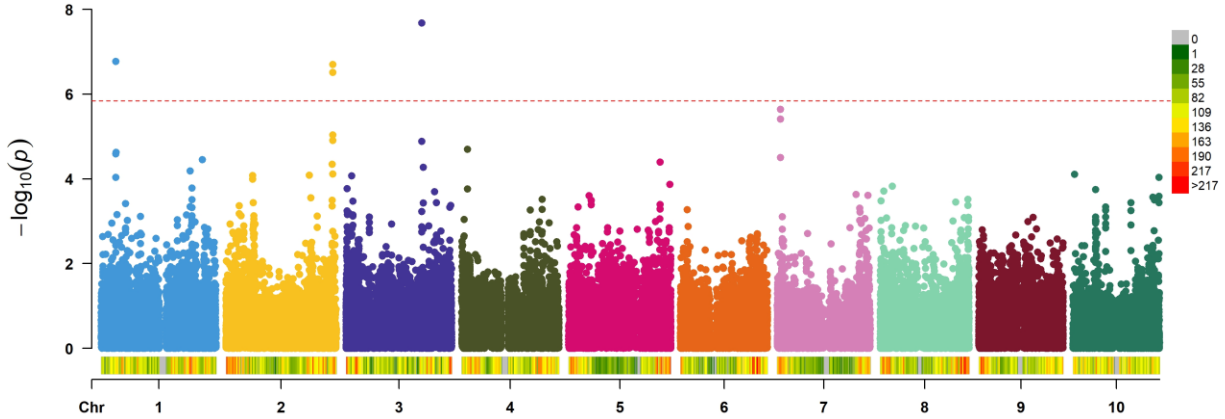
Results – Genome-Wide Association Study (GWAS)

What about ridge and lasso regression?

$$\log\left(\frac{\Pr(Y=1\,|\,X=x)}{\Pr(Y=0\,|\,X=x)}\right)=\mu+x\beta.$$

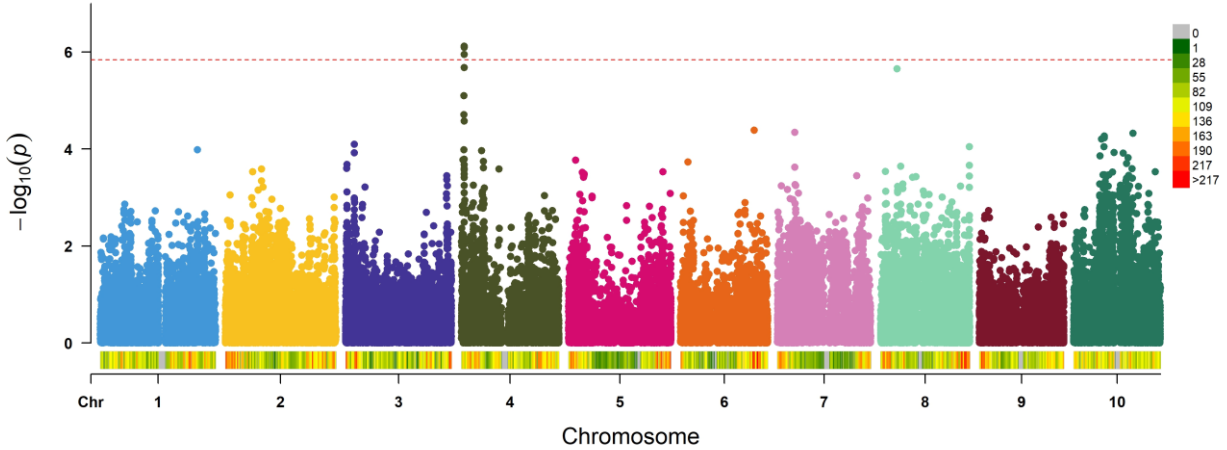


Compact elliptic



SNP 9933:5:+, CHROM1, position 10319147
SNP 124195:45:+, CHROM2, position 75185962
SNP 124229:353:-, CHROM2, position 75221846
SNP 168618:68:+, CHROM3, position 53254930

