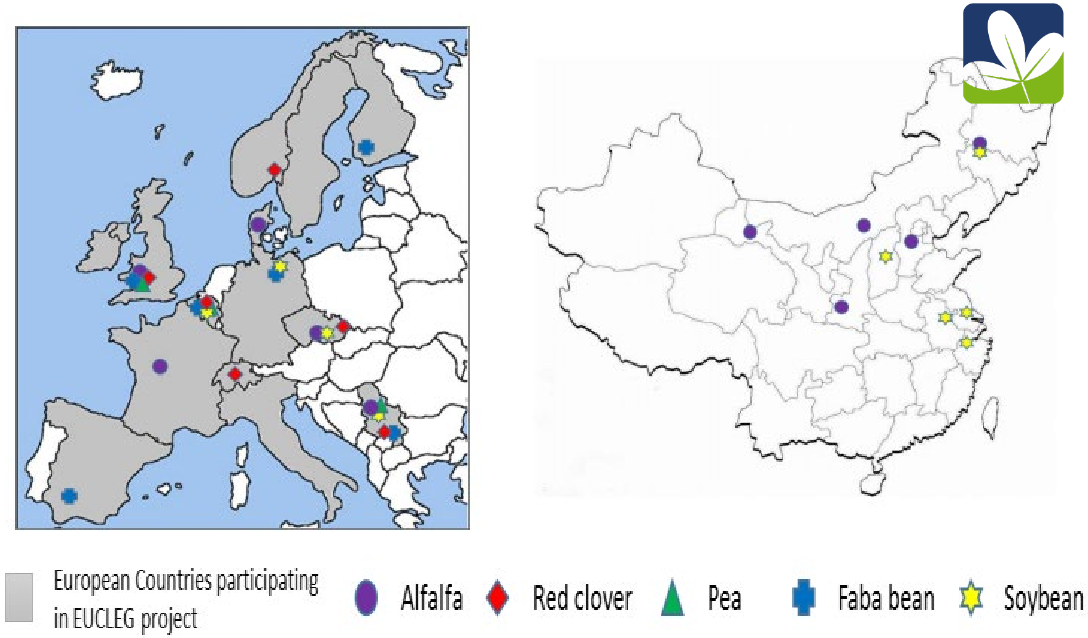
Genomic prediction for phenology, forage yield and quality in alfalfa

Marie Pégard, Julien Leuenberger, Bernadette Julier and Philippe Barre

North American Alfalfa Improvement Conference – June 7th to 9th 2022

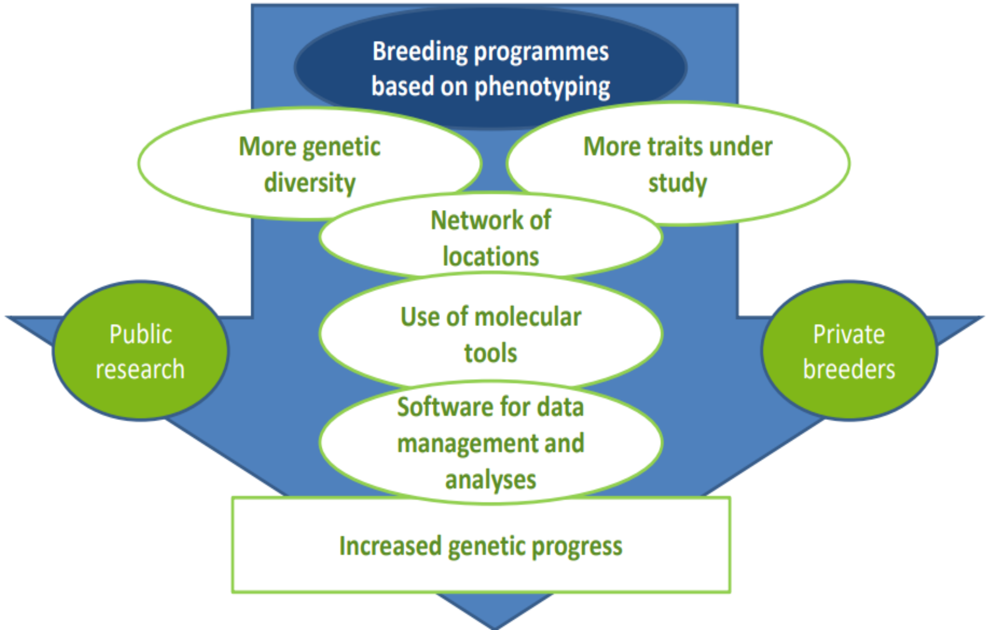
EUCLEG

Breeding forage and grain legumes to increase EU’s and China’s protein self-sufficiency



- 26 partners in 13 EU countries, and 12 partners in China
- **Overall objective:** To develop efficient breeding strategies for the legume crops and improve:
  - diversification of crops
  - crop productivity
  - yield stability
  - protein content

