



**HAL**  
open science

## The application of genomic technologies in the breeding of legume species

Catherine Howarth, Sarah Clarke, Bernadette Julier, Isabel Roldán-Ruiz, David Lloyd, Hilde Muylle, Radu Grumeza, Ana Maria Torres, Leif Skot, Roland Kölliker

### ► To cite this version:

Catherine Howarth, Sarah Clarke, Bernadette Julier, Isabel Roldán-Ruiz, David Lloyd, et al.. The application of genomic technologies in the breeding of legume species. EUCLEG online workshop, Sep 2021, En ligne, 182 p., 2021. hal-03655588

**HAL Id: hal-03655588**

**<https://hal.inrae.fr/hal-03655588>**

Submitted on 29 Apr 2022

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



**EUCLEG project “Breeding forage and grain legumes to increase EU’s and China’s protein self-sufficiency” ([www.eucleg.eu](http://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 30<sup>th</sup> September and 1<sup>st</sup> 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](http://www.eucleg.eu)).

## Contents

<b>1. Introduction to EUCLEG</b> .....	3
Bernadette Julier .....	3
<b>2. Lessons learned on the design and planning of multi-location trials and phenotypic assessment for association studies</b> .....	14
Isabel Roldán-Ruiz .....	14
<b>3. Selection of genotyping platforms: GBS and SNP arrays for individuals and populations</b> .....	31
Leif Skot .....	31
<b>4. Introduction to inbreeding species: traditional breeding methodologies</b> .....	52
David Lloyd .....	52
<b>5. Genomics assisted breeding in soybean</b> .....	59
Hilde Muylle.....	59
<b>6. Genomics assisted breeding in pea</b> .....	86
David Lloyd and Radu Grumeza.....	86
<b>7. Genomics assisted breeding in faba bean</b> .....	95
Dr Ana M <sup>a</sup> Torres.....	95
<b>8. Introduction to outbreeding species: traditional breeding methodologies</b> .....	127
David Lloyd .....	127
<b>9. Genomics assisted breeding in alfalfa</b> .....	135
Bernadette Julier .....	135
<b>10. Genomics assisted breeding in red clover</b> .....	160
Roland Kölliker.....	160





## 1. Introduction to EUCLEG

Bernadette Julier

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

Horizon 2020 of European Union: Call 2016-5F3-44: "A joint plant breeding programme to decrease the EU's 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.

**EUC  
LEG**

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

INRAE

Bernadette Julier

[www.eucleg.eu](http://www.eucleg.eu)



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

**EUCLEG.eu**

## Horizon 2020 of European Union



Call 2016, SFS 44 : "A joint plant breeding programme to decrease the EU's and China's dependency on protein"



EUCLEG: 09/2017 – 12/2021



This project EUCLEG was prepared after a call launch by the EU in 2016 entitled "a joint plant breeding programme to decrease the EU's and China's dependency on protein". The call included many important words requesting consideration of many topics such as forage and animal feed, productivity, climate change, diversification, stresses etc. and also called for collaboration with Chinese colleagues.

## Protein imports in Europe and China



Europe dependency : 69%

China imports 60% of soybean world market trade



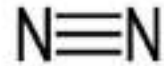
This project has received funding from the European Union's Horizon 2020 Programme for Research & Innovation under grant agreement n°727312.

**EUCLEG.eu**



## From Nitrogen (N<sub>2</sub>) to proteins

Dinitrogen: very stable molecule, 78% of the atmosphere



N is a component of proteins, vital molecules

Two ways to transform N<sub>2</sub> into reactive Nitrogen:

- Industrial chemical synthesis



- Symbiosis plant + *Rhizobium*



Plant amino acids

Plant proteins

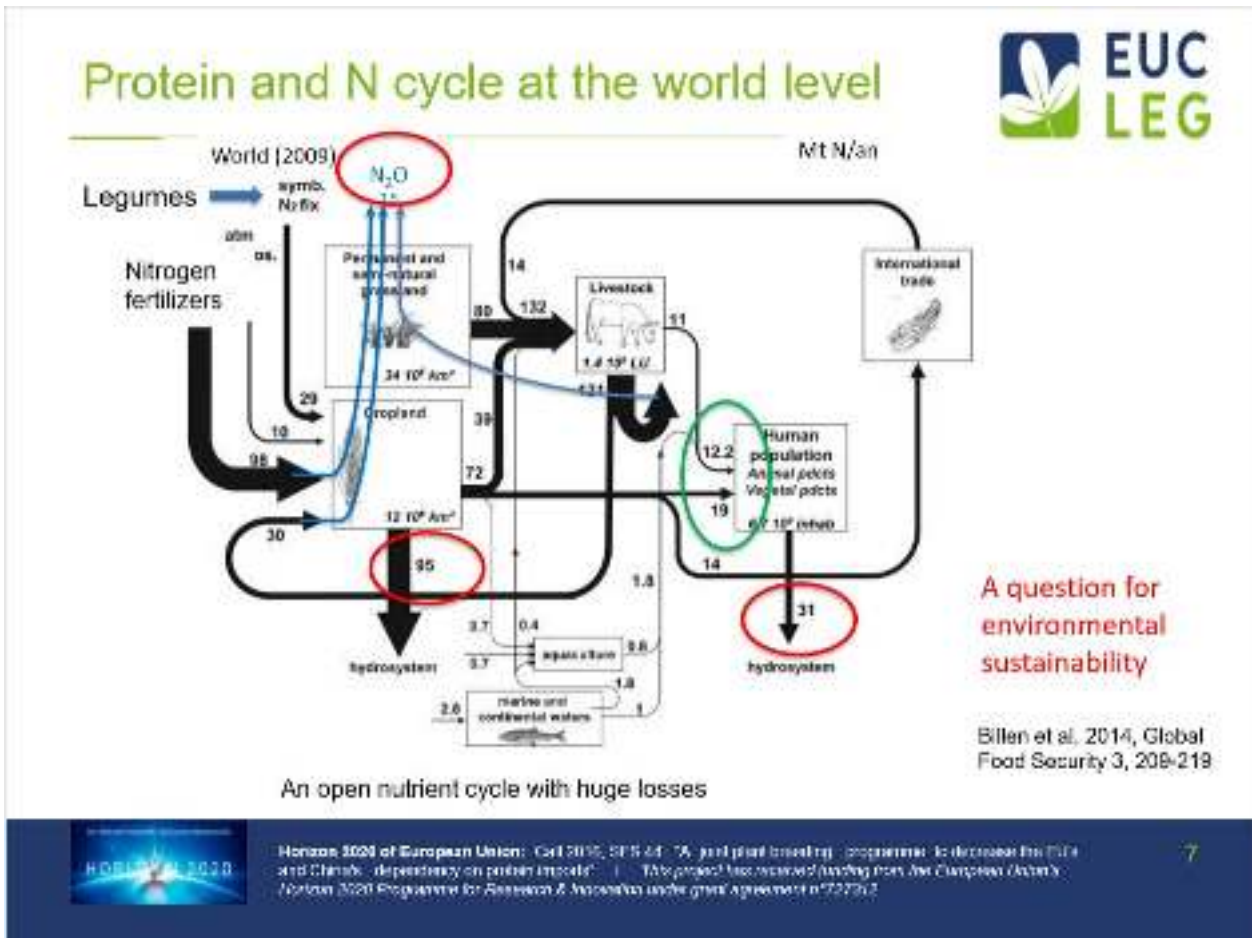
Legume species  
(*Fabaceae*)



Horizon 2020 of European Union: Call 2016, SFS 44: "A joint plant breeding programme to reduce the 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.

6

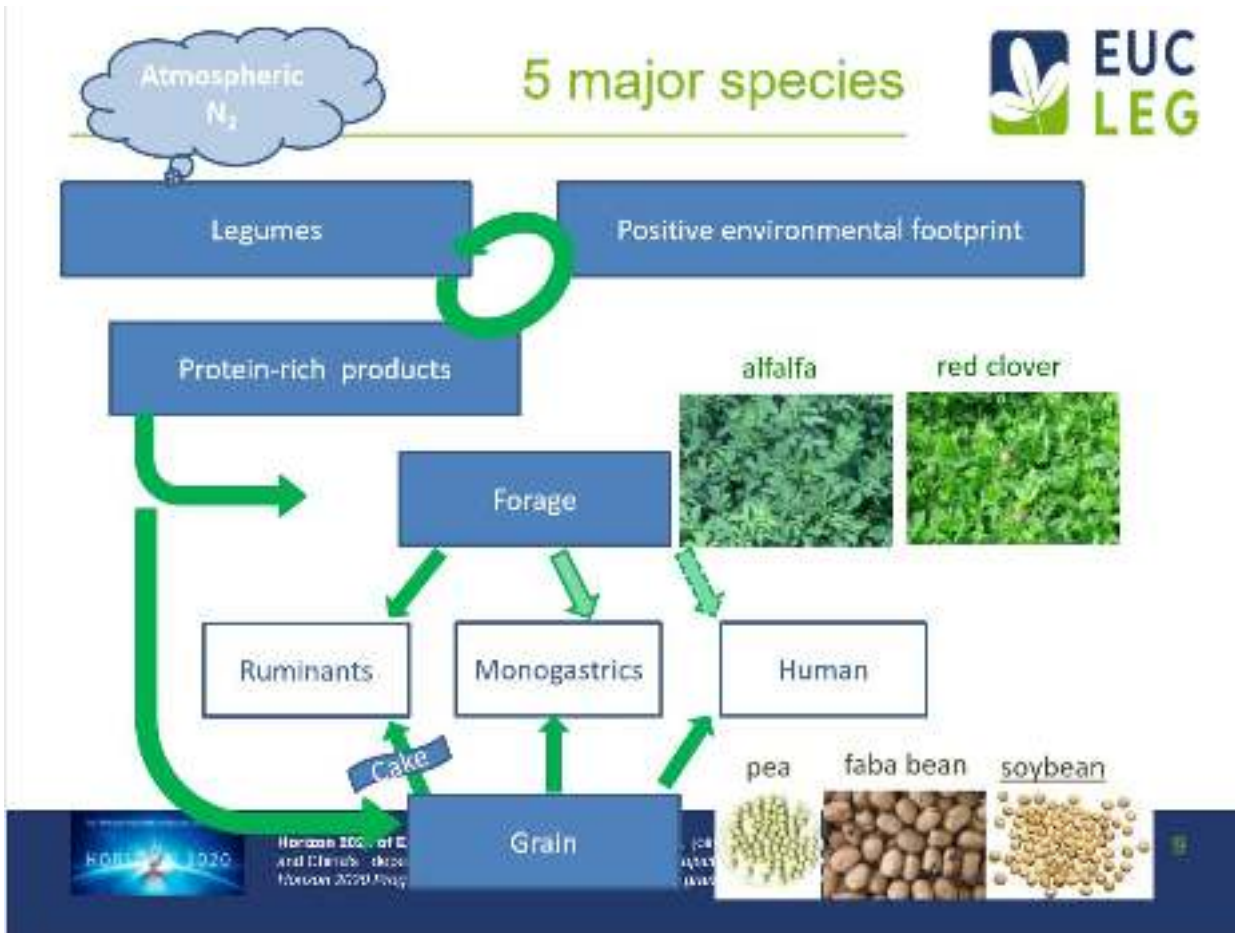
What about the proteins? In fact the question here is how to transfer atmospheric nitrogen gas (more exactly dinitrogen) into protein. Dinitrogen is a very stable molecule and makes up 68% of the atmosphere. It is a molecule composed of 2 atoms of nitrogen which bond with 3 very strong covalent bonds. Nitrogen is a component of proteins. There are two ways to transfer dinitrogen into reactive nitrogen, the first one is chemical synthesis carried out in industries, the process turns dinitrogen into ammonia by using large quantities of fossil gas energy. The second way is to use the symbiosis between legumes and specific soil bacteria called *Rhizobium* that are able to carry out the same reaction with the help of the nitrogenase enzyme. Plants are then able to assimilate and absorb the ammonia and transform it into amino acids and then plant proteins.



We also have a problem with the nitrogen cycle at a world level. Nitrogen enters into the plant system, mostly from nitrogen fertilisers that are applied on both crop and grasslands (72%). The second way that nitrogen enters the system is from nitrogen fixation with legumes, but this represents only 21%, and the third way is from acquisition from the atmosphere (7%). By using this nitrogen, the plant grows to produce grains, forages, fruits or vegetables but there is important leaching towards the hydrosystem. The plant products are eaten either by livestock or the human population and again there is some waste from livestock or human effluents. As a whole, the nitrogen cycle is completely open with huge losses that cause pollution and question environmental sustainability of agricultural system. We also see here the part that the entrance of nitrogen into the system of crop production plays and the small part of nitrogen currently coming from plant based nitrogen fixation.

As a consequence, the current situation indicates that we need to increase atmospheric nitrogen fixation by expanding legume cropping and produce more plant proteins that originate from nitrogen fixation.





In EUCLEG, we have focused on the five major agricultural legume species in Europe. These species are able to produce protein rich products, either forage or grain, whilst providing a positive environmental footprint. We have studied alfalfa (or lucerne) and red clover as forage crops, pea, faba bean and soybean as grain crops. Our Chinese colleagues have worked on alfalfa and soybean, these two species are the most important legume crops in China.

Forages are used to feed ruminants mostly, monogastrics marginally and very rarely humans. Grain legumes are used to feed monogastrics and human and partly ruminants, mostly as soybean meal.

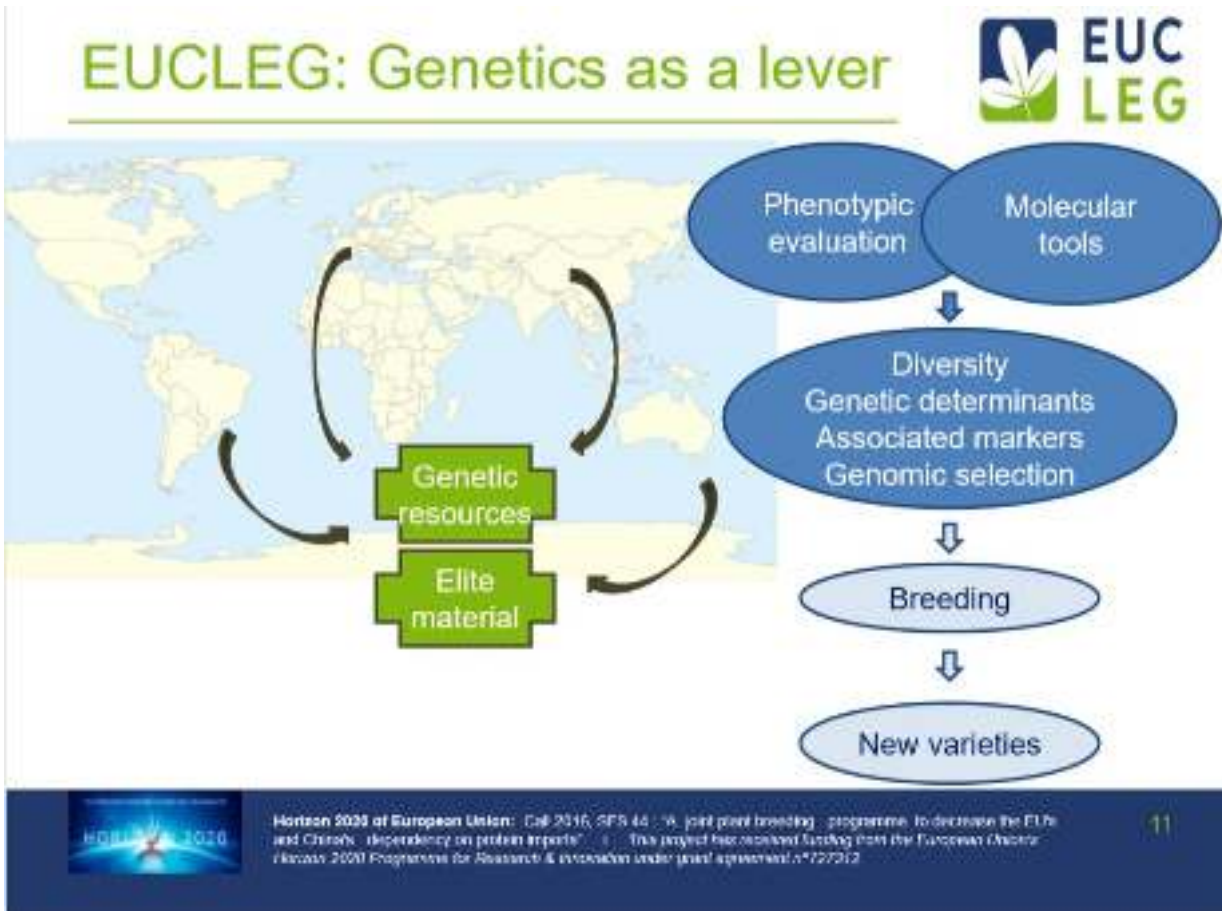
## Eucleg impacts

To increase protein production where legumes are already grown

To increase adaptation of legumes to more pedoclimatic regions



The expected impact of EUCLEG is to increase protein production in the regions where legumes are already grown, but also to increase the adaptation of legumes to regions where we are not currently able to grow legumes at a competitive level. At the end, the production of feed and food must be achieved with an improved yield and yield stability. This requires many improvements related to general adaptation, resistance to biotic and abiotic stresses, adaptation to climate change, also quality traits that, depending on the species, refer to protein content and protein composition, forage quality and anti-nutritional components.



EUCLEG uses genetics as a lever to achieve these goals, it is based on the use of genetic resources and also elite material coming from worldwide sources, with an emphasis on material from Europe and China. A large phenotypic evaluation was carried out with a long list of traits established by the species experts. We have also developed molecular tools, with which the accessions of the five species were genotyped. We have studied genetic diversity, genetic determinants of traits and looked for markers associated with trait variation i.e quantitative trait loci (QTL). We have also worked on the potential of genomic selection to improve breeding programme efficiency. The outputs and impacts will be for breeding of new varieties in the future.



## EUCLEG: Genetics as a lever



### At the scientific level:

- **Broaden the genetic base of legume crops and analyse the genetic diversity** of European and Chinese legume accessions using phenotypic traits and molecular markers
- **Analyse the genetic architecture of key breeding traits** using association genetics (GWAS)
- **Evaluate the benefits brought by genomic selection (GS)** to create new legume varieties

### At the technological level:

- **Develop searchable databases** containing passport data, as well as agronomic and genetic features
- **Develop molecular tools and data**

### At the applied level (breeding):

- **Develop tools for genotyping**
- **Implement data management and analysis**



Horizon 2020 of European Union: Call 2016, SES-11 - A joint plant breeding programme to decrease the EU's 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.

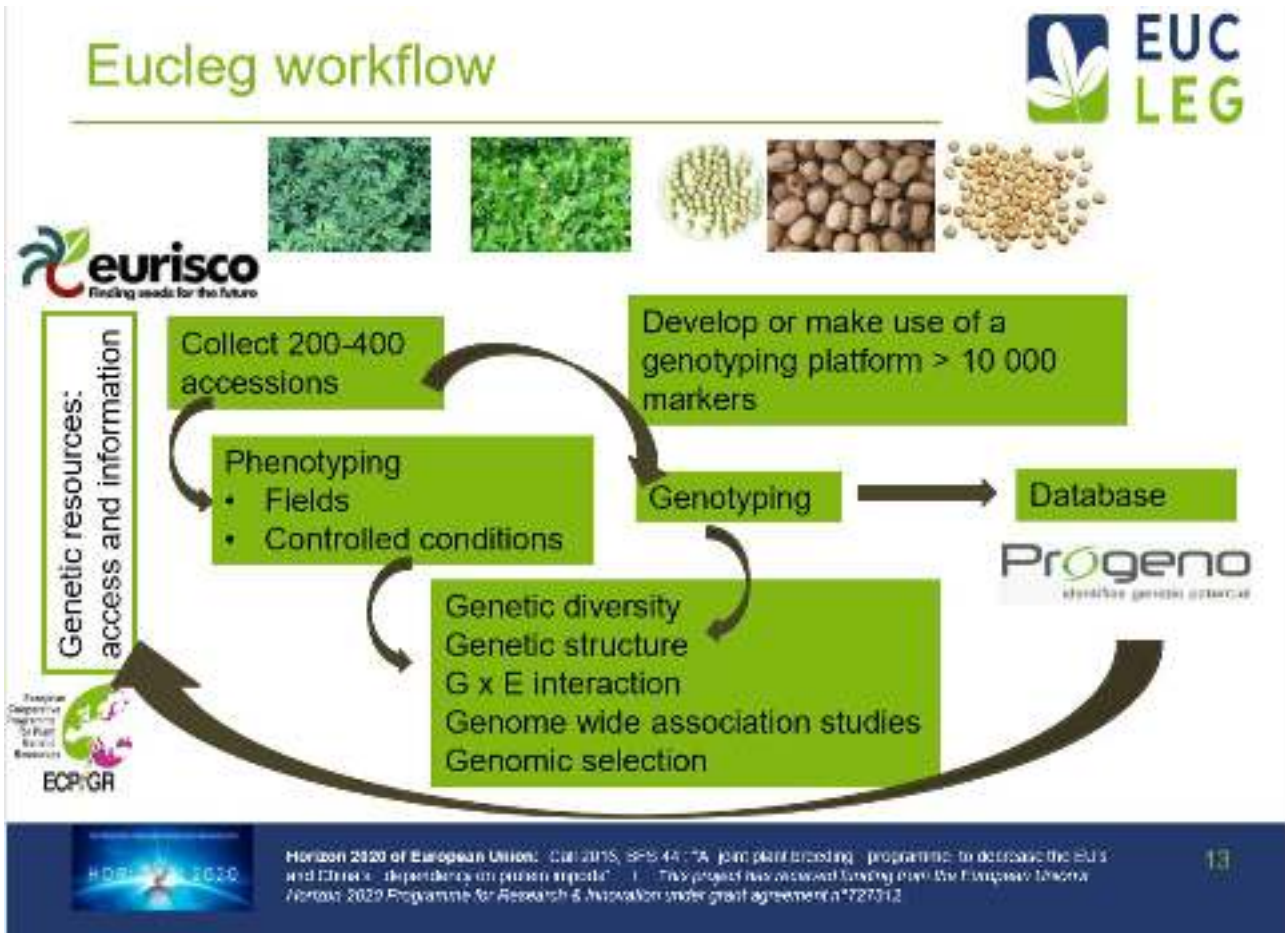
12

EUCLEG had ambitious objectives at different levels. At the scientific level, a first objective was to expand the genetic basis used by breeders in legume breeding programmes, after the analysis of the genetic diversity of European and Chinese legume accessions, using phenotypic traits and molecular markers. We also had an objective to analyse the genetic architecture of key breeding traits using genome wide association and analysis (GWAS) and to evaluate the benefits for genomic selection to create new legume varieties. At the technological level, we have developed searchable databases containing passport data as well as agronomic and genetic features, as well as developed molecular tools and data that will be available for future programmes. At the applied level, that is breeding in this case, the objectives were to develop tools for genotyping and also to implement data management and analysis that is now essential in genetic and breeding programmes.



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

**EUCLEG.eu**



The EUCLEG workflow illustrated here is the basis of the presentations of this workshop and booklet. For each of the 5 species, we have adopted the same general scheme. Based on collections of genetic resources we have collated between 200 and 400 accessions. We have developed or made use of existing genotyping platforms, depending on the species, with the objective was to obtain more than 10 000 markers. We have genotyped accessions with the chosen genotyping platform. Accessions were phenotyped in different conditions, either in multi-location field conditions, but also in controlled conditions, especially for diseases or drought stresses. All these results have been transferred to databases on Progeno and this programme is available for the analysis of genetic diversity, genetic structure and genetic control of traits.

The objective of this workshop and booklet was to disseminate the results obtained so far for both forage and grain legumes. We also wish to share general considerations about the design of the experiments, genotyping and breeding methodology. The idea is to talk and discuss with you, with legume breeders and scientists to imagine the future of breeding of legumes, what we could call post-EUCLEG breeding.





## About the author

### Bernadette Julier

Dr Bernadette Julier is Research Director at INRAE, Unité de Recherche Pluridisciplinaire Prairies et Plantes Fourragères (URP3F), in Lusignan, France. Since her PhD, she has been continuously working on legume genetics and mostly on alfalfa or lucerne, the most famous, and protein producing forage species. Her main topic was first to evidence genetic variation for energy value and to combine it to forage production. She has been involved in projects on seed production and protein degradation too. More recently, her research is focused on the genetic bases of aerial morphogenesis, either in pure stand or in mixtures with forage grasses. The use of molecular markers to assist breeding is a major topic to promote genetic progress on this autotetraploid species. She is currently leading EUCLEG, an European project (H2020, 2017-2021) “Breeding forage and grain legumes to increase EU’s and China’s protein self-sufficiency” that aims to use more genetic resources and develop molecular tools to be able to create improved legume varieties ([www.eucleg.eu](http://www.eucleg.eu)) and thus promote protein production. She is a member of the Permanent Technical Committee of Selection (CTPS) in France, in the section “Forage and turf plants” since 1998, in charge of the variety registration.

**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 1<sup>st</sup> October 2021**

A recording of the presentation is available at <https://youtu.be/z6AWKmKwXJ0>



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

**EUCLEG.eu**

## 2. Lessons learned on the design and planning of multi-location trials and phenotypic assessment for association studies

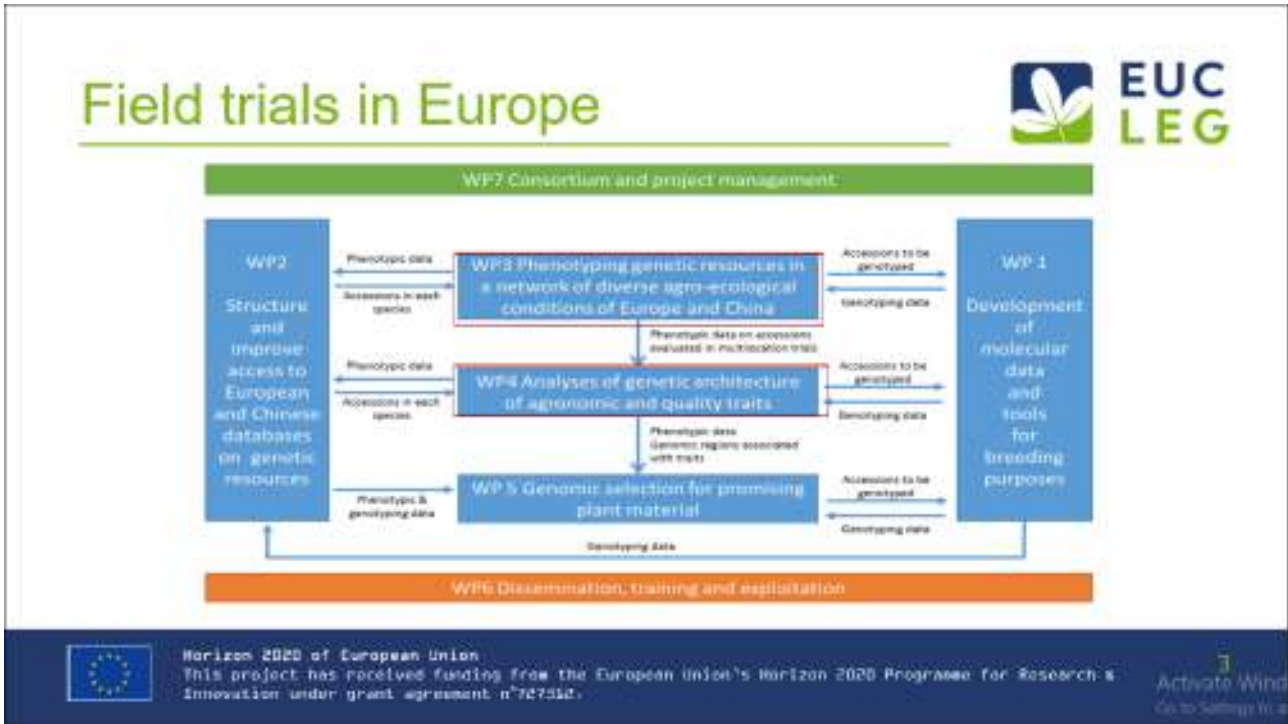
Isabel Roldán-Ruiz

Scientific Director, Plant Sciences Unit, Institute for Agricultural, Fisheries and Food Research (ILVO), Melle, Belgium and Professor, Ghent University, Department of Plant Biotechnology and bioinformatics



Hello, good morning. You have had a very nice presentation of the general objectives and the structure of the project, and I would like to elaborate further on one of the slides presented by Bernadette in the introduction, regarding the EUCLEG workflow, which things we did, how we did those things and why? Especially I will try to illustrate why we think that the approach, the EUCLEG approach, worked.

So, my presentation will be quite atypical, I will not speak about research objectives. I will not show you any research results, this is something that is kept for the other presentations of this workshop, and my colleagues will take care of that. I just want to stress and to provide information, and discussion points perhaps, on how we organize things within EUCLEG. Especially why we think that it was okay to look at it that way and why we think that it worked. Stressing also why we believe that we have produced quite high-quality data sets, to be used in the context of this project, by the running. But I'm sure that a lot of work still must be done in the future and we are quite optimistic about what can be done with the data sets that we have generated.



This is the general structure of the EUCLEG project. I will be concentrating my talk mostly on the work packages in which we were carrying out field trials (WP3 and WP4). We will deal today with the design of the trials and how we approached the establishment and the follow up of the field trials in Europe for the five crops that we are investigating. In WP4 our focus was the analysis of yield components and also quality parameters, including anti nutritional aspects of the crops that we investigated. Also in WP3 we had to establish field trials, to define and to study the phenotype and to analyse the genetic diversity present in the five crops. So, diversity can be used into the future, also for breeding purposes. What we did in the end was to combine the field trials of work package 4 with work package 3. This will become clearer when I give further explanations.



## WP4

*To identify genes, alleles and molecular markers that explain a large part of the phenotypic variation available for traits of agronomic and economic relevance using GWAS*

### Multi-location, multi-year field trials

- For five crops
- Crop yield and quality (protein content, anti-nutritional components)



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

4  
Activate Windows  
Go to Settings to activate Windows.

So, coming back to work package 4. The objective was to identify genes, alleles and molecular markers that explain a large part of the phenotypic variation available for traits of agronomic and economic relevance using GWAS. Quite broad objectives. As explained in the introduction, some traits were evaluated in controlled conditions and some other traits especially yield related traits, but also compositional aspects were evaluated using field trials. But remember that we used the same set of accessions per species, so between 200 and 400 accessions, common to all the experiments that were included in work package 4. When we are presenting here the results for, for example red clover association mapping for a given trait and then association mapping for a different trait, remember that we were dealing with the same set of accessions. So, for us it was very important, because in this way we were able to generate quite detailed and in-depth information for a common set of accessions that were included in all the experiments for the European partners. There were some variations, but it was the main objective. So, this has contributed of course to quite valuable data sets that will also become public, once we have published the results. They will also be available for other researchers if they want to investigate further, because the data will become available in public databases. So, this was the topic of many discussions at the start of the EUCLEG, to establish multi location field trials, but also multiyear field trials. This had to be done for five crops and the main focus of the field trials was crop yield and crop quality.



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

**EUCLEG.eu**

## WP4: five crops – two locations



Alfalfa (perennial)  
Red clover (perennial)

Pea (annual)  
Faba bean (annual)  
Soybean (annual)



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

5  
Activate Windows  
Go to Settings to activate Windows.


This is an overview of the locations that were relevant for work package 4. So you see here the five species. Two perennial crops, alfalfa and red clover, and three annual grain crops, pea, faba bean and soybean. As you can see, for each of those species there are two dots on the map. Please don't look at the specific location of the dots, it is the country that is relevant, as the map is quite small to put the dots accurately in the right location. So, we had to coordinate the work at least two locations, to be sure that the data could be comparable at the end, because our main purpose was to combine everything together per species, to have a joint interpretation of the data and a joint analysis of the data. But as I told you before we needed also to coordinate with work package 3. What was the difference between these two work packages? In work package 3, we were interested in analyzing the performance of a subset of the accessions, not all of them. It was a smaller set, but common to work package 4, which was to analyze their phenotypic behavior in an extra set of locations, because we wanted to know what diversity was available for the different crops, but also to investigate in depth the phenotype and its interaction with the environment. So, it was therefore necessary to consider a larger set of environments. In what follows, when I'm speaking about the design of the field trials of EUCLEG for different species, I'm referring to the field trials that were prepared jointly for work package 3 and work package 4.




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

**EUCLEG.eu**



## Joint field trials WP3 and WP4





WP3: extra locations, subset accessions WP4

- Alfalfa (perennial)
- Red clover (perennial)
- Pea (annual)
- Faba bean (annual)
- Soybean (annual)


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


So how did we approach this? I will now go through a number of requirements and provide some examples and explanation of aspects that I think are relevant. I will now take the role of the experienced scientist; I have been working in research already for almost 40 years and have been collaborating in many European projects, which are collaborative projects in which in many cases we had the same objectives as in EUCLEG, but the main mistake that we make in several projects in my experience is that there is not sufficient coordination from the beginning. So, the strength of EUCLEG as I already explained, or one of the main strengths, is that we were working with common sets of accessions to analyze the different traits. Per crop we had a common set of accessions to analyze for different aspects, and the field trials that were established at different locations were established with common sets of accessions. If this is the objective of the project, partners should not start making sub selections, as this would not contribute to the general goal of the project. This was quite a strong control point at the start of EUCLEG, in which we had many discussions on how to arrange things and you will see how we did it with several examples that I will illustrate or try to illustrate today. Remember I have said it already several times, we wanted to have a common set of accessions in different locations and in different years. So, for the perennial crops alfalfa and red clover, the field trials were established early in the project, and they were maintained and phenotyped for two to three seasons. What you need is one seed lot to establish the field trials at the different sites and then do the follow up. But remember we also use the seed lots for experiments in controlled conditions. Using a common seed lot per accession we were sure of the genetic identity of the plants materials that were evaluated in the fields and also in controlled environment experiments. It was important to have control over this. In the annual crops it was even more complicated, in the sense that we had to establish the field trials in at least two different years. If we wanted to combine the data, of different

years, it was necessary also to keep control on the genetic identity of the seed lots that were used, not only in the different locations, but also in the different years in which the field trials were established.



**Requirements**

Common set of accessions in different locations and years  
=> what's in a name?

				
Alfalfa France	Red clover Norway	Pea Serbia	Faba bean Spain	Soybean Belgium

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

Here you have an illustration of what I'm speaking about. This is an example of a field trial in alfalfa in France, red clover in Norway, pea in Serbia, Faba bean in Spain, and soybean in Belgium. You can see large field trials and it is not so simple to coordinate all this work. Constructing a list of accessions to evaluate in the different locations is not sufficient. We need to be sure that what is called alfalfa variety A in France, is genetically the same as in another location let's say in Serbia. So if the sources of the seed are different, we are not sure about the genetic identity.

And that's why at the project preparation step, before beginning, we already defined, and this was a very good choice to work with group experts. So, this is one of the projects, the first one in my experience in which the work packages were important, but the most important role was played by the crop experts. They had a very hard job to do, because they had a very strong coordinating task, to be sure that all the field trials and also that the seed lots and all the preparation steps that were necessary to warrant the quality of our data sets were taken.

## Five crops => five experts



- Seeds centralized at one partner-location, managed and distributed from this location  
⇒ if needed, multiplications ideally at one location (even before start project)
- One unique identifier per accession – linked to databank; no other labels allowed



Faba bean multiplication  
Spain



Horizon 2020 of European Union

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



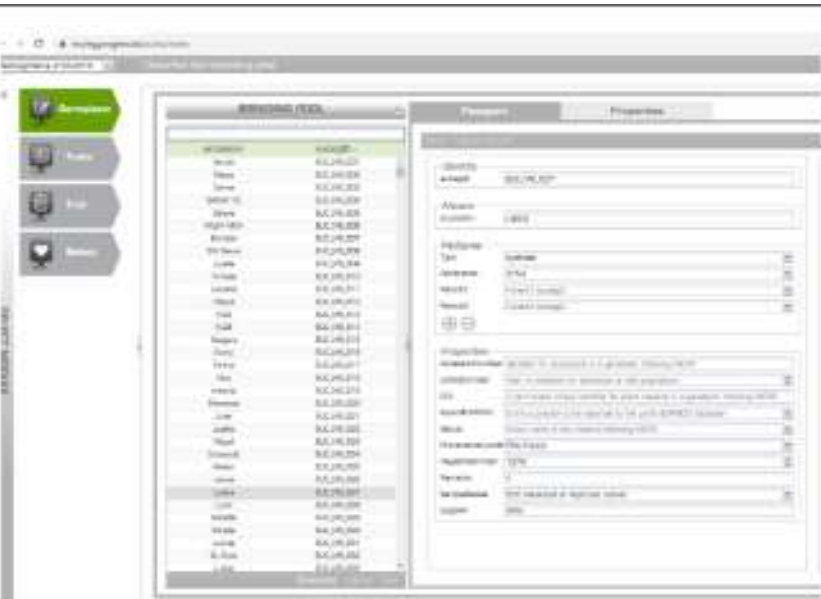
One of the main roles of the crop experts was to share their knowledge about the crops themselves. In addition, the seed lots were centralized at one partner location, and this was the partner location of the crop expert. So, it means that all the seed lots that were collected coming from different sources (other companies, other partner institutes, germplasm collections and many other sources), were centralized at one single location. And from this central location the seeds were distributed to the partners who needed them for field trials and also for the evaluation of other traits in controlled conditions. So, if multiplications were needed (in some crops this was necessary because the seed viability was not sufficient), we avoided combining different sources with different provenances. So, we tried to establish our own collection of seeds to be sure that we were always working with materials that had the same genetic identity. In some cases, we even started with seed collection and multiplication before the project EUCLEG officially got started, because it was for us a very important point to be sure that we could always work with the same or to keep track of the genetic identity of the seed lots that we were using our experiments.

Another very important thing that we did at the beginning of the project was to assign a unique identifier to each accession by the start of the project. This was the only identifier used for labelling, communication and data storage. No other labels were allowed, at least not in communication. So, this was also very important to avoid problems that I have seen occurring in the past.



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

**EUCLEG.eu**



**EUC LEG**

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

This is an example of what I am speaking about. This is one example of the database that we were using for communication and for the storage of all the information of EUCLEG. It is a commercial product, 'Progeno', but it was a very good thing from the beginning up to the end of the project, because here again the crop experts defined for each accession the right identity. So, we had here a full description and unique EUCLEG identifier of the different accessions.

## Five crops => five experts



- Seeds centralized at one partner-location, managed and distributed from this location (multiplications at one location; even before start project) – partner in charge of genetic characterization
- One unique identifier per accession – linked to database; no other labels allowed

⇒ *Sure that we worked exactly with the same genetic materials*  
 ⇒ *Labeling mistakes prevented*  
 ⇒ *One partner responsible for setting up MTAs and other agreements with seed providers*

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



So, we had our common set of accessions and assistance to trace them and to follow up these accessions, but then we needed to establish the field trials. Something that I have seen happening in many projects is that we try to standardize as much as possible, and sometimes we even go too far. For some projects not too much and in other projects it is too much. It is very important to think about what is meant by standardization in the context of a field trial, especially when you are dealing with different locations that have very different environmental conditions. So, remember for some species we were working in South Europe up to Norway. So, what does standardization mean when you are working with locations which are by nature so different in their environment? So, at the start of the project there were also a lot of discussions about which things needed to be standardized and at which level.

## Requirements



Common set of accessions in different locations and years  
=> what's in a name?

Field management  
=> Standardization OK, but to which level?

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

 Activate Windows  
Go to Settings to activate Windows.

So, this is also something, I don't say that was perfect in our case, but it was something that required a lot of discussions and I think that that we found in most cases in the end quite good compromises from participants. So, for example, dealing with fertilisation. So, you can say we will apply the same level of fertilisers in all locations or years. That is an option, but then you can be sure that the fertiliser availability for the crops in different locations will be different, just because you start from different starting points in the soil, because the environmental conditions are different, etc. So, in EUCLEG the approach was different. We started by analyzing the soil characteristics before the field trials were established, and this information was used to determine the doses of fertilisers, according to the practices that are used in the area in which the field trial is established.

## Agreed in advance



- Establishment of field trials  
⇒ soil analyses, weather data, field preparation, sowing density, sowing depth
- Plot dimensions and trial design

	common 4 rep (WP3) & rep (WP4)				common 2 rep				Total accessions	WP3 1 rep				WP4 1 rep				Total # accessions	Total # plots				
	M000	M600	M80	M81- M84	M000	M600	M80	M81- M84		M000	M600	M80	M81- M84	M000	M600	M80	M81- M84		M000	M600	M80	M81- M84	Sum
loc 1: RW06	2	2	2	2	8	8	8	8	32					80	80	80	80	320	108	108	108	108	432
loc 2: ART	2	2	2	2	4	4	4	4	16	24	24	24	24					100	40	40	40	160	
loc 3: R0	2	2	2	2	4	4	4	4	16									100					100
loc 4: AN0	2	2	2	2	8	8	8	8	32					80	80	80	80	320	108	108	108	108	432

Soybean



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

Activate Windows  
Go to Settings to activate Windows.

Another example: remember we were dealing with forage crops, so the mowing regime, the number of cuts and timing of the cuts, that is applied in alfalfa or red Clover in different locations should be different, because the growth path of the crop is different. Apply four compulsory cuts per year in red Clover might not be the optimal solution in some years in some locations. But we tried to standardize as much as we could in other things. So, in sowing densities, sowing depth, if possible, we did it, but also for the plot dimensions on the trial designs. Again, randomization was different in different years, at different locations, but we agreed on which checks to sow in the different locations. This is an example of soybean: you can see here the four locations where field trials were established for work packages 3 and 4. So, two locations were specific for WP3, and two other locations were also used for WP4 package four. This is a complicated scheme, that I will not explain, with the number of plots at each location, but here you see that some accessions were common to all locations and some of them were specific for some of these locations.



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

**EUCLEG.eu**



## Row-column augmented design



33	289	304	119	72	155	30	35	25	129	306	289	157	44	306	76	142	118	272
400	182	274	183	7	317	298	335	88	288	98	114	287	324	71	62	181	89	388
177	283	177	184	14	128	246	373	234	307	238	260	273	748	206	84	263	236	21
260	273	170	184	54	258	175	324	282	87	294	268	84	207	225	297	300	277	343
350	219	62	207	183	308	145	318	245	325	225	333	257	305	185	274	152	204	87
391	242	57	400	286	284	277	543	220	308	249	282	285	0	281	308	284	302	86
11	284	266	311	331	76	241	147	309	307	327	321	280	348	76	237	76	302	284
53	287	144	102	38	260	232	83	34	5	237	320	186	188	230	252	343	304	4
213	225	106	162	165	75	304	368	342	15	333	25	480	245	248	166	362	225	157
258	215	307	8	380	77	43	95	214	6	332	118	122	136	330	33	123	256	25
18	26	13	225	388	213	282	282	317	288	38	312	288	43	337	18	157	287	288
112	274	23	125	188	267	254	424	144	388	223	352	289	218	252	218	125	288	288
228	274	278	184	226	263	63	313	250	324	27	348	397	272	264	126	143	400	218
17	34	162	263	277	82	64	273	236	67	374	24	2	156	226	52	183	232	287
121	307	12	117	107	211	211	133	271	388	240	89	182	338	188	308	87	178	13
110	0	386	287	270	243	181	224	249	41	285	188	285	238	180	284	173	20	80
11	182	238	223	48	262	215	35	289	345	120	211	170	22	134	120	400	237	154
17	282	307	224	280	2	158	402	234	1	66	87	181	13	282	138	164	236	152
81	386	308	36	272	280	18	29	332	82	338	291	291	148	368	21	287	186	18
208	0	308	288	147	288	218	318	326	148	67	85	80	386	226	288	315	212	388
257	283	48	180	221	223	135	303	230	328	317	298	25	221	130	127	25	282	284
352	248	190	287	275	74	236	124	247	143	180	128	176	247	187	217	287	248	282

Faba bean

### Incomplete experimental design

- maximizes the number of accessions to test; ideal for GWAS
- should be planned carefully to achieve high statistical power



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



Because one of the ideas in GWAS is that you try to analyze as many accessions as possible. So, this is the power of your analysis, you rely for the large part on the number of accessions that you include in your analysis. This is always a compromise between the space, the resources that you have for your project (these are always limited) and the quality of your data. That's why in EUCLEG for many of the field trials, we work with what we call row column augmented designs. What is this? These are in fact incomplete experimental designs in which only a subset of accessions that we call the checks, or the reference accessions are replicated several times in the try out. So, this is an example of Faba bean for one specific location. You see that some accessions like accession 400 highlighted above has been replicated several times in the trial but the ones that have no color here were only present one time. So, this was a way that can be used to maximize the number of accessions that you have in your field trial. Thus, we optimized the designs to analyze as many accessions as possible because this is important for association mapping. So of course, this is something that you don't do just by yourself, it is necessary to take into account the advice and the experience of experts on this kind of design and this is also what we did at the start of EUCLEG. So, this is something I think that also worked at EUCLEG, the design of the different field trials.