Half broomcorn



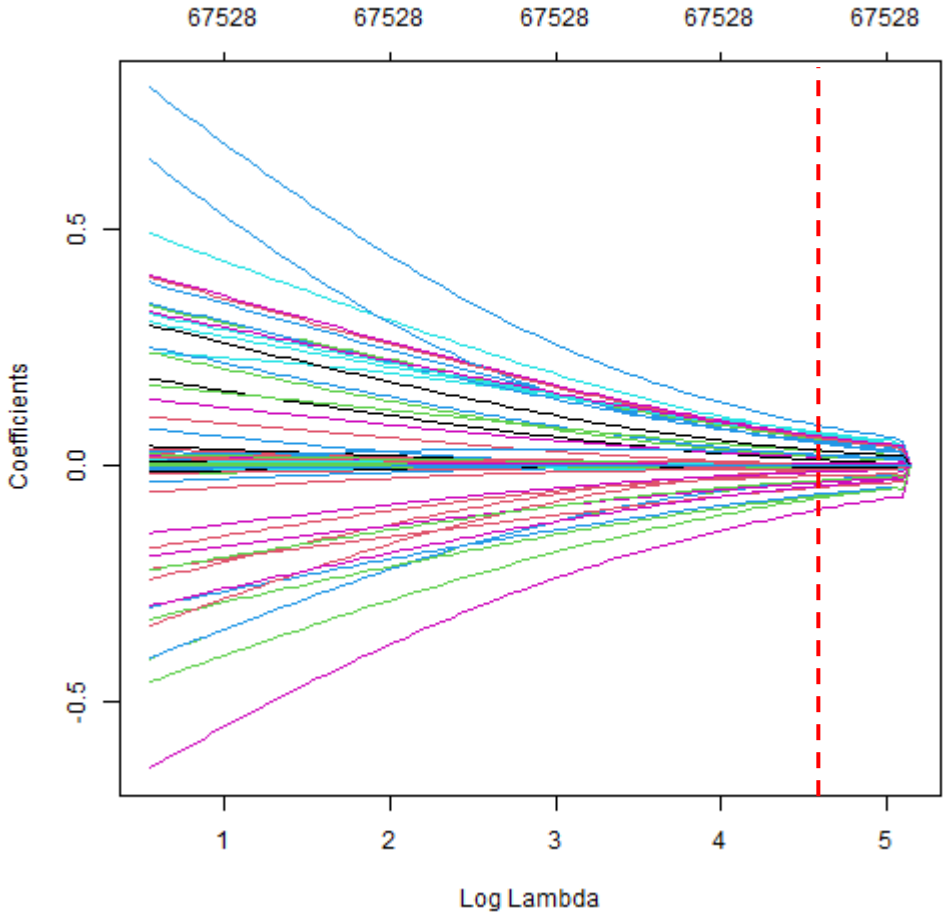
SNP 191353:132:-, CHROM4 position 2076595
SNP 191353:117:-, CHROM4, position 2076610
SNP 191353:94:-, CHROM4, position 2076633

