Genomics assisted breeding in alfalfa

Bernadette Julier

EUCLEG project “Breeding forage and grain legumes to increase EU’s and China’s protein self-sufficiency” (www.eucleg.eu).

The application of genomic technologies in the breeding of legume species

Techical booklet based on the EUCLEG online workshop on held on the 30th September and 1st October 2021

Thank you to the organisers, contributors, and sponsors of this event

Contributors: Bernadette Julier; Isabel Roldán-Ruiz; David Lloyd; Hilde Muylle; Radu Grumeza; Ana Maria Torres; Leif Skot; Roland Kölliker.

Editors: Catherine Howarth, Sarah Clarke, Aberystwyth University

Sponsors: European Union’s Horizon 2020 Programme for Research & Innovation under grant agreement n°727312.

This project has received funding from the European Union’s Horizon 2020 Programme for Research & Innovation under grant agreement n°727312.

EUCLEG.eu

The application of genomic technologies in the breeding of legume species

EUCLEG project "Breeding forage and grain legumes to increase EU's and China's protein self-sufficiency" (www.eucleg.eu).

Contents

This project has received funding from the European Union's Horizon 2020 Programme for Research & Innovation under grant agreement n°727312.

9. Genomics assisted breeding in alfalfa

Bernadette Julier

Research Director at INRAE, Unité de Recherche Pluridisciplinaire Prairies et Plantes Fourragères (URP3F), Lusignan, France

EUCLEG alfalfa species leader

This project has received funding from the European Union's Horizon 2020 Programme for Research & Innovation under grant agreement n°727312.

Alfalfa - Lucerne

A major legume species

- Highest protein production/ha in temperature climates
- Drought tolerant
- Protein/energy
- Ruminant health
- Positive effects in the rotation

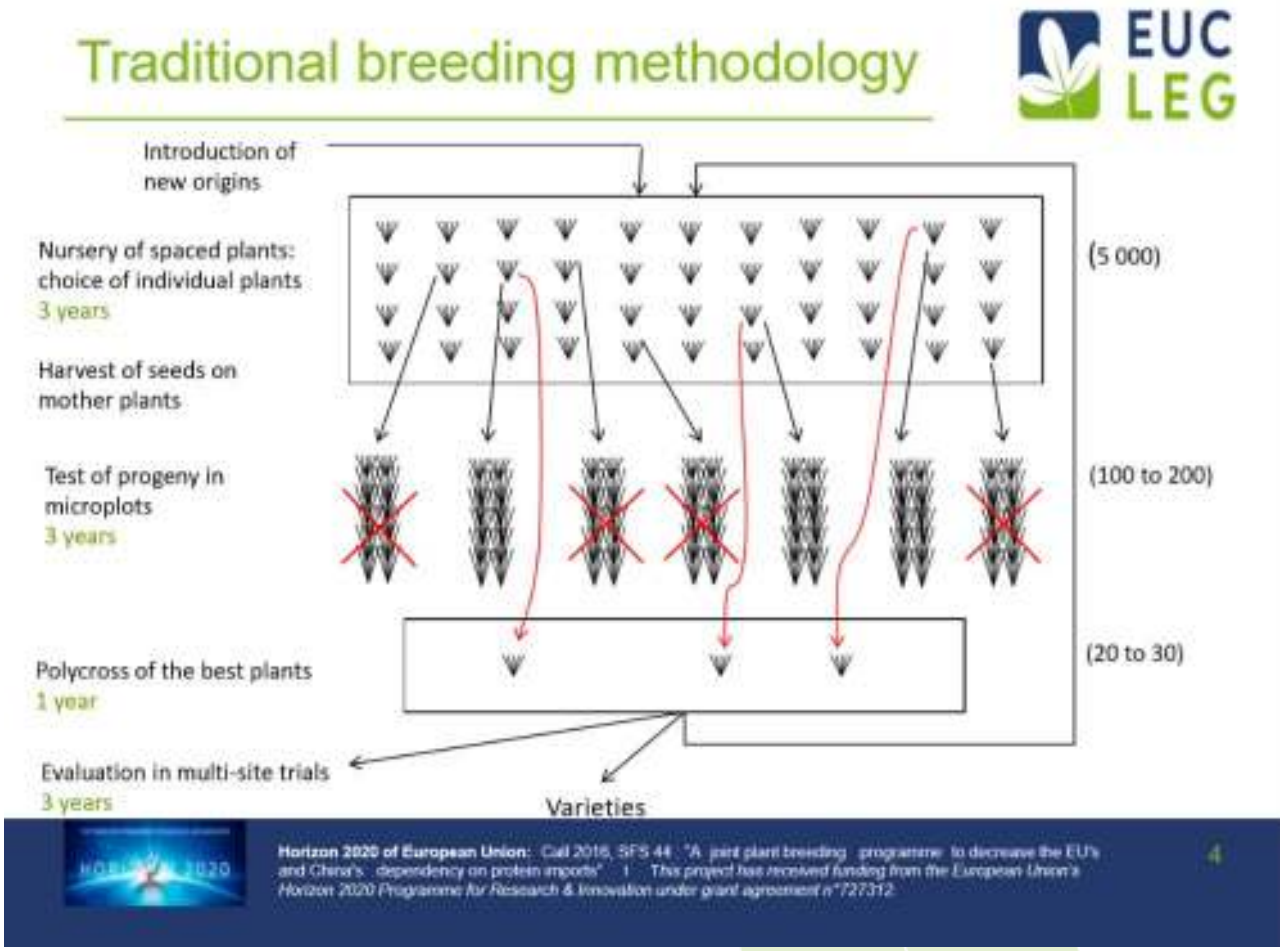
Allogamous reproduction, synthetic varieties

2n = 4x = 32

Julier et al. 2017, CABI Publishing

Alfalfa, or lucerne, is a major legume species that gives the highest protein production per acreage among all legume species in temperate regions. It is quite drought tolerant; it has a convenient protein/energy ratio. It provides some advantages to ruminant health, and it has positive effects in the rotation. It is an allogamous species and the varieties are synthetic populations. In addition to that, it is an autotetraploid species with 32 chromosomes.

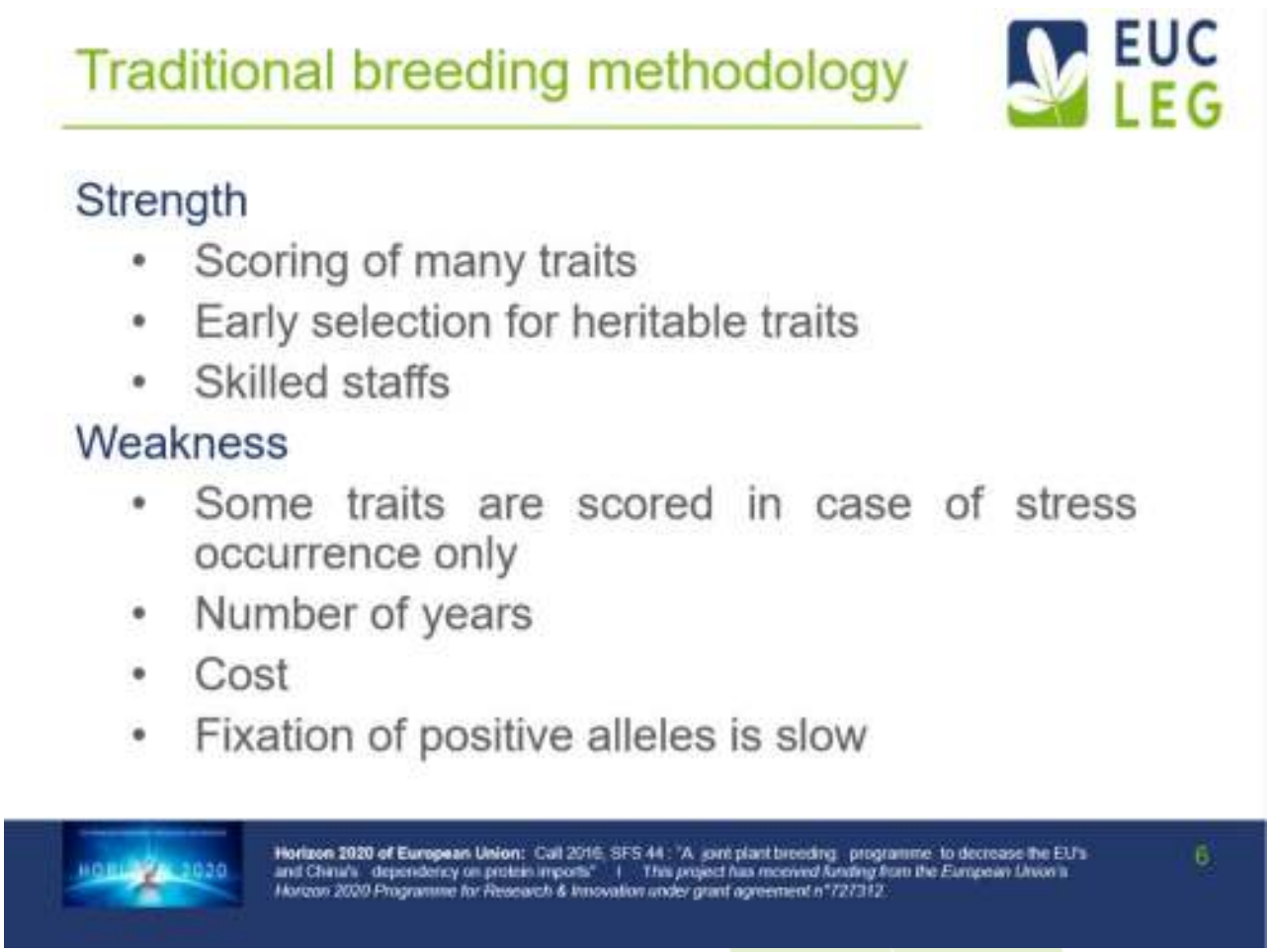
This project has received funding from the European Union's Horizon 2020 Programme for Research & Innovation under grant agreement n°727312.