Genetics as a major lever



Breeding programmes based on phenotyping

More genetic diversity

More traits under study

Network of locations

Use of molecular tools

Software for data management and analyses

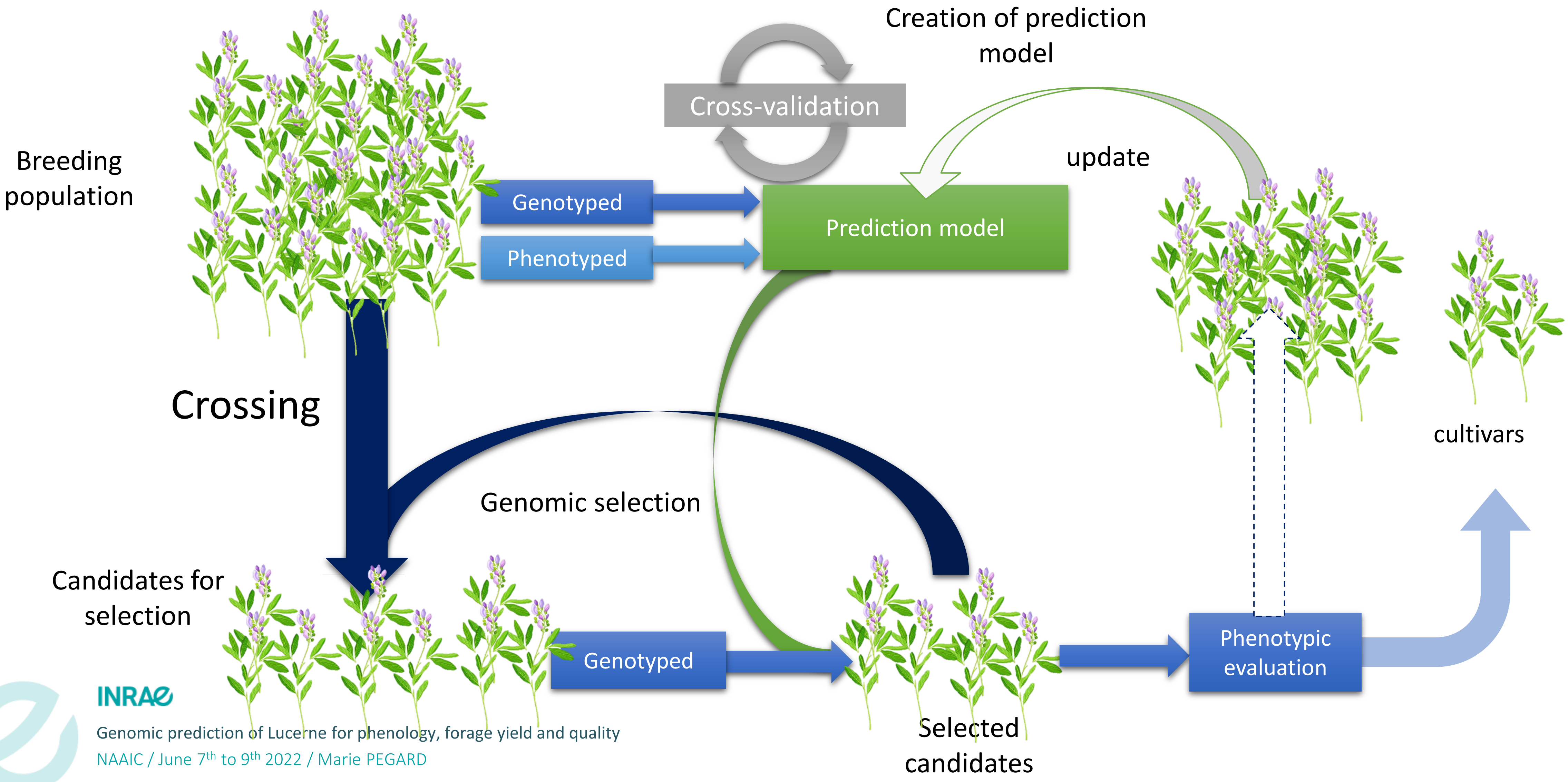
Public research

Private breeders

Increased genetic progress

Genomique Selection (GS)

GS is the direct descendant of marker-assisted selection (SAM)



Aim of the study

To evaluate the ability of genomic prediction models to predict phenology, yield and quality traits

Different parameters tested :

- The size of the training population
- Model

Previous work on Alfalfa

Found a low or moderate accuracies (0-0.65) for yield and quality traits:

- A number of SNPs from 8K to 44K
- A number of individuals between 75 and 274

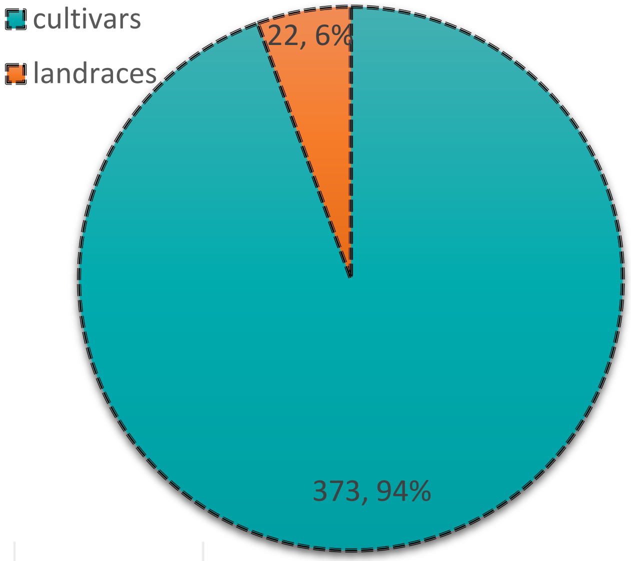
Annicchiarico *et al.*, 2015
Biazzi *et al.*, 2017
Jia *et al.*, 2018
Li *et al.*, 2015

What is new in our study?

- a larger population covering a large diversity, fall dormancy 2-8
- more SNPs
- the information from QTLs in the model prediction.

Material and methods

Research population

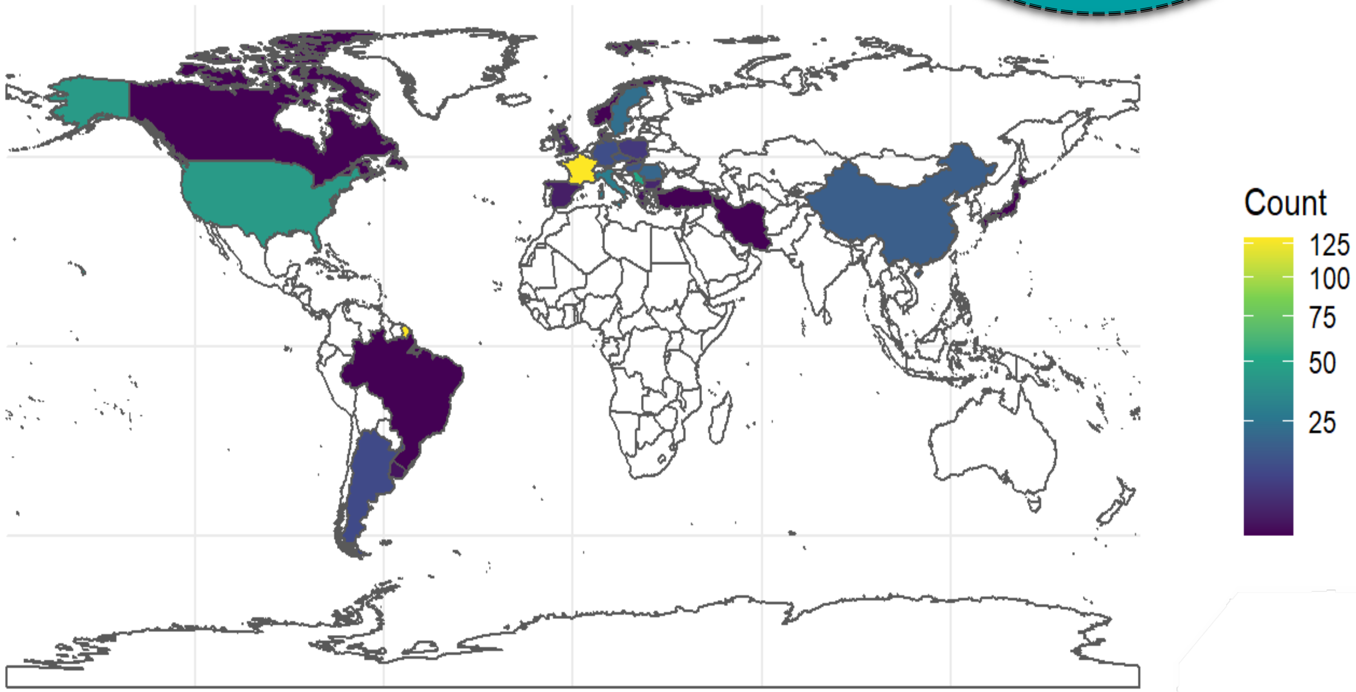


395 accessions comprised of 373 cultivars and 22 landraces

One accession consists of a population

Origins (landraces : collected), cultivars (registered) :