Results - Supervised learning: Ridge and LASSO

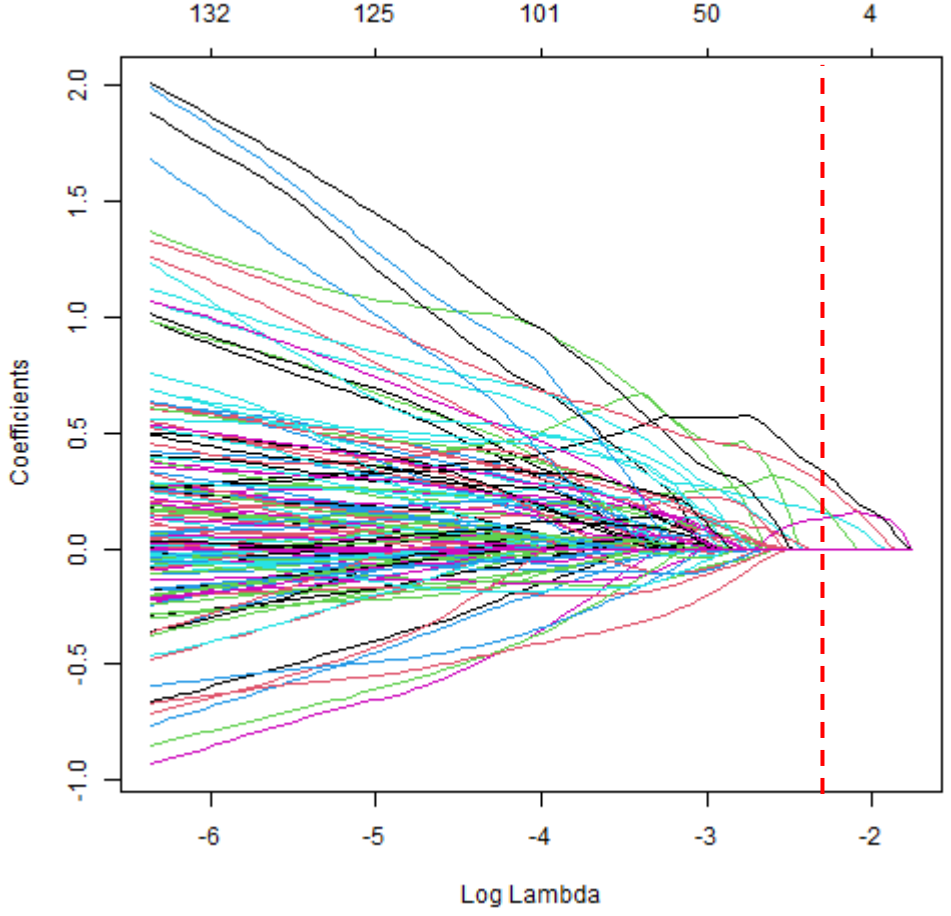
Half broomcorn

Change in size of coefficients depending on Lambda

Ridge regression



Lasso regression



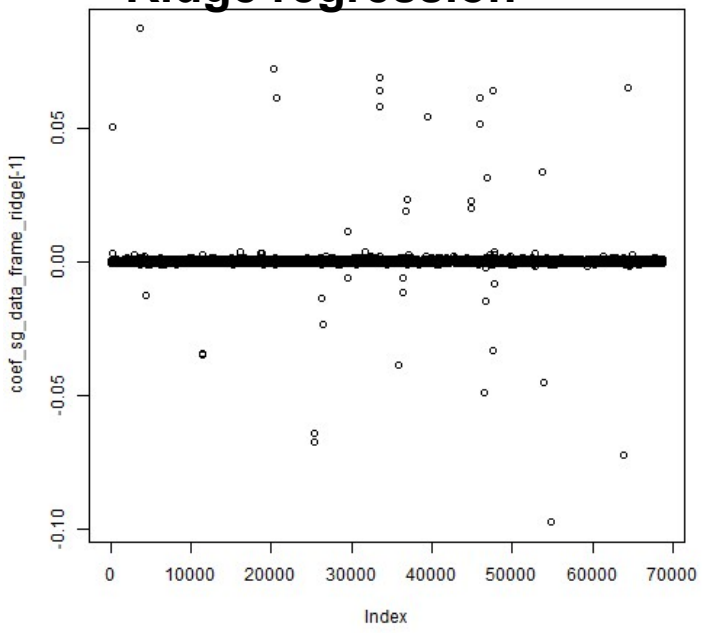
CV result: Lambda.min 94.104, log: 4.544

Lambda.min 0.089 , log: -2.409

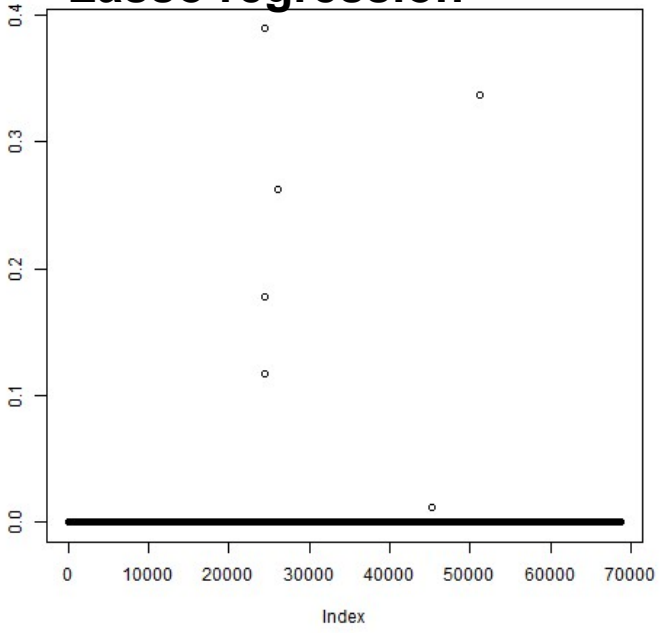
Results – GWAS, Rige and Lasso….