What about traditional breeding methodology? It is based on the evaluation of phenotypic traits of course; the first step of selection takes place in a nursery of spaced plants and the second step is applied in progeny testing.

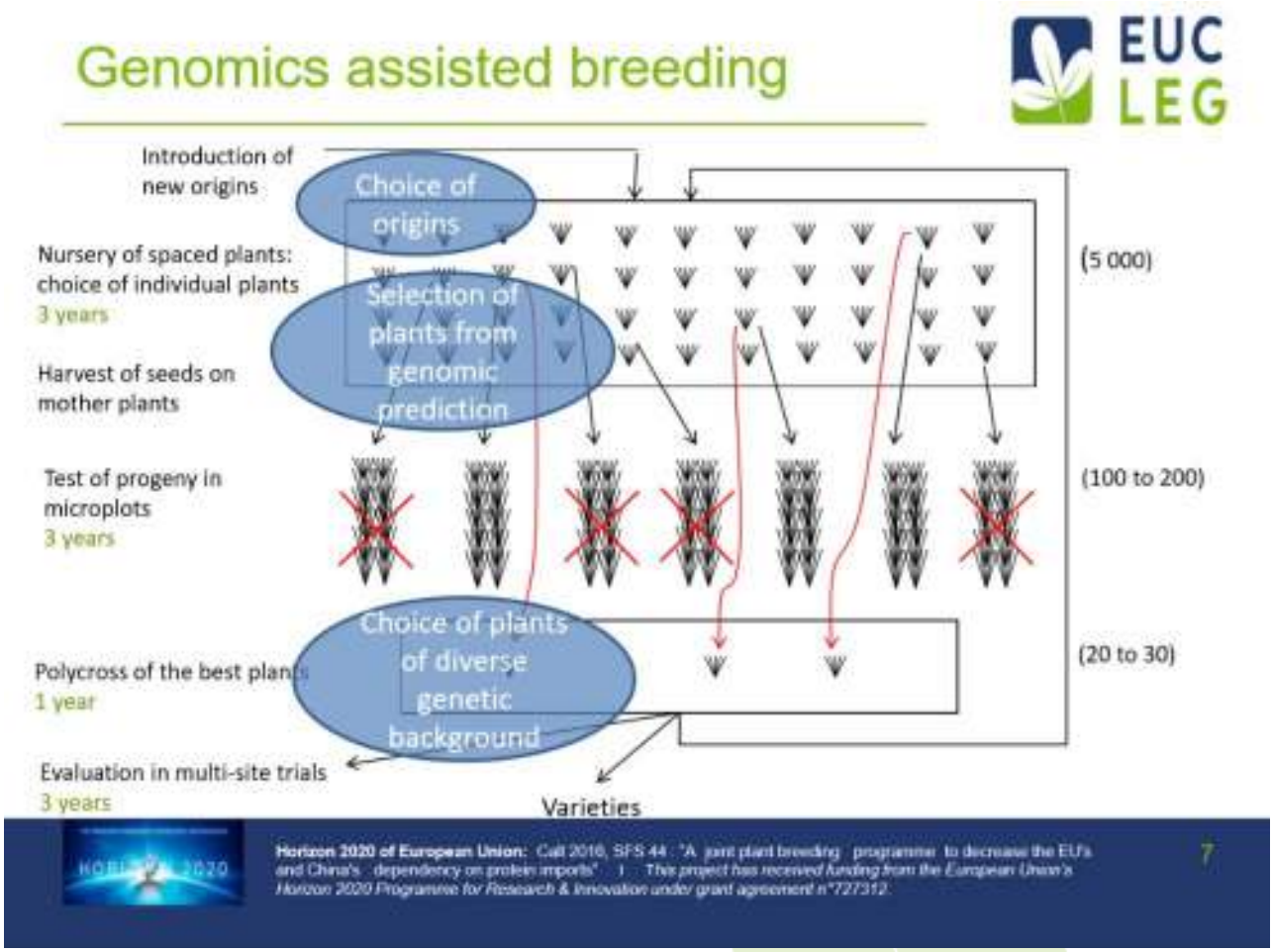
In your breeding pool, you may have introduced new origins to expand genetic diversity. You study this breeding pool in a nursery of spaced plants, where you can choose individual plants on their value for heritable traits. The spaced plant nursery is studied for 3 years, and you can study about 5000 plants or more. At the end of the 3 years, you harvest the seeds from the selected mother plant and these progenies are tested in a micro plot design for 3 more years. Here you can have from 100 to 200 progeny testing and depending on the value of the progeny, you go back to the mother plant and polycross the best plants during the following year. You can have 20 to 30 polycross a year. You then study the progeny in multi-site trials for 3 years and the best polycross goes to a variety registration test. These progenies are also the basis of a new cycle of recurrent selection. This is a theoretical breeding scheme and quite often, the mother plants no longer exist when you have the result of the progeny test. In that case, the polycross of the best plant is based on plants or seeds collected in the progeny test. As a consequence, you lose part of the genetic progress.

This project has received funding from the European Union's Horizon 2020 Programme for Research & Innovation under grant agreement n°727312.



What are the strengths and the weaknesses of this methodology? The strengths include that you can score many of the traits, you can have early selection for heritable traits in the nursery, and for most breeding companies the staff are already skilled to be able to do this work. The weaknesses include that you cannot score stress tolerance if this stress doesn't occur every year, so you need to test this in controlled conditions. All this evaluation, in nursery or controlled conditions require several years to carry out one cycle of selection, the cost is related to this number of years. In addition, the fixation of positive alleles is slow, especially for such a heterozygous and autotetraploid species.

This project has received funding from the European Union's Horizon 2020 Programme for Research & Innovation under grant agreement n°727312.



Where could genomics assist breeding? It could assist breeding in the choice of origins to be introduced into a breeding scheme. It could help with the selection of plants from genomic prediction, and it could also help with the choice of plants used in the polycross of the best plants based on the diversity of genetic background. We will discuss these three points.

Before describing genomic assisted breeding, I will provide an overview of different marker developments. Then I will explain a bit more about the management of genetic diversity, genome wide association study and genomic selection from EUCLEG results.

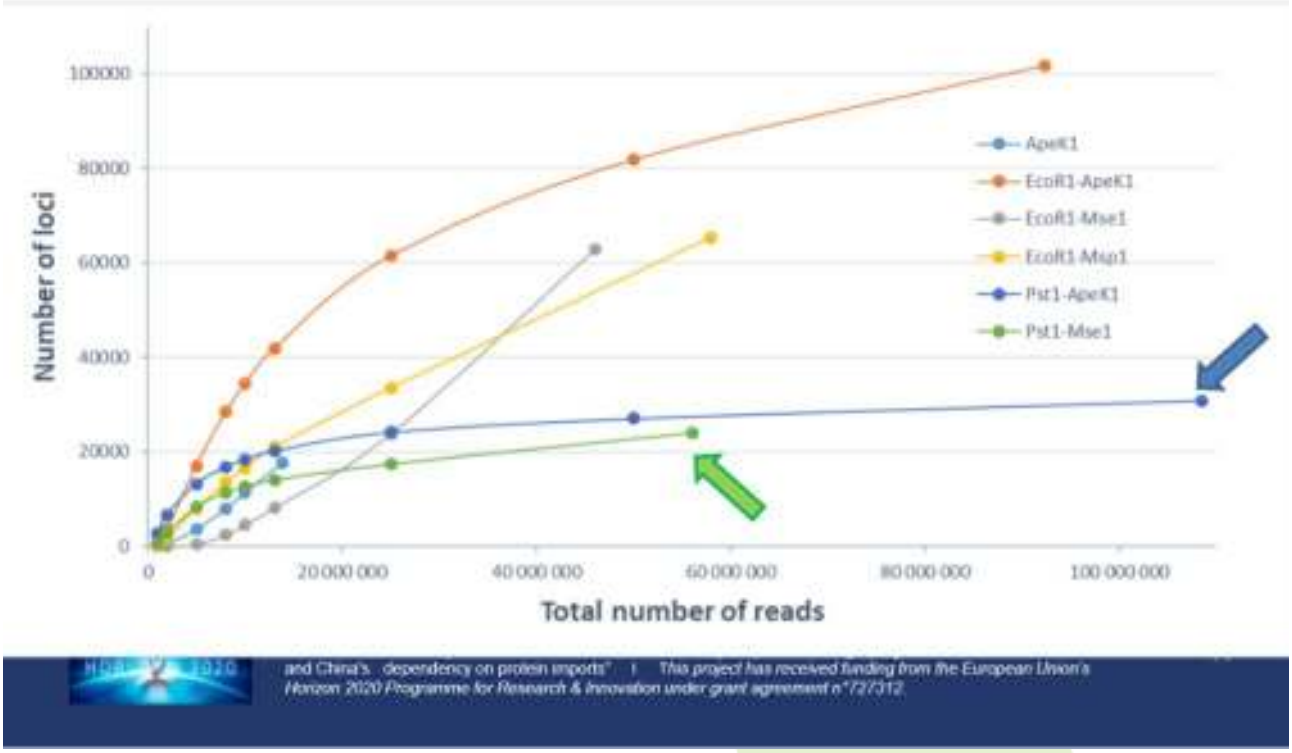
This project has received funding from the European Union's Horizon 2020 Programme for Research & Innovation under grant agreement n°727312.

What was the situation of marker development, before EUCLEG? In the past, we had low throughput markers such as SSR and AFLP. Then a 10K single nucleotide polymorphism (SNP) array was developed, but it was too expensive, especially when you want to study a population represented by at least 20 or 30 plants. The array was not that big with only 10K SNPs. More recently, genotype-by sequencing (GBS) was developed in heterozygous species, and it was interesting to see that it was quite good for these species. In most cases, we had less than 40K markers and in many cases we have seen quite a lot of missing data and this is an issue.

In EUCLEG, we have developed an improved GBS protocol, by testing different restriction enzymes to reduce missing data and thus optimize the protocol of GBS.

This project has received funding from the European Union's Horizon 2020 Programme for Research & Innovation under grant agreement n°727312.