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

**EUCLEG.eu**



**Agreed in advance**

- Management (plant protection, fertilizer, number of harvests, ...): standard vs local practices

**In 2018:** Cutting regime as locally used to guarantee best establishment of plant stands.

**Number of cuttings per year** (2019 – 2020): 3 to 5 depending on the site and climatic conditions.

An unique cutting date for all plots in one site: 10% flowering of a well-known variety in your condition.  
Record date and stage.

After establishment, no herbicide, no insecticide, no irrigation except if the trial is severely threatened.  
If a treatment is applied, record pest, product and date of application.

After establishment, no fertilization.

Red clover

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

Activate Windows  
Go to Settings to activate Windows.

These are all things that we agreed on in advance. Here I'm just showing an example of red clover to illustrate what I mean. You see in the slide a number of things, the way in which things were discussed and agreed upon to do:

- The number of cuttings: three to five depending on the site and climatic conditions, but we defined as much as possible, how to determine the cutting dates or how to determine the moment at that this should be done.
- We agreed that after establishment, no insecticide and no irrigation would be applied; only if the trial was severely threatened.

This is just an illustration for one of the crops. The same was done for all the crops, I'm just using examples here.

## Requirements



Common set of accessions in different locations and years

=> what's in a name?

Field management

=> Standardization OK, but to which level?

Consistency of data over years and locations

=> clear definition of variables

=> biochemical analyses centralized per crop



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

Activate Wind  
for the future

But then we also look at the evaluation of the different traits, because sometimes that is forgotten. We say OK we will evaluate yield, or we will evaluate protein content of the cuts or the protein content of the seeds, in the case of grain legumes. But then each partner carries out his or her own biochemical analysis. If you take this choice from the beginning, you can be sure that it will be difficult to combine the results from the different partners, because different labs deliver data that might not be easy to combine because of the use of different instruments, slightly different protocols and sometimes just because sometimes the lab has an influence on the outcome of your analysis. And that's why we decided also from the beginning to centralize all analyses of each crop as a single location. I told you we centralized all the management of the seed lots, so the distribution of the seed lots was done by one partner lab. The same thing was passed on, so the biochemical analysis of all the samples, if even if they were collected in five different locations, were done by one partner. This was quite a lot of logistic work, so a lot of effort, but at the end we are very happy that we did it in that way, because we are sure that we are comparing things that are comparable.



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

**EUCLEG.eu**

## Agreed in advance



- Overview of traits to score with detailed description in phenotyping protocol
  - Establishment and development (V and R stages, canopy closure, ...)
  - Disease incidence
  - Abiotic stress
  - ....
- Harvest (e.g. soybean)
  - Yield (kg, #seeds, seed irregularities)
  - Other morphological characteristics of subset of plants (number of pods, number of branches, ...)
  - Thousand seed weight
  - Subsample for quality parameters

+ Workshops to discuss and check that everything is clear to all partners



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

Activate Windows  
Go to Settings to activate Windows.

Using some examples, we prepared from the start of the projects an overview of the traits that we wanted to evaluate; not only the name, but we also provided very detailed descriptions of the phenotyping protocol. This described everything, from establishment, developmental stages, how to determine canopy closure, canopy development, canopy height, etc.

## Example soybean



- |                               |   |
|-------------------------------|---|
| 1. Plant emergence            | 13. Protein content & composition (ILVO)      |
| 2. Plant vigour at emergence  | 14. Seed weight                               |
| 3. V2 stage                   | 15. Plant length                              |
| 4. Plant height at V2         | 16. Height first pod                          |
| 5. R1 stage (start flowering) | 17. Mottled seeds                             |
| 6. R2 stage (full flowering)  | 18. Node number on the main stem of the plant |
| 7. Diseases, pests            | 19. Number of branches with pods              |
| 8. Abiotic stress             | 20. Distribution score                        |
| 9. R8 stage (maturity)        | 21. Seed number per plant                     |
| 10. Lodging at R8             |   |
| 11. Seed yield                |   |
| 12. Moisture content          |   |



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

Activate Windows  
Go to Settings to activate Windows.

This is an example for soybean, not just yield but also about other morphological characteristics that were determined during the winter in the lab. How to determine the thousands seed weight, or how to take a subsample to determine the quality parameters, because it is not only important that the quality



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

**EUCLEG.eu**

parameters for biochemical analysis are done at one central location, it is also important that the sub samples that are taken for that are representative and are comparable. So, we had many workshops to discuss and check that everything was clear to all the partners.



**Phenotyping protocols**

One version ready before start of experiment – preparation led by crop expert in close collaboration with all partners involved

Revision after first year

- Traits difficult / impossible to record?
- Adjustments of protocol when needed

Fig. 1. Parts of a soybean plant of the commercial cultivar...

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

Active Wind  
Go to Settings for a...

At the end we defined a very long list of traits to investigate. This is just an example for soybean, and you see 21 with detailed descriptions that were described.

Phenotyping protocols describing all these aspects were made available through a central communication system to all the partners, but specifically to the ones that were involved in the evaluation of the corresponding crop. So we had already one version before the start of the experiment, before the measurements needed to be taken. I come back to the important role of the crop experts, because they led in the preparation of this phenotyping protocol in collaboration with all the partners involved. Here I show an example of a detailed description, because the protocols are full of illustrations of how to determine particular traits. These phenotyping protocols can also be relevant for the future for other projects in which these crops are investigated. Of course, after one year of experience, there was an evaluation of the different protocols and if necessary, some traits were revised, or partners made new agreements on this.

Again, very important, not only was the phenotyping protocol very well defined, but also in our Progeno database the variables were clearly defined. The crop experts defined the variables and it was not possible for other partners to change the definition of the variables. This forced everybody to work on the same terms, and here you see an example of several variables defined not only with the name but also with a description. Units that should be used and also which values are allowed was important also to prevent the occurrence of mistakes. It was very interesting as well that we had a team of people that helped to import all the data sets into the database in a semi-automatic way.

## Conclusions



### EUCLEG strengths

- Common set of accessions per crop (not only for field trials, but for all experiments in Europe)
- Data organized in central database
- Strongly coordinated work

### Huge scientific and **LOGISTICS** work

**BUT**, datasets of high quality constructed, also for future work

=> illustrations in this workshop



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

20  
Activate Wind  
for the Settings of a

I come now to the discussion and the conclusions or what the main message of what I wanted to say is. I hope that I have convinced you about some of the strengths of our project. Things that I certainly would like to takeover to other projects in the future. So, it was very interesting, and I think that is one of the main strengths of this project, at least for the European experiments (the idea was to do it also at the Chinese level, I mean Europe and China together, but from the beginning it became clear that was very difficult to exchange plant materials between the two continents). But at least we succeeded in setting up very well coordinated experiments within EUCLEG for all the European partners. And using a common set of accessions per crop, it was very interesting, not only for the field trials but also for the other experiments in more control conditions, because now we have a very in-depth description of all the accessions that we have investigated in EUCLEG.

It was very good to get all the data organized in one central database, but you should do it from the beginning and organise that from the beginning not at the end, as I have seen happening in several other projects.

It was important to coordinate the work. I was work package 4 leader, and I thought this would be a huge amount of work. In fact, a huge work was for the crop leaders and not for the work package leaders in my experience, at least not for me in this project. So, this two-dimensional work for work packages and crop leaders was a very good choice.

EUCLEG was a huge scientific and logistical undertaking, but the data sets are very high quality, also for future work and this will be illustrated in this workshop today and tomorrow by my colleagues. Thank you very much for listening.



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

**EUCLEG.eu**

Thank you very much Isabel, that was an excellent presentation and very timely for new projects that may be starting out, to learn the lessons that you have very clearly put in that presentation.

There is one question about the numbers of accessions that were used, and you mentioned that there were subsets of the whole collection of crops that were tested at additional geographical locations. So how did you decide the numbers that were used and what numbers were used?

I will take the example of soybean, to this table that I said I will not explain that, so you see here the complicated scheme, so we have a common set of accessions and in soybean it was complicated, because we had different maturity groups. You have accessions ranging from maturity groups 000 to I to II. So, this was also a complicating factor for the design of the experiments, because for practical reasons it was sometimes necessary to not have a complete randomization of the field trials, but to organise the accessions into subset even at a field level, but also when looking at the locations it made absolutely no sense to put some accessions at some locations. And this is what you see here, so for WP4, indeed we had common set of accessions, but not for WP3. So, when you will hear the results of our WP 4, we are always working with a common set of accessions. So, this is the total number that you get here, but when speaking about WP3, there were at least 100 accessions that were common to all the locations, and these are the focus of work package 3. Important to say that you can analyze all the data together, because it was the same (multi-location, multi-year) trial. I hope that this answered the questions.

Another question that's come up is, you stressed the importance of the consistency of data over years and locations, so was there an excel file or something equivalent distributed to all partners with details of the phenotyping?

Yes. So, we had phenotyping protocols. These were word documents and pdfs with the description. Then we have the Progeno database that was established quite soon after the start of the project, with the description of the variables. And for communication purposes and for export and import of data and to manage the work at the different partner locations, there were templates created in Excel and which the partners could easily input the data and then make them ready to be imported into the database.

## About the author

**Dr Isabel Roldan-Ruiz**, Scientific Director at Institute for Agricultural, Fisheries and Food Research ILVO, Plant Sciences Unit in Melle Belgium from 2008 and Professor at Ghent University, Department of Plant Biotechnology and bioinformatics from 2017. She coordinates research of agronomists, plant geneticists, breeders, ecophysiologicalists and modellers with a main focus on grassland grass and clover species, and protein crops such as soybean. Today her main research area is molecular plant breeding. Isabel Roldan-Ruiz is leading EUCLEG WP4.

**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 30<sup>th</sup> September and 1<sup>st</sup> October 2021**

Recording link to the presentation: <https://youtu.be/pyKe069u-ng>



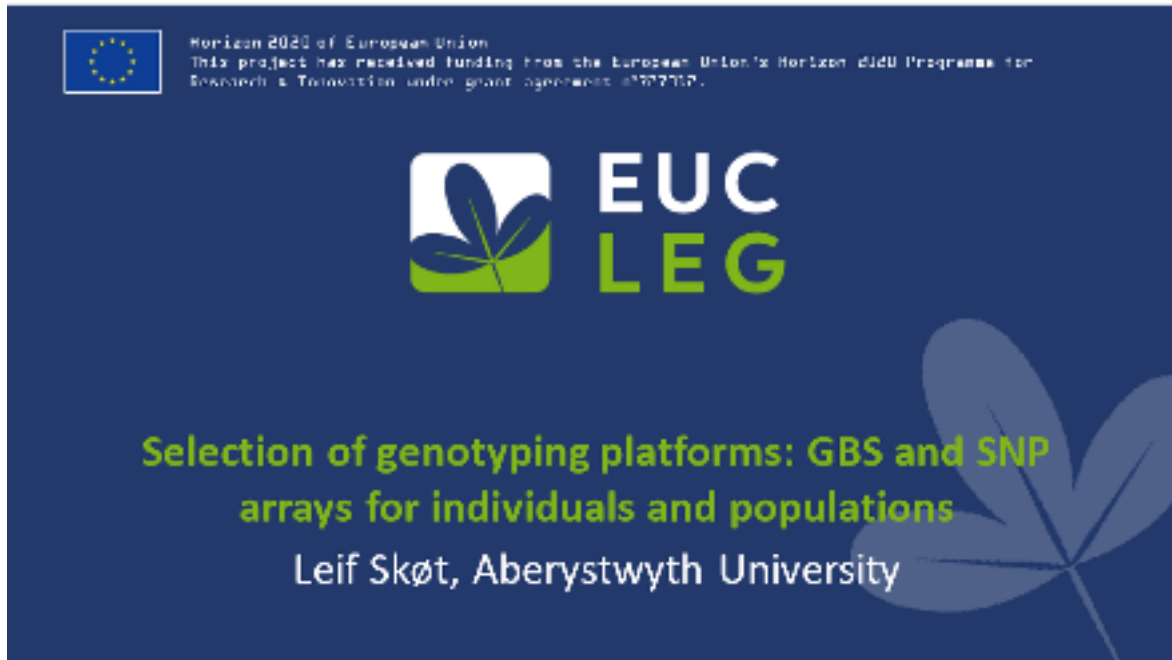
This project has received funding from the European Union's Horizon 2020 Programme for Research & Innovation under grant agreement n°727312.

**EUCLEG.eu**

### 3. Selection of genotyping platforms: GBS and SNP arrays for individuals and populations

Leif Skot

Emeritus Professor, IBERS, Aberystwyth University, U.K.



#### Nucleotide polymorphisms



Base substitutions – one or more bases of DNA get changed

Individual 1: GATTACCGTAATC

Individual 2: GATGACCAACAATC

Insertions or deletions – one or more (sometimes hundreds!) of bases are added or removed from a stretch of DNA

Individual 1: GATTACCGTAATC

Individual 2: GATTACCAATGTAATC

Individual 3: GATCCGTAATC



This chapter concerns the genotyping platforms that are especially relevant for the EUCLEG project. All genotyping nowadays start with single nucleotide polymorphisms as the basis for all genotyping. As you can



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

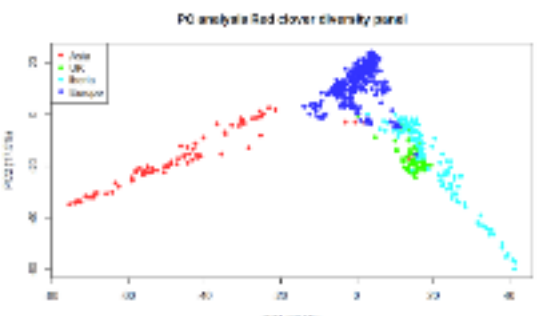
**EUCLEG.eu**



see here, it is either a change in a base at certain positions or it can also be insertions of some sequences or deletions of them.

**Why are we interested in genotyping in crops?**

**Genetic diversity in germplasm**



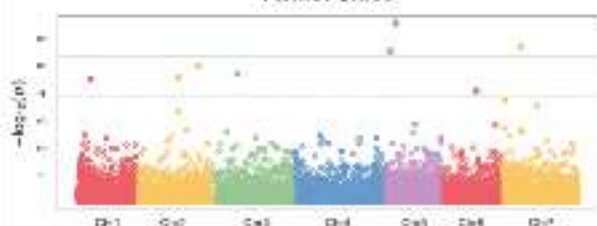
- Genetic variation is the basis for breeding
- Source of novel variation
- Climate resilience
- Pest and disease resistance
- Heterosis
- Transgressive segregation

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

So why are we interested in genotyping in crops? Firstly, it is to do with identifying genetic diversity in germplasm. Because genetic variation is the basis for breeding, without that you might as well give up go home. Secondly you can measure the source or you can get sources of new novel variation, that you need to improve your breeding programme or introduce new traits. It is also important in terms of improving varieties resilience to climate change, and of course pest and disease resistance is important. It could be used also in determining which parents to use to improve or to get heterosis and identify parents that may give transgressive segregation, traits that are beyond what was there in the parents.

**Why are we interested in genotyping in crops?**

**Genome wide association studies (GWAS)**



- Associate genetic markers with phenotype
- Identification of QTL
- Genetic basis of phenotype
- Identification of markers for breeding

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

The other reason why we are interested in genotyping is for genome wide association studies (GWAS) to associate genetic markers with phenotype. You want to identify QTL and it also gives you the opportunity



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

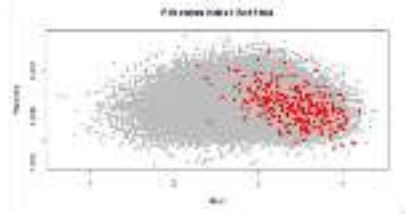
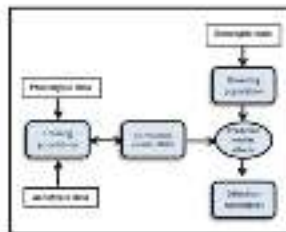
of diving into the genetic basis of certain phenotype that you're interested in and to identify markers for breeding i.e., for marker assisted selection and so on.

## Why are we interested in genotyping in crops?



### Genomic selection

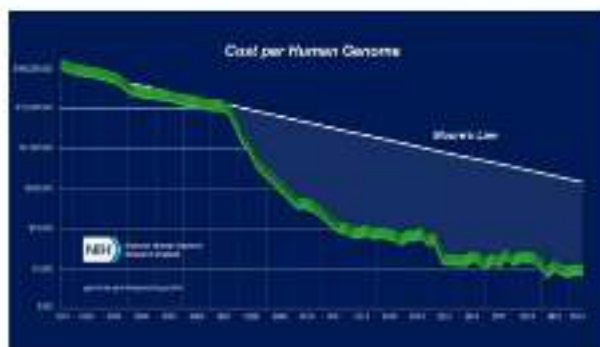
- Identify breeding value based on all marker effects
- Predict the best candidates for crossing
- Shorten the breeding cycle



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

Finally, another reason is for genomic selection, where you identify breeding values of genotypes or populations based on all the marker effects identified throughout the genome. You can then predict the best candidates for crossing, based on the estimated breeding value and thereby try to shorten the breeding cycle and thereby save time and money in the breeding programme.

## Why are we interested in genotyping in crops?



<http://www.genome.gov/about-genomics/fact-sheets/Sequencing-Human-Genome-cost>

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

The final reason I wanted to mention is of course the price. I am sure many of you have seen this slide before. Since 2007 when next generation sequencing was introduced and really started to gain in



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

**EUCLEG.eu**

popularity, there's been a dramatic drop, until you reach the \$1000 per genome, for the human genome it seems to have flattened out since then, but maybe that's just because this was the goal that people were after.

## Current genotyping platforms



- Whole genome resequencing
- SNP arrays
- Reduced representation sequencing



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

The first genotyping platform that I'm going to mention is whole genome resequencing although this wasn't used in the EUCLEG project. Then there are also SNP arrays of which you have seen examples of three crops yesterday. Finally, reduced representation sequencing, which is what was used for genotyping in red clover and alfalfa

## Considerations in choosing genotyping platform



- Linkage disequilibrium
- Inbreeding or outbreeding lifecycle of crop
- Price per sample
- Ploidy



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

Some of the considerations for which genotyping platform to use are listed here, for example Linkage disequilibrium. This is important because it determines how many markers you are going to need. Linkage disequilibrium is quite often linked quite closely to whether you have an inbreeding or outbreeding crop. We have examples of the two types in the five crops that we have been working with here in the EUCLEG

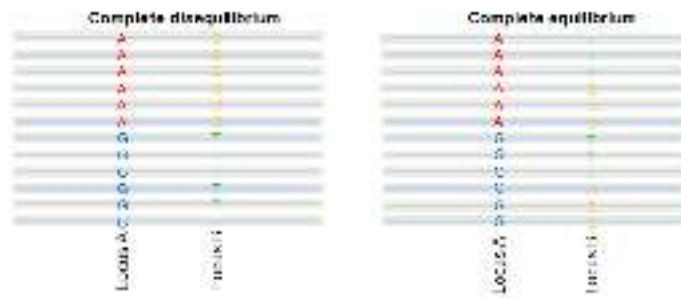


This project has received funding from the European Union's Horizon 2020 Programme for Research & Innovation under grant agreement n°727312.

**EUCLEG.eu**

project. Price per sample is very important for everybody. And finally, ploidy as many of the important crops that we are working with are polyploid and this makes things a little more complicated usually.

## What is linkage disequilibrium?



$$D = p_{AB} - p_A \times p_B$$

I will go back to basics a little bit, on the left you can see an example of complete disequilibrium. You have a number of genomes illustrated there, and there are two loci where there is polymorphism. But if you look at locus A there is polymorphism between A and G, and locus B, C and T. but no matter which of the haplotypes or which of the genome's you choose, which have A at locus A you also have C at locus B, so the probability of finding C at the locus B, provided you have A and locus A is one. And on the right-hand side you will see examples of complete equilibrium, where there's a 50:50 chance of finding the T or a C given that you have A at locus A and likewise if you have G at locus A there's a 50:50 chance. So, the two loci are independently segregating.

### Measures of LD

$$D = p_{AB} - p_A \times p_B$$

- D is dependent upon allele frequencies
- We wish to normalize the expression (0 – 1)

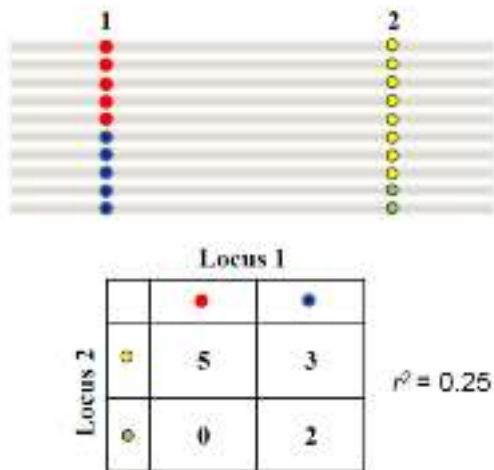
$$r^2 = D^2 / (p_{A_1} p_{B_1} p_{A_2} p_{B_2})$$

The basic measure of this linkage disequilibrium is at the top of the next slide. What is the probability of finding A and B together, minus the multiplication of the probabilities of each of the two allele frequencies. So, if they were independent it's the product of the two, but that measure depends on allele frequencies. So there was a wish to normalise the expression, so it varies between zero and one. That resulted in the



equation at the bottom, which is the one that almost exclusively used or certainly in the presentation series in EUCLEG, which is sort of the correlation between LD and the each of the four possible allele frequencies.

## $r^2$ in association mapping



Example: Marker 1 is a QTL explaining 10% of total Phenotypic variance, so it is not the causative SNP

SNP marker 2 explains 25% ( $r^2 = 0.25$ ), but it only Explains 2.5% of total phenotypic variance.

This requires large sample sizes for sufficient power. Or higher value of  $r^2$

An example here is where you have partial disequilibrium between 2 loci, so let's say that marker one is a QTL, and that explains 10% of the phenotype. So it may not be the causative SNP. So SNP marker 2, which maybe the one that you have in your genotype platform that explains 25%. The linkage disequilibrium is 0.25, but it only explains 2.5 percent of the total phenotypic value, because that's only explained by 10% at locus one. So it requires quite a large sample size to gain sufficient power or higher value of linkage disequilibrium. That's possibly why there have been so many attempts or attempts that have not succeeded in finding single markers that can be used in marker assisted selection for some of their complex traits that are really important in breeding programmes. This is because it requires a really high density of markers.

## Polyploidy

### Allopolyploidy:

The combination of genomes from two or more related species through hybridisation and subsequent chromosome doubling (Bread wheat, White clover). Behave like diploids

### Autopolyploidy:

The failure of meiosis or mitosis, and the fusion of unreduced gametes (Potato, Alfalfa)

Paleopolyploids are formed at least several million years ago with ancient and diploidized genomes (Maize, Soybean)



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

12

Now I am going to look at polyploidy, which is also an important factor in choosing a genotyping platform. There are two types: Allopolyploidy, this the combination of genomes from two related species and through some sort of hybridization and subsequent chromosome doubling. Examples of that include bread wheat and white Clover. They behave like diploids when it comes to assortment of chromatids.

## Polyploidy

- Impacts on genotyping
- How to tell the difference between allelic SNPs and homeologous SNPs
- Competition for primers in allele-specific primer methodology
- In NGS technology read depth needs to be larger in polyploids
  - In diploids 7.7x coverage needed for sequencing of both alleles 99% of the time
  - In polyploids 50x to 100x needed

Clewley M. et al. (2015) Mol. Plant 8:833-846. <https://doi.org/10.1016/j.molp.2015.02.002>



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

13

The other one is autopolyploidy and that's basically when there's a chromosome doubling going on, as a result failure of meiosis or mitosis. Examples of that are potato and alfalfa. Paleopolyploids are formed a



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

long time ago and are ancient polyploid and they have a diploidised over the millennia and examples of that are maize and soybean.

So the way in which it impacts on genotyping, is how do we tell the difference between allelic SNPs and homeologous SNPs on the other genomes in the crop, and there is also the competition for primers, when you talk about allele specific primer methodology, as you do in the array technology. In next generation technology such as the reduced representation sequencing that I will talk about shortly. You need a larger coverage compared to this- these are numbers I've taken from the paper quoted at the bottom of the slide. They mentioned 7.7 times coverage, which is the number of times a certain base is sequenced in the sequencing effort. In diploids this coverage is needed for sequencing of both alleles 99% of the time, and in polyploids that number goes up considerably, as you can see. It becomes more complicated technically, but also much more expensive.

## Polyploidy



- Divergence between subgenomes in allopolyploids is important
- If more than 2% divergence read mapping can be done with less than 2% mismatches per read.



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

34

This is mostly for example in the allopolyploids where there can be a certain amount of divergence between the two genomes that merged and for example if more than 2% divergence is present in the two genomes. If there is a tetraploid for example then the read mapping can be done with less than 2% mismatches per read. This is the sort of a theoretical calculation that you can do when you plan your sequencing or genotyping.



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

**EUCLEG.eu**

## Whole genome resequencing



- Price is still an important issue
- Sequence coverage needed can vary from 1x in inbred species to 15x in highly heterozygous species, or >30x in pooled populations
- If high coverage is needed and the genome is large, the cost for a GWAS panel of a few hundred can be several hundred thousand \$.

Powell et al. *Front. Genet.*, 05, 2020 | <https://doi.org/10.3389/fgene.2020.00444>



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

14

Now I'll mention one of the first of the three platforms. They all use next generation technology. Whole genome resequencing, as the name suggests you basically resequence the genomes of a whole panel of genotypes of populations that you are interested in characterising. It is especially useful for species where this linkage disequilibrium (LD) decays very, very rapidly, such as in maize, olive, perennial ryegrasses, red clover and so on. So for many of our breeding species that is the case, because then when you do whole genome sequencing in theory you get every single nucleotide polymorphism that are there. So you will have a SNP, the SNPs very close together, so have a high chance of being in big LD with your trait of interest. There is a wide range of variation between species in terms of LD. It can go down to 25 bases in olive, up to 8 megabases in wheat and in the 5 species we are working on somewhere in between that.

## Whole genome resequencing



- Price is still an important issue
- Sequence coverage needed can vary from 1x in inbred species to 15x in highly heterozygous species, or >30x in pooled populations
- If high coverage is needed and the genome is large, the cost for a GWAS panel of a few hundred can be several hundred thousand \$.

Powell et al. *Front. Genet.*, 05, 2020 | <https://doi.org/10.3389/fgene.2020.00444>



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

15



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

**EUCLEG.eu**



Price is an important issue for genome resequencing, and it can get quite expensive if you have large panels and high coverage needed. For example choose 1-x coverage in inbred species, which might be a possibility and if you have a reference genome, you can impute missing data, but then you need 5 times, 15 times or even 30 times in pooled populations of tetraploid species such as alfalfa. But if you need high coverage and the genome is large, so the cost of GWAS panel for example of a few 100 genotypes or samples can be several \$100,000. So it can escalate quite rapidly with whole genome resequencing. Nevertheless, this strategy is used more and more in many of the major crop species.

## SNP arrays



- Affymetrix and Illumina are the main manufacturers
- 46 arrays
- 25 crop species
- From 3K to 820K markers

Combeiro, (2017), *Adv. Plant. 10*, 1047–1054. doi: 10.1016/j.advpl.2017.03.008



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

37

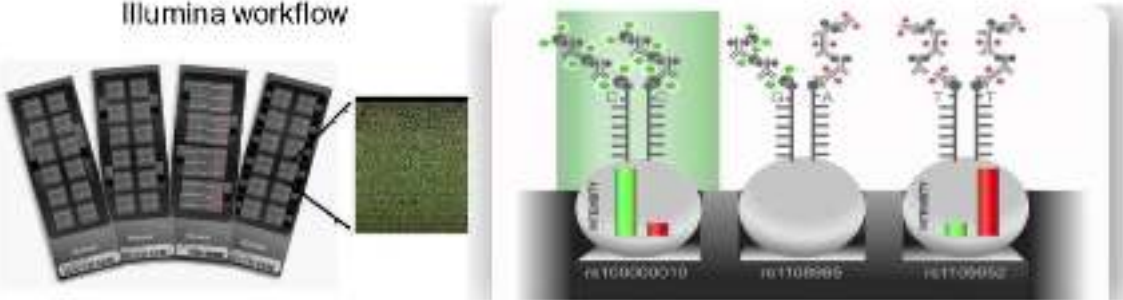
The next platform which was used in the species that were talked about in detail yesterday. Affymetrix and Illumina are the companies that are the main manufacturers of these SNP arrays. The numbers are taken from the paper quoted, so the numbers may have changed slightly, but at that time there were 46 arrays available in 25 crop species and they varied in size from 3000 markers to up to nearly a million markers in certain species.



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

**EUCLEG.eu**

**Illumina workflow**



Each chip does 12-24 samples across 3000 to 1 million SNPs per sample

Chip surface covered in DNA probes next to known SNPs in that genome

Sample DNA binds to probes and is used as a template to copy. When the SNP is added, the different bases are labelled with different dyes.

So, if the SNP is A or G, A is labelled green and G is labelled red.

Computer scans and records colour for each probe: Green = AA, Red = GG, Yellow = AG

Contributed by Mark Hildebrand

This is an example of the Illumina workflow. Each chip shown on the left can do 12 to 24 samples with anything varying from 3000 to 1,000,000 SNPs per sample. The chip surface is covered in DNA probes, which are designed based on known SNPs and the surrounding sequences in that genome sample. And then the sample DNA binds to the probes and used as a template to copy. The probe goes right up to where the SNP is, then you add the SNP. The different bases are labelled with different dyes, so the SNP is A or G. A is labelled green and G is labelled red and then the computer scans and records the colours for each of the probes. So you get green for AA, red for GG and yellow for AG, which is heterozygote. So that's a very simplified explaining how a SNP chip or SNP array works.

One of the advantages of using a SNP array is that it is quite an accurate method of identifying SNPs in your panel of samples and subsequent bioinformatics is relatively straight forward. That's my experience. There are some disadvantages, the probes that you use to hybridise to your DNA sample are based on sequences that you already know from the panel you used initially to find your SNPs and that is a finite number of samples. So your SNP array suffers from what is called ascertainment bias, so you will never discover any novel information that is not present in that initial panel where you identified your SNPs. And you also need some knowledge, sequence information. You need knowledge of where the SNPs are in the surrounding sequence. And you need to genotype many individuals to obtain allele frequencies in highly heterozygous species. So, these are some of the pros and cons in SNP arrays.

## SNP arrays

### Advantages

- Accurate calling
- Bioinformatics relatively straightforward

### Disadvantages

- Ascertainment bias (no discovery of novel information)
- Prior sequence information required
- Many individuals need to be genotyped to obtain allele frequencies in highly heterozygous species



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

19

I will now go on to say some words about reduced representation sequencing, which was a technology or methodology used here in EUCLEG projects, in the two forage species alfalfa and red clover. There are three types, two of which are really important and they are very closely related. The genotyping by sequencing (GbS), which was first introduced by Elshire et al 2011, and RADseq which goes back to 2008, but they're very closely related and I will come back to this. There is a third method, which I feel I ought to mention, and that is DARTseq, which is based on a library that you make of genome sequences or fragments of your panel and then you hybridise to make a chip or an array as well. But that suffers from not being very high throughput, so I think there's less and less usage of that compared to the next Gen sequencing that I will go onto now.

## Genotyping by sequencing

### Overview of the technology

- Sequencing only from restriction sites
- SNP discovery and genotyping simultaneously
- No DNA size fractionation (unlike RADseq)
- Presence/absence of restriction site
- Small insertions/deletions



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

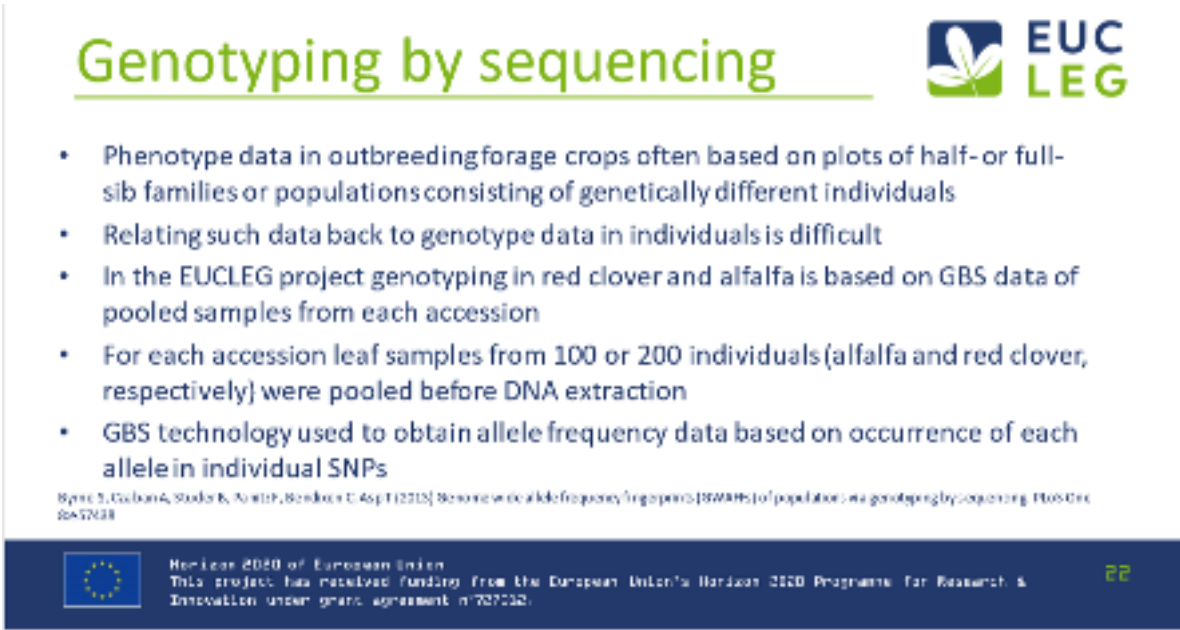
20




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

**EUCLEG.eu**

Genotyping by sequencing means you reduce the amount of sequencing you do, by cutting your genome with restriction enzymes, which recognise very specific small sequences and then you get some fragments and you sequence little bit from each of those fragment ends. Then you can detect SNPs and you do genotyping simultaneously, so you don't need any prior sequence information, and for genotype sequencing, the difference to RADseq, as far as I understand it, is mainly that in RADseq, in the traditional one, you normally include a size fractionation step which you don't in the GbS published by Elshire. It will also detect presence or absence of restriction sites and small insertions and deletions as well.



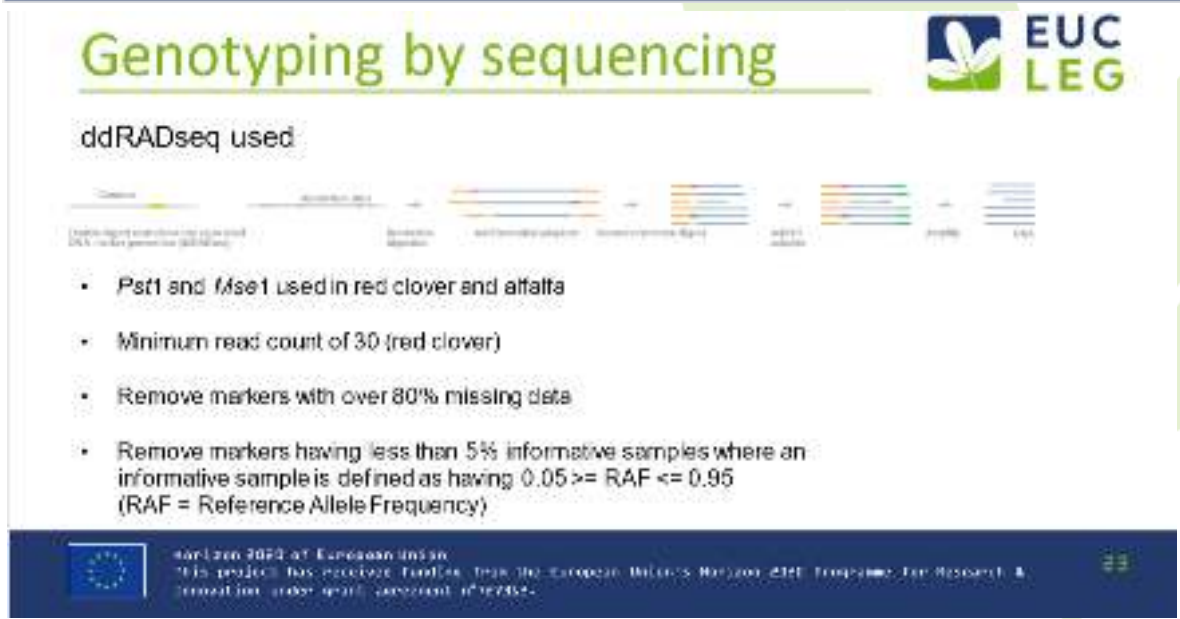
## Genotyping by sequencing




- Phenotype data in outbreeding forage crops often based on plots of half- or full-sib families or populations consisting of genetically different individuals
- Relating such data back to genotype data in individuals is difficult
- In the EUCLEG project genotyping in red clover and alfalfa is based on GBS data of pooled samples from each accession
- For each accession leaf samples from 100 or 200 individuals (alfalfa and red clover, respectively) were pooled before DNA extraction
- GBS technology used to obtain allele frequency data based on occurrence of each allele in individual SNPs

Byrne S, Galbraith A, Studer B, Forthofer B, Baskin C (4 Sept 2018) Genome-wide allele frequency (genotype) (GWAFs) of populations via genotyping by sequencing. H2020 845723


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



## Genotyping by sequencing



### ddRADseq used



- *PstI* and *MseI* used in red clover and alfalfa
- Minimum read count of 30 (red clover)
- Remove markers with over 80% missing data
- Remove markers having less than 5% informative samples where an informative sample is defined as having  $0.05 \geq \text{RAF} \leq 0.95$  (RAF = Reference Allele Frequency)

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

That's all very well and you can do genotyping by sequencing on individual samples, as has been done in countless of examples. However, when you work with out-breeding forage crops, you often base your

phenotypic data on plot trials of full or half sib families or populations, each consisting of genetically different individuals. So relating such plot related data back to genotypic data on individuals is difficult. In the EUCLEG project, the genotyping in alfalfa and red clover is therefore based on genotyping by sequencing technology of pooled samples from each of the accessions, that we are including in the panel. For each accession leaf samples from 100 or 200 individuals in alfalfa and red clover respectively, because you don't need as many pooled samples in alfalfa because it's tetraploid and you get two genomes for the price of one, whereas red clover is diploid. So those numbers of individuals were pooled before DNA extraction, and the GBS technologies is used to obtain allele frequency data based on the occurrence of each allele in individual SNPs. That was the rationale for using pooled data.

## Alfalfa genotype data



- 1016 accessions sampled and genotyped
- Samples from each of 100 plants per accession were pooled and sequenced
- At least 9 million reads per sample
- Various filtering processes resulted in 227,092 SNPs available for analysis

### Reference sequenced:

Chen, H., Zeng, Y., Yang, Y. et al. Allele-aware chromosome-level genome assembly and efficient transgene-free genome editing for the autotetraploid cultivated alfalfa. *Nat Commun* 11, 2484 (2020). <https://doi.org/10.1038/s41467-020-19338-x>



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

24

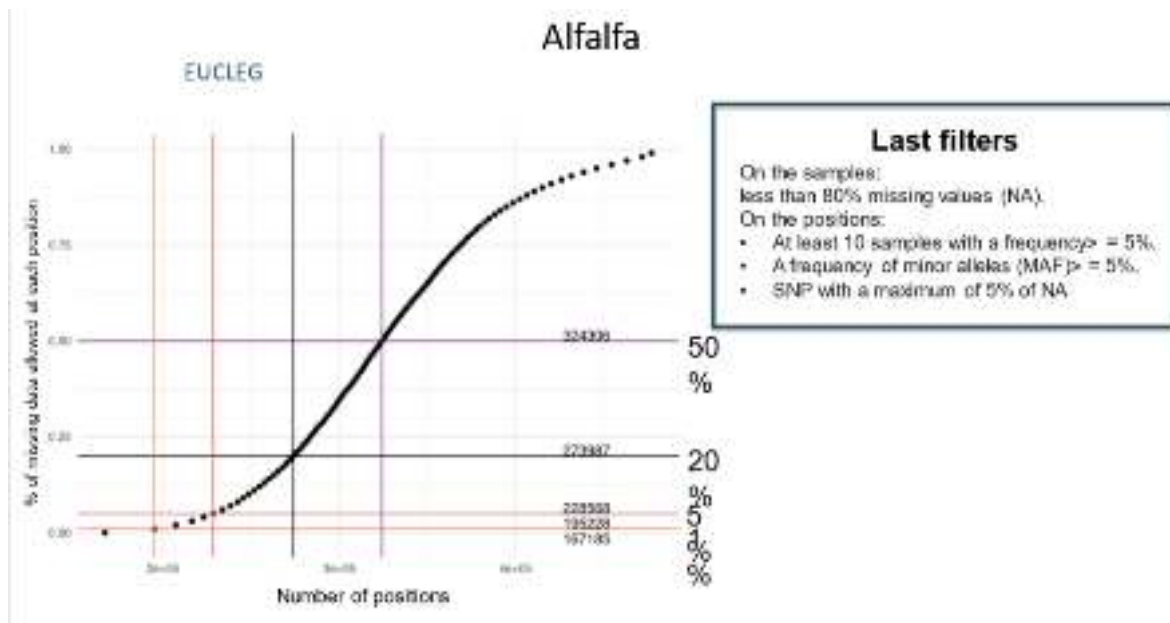
This is a brief description of what we did with red Clover and also with alfalfa. So you have your genome and you cut with restriction enzymes, and after pilot experiments the enzyme Pst1 and Mse1 were used in both species. There's a minimum read count of 30, certainly in red clover. I think it was reduced to 27. You remove markers with over 80% missing data and then you also remove markers having less than 5% informative samples. Such samples are defined as having a reference allele frequency from between 5 and 95%. So basically you cut with your restriction enzymes, you add some barcoded adapters, so you can multiplex your genotyping effort and you introduce a second restriction digest and then you add barcode adapters to those sites and then you amplify your DNA. I should say that in the case of red Clover Tom Ruttink from ILVO developed a bioinformatics pipeline that was used here, so he is really the person who has dealt with the bioinformatics after we had the samples sequenced. So we have a lot to thank him for certainly in red clover. In alfalfa it was the people from INRAE, Philippe Barre, Bernadette Julier and Marie Pegard. So for alfalfa genotype data, which you will hear much more about later, 1016 samples were collected and then were genotyped with 100 plants from each of those accessions were pooled and sequenced. At least 9,000,000 reads per sample were secured and after a number of filtering processes



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

**EUCLEG.eu**

over 200,000 markers resulted that are available in the in the Progeno database platform that everybody has used.