Variable and coefficient plot

Ridge regression



Lasso regression



Rige		Lasso		GWAS
SNP	Coef	SNP	Coef	SNP
30521:33:-	0.087	402529:102:-	0.337	191353:132:-
157423:245:-	0.072	350017:27:-	0.013	191353:117:-
255086:115:+	0.069	202300:19:+	0.262	191353:94:-
499907:7:-	0.065	191353:94:-	0.392	
255086:127:+	0.064	191353:117:-	0.113	
375622:123:-	0.064	191272:62:+	0.178	
354300:398:-	0.061			
159639:64:-	0.061			
255086:23:+	0.058			
255086:75:+	0.058			
306530:101:-	0.054			
354300:30:-	0.051			
3331:118:+	0.050			
197732:68:+	-0.064			
197732:41:+	-0.068			
430039:59:+	-0.098			

Half broomcorn

SNP 191353:132:- Chromosome 4 Ankyrin repeat family protein **GWAS**

SNP 191353:117:- Chromosome 4 Ankyrin repeat family protein **GWAS & Lasso**

SNP 191353:94:- Chromosome 4 Ankyrin repeat family protein **GWAS & Lasso**

SNP 191272:62:+ Chromosome 4 microtubule associated protein, putative, expressed **Lasso**

SNP 202300:19:+ Chromosome 4 ---------- **Lasso**

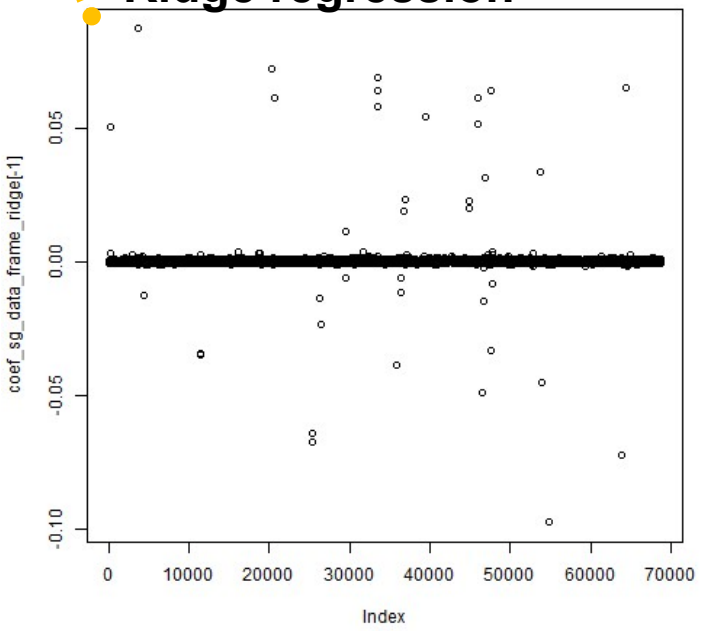
SNP 350017:27:- Chromosome 7 ---------- **Lasso**

SNP 402529:102:- Chromosome 8 ---------- **Lasso**

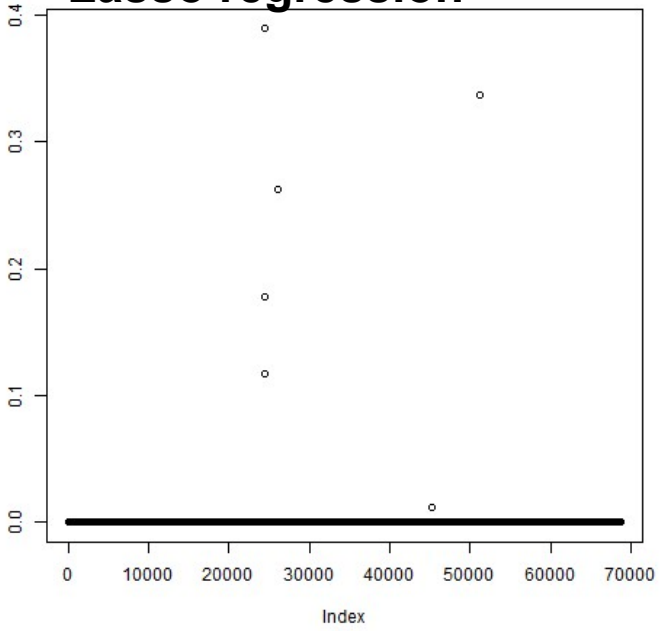
Results – GWAS, Rige and Lasso….

Variable and coefficient plot

Ridge regression



Lasso regression



Rige		Lasso		GWAS
SNP	Coef	SNP	Coef	SNP
30521:33:-	0.087	402529:102:-	0.337	191353:132:-
157423:245:-	0.072	350017:27:-	0.013	191353:117:-
255086:115:+	0.069	202300:19:+	0.262	191353:94:-
499907:7:-	0.065	191353:94:-	0.392	
255086:127:+	0.064	191353:117:-	0.113	
375622:123:-	0.064	191272:62:+	0.178	
354300:398:-	0.061			
159639:64:-	0.061			
255086:23:+	0.058			
255086:75:+	0.058			
306530:101:-	0.054			
354300:30:-	0.051			
3331:118:+	0.050			
197732:68:+	-0.064			
197732:41:+	-0.068			
430039:59:+	-0.098			