Marker development



and China's dependency on protein imports" | This project has received funding from the European Union's Horizon 2020 Programme for Research & Innovation under grant agreement n°727312

We have tested a number of enzymes or pairs of enzymes and we have sequenced the GBS reads, or fragments. We have done some bioinformatic analysis that included mapping of the reads on the reference genome. The number of the loci is here represented as a function of the number of reads. We have seen in 2 cases, Pst1-ApeK1 in blue and Pst1-Mse1 in green, that we have a clear plateau meaning that with about 10 million reads we can achieve a stable number of loci. This means that we have less risk of missing data. We have chosen Pst1-Mse1, because it was a pair of enzymes that were already chosen for red clover, meaning we may be able to compare the markers that can be important for trait variation.

This project has received funding from the European Union's Horizon 2020 Programme for Research & Innovation under grant agreement n°727312.

Marker development

EUCLEG: an improved GBS protocol

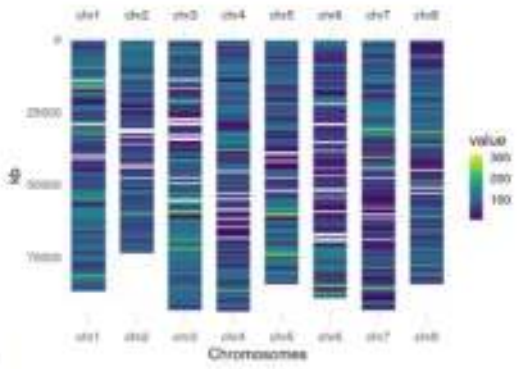
- Choice of restriction enzymes to reduce missing data
- Use of a reference genome sequence: Chen et al. 2020
- Allele frequency of each accession

After we have chosen these restriction enzymes, we have used a reference genome sequence delivered by Chen et al. in China to map the GBS reads and we have obtained allele frequency of each accession.

Marker development

On 1 061 accessions:

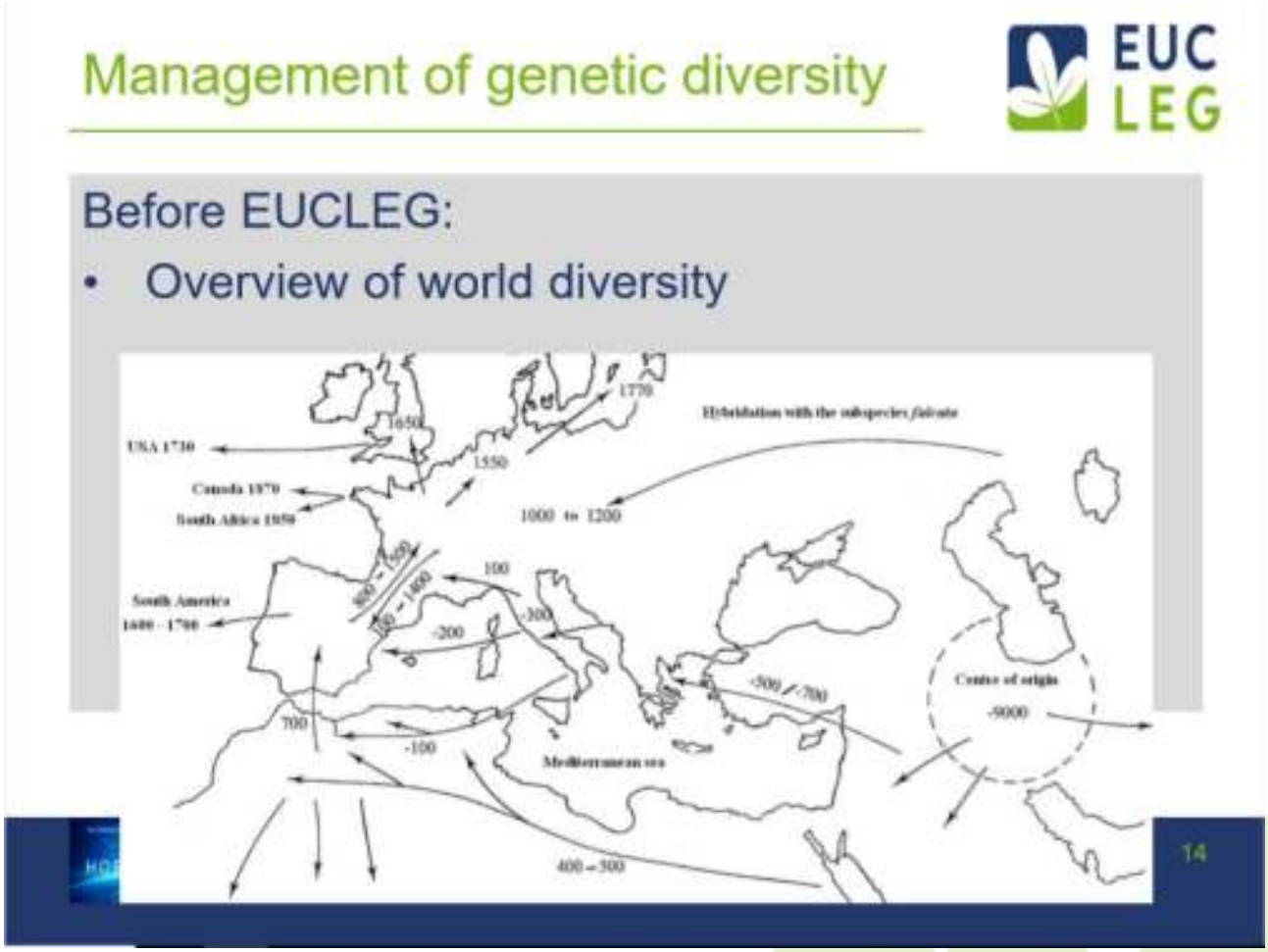
- 31 743 loci
- 228 568 SNP with less than 5% missing data per SNP
- 118 421 SNP without missing data



At the end, we were able to sequence more than 1000 accessions, we obtained more than 30 000 loci and at each locus we had several SNPs, so we obtained more than 200 000 SNPs with less than 5% missing data

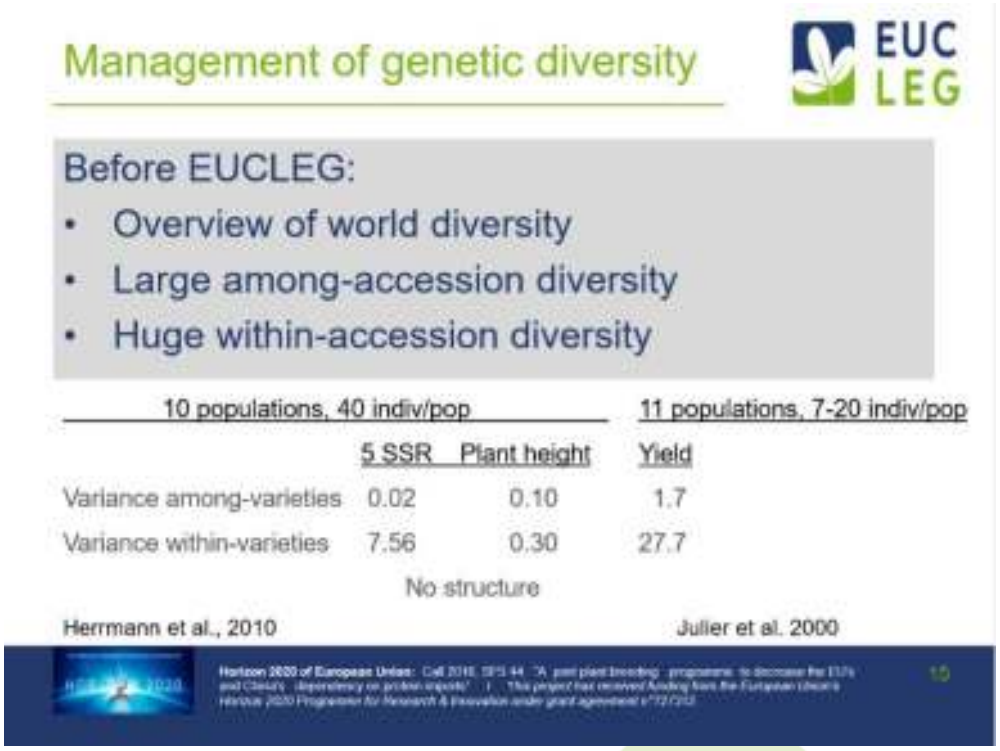
This project has received funding from the European Union's Horizon 2020 Programme for Research & Innovation under grant agreement n°727312.

per SNP: a huge number of SNP with a low ratio of missing data. If we chose to add a new threshold without any missing data, we have more than 100 000 SNPs. As shown on the right, these SNPs cover the genome very well. This genotyping tool is now available and this is great progress for alfalfa genetics.