This slide describes some of the last filters that were used. For example, you want less than 80% missing values on the samples, and on the positions you want at least 10 samples with a frequency of at least 5%, and a frequency of minor alleles of at least 5% and SNPs with a maximum of 5% missing data.

## Red clover genotype data

- 12,251 GBS-tag loci are present in over 600 of the 641 samples or populations
- Those loci are used to identify polymorphic sites and extract SNP allele frequencies.
- Average distance between neighbouring GBS-tag loci is 34Kb. 3.7 SNP markers per GBS-tag on average reveal high levels of sequence diversity (1 SNP / 40 bp) across the populations
- 65014 SNP available for analysis

### Reference genome

De Vega et al. (2016) Red clover (*Trifolium pratense* L.) draft genome provides a platform for trait improvement. *Sci. Rep.* 6, e17384

### Evolutionary analysis

Ramirez-Parra et al. (2015) Pleistocene climatic changes, and not agricultural spread, accounts for range expansion and admixture in the dominant grassland species *Lolium perenne* (L.) J. *Biogeography* 48: 1461-1485



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

24

For the Red Clover genotype data we didn't get quite as many markers as they did in alfalfa. There were 12,251 GBS tag loci identified that were present in over 600 of the 641 samples that were used in red



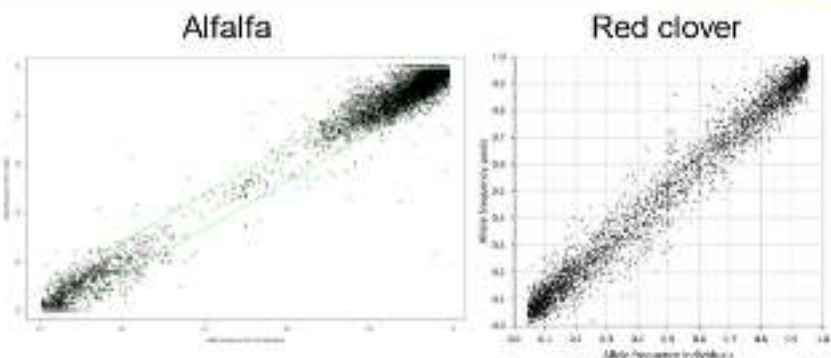
This project has received funding from the European Union's Horizon 2020 Programme for Research & Innovation under grant agreement n°727312.

**EUCLEG.eu**

Clover. Those loci were used to identify polymorphic sites and to get the SNP allele frequencies from those. It was found that the average distance between neighbouring GBS tag loci was 34 KB and there were 3.7 SNP markers per GBS tag on average. This is evidence of high sequence diversity, basically one SNP per 40 base pairs across the populations. All in all, over 65,000 SNPs are available in the ProGene database for our colleagues in the EUCLEG consortium to work with.

## Alfalfa and red clover

Correspondance between allele frequency pools and allele frequency based on data from individuals



Figuer et al. (2019), Allele-Frequency Changes Provide Evidence for Selection and Identification of Candidate Loci for Survival in Red Clover. *Evolutionary Bioinformatics Online*, 14(1), 1-11.



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

27

Theoretically you could genotype each individual, each of the 200 individuals or 100 individuals within whatever species you're talking about and get very accurate allele frequencies in the population. When you do this pooled strategy, you do lose a little bit of accuracy, but as you can see here there is quite good correspondence between allele frequencies, as measured when using individual samples compared to the pooled samples.



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

## Vicia faba genotype data



- DNA from a set of 400 faba bean accessions was sent to the University of Reading, UK, for genotyping. The faba bean 50K Axiom Array (Affymetrix) was used (Donal O'Sullivan pers. Comm.)
- 34,354 SNPs identified
- 11,387 (33.4 %) out of the 34,354 genes analysed were monomorphic (non informative).
- The final score matrix include 352 samples with 22,867 SNPs



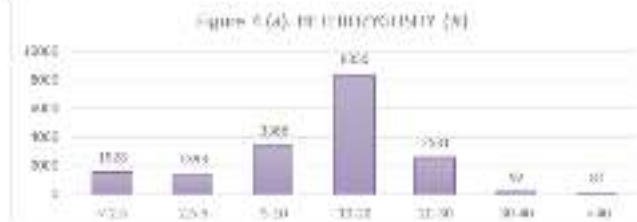
Horizon 2020 of European Union  
This project has received funding from the European Union's Horizon 2020 Programme for Research & Innovation under grant agreement n°727312.

28

## Vicia faba genotype data



- Significant heterozygosity shows outbreeding to some extent
- A GBS strategy was initially selected for this species, but as a SNP array became available in time and within our budget it was considered a superior option



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

29

Now I will go onto three of the crops where we use SNP arrays. There were 400 faba bean accessions genotyped with a 50K Axiom array with over 34,354 SNPs identified. The final score matrix included 352 samples with over 22,000 SNPs. There was some heterozygosity in some of the samples, and initially there was a GBS strategy selected for this species, but a SNP array became available and was considered a superior option so that was used.



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

**EUCLEG.eu**



## Pea

- 260 accessions were sampled
- A 13K SNP array was used for genotyping
- R package “Argyle” (Morgan) used for initial QC
- 12365 markers available for analysis

Taylor et al. 2015, Plant J. 84: 1257-1273



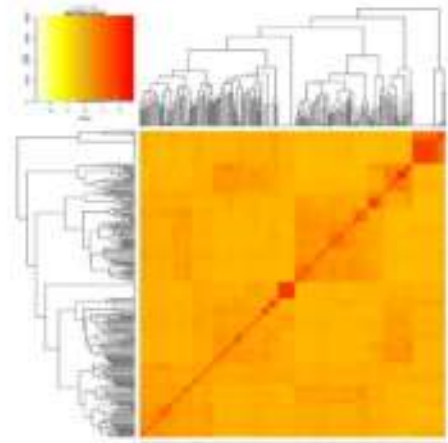
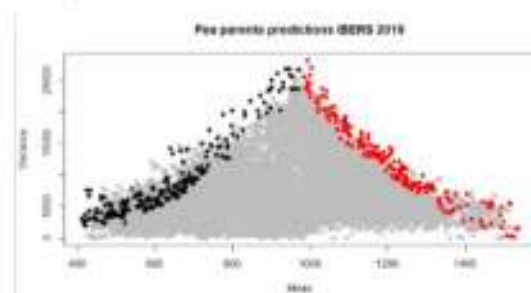
Horizon 2020 of European Union  
This project has received funding from the European Union's Horizon 2020 Programme for Research & Innovation under grant agreement n°727312.

30

## Pea

Genomic relationship illustrated  
by heatmap

Genomic prediction of RIL  
performance



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

31

For peas 260 samples were used and there was a 13,000 SNP array for genotyping and over 12,000 markers were available for analysis for Progeno. Also here there are some examples of initial analysis of genomic relationships and genomic predictions of potential RIL performance in pea.



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

## Soybean



- Soybean EUCLEG collection has been genotyped using the 355K SoySNP microarray
- And 394 Chinese accessions (the NJAU collection)

Wang et al. (2016) Development and application of a novel genome-wide SNP array reveals domestication history in soybean. *PLoS One* 11: e0157112



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

32

Finally soybean, the 355K soySNP micro array was used which was developed by our Chinese colleagues. 480 samples European samples were used and 394 Chinese accessions. There has already been published a paper with a detailed analysis of the genetic diversity of these panels here and more discussion in the chapter on soybean so I won't go into that any further here.

## Soybean



- Quality control resulted in 805 good samples with 229,557 SNPs
- We have positioned the SNP coordinates onto the novel genome assembly Glyma.Wm82.a2 resulting in 224,993 SNPs for further genetic analysis.
- This dataset was divided in three subsets: EUCLEG, NJAU-Wild and NJAU-Cultivated, comprising 477, 82 and 246 accessions respectively (NJAU subsets according to Wang et al. (2016)).

Saleem, A. Muyle, H. Ager, J. Ruzick, T. Wang, J. Yu, D. Rodríguez-Puliz, I. (2021) A Genome-Wide Genetic Diversity Scan Reveals Multiple Signatures of Selection in a European Soybean Collection Compared to Chinese Collections of Wild and Cultivated Soybean Accessions. *Frontiers in Plant Science* 12: 631767. <https://doi.org/10.3389/fpls.2021.631767>



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

33



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

**EUCLEG.eu**

## Conclusions



- Choice of genotyping strategy
  - Inbreeding or outbreeding
  - Linkage disequilibrium
  - Size of genome
  - Reference genome availability
  - Purpose – GWAS, GS
  - Ploidy
  - Budget



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

34

These are some of the conclusions that you can take home from this lecture. So when you choose your genotyping strategy you have to consider whether you are working inbreeding or outbreeding crops. The linkage disequilibrium if it is known, possibly the size of the genome, and whether or not there is a reference genome available, and the purpose for which you want it. For GWAS you may need a higher density than you do for GS as was shown in the soybean talk yesterday, and of course ploidy and budget are important considerations. Thank you very much.

Questions:

1) Was there a specific reason why you used 100 sample per accession for alfalfa but 200 for the red Clover?

I mentioned it briefly, basically because alfalfa is tetraploid, you have double number of genomes. I know people who are working with fewer numbers when they pool data, but certainly I would say that the higher the number, the more accurate allele frequency data will be, everything else being equal.

2) This concerns the high rates of missing data points you get with the GBS technologies. So do you recommend these for the routine application of genomic selection?

Well as was alluded to yesterday, especially for genomic selection, in theory it is true that you need an lot of SNPs to cover the genome, so that in theory you have a SNP close to one of your traits of interest. But there is evidence to suggest that you can get away with much fewer SNPs when you do genomic selection, because the genomic relationship matrix that you can use, for example in GBLUP can be used to calculate your breeding value and is also accurate with much much fewer SNPs, as shown by Hilde Muylle yesterday



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

**EUCLEG.eu**



with soybean, and I've had similar results when I have looked at data from perennial ryegrass. I haven't yet checked it out with red Clover and I don't know the experience of people who have worked with alfalfa here. Certainly for GS you may be able to get away with much fewer markers, so in a sense the GBS technology worked well for us in this project.

### **About the author**

Professor Leif Skøt was head of the Forage Plant Breeding team at IBERS, Aberystwyth University until his retirement at the end of 2020, and is now Emeritus. He has been closely involved in the introduction of genomics assisted approaches to assist IBERS ryegrass and clover breeding programmes, through his research interest in genetic characterisation of germplasm, association genetics and genomic selection.

**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 1<sup>st</sup> October 2021**

Recording link to the presentation: <https://youtu.be/H414E65JI-w>



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

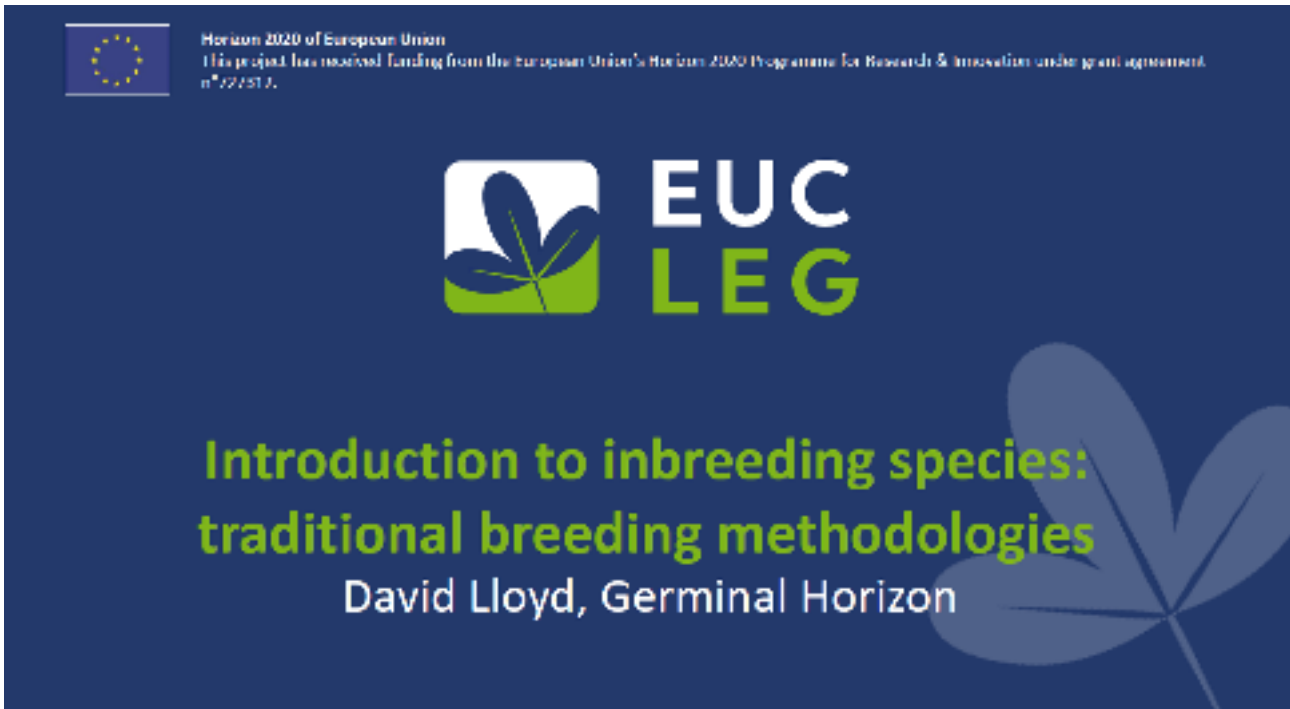
**EUCLEG.eu**



## 4. Introduction to inbreeding species: traditional breeding methodologies

David Lloyd

Head of Forage Breeding, Germinal Horizon, based at Aberystwyth University, IBERS, UK



I'm going to discuss what breeding is from a traditional point of view. What we're trying to achieve in plant breeding, focusing on agriculture species and specifically on legumes, but these principles can be extended to other species. What inbreeding species are. What pure line cultivars are, which is the most basic way that inbreeding species are bred. I'm going to give a couple of examples of how we traditionally breed pure line cultivars, and a couple of examples of how we can speed things up. It's important to note this is a brief introduction, as full coverage of these concepts would easily fill a semester worth of undergraduate lectures, so apologies if I skip over some of the more nuanced aspects of the subject.



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

**EUCLEG.eu**

## Plant Breeding

The development of cultivars that are better suited for our needs

- **Breeding objectives**
  - Higher yield
  - More reliable yields
    - Biotic stress tolerance
      - Pest and disease resistance/tolerance
    - Abiotic stress tolerance
      - Temperature, water availability, soil conditions, wind, etc.
  - Adaptation to environments
    - Nutrient use efficiency
  - Differences in maturity
  - Improved quality
    - Digestibility, Fibre content, Protein content
    - Appearance, Palatability
    - Storage
  - Etc.



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

2

Plant breeding is essentially the development of cultivars that are better suited for our needs, through hybridization and selection. Our main objective is usually higher yields, higher yield potential and also more reliable, more sustainable yields and yield stability. This can be through improved pest and disease tolerance or resistance and abiotic stress tolerance. Arguably the biggest advance in plant breeding in the twentieth century was the introduction of dwarf wheat varieties which made them less susceptible to lodging in wind.

Of particular interest at present is the improvement of cultivars adaptation to environmental conditions, particularly in relation to nutrient deficiency. Phosphorous, for example, is a finite resource, one which is being depleted at a rate that is of some concern. Legumes are notoriously phosphorus hungry and new varieties that are able to thrive on the lower inputs would be a major advance.

Differences in maturity and dormancy is another area of interest to breeders. Soybean, for example, is not grown to any great extent in the UK. Even triple zero maturity groups, those that are quickest to mature, often fail to mature in time to be reliably harvested in some northern parts of Europe like the UK. The development of faster maturing varieties would enable this crop to be grown over a wider range.

We also breed for quality, which encompasses a huge range of traits and I will go into some more detail later.



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

**EUCLEG.eu**

## Cultivars (varieties)

What is a cultivar?

- **Cultivar = cultivated variety of plant (sometimes just called "variety")**
  - A variety of plant that has been developed for specific use in agriculture/horticulture etc.
  - Satisfies requirements of distinctiveness, uniformity and stability (DUS)
    - Distinctiveness: it must be distinct from other, already available cultivars
    - Uniformity: individual genotypes of the cultivar should conform to a prescribed degree of uniformity.
    - Stability: it must stay true to its description when reproduced
  - Generally measured as per guidelines published by International Union for the Protection of New Varieties of Plants (UPOV)
    - 77 member states worldwide
  - Also oversee legal protection of varieties
- **Cultivars defined to some degree by their reproductive biology**
  - Pure line (inbreeding) cultivars
  - Open-pollinated population cultivars (mostly outbreeding)
  - Also hybrid, clonal, multiline etc. cultivars.



Horizon 2020 of European Union

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

3

Cultivars are essentially a “cultivated variety” of plant. While “cultivar” is technically the most correct way to refer to commercialized strains of plants developed for a specific use, it is often used interchangeably with “variety”. Cultivars generally are protected under plant breeder’s rights, which can be thought of as analogous to patenting. To qualify, a cultivar must undergo statutory testing to determine that it satisfies requirements of distinctiveness, uniformity and stability (DUS). Distinctiveness means it must be distinct from established cultivars. Uniformity means that individual genotypes of the cultivar need to conform to a prescribed degree of uniformity: The plants within that cultivar need to be very similar to each other. Stability means it must stay true to this description when it is reproduced. Requirements for DUS are determined by the International Union for Protection of new varieties of plants (UPOV) an international group that has 77 members, comprising roughly half of the countries in the world, including most of the developed nations.

The way that we breed plants is defined by the reproductive biology of that species. For now I’m going to cover pure-line cultivars and will cover open-pollinated cultivars later. There are various other types including hybrid cultivars that I am not going to cover.



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

**EUCLEG.eu**

## Pure line (inbreeding) cultivars

- Tend to be species that naturally self-pollinate
  - Often happens before the flower opens
    - Little or option for plant to cross pollinate naturally
  - Many cereals
  - Some grain legumes
    - Peas, soybean
  - As name suggests, genotypes are advanced to a mostly homozygous state
    - Generally 95% + homozygosity
    - Achieved at around  $F_8$  generation... sometimes taken further
    - Makes the "U" and "S" part of DUS relatively easy
  - Number of approaches taken,
    - Start with hybridization of two divergent lines by hand crossing
    - Take to  $F_{10}$  through single seed descent, then multiply
    - Various selection strategies



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

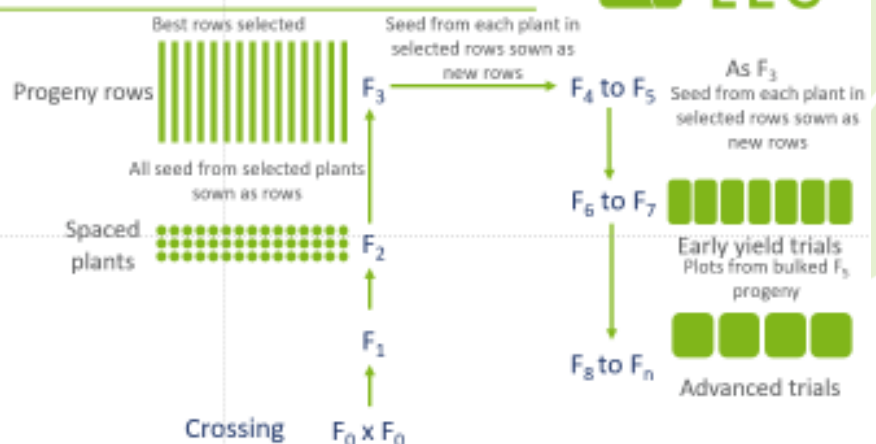
4

Pure line inbreeding cultivars are developed in species that naturally self-pollinate. This includes most cereals, wheat and barley for example, and most grain legumes. They tend to self-pollinate before the flower opens. This means there's little prospect for inter-pollination with surrounding plants of the same species. Pure-line cultivars are so-called because they consist of highly homozygous populations. They are developed through hand crossing of unrelated lines and the progeny taken through various forms of single seed descent selection

## Pedigree method

### Crosses made to produce $F_3$ generation

- $F_1$  selfed to produce  $F_2$
- $F_2$  sown as spaced plants
- $F_3$  seed from selected plants harvested separately, sown as progeny rows
- $F_3$  rows selected.  $F_4$  seed from individual plants sown as new progeny rows
- Repeat for  $F_4$  and  $F_5$  generations
- Bulk harvest from  $F_5$  rows sown as plots for early yield trial
- Repeat n times, with increased selectivity and increasing size
- Select prospective varieties



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

5



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



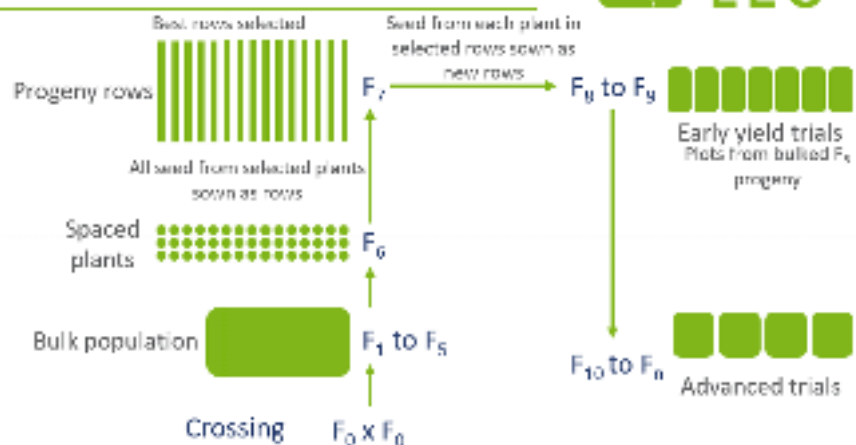
A classic way of breeding pure line cultivars is known as the Pedigree method. F1 hybrids are produced by hand crossing and the progeny selfed to produce an F2 generation. These are sown or planted in a spaced plant nursery. The resulting plants are assessed through the growing season and, based on the criteria set by the breeder, the “best plants” are selected. F3 seed selected from these plants are sown as progeny rows in the next season. Selection is repeated in subsequent generations, taking harvested seed from individual plants within selected rows to form new progeny rows each season.

Once an acceptable level of homozygosity is achieved, typically at the F6 generation or higher, seeds are harvested from the whole row as a bulk and sewn in yield trials that may be carried out over two or more seasons and at increasing numbers of sites, using seed harvested from the initial yield trial. Data gathered over the course of these trials informs the breeder which of their lines are most likely to become a commercially successful cultivar. These are then entered into statutory trials where they are measured independently for DUS and for “value for cultivation and use”. If these criteria are met, then the cultivar can be marketed.

## Bulk method

### Crosses made to produce F<sub>1</sub> generation

- F<sub>1</sub> grown as a population, seed harvested as a bulk
- F<sub>2</sub> to F<sub>5</sub> again grown as populations with no selection
- F<sub>6</sub> grown as spaced plants, best plants selected
- Progeny of selected plants grown as F<sub>7</sub> rows. Rows selected as bulks and put into early yield trials.
- Remainder as per pedigree method



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

6

An alternative approach that provides some advantages over the pedigree method is known as Bulk method. Again, crosses are made to produce an F1 generation. These are then advanced for several generations as bulk families, hence the name. All F2 seed harvested from each of the F1 progeny is harvested as a bulk and sown in F2 plots representing each family (resulting from each initial F1 cross). Again, all F3 seed harvested from each F2 plot is maintained as a family and sown in plots in the next generation. No selection is carried out for the first few generations. Once the families get to a suitable level of homozygosity whereby recessive traits can be reasonably expected to be fully expressed, typically the F6 generation, then selection is carried out. At this point the method progresses in much the same way as in the pedigree method.



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

**EUCLEG.eu**

## Pros/cons of methods

### Pedigree method

- + Intuitive... easy to understand
- + Relatively quick
- + Laborious and arguably inefficient
- + High level of selection at early stages when heterozygous: poor selection for quantitative traits

### Bulk method

- + Simple and efficient
- + No selection until a high degree of homozygosity has been achieved
- + Breeding cycle can be long

Possible to use hybridized approaches

### Level of heterozygosity in segregated alleles

$F_1$ : 100% Heterozygous	
$F_2$ : 50% Heterozygous	
$F_3$ : 25% Heterozygous	← Pedigree method starts selection
$F_4$ : 12.5% Heterozygous	
$F_5$ : 6.3% Heterozygous	
$F_6$ : 3.1% Heterozygous	← Bulk method starts selection
$F_7$ : 1.6% Heterozygous	
$F_8$ : 0.8% Heterozygous	
$F_9$ : 0.4% Heterozygous	
$F_{10}$ : 0.2% Heterozygous	



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

7

Both methods have pros and cons. The pedigree method is intuitive and easy to understand, and has a long history of use by breeders. It's a relatively quick method but also is very labor intensive. A considerable amount of selection takes place at a very early generational stage, before recessive traits have fully segregated. For that reason it's a poor selective method for quantitative traits. The bulk method is very simple, no selection is made until a high degree of homozygosity has been achieved, but the long period of development before any selection can prolong the process considerably, slowing the production of cultivars. It is possible to use combinations of the two approaches and various refinements can be used. But these are the two main approaches used for producing pure line cultivars.

There are a number of ways we can speed things up. For example "speed breeding" has been much discussed in recent years. This is the process of growing plants under controlled lighting, often with very long day lengths, to reduce the length of time between sowing the seed and harvesting progeny seed. Thus it is possible to get two or more generations in a year. It can be used with the bulk method to speed up generation time, but care needs to be taken to avoid having glasshouses running at capacity 12 months of the year. Some down time in glasshouse use is beneficial for controlling pests and pathogens.

It's important to be aware that speed breeding runs some risk of unintended selection. Any variation for traits that could favour some genotypes over others under long day lengths in glass houses may result in selection that is at odds with the breeder's overall intention.



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

**EUCLEG.eu**

## Speeding things up

### Speed breeding

- Growing under lights with optimized daylength, possible to get 3+ generations per year
  - Can be labour intensive
  - "Down time" in glasshouses important for pathogen control

### Off season multiplication

- Can be particularly useful in bulk approach
  - Care needed to avoid mass selection in non-target environment

### Marker assisted selection (MAS)

- Molecular markers can be used to select without phenotyping
  - Need considerable investment in QTL analysis/GWAS
  - Not ideal for polygenic traits with multiple small effect QTLs

### Genomic selection

- Similar basis to MAS, but genome wide... see other talks in training course.



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

8

Likewise, off-season multiplication can be used if the growing season is short enough. Generations can be alternately grown in the northern and southern hemispheres to get twice as many multiplications in a year. This can work very well with bulk method mentioned above but can be confounded by issues relating to importation and border controls. Again care is needed to ensure that no selection is happening in non-target environments.

Other ways of bringing efficiency to breeding include marker assisted selection, which I'm not going to cover at length in this discussion. This can reduce the amount of effort spent on phenotyping particularly with traits that take time to be expressed. Screening can be carried out an early stage, with seedlings that don't have alleles of interest culled so that further effort can be spent on relevant material. This method, which works very well in combination with speed breeding, does need a considerable investment in characterizing of markers through QTL analysis or GWAS. It works very well on simple Mendelian traits but less so on polygenic traits, where there are a lot of QTLs with small effects. Genomic selection is a refinement of marker assisted selection and is discussed at length elsewhere in this booklet.

## About the Author

Dr David Lloyd is head of Forage Breeding for Germinal Holdings. Before taking on his role at Germinal, he was a Senior Legume Breeder at Aberystwyth University, specialising on clovers, peas and faba beans.

**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 30<sup>th</sup> September and 1<sup>st</sup> October 2021**

Recording link to the presentation: <https://youtu.be/H4TWfOInfcg>



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

## 5. Genomics assisted breeding in soybean

Hilde Muylle

Senior research scientist, Plant Sciences Unit, ILVO, Melle, Belgium.



**Soybean – a crop to grow in Europe** 



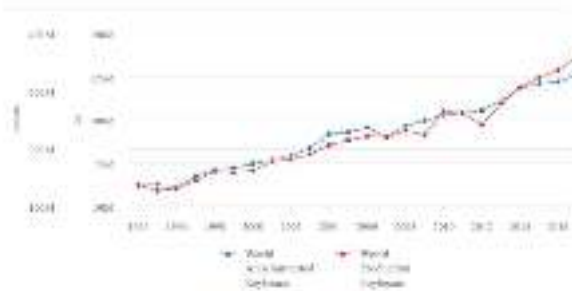
- Important oil and protein crop
- Food and feed applications
- Originated from East Asia
- Domesticated 7000 – 9000 yrs ago

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

Soybean has a high protein content of up to 40% in the seeds as well as 20% of oil making it a very important crop for food and feed applications. It originated from East Asia and has been domesticated for a long time, with domestication started 7000-9000 years ago.



## Soybean – a crop to grow in Europe



If you look to the worldwide production area and worldwide production volume, this is still increasing. There is a high demand of soybean and if you then look to the proportion of soybean that has been produced or is produced in Europe, this is only a very small proportion. It is only 0. 3% of the total volume that is produced within Europe.

## Soybean – a crop to grow in Europe



- Current soybean acreage in Europe is 5.5 mio ha
- Meet only 34% of current European need for soybean

=> Increase European's protein selfsufficiency



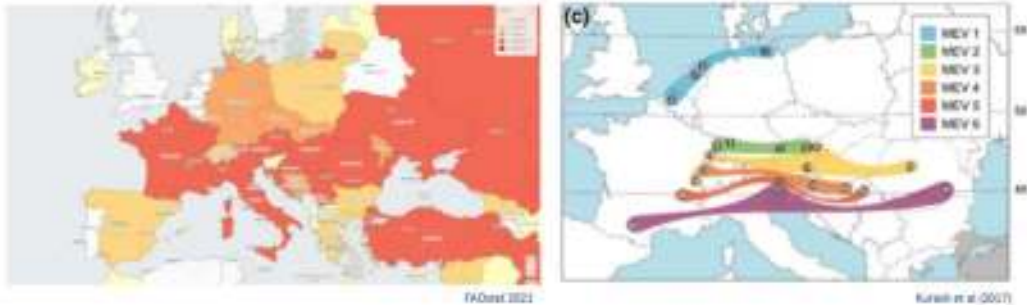
However, we consume quite a lot of it and need quite a lot of soybean within Europe. Currently we produce only 5.5 million hectares of soybean in Europe, and this meets only one third of our needs for soybean. Now with this increasing demand for protein, and the advance to get Europe protein self-sufficient, there is an interest to increase the production of soybean in Europe.



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

**EUCLEG.eu**

## Soybean – a crop to grow in Europe




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

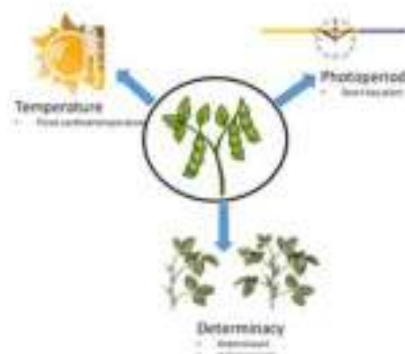
When we look where those 5.5 million hectares are situated within Europe, you can see it's mainly in the belt at a latitude of 45° and indeed this is the region where current varieties can grow and where they are profitable. But if you want to increase the production area within Europe, we should aim for varieties which are suitable to produce within the European Community. So when we want to increase the acreage we have to move more up north and to look for varieties that can grow in those conditions, which are mainly characterized by colder climates, but also shorter growing season. So indeed we have to adapt those varieties to those conditions.

## Soybean – European breeding goals

- Yield improvement
- Increased protein content

by acting on optimal use of season

- Cardinal temperature
- Photoperiodicity
- Determinacy




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

This means that first of all we have to improve yields and to make varieties that really can profit from this short growing season. How can we achieve varieties that can grow in those shorter regions in the Nordic areas? The answer is to select for varieties where the cardinal temperatures are more adapted to those

regions. So we should have varieties that can germinate at colder temperatures. A second characteristic we should breed for is photoperiodicity. Soybean is a short-day crop, so it should sense short days before it turns to a generative status, and it is known that there are some varieties that already can shift to generative status in conditions at higher latitudes. But if you want to go further up, we need photo periodic insensitive varieties. A third item which requires improvement for north-western European adapted varieties, is on the determinacy, because we have those cold conditions or cold spells during the season we should aim for plants which are not determinant. So, if there are cold spells and flowers are getting destroyed, we can still expect a second flush of flowers and to have a production after the cold spell. These are the three major breeding goals within European soybean breeding programs.

## Soybean – European breeding goals



### Yield improvement through abiotic and biotic stress tolerance

- Drought tolerance (slow wilting, fibrous rooting)
- Cold tolerance during germination
- Pseudomonas* resistance



Further we need as well to have adaptation towards biotic and abiotic stress tolerance. Indeed we have been suffering in the past few years with drought spells, so tolerance to drought periods, and as mentioned above cold periods, are very important characteristics that we should breed for, as well as disease resistance. We know for some breeding lines there is susceptibility to *Pseudomonas* and we should prevent this entering into the germplasm of north western European soybean.



## Soybean – Current breeding strategy



How to increase genetic gain in soybean?

$$\Delta G = h^2 i \sigma_p / L$$

$h^2$  = narrow sense heritability  
 $i$  = selection intensity  
 $\sigma_p$  = phenotypic standard deviation  
 $L$  = breeding cycle time



So having those breeding goals, how can we achieve those new varieties that meet these breeding goals? This is expressed or summarized in the formula above. So how to get genetic progress? In fact we should have traits that we can select for, so traits which show higher heritability, we should be able to imply a selection index, so that  $I$  stands for the selection intensity and we can only increase the intensity when we have a large pool where we can do the selection inside, together with a lot of phenotypic diversity. The  $I$  and the  $\sigma$  are really only the variation which is needed to select and then to achieve a quick genetic gain, we should aim for a reduced breeding cycle time.

## Soybean – Current breeding strategy



Self - pollinating species



Soybean is a self-pollinating species. If you look at the current breeding programme, you take 2 homozygous lines, which are crossed. Then you go to an  $F_1$  and through single seed descent you go up to  $F_4$  and then from then on you can start to do phenotyping at the plot or row levels, where you can assess the

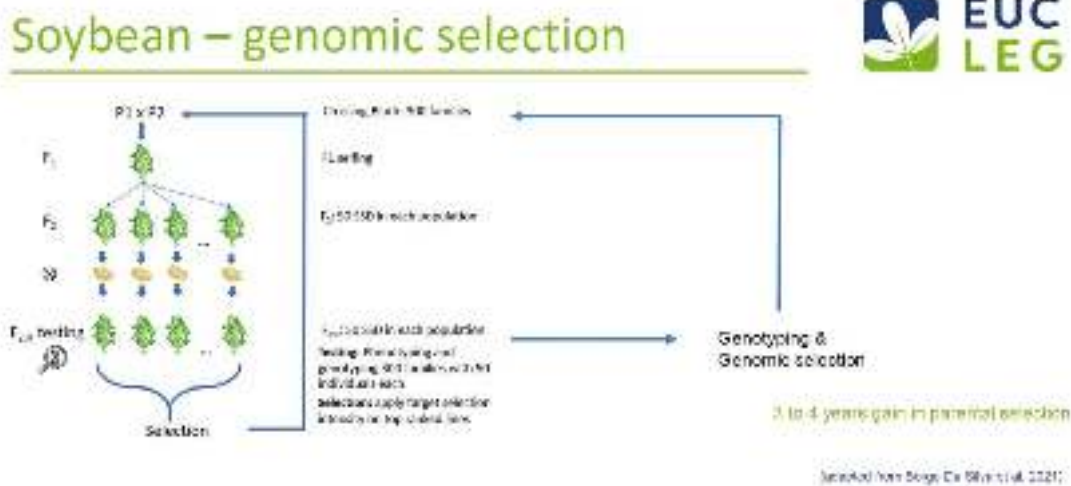


This project has received funding from the European Union's Horizon 2020 Programme for Research & Innovation under grant agreement n°727312.

**EUCLEG.eu**



new genetic combinations. After further selfing and further up scaling of the seeds you get an F9 line, which can be released as a variety. It takes up to 9 to 10 years before a variety is produced. If you could impact the breeding cycle time we could achieve a bigger or quicker genetic gain. That's what we aimed for in EUCLEG to see whether we can use genomic selection to speed up that breeding cycle.



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

So here again you have the breeding scheme of soybean and if we could, instead of phenotyping from F4 on, we could do the selection on the basis of genotyping instead of phenotyping, we could speed up the breeding cycle by three to four years. That's what we investigated within EUCLEG, whether we could set up genomic prediction equations to predict the potential of a given plant on their phenotype.



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

The way we did this was first was by collecting the correct gene pool, which is necessary for breeding. As I already said in my introduction on soybean, we have to use early maturing or very early maturing material



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

to be able to produce seeds in time in the short season we get that Europe. So worldwide, we categorized varieties in maturity classes. There are about 14 defined classes and in fact in Europe only the very early ones, featured on the slide can be used for growing in Europe. So in fact within that gene pool we selected the material for EUCLEG. Also as I mentioned previously, to prevent the risk of a total loss of the yields we aim for semi determinate growth, so we can have consecutive flashes of flower. And it should be a material which is cold tolerant for growth in north-western European conditions.

## Soybean – Gene pool for Europe



Current European varieties show a narrow genetic base

(Miki and Wiershan, 2014; Raj, Mihaljević et al., 2010)

Due to the use of few ancestors from Canada/North America,  
Japan and China

(Kishino et al., 2010; Miki and Wiershan, 2014; Mihaljević et al., 2010)

=> Pay attention to selective sweeps!!

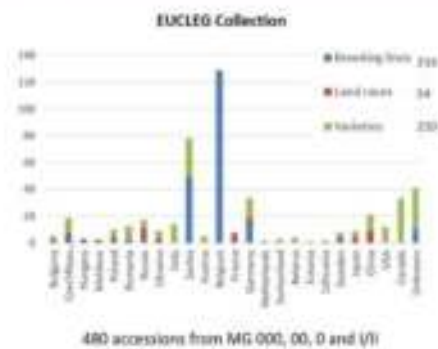


We already knew from beforehand that the current European varieties are based on a very narrow genetic pool, and that have been selected from already pre-selected material and we already knew beforehand that we should enlarge our collection with additional material. So that narrow genetic base is indeed caused by the number of ancestors that were used to build those varieties and we should take care of potential selective sweeps that have occurred during the previous selection cycles.





## Soybean – EUCLEG collection



The first thing we did is to assess whether this collection was harbouring enough genetic diversity for the work we intended to do. We compared the collection of EUCLEG with an already described collection from China of 394 accessions that contained about 120 wild accessions of soybean and 270 cultivated varieties.

## Soybean – EUCLEG collection



Genetic diversity inspected with 355K SoySNP array on 805 genotypes

EUCLEG + NJAU collection (wild and cultivated Chinese material)

- 285,953 SNP markers (80 till 90% polymorphic)
- SNP density = 23 SNPs/100 kbp
- LD decay = 55 till 188 kb
- Nucleotide diversity  $\pi$  between 0.23 and 0.31

Seem et al (2021)



We used the 355K soybean SNP chip to analyse the genetic diversity of 805 genotypes. The SNP chip was a highly useful EUCLEG tool, in fact this yielded the 285,000 SNPs available for genetic diversity studies. We ended up with about 23 SNPs on a 100K KB region on average. We observed that LD decay was about 55 to 188 KB and we had very nice nucleotide diversity between 0.23 and 0.31.

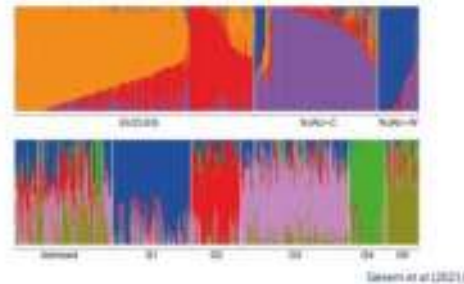


## Soybean – EUCLEG collection



Population structure :

- 3 major groups :
- EUCLEG
- NJUA – Cultivated material
- NJUA – wild material

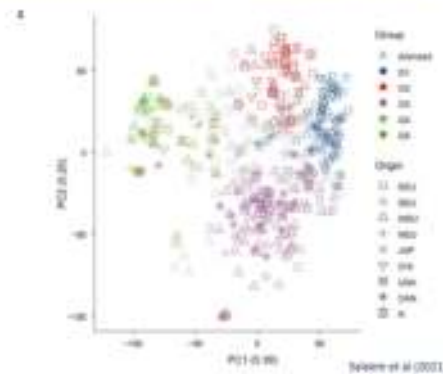


Looking at the population structure and using those markers, we were able to identify the three different collections; the wild accessions, clearly differentiated from the cultivated Chinese material and the EUCLEG material was also clearly differentiated from the Chinese cultivated material. So indeed we had quite unique material which we could differentiate from the wild.

## Soybean – EUCLEG collection



- G1 : South European - Medium late to late
- G2 : Eastern European
- G3 : Western Europe – Early to very early
- G4 : Edamame types
- G5 : Distinct very early material
- Admixed group



Looking within the EUCLEG collection, we could distinguish about five groups, and they were mainly related towards the breeding programs they were coming from, or the region they were selected in. The first one was mostly southern European material, which was later than for example the group five, which was very early which is indeed very early material. The 4<sup>th</sup> group is mainly consisted of Edamame types, so this is the vegetable soybean which is eaten fresh. This material was distinct within the EUCLEG collection. Then groups two and three: two is really originating from Eastern Europe, and the third group from Western Europe containing early to very early material.



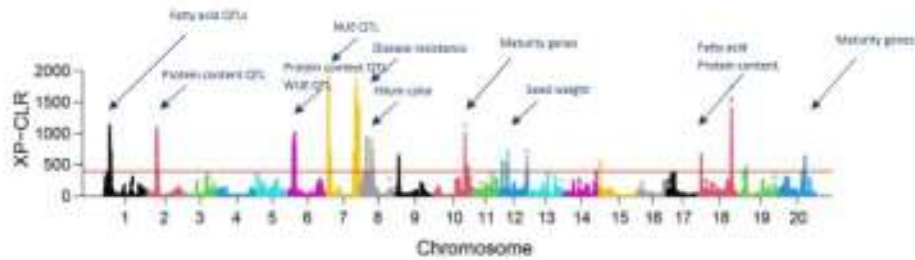
This project has received funding from the European Union's Horizon 2020 Programme for Research & Innovation under grant agreement n°727312.

**EUCLEG.eu**

## Soybean – EUCLEG collection



23 selective sweeps affecting 4% of total genome length



Saareniemi et al. (2021)



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

19

From this marker study and genetic diversity study we used this to also look to see whether we have selective sweeps. We looked through the diversity in all the collections compared to the Chinese wild material. We identified 23 selective sweeps in our collection, in fact affecting 4% of the total genome length, which we should take care of. For some of the selective sweeps we already knew beforehand that they might have been there in our collection, for example we knew that we had very early maturity material in our collection and that those genes coding for early maturity might have been fixed in our collection. Also other selective sweeps were identified and they were mainly located where QTLs for important agronomic traits that have been selected for already are situated.

## Soybean – EUCLEG collection



Conclusion on genetic diversity EUCLEG collection :

- Reduction in genetic diversity
- Longer LD than in wild material
- Selective sweeps (eg on E2 and E4)
- No selection signature for determinacy

=> Enough diversity for breeding European soybean

Saareniemi et al. (2021)



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

20

So that was a conclusion from our genetic diversity study on the EUCLEG collection. Indeed there is a reduction in genetically diversity in the EUCLEG collection compared to the wild collection. We also



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

**EUCLEG.eu**

observed a longer LD, than in the wild material. We observed some selective sweeps, however determinacy we also expected to be fixed in our material, but it wasn't, so we couldn't observe any selective sweep on that region where the determinacy is coded. So for us, although 4% of the genome is affected by selection, we assume that there's still enough diversity for breeding within this collection.