- Europe : 313
- North America : 45
- South America : 16
- China : 17
- Middle East : 3
- Japan : 1



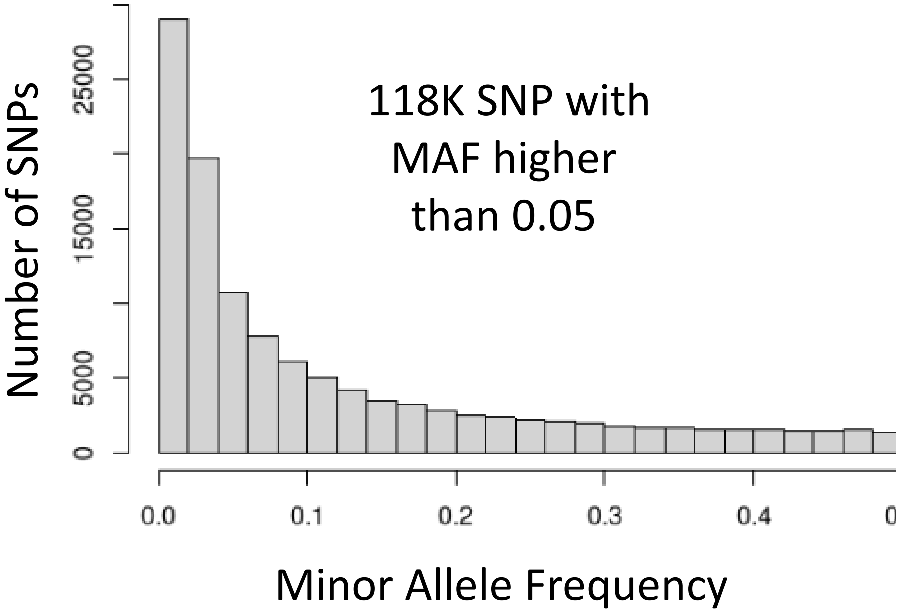
Genotyping

395 accessions genotyped – 227K SNP with 0.57% of missing values

92.5 % of the SNPs are on the chromosomes

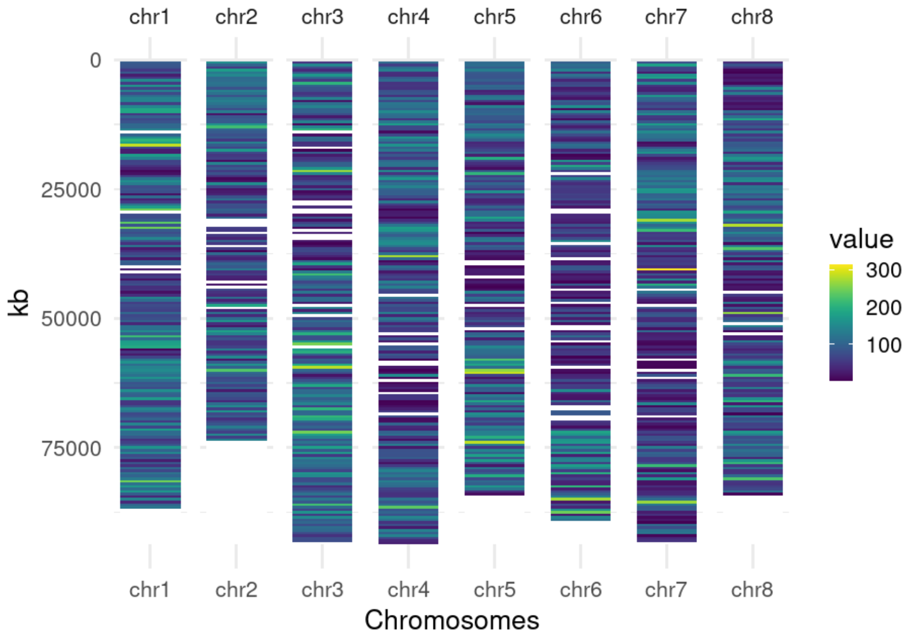
One accession consists of a population for which allele frequencies are estimated

Allelic frequency distribution

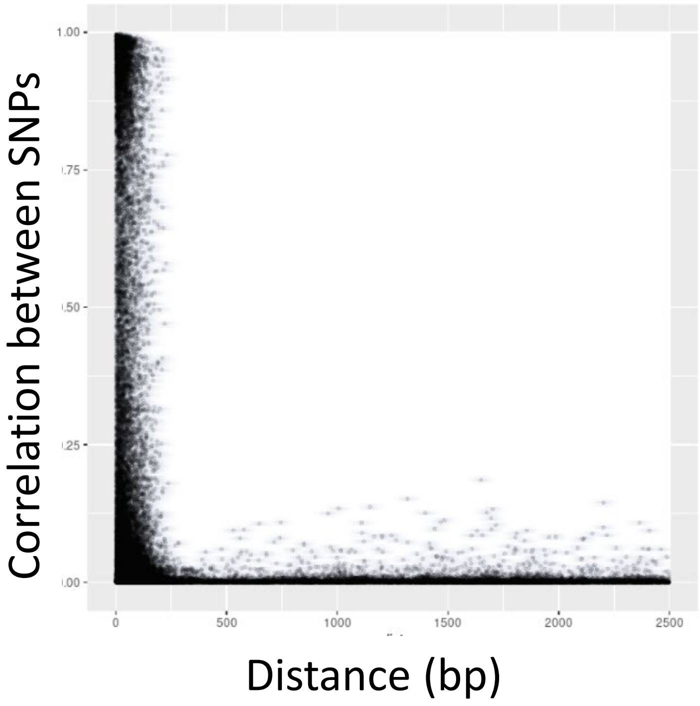


118K SNP with MAF higher than 0.05

Marker density over the genome (500 kb window size)



Very low LD even at short distances



Reference sequence : Chen *et al.* 2020

Phenotyping

Lignin + cellulose

Many traits phenotyped (187) : dry matter yield, ligno-cellulose content (acid detergent fiber ADF), protein content , flowering date and autumn dormancy