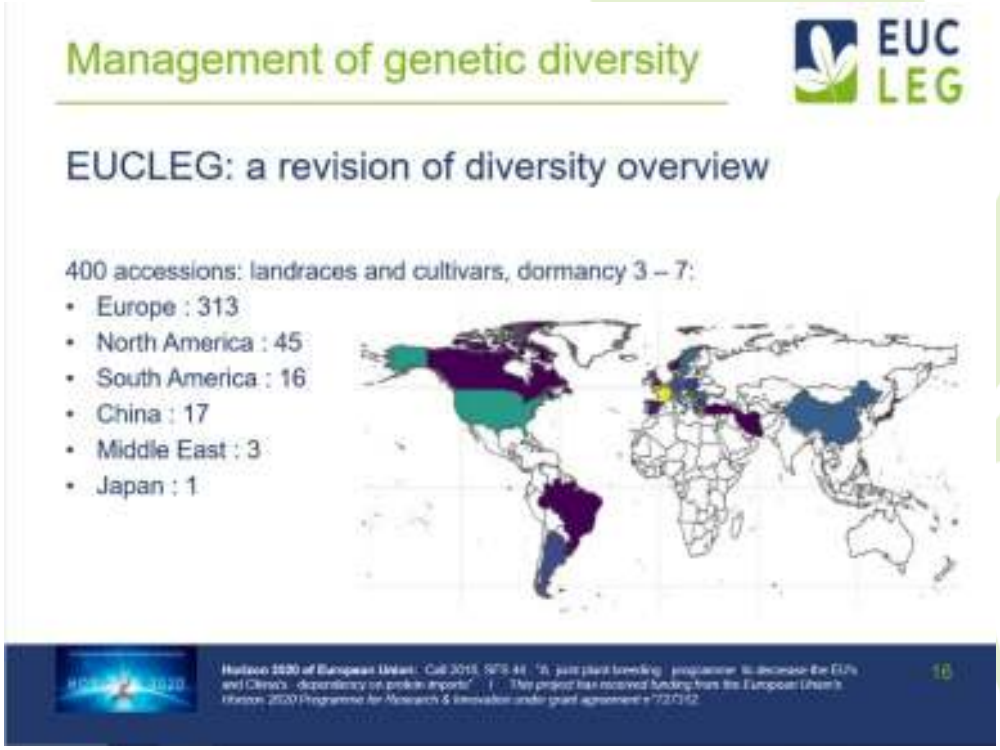


Let's move to the management of genetic diversity. Before EUCLEG, we had quite a bit of knowledge of course. We had an overview of world diversity with the Centre of Origin in the Middle East and the trace of its introduction in Western Europe and North Africa. Alfalfa, of *sativa* subspecies origin, followed the migrations with Greeks, Romans and it hybridized with *falcata* subspecies populations from Northern Eurasia. Alfalfa moved towards the Americas and Australia from 1600 on. There are also historical traces of its movement towards China about 2000 years ago.

This project has received funding from the European Union's Horizon 2020 Programme for Research & Innovation under grant agreement n°727312.

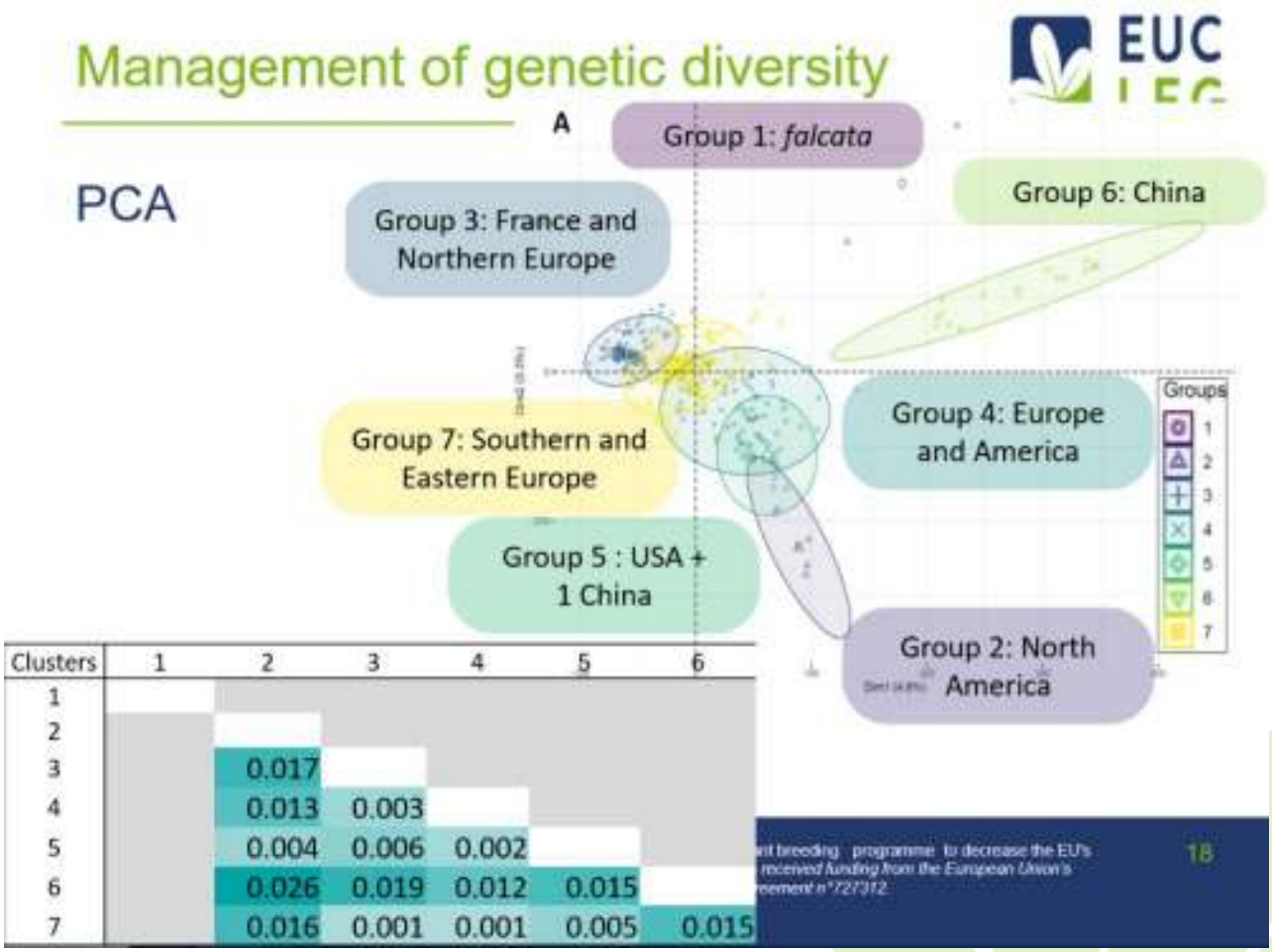


In the past, we had quite a nice overview of among and within-accession diversity. A large among accession diversity was evidenced for phenotypic traits and molecular markers, but a huge within accession diversity was also shown with the phenotypic traits or SSR. Alfalfa thus offers a very large diversity within the varieties with phenotypic traits and even more with markers.



This project has received funding from the European Union's Horizon 2020 Programme for Research & Innovation under grant agreement n°727312.

With the results from EUCLEG we have obtained a revision of genetic diversity. We have studied most extensively 400 accessions, landraces and cultivars. The dormancies are from type 3 to 7 mostly. Most of the accessions come from Europe, but we also have accessions from North and South America, China, the Middle East and 1 accession from Japan.

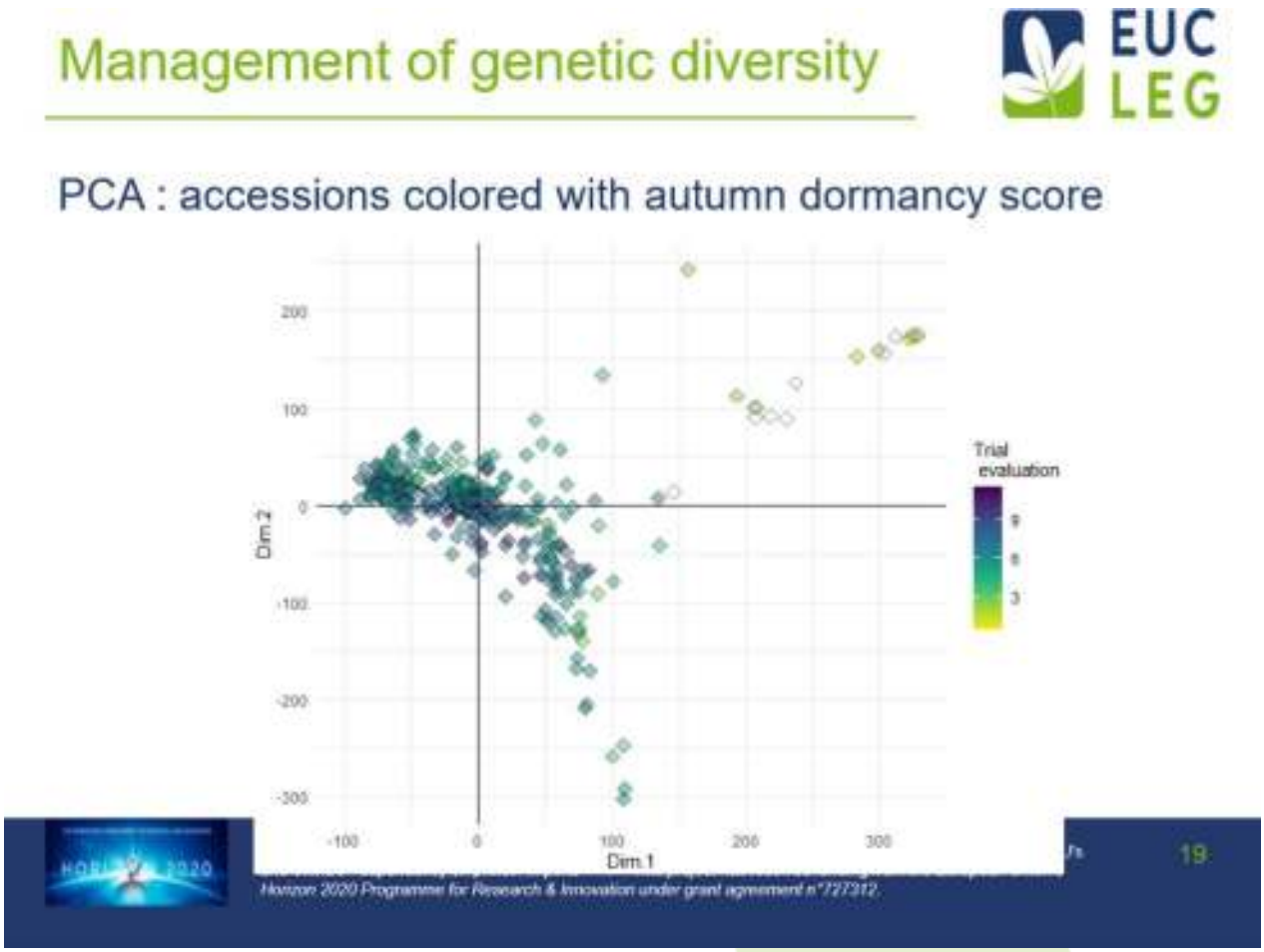


Clusters	1	2	3	4	5	6
1						
2						
3		0.017				
4		0.013	0.003			
5		0.004	0.006	0.002		
6		0.026	0.019	0.012	0.015	
7		0.016	0.001	0.001	0.005	0.015

We have obtained GBS genotypes for all these accessions and after PCA with these markers, we have identified several groupings as seen above. The first group represented with only 2 accessions is close to the subspecies *falcata*. We have quite a clear different group of accessions coming from China (group 6). We have also five groups, more or less overlapping, branching from Europe to America. Group 3 with accessions from France and Northern Europe is quite far from group 2, composed of accessions from varieties from North America. This means that there is a structure that is partly related to the geographic origin of the varieties. Another point about diversity, we have calculated F_{ST}, a diversity index, among groups, and you can see that the F_{ST} overall are quite low. The group giving the highest F_{ST} is group 6 with the accessions of China.