## Phenotyping



### Connected multi location field trials

Augmented design

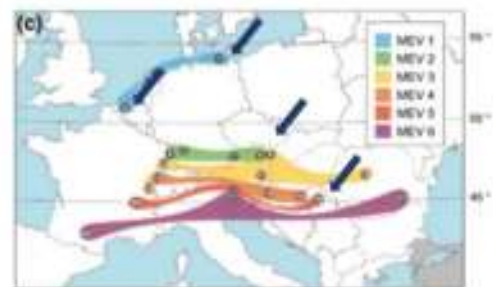
4 locations

### Controlled conditions :

Cold tolerance

Drought tolerance

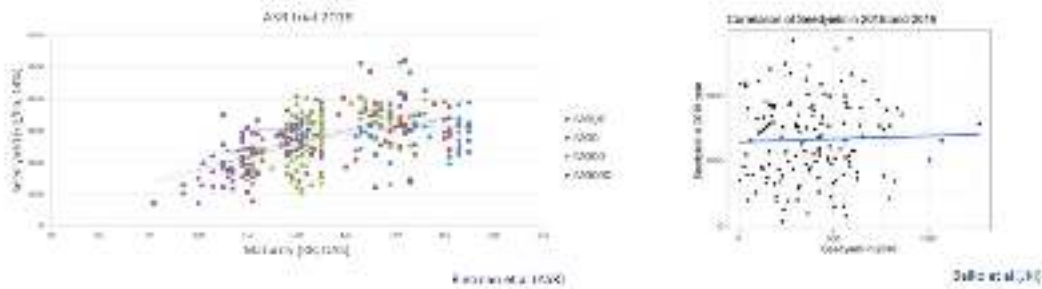
Disease tolerance



The gene pool established within EUCLEG was phenotyped in multi-location field trials using an augmented experimental design. The experiments were conducted at four different locations. In Germany, a subpopulation was tested containing mainly the early maturing genotypes. While in Southern Europe a subcollection with late maturing genotypes was tested. The complete collection was tested on the two main sites situated at a latitude of 50°degrees. Phenotyping under controlled conditions was done for cold tolerance and germination, drought tolerance and disease tolerance.

## Phenotyping - Yield and Protein

- Late maturing do not reach full maturity => Lower yield
- Big gap between 2018 and 2019 => Low “yield stability”



Horizon 2020 of European Union

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

22

In the image above you can see the effect of maturity in relation to seed yield; the later the maturing material is, the higher the yield. The early maturing material is yielding less, so there we need to be sure to use proven material to have a profitable crop, but what we observed as well was that we could not move forward to later maturing material, because those genotypes don't reach full maturity or full potential in our conditions. So we have to take into consideration the maturity of the material, to obtain sufficient yields. Another observation we made in our collection, was that we have quite low yield stability. We phenotyped over two years and if we look to the correlation between 2018 and 2019, we see that there is a very low correlation between both years. Indicating that yield stability is also a problem within our material, and that we should aim to breed for yield stability as well in our conditions.

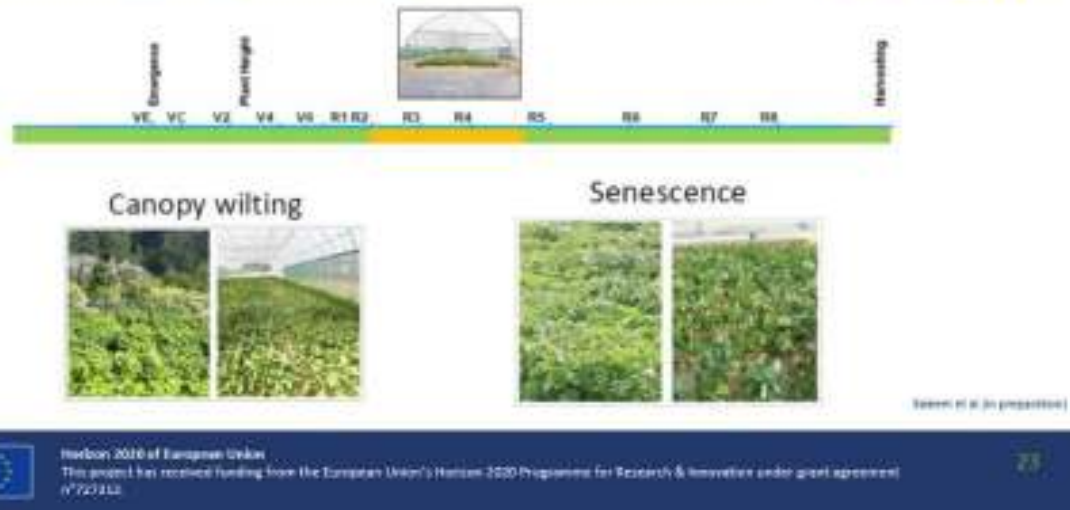


This project has received funding from the European Union's Horizon 2020 Programme for Research & Innovation under grant agreement n°727312.

**EUCLEG.eu**



## Phenotyping – Drought tolerance



We also observed phenotypic differences for drought tolerance. We looked at the reaction to drought during the generative phase, at canopy wilting, and leaf senescence as a reaction to drought.

## Phenotyping – Drought tolerance

- 2018 drought was less severe than 2019
- Phenotypic diversity in canopy wilting and leaf senescence ( $h^2 = 0.12 - 0.54$ )
- Narrow diversity for crop water stress index (CSWI)

We could see that 2018 was a different season to 2019, but we could observe phenotypic diversity in canopy wilting and also in leaf senescence, showing a nice heritability especially in the later observations after drought. But we didn't see a lot of diversity in canopy temperature. So in the crop water stress index we didn't see a lot of phenotypic diversity.

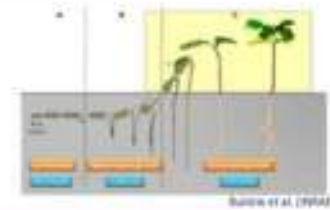
## Phenotyping – Tolerance to cold imbibition



To select for genotypes that can be early sown:

⇒ Seedling emergence

⇒ Germination and seedling growth (control and cold)



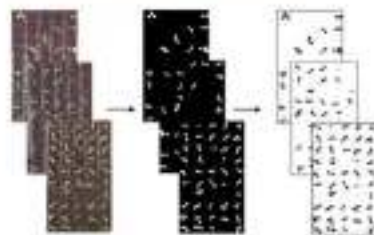
Another abiotic stress that was assessed during EUCLEG was tolerance to cold imbibition. When we want to grow soybean at higher latitudes, we want the germination to go smoothly even under cold conditions and this was assessed under cold conditions.

## Phenotyping – Tolerance to cold imbibition



### Use of High Throughput Phenotyping platforms

- Percentage of emergence
- Surface area after 10 days [1st leaf]
- Automated pipeline to turn, crop, filter and analyse 'green' surface area



SUN ET AL. (2014)



We could use high throughput phenotyping platforms, because this is also a message that I want to address within EUCLEG. Some high throughput phenotyping platforms were established, allowing us to assess a higher number of plants and allowing us to assess the higher phenotypic diversity. This is also one of the factors in the famous formula of genetic gain, and by means of those phenotyping platforms we can assess higher number of individuals.

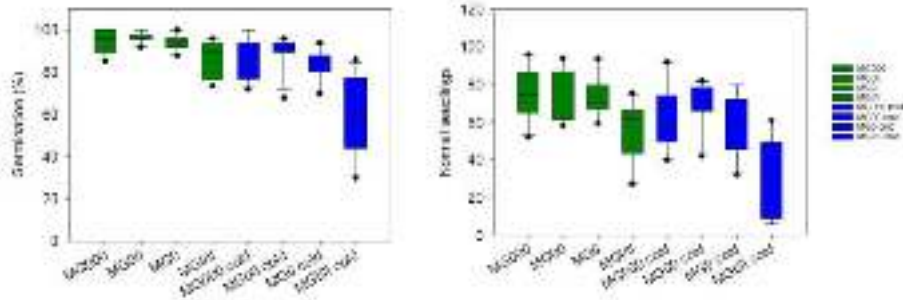


This project has received funding from the European Union's Horizon 2020 Programme for Research & Innovation under grant agreement n°727312.

**EUCLEG.eu**

## Phenotyping – Tolerance to cold imbibition

### Use of High Throughput Phenotyping platforms



- Link between MG and imbibitional cold injury?

Balasko et al. (2018)



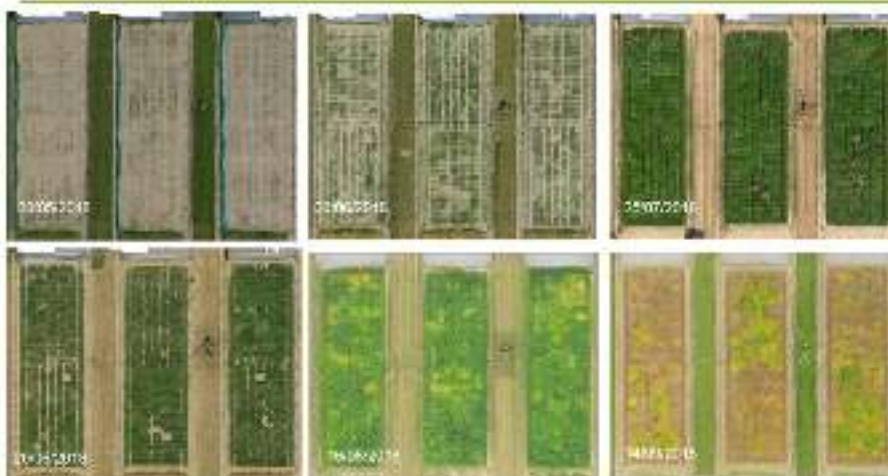
Horizon 2020 of European Union

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

27

If we look to the tolerance to cold imbibition indeed we could see phenotypic diversity in the reaction towards cold. Germination and also the formation of a normal seedlings was impacted, and we could see that there was a link with maturity groups indeed the ones which were late maturing were more affected by cold during imbibition during germination. However, we saw phenotypic diversity, so there is potential for breeding.

## Phenotyping – UAV methods



Bello Schiavon et al. (2020)



Horizon 2020 of European Union

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

28

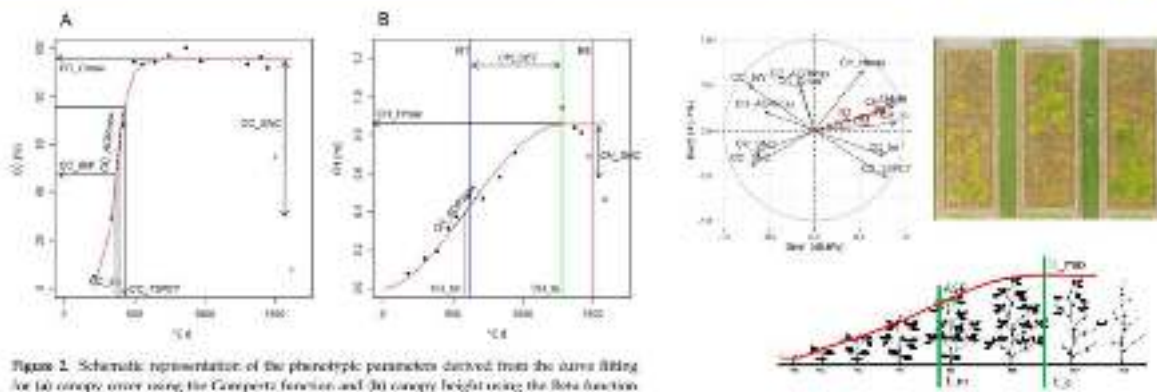


This project has received funding from the European Union's Horizon 2020 Programme for Research & Innovation under grant agreement n°727312.

**EUCLEG.eu**

Another high throughput phenotyping platform I would like to illustrate during EUCLEG is the use of unmanned aerial vehicle (UAV) methods to assess phenological parameters of soybean. Here you can see the different pictures during the growth of a control experiment at ILVO. You can see the germination and canopy closure in the upper series of photos and at the lower end they move towards senescence and you can really see diversity already by eye.

## Phenotyping – UAV methods



Morel-Santoni et al. (2016)



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

29

By means of flying at different times over the control fields, we can simulate the growth curve of each genotype and from that growth curve we can derive different phenological parameters that are interesting to select for. Because yield is a very complex trait, by means of UAV methodology we can dissect yield and different yield components and different aspects. And from there we can assess, for example the growth of canopy closure, the time to close a canopy, the time needed to senesce, final plant length and so on. So this also provided a means to assess quite a lot of phenotypic traits, in which we observed phenotypic diversity.



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

**EUCLEG.eu**

## Soybean – GWAS

Adaptation to NW-Europe => focus on genes related to :

- Phenology
- Architecture
- Tolerance to biotic and abiotic stress

**BUT** we can expect significant heritable covariation with other traits!

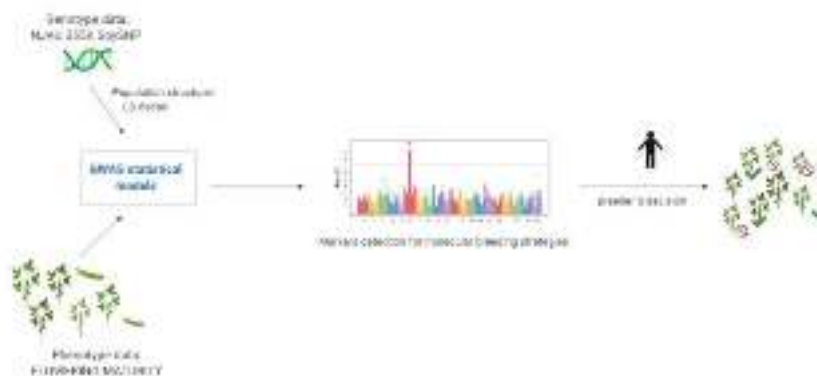


Horizon 2020 of European Union  
This project has received funding from the European Union's Horizon 2020 Programme for Research & Innovation under grant agreement n°727312.

30

So assessing the collection revealed that there was quite a lot of phenotypic diversity in which we could select for. But we wanted to know as well the genetic control of those traits. What genes are related to phenology, which genes were related to architecture, tolerance to biotic and abiotic stress? And as well what is a covariation with other traits, because for example if you could select for tolerance for cold inhibition, you might select for the early ones, but preventing selection of the late maturing ones. So there's some covariation with important traits and we should know the genetic control of each trait and the interrelationship with other phenotypic traits.

## Soybean – GWAS



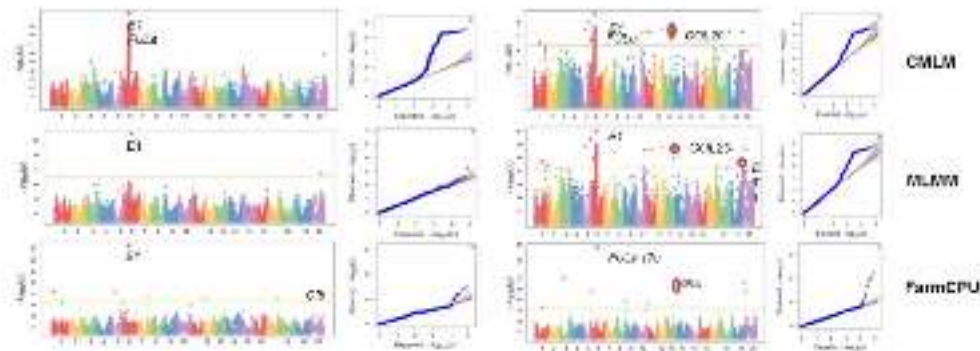

Horizon 2020 of European Union  
This project has received funding from the European Union's Horizon 2020 Programme for Research & Innovation under grant agreement n°727312.

31



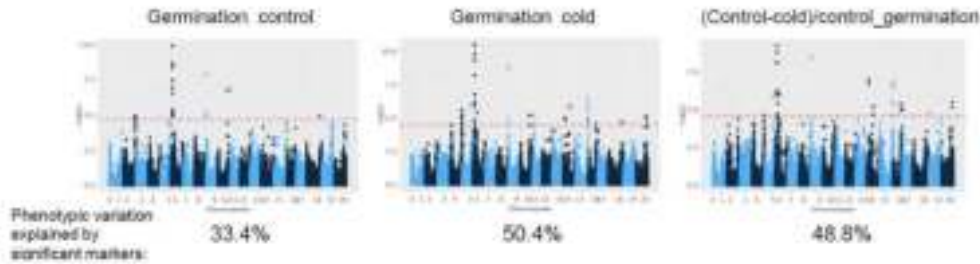
The GWAS studies which were performed within EUCLEG was on the Soybean collection that I have been describing using the 355K soybean SNP chip. We checked the association with given genomic regions for each trait. I will go through some of the results. For flowering and maturity, although we had some selective sweeps on the E2 and E4 genes, we could identify some other genes related to flowering and maturity. To the left shows flowering, we clearly identified an association with the E1 and for majority we had some interesting candidate genes which we can study more depth and to see what the role is on maturity within the collection.

## GWAS – flowering and maturity



For flowering and maturity we had some interesting candidates, similarly to tolerance to cold imbibition. We see some regions popping up which explain quite a lot of phenotypic variation for those markers. So further research is going on to identify what genes are behind those significant associations. They might shed light onto the biological processes behind it.

## Phenotyping – Tolerance to cold imbibition



Sassi et al. (2020)



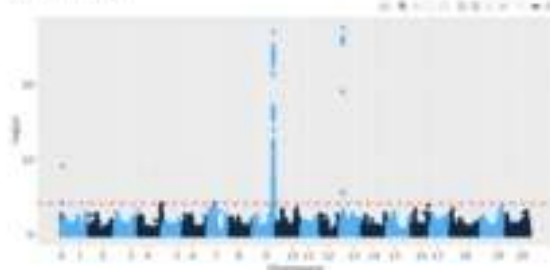
Horizon 2020 of European Union  
This project has received funding from the European Union's Horizon 2020 Programme for Research & Innovation under grant agreement n°727312.

33

## GWAS – disease resistance

- *Pseudomonas savastanoi* pv *glycinea*

Manhattan plot



Sierman et al. (2020)

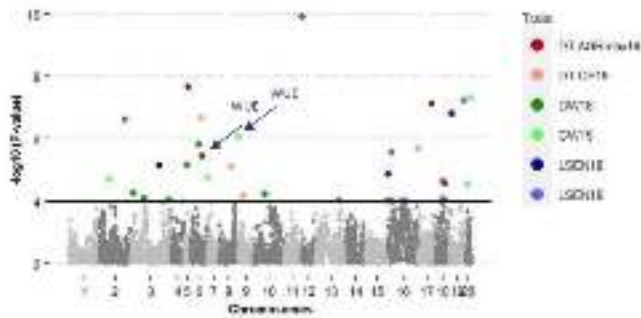


Horizon 2020 of European Union  
This project has received funding from the European Union's Horizon 2020 Programme for Research & Innovation under grant agreement n°727312.

34

Similar for disease resistance, here we see two regions popping up for *Pseudomonas* resistance.

## GWAS – drought tolerance

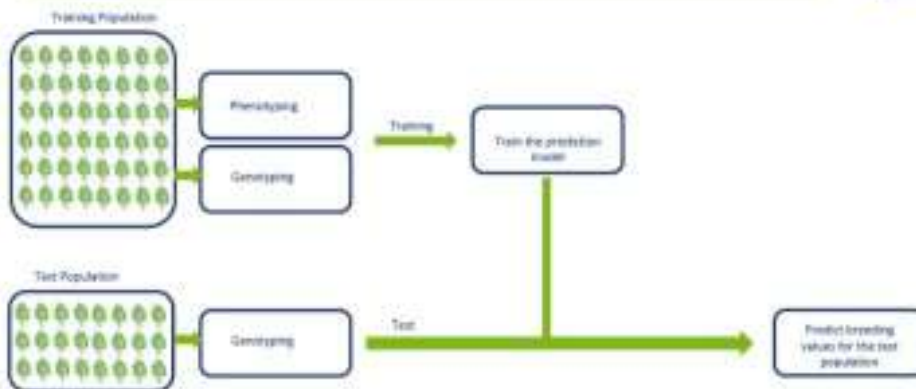


www.nature.com/nature


 Horizon 2020 of European Union  
 This project has received funding from the European Union's Horizon 2020 Programme for Research & Innovation under grant agreement n°7272812.
 35

And finally, onto drought tolerance. We identified some significant associations and behind some of those significant associations we have some reported QTLs on water use efficiency, for example.

## Soybean – genomic selection



Adapted from Borjesson, de Silva et al. (2022)


 Horizon 2020 of European Union  
 This project has received funding from the European Union's Horizon 2020 Programme for Research & Innovation under grant agreement n°727312.
 36

So indeed we have some quite interesting association significant associations, with important traits, but how can we implement this now in the soybean breeding programme. One way is to use this knowledge in our genomic prediction procedures. So in fact that's what we tried as well within EUCLEG, to build prediction formulas to predict yield and protein. We divided our collection into a training population which was phenotyped and genotyped. We made a prediction model and then we tested whether this prediction model was able to be applied on the test population and to see how good the prediction could, be solely on the basis of genotyping.



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

**EUCLEG.eu**



## Soybean – genomic prediction

### Size effect of validation set on predictability

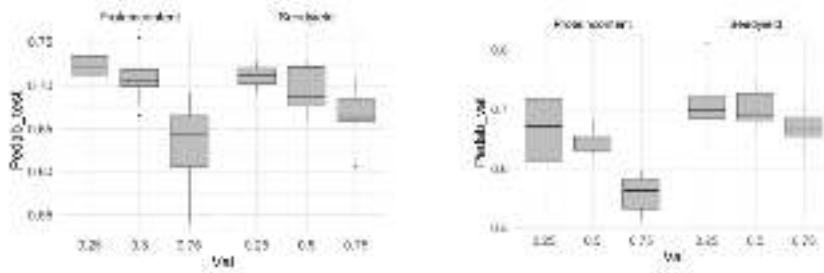


Figure 14 (18/17)



37

Some of our results from the soybean collection are shown above. We tested how big the training population should be. If we divide the EUCLEG gene pool and we take only 25% as the test population, 75% as the training population, we get a good prediction for protein and seed yield. But you can see the further we go, so if we have 75% of our population as a test population and only 25% as training, we see that predictability is lower, although it's quite high in the case for protein and in seed yield. These are quite promising results here, to see that we have quite good predictabilities.

## Soybean – genomic prediction

### Effect of structure in validation set on predictability

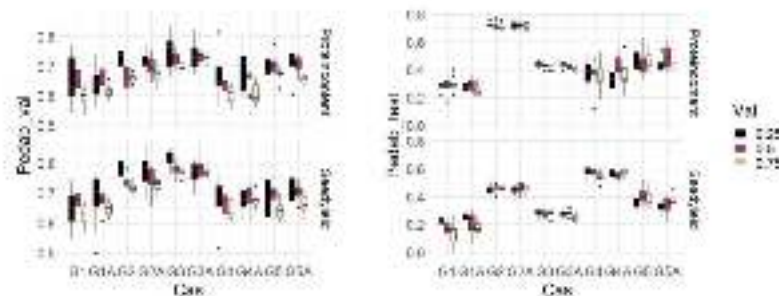


Figure 14 (18/18)



38

As you know that in the EUCLEG collection we had some genetic structure, so there were quite clear distinct groups. We tested whether the prediction equation that was built within a group could be used to predict phenotype for the rest of the population. You can see that depending on which group was used to

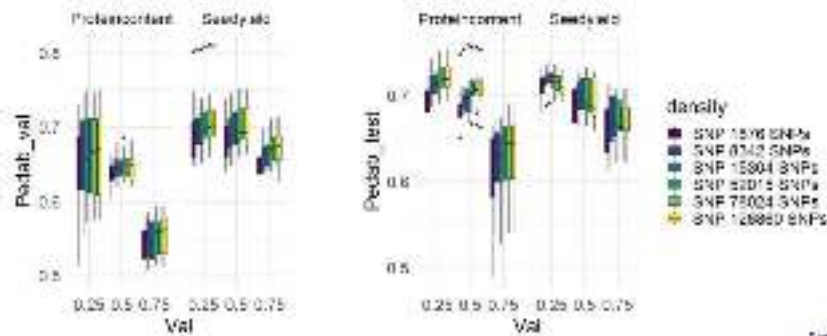


build them, the predictability of the prediction equations is changing. So some of the groups are better suited to build a model and they are providing a more robust prediction equation, to predict the phenotype. It is very important to know the structure within the population, and to see how this can be used to make your predictions

## Soybean – genomic prediction



### Effect of marker density on predictability



Figures 1-18642



Horizon 2020 of European Union

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

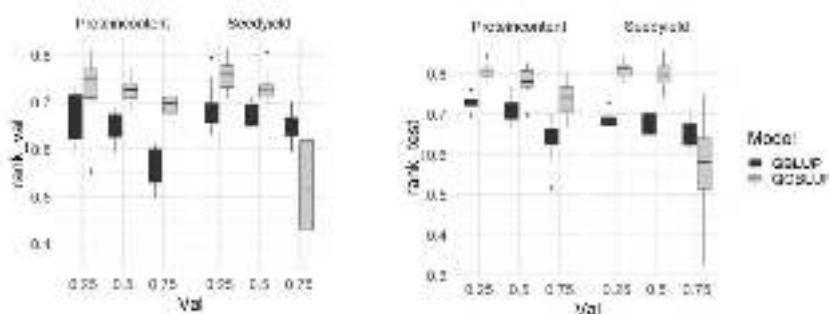
39

Another test that was performed within EUCLEG was to look to the marker density. We had about 285,000 polymorphic SNPs. In fact we could reduce it up to 2000 SNPs, to still have quite a good predictability. So the size of the training site was much more important, to have good predictability compared to the numbers of SNPs taken to build the prediction equation. It could be reduced to 2000 SNPs whilst having a quite similar predictability.

## Soybean – genomic prediction



### Inclusion of associated markers in prediction model



Figures 1-18642



Horizon 2020 of European Union

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

40



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

**EUCLEG.eu**

Finally, as we had the GWAS methodology within EUCLEG we tried to put associated markers with protein or seed yield as fixed factors within our prediction model. And what we saw, was that if we put already known associated markers or already known genomic regions as fixed factors within our prediction model, we could achieve higher predictability compared to a completely random driven model.

## Soybean – genomic prediction



Genomic prediction in EUCLEG collection :

- Composition and size of training set is important
- Marker number needed is low
- Associated markers as fixed effect aid
- Difference between traits

Reger et al. (2014)



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

41

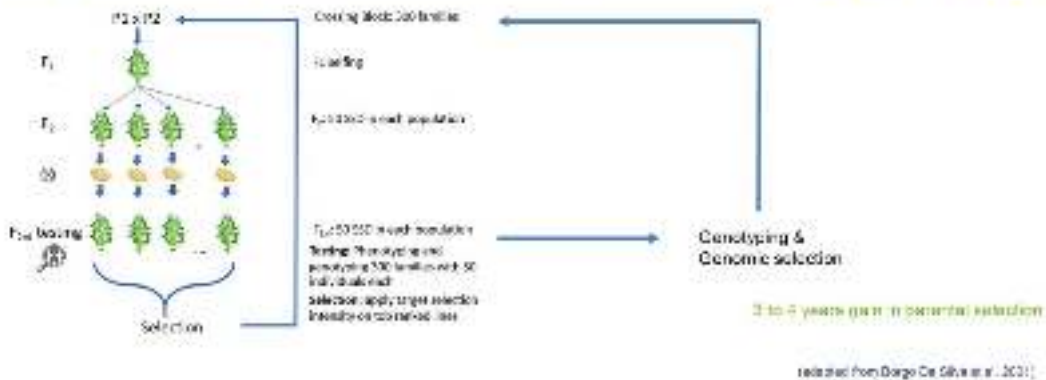
So coming to the conclusion on the genomic prediction models, indeed within EUCLEG the size of the training set is quite important, as well as structure within the population. So take this into account. It was very interesting to see was that the marker number needed can be quite low and that associated markers already known to be associated with the traits, if you put them as a fixed effect, this aids the predictability. And of course, we see a difference between traits, some traits have a higher heritability than others and this is reflected in the prediction efficiency.



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

**EUCLEG.eu**

## Soybean – genomic selection



So how can we put this into the breeding programme? In fact, if you have those prediction equations available, we can test already at an  $F_4$  stage, to genotype it and to predict the phenotypes. And in fact at that time we can already have an idea of the phenotypic performance of that genotype and we can already select new parental lines that feed into the parental crosses. So instead of waiting up till 8 or 9 years, the end of the selection cycle, we can shorten the cycle by three to four years, by selecting already at  $F_4$ .

## Soybean – genomic selection

How to implement in current breeding programs :

- Depends on breeding scheme and related training set  
Within population selection being superior to across family selection
- Renewal of calibration curves?  
⇒ Highest gain is achieved after 10 cycles
- Maintaining genetic variance

(Dorge De Silva et al. 2021)

So, concerning genomic selection and how to put it into the breeding programme? It's important to have the correct and suitable training set within your breeding programme. There are different ways to set up the breeding programme. You can select within populations, so if you make crosses you continue to select within those ones, where you can select between them. Simulations already published by other authors, within population selection are assumed to be superior to cross family selection. Which is quite logical!

think, because then you really select within your training population. Concerning the renewal of calibration curves, so it takes some time to build up those calculation for prediction equations. Within simulations it's been shown that the real potential of a prediction curve, is reaching its maximum or the highest gain after 10 cycles of breeding, and after that the training set should be updated. Of course, it can be updated during the consecutive breeding cycles, but after 10 cycles the maximum benefit will be gotten out of it. A third important issue is when you select within the population, you have to take care that you continue to maintain genetic diversity within your collection, because indeed after genomic selection you go in a certain direction and you should take care to not lose too much diversity within the breeding programme.

## Soybean – Future perspectives



### Breeding soybean for NW – Europe using GS models

⇒ Choice of training set (and renewal)

⇒ Exploit knowledge on target candidate genes (Kim et al, 2021)

⇒ Haplotype based GS

⇒ Whole genome sequencing (Kim et al, 2021)



Horizon 2020 of European Union

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

44

Looking also to soybean and going back to the start of EUCLEG and where we are now, I think a lot of new tools have become available. A lot of knowledge is becoming available within soybean. Mainly outside Europe, but I think also within Europe quite a lot of knowledge has been generated. Quite a lot of whole genome sequencing has been done using shallow genome sequencing, but there's quite a lot of data available which can be used to haplotype discovery within those genotypes. Also next target candidate genes, quite a lot of genes are involved in those important traits, which were selected for. They have been characterized, we know the different haplotypes available in those candidate genes, so this knowledge can also be integrated within the breeding programs. Also the training set, how to build this and how to renew it. This is quite important, especially for north-western Europe, because we don't want to be too narrow within our collection. So during the duration of the four to five years of EUCLEG, quite a lot of new technologies and knowledge have become available and we have a good perspective for genomic selection within soybean.



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

**EUCLEG.eu**



## About the author

Dr. Hilde Muylle is a senior research scientist at the Plant Sciences Unit of ILVO in Melle, Belgium. Her research focuses on molecular genetics and genomics of ryegrasses and arable crops as triticale and soybean. Specific focus is on understanding the genetics of biomass quality in relation to the crop use by means of GWAS and genomic selection.

**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 1<sup>st</sup> October 2021**

Recording link to the presentation Genomics assisted breeding in soybean: <https://youtu.be/X10cMmO0joY>



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

**EUCLEG.eu**

## 6. Genomics assisted breeding in pea

David Lloyd and Radu Grumeza



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



# Genomics assisted breeding in pea

Storm Seeds: Radu Grumeza  
Germinal Horizon: David Lloyd

Activate Windows  
Go to Settings to activate Windows.

### Usage of dry pea seed



- **Usage of dry seeds for food**
  - High value protein and starch fractions use have expanded substantially for food
    - Plant based burgers and meat replacement products
    - Beverages
    - Pasta
    - Extruded snacks
  - Direct use as food: cracked pea, snacks, mushy pea, canned dry pea
- **Usage for animal feed**



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

Activate Windows  
Go to Settings to activate Windows.

There has been a large increase in the acreage of peas in recent years, largely due to expansion in the human consumption markets. This reflects the increase in vegetarianism and veganism, but also in the expansion of 'flexitarianism', where consumers are reducing the amount of meat they are eating, replacing



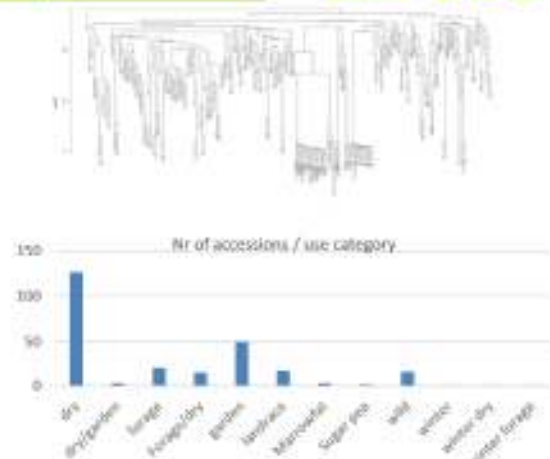
it with vegetarian meals, but not necessarily going fully vegetarian. There are also novel ways in which pea derived products are influencing the human consumption market. The use of extracted protein has contributed to this, with the development of pea protein based meat replacement products. There is unprecedented demand for nondairy milk replacements, with “pea milk” increasingly being stocked by supermarkets along with pea flour based pasta and extruded snacks. Pea protein is also being used increasingly as a filler in traditional meat products, such as sausages, reducing their meat content.

There has also been an increase in use of peas for animal feed. Over the past few decades, we have built a dependence on the importation of protein for animal feed. This is mostly soy derived, sourced from North and South America, comprising about 70% of protein used in feed rations. Pea has unsurprisingly been identified as a potential replacement for imported soy derived protein. This is particularly true for the northern parts of Europe, where home grown soybean is currently not an option.

## Pisum genetic diversity



- 260 varieties analysed in Eucleg
- A mix of core collections from JIC and Nordgene
- Majority are dry pea cultivars with different origins
  - Europe (UK, Nordic countries, France, Belgium, Serbia, Ukraine)
  - America (Canada USA)
  - New Zealand and Australia



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

Activate Windows  
Go to Settings to activate Windows.

The goals within EUCLEG were to use genome wide association to identify and genomic selection to attempt to speed up the rate of genetic increase within the pea germplasm. In EUCLEG we looked at 260 different varieties to assess pea genetic diversity. The majority of cultivars used were selected from core collections from the John Innes Centre Norwich in the UK and NordGen, the Nordic Genetic Resource Centre. By and large these were dry pea cultivars from Europe, America and New Zealand, as well as other types including garden peas, landraces and wild peas.

At the beginning of the project there was a relative lack of material in terms of bulk quantities for many of the lines we wanted to study. Consequently, there was a requirement for some multiplication to be carried out before we initiated trialling. As a result of this, the pea work is at an earlier stage of data analysis than, the other grain legumes. The genetic analysis of traits is, as such, at the beginning stages.



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

**EUCLEG.eu**



## Breeding objectives



- **Grain Yield and yield components:**
  - Agronomic, novel plant morphology and phenology; improved lodging resistance etc
  - Environment and rhizobia interaction adaptations
- **Protein content and composition**
- **Biotic factors:**
  - diseases are one the most important constrain for global pea production
    - Root rot complex (*Aphanomyces eutiches*, *Fusarium solani*, *F. oxysporum* etc.)
    - Aschochyta blight complex (3 pathogens), *Sclerotinia*, *Botrytis* spp., *Septoria* sp., often form a canopy rot complex
    - Powdery mildew, Downy mildew, Rust
    - Viruses (PEMV, PSBMV, BYMV, BLRV and the recently identified PNYDV)



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



In terms of traits of interest, breeders are mostly concerned with grain yield and yield components including agronomic traits, plant morphology and phenology. Of particular interest are things like improving harvest indices, improvement of standing abilities, environmental and rhizobial interaction adaptations. Breeders are also interested in quality traits, particularly protein content and composition, for instance improving the amino acid profile of the seed in order to better match end user requirements. Within the diversity collection used within EUCLEG collection there are several lines of germplasm that have promising protein traits from which to breed “better” varieties.

In terms of biotic stresses, a lot of work has been carried out on disease tolerance in peas. The main diseases of interest are the so-called root rot complexes, including *Aphanomyces* and *Fusariums*. *Aschochyta* blights affects leaves along with various other fungal pathogens like *Sclerotinia*, *Botrytis* and *Septoria*, which form a canopy rot complex. It is frequently difficult to visually distinguish between the contributing pathogens. Powdery mildew, Downey mildew and rusts tend to strike later in the growing season. Downy mildew can be a real problem in terms of delaying the maturity of the pea crop and it has a major impact on harvestable yields.

Viruses, that particularly affect peas include pea enation mosaic virus (PEMV), pea seed borne mosaic virus (PSBMV), bean yellow mosaic virus (BYMV), bean leaf roll virus (BLRV) and more recently identified pea necrotic yellow dwarf virus (PNYDV). Robust resistance to these pathogens is a major focus of pea breeding programmes.



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

## Breeding objectives



- **Abiotic stress**

- Drought and water deficit stress
- Waterlogging (WL)
  - Peas are more sensitive to WL than other pulses
- Heat and temperature stress
  - Temperature above 25 °C can lead to seed abortion and great yield losses
  - Low temperature decrease growth but have lower impact on yield unless there is frost
- Nitrogen stress.
  - Quantity of nitrogen accumulated = combination of absorbed and fixed nitrogen
  - Mostly linked with drought and water deficit



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

Activate Wind  
Go to Settings to a

A key element of yield stability is the ability of a variety to cope with the range of abiotic stresses that they can typically be expected to encounter. Water deficits are a major contributor to reductions in yields, but equally peas can be very sensitive to surfeits of water, particularly in regards to water logging. In the UK, for example, the area for pea cultivation is concentrated in the eastern part of the country, which has far lower rates of precipitation that in the west of the country. Much of this sensitivity to water logging is due to the associated increase in disease pressure, but there are also issues of hypoxia, a reduced availability of oxygen to the roots that is caused by increased water content in soils.

Peas are also prone to temperature stress. Prolonged temperatures above 25 degrees can lead to seed abortion which obviously reduces the yield of the crop. Low temperatures can result yield losses if frost occurs during the growing season. Nitrogen can also be a limiting factor. Soil conditions can result in reduced rhizobia activity that result in nitrogen deficits.

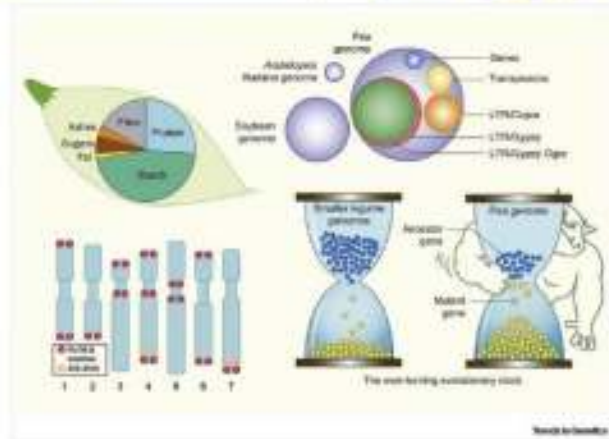


This project has received funding from the European Union's Horizon 2020 Programme for Research & Innovation under grant agreement n°727312.

**EUCLEG.eu**

## Pisum sativum genome

- **Self pollinated crop with a big genome 4.063 Mb**
  - 7 chromosomes ( $2n=2x=14$ )
  - The sequence assembly represents 88% of the genome and contains 44 756 annotated genes
  - 2 225 175 repetitive elements most of which are Transposable elements (TE's)
  - The *Pisum* genus comprises of two species
    - *Pisum fulvum*
    - *Pisum sativum*
    - Several subspecies (wild)
      - *Pisum sativum* subsp. *elatius*
      - *Pisum sativum* subsp. *abyssinicum*



Bursin, Knapik Trends 2019 ->



Horizon 2020 of European Union

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

Activate Win  
Go to Settings to

The pea genome is approximately 4 Gb: extremely large compared it to the likes of *Arabidopsis* at around 100 Mb, slightly larger than soybean at around 1 Gb, but somewhat smaller than Faba bean (13 Gb). The currently available sequence assembly represents 88% of the genome. This is work that has predominantly been carried out in INRAE in Dijon and contains 44,756 annotated genes. There's lots of repetitive elements within this, including various retro transposons. The genome comprises 2 species, *Pisum fulvum* and *Pisum sativum* as well as the subspecies, *P. sativum* subsp. *elatius* and *P. sativum* subsp. *abyssinicum*.

## Genomic resources

- **SSR and RBTP markers**
- **Genetic maps and QTL mapping**
- **Specific markers linked to major genes and QTL's (some examples)**
  - Trypsin inhibitors
  - Flowering
  - Lodging
  - Resistances to Powdery mildew, PEMV, PSBMV, Rust
  - Seed composition genes
  - Frost tolerance NIL
  - Sclerotinia QTL's
  - Fusarium solani, F. avenaceum
  - Aschocytia blight
  - Bruchus pisorum
  - Vicillins/Convicillins



Horizon 2020 of European Union

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

Activate Win  
Go to Settings to



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

**EUCLEG.eu**

Marker technology is increasingly being used to aid selection in pea breeding. Various markers are available for major genes and QTLs, including trypsin inhibitors, flowering dates, lodging, diseases susceptibility, seed composition traits. For instance, we are using markers to identify non-functional vicillin and convicillin alleles, which code for major seed proteins that are antinutritional and have poor amino acid profiles. By reducing the vicilin/convicilin content we are hoping to rebalance the storage protein composition by increasing legumin content, a more desirable storage protein.

## Genomic resources



- **MAS , MABC (Marker assisted backcrossing) are commonly used to introgress QTL's and genes**
  - Is inefficient for significantly improving quantitative/complex traits controlled by a large number of QTLs
  - Efficiency of these strategies is limited due to use of low density marker systems
- **SNP markers and High throughput genotyping (HTG)**
  - Generated by transcriptome sequencing, GBS or Radseq
  - Creation of 2 SNP arrays: 13.2 k SNP array and recently 90k SNP array
- **GWAS (Galli 2019, etc.) based on the newly published genome sequence + phenotyping observations**
  - Agronomic and seed quality traits of field pea



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



Marker assisted selection can be used to identify and introgress QTLs and genes of these types into commercial varieties. It is inefficient, however, for improving quantitative traits. Yield, for instance, is a highly polygenic trait (or set of traits), having many contributory genes. Minor QTLs can be very difficult to select using marker assisted selection. To date this has been compounded by low marker density and availability. Recently developed SNP chip arrays have addressed the issue of marker density to some extent. In EUCLEG we are using a 13.2 K SNP array and we have also used a more recent 90 K SNP array, which gives far better coverage of the genome. Genome wide association studies (GWAS) has, as a result, become the go to method to conduct genetic studies of pea, providing a more powerful way of using a genetic approach to breeding.