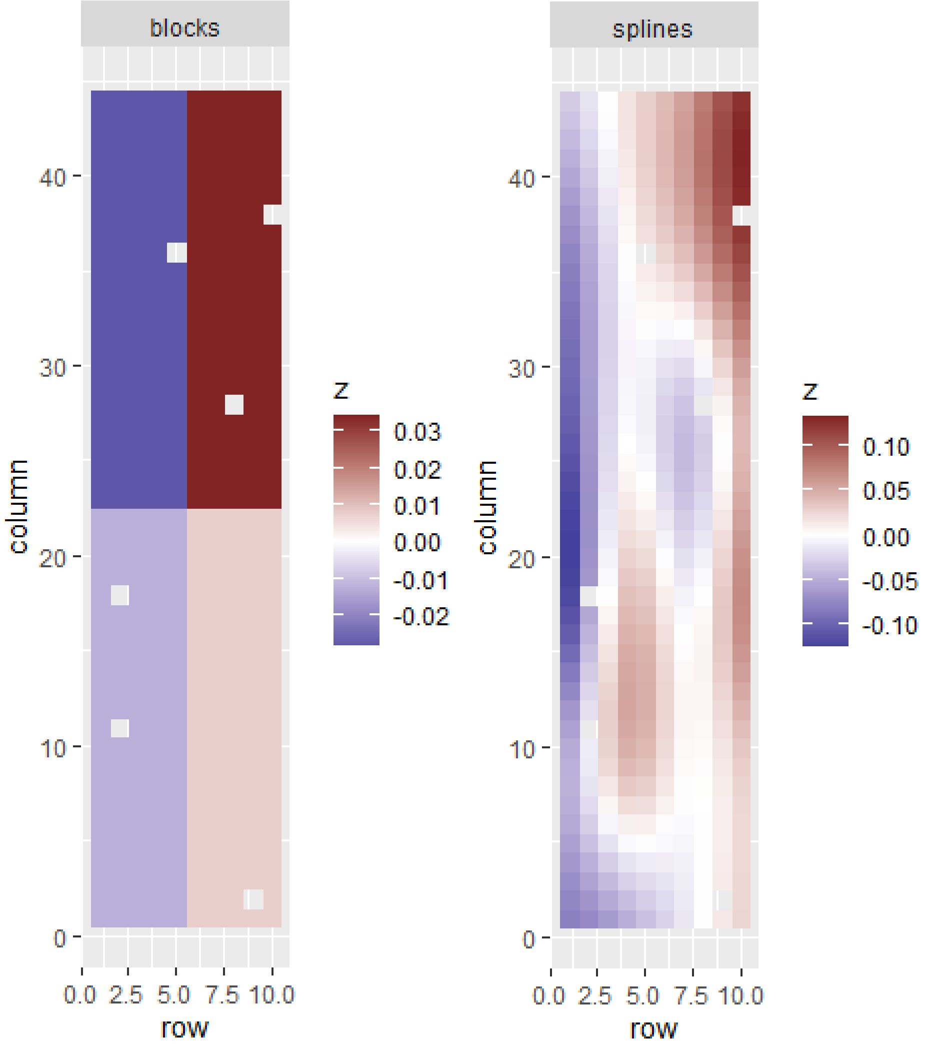
Two trials established in 2018 :

- Lusignan – FRA (382 accessions)
- Novi Sad – SRB (395 accessions)

Scored during two years (2019 and 2020),4-5 cuts/year.

Phenotypes were adjusted :

- For the micro-environmental variation inside each trial by a bi-splines model
- For the year and the location effect.



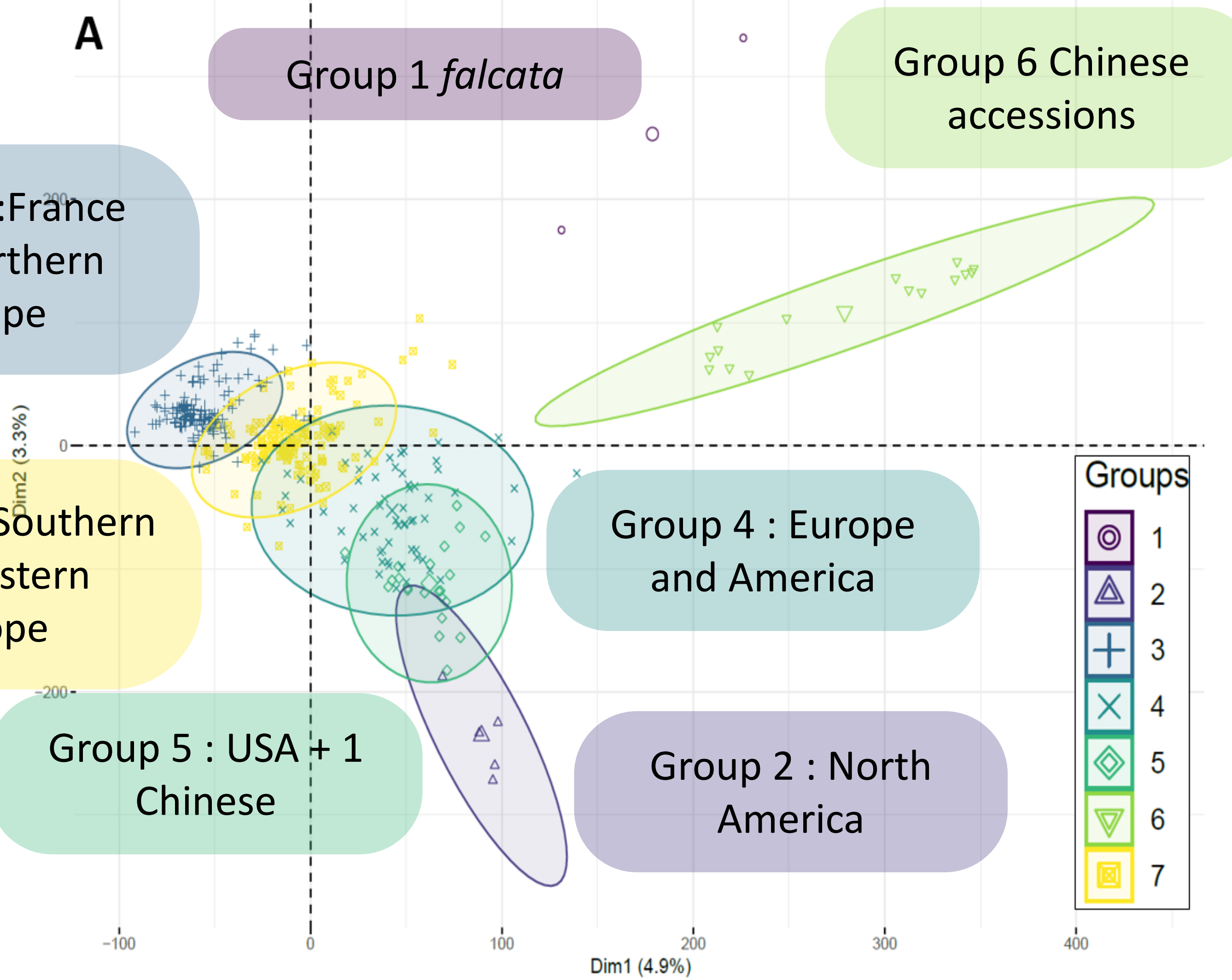
Genetic diversity

PCA : ellipse based on the clustering

- Two groups were clearly separated from the others
- The five other groups showed a genetic continuum
- The European accessions as well as the American accessions were splitted into several groups.
- Fst values are really small
- The group 6 and 2 are the most distinct ones