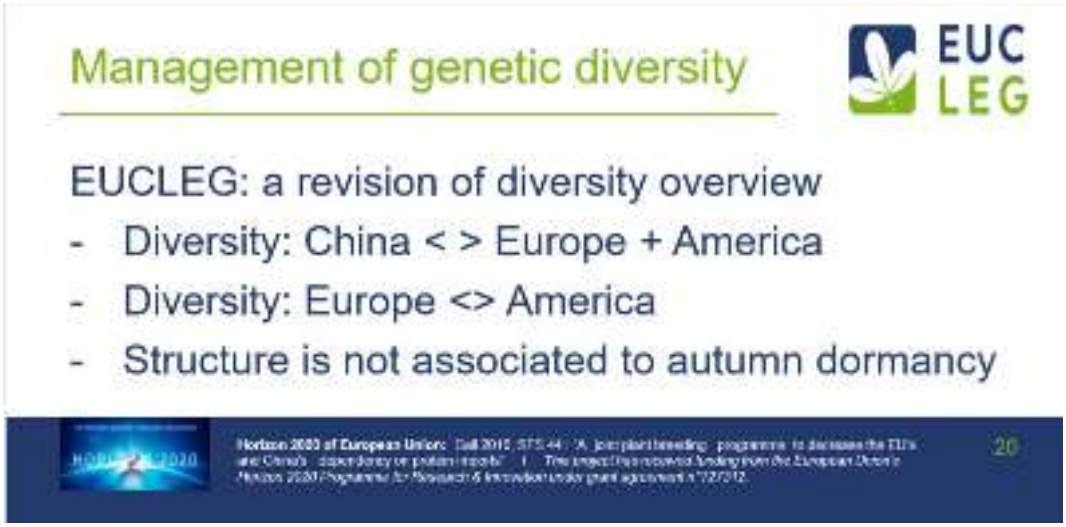
This project has received funding from the European Union's Horizon 2020 Programme for Research & Innovation under grant agreement n°727312.

Management of genetic diversity

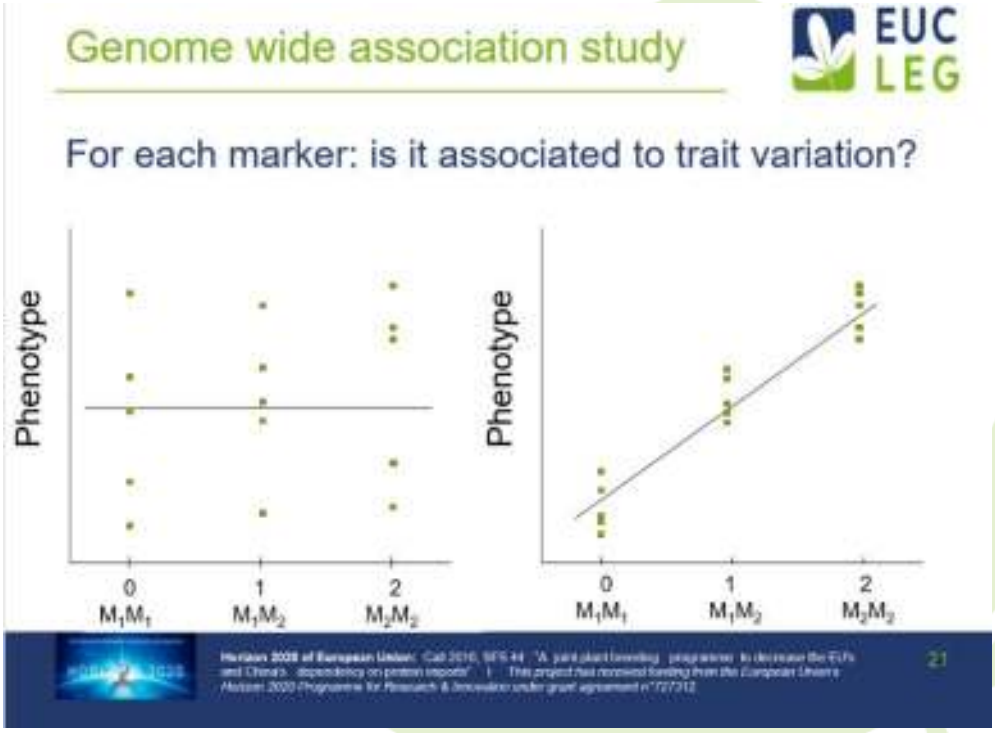


For most people in charge of alfalfa breeding, autumn dormancy is a very important trait and it gives a strong structure to the breeding programmes because breeders are usually working within a certain autumn dormancy. Here if you look at the image where we put a dormancy score on the PCA plot, you can see that dormancy is not a way for sorting the varieties. Thus diversity structure is not linked to the autumn dormancy score and these are quite new results.

This project has received funding from the European Union's Horizon 2020 Programme for Research & Innovation under grant agreement n°727312.

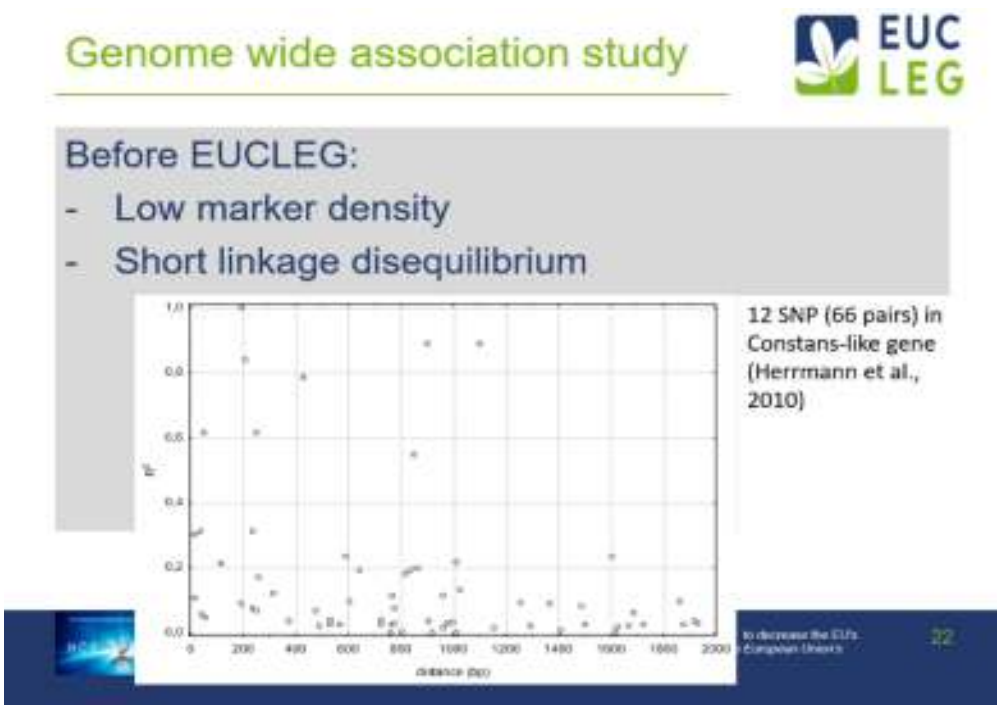


With this project, we have revised the overview of genetic diversity. We have a diversity in China which is quite different to the diversity which is present in Europe and America. Also the diversity in Europe and America is different even if most American accessions originate from Europe. The structure of the diversity is not associated with autumn dormancy.

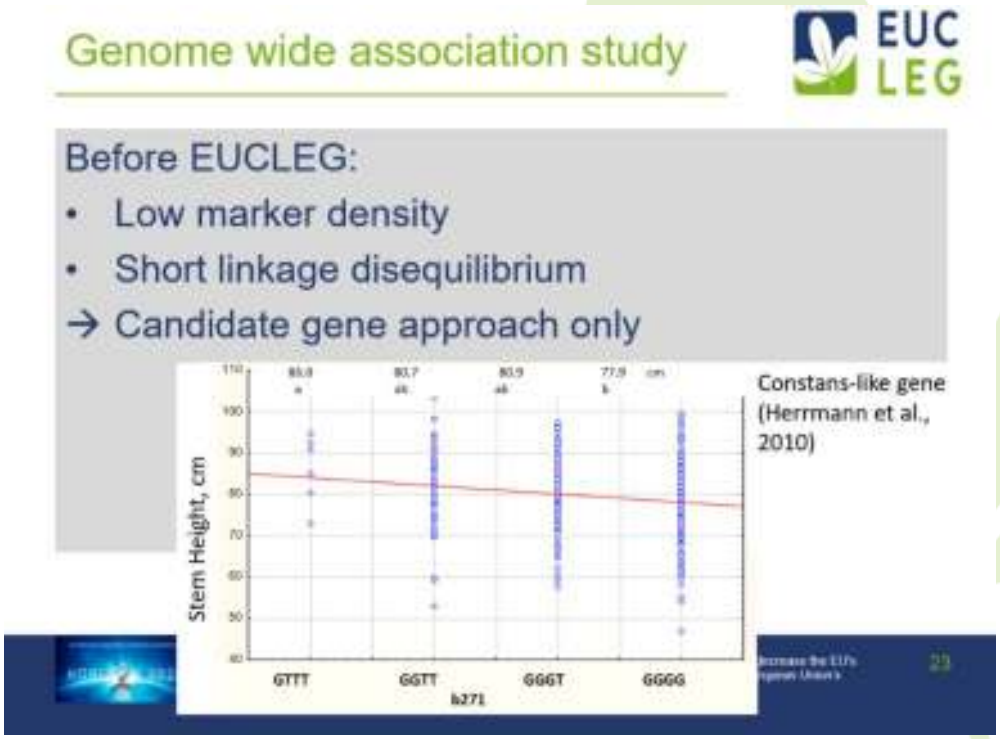


Now I will discuss the genome wide association study. Briefly, the question is to test if each marker is associated to trait variation. Here on the left hand side, the marker is not associated to the phenotype and on the right hand side, the marker is associated with phenotypic trait.