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

# Genomic selection

## Genomic selection (GS) and prediction

- Proposed in 2001 by Meuwissen et al
- Is a form of MAS with extended scope and advantages
- Genome wide markers are used to capture genetic variations in the population and assign genome estimated breeding values (GEBV)
- Is dependant on the design of the Training population and on marker density
- Validation population is both genotyped and phenotyped in Multi Environmental Trials (MET)
- GEBV is compared to actual true breeding values
- Choice of statistical model and quality of phenotyping are contributing factors to success of GS

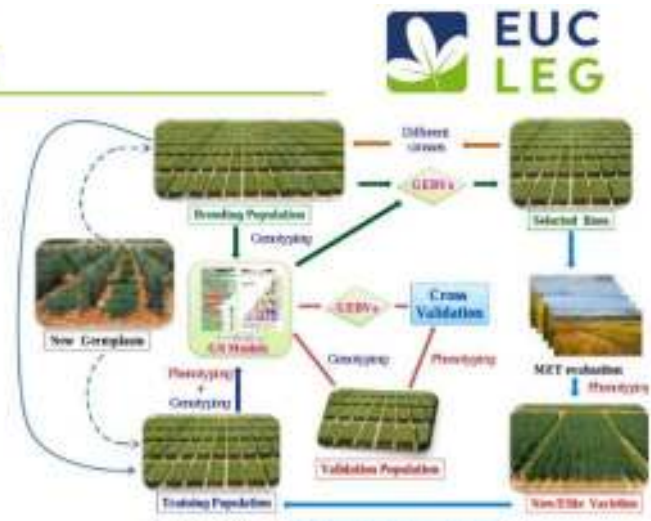


Fig. 1. Schematic representation of the basic genomic selection (GS) methodology.

Krishnasopa 2021, Genomics



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

Activa Win  
Go to Settings to...

Genomic selection was proposed in 2001 by Meuwissen et al. It has taken a while to get up and running because the density of coverage of markers that is needed for genomic selection is very high and until recently the financial cost of achieving this coverage has been prohibitive. As technology has improved, costs have reduced rapidly. It is still expensive to genotype with sufficient coverage, but we are at a stage where it is feasible. Genomic selection can basically be looked at as a form of marker assisted selection. Markers across the genome are used to capture genetic variations within the population and mixed modelling used to assign genome estimated breeding values to genotypes. The approach that we have found most useful is to use best linear unbiased prediction (BLUPs) which is covered elsewhere in this booklet. A training population is genotyped and phenotyped to input data into the model and used to predict breeding values.

Speed breeding is a concept that has been getting a lot of attention recently. It can be applied very successfully in peas, and it is possible to get up to six generations per year for very rapid cycling types. Three or four generations is perhaps more realistic for commercial types. This can be used with the genomic selection approach to really maximise genetic gain. This needs to be combined with classical pedigree selection or bulk selection techniques, and multi environment trials especially for qualitative traits. In combination with functional omics we can better understand the genetic basis of phenotypes and approach breeding in a far more targeted way.

Genomic assisted breeding and marker assisted selection is ultimately the future of pea breeding, which in combination with speed breeding will maximise genetic gain. Genome editing is currently proscribed within the EU but holds enormous potential for the future. It's progressing outside of the EU, particularly in North America and China. This is an area that will need regular review in the EU.



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

**EUCLEG.eu**

## Genomic resources

- **Speed breeding in pea and Speed GS**
  - Up to 6 generation per year (Watson 2018) for early varieties. 4 generations for late pea varieties
  - GS application to segregating populations leads to maximal genetic gain
  - Needs to be combined with classical pedigree selection and MET especially for qualitative traits
- **Functional omics studies**
  - Gene expression atlas
  - RNA seq used for plant pathogen interactions
  - Gene expression studies: Phoma, Powdery Mildew, *Fusarium solani*



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

10  
Activate Windows  
Go to Settings to activate Windows.

## Conclusion: Future breeding in Pea

- **Genomic assisted breeding (GAB)**
- **Marker assisted selection and integration of GAB with speed breeding**
- **Genomic selection (GS) based combined with high throughput SNP assays and High throughput phenotyping**
  - GS + Genomic prediction and Genomic estimated breeding values
- **Microbiome interaction**
- **Climate ready breeding methods**
- **Crop pan-genome assemblies**
- **Haplotype based breeding**
- **Genomic editing (outside of EU)**



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

Activate Windows  
Go to Settings to activate Windows.

To fully realise the potential of genomic assisted breeding there exists a need to fully bridge the gap with high throughput phenotyping, or phenomics. The associated increase in data accumulation will require new approaches to data analysis and modelling, requiring the application of AI. Genomic approaches will furthermore allow for more efficient mining of genome resources for new sources of disease resistance and novel traits allowing them to be effectively incorporated into commercial cultivars, giving greater yield stability in the face of biotic and abiotic stress.



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

## Conclusion: Future breeding in Pea

- Phenomics: bridging the gap between genomics and phenomics.
  - High throughput phenotyping
  - Phenomics coupled with AI and machine learning
- New resistance sources (landraces and wild material....)
- Pyramiding of different resistance genes using NGS methods
- Functional genomics and discovery of new candidate loci
- Resistances and adaptations to the changing climate
- Yield stability by breeding more resistant varieties

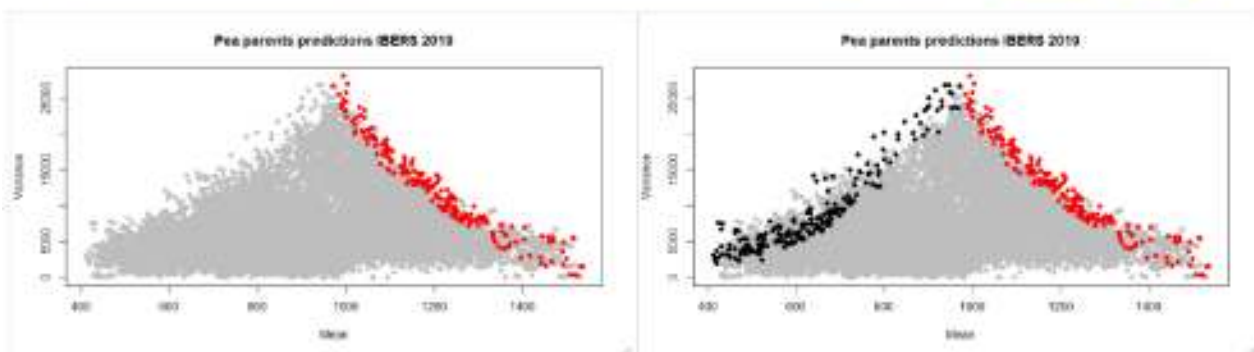


Horizon 2020 of European Union  
This project has received funding from the European Union's Horizon 2020 Programme for Research & Innovation under grant agreement n°727312.

12  
Activate Wind  
Go to Settings for a

Genomic prediction will allow for better targeted crossing, eliminating the “cross the best with the best and hope for the best” approach to breeding. Modelling will allow predicted phenotyping of crosses between different genotypes, allowing breeders to pre-select on the basis of their breeding values.

## Genomic prediction Eucleg



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

13  
Activate Wind  
Go to Settings for a

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 1<sup>st</sup> October 2021

Recording link to the presentation: <https://youtu.be/FTiZy33gSyE>



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

**EUCLEG.eu**

## 7. Genomics assisted breeding in faba bean

Dr Ana M<sup>a</sup> Torres

Senior Research Scientist, IFAPA, Córdoba, Spain

EUCLEG Faba Bean Species Expert



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



# Genomics assisted breeding in faba bean

*EUCLEG Workshop 30th September - 1<sup>st</sup> October 2021*

Ana M. Torres  
IFAPA, Centro Alameda del Obispo, Córdoba, SPAIN

### The crop: faba bean



- Fourth most widely grown cool season legume
- Worldwide cultivated area: 2,5 M has
- World production: 5 Mt
- Main producers:
  - China, Ethiopia, Australia, France



- Partially allogamous annual crop
- Small chromosome number:  $2n = 12$  ( $n = 6$ )
- No wild relative known
- Great variability within the domesticated gene pool



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

2

Faba bean is one of the oldest crops grown by man since the beginning of agriculture. Due to the high protein content and high yield potential, faba bean is today the fourth most widely grown cool season



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

**EUCLEG.eu**

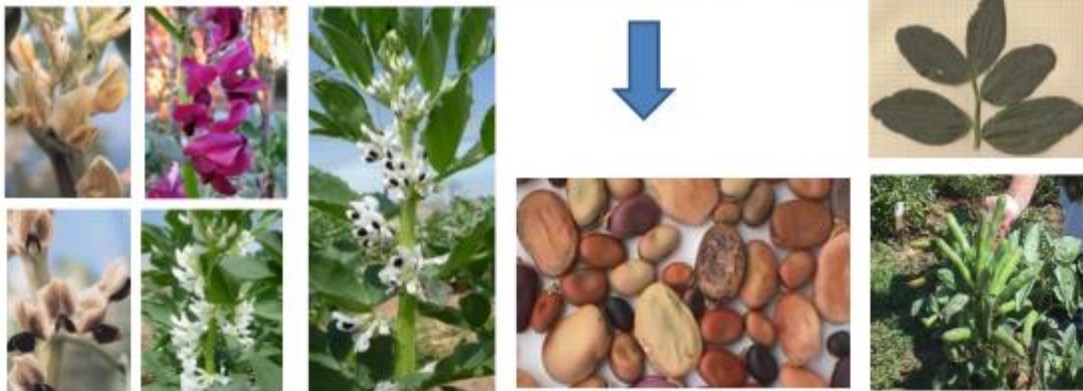


legume. The worldwide cultivated area covers 2.5M hectares and the world production accounts for 5Mt. The main producers are China, Ethiopia, Australia, and France.

Faba is partially an allogamous crop. The chromosome number is very small, only 6, and it has very large chromosomes that are easily observable and because of this was a model species for plant cytogenetics in the last century. No wild relatives are known and all the crosses that have been tried with closely related species have been unsuccessful. Nevertheless, there is a great variability within the domesticated gene pool, with the main centre located in the Middle East and with secondary centres in South America and Asia.

## Faba bean variability

Long history of cultivation, wide distribution, mating system and human selection



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

3

The long history of cultivation, the wide distribution, the mating system and the response to human selection has led to faba bean being one of the most variable crop species, with a wide spectrum of colour in flowers and also variation in seed shape, colour and size, as well as many other agronomical characteristics.



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

**EUCLEG.eu**

## Faba bean variability

Long history of cultivation, wide distribution, mating system and human selection

• Botanical groups:

- **Major** (1,11 – 1,70 g)
- **Equina** (0,61 – 1,10 g)
- **Minor** (0,41 – 0,60 g)
- **Paucijuga** (0,31 – 0,40 g)



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

4

After selection we can distinguish four botanical groups. The Major type on the right, the Equina, and Minor types, which are mainly used for animal consumption and the Paucijuga types which are thought to be very close to the wild type and they are very small, round, black seeds, mostly cultivated in India.

## Faba bean challenges

- |  |   |                                       |
|--|---|---------------------------------------|
| <ol style="list-style-type: none"> <li>1. Resistance to biotic and abiotic stresses</li> <li>2. Yield and appropriate phenology related traits</li> </ol>          | } | <b>Quantitative trait loci (QTLs)</b> |
| <ol style="list-style-type: none"> <li>3. Enhance seed quality<br/>Low tannin and vicine-convicine</li> <li>4. Crop architecture<br/>Plant growth habit</li> </ol> | } | <b>Qualitative loci</b>               |



As in many other crops, the main challenges for faba bean are the resistance to biotic and abiotic stresses, yield and to find the appropriate phenology adapted to different environments. Other problems are related to the enhancement of seed quality, especially to eliminate low tannin and vicine-convicine. Finally, there have been some efforts identify the genes which control plant growth habit.



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

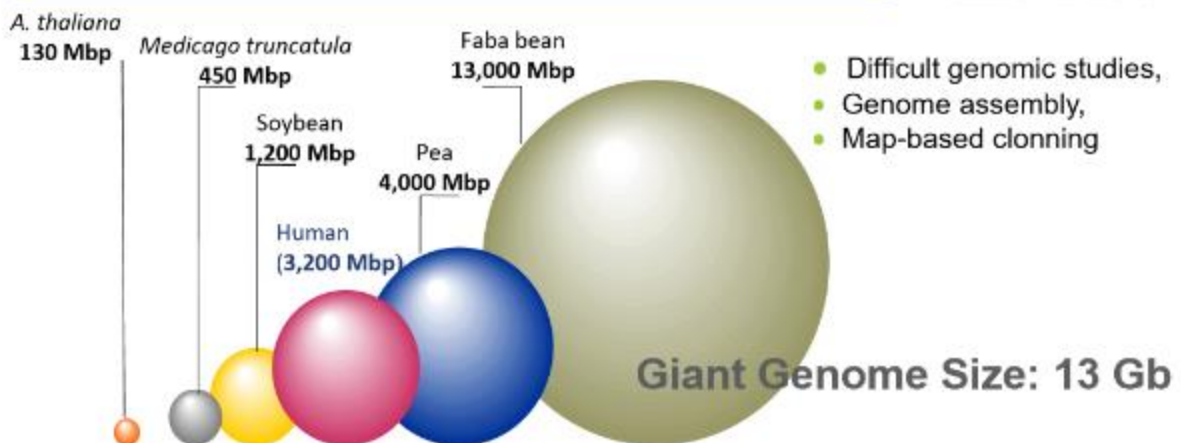
## Biotic and abiotic stresses

- **Broomrape** (*Orobanche crenata*): parasitic weed, damaging in Mediterranean basin, Northern Africa
- **Fungal diseases:**
  - ✓ Ascochyta blight (*Ascochyta fabae*)
  - ✓ Rust (*Uromyces viciae-fabae*)
  - ✓ Chocolate spot (*Botrytis fabae*)
  - ✓ Downy mildew (*Peronospora viciae*)
  - ✓ Foot rots (*Fusarium* spp.)
- **Insects:**
  - ✓ Black bean aphid (*Aphis fabae*)
  - ✓ Bruchid seed beetle (*Bruchus rufimanus*)
- **Abiotic stresses:**
  - ✓ Cold/frost tolerance
  - ✓ Heat/Drought tolerance
  - ✓ Salinity tolerance



Here I will explain in more detail the main biotic and abiotic stresses for this crop. In the Mediterranean basin, and Northern Africa the main problem are parasitic weeds, the broomrapes, that results in severe losses every year together with fungal diseases such as Ascochyta, Rust, Chocolate spot and rots. The main abiotic stresses are related with cold and frost tolerance, heat and drought tolerance and salinity tolerance. Insect pests include *Aphis* and *Bruchus*.

## Molecular breeding approaches



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

7

Faba bean has a very big genome size, 13Gb, one of the biggest among the legume crops, 3 times bigger than pea, 10 times bigger than soybean and 30 times bigger than the model *Medicago truncatula*.

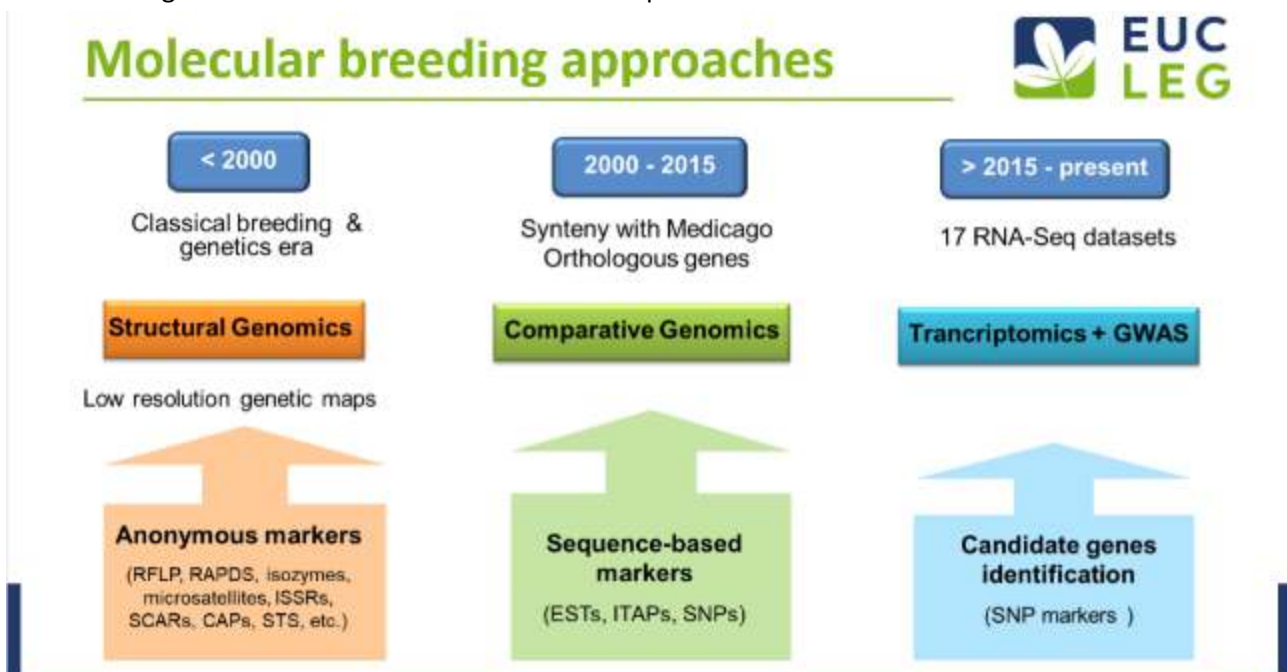


This project has received funding from the European Union's Horizon 2020 Programme for Research & Innovation under grant agreement n°727312.

**EUCLEG.eu**

Obviously, this has significant impact on genomic studies and also on genome assembly and map-based cloning.

The molecular breeding approaches and the genomic tools used in faba bean have been parallel to the development of molecular markers and techniques. At the beginning, before the 2020's there were genetic maps with very low resolution and density, mainly using anonymous markers such as RFLP, RAPDS, isozymes, microsatellites etc. Later on, the use of comparative genomics using synteny with Medicago and other crops allowed the use of orthologous markers and the development of sequence-based markers. Finally, the decrease in the cost of sequencing has allowed the development of datasets and identification of candidate genes that have been included in the maps as SNP markers



## Genetic maps (1990 - 2015)

- Selection of contrasting parentals
- Development of F<sub>2</sub> or RIL populations for mapping



- QTL validation: stability (environment, backgrounds)
- QTL saturation (positional or functional markers)



**Diagnostic markers or identify candidate genes for MAS (Marker Assisted Selection)**



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

12

Among the genomic resources I would just like to mention that as in many other crops genetic map were developed with a selection of contrasting parentals for the development of F<sub>2</sub> and RILs. These populations were used to identify QTLs, which were further validated in different environments and genetic backgrounds followed by QTL saturation using positional and functional markers with the final aim of identifying candidate genes or markers useful for Marker Assisted Selection approaches.

## Genetic maps (1990 - 2015)

References	Cross	Mapping population	No. individuals	No. markers	No. LGs <sup>a</sup>	Length (cM)	Uses <sup>b</sup>
Van de Ven et al. (1991)		BC		17	7 (-)	231	
Tomes et al. (1993)		2 F <sub>2</sub>	20	51	11 (1)	~300	
Ramsay et al. (1995)		BC		23	7 (-)	~300	
Sotercic et al. (1999)		3 F <sub>2</sub>	813	157	48 (6)	~850	T / C
Vaz Patto et al. (1999)	V8 × V27	3 F <sub>2</sub>	175	116	13 (7)	~1200	T / C
Román et al. (2007, 2007)	V8 × V1136	F <sub>2</sub>	190	121	16 (9)	1446	QTL
Román et al. (2004)		11 F <sub>2</sub>	654	192	14 (5)	1559	T / C
Ávila et al. (2004, 2005)	29H × V1136	F <sub>2</sub>	150	103	18 (6)	1308	QTL
Elwood et al. (2008)	V8 × V27	RIL	98	133	12 (-)	1685	
Arbaoui et al. (2008)	Côte d'Or × BPL14628	RIL	101	131	21 (-)	~980	QTL
Diaz et al. (2009, 2010)	V8 × V1136	RIL	165	277	21 (9)	2857	QTL
Cruz-Loquendo et al. (2012)	V8 × V27	RIL	124	258	16 (8)	1874	QTL
Ma et al. (2013)	91825 × K1503	F <sub>2</sub>	129	128	15 (-)	1587	
González et al. (2013)	29H × V1136	RIL	119	171	20 (15)	1402	QTL
CONSENSUS MAP (2013)		3 RIL	408	587 <sup>c</sup>	6 (6)	3515	C
El-Badry et al. (2014)							
Kaur et al. (2014)							



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


13



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

## Genetic maps (1990 - 2015)

Reference	Cross	Mapping population	No. individuals	No. markers	No. LGs <sup>a</sup>	Length (cM)	Uses <sup>b</sup>
Van de Ven et al. (1991)		BC		17	7 (2)	231	
Torres et al. (1995)		F <sub>2</sub>	20	51	11 (1)	~300	
Ramsay et al. (1995)		BC		23	7 (2)	~300	
Satovic et al. (1996)		F <sub>1</sub>	813	157	48 (6)	~850	T
Vic Pardo et al. (1999)	V16-V127	F <sub>1</sub>	175	118	15 (7)	~1200	T
Bonin et al. (2002; 2003)	V16-V136	F <sub>1</sub>	196	121	16 (9)	1446	QTL
Bonin et al. (2004)	a	F <sub>1</sub>	634	192	14 (5)	1559	T
Arilla et al. (2004; 2005)	2001-V136	F <sub>2</sub>	159	103	18 (6)	1308	QTL
Ellwood et al. (2008)							
Arbañal et al. (2008)							
Diaz et al. (2009; 2010)							
Carrizosa et al. (2012)							
Mao et al. (2013)							
García et al. (2013)							
CONSENSUS MAP (2013)							
El-Rasky et al. (2014)							
Kan et al. (2014)							




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

14

For example, highlighted here are the first maps used to assign the first linkage groups to chromosomes, using the progeny of trisomic plants.

## Genetic maps (1990 - 2015)

Reference	Uses <sup>b</sup>
Van de Ven et al. (1991)	
Torres et al. (1995)	
Ramsay et al. (1995)	
Satovic et al. (1996)	T
Vic Pardo et al. (1999)	T
Bonin et al. (2002; 2003)	QTL
Bonin et al. (2004)	T
Arilla et al. (2004; 2005)	QTL
Ellwood et al. (2008)	
Arbañal et al. (2008)	QTL
Diaz et al. (2009; 2010)	QTL
Carrizosa et al. (2012)	QTL
Mao et al. (2013)	QTL
García et al. (2013)	QTL
CONSENSUS MAP (2013)	
El-Rasky et al. (2014)	
Kan et al. (2014)	





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

15

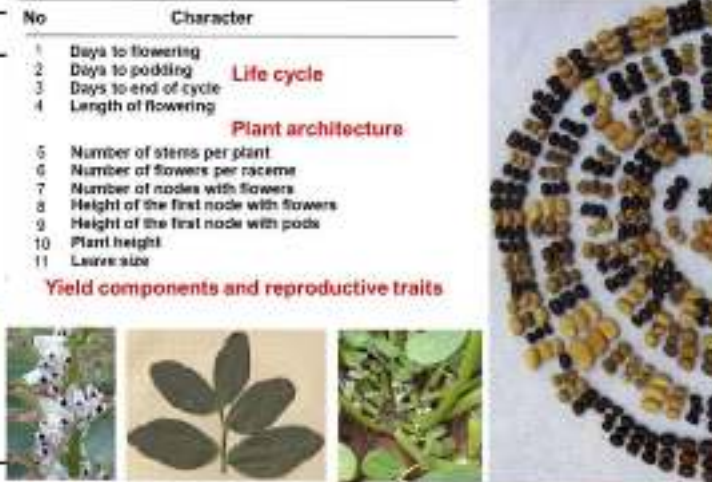
The next ones were used for the identification of QTLs for Ascochyta, Broomrape or frost resistance. Some of them have been validated in different environments. Life cycle, plant architecture, yield components and reproductive traits have also been studied in faba bean.



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

## Genetic maps (1990 - 2015)

Reference	No	Character
Monte-Vin et al. (1991)	1	Days to flowering
Torres et al. (1993)	2	Days to podding
Ramos et al. (1995)	3	Days to end of cycle
Saiz et al. (1996)	4	Length of flowering
Naz Pinto et al. (1999)	5	Number of stems per plant
Bonato et al. (2002, 2003)	6	Number of flowers per raceme
Bonato et al. (2004)	7	Number of nodes with flowers
Avila et al. (2004, 2005)	8	Height of the first node with flowers
Hilbrød et al. (2005)	9	Height of the first node with pods
Arbesi et al. (2008)	10	Plant height
Diaz et al. (2009, 2010)	11	Leaf size
Cruciani et al. (2012)		
Mu et al. (2017)		
Datiency et al. (2011)		
CONSENSUS MAP (2011)		
El-Rodeny et al. (2014)		
Kaur et al. (2014)		



**Yield components and reproductive traits**

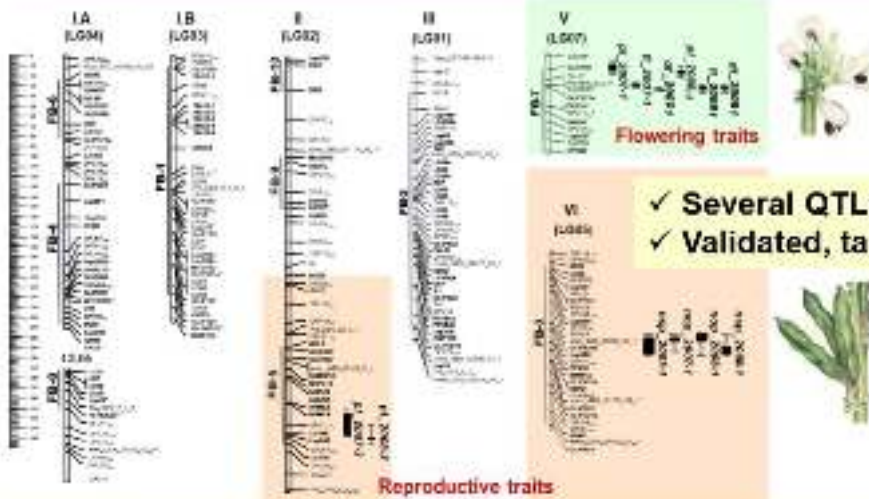


Horizon 2020 of European Union  
This project has received funding from the European Union's Horizon 2020 Programme for Research & Innovation under grant agreement n°727312.

16

Several QTLs have proved to be stable across years, these QTLs have been validated and the targeted QTL regions have been further saturated with more candidates.

## Genetic maps



**✓ Several QTLs stable**  
**✓ Validated, target regions saturated**

- Avila et al. 2017
- Catt et al. 2017
- Aguilar et al. 2021



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

17



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

## Genetic maps (1990 - 2015)

Reference	Cross	Mapping population	No. individuals	No. markers	No. LGs <sup>a</sup>	Length (cM)	Uses <sup>b</sup>
Van de Ven et al. (1991)		BC		17	7 (-)	231	
Tomes et al. (1995)		2 F <sub>2</sub>	20	51	11 (1)	~300	
Ramsay et al. (1995)		BC		23	7 (-)	~300	
Satovic et al. (1996)		7 F <sub>2</sub>	813	157	46 (6)	~850	T / C
Van Pelt et al. (1999)		5 F <sub>2</sub>	175	116	13 (7)	~1200	T / C
Román et al. (2002, 2003)		F <sub>1</sub>	196	121	16 (9)	1446	QTL
Román et al. (2004)		11 F <sub>1</sub>	654	192	14 (5)	1579	T / C
Ávila et al. (2004, 2005)		F <sub>2</sub>	159	103	18 (6)	1308	QTL
<b>Ellwood et al. (2008)</b>	→ <b>The first gene-based genetic map anchored with orthologous markers from <i>Medicago truncatula</i></b>						
Crujeira-Queiroz et al. (2012)		RIL	124	258	16 (8)	1874	QTL
Mé et al. (2013)		F <sub>1</sub>	129	128	15 (-)	1587	
Gutiérrez et al. (2013)		RIL	119	171	29 (15)	1402	QTL
Satovic et al. (2013)		3 RIL	408	837	6 (6)	5515	C
<b>Kaur et al. (2014)</b>	→ <b>The first exclusively SNP-based genetic map</b>						



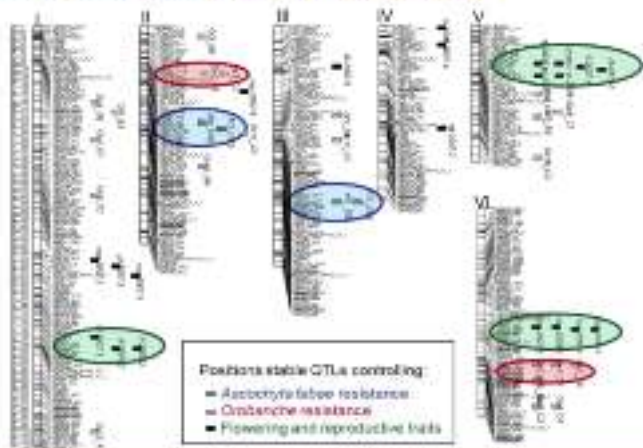
Horizon 2020 of European Union  
This project has received funding from the European Union's Horizon 2020 Programme for Research & Innovation under grant agreement n°727312.

18

In green are highlighted the first gene-based genetic map, anchored with orthologous markers from *Medicago truncatula* and next, the first exclusively SNP-based genetic map.

## Consensus maps

**Satovic et al. 2013: 3 RIL**  
- 729 anonymous and gene-based markers



Development of sequences  
+ marker datasets



Increase in density and utility  
of gene-based genetic maps



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

19

The development of new sequences and marker datasets has increased in density and utility of gene-based genetic maps. Above is the first consensus map that was developed using 3 RIL populations and also shows




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



the position of stable QTLs controlling *Ascochyta* or *Orobancha* resistance and flowering and reproductive traits.

### Consensus maps

**Webb et al. 2016: SNP-based consensus map (6 RILs) →**  
**- 687 SNP markers on six linkage groups**



Albus = BPL10,  
 Albus = 25%,  
 Hedra = CGN07715 cl-3,  
 Nyb44-1 = IC 12658,  
 Mändle/2 × il B 538/2,  
 Côte d'Or/1 = BP14528/1521

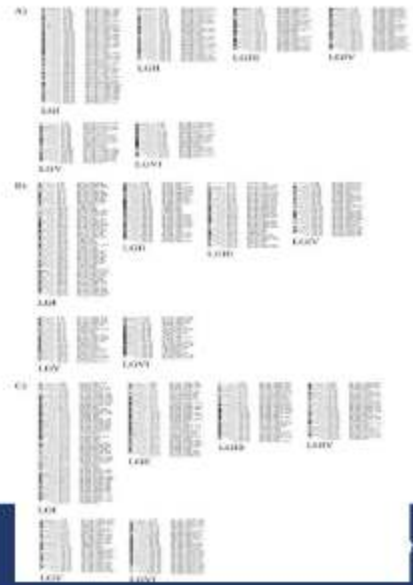
- ✓ Mine and validate > 800 SNPs from transcriptomes + most informative SNP assays from previous maps
- ✓ **KASP** (Kompetitive Allele-Specific PCR, LGC genotyping technology → enables accurate bi-allelic discrimination of known SNPs and InDels

20

Webb et. al. in 2016, reported a SNP-based consensus map using 6 RIL populations. Markers were obtained after mining and validating the SNPs from different transcriptomes and included SNP assays from previous studies. This effort was very useful for the crop allowing the development of a genotyping technology that is being widely used by the faba bean community today, to perform accurate bi-allelic discrimination of SNPs and InDels in different mapping studies.

### Consensus maps

**Carrillo-Perdomo et al. (2020)**



- Most saturated consensus map (3 F<sub>3</sub>s)
- **1728 SNP markers**
- Transcriptome data from **4 accessions** (HIVERNA, NOVA GRADISKA, SILIAN, QUASAR)

↓


- **105,828 gene-based SNPs**
- Identification of candidate genes of agronomic interest through synteny-based approaches

1



The last consensus map was published by Carrillo-Perdomo et al. in 2020, using many more SNP markers from the 3 F<sub>3</sub>s population. In addition to the map they have also developed a transcriptome from 4 accessions providing gene-based SNPs which will be very useful for the identification of candidate genes for agronomic interest for this crop.

## Other relevant maps



Cross	Population type	No. Individuals	Map length	Mapped traits	References
Icarus × Ascot	F <sub>2</sub>	95	1217	Ascochyta blight resistance, flowering time	Kaur et al. (2014) Catt et al. (2017)
Mélodie/2 × ILB 998/2	F <sub>2</sub>	211	928	Drought adaptation, vicine-convicine	Khazaei et al. (2014a, 2014b, 2015)
Fjord × Dazza#12034	F <sub>2</sub>	104	1027	Rust resistance	Jaz (2018)
Nura × Farah	F <sub>4</sub>	145	1022	Ascochyta blight resistance	Sudheesh et al. (2019)
VF6 × VF27	F <sub>2</sub>	124	4421	Pod dehiscence	Aguilar-Benitez et al. (2020)
VF6 × VF27	F <sub>2</sub>	124	4421	Flowering time	Aguilar-Benitez et al. (2021)

**Bi-parental populations → QTL number and resolution LOW, limited recombination (two alleles/locus, low genetic diversity between two parents)**



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

22

Finally, a number of other genetic maps have been developed specifically looking at traits related to drought adaptation, rust resistance, pod dehiscence and flowering time. All of them derive from the bi-parental populations that are very easy to construct and represent a powerful tool for QTL detection although the number of QTLs and their resolution is very low since we have limited recombination (only 2 alleles and relatively low genetic diversity between two parents).

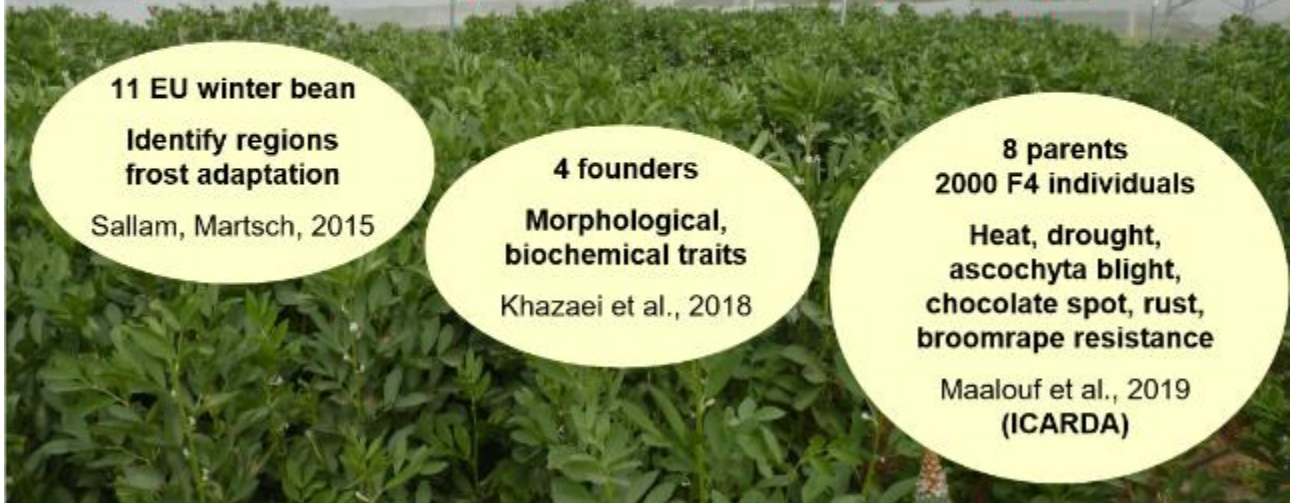


This project has received funding from the European Union's Horizon 2020 Programme for Research & Innovation under grant agreement n°727312.

**EUCLEG.eu**

## MAGIC populations

### Multiparent advanced generation intercross (MAGIC) populations



For this reason, in the past years there has been a great interest in developing multiparent advanced generation intercross (MAGIC) populations. For example in 2015 using 11 EU winter faba beans, to identify regions for frost adaptation; the second example in 2018 looking at morphological and biochemical traits and the last one in 2019 represents a quite important effort because it used 8 parents segregating for both biotic and abiotic stresses. Of course the use of a wider genetic diversity existing in the multiple parents and the recombination along several generations produce a population useful for high-resolution mapping, as compared with the traditional bi-parental mapping populations.

## Faba bean transcriptomes

References	Aim of study	Tissue	Output	NGS platforms
Ray and Gougeon (2002)	Development ESTs	embryo	5000 ESTs	454 sequencing
Kear et al. (2012)	EST-SSRs	leaf, stem, flower, pod, seed	600 SSRs	454 Fabae GS TM Illumina
Kear, Kishor, et al. (2014)	SNP markers	leaf	100 SNPs	Illumina GS, bead array
Kay et al. (2015)	SNP markers for agronomical traits	developing seed coat, root, shoot	1,200 SNPs	454 sequencing
Arora-Chinnappa and McCurdy (2015)	Transcriptome	leaf, stem, flower, root, seedling	17,162 unigenes	Illumina HiSeq2000
Coats et al. (2015)	Transcriptome under waterlogging condition	leaf three other tissues	21,248 transcripts, 22,062 SNPs, 360 lncRNAs	Illumina
Went et al. (2017)	SNP discovery	seedling	668 SNP markers	454 FUSION reads
Brodthorn et al. (2017)	single cell	pod, flower	16,259 transcripts	RNA-seq, Illumina
Compadri et al. (2017)	genome resources	embryo	17,500 unigenes	RNA-seq, Illumina HiSeq
Alghamdi et al. (2018)	drought stress	root, flowering	15,327 genes	RNA-seq, Illumina HiSeq
Caicedo et al. (2018)	seedling	seedling	6522 genes	RNA-seq, Illumina HiSeq
Wang et al. (2018)	seedling stress	seed	2496 differentially expressed genes	RNA-seq, Illumina HiSeq
Camillo-Ferdone et al. (2020)	SNP discovery	leaf	22,423 transcripts	RNA-seq, Illumina HiSeq
Agamotom et al. (2020)	biophysical and metabolic	leaf, flower, seed, seedling	48,077 transcripts	RNA-seq, Illumina HiSeq

### 17 transcriptomes



### Reference gene set

for differential gene expression analysis and genome annotation

• NGS platforms, RNA-seq technology → enhanced speed of faba bean gene discovery



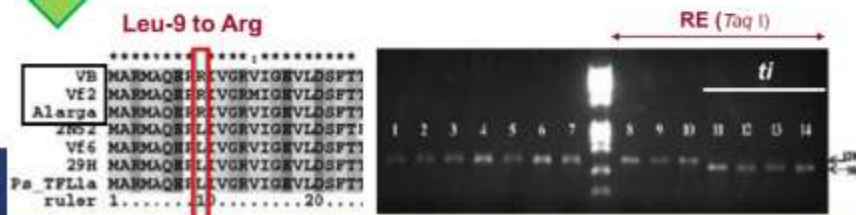
This project has received funding from the European Union's Horizon 2020 Programme for Research & Innovation under grant agreement n°727312.

I would like also to highlight the importance of the Next Generation Sequence platforms (NGS), especially the RNA-seq technology to enhance the speed of the faba bean gene discovery. There have been 17 important transcriptomes in faba beans for different varieties, different issues, different developmental stages etc., which have been very useful for developing new markers, ESTs, SNP markers and also to identify genes that were differentially expressed for *Ascochyta*, drought stress, salinity. This has produced an excellent reference gene set, which is very useful for differential gene expression analysis and genome annotation.

## Gene discovery: *ti* (terminal inflorescence)



- ✓ *ti* facilitates crop management and mechanical harvesting
- ✓ Translational genomics proved **TERMINAL FLOWER (TFL1)** controls determinacy in *Vf* as in *Medicago*, *Arabidopsis*, soybean and other legume and non-legume species
- ✓ Sequence alignment → a non-synonymous aa change
- ✓ **dCAP: diagnostic marker for determinate growth habit**



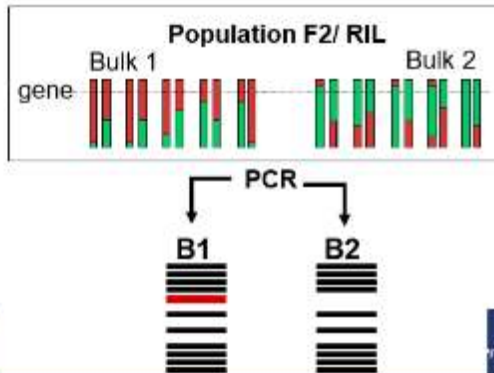
26

In the next few slides I am going to show you some examples of gene discovery that have been developed so far in faba bean. The first one was the terminal inflorescence, a trait which facilitates crop management and mechanical harvesting. Using a translational genomics approach was proved that the Terminal flower, TFL1, controls determinacy in faba bean, as happens in *Medicago*, *Arabidopsis* and many other species. So after the sequence alignment of contrasting genotypes, the authors (Avila et al. 2007) were able to find a non-synonymous amino acid change, which allowed the development of dCAP, the first diagnostic marker useful to select for determinate growth habit.

## Gene discoveries: quality traits

1. Tannin content (2 genes *zt1* and *zt2*, chr. II and III)
2. Vicine-convicine content (v-c, chr. I)

- Lower protein digestibility and energy content in animal feeding
- V-C produce **favism** in genetically predisposed humans

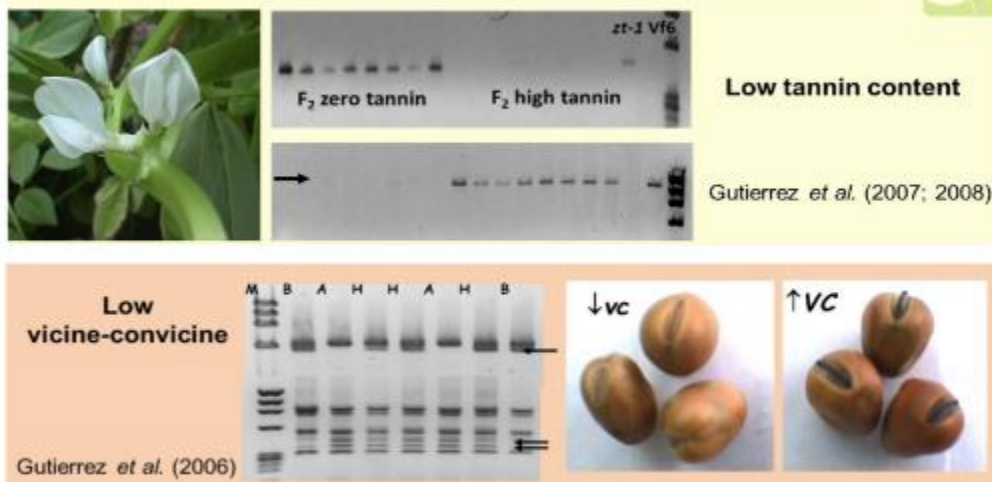


programme for Research & Innovation under grant agreement

27

The second example was focussed in quality traits with the aim of eliminating anti-nutritional compounds (such as tannin and vicine-convicine (V-C) content) from the seeds. Tannin content is controlled by 2 genes, which are located on chromosome 2 and 3, while the vicine-convicine content is located on chromosome 1. Both compounds lower the protein digestibility and energy content when faba bean seeds are used in animal feeding. Moreover, V-C produces favism, a type of anaemia present in genetically predisposed humans. So in both quality traits different molecular studies (based on bulk segregant analysis), were carried out to identify markers linked to these compounds.

## Gene discoveries



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

28

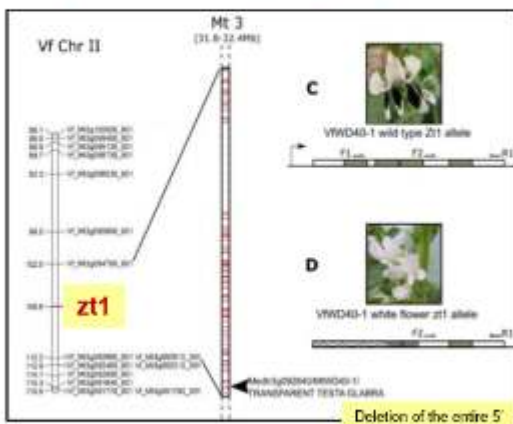


This project has received funding from the European Union's Horizon 2020 Programme for Research & Innovation under grant agreement n°727312.