FST values

Clusters	1	2	3	4	5	6
1						
2						
3		0.017				
4		0.013	0.003			
5		0.004	0.006	0.002		
6		0.026	0.019	0.012	0.015	
7		0.016	0.001	0.001	0.005	0.015



Group 1 *falcata*

Group 6 Chinese accessions

Group 3 :France and Northern Europe

Group 7 : Southern and Eastern Europe

Group 4 : Europe and America

Group 5 : USA + 1 Chinese

Group 2 : North America

Statiscal Methods & Results

Genome Wide Association Study (GWAS)

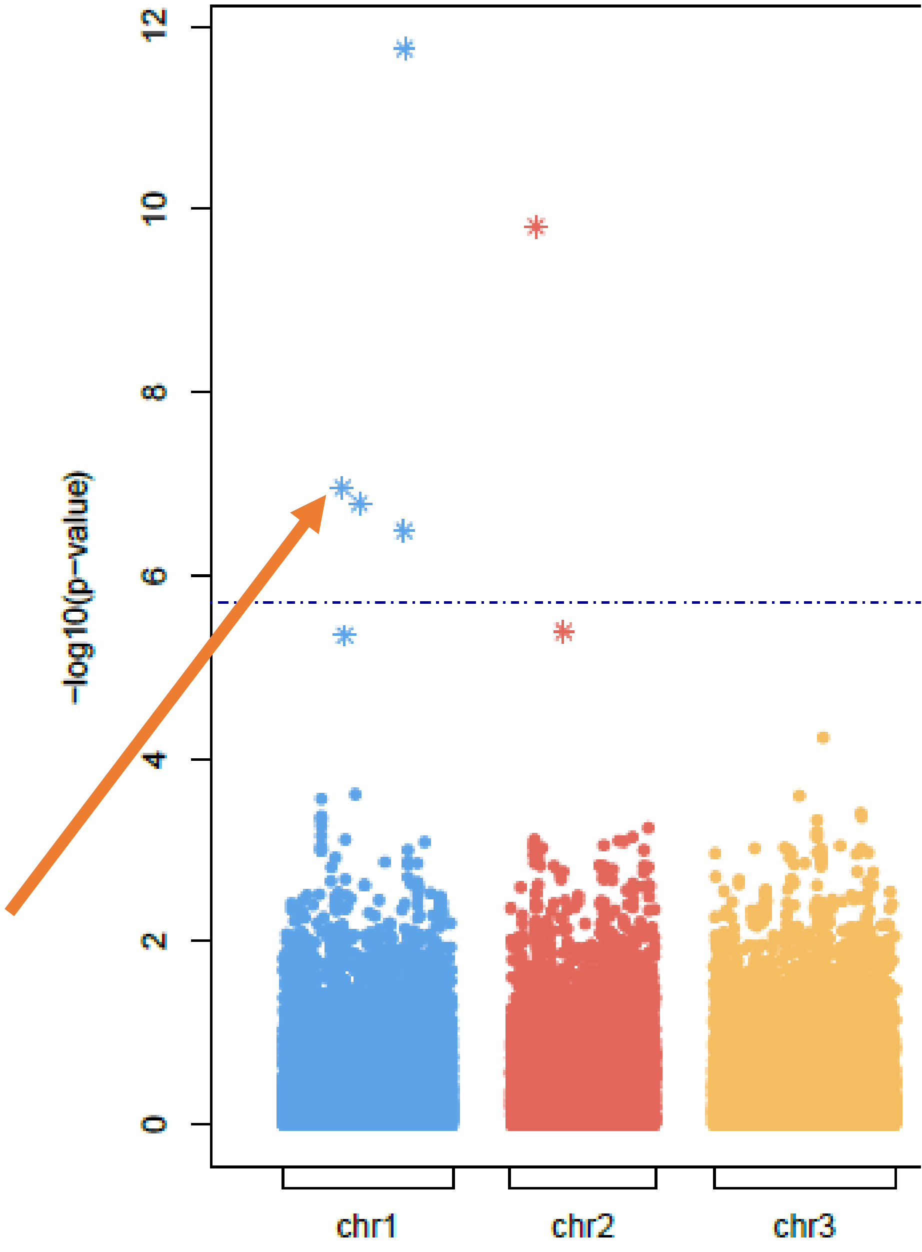
GWAS aims at determining the association between the phenotype and genotype

Some parameters influence the GWAS results :

- Linkage disequilibrium
- Population structure

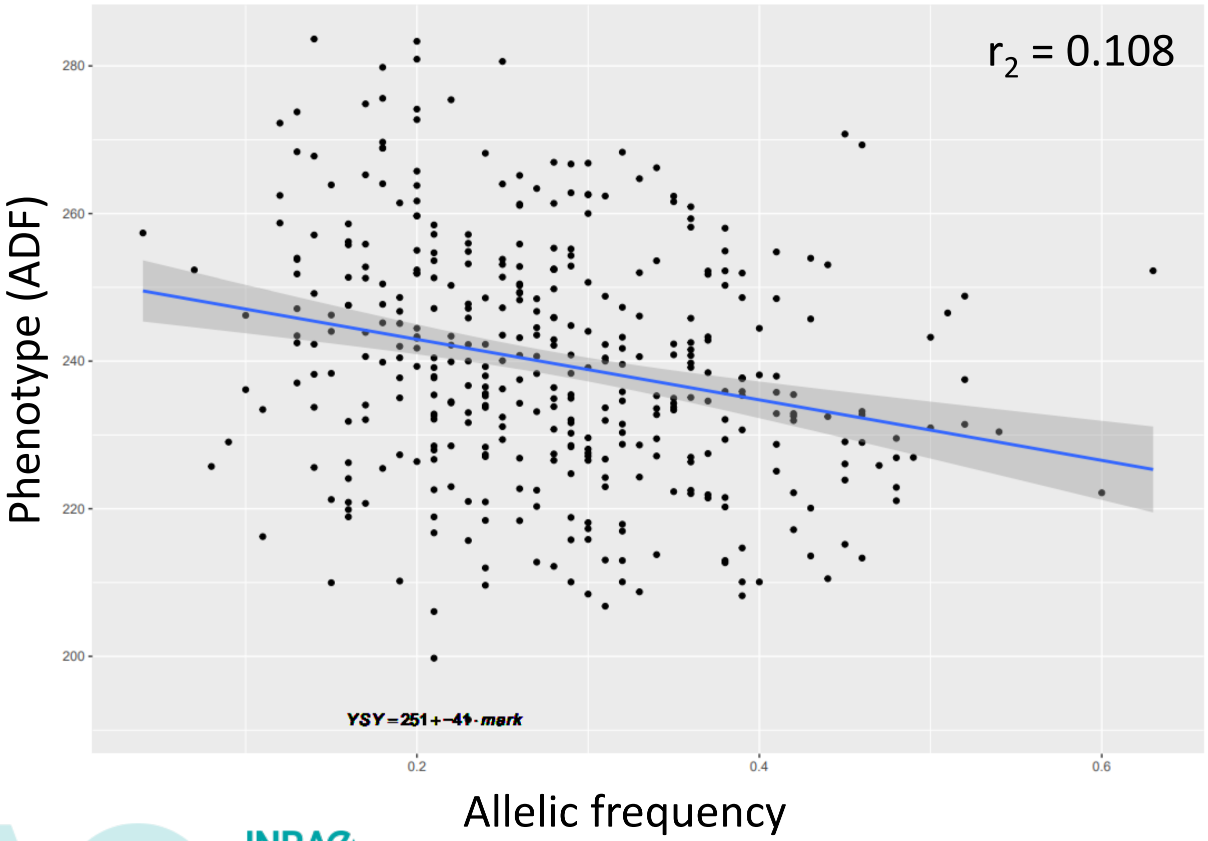
Multi Locus Mixed Model (MLMM) :