This project has received funding from the European Union's Horizon 2020 Programme for Research & Innovation under grant agreement n°727312.

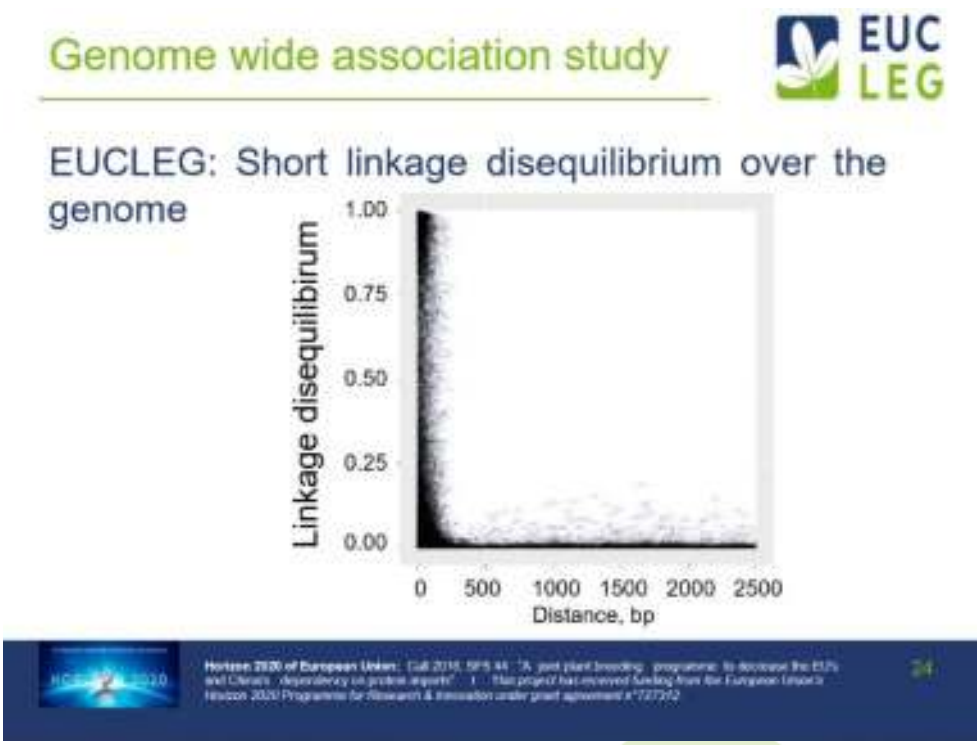


Before EUCLEG, we had low marker density, but we had shown that linkage disequilibrium was very short in this species, here studied at the level of a single gene, and linkage disequilibrium over 1000 base pair is broken with some exceptions in this gene.



With this short linkage disequilibrium and low throughput genotyping, only a candidate gene approach could be acceptable for an association study. It can work, as shown here in a Constans-like gene and the association of this gene to a phenotypic trait, stem height.

This project has received funding from the European Union's Horizon 2020 Programme for Research & Innovation under grant agreement n°727312.



In EUCLEG, we have found again a very short linkage disequilibrium over the whole genome. After 500 base pair, there is no more linkage disequilibrium, a very short linkage disequilibrium as expected in the allogamous species. It means that, because we have set up the GBS methodology that yields many markers, genome wide association studies are now possible. The candidate gene approach is of course still available for association studies.

We have done extensive phenotyping during EUCLEG, we have studied yield and quality described by protein content, fibre content and saponins. We have studied 400 accessions at 2 locations, 2 years after the establishment year. In addition, we have also studied 100 accessions within the 400 accessions, in 3 locations across 2 years. We have also studied germination, diseased resistance, drought tolerance and phosphorous tolerance and interaction between drought and *Fusarium*.

This project has received funding from the European Union's Horizon 2020 Programme for Research & Innovation under grant agreement n°727312.

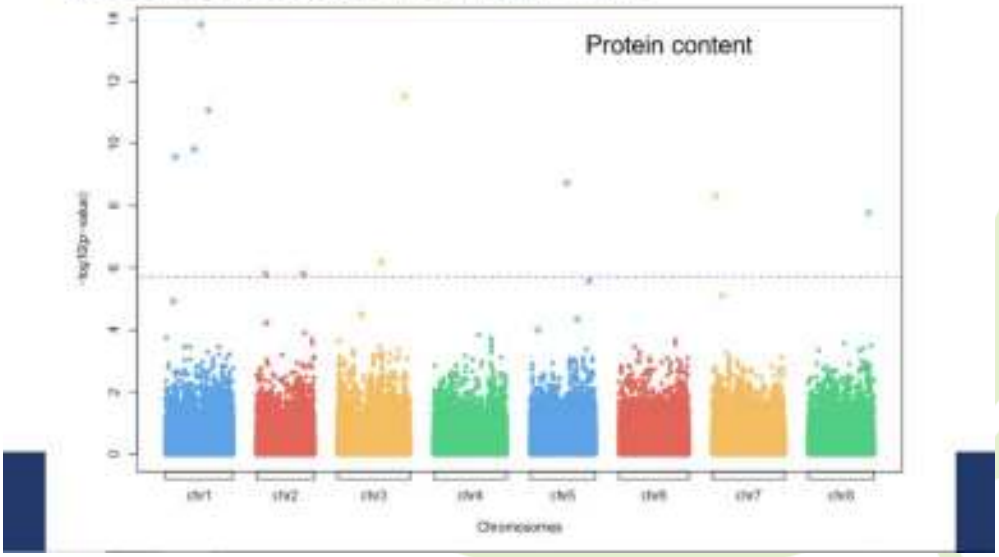
Phenotyping

Yield and quality (proteins, fibres, saponins)
400 accessions x 2 locations x 2 years
\+ 100 accessions x 3 locations x 2 years
Germination
Disease resistance: fusarium, anthracnose
Drought and P tolerance
Drought x fusarium

In EUCLEG, we were able to detect major QTLs in these accessions using a multi locus mixed model (MLMM).