White flowered plants produce seeds without tannins. The evaluation of F2 progenies segregating for flower colour or for V-C content allowed the identification of markers linked to both antinutritional traits. These markers were transformed into SCARs, which were used for marker assisted selection, in the same genetic background. At this moment no candidate genes were identified.

## Gene Discovery: *zt1*

### Comparative mapping approach with *Medicago truncatula*



Transparent Testa Glabra 1 (TTG1) a WD40 TF determining flower color in *Mt* is the **faba bean *zt1*** (Webb et al. 2016)

*zt1* characterized and **allele-specific diagnostic marker** developed to differentiate *zt1* in MAS (Gutiérrez and Torres, 2019)



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

30

Much later using a comparative mapping approach with *Medicago truncatula* Webb et al. 2016 were able to identify TTG1, a WD40 TF which determine the flower colour in *Medicago* as the gene controlling white flower in faba bean. This TTG1 was later characterised and an allele-specific diagnostic markers was developed that allow to differentiate *zt1* from *zt2* genotypes and from the wild genotypes with normal coloured flower.



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

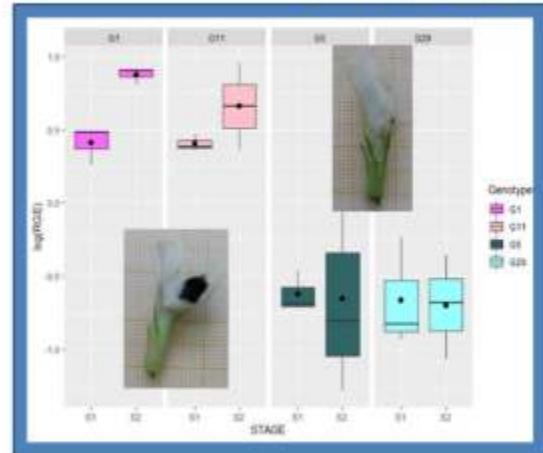
## Gene Discovery: *zt2*

**Gutiérrez N, Avila C, Torres AM (2020):**

- ✓ Comparative genomics *Mt*
- ✓ Candidate gene approach
- ✓ Linkage mapping (3 pop)
- ✓ Gene expression



- ✓ Fine map *zt2*
- ✓ Identify **TT8** (TRANSPARENT TESTA8), a basic helix-loop-helix (**bHLH**) TF as the locus underlying *zt2*



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

31

Another example of gene discovery is *zt2*. Using a combination of comparative genomics, candidate gene approach and linkage mapping (in 3 populations). Using this combined approach, the authors were able to fine map *zt2* and to identify TT8, a basic helix-loop-helix bHLH, as the locus underlying the *zt2* gene, which also determines the white flower in faba bean.

## Gene discovery: *vc*

- Most difficult faba bean quality trait to discover
- Numerous **SCARs** and **KASP markers** developed (Gutiérrez et al 2006, 2007, 2016; Khazaee et al 2015, 2017, 2019, 2020, etc.),
- Large intervals + v-c biosynthetic pathway unknown  
→ **hampered saturation** with functional candidates

**Björnsdotter et al., 2020:** the first enzyme associated with v-c biosynthesis (**VC1**) encodes a **bifunctional riboflavin biosynthesis protein (RIBA1)**



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

32

Finally the second most important antinutritional compound in faba bean (vicine-covicine), has been the most difficult trait to discover. Despite the numerous SCARs and KASP markers developed and the many publications related with the V-C, the position of this gene in the map still had very large intervals and the



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

v-c biosynthetic pathway was unknown because this compound is not present in model species. Both facts were hampering the saturation of the target region with functional markers.

Only recently Björnsdotter et al. 2020 reported the identification of the first enzyme associated with the vicine-convicine biosynthesis. The authors show that VC1 co-locates with the major QTL for vicine and convicine content and that the expression of VC1 correlates highly with vicine content across tissues. VC1 encodes an enzyme normally involved in riboflavin biosynthesis from the purine GTP.

## Faba bean genomics before EUCLEG

### SUMMARY:



- **ONLY** genes for **quality (tannins, v-c)** and ***ti*** identified /**markers available for MAS**
- **QTLs abiotic and biotic stresses reported**, but validation + saturation genomic regions needed to uncover reliable marker-trait associations → **NO gene/markers available MAS**
- **NO reference genome**
- **Syntenly with *Mt*** (translational genomics) → identify candidates in colinear regions
- **Transcriptomes** → high density gene-based maps
- → genes differentially expressed, candidates for genomic-assisted breeding



To summarise the position of faba bean genomics before the EUCLEG project, only 4 genes for quality traits and *ti* had been identified and markers for marker assisted selection (MAS) available. Several QTLs for biotic and abiotic stresses had been reported. Some of them have been validated but the saturation of the target genomic regions is still needed to uncover reliable marker-trait associations. As a result, no responsible genes have been identified so far and no markers are available for MAS.

As I mentioned before faba bean doesn't have a reference genome. Syntenly with *Mt* has now allowed to identify candidates in colinear regions and the recent transcriptomes have facilitated the development of high density gene-based maps and the identification of differentially expressed genes that are candidates for genomic-assisted breeding.



## EUCLEG (2017-2021)



**AIM:** reduce Europe and China's dependency on protein imports by developing efficient breeding strategies for major economic legume crops (alfalfa, red clover, pea, **faba bean**, soybean)

Crop diversification, crop productivity, yield stability, protein quality



### OBJECTIVES:

- **Broaden genetic base of the crop and analyse the genetic diversity**
- **Analyse the genetic architecture of key breeding traits** using genome-wide association studies (**GWAS**) → Markers related to phenotypic traits
- **Evaluate the benefits brought by genomic selection (GS)** to create new varieties.



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

34

The aim of the EUCLEG project for faba bean is to improve crop diversification, crop productivity, yield stability and protein quality.

The objectives are to broaden the genetic base of the crop and to analyse the genetic diversity. To analyse the genetic architecture of key breeding traits using GWAS, and to evaluate the benefits brought by genomic selection to create new varieties.

## GWAS in faba bean



- **Sallam et al. 2016. Identification and verification of frost tolerance QTLs**
  - 188 winter faba bean lines
  - 156 SNPs (KASPar)
- **Faridi et al. 2021. GWAS of faba bean resistance to *Ascochyta fabae***
  - 188 winter faba bean lines
  - 1829 AFLP and 229 SNP markers
- **EUCLEG: step forward the faba bean GWAS analysis**
  - **400 faba bean lines** / key agronomic traits / environments
  - **Genotyping: Vfaba\_v2 Axiom array** (Angra & O'Sullivan, 2017)



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

35

At this time only 2 GWAS studies have been reported for faba bean, the first one to identify frost tolerance and the second to identify genes for resistance to *Ascochyta*. In both cases they used the same faba bean

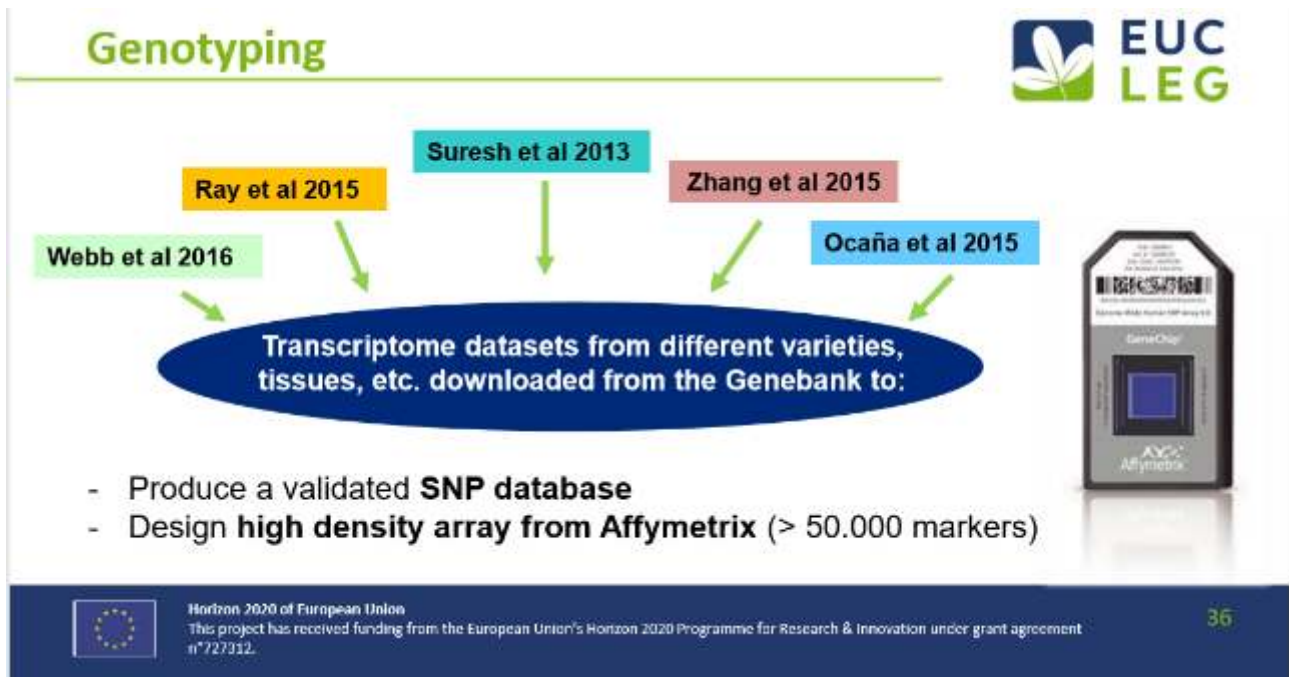


This project has received funding from the European Union's Horizon 2020 Programme for Research & Innovation under grant agreement n°727312.

**EUCLEG.eu**

population, 188 winter faba beans and a number of SNPs of 156 in the first case and nearly 2000 in the second case.

The EUCLEG project represents a step forward for faba bean GWAS analysis, because we are analysing 400 faba bean lines for many key agronomic traits which have been analysed in different environments. In addition to this, genotyping has been performed with the new Vfaba Axiom array that has been developed by Angra and O’Sullivan, 2017.



Dr O’Sullivan collected the transcriptomes datasets from 5 different sources, from different varieties, tissues etc. in order to produce a validated SNP database and thus design a high-density array from Affymetrix, which has more than 50 000 markers. This is the array that has been used to genotype the samples used in the EUCLEG project.

## Seed multiplication and shipping

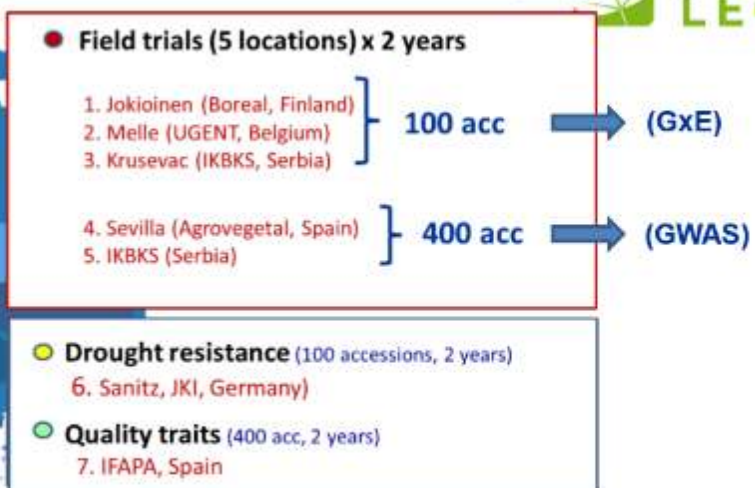


Horizon 2020 of European Union  
This project has received funding from the European Union's Horizon 2020 Programme for Research & Innovation under grant agreement n°727312.

37

The 400 faba bean lines were first multiplied for 2 years in order to produce the required amount of seeds to be distributed among partners. In order to avoid cross-pollination by insects, the multiplication was performed in insect proof cages.

## Partners and activities



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

38


We then sent these materials to the different partners for the different activities. We have set up field trials in Finland, Belgium and Serbia in which genotype and environmental interactions were performed with 100 accessions. Then 400 accessions were set up in Spain, Serbia and Wales and the results were analysed using



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

GWAS analysis. In addition to that drought resistance was studied in Germany and quality traits were centralised at IFAPA, Spain where the 400 faba bean accessions were analysed over 2 years.

## Phenotyping



**43 agronomic and adaptive traits evaluated**

**Architecture**

- Branch number
- Plant height
- Height first pod

**Morphology**


- Leaflet size
- Seed color
- Hilum color

**Phenology**

- Flowering date
- Flowering
- Flower color
- Flowers/node
- Flowered nodes/plant
- Height of the first pod
- Pods/node
- Pod length
- Maturity date

**Yield**

- Pods per plant
- Seeds per pod
- Seed weight
- 100 seed weight
- Plot yield




from the European Union's Horizon 2020 Programme for Research & Innovation under grant agreement

39

Here I have summarised the 43 agronomic and adaptive traits that we have evaluated. The traits are joined in four groups: Architecture, Morphology, Phenology and Yield.





## Phenotyping




**Stress resistance**

- *Orobanche crenata*
- Chocolate spot
- Ascochyta blight
- Rust (*Uromyces fabae*)
- *Fusarium* spp. (root diseases)
- *Sitona* spp.
- *Bruchus rufimanus*
- Aphids
  
- Lodging tolerance
- Shattering
- Drought

- Natural occurring pest and diseases



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

40

In addition, partners recorded the stress resistance to naturally occurring pests and diseases.

## Phenotyping



### Quality

- Protein content (%)
- Tannin (vanillin test and HPLC-MS)
- Vicine-convicine (HPLC-MS)

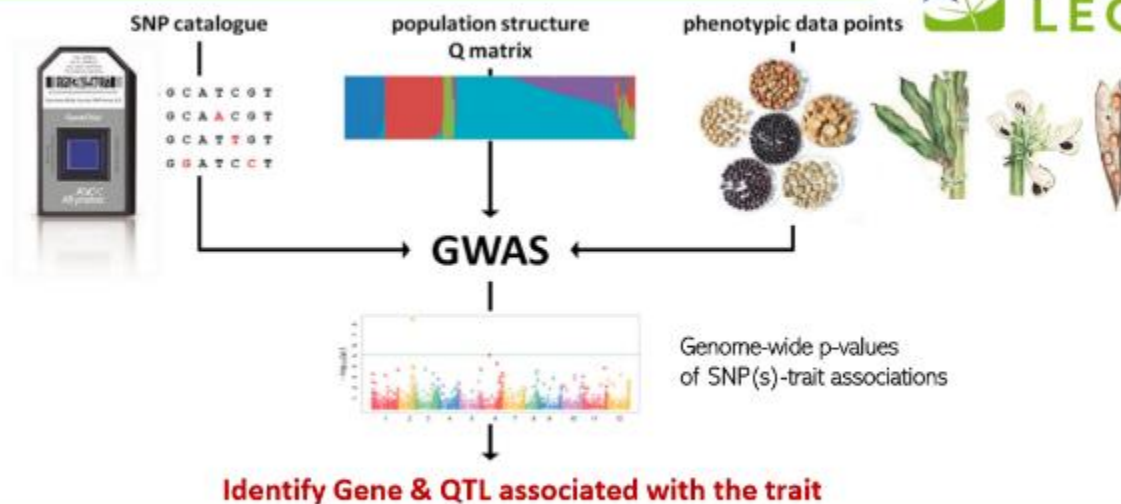


Horizon 2020 of European Union  
This project has received funding from the European Union's Horizon 2020 Programme for Research & Innovation under grant agreement n°727312.

41

Concerning the quality traits, we have analysed the protein content, the tannin and the vicine-convicine content of these accessions.

## GWAS



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

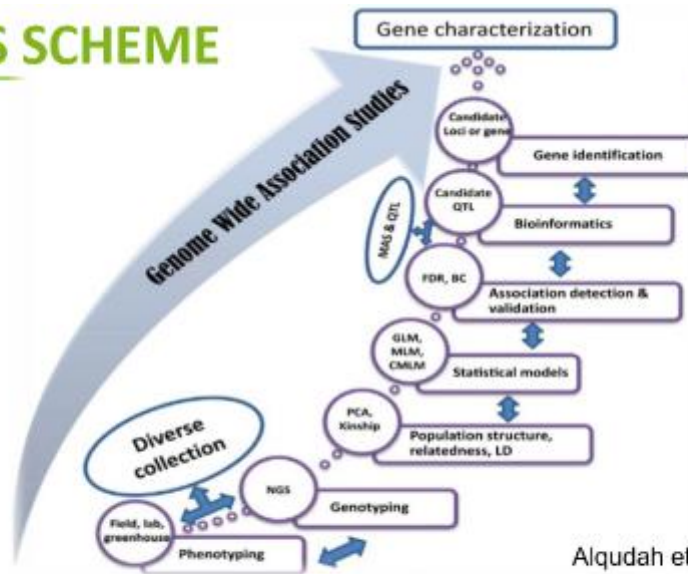
42

Once we had available the phenotypic field assays data and the genotypic information, we were able to start our first GWAS analysis. GWAS is based on a high density SNP catalogue, the population structure and the information from the phenotypic data points. So by doing GWAS you are able to identify genes and QTL associated with the traits.



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

## GWAS SCHEME



Alqudah et al. 2020



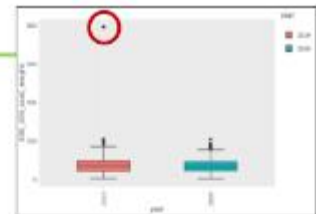
Horizon 2020 of European Union  
This project has received funding from the European Union's Horizon 2020 Programme for Research & Innovation under grant agreement n°727312.

43

We have followed this scheme, after phenotyping and genotyping we have analysed population structure and relatedness and LD, and used statistical models in order to identify associations and validating them to finally identify and characterize the genes linked to specific key traits.

## 1. Faba bean phenotyping

- Boxplots for raw phenotypic were filtered from **outliers**
- **Broad-sense heritability estimated**
- **Best linear unbiased predictor (BLUP)** used to adjust data across locations or years → better estimates of phenotypic values considering G x E



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

44

Once we received the phenotypic data from the farmers we filtered the outliers and established the broad sense heritability and also used the BLUP to adjust the data across locations and years.



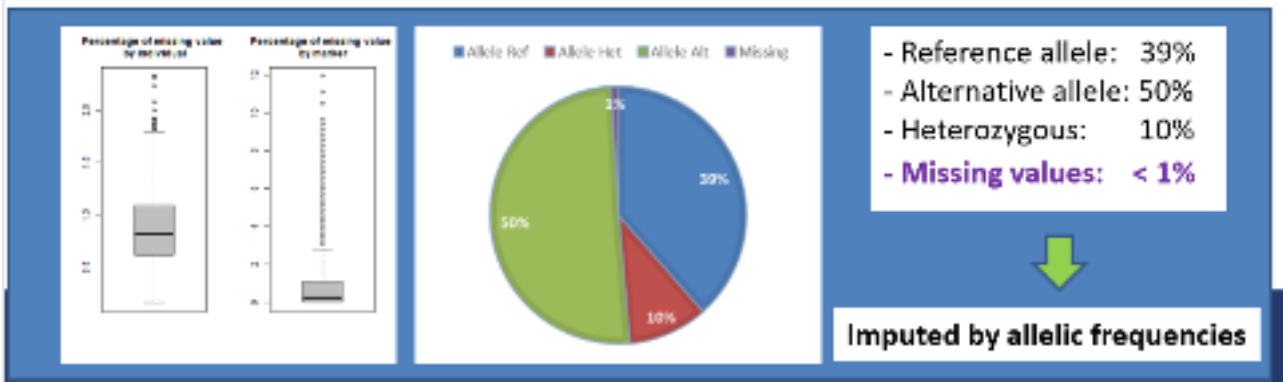
This project has received funding from the European Union's Horizon 2020 Programme for Research & Innovation under grant agreement n°727312.

## 2. Faba bean genotyping

**QUALITY CONTROL ANALYSIS** → Remove:

- Markers with MAF < 0.05
- Monomorphic markers (1 allele or 99% equal)

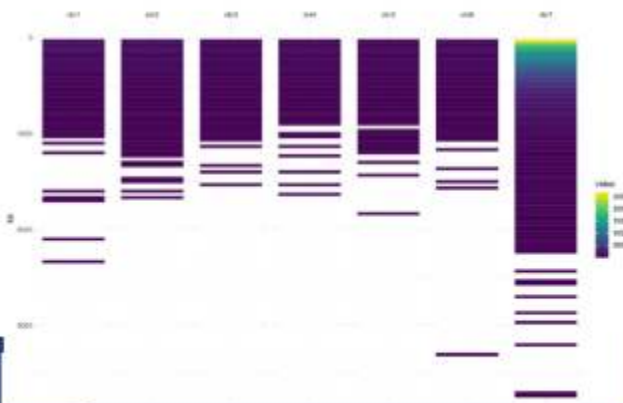
**Score matrix include 352 samples with 22.867 SNPs**



The faba bean genotyping data was also subjected to quality control analysis by removing markers with MAF allele frequencies of 0.05 and any monomorphic markers. The score matrix matrix included 352 samples with 22.867 SNPs. Missing values were inputed by allelic frequencies.

## 2. Faba bean genotyping

- **Genome Zipper:** exploits the **synteny with *Medicago truncatula*** to identify, order and structure chromosomal sequences of large genomes that lack physical maps
- **Genetic maps positions** in 3 RIL populations



**Score matrix: 22.867 SNPs**

Nº Chromosome	Nº SNPs markers
Vf01	1091
Vf02	930
Vf03	545
Vf04	655
Vf05	533
Vf06	705
<b>TOTAL</b>	<b>4,459 (17%)</b>

programme for Research & Innovation under grant agreement.

46

As I mentioned before and due to the huge genome size, we lack genome sequence and physical maps, so we use the Genome Zipper, which exploits the synteny with *Medicago truncatula* to identify, order and structure chromosomal sequences of chromosomes. Moreover, we used the genetic map positions



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

**EUCLEG.eu**

observed in 3 RIL populations genotyped with the same Vfab array. By doing so we were able to assign 17% of the genes to the 6 faba bean chromosomes.

### 3. Population structure

- Calculate **relatedness correlation** among individuals due to mixture/ historical structure.  
**Methods** used:
  - **Principal components analysis (PCA):** considers the genotypic data to deduce genetic variation that can be explained by a small number of dimensions.
  - **Kinship matrix (K):** calculate the relatedness between pairs of individuals using genotypic information.
  - **STRUCTURE:** identify subsets of the whole sample by detecting allele frequency differences within the data and assign individuals to those sub-populations based on analysis of likelihoods.



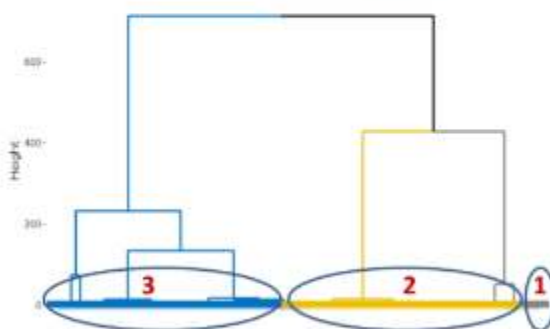
Horizon 2020 of European Union  
This project has received funding from the European Union's Horizon 2020 Programme for Research & Innovation under grant agreement n°727312.

47

We performed population structure that as you know calculates the relatedness correlation among individuals due to mixture and historical structure. We used different methods including Principal Component Analysis, Kinship matrix and Structure.

### 3. Population structure

- **Principal Components Analysis (PCA):**
  - The first 3 PCs explained 88 % of the total variance



#### Cluster dendrogram

1. Exotic lines
2. Mediterranean accessions
3. Northern types



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

48

As you can see here the PCA divided the population in 3 groups: Exotic lines, Mediterranean accessions and Northern types. The 3 main PCs explained 88% of the variance.

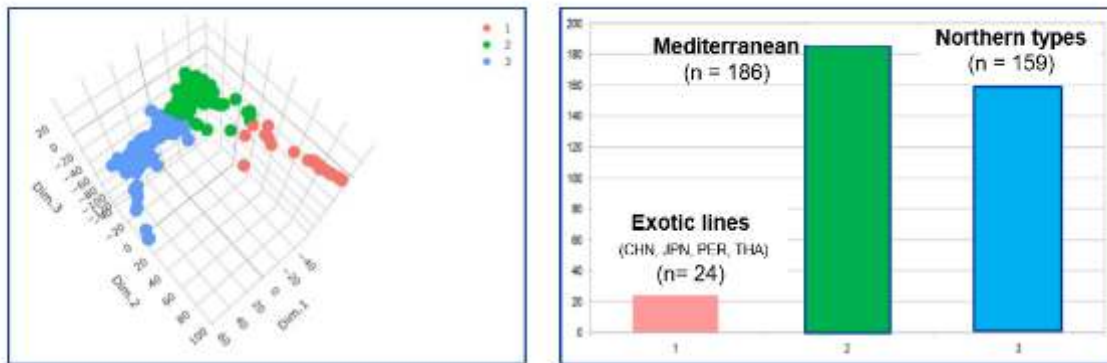


This project has received funding from the European Union's Horizon 2020 Programme for Research & Innovation under grant agreement n°727312.



### 3. Population structure

- **Principal Components Analysis (PCA):** 3D scatter plot



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

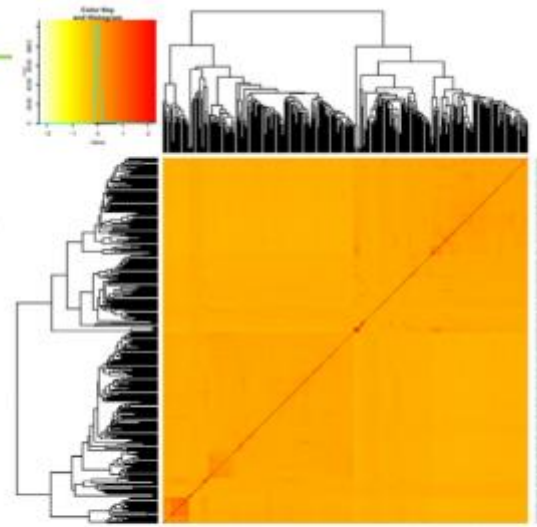
49

Here are same results, but using a 3D scatter plot and we can clearly see the 3 groups described above.

### 3. Population structure

- **Kinship analysis (GAPIT)**

- Genetic clustering heat map for evaluating the genetic differences among accessions
- The association mapping panel was divided in groups, with considerable genetic differences among lines



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

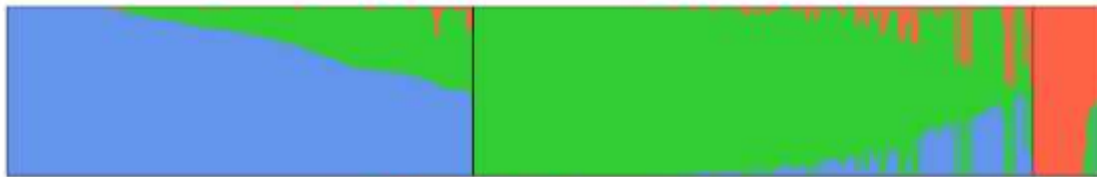
50

GAPIT also displays a heat map. This heat map shows the genetic differences between the accessions and also distinguishes the differences between the groups.

### 3. Population structure

- **FastSTRUCTURE**

Algorithm for inferring population structure from large SNP genotype data, based on a variational Bayesian framework



1. Exotic lines
2. Mediterranean accessions
3. Northern types



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

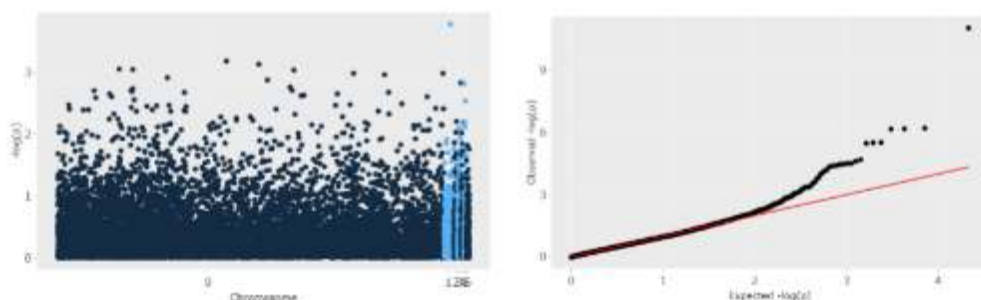
51

Finally, the FastSTRUCTURE based on variational Bayesian framework also distinguishes the 3 groups as I mentioned before, Exotic lines, Mediterranean accessions, and Northern types.

### 4. Statistical models (GWAS)

- **Trait: Plot yield**
- **Multi-Locus Mixed Model (PROGENO)**

- Manhattan and Q-Q plots → No significantly SNPs associated



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

52

Then we started to analyse which statistical models would provide the best results. So we first analysed plot yield using the Multi-Locus Mixed Model (PROGENO), but after the Manhattan and Q-Q plots no significant SNPs were associated.

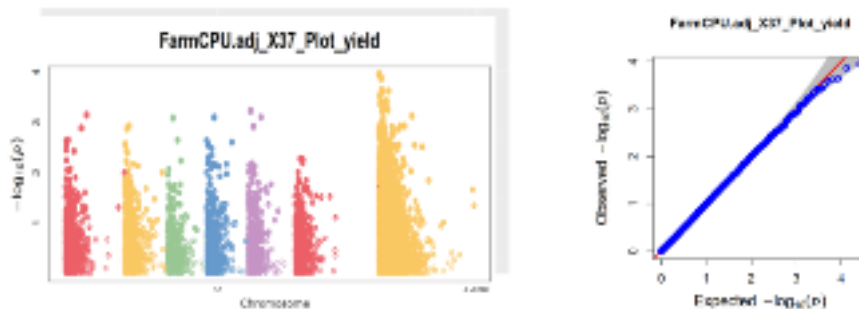


This project has received funding from the European Union's Horizon 2020 Programme for Research & Innovation under grant agreement n°727312.

## 4. Statistical models (GWAS)

- **Trait: Plot yield**
- **FarmCPU method (GAPIT)**

- Manhattan and Q-Q plots → **No significantly SNPs associated**



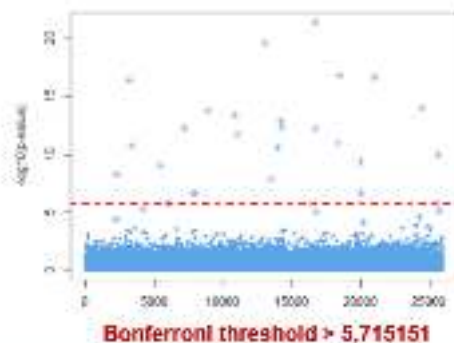
Horizon 2020 of European Union  
This project has received funding from the European Union's Horizon 2020 Programme for Research & Innovation under grant agreement n°727312.

53

The same happened when we used the GAPIT FarmCPU method. No significant SNPs were associated.

## 4. Statistical models (GWAS)

- **Trait: Plot yield**
- **Multi-Locus Mixed Model (mlmm.gwas, R package)**



- **25 SNPs significantly associated**
- **83,7 % phenotypic variance**



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

54

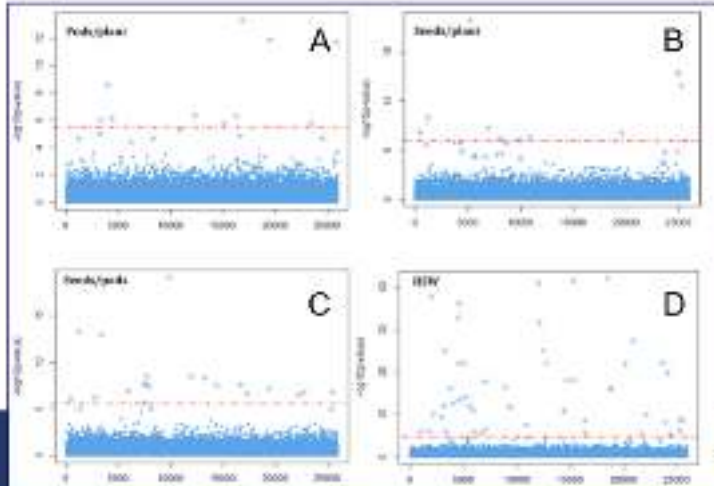
Later on we used the Multi-Locus Mixed Model, implemented in the R package and we were able to identify 25 SNPs significantly associated, able to explain nearly 84% of the phenotypic variance. Once we established this, we started to use the same statistical model for other phenotypic traits such as pods per plant, seeds per plant, seeds per pod and hundred seed weight. As you can see we identified significant SNPs for all these traits.



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

## 4. Statistical models (GWAS)

- **Multi-Locus Mixed Model (mlmm.gwas, R package)**



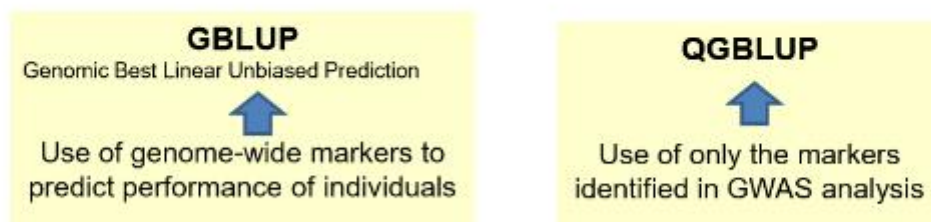
### RESULTS :

- A. Pods per plant → 10 QTLs
- B. Seeds per plant → 11 QTLs
- C. Seeds per pod → 19 QTLs
- D. Hundred seed weight (HSW) → 31

Manhattan plot of related yield components  
SNPs significantly associated are showed above the Bonferroni threshold (red line).

## 5. Genomic selection (GS)

- **GS uses large marker sets (SNPs) to predict the breeding value and select the best candidates for further breeding**
- Prediction model is based on the **genotypic** and **phenotypic** data of training population → **Genomic Estimated Breeding Value (GEBV)**. Tested different **METHODS**:



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

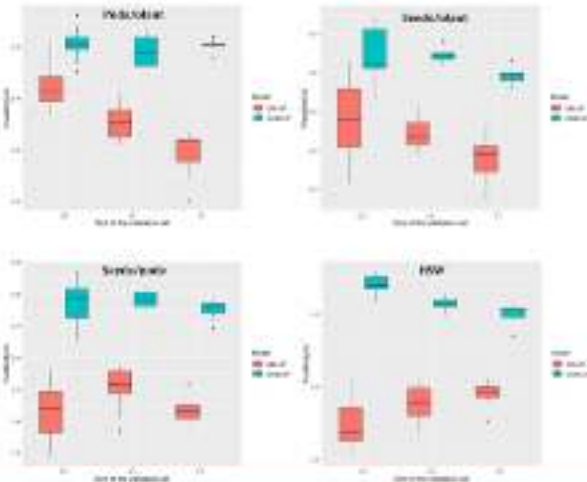
56

We have also started to perform some genomic selection studies. We used large marker sets to predict the breeding value and to select the best candidates for further breeding. So the prediction model is based on genotypic and phenotypic data of training populations, in order to estimate the genomic breeding value. We tested 2 methods: the GBLUP and the QGBLUP. In the first case we used the genome wide markers and in the second case we used only the markers identified in GWAS analysis.



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

## 5. Genomic selection (GS)



- For all the QTLs, **QGBLUP** gave better prediction than **GBLUP**.
- Decrease in size of the **training population** tended to decrease the quality of the genomic prediction

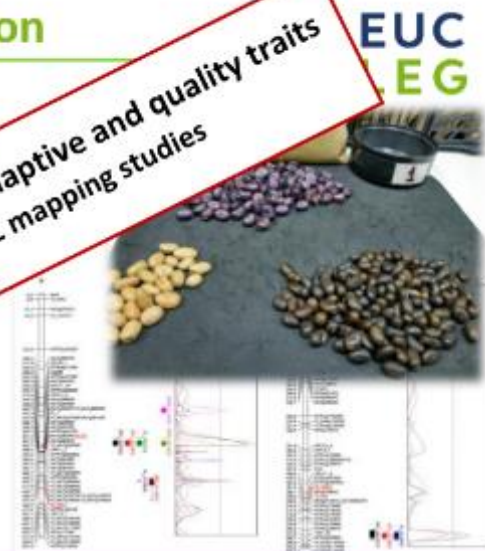
Effect of the genomic model (**GBLUP/QGBLUP**) and size of the validation set on the predictive ability for related yield components in faba bean

As you can see above for all the traits mentioned before and for all the QTLs, the QGBLUP gave better prediction than GBLUP. And also as previously mentioned for soybean, a decrease in the size of the training population tended to decrease the quality of the genomic prediction.

## 6. Association detection/validation



**- Workflow applied in the rest of agronomic, adaptive and quality traits**  
**- Significant SNPs (GWAS) → validated with QTL mapping studies**



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

58

We selected this workflow to be applied to the rest of the agronomic, adaptive and quality traits. All the significant SNPs identified with GWAS will be validated, if possible, with QTL mapping studies.

In summary, GWAS has become the driver of gene discovery in faba bean. Once the GWAS output passes the statistical criteria and the validation, the next step would be candidate gene identification, by defining



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

**EUCLEG.eu**

## Conclusions and future prospects

- **GWAS is a driver of gene discovery in faba bean**
- GWAS output → pass **statistical criteria** → validation:  
**Candidate gene identification** by defining physical interval



- ✓ **Reference genome assembly effort** (NORFAB consortium) → underway
- ✓ **Pan-genome** (University of Helsinki and Luke, Finland) → underway
- ✓ **Consensus map** 5x biparental populations → underway
- ✓ **Annotated reference faba bean transcriptome** (Escobar-Herrera et al 2020)



- **Aid genome assembly and advance faba bean genomics and breeding**



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

59

the physical interval. This is not yet possible as in faba bean as the physical map is not yet available. Nevertheless, a reference genome assembly effort is underway, as well as the development of a pan-genome and a consensus map using several bi-parental populations together with an annotated reference faba bean transcriptome. These are all excellent tools which will aid in the following genome assembly and will boost the advances in faba bean genomics and breeding.

## Acknowledgments



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

60

I would like to acknowledge the work performed by my colleagues at IFAPA. On the left are those involved in the multiplication and GWAS analysis. On the right are those in charge of the quality trait evaluation.



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

**EUCLEG.eu**



Especially thanks to Natalia Gutierrez who was mostly involved in the GWAS and the excellent support from Marie Pégard, INRAE, France as well as the two colleagues Ignacio Solis in Spain and Dejan Sokolovic in Serbia who performed the field evaluations

## About the author

**Dr Ana M<sup>a</sup> Torres** is Senior Research Scientist at IFAPA (Institute for Agricultural and Fisheries Research and Training) in Córdoba, Spain. She graduated in Biology from the university of Barcelona and completed her PhD at the University of Córdoba studying the genetics of the self-fertility in faba bean. Since 1991 she spent several training periods in international laboratories (Geneva, New York and Davis, California, USA) in order to learn basic techniques of plant molecular biology. With more than 30 years of experience in legume crops, her main research interests are: classical and marker-assisted breeding, development of genetic maps, Quantitative Trait Loci (QTL) analysis for disease resistance, yield and quality parameters, genetics, genomics and transcriptomics in crop legumes. Her group contribution has led to remarkable advances in the development of the faba bean genetic maps available so far. This research focuses specifically on improving agronomically important traits such as resistance to broomrape (*Orobanche crenata*), ascochyta blight (*Ascochyta fabae*) and rust (*Uromyces viciae-fabae*), as well as on yield components and nutritional aspects such as tannins and vicine-convicine content. As a result, several molecular markers for improving faba bean breeding programs are already available. Dr. Torres participates in several national and European projects on legumes, acting as coordinator of faba bean molecular tasks.

**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 1<sup>st</sup> October 2021**

Recording link to the presentation Genomics assisted breeding in faba bean :

<https://youtu.be/UmXQg2Y-Jps>



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

**EUCLEG.eu**

## 8. Introduction to outbreeding species: traditional breeding methodologies

David Lloyd



In this chapter I will cover methodologies traditionally used for breeding for outbreeding species. While the goal for inbreeding species tends to be to produce pure-line cultivars as discussed previously, this approach isn't suitable for outbreeding species. They tend to have a robust self-compatibility mechanisms, so they naturally out-cross. These self-incompatibility mechanisms can be genetic, morphologic or phenologic. Populations tend to have high level of heterozygosity and we want to maintain a level of heterozygosity because outbreeding species tend to be susceptible to inbreeding depression, carrying high loads of deleterious recessive alleles.

The process used for variety development tend to use population improvement, to produce open-pollinated, population cultivars. The selection process places some focus on the production genetically distinct, but phenotypically similar genotypes within the breeding population. A good analogy is that of animal breeds, where each individual animal is genetically distinct from other animals of that breed, but they are also visually similar.



## Outbreeding cultivars

- **In contrast to inbreeding species, outbreeding species rely heavily on open pollination**
  - Comprised of populations of genetically non-identical genotypes
    - Can be thought of as population cultivars
    - Individuals are, however, phenotypically similar
      - A valid analogy is the case of animal breeds
  - High levels of heterozygosity
    - Often susceptible to inbreeding depression
    - Often carry deleterious recessive alleles
  - Still have to adhere to DUS
    - More scope for DUS failures
    - Uniformity requirement sometimes "looser"
  - Forage legumes:
    - Clovers, Alfalfa
  - Some grain legumes have significant levels of open pollination and can be treated as either pure line or population cultivars
    - Faba bean is a good example of this
  - Number of approaches taken.



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

Activate Wind  
Go to Settings to e

Open pollinated cultivars still have to adhere to DUS requirements. There is more scope for populations to fail the uniformity requirements, so sometimes these criteria are looser than in inbreeding species. This can be a double edged sword, as broadening requirements for uniformity can also increase the potential for failing the criteria for distinctiveness.

Open pollination is common in forage legumes, such as clovers and alfalfa. Some grain legumes also have some degree of open pollination. For example, faba beans can be treated as either as pure line cultivars or open line cultivars. In practice combinations of the two approaches are used for faba bean. Pedigree methods are often used to produce inbred lines that are subsequently combined as pollinated synthetic varieties.



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

**EUCLEG.eu**

## Breeding objectives



- **Similar to those of inbreeding species**
  - Forage yield
  - Yield stability
    - Abiotic stress tolerance
    - Biotic stress tolerance
      - Grazing, disease, pest tolerance
  - Quality traits
    - Digestibility
    - Protein content
    - Isoflavone content (red clover)
  - Persistence
    - Unlike grain legumes, forage legumes are harvested repeatedly
  - Seed yield
    - Sometimes (mistakenly) an afterthought




Horizon 2020 of European Union  
This project has received funding from the European Union's Horizon 2020 Programme for Research & Innovation under grant agreement n°727312.
Activate Windows  
Go to Settings to activate Windows.

Breeding objectives for outbreeding species are similar to those for inbreeding species. Yield and potential yield are often the largest concern. With forage legumes, the main yield component that is considered is that of the above ground vegetative parts of the plant, so-called herbage yield, and as they are often perennial species consideration has to be made of the yield of multiple cuts taken over multiple years, rather than that from a single harvest.


Yield stability is important, so attention is paid to the populations' response to biotic and abiotic stresses. Abiotic stresses includes grazing tolerance and resistance to disease. The emphases put on disease resistance can be geographically determined. In the UK these include *Sclerotinia* crown rot, *Ditylenchus* nematodes, Anthracnose and various fungal and bacterial wilts.

Quality traits of importance in forage legumes include digestibility, fibre content and protein content. In red Clover, isoflavones (often called phytoestrogens) are of some concern. These can have these oestrogenic effects on livestock causing fertility problems.

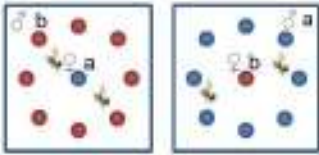
Persistency is a major goal of breeding with forage crops. This is yield stability in response to cutting multiple times over multiple seasons. Sometimes initially high-yielding varieties can have a tendency to suffer persistency issues later on in their life cycle.