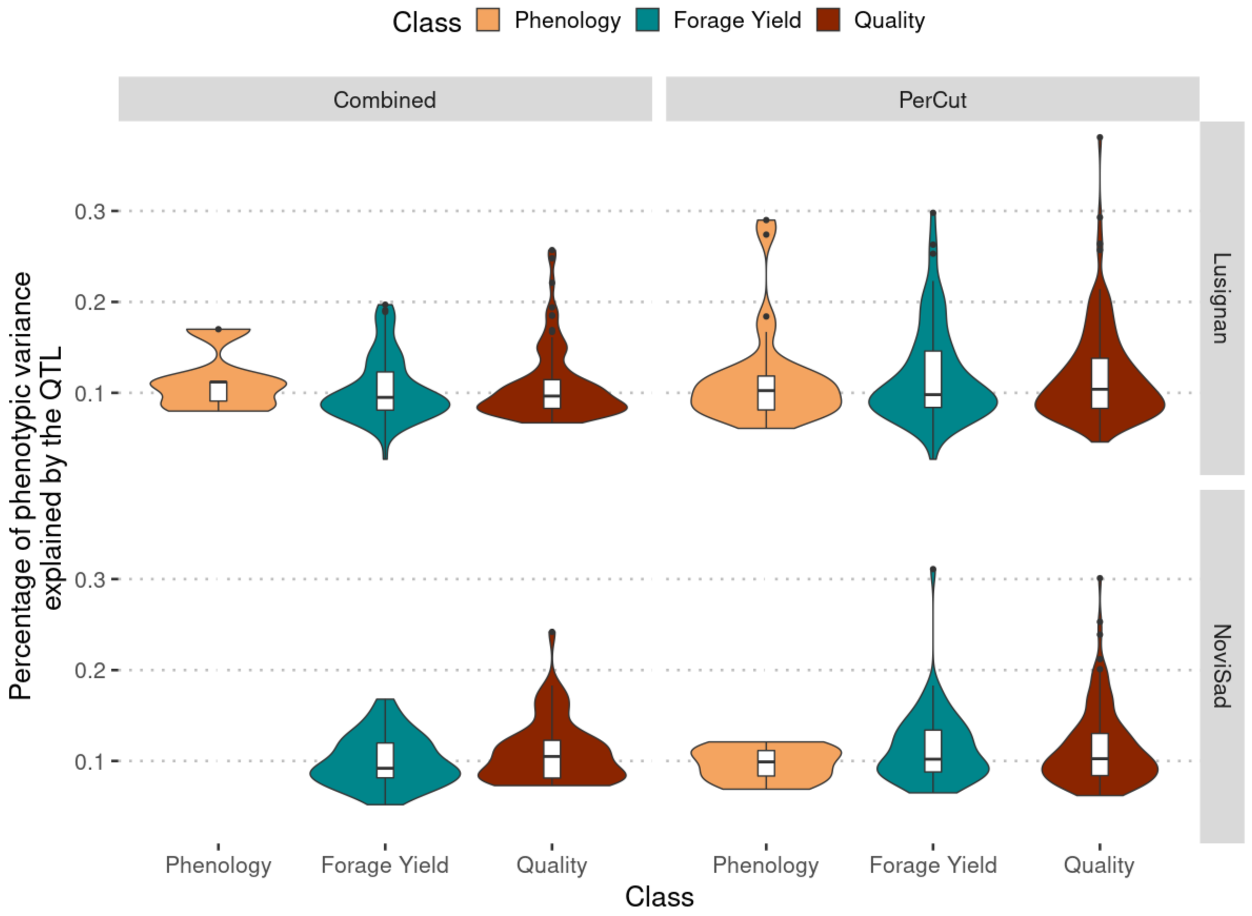
- Includes the genomic relationship matrix to take into account the genetic structure in the population
- Stepwise model regression with a forward inclusion of the QTLs as co-factors and a backward elimination.