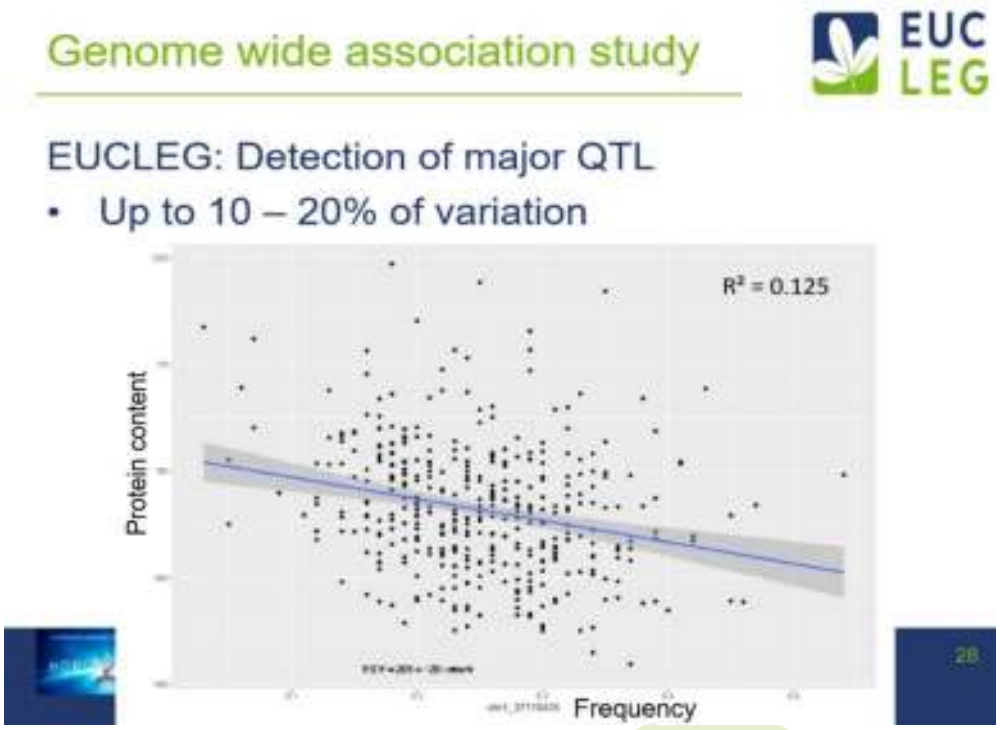
Genome wide association study

EUCLEG: Detection of major QTL

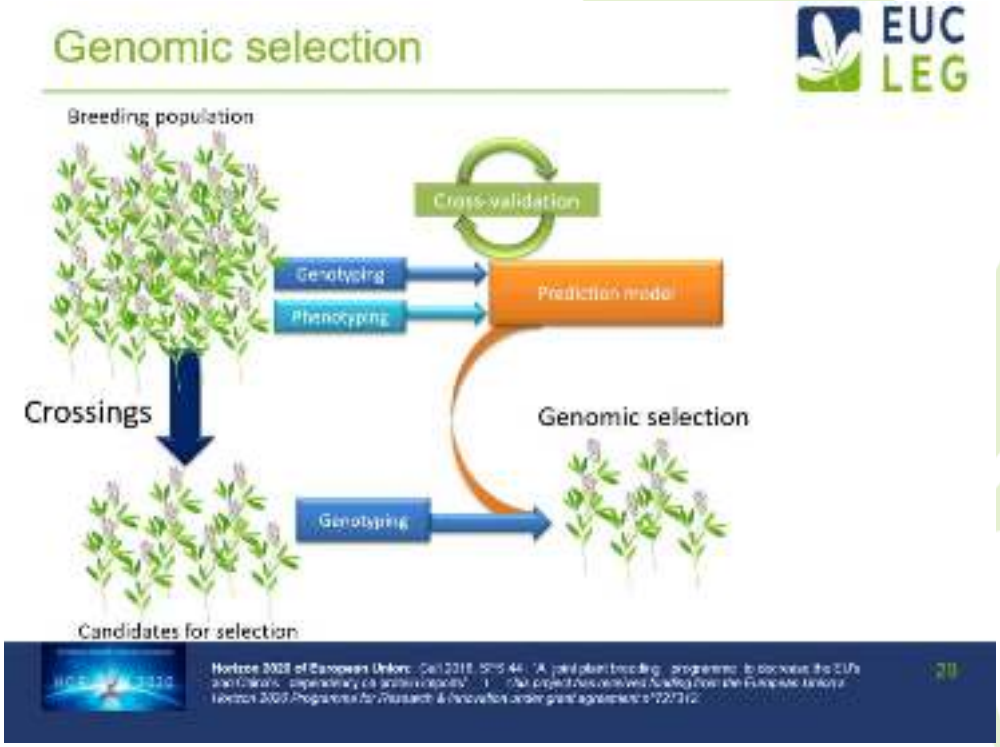


The results here show an example on protein content for which we were able to identify some QTLs with a strong effect and a significant p-value.

This project has received funding from the European Union's Horizon 2020 Programme for Research & Innovation under grant agreement n°727312.



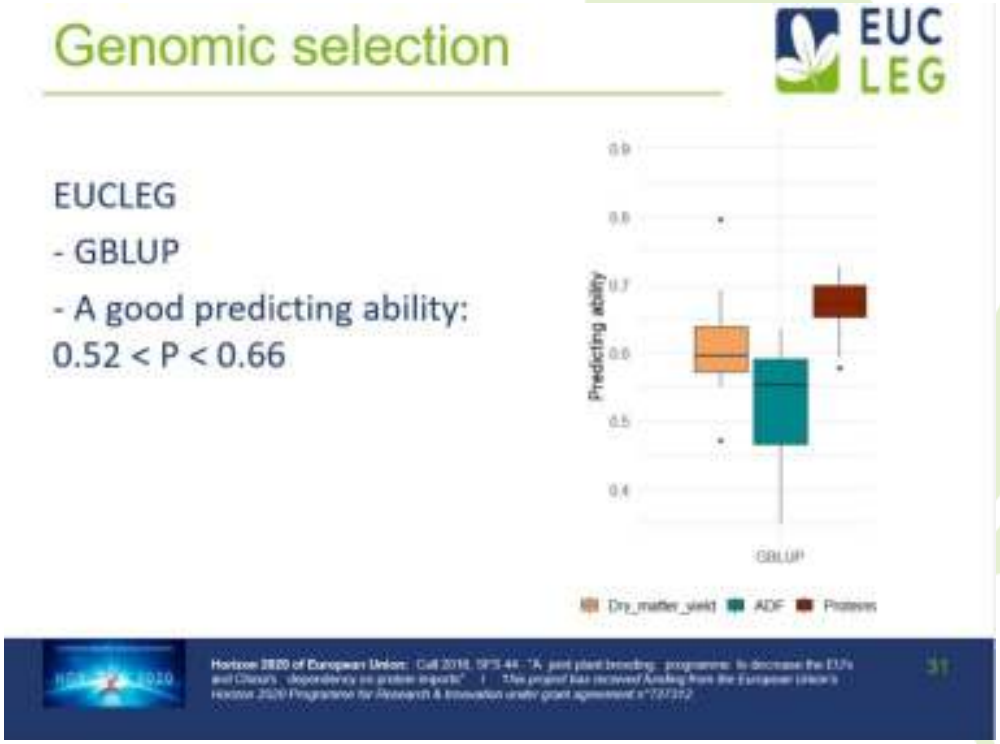
If we look at the specific QTL amongst this data for protein content, we were able to see quite a nice explanation of variation, here with an explanation of 12.5% of variation for protein content. Depending on the traits, we were able to identify QTL explaining 10-20% of the variation.



This project has received funding from the European Union's Horizon 2020 Programme for Research & Innovation under grant agreement n°727312.

Now we will look at genomic selection. As explained before, it is based on a breeding population on which we have both genotype and phenotype. First, you establish a prediction model, you test this with cross validation. Then the plants obtained after crossing - that are candidates for selection, are genotyped. You apply your prediction model on this genotyping data and, from the predicted values, you select the plants you prefer.

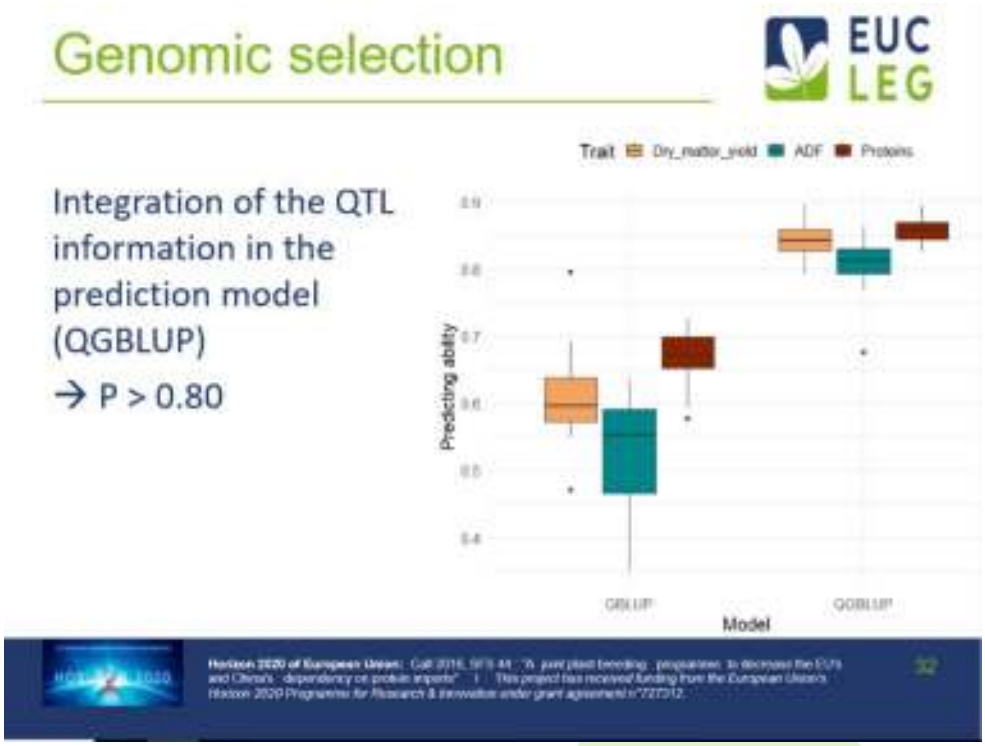
Before EUCLEG, some attempts were published on genomic selection. The number of SNPs was not so high and the number of individuals used was not very high, but provided some promising results with the predictive ability averaging 30%.



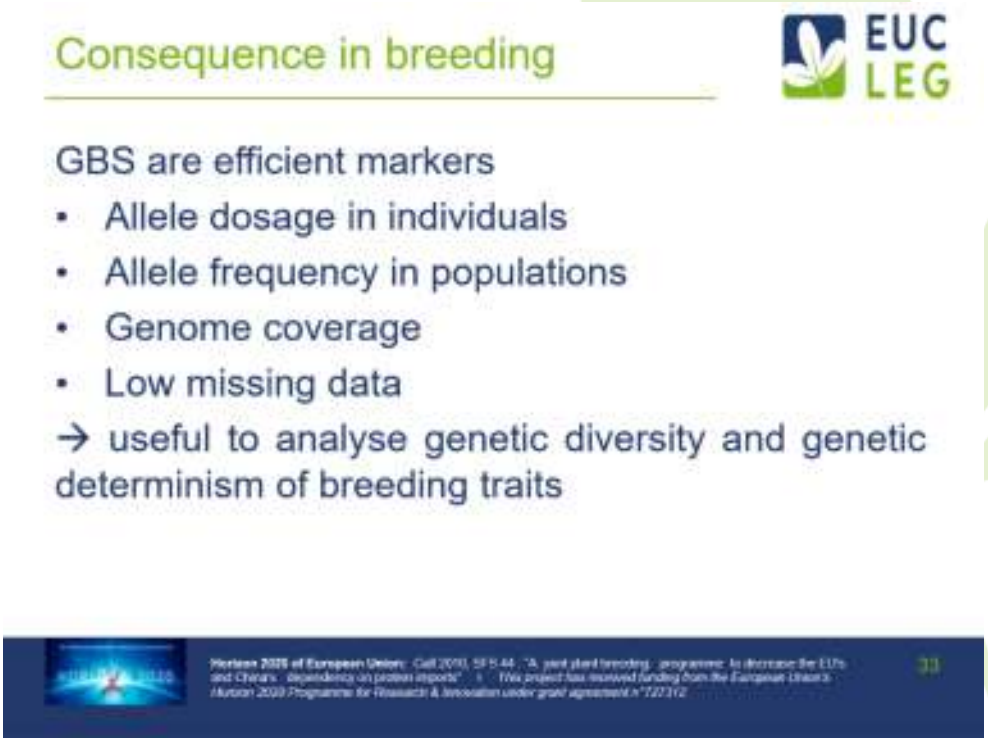
In EUCLEG, based on our 400 accessions and more than 200 000 markers, we carried out GBLUP prediction and we obtained quite a good predicting ability, between 0.5 and 0.8 for the p-value, shown here for dry

This project has received funding from the European Union's Horizon 2020 Programme for Research & Innovation under grant agreement n°727312.

matter yield, ADF content (fibre content) and protein content. The predicting ability was better for protein content than for the two other traits.