There is a natural tendency to concentrate on herbage yield at the expense of seed yield. This can be problematic as it is possible to produce cultivars that are agronomically excellent but difficult to produce seed from. It doesn't matter how well a variety does in terms of its agronomic performance, if you can't produce sufficient seed it will not make it in the commercial sector.

## Founder populations




- **First step is usually one of creating a breeding population**
  - Population needs to be genetically diverse
    - Existing population cultivars can be used as the founder population, but care needed to avoid DUS issues
  - A better approach is to hybridise divergent material
    - If crossing two existing varieties, handcrossing can be used to ensure crosses are true hybrids, but very laborious to get sufficient numbers to avoid inbreeding.
    - Insect pollinators can be used... best approach is to use top crossing, where one mother plant of one population is crossed with many pollen donors of another population, and seed only taken from the mother plant
      - Use insect proof crossing chambers
    - Reciprocal top crosses can be combined in subsequent polycrosses to give maximum genetic diversity
      - Combine F<sub>1</sub> seed from both top crosses in a pollen proof chamber; allow insects to freely pollinate
      - More generations of polycross are better to provide HW equilibrium





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



The first step in breeding outbreeding populations is to create or identify sufficiently diverse breeding population, or founder population. This could be a pre-existing population, such as ecotype selections involving collections of wild material, or it can be refinement of existing varieties. Some old deleted varieties may have very promising qualities that can be harnessed, but may fail to meet other aspects of modern varieties, such as the uniformity requirement for DUS. An alternative and arguably a better approach is to take divergent material and hybridize it, similar to the approach used in inbreeding populations. Hand crosses can be used effectively, but flowers are generally a lot smaller than they are in the grain legumes and sufficient numbers need to be made to create a population that is not prone to inbreeding depression.

Insect pollinators can be used effectively, although they're not very discerning about what they pollinate. A top crossing approach can be very effective. A mother plant of population A can be surrounded by many pollen donors of population B in an insect proof isolation chamber. Pollinating insects are introduced and allowed to freely pollinate. Only seed from the mother plant is collected and all seed collected will theoretically be F<sub>1</sub>s crossed between the two populations with self incompatibility preventing selfing. Balanced bulks of multiple reciprocal top crosses can then be polycrossed one or more times to sort alleles into Hardy Weinberg equilibrium.

## Spaced plant selection

- **Spaced plants are still the best initial way to refine breeding population**
  - Need to be aware that spaced plants behave differently to closely packed plants in a ley
    - Very difficult to phenotypically characterise closely packed plants
  - A balance of selection intensity versus potential for inbreeding depression needs to be made
    - Around 30 genotypes is a good rule of thumb, but this is dependent on the genetic diversity of the founder population.
    - Work back from this point with desired selection intensity to give spaced plant nursery size
      - If 30 plants are to be selected at an intensity of 1%, need a starting point of 3,000 plants from founder population
  - Phenotypic selection carried out on desired agronomic traits etc. and on LIPOV traits
    - Aim to produce a population that fulfils requirements but will pass DUS.



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

Activate Wind  
Go to Settings to activate

Selection of spaced plants from nurseries is problematic to some degree as their growing conditions do not perfectly imitate growing conditions in a ley. It is however still the most efficient way to phenotype at a large scale and breeding is a numbers game. To achieve a selection intensity that is high enough to achieve sufficient genetic gain while avoiding inbreeding depression and achieving reasonable uniformity, the population that is being selected from needs to be large. If the aim is to select 30 plants, a realistic figure to maintain heterozygosity, at a 1% selection intensity, the starting point needs to be 3000 plants. It is inefficient phenotype numbers of this magnitude in closely packed plants. It's also worth mentioning that most DUS testing is done on spaced plants, so there is a valid reason to select for conditions where the plants are actually assessed.

The process is often started with plants growing in pots. Depending on how uniform the initial population is expected to be an initial selection is made from perhaps two or three times the number of plants to be transplanted to the spaced plant nursery in the field. Assessment is made in the glass house for various visual traits. Cotyledon size, leaf colour, markings, and the like. General vigour is considered and any outliers removed.

Plants are then transplanted to the field, typically in springtime for forage legumes, and in that establishment year various traits are scored including plant size, vigour and its tendency to flower without vernalisation. In the second year, following the first winter, more data is collected. This includes growth habit, whether the plants are erect or prostrate, flowering dates, leaf colour, leaf size and shade, plant sizes and vigour, height, stem thickness, etc. Subpopulations are made that combine sufficient uniformity with performance and ideally several subpopulations will be made that diverge, for example, for flowering time.



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

**EUCLEG.eu**

Selected plants are then polycrossed, to produce the next generation. This involves putting plants into separate glasshouse and allowing insects to freely pollinate all plants and seed harvested as a bulk.

## Mass selection



- **Simplest method available**
  - Similar principle to natural selection
  - Founder population is exposed to a challenging environment
    - Abiotic stress
      - Soil pH, low nutrient availability, temperature extremes etc.
    - Biotic stress
      - Disease and pest load, grazing intensity
  - Survivor plants collected at end of growing period
    - Excellent way of selecting plants to fit a specific environmental requirement
    - Difficult to control
      - If environmental challenge is too minor, insufficient selection intensity occurs
      - If challenge is too harsh, you have insufficient selections to make viable variety





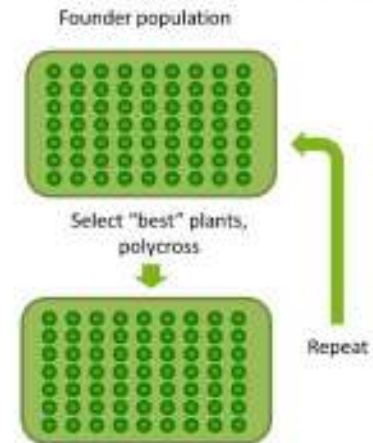
Horizon 2020 of European Union  
This project has received funding from the European Union's Horizon 2020 Programme for Research & Innovation under grant agreement n°727312.



Mass selection is one of the oldest forms of selection used within plant breeding. This essentially uses the same principles as natural selection. The breeding population is subjected to stresses that are theoretically controlled by the breeder. These can be abiotic, subjecting the population for, say, lower soil pH than is optimal, or for low nutrient availabilities. This approach has been used successfully for example for increasing phosphorous use efficiency in white Clover. Mass selection can be used to improve tolerance for biotic stress, such as disease and pest load, grazing intensity. We can impose grazing stresses to select for grazing tolerance. After a certain length of time, survivor plants are taken and polycrossed to produce an improved population. Mass selection can be very successful but the selective pressures can be difficult to control precisely. A balance between the challenge of the environment and selection intensity has to be made. If the selection pressure is too small there is insufficient selection intensity and if it is too harsh insufficient plants are available to produce a viable population.

## Recurrent selection

- **Refinement of phenotypic selection**
  - Spaced plants are selected as normal
    - Spaced plant nurseries
    - Mass selection
  - Selected plants are polycrossed in insect proof glasshouses
  - Progeny from polycross are planted back in spaced plant nursery
    - Same selection pressure imposed on a cyclic basis
- **"Two pronged attack" can be used**
  - Smaller set of selections can be taken for variety production
    - Maximises genetic gain
  - Larger set can be cycled into population improvement
    - Lower genetic gain
    - Lower risk of inbreeding depression in later generations



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

Activate Wind  
Go to Settings for more

Recurrent selection is a commonly used approach for a population improvement. This can use previously described, conventional selection methods, whether mass selection or phenotypic selection. Selections are made and the best plants from the breeding population polycrossed to form an improved population. This is repeated over and over to produce improvements in each subsequent generation. A balance needs to be made between genetic gain and maintaining genetic diversity within the populations. A harsh selection intensity gives the best genetic gain in a single generation but loss of genetic diversity will cause improvement to plateau in subsequent generations.

In practice it's preferable to select two populations. One with a harsher selection intensity, used for variety development and a larger selection of plants with less stringent selection intensity for population improvement. Typically the smaller population will be a subset of the larger population.



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

## Progeny selection

- **Excellent way of testing compatibility of parent plants for varieties**
  - Works very well for perennial species, some adaptations can be made for annual species
  - Polycross of selections made
    - Usually a large number
  - Seed harvested from each "mother plant" and kept separate
    - Called half-sib families: each seed has same mother, but pollen derived from all other plants in polycross
  - Mother plants kept alive
  - Half sib families trialled and best performing families identified
  - Elite mother plants then polycrossed in a smaller scale to produce potential varieties
  - Works extremely well in combination with recurrent population improvement methodologies



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

Activate Wind  
Go to Settings for a

Progeny selection uses test crossing to select genotypes to use for variety production. Crossing the genotypes that appear to be the best phenotypically is often not the most effective approach. Visual appearance does not guarantee how well a group of plants will combine in a cross. Half sib families are thus used to assess the breeding value of genotypes. The breeding population is polycrossed and mother plants retained. Seed from the mother plant is harvested separately from other plants and retained. The seed harvested from each plant is a distinct half sib family.

The half sib families are then sown in trial plots and the best performing half sib families are identified. The assumption is that the best performing half sib families are from mother plants that have the best of the best combining ability with other genotypes. The elite mother plants are then polycrossed to produce the new variety. This is a very effective approach to producing outcrossing population cultivars.

### About the Author

Dr David Lloyd is head of Forage Breeding for Germinal Holdings. Before taking on his role at Germinal, he was a Senior Legume Breeder at Aberystwyth University, specialising on clovers, peas and faba beans.

**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 1<sup>st</sup> October 2021**

Recording link to the presentation: <https://youtu.be/fe-KF6R7gRM>



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

**EUCLEG.eu**



## 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

**HORIZON 2020**

Horizon 2020 of European Union; Call 2016, SFS 44 - "A joint plant breeding programme to decrease the EU's 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.

**EUC LEG**

**Genomics assisted breeding in alfalfa**

**INRAE**

Bernadette Julier  
Marie Pégard, Julien Leuenberger, Philippe Barre  
[www.eucleg.eu](http://www.eucleg.eu)



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

**EUCLEG.eu**



## 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



Horizon 2020 of European Union: Call 2016, SFS 44: "A joint plant breeding programme to decrease the EU's 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.

2

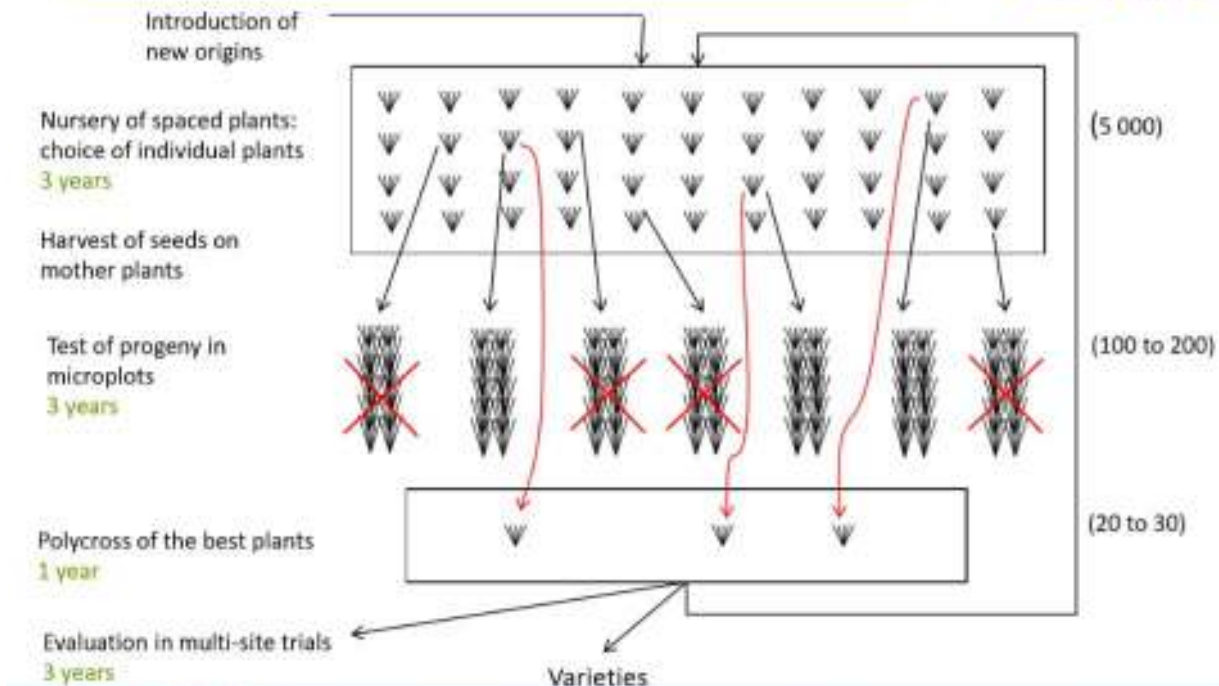
Alfalfa, or lucerne, is a major legume species that gives the highest protein production per acreage among all legume species in temperate regions. It is quite drought tolerant; it has a convenient protein/energy ratio. It provides some advantages to ruminant health, and it has positive effects in the 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.

**EUCLEG.eu**

## Traditional breeding methodology



Horizon 2020 of European Union: Call 2016, SFS 44: "A joint plant breeding programme to decrease the EU's 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.

4

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.

## Traditional breeding methodology

### Strength

- Scoring of many traits
- Early selection for heritable traits
- Skilled staffs

### Weakness

- Some traits are scored in case of stress occurrence only
- Number of years
- Cost
- Fixation of positive alleles is slow



Horizon 2020 of European Union: Call 2016-SFS-44: "A joint plant breeding programme to decrease the EU's 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.

6

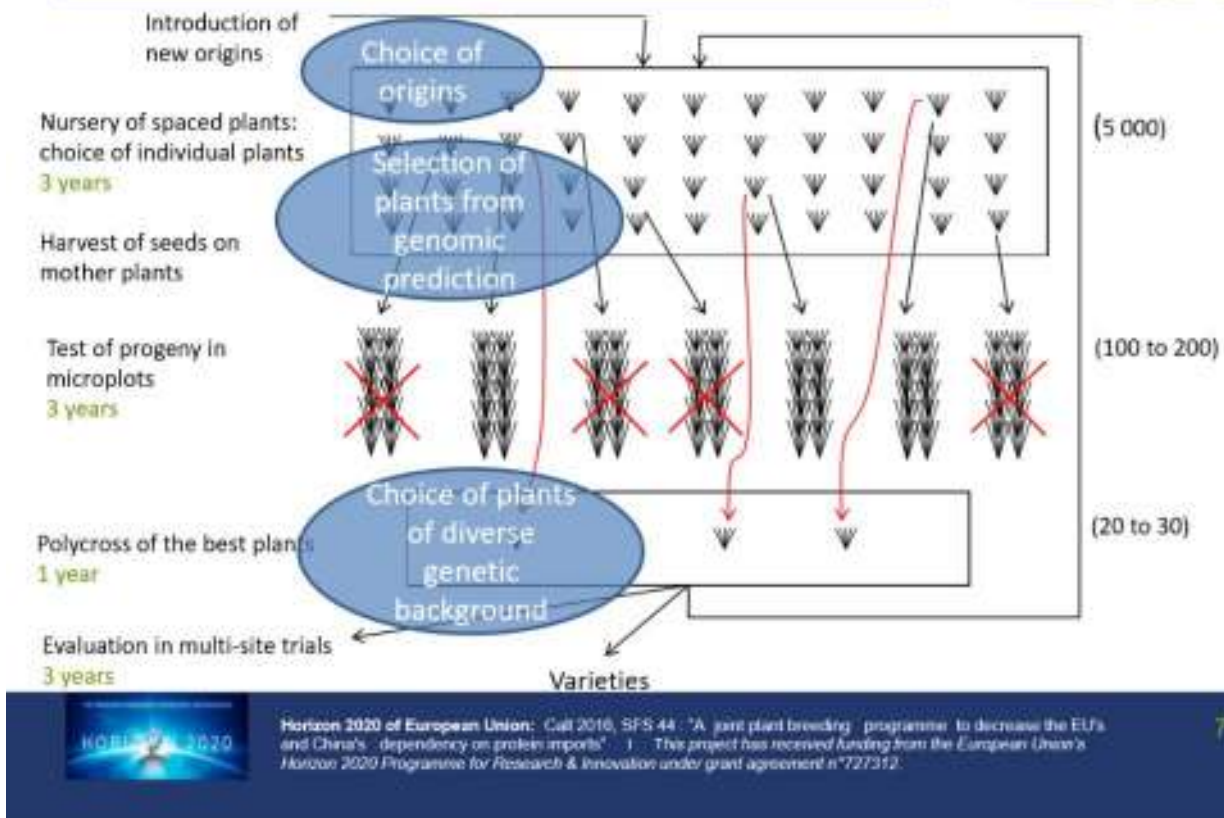
What are the strengths and the weaknesses of this methodology? The strengths include that you can score many of the traits, you can have early selection for heritable traits in the nursery, and for most breeding companies the staff are already skilled to be able to do this work. The weaknesses include 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.

**EUCLEG.eu**

## Genomics assisted breeding



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.

## Marker development

### Before EUCLEG:

- Low throughput markers: SSR, AFLP...
- 10k SNP array: too expensive
- GBS: < 40K markers, risk of missing data



Horizon 2020 of European Union: Call 2018, SFS 44, "A joint plant breeding programme to decrease the EU's 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.

9

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.

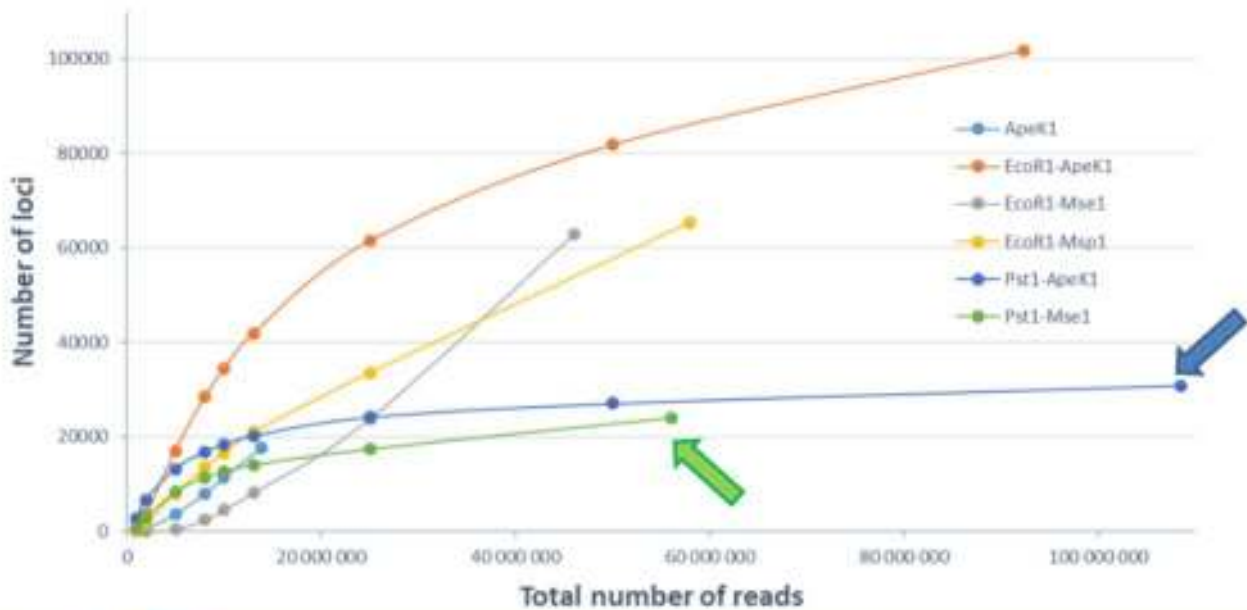
In EUCLEG, we have developed an improved GBS protocol, by testing different restriction enzymes to reduce missing data and thus optimize the protocol of GBS.



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

**EUCLEG.eu**

## 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.

**EUCLEG.eu**

## 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



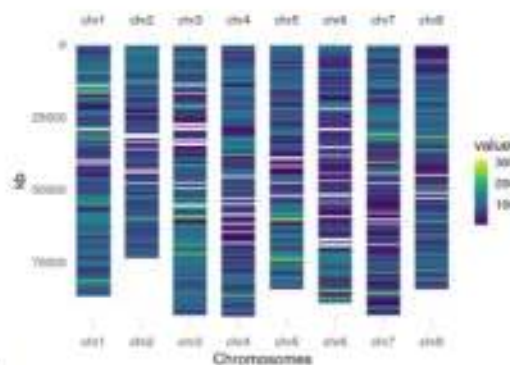
After we have chosen these restriction enzymes, we have used a reference genome sequence delivered by Chen et al. in China to map the GBS reads and we have obtained allele frequency of each accession.

## 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.

**EUCLEG.eu**

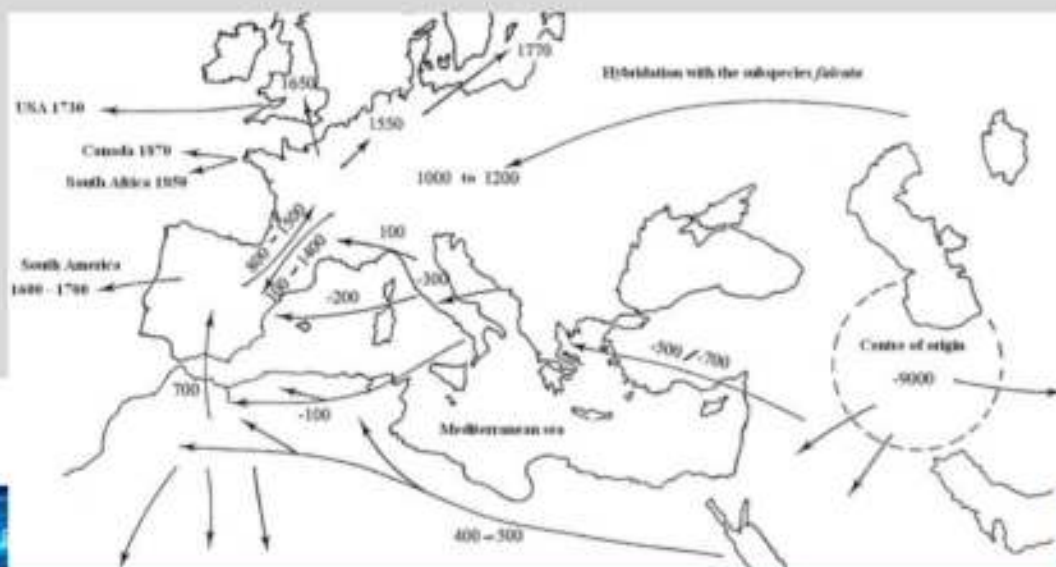
per SNP: a huge number of SNP with a low ratio of missing data. If we chose to add a new threshold without any missing data, we have more than 100 000 SNPs. As shown on the right, these SNPs cover the genome very well. This genotyping tool is now available and this is great progress for alfalfa genetics.

## Management of genetic diversity



### Before EUCLEG:

- Overview of world diversity



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.





## Management of genetic diversity



### Before EUCLEG:

- Overview of world diversity
- Large among-accession diversity
- Huge within-accession diversity

	10 populations, 40 indiv/pop		11 populations, 7-20 indiv/pop
	5 SSR	Plant height	Yield
Variance among-varieties	0.02	0.10	1.7
Variance within-varieties	7.56	0.30	27.7
	No structure		
	Herrmann et al., 2010		Julier et al. 2000



Horizon 2020 of European Union - Call 2015, SFS 44 - "A just plant breeding programme to increase the EU's 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

15

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.

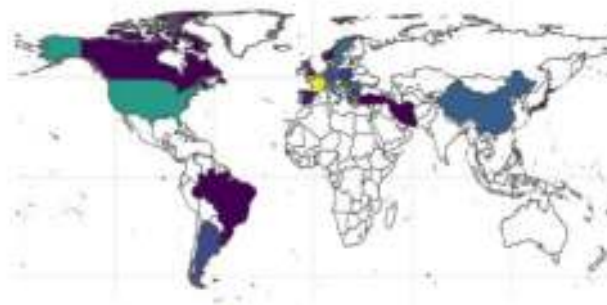
## Management of genetic diversity



### EUCLEG: a revision of diversity overview

400 accessions: landraces and cultivars, dormancy 3 – 7:

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



Horizon 2020 of European Union - Call 2015, SFS 44 - "A just plant breeding programme to increase the EU's 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

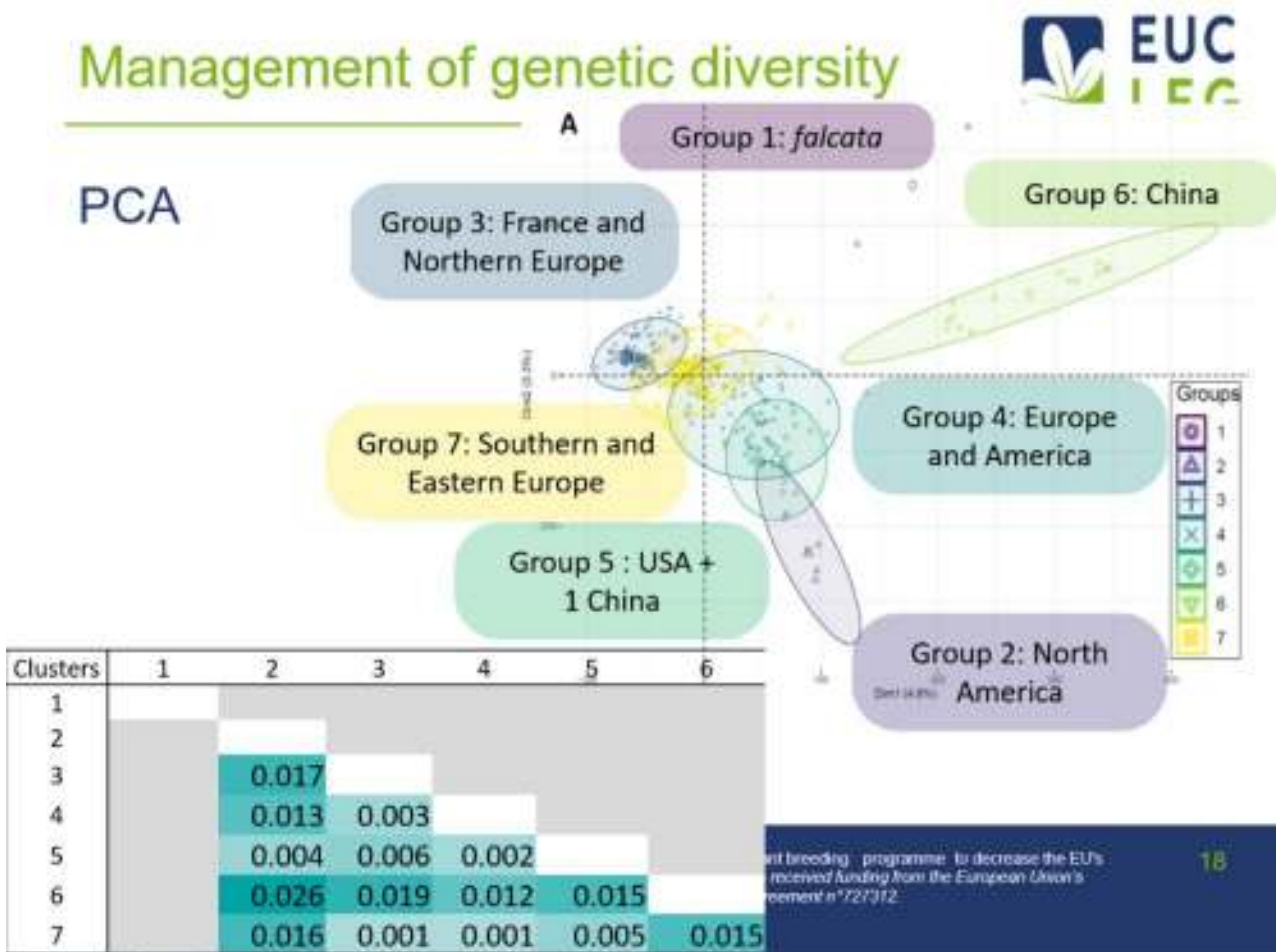
16



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

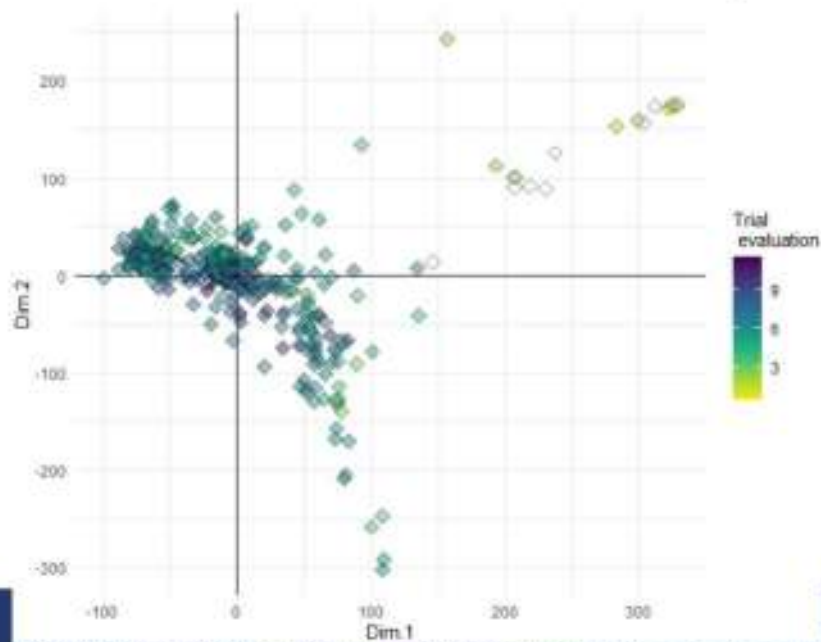
**EUCLEG.eu**

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.



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.

PCA : accessions colored with autumn dormancy score



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.

## Management of genetic diversity



### EUCLEG: a revision of diversity overview

- Diversity: China < > Europe + America
- Diversity: Europe < > America
- Structure is not associated to autumn dormancy

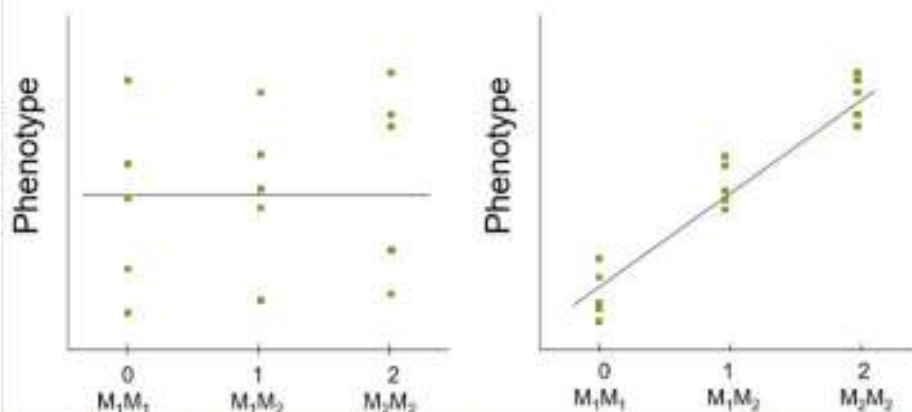


With this project, we have revised the overview of genetic diversity. We have a diversity in China which is quite different to the diversity which is present in Europe and America. Also the diversity in Europe and America is different even if most American accessions originate from Europe. The structure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

## Genome wide association study



For each marker: is it associated to trait variation?



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.

**EUCLEG.eu**



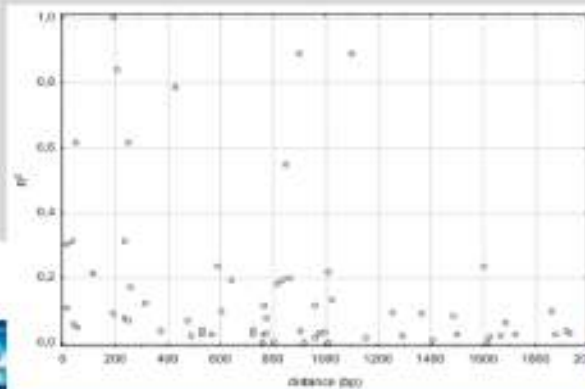
# EUC LEG

## Genome wide association study



Before EUCLEG:

- Low marker density
- Short linkage disequilibrium



12 SNP (66 pairs) in Constans-like gene (Herrmann et al., 2010)



To decrease the LDs: European Union's

32

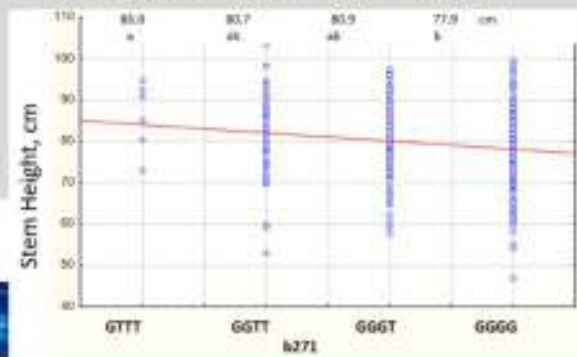
Before EUCLEG, we had low marker density, but we had shown that linkage disequilibrium was very short in this species, here studied at the level of a single gene, and linkage disequilibrium over 1000 base pair is broken with some exceptions in this gene.

## Genome wide association study



Before EUCLEG:

- Low marker density
  - Short linkage disequilibrium
- Candidate gene approach only



Constans-like gene (Herrmann et al., 2010)



To decrease the LDs: European Union's

33

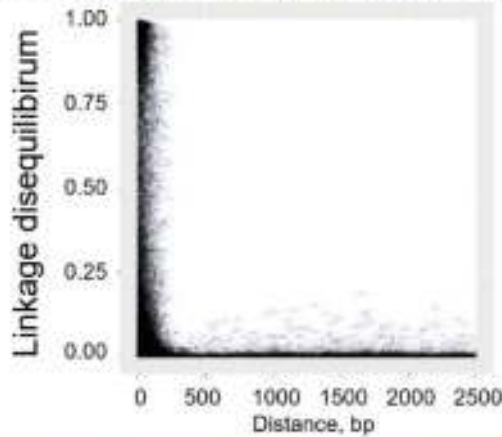
With this short linkage disequilibrium and low throughput genotyping, only a candidate gene approach could be acceptable for an association study. It can work, as shown here in a Constans-like gene and the association of this gene to a phenotypic trait, stem height.



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

**EUCLEG.eu**

EUCLEG: Short linkage disequilibrium over the genome



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*.

## 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

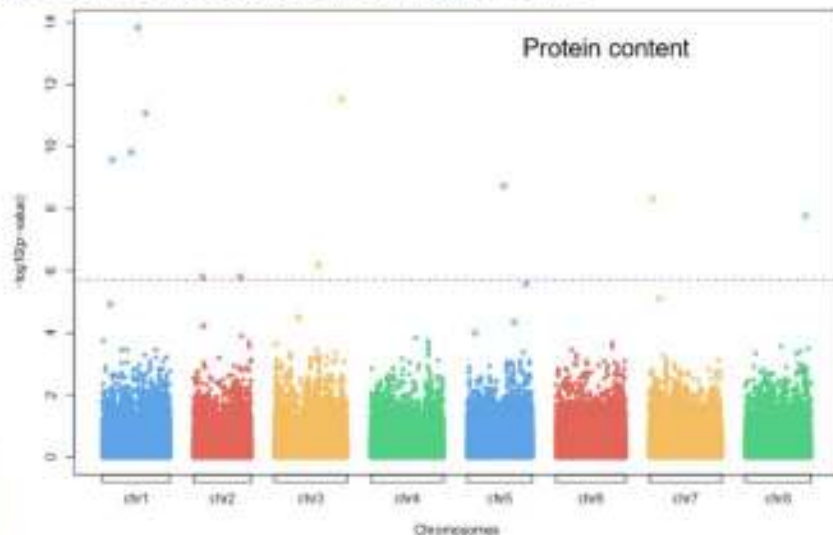


In EUCLEG, we were able to detect major QTLs in these accessions using a multi locus mixed model (MLMM).

## 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.



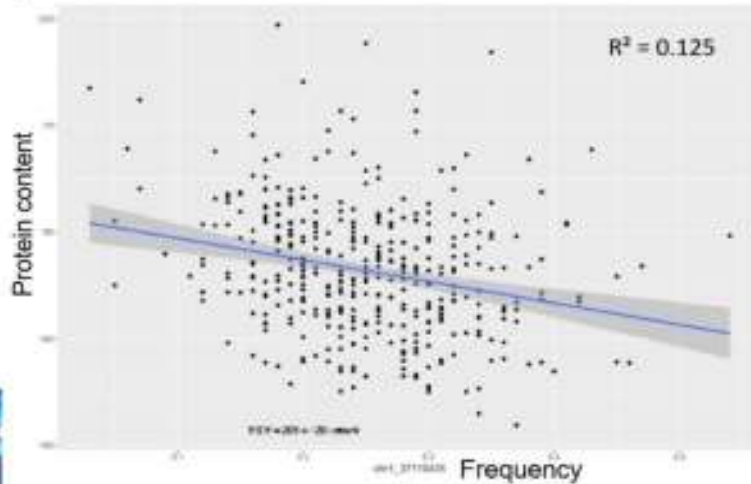
This project has received funding from the European Union's Horizon 2020 Programme for Research & Innovation under grant agreement n°727312.

**EUCLEG.eu**

## Genome wide association study

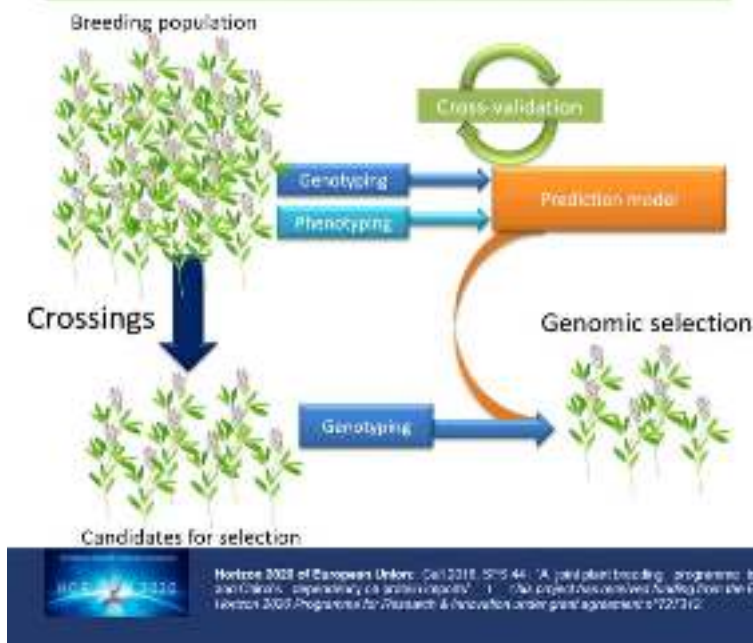
### EUCLEG: Detection of major QTL

- Up to 10 – 20% of variation



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.

## Genomic selection



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

**EUCLEG.eu**



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.


**Genomic selection** 

**Before EUCLEG**

- 8 – 44 K SNP, 75 – 244 individuals
- Promising results, predictive ability ~ 30%

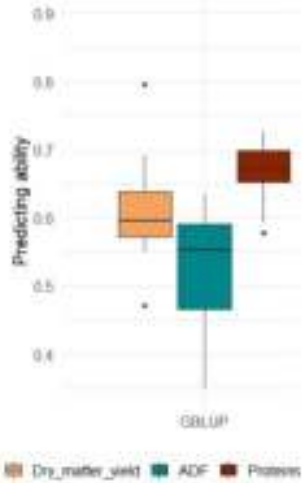
Horizon 2020 of European Union: Call 2015, 973 44 "A joint plant breeding programme to decrease the EU's and China's dependency on protein imports" | This project has received funding from the European Union's Horizon 2020 Programme for Research & Innovation under grant agreement n°727312

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%.

**Genomic selection** 

**EUCLEG**

- GBLUP
- A good predicting ability:  
 $0.52 < P < 0.66$

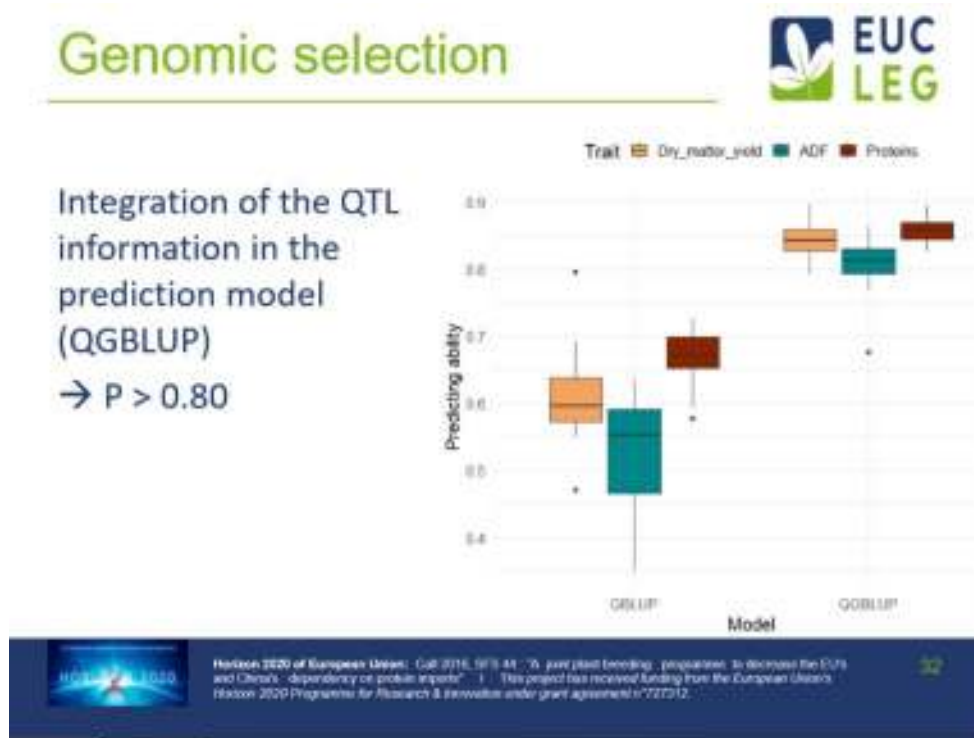


Horizon 2020 of European Union: Call 2015, 973 44 "A joint plant breeding programme to decrease the EU's and China's dependency on protein imports" | This project has received funding from the European Union's Horizon 2020 Programme for Research & Innovation under grant agreement n°727312

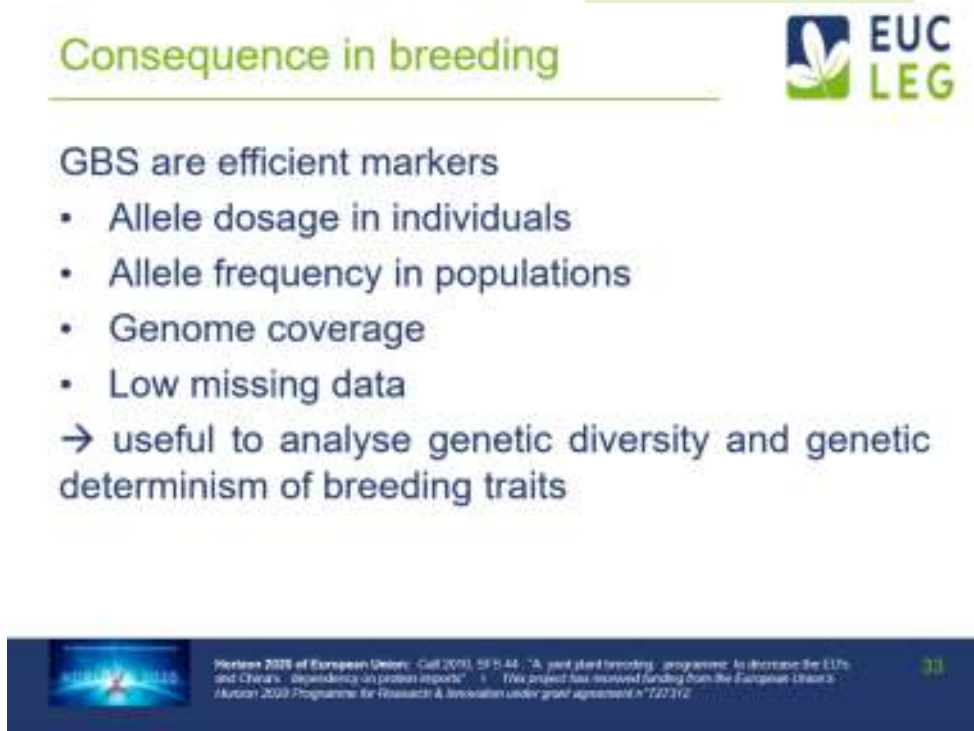
In EUCLEG, based on our 400 accessions and more than 200 000 markers, we carried out GBLUP prediction and we obtained quite a good predicting ability, between 0.5 and 0.8 for the p-value, shown here for dry



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.