GWAS -results

Number of QTLs and the percentage of phenotypic variance explained





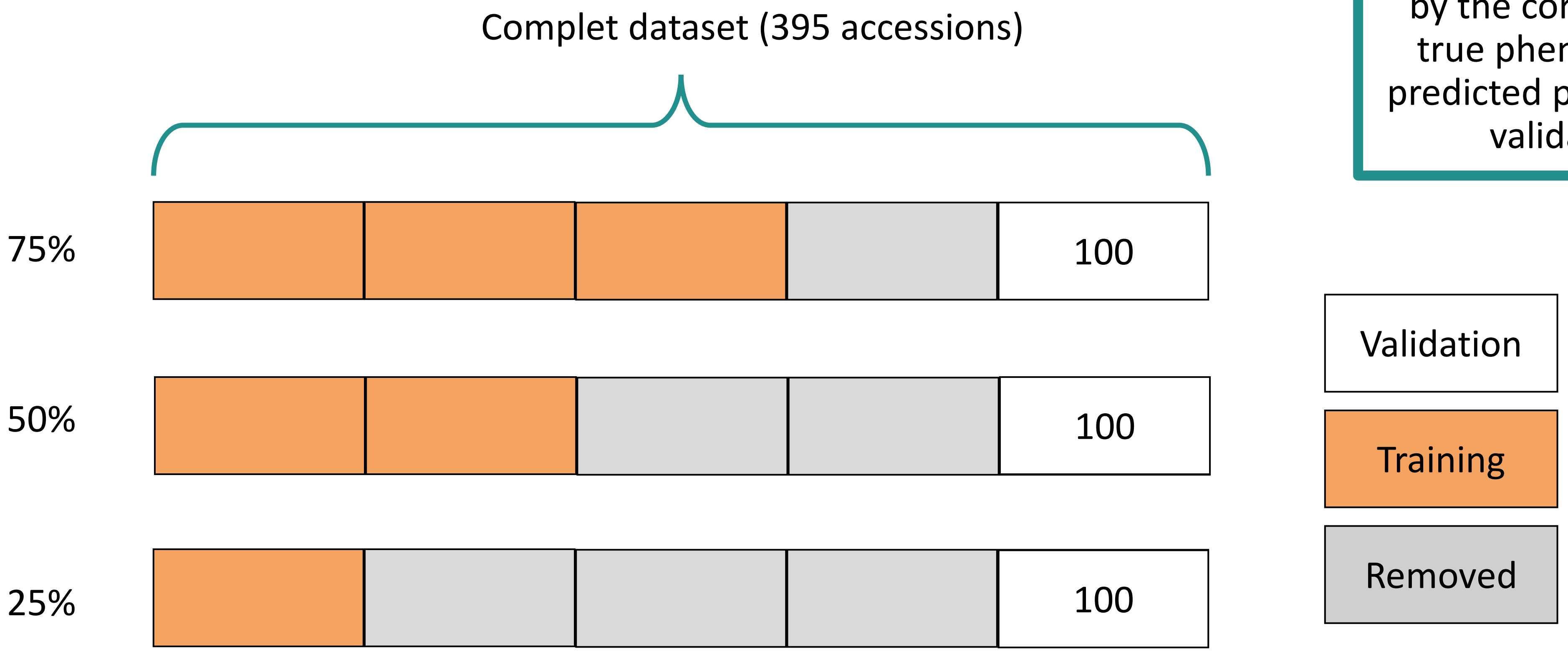
Values for both locations and all traits together (combined or per cut)

	Nb of traits	Nb of QTLs
Phenology	16	79
Forage yield	37	438
Quality	55	564

Genomic selection (GS)

Test of the training/validation population size

We tested the effect of different Training population sizes :



The predicting ability is evaluated by the correlation between the true phenotypic value and the predicted phenotypic value of the validation population

GS - results

Test of the training/validation population size

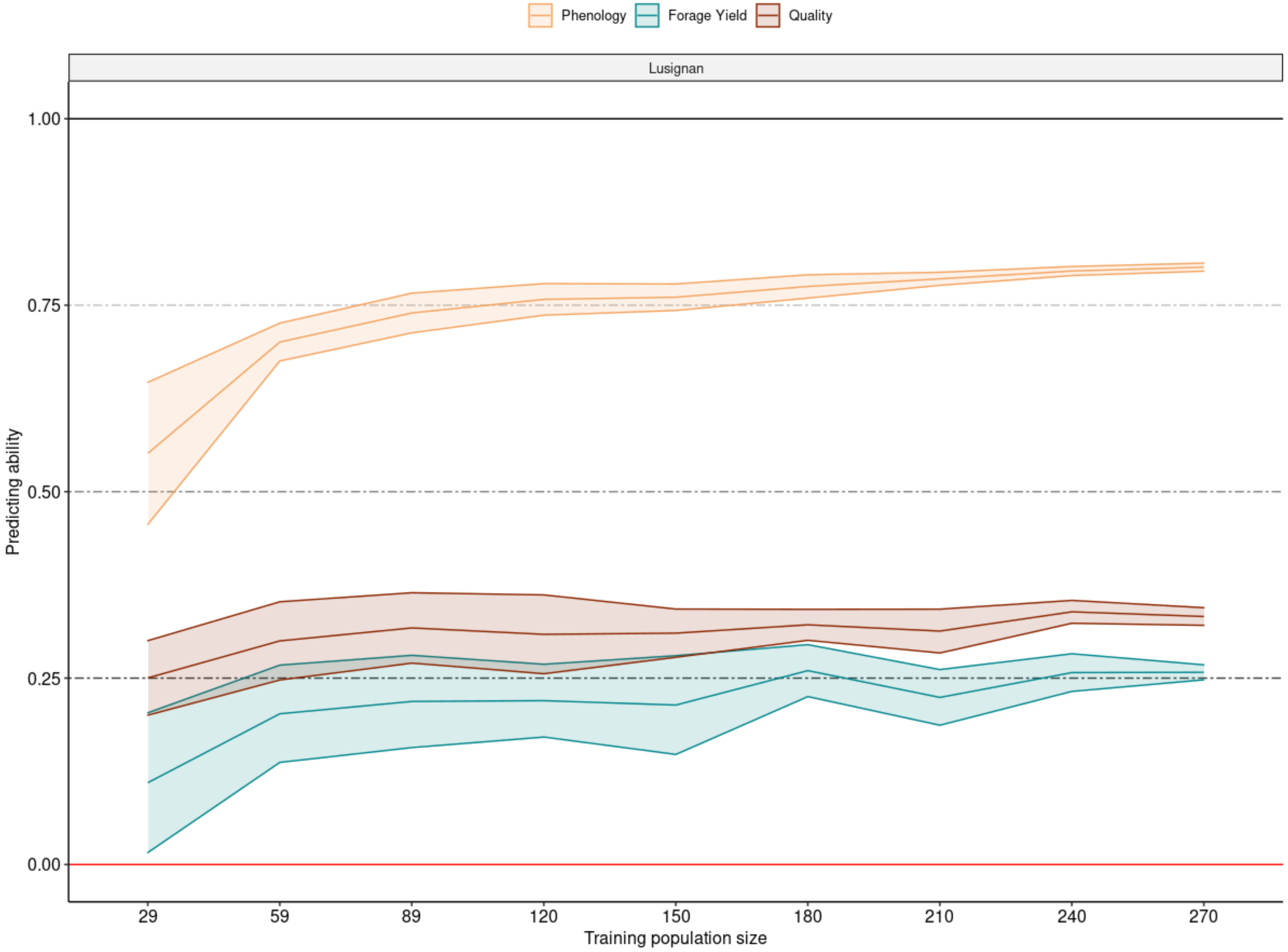
When the number of accessions in the training population increases:

- The predicting ability increases
- The variability between repetitions decreases

Phenology > Quality > Forage Yield

In the next steps we used Training with 270 individuals

Model GBLUP for three independent traits belonging to the classes



GS

Models

Four models :

- With the relationship matrix only → GBLUP
- With the QTLs as fixed effects and the relationship matrix only → QGBLUP
- Ridge regression → RR
- Bayesian Lasso → Lasso