We then integrated QTL information in the prediction model with QGBLUP and you can see that we obtained a very high p-value over 0.8 and this is very interesting.



This project has received funding from the European Union's Horizon 2020 Programme for Research & Innovation under grant agreement n°727312.

As a consequence, in breeding, we have GBS as an efficient technique to reveal markers. We can have allele dosage in individuals and we can have allele frequencies in populations. We have a high coverage of the genome with low missing data. This will be useful to analyse genetic diversity and genetic determinism of breeding traits.

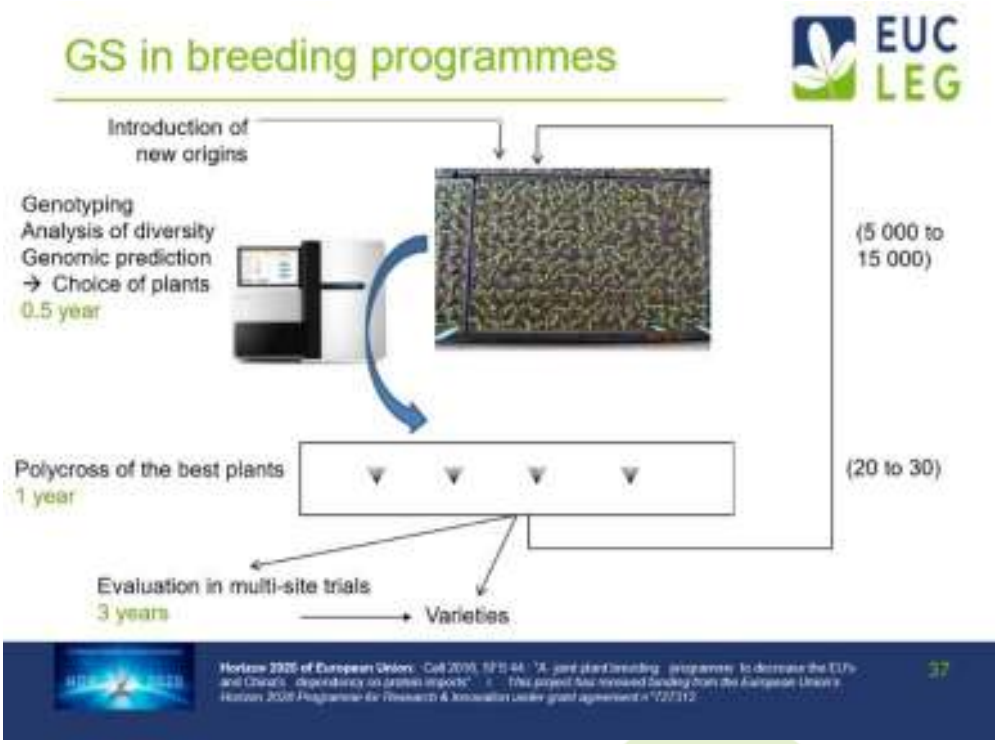
These markers can be used to manage genetic diversity. We have evidenced some specialization of the breeding pools in the EU, America, and China. We can also use these markers to decide on the introduction of new genetic diversity in the breeding pool.

We have seen that GS models provide high predictive ability, with even higher predictability when we include the QTL effects. These models are ready to be used to select promising individuals in breeding pools.

This project has received funding from the European Union's Horizon 2020 Programme for Research & Innovation under grant agreement n°727312.

What do we still have to do? We need to extend the analysis of alfalfa diversity from dormancy 3-7 to the whole species complex, including wild populations if possible. We must improve cost-efficiency of genotyping, we need to see if we can reduce the cost to be applied in a breeding programme. We also have to calculate genetic gain with GS prediction, which will also depend on the cost of genotyping. And then estimate the cost efficiency of GS. And of course, we need to implement genomic selection of breeding programmes to go from the theoretical to the practical aspect.

This project has received funding from the European Union's Horizon 2020 Programme for Research & Innovation under grant agreement n°727312.



Here I propose implementation of genomic selection in breeding programmes. Starting from the introduction of new origins and the current breeding pool, the breeding programme starts by growing seeds in the greenhouse and as soon as possible collect leaflets on each seedling, extract DNA and obtain the genotypes. Then, the genomic prediction model is applied and at this stage after a few months only, you are able to choose the best plants to be established in one or several polycross. From the polycross and after 1 year, you have seeds to test the polycross. Some of these can be a candidate for registration and you continue the recurrent selection. You can imagine evaluating many more candidates with genotyping than with phenotyping, moving from 5000 to 15000 plants depending on the cost.

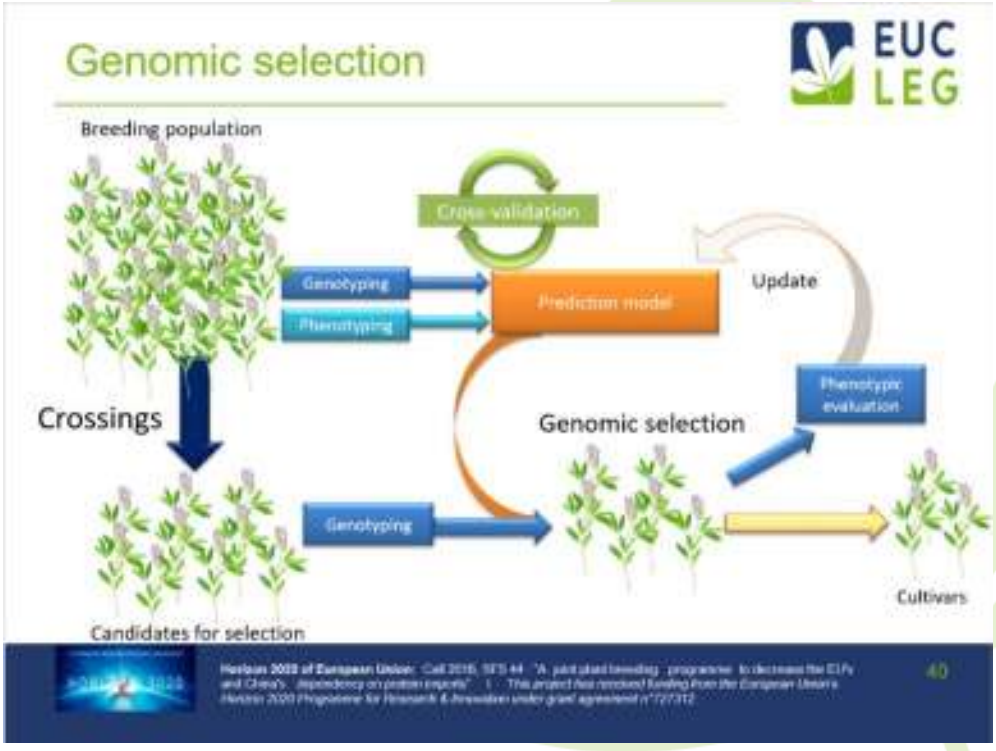
This project has received funding from the European Union's Horizon 2020 Programme for Research & Innovation under grant agreement n°727312.

What are the strengths of this breeding programme? Reduced field work, a very early selection for all predicted traits, a reduced number of years of the breeding cycle, and a very quick fixation of positive alleles, especially important for the autopolyploid species. There are some weaknesses, firstly if you have no prediction for some traits, you are not able to select for this trait, and secondly, issues with staff having to be trained to develop new skills to adopt these new breeding programme. The question is now: what is the genetic gain and the cost efficiency?

This project has received funding from the European Union's Horizon 2020 Programme for Research & Innovation under grant agreement n°727312.

Still to be done

- Extend the analysis of alfalfa diversity from dormancy 3-7 to the whole species complex
- Improve cost-efficiency of genotyping
- Calculate genetic gain with GS
- Estimate cost-efficiency of GS
- Implement genomic selection in breeding programmes
- Imagine the updating of GS equations



The next step will be to imagine the updating of GS equations, because up to now we have an equation, but we need to make it living, including new genetic resources and new phenotyping data. The question is how to use the data obtained on new progeny or new polycross. Phenotypic evaluation and genotypic data could be used to update the existing prediction model

This project has received funding from the European Union's Horizon 2020 Programme for Research & Innovation under grant agreement n°727312.

Questions and answers from the presentation.

(Q1) What was the cost per sample in your GBS approach?

Approximately 50 Euros for one population. Does that include the DNA extraction? Yes, everything apart from staff time

(Q2) How many years of phenotyping does it take to build the genomic selection model and calculate the prediction accuracy?

In our case we used data for 2 years. We didn't use the data from the first year.

(Q3) How often do you have to renew the prediction model?

Good question. We don't know in fact, maybe the first thing we have to establish is the efficiency (the quality) of the predictions, depending on the accessions you are studying. I have shown some groups of accessions, we have to test if a prediction model is valid for all types of accessions or not. This is the first part of the answer. Once you start using the prediction model, you select plants so the genetic bases of the material may change; of course we have to check this and to learn from experience. My idea is that maybe we could not start again from zero, meaning we don't have to collect so many accessions and study them again in field trials. Maybe we could use this first set of information and then add new information coming from new trials and new accessions. It is not simple at a mathematical level and we also have to find an organisation to do that.

(Q4) How much of the variability between each trait varied between years?

We had some changes in the variability and we also had some interaction between the environment and genotype. Here we have tried to predict the mean values of the populations, for example for annual yield. We are able to also predict the traits in each environment. We have obtained a better evaluation if we look at the average value of the populations over all the sites that were available.

(Q5) Were the predictions accuracies that you showed cross validation results?

The equation was done on a subset of accessions and used to predict another subset. There was also a test set to calculate the p-value.

This chapter is based on a presentation given to the EUCLEG online workshop on the application of cutting-edge genomic technologies in the breeding of legume species held on the 30th September and 1st October 2021

Recording link to the presentation: https://youtu.be/l6QEXn5Uhd0

This project has received funding from the European Union's Horizon 2020 Programme for Research & Innovation under grant agreement n°727312.