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.

## Consequence in breeding



### Management of genetic diversity

- Some specialisation of the breeding pools in EU, America, China
- GBS markers to decide on the introduction of new genetic diversity in a breeding pool



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.

## Consequence in breeding



### GS models provide high predictive ability

- Even higher with the inclusion of QTL effect
- To be used to select promising individuals in breeding pools



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.



## 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



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.





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.



## GS in breeding programmes



### Strength

- Reduced field work
- Early selection for all predicted traits
- Reduced number of years
- Fixation of positive alleles is quick

Genetic  
gain?

### Weakness

- No prediction for some traits
- Staffs have to get new skills

Cost  
efficiency ?



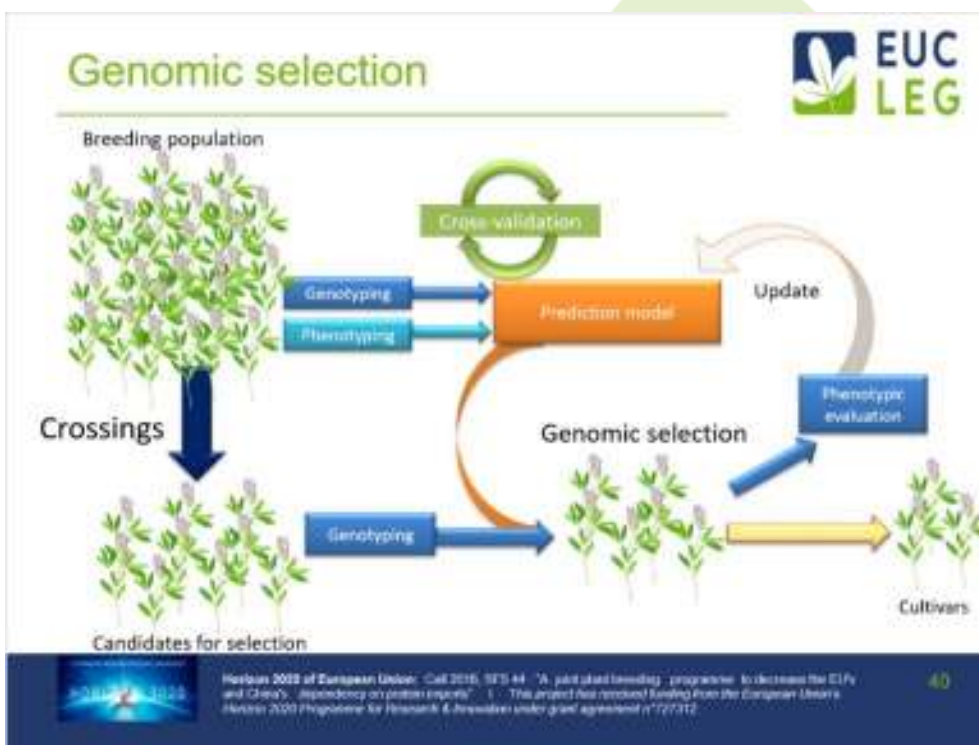
What are the strengths of this breeding programme? Reduced field work, a very early selection for all predicted traits, a reduced number of years of the breeding cycle, and a very quick fixation of positive alleles, especially important for the autopolyploid species. There are some weaknesses, firstly 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?



## 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



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 1<sup>st</sup> October 2021**

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





## 10. Genomics assisted breeding in red clover

Roland Kölliker

Senior Scientist at ETH Zurich, Molecular Plant Breeding, Institute of Agricultural Sciences, Zurich, Switzerland

**EUCLEG red clover species leader**



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

**ETH zürich**  
Agroscope

**EUC  
LEG**

**Genomics assisted breeding in red clover**  
Roland Kölliker, ETH Zurich

### A joint effort



- AFFRC, JP (Hiroko Sato)
- AgResearch, NZ (Michelle Williamson)
- Agroscope, CH (**Christoph Grieder**, Franz Schubiger, Michelle Nay, Philipp Streckelsen)
- DLF, CZ (**Libor Jalůvka**)
- ETH Zurich, CH (Lea Frey, Bruno Studer, Ingrid Stoffel, Verena Knorst, Lukas Kronenberg)
- Graminor, NO (**Helga Amdahl**)
- IBERS, UK (**Leif Skøt**, David Lloyd)
- IFVCNS, RS (Sanja Vasiljević)
- IKBKS, RS (**Jasmina Radović**)
- ILVO, BE (Tim Vleugels, Tom Ruttink, Hilde Muylle, Aamir Saleem, Reena Dubey)
- INRAE, FR (Julia Butink, Philippe Barre, Marie Pégard, Bernadette Julier)
- Lantmännen, SE (Linda Ohlund)
- NMBU, NO (Åshild Ergon, Stefano Zanotto)
- NordGen, SE (Anna Palme)
- NPZ, DE (Wilbert Luesink)
- RAGT2n, FR (Marie-Christine Gras)
- USDA, USA (Heathcliffe Riday)
- VUPT, CZ (Tomáš Vymyslický)



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

2



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

**EUCLEG.eu**

## Red clover – *Trifolium pratense* L.



Important forage legume grown as roughage for ruminants in pure stands or in mixture with forage grasses



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

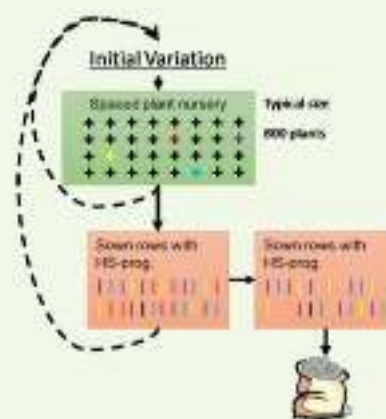


4

Red clover is an important forage legume, grown as a roughage for ruminants in pure stands or in mixture with forage grasses. It is of particular importance in areas unfavourable for arable crops. Another big advantage of red clover is its ability to fix atmospheric nitrogen and to transfer this nitrogen to companion species, which then can lead to transgressive overyielding. Red clover is also a very important component in crop rotations to improve soil fertility and is especially valued in organic agriculture.

## Breeding red clover

- Breeding aims
  - Yield and quality
  - Persistence, disease resistance
  - Emerging traits...
- Breeding strategies
  - Population improvement
  - Open-pollination of elite plants
  - Pair-wise and poly-crosses



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



5

Yield, quality, persistence, disease resistance and seed yield constitute the most important breeding targets in red clover.



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

## Breeding red clover - challenges



- Population based cultivars
  - Complex pedigrees, many parents
  - Fixation of traits difficult
- Pluri-annual crop
  - Improvement slow for many traits
- Changing requirements
  - Emerging pathogens
  - Environmental conditions



Horizon 2020 of European Union

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



Red clover breeding is based on population breeding and the resulting cultivars are based on many different genotypes. Consequently, we have very complex pedigrees with many parents, illustrated on the right here on the diagram. This makes the fixation of traits very difficult. Also, red clover is a pluri-annual crop, and for some traits such as for example persistence improvement is quite slow, because it takes many seasons of phenotyping. In addition, changing requirements, such as emerging pathogen populations or emerging pathogen species, or a general change in environmental conditions, pose a big challenge to red clover breeding.

## Breeding red clover - opportunities



- Large pool of genetic resources
    - Wild populations, ecotypes, landraces and cultivars
    - Efficient breeding schemes and knowledge in phenotyping
  - Increasing availability of genomic resources
    - Reference genomes, low-cost sequencing technologies, statistical concepts
- Genomics-assisted breeding



Horizon 2020 of European Union

This project has received funding from the 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**

On the other hand, we also have a large pool of genetic resources and have the advantage of wild populations and ecotypes growing simultaneously as the cultivated crops. So, there's quite a big gene pool that can actually be drawn upon when breeding red clover. We have quite efficient breeding schemes and a lot of knowledge in genotyping. With the advances in genome sequencing technologies, there's an increasing availability of genomic resources and also the appropriate statistical concepts are constantly being developed.

## Aims



- Establish a diverse collection of red clover germplasm
- Create genotypic and phenotypic information
- Elucidate the genetic control of key traits
- Develop concepts / models for genomics-assisted breeding



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



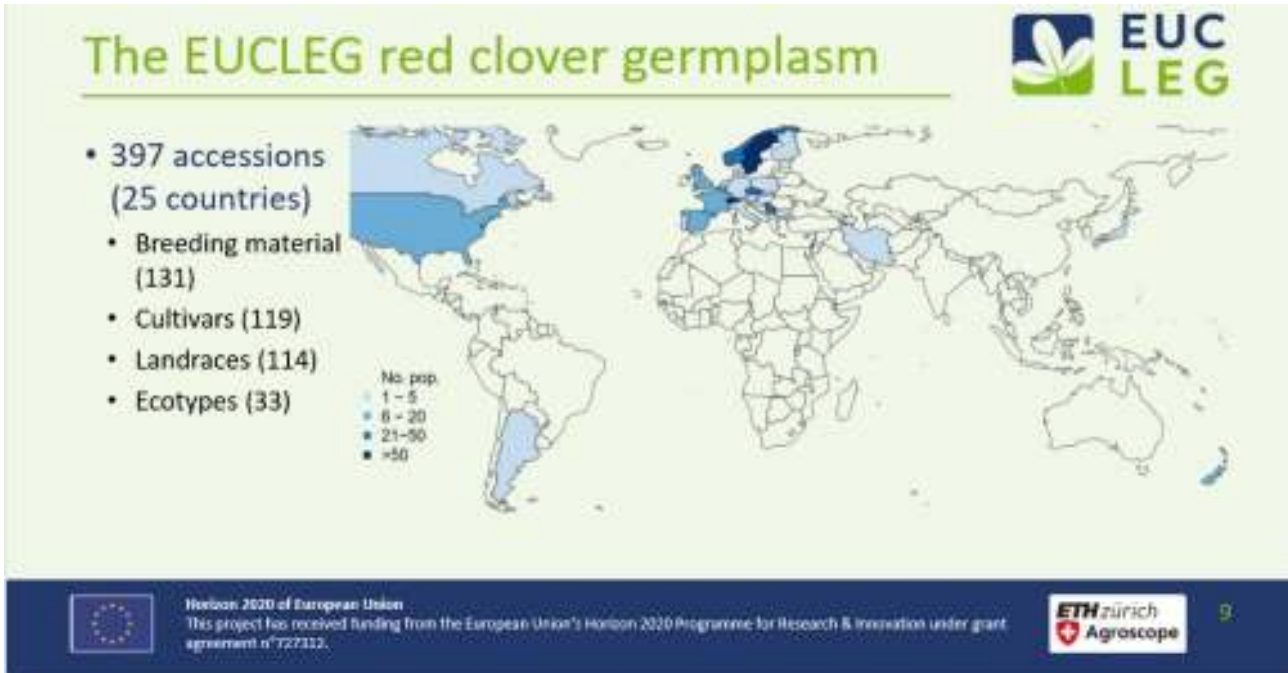
8

The aim of the EUCLEG project for red clover in particular was the establishment of a diverse collection of red clover germplasm and the generation of genotypic and phenotypic information, with the aim to develop concepts and models for genomics assisted breeding, but also with the aim to elucidate the genetic control of specific traits, to enable efficient breeding methods.

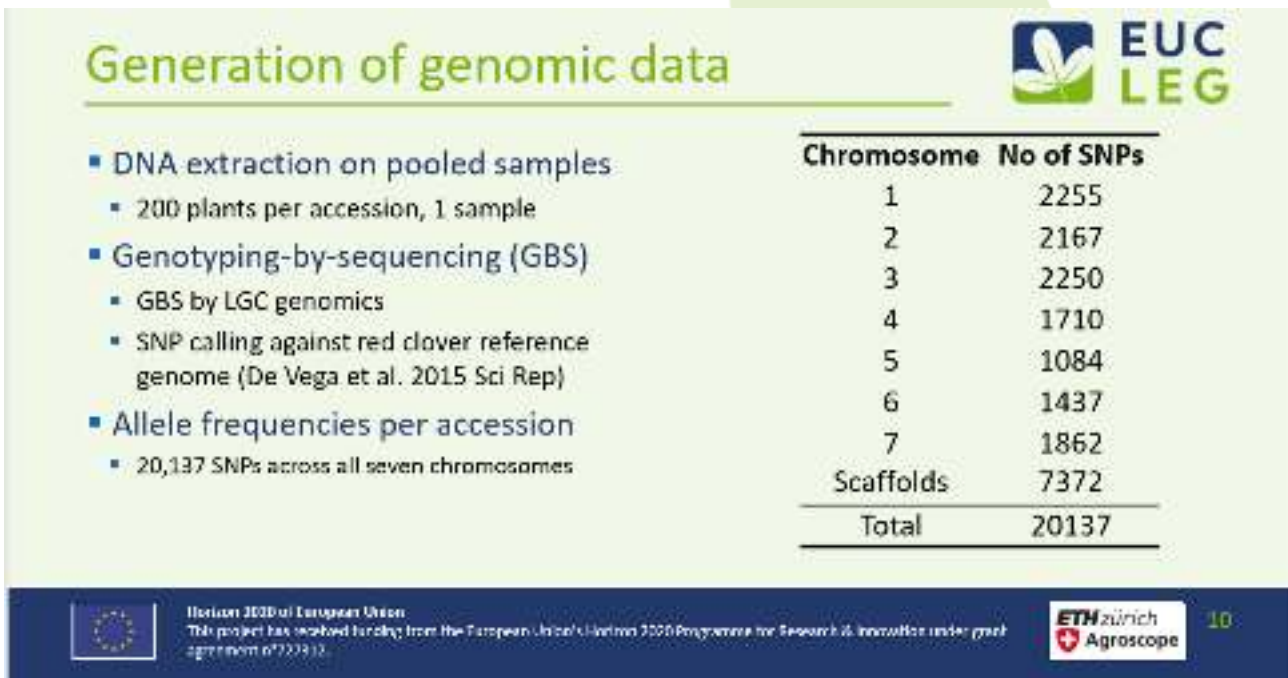


This project has received funding from the European Union's Horizon 2020 Programme for Research & Innovation under grant agreement n°727312.

**EUCLEG.eu**

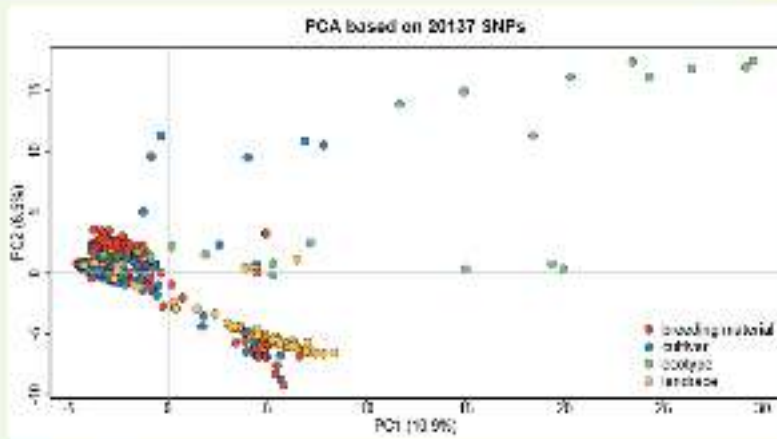


The EUCLEG red clover germplasm finally consisted of 397 accessions that were derived from 25 different countries. Switzerland and Sweden contributed the most accessions, around 100, Serbia, Norway and the UK contributed around 25. The Czech Republic contributed 50 and the rest were more or less evenly distributed among the countries coloured in blue on this map. The germ plasm could be divided into different categories: breeding materials, cultivars, landraces and ecotypes.



We used 200 plants per accession to genotype by sequencing and came up with a set of around 20,000 reliable SNPs. These were evenly distributed among the seven chromosomes of red clover.

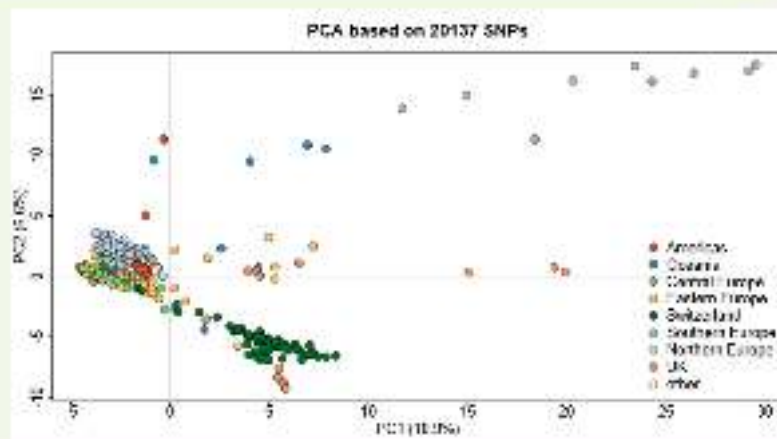
## Genetic diversity among accessions



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

We used principal component analysis based on these 20,000 single nucleotide polymorphisms to get an idea about the genetic diversity present in our germplasm collection. What you see here is a biplot of the first two principal components, and each dot represents one of the 397 accessions. We have some sort of structuring according to different breeding materials, so that the landraces seem to be quite a prominent group and also the ecotypes seem to form a distinct group on the top right of this graph.

## Distinction according to origin



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

Most of the land races we have in this set were derived from Switzerland and they all nicely cluster in one of the corners of the principal component analysis and also the accessions from northern Europe form quite a distinct cluster. We also have some outliers that form a distinct group of ecotypes coming from



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

southern Europe.

## Genetic characterisation



- 397 accessions genotyped with > 20,000 SNPs
- Clear genetic structure among accessions observed
- Genetic differences between accessions partially correspond to their country/region of origin
- Valuable resource for further analyses



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



17

We have a nice set of accessions genotyped with a large number of SNPs and we saw some clear genetic structure among these accessions, mainly based on the regions where they were collected.

## Phenotypic characterisation



Field trials at five locations: **Switzerland** (Agroscope); **Czech Republic** (DLF), **Wales** (IBERS), **Serbia** (IKBKS), **Norway** (Graminor)



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



18

We set out to do a large scale phenotypic characterization of these accessions and established field trials at five different locations: in Switzerland, in the Czech Republic, in Wales, in Serbia and in Norway.



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

**EUCLEG.eu**

## Phenotyping in field trials

- 100 – 395 accessions per site
  - P-rep design
- Phenotyping during two growing seasons
  - Establishment / stand density / persistence
  - Time of flowering
  - Forage yield and quality (3-5 cuts per season)
  - Disease occurrence
  - Dynamics of regrowth



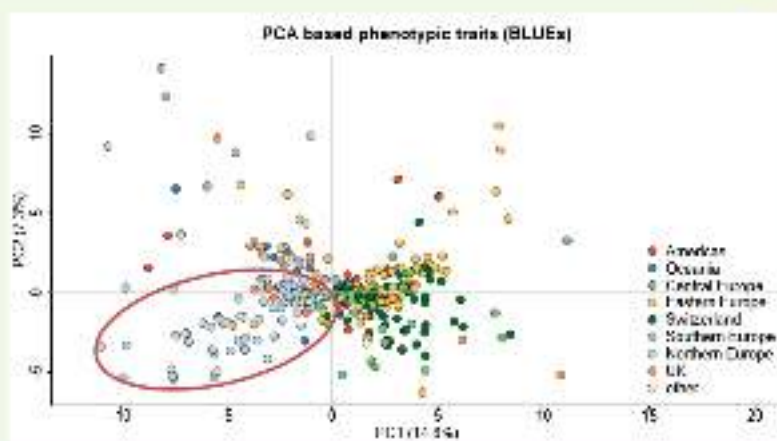
Horizon 2020 of European Union  
This project has received funding from the European Union's Horizon 2020 Programme for Research & Innovation under grant agreement n°727312.



19

The phenotyping was conducted on the entire set of the 395 accessions in Switzerland and in the Czech Republic. 200 accessions were phenotyped in a further three locations as well as 20 accessions which were phenotyped in all 5 locations. All this was arranged in a P-rep design and carried out across two growing seasons. We looked at establishment, stand density, persistence, time of flowering and of course forage yield and quality, disease occurrence was scored in the field along with dynamics of regrowth analysis.

## Phenotypic variability of accessions



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



21

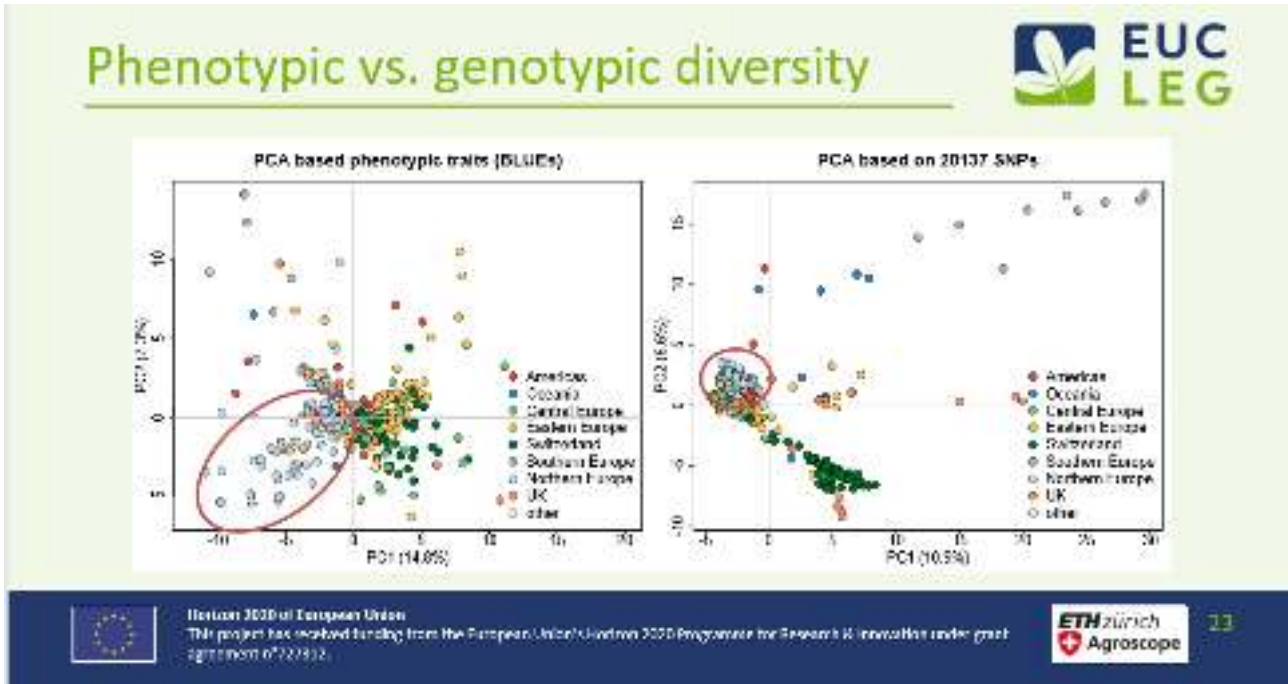
Principle component analysis based on BLUEs calculated from the field trial data from all the five sites shows not very distinct clustering, but you can identify quite a few groups. Again, these occur according to



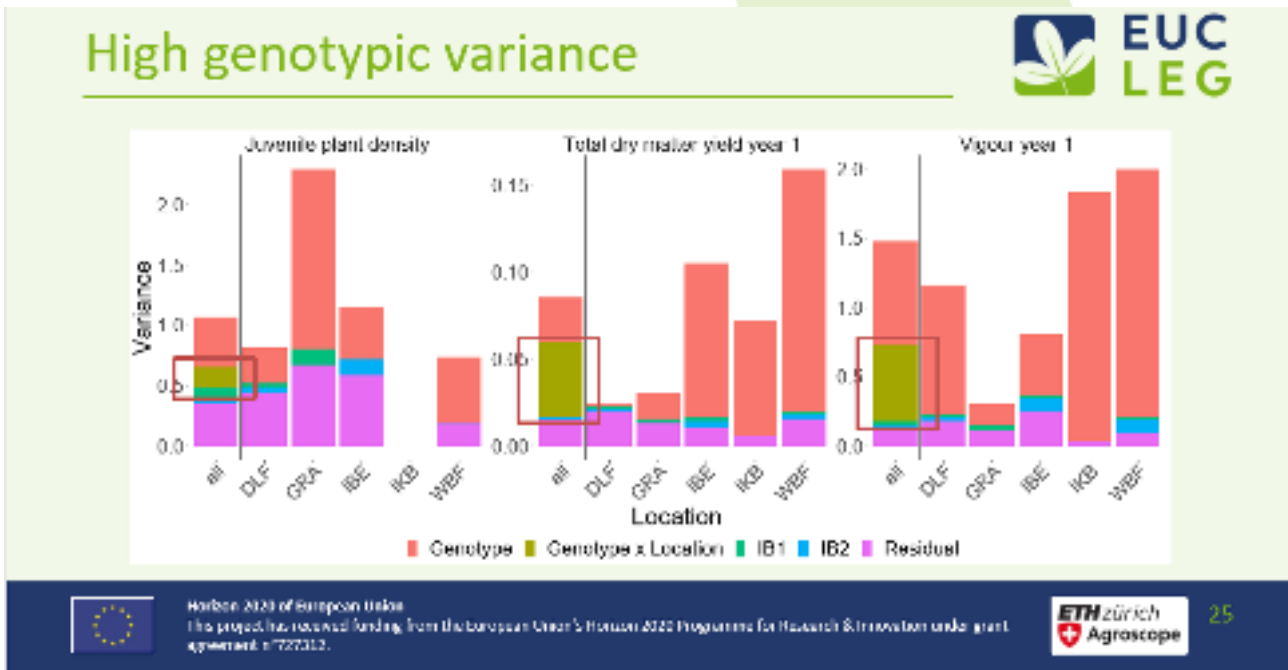
This project has received funding from the European Union's Horizon 2020 Programme for Research & Innovation under grant agreement n°727312.



the origin of the accessions and probably the most remarkable grouping is the one of the accessions from the northern European countries. You can also see some structure from Eastern Europe, Switzerland and central European countries.



If you compare this to the genetic data on the right, you can see a little bit of congruence but maybe not completely. You can clearly see that the genetic structure detected with SNPs is also to some extent reflected in the phenotypic diversity observed, based on traits evaluated in the field.

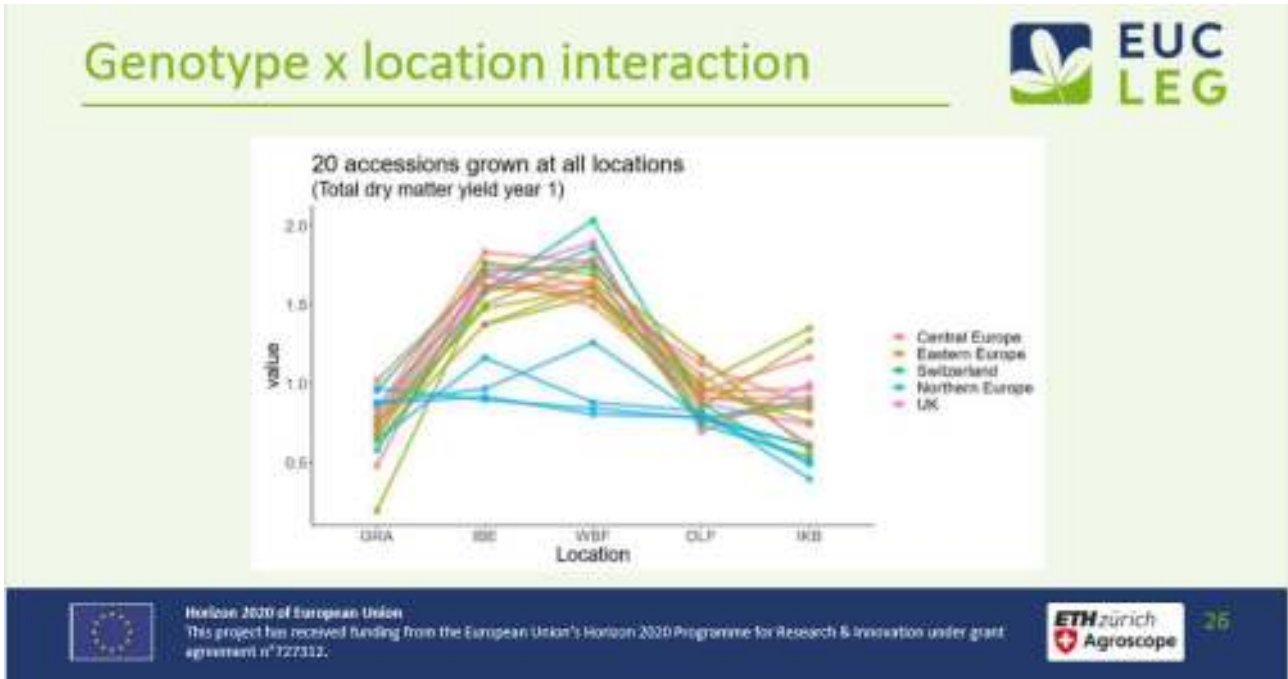


On the X axis are the five different sites. You can see for all three traits, we have quite a strong effect of the genotype, and therefore this is certainly a very valuable basis for any further analysis. What was also



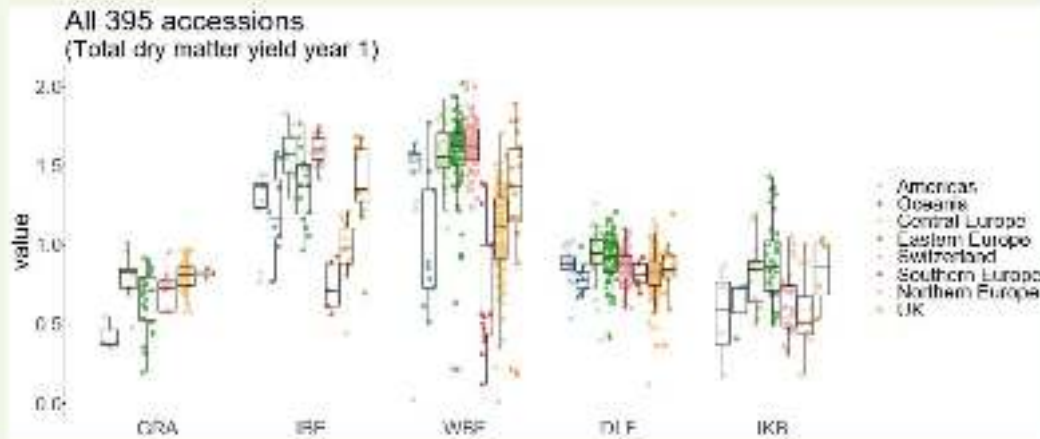
This project has received funding from the European Union's Horizon 2020 Programme for Research & Innovation under grant agreement n°727312.

quite clear from this analysis is that you have quite a large effect of genotype by location interaction if you look at the combined analysis across all five sites.



First, if we only look at the 20 accessions which were grown at all five locations (each line here is one accession) you can compare the phenotypic data for the five different sites. On the X axis on the left is Graminor in Norway, in the middle is Switzerland (WBF), and IKB in Serbia, is on the far right. If for example we look at the Nordic accessions, you can see that these accessions do fairly well in Norway, at the Graminor site, but they performed quite poorly at all the other locations, which can of course be explained by the adaptation to northern climates. Conversely, if you look at the accessions from Switzerland, you see that they performed poorly under the Nordic conditions.

## Best performance “at home”



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



29

If we look now at all the 395 accessions which were evaluated in the five locations, most accessions seem to perform best at home. This clearly indicates the importance of course of breeding red clover under conditions where it's actually intended to be used later on, and this is something which also has to be accounted for in any prediction models for genomic selection.

## Phenotypic characterisation

- Significant effect of accession at all locations
- Significant accession x location interaction
- Breeding for specific locations / environments
- Valuable dataset for further analyses



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



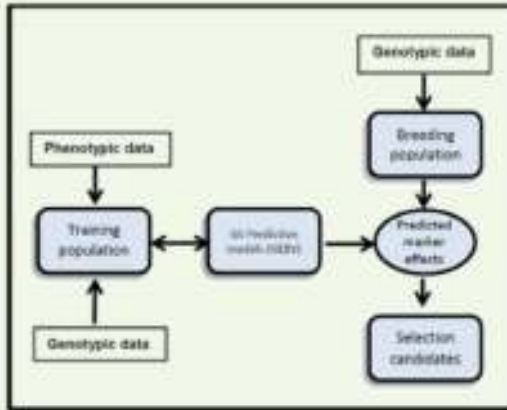
32

So we observed a significant effect of accession at all the locations and we had significant accession by location interactions.



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

## Genomic prediction



- Prior to EUCLEG not many studies on GS or GWAS in red clover
- Platform for predictive breeding in red clover

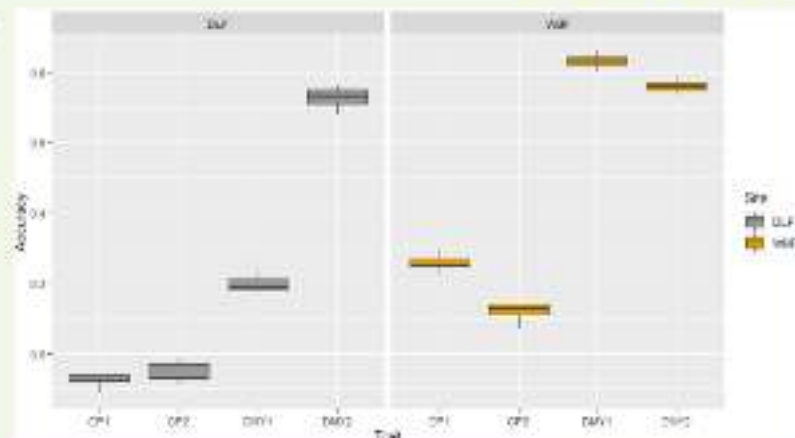


Horizon 2020 of European Union  
This project has received funding from the European Union's Horizon 2020 Programme for Research & Innovation under grant agreement n°727312.

Prior to EUCLEG, where there had not been many studies on GWAS or genomic selection in red clover, but this dataset would allow for the development of a platform for predictive breeding in red clover.

## Genomic prediction

- Substantial accuracy of prediction for some traits
- Crude protein content (CP) and dry matter yield (DMY) show high effect of location and year on prediction ability



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

First attempts in looking at the ability to predict resulted in quite substantial prediction accuracy for some traits. You can see this on this graph, showing the prediction accuracy for crude protein (CP) on the left in each panel, and dry matter yield in year one and year two. We have quite distinct differences in predictive ability, depending also on the location where the field trials were established.



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

## Genomic prediction



- First predictions for biomass yield and protein content and a range of other traits
- Models / training populations need to be optimised
- Incorporation of significant QTL in the analysis from GWAS analyses



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



37

This needs to be analysed in more detail, but of course we now have first predictions and this can basically be used for further development of prediction models.

## Genetic control of specific traits



- Experiments under controlled conditions
- GWAS / QTL analyses to identify candidate genes and to improve prediction models
- Traits:
  - Disease resistance
  - Persistence / winter survival
  - Seedling emergence



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



38

In addition, we conducted a range of experiments under controlled conditions, in order to detect QTL for some key traits, which can be used to identify candidate genes and to better understand the control of these traits but also can then later be used to improve prediction models.



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

**EUCLEG.eu**

## Resistance to southern anthracnose



Caused by *Colletotrichum trifolii*, increasingly problematic due to rising temperatures



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



39

Southern anthracnose, caused by *Colletotrichum trifolii*, is increasingly problematic due to rising temperatures. It can cause very severe yield losses, especially in warmer climates. Currently it's not a problem in Nordic countries, but in Switzerland for example it has become a major threat to the red clover production.

## Artificial inoculation experiment



- 397 accessions
  - 24 plants per accession
  - 4 biological replicates
- Artificial inoculation
  - Single spore isolate of *C. trifolii*
  - Survival rate (%)
- Association analysis



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

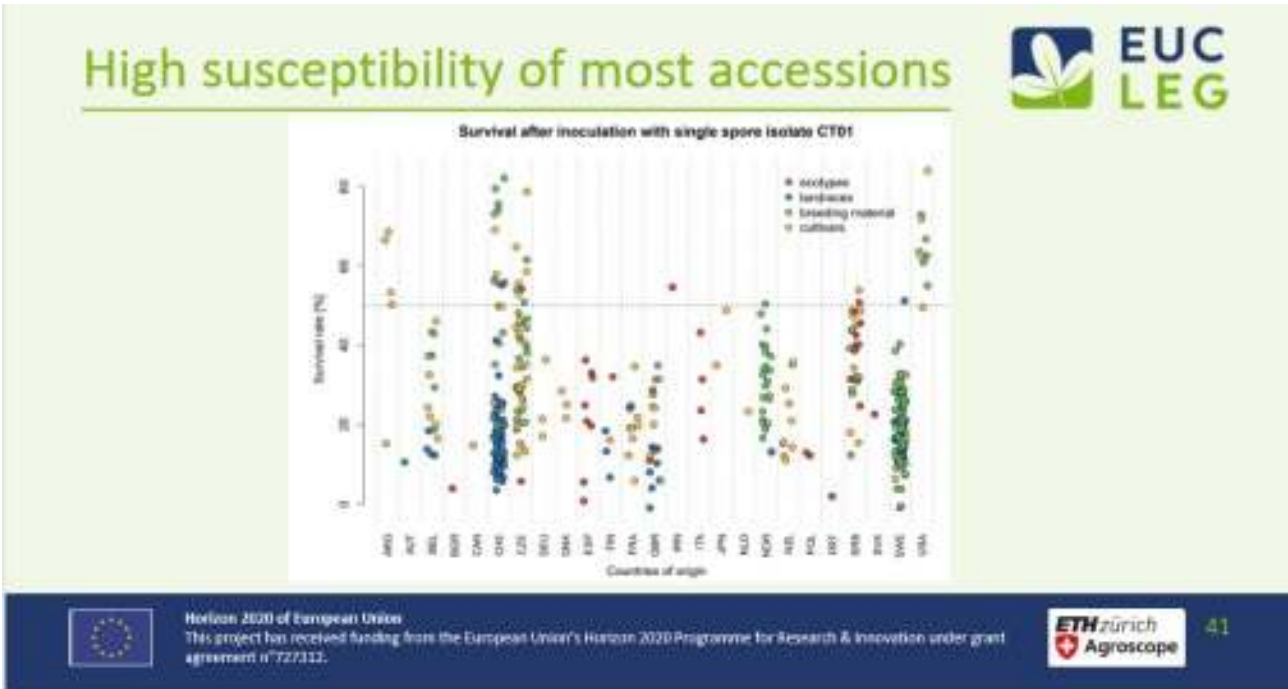


40

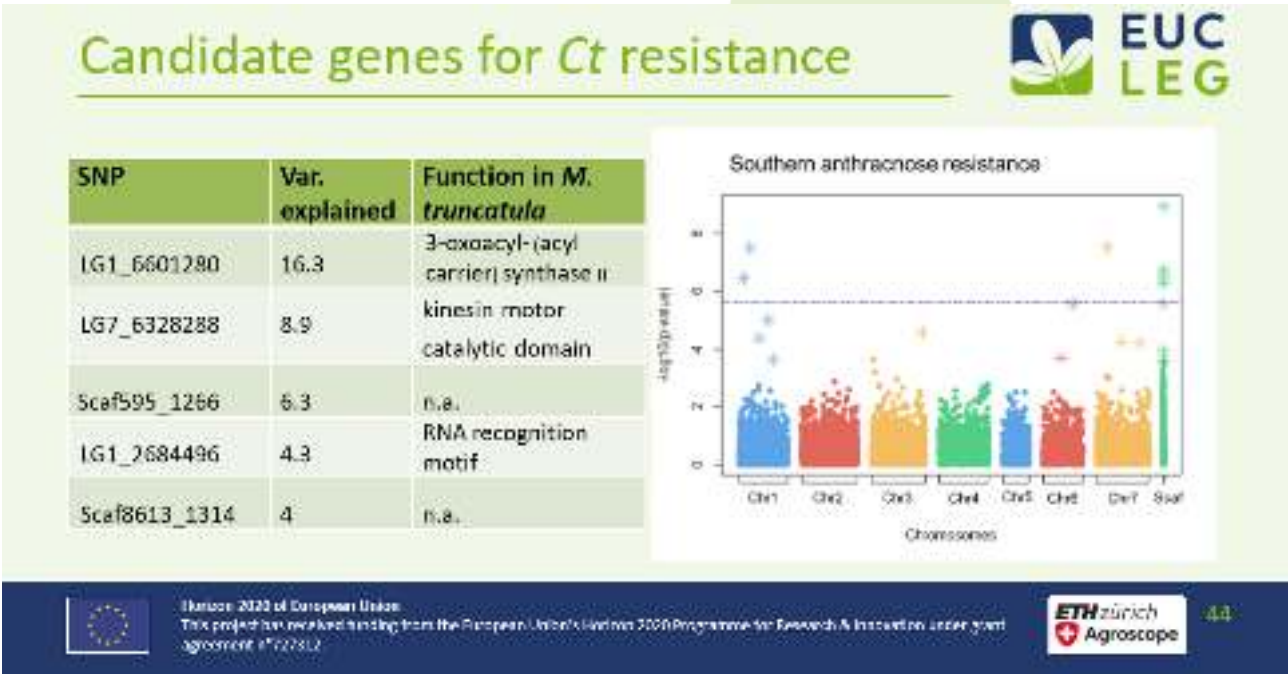
We used all the accessions we had available and performed artificial inoculation, using single spore isolates or one single spore isolate of *Colletotrichum trifolii*, and we analysed the survival rate and used this data for association analysis.



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



Most of the accessions we tested in the greenhouse were highly susceptible to southern anthracnose. There were a few accessions, mainly from Switzerland and the US, which showed quite a high degree of resistance, but a lot of the accessions showed very poor resistance to southern anthracnose.



GWAS using the panel of SNP described before identified a number of significant QTL and some of them explained quite a large proportion of the variance. For example, the one on chromosome one which explained up to 16% or the other one on linkage group 7 which explained around 9%. These are certainly

interesting candidates for further investigations and also for the development of marker assisted breeding strategies.

## Resistance to clover rot







Caused by *Sclerotinia trifoliorum* Erikks., reduces persistence





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


47


Another important disease in red Clover is clover rot caused by *Sclerotinia trifoliorum* and this work was conducted at ILVO.

## Artificial inoculation experiment






- 395 accessions
- 36 plants per accession
- 3 biological replicates
- Artificial inoculation
  - Single spore isolate of *S. trifoliorum*
  - Survival rate (%)
- Association analysis