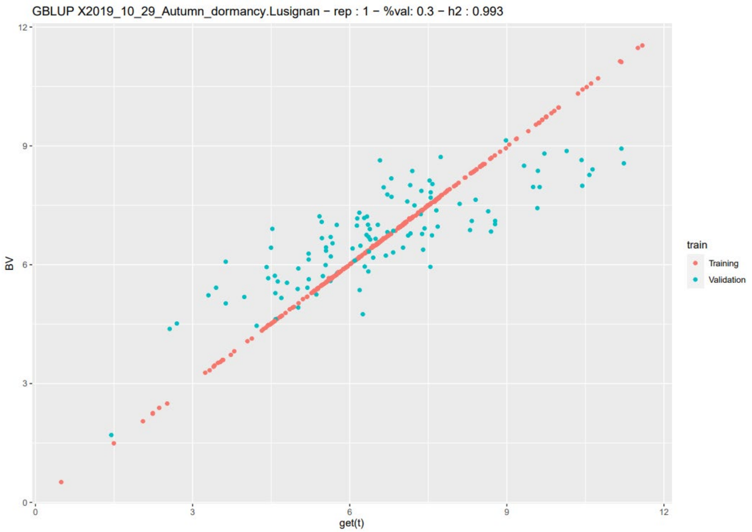
To avoid overfit, the QTLs detection by GWAS is performed at each repetition on the training population

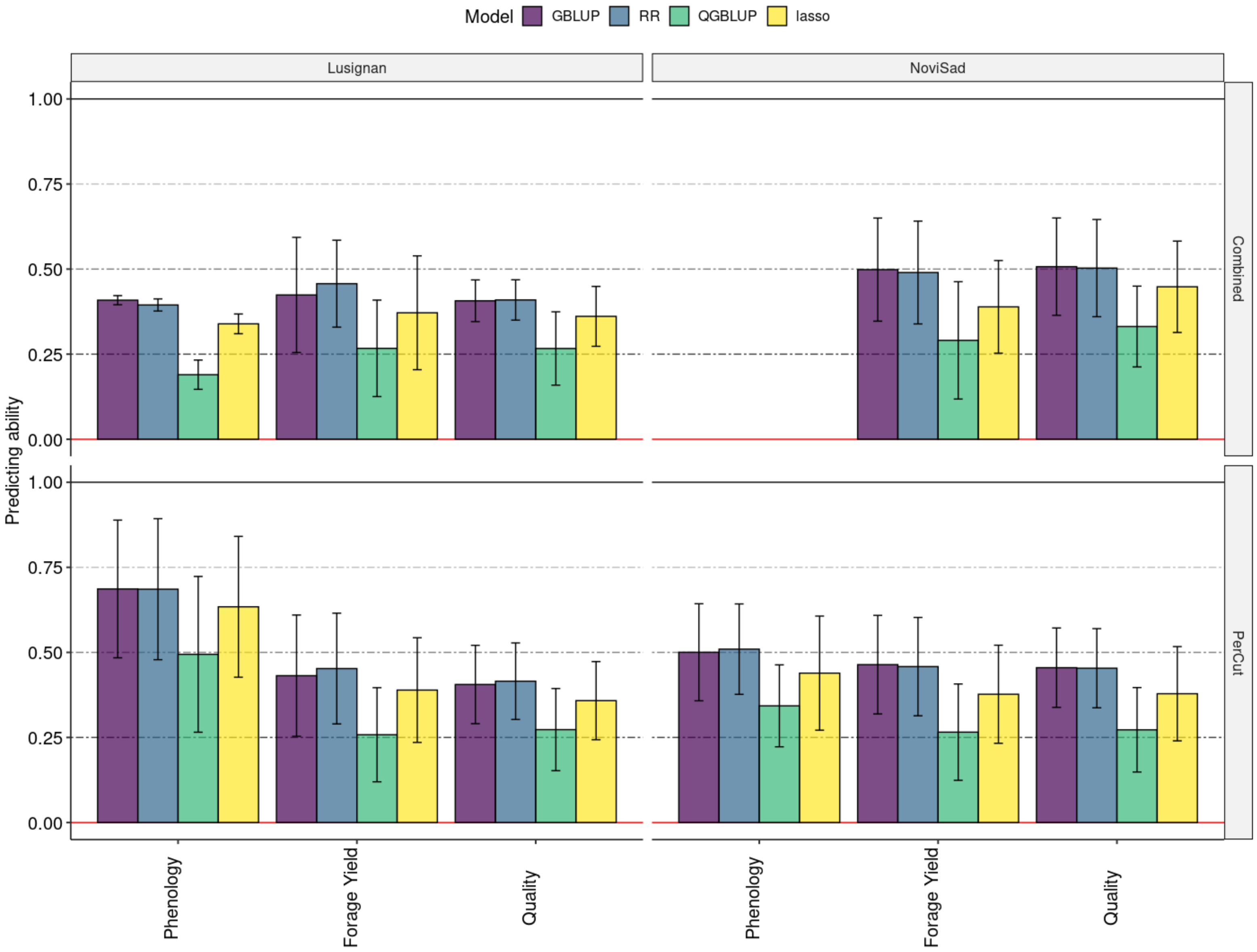
GS - results

High predicting ability (> 0.7) depending on the traits

- GBLUP and Ridge Regression are equivalent
- QGBLUP and BayesLasso are less performant in most cases

Phenology > Quality > Forage Yield





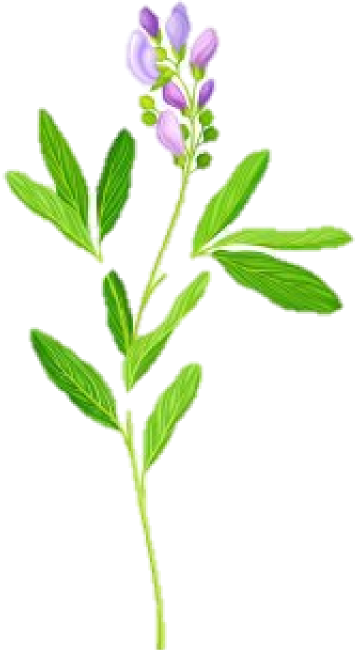
Why the models without QTLs performed better ?

- Some false positive QTLs remain

→ higher threshold

- Not enough accessions so the QTL effect is over-estimated

→ More data will be add with future projects



Conclusion

Take home message

1. The infinitesimals methods showed higher quality of prediction
2. The models integrating the QTLs information (QGBLUP) or based on feature selection were less efficient
3. The traits linked to :
    1. phenology tended to be better-predicted (0.62 ± 0.17) than
    2. forage yield (0.45 ± 0.19) and quality (0.43 ± 0.14).
4. Some QTLs explaining a high percentage of phenotypic variation were found, and represents good candidates for further investigation
    1. Some QTLs match for several traits or in less than 1000 bp between locations

With a population with a larger genetic diversity we have reached higher predicting abilities than in the previous studies in alfalfa.

Thanks for your attention !