Half broomcorn

SNP 191353:132:- Chromosome 4 Ankyrin repeat family protein **GWAS**

SNP 191353:117:- Chromosome 4 Ankyrin repeat family protein **GWAS & Lasso**

SNP 191353:94:- Chromosome 4 Ankyrin repeat family protein **GWAS & Lasso**

SNP 191272:62:+ Chromosome 4 microtubule associated protein, putative, expressed **Lasso**

SNP 202300:19:+ Chromosome 4 ---------- **Lasso**

SNP 350017:27:- Chromosome 7 ---------- **Lasso**

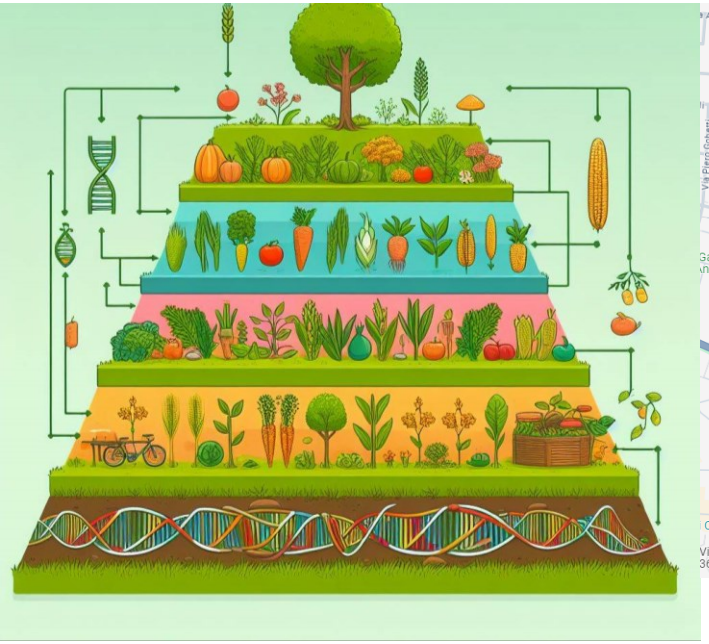
SNP 402529:102:- Chromosome 8 ---------- **Lasso**

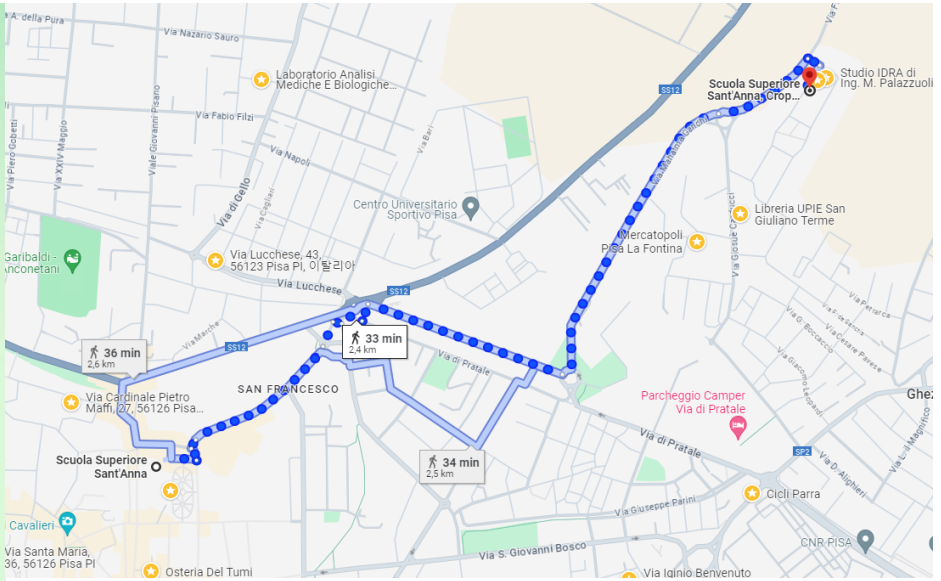
What next?

- More sorghum genetic and phenotypic characteristics need to be explored.
- Increase results reliability by using bootstrap or…
- Calculate Genetic prediction model: ridge regression BLUP.
- Analyzing environment and social effect for variety selection

CENTER OF PLANT SCIENCES

Sant'Anna
Scuola Universitaria Superiore Pisa





Email

tobias.recha@santannapisa.it
geon.kang@santannapisa.it

Adress

Via L. Alamanni, 20, 56010 Ghezzano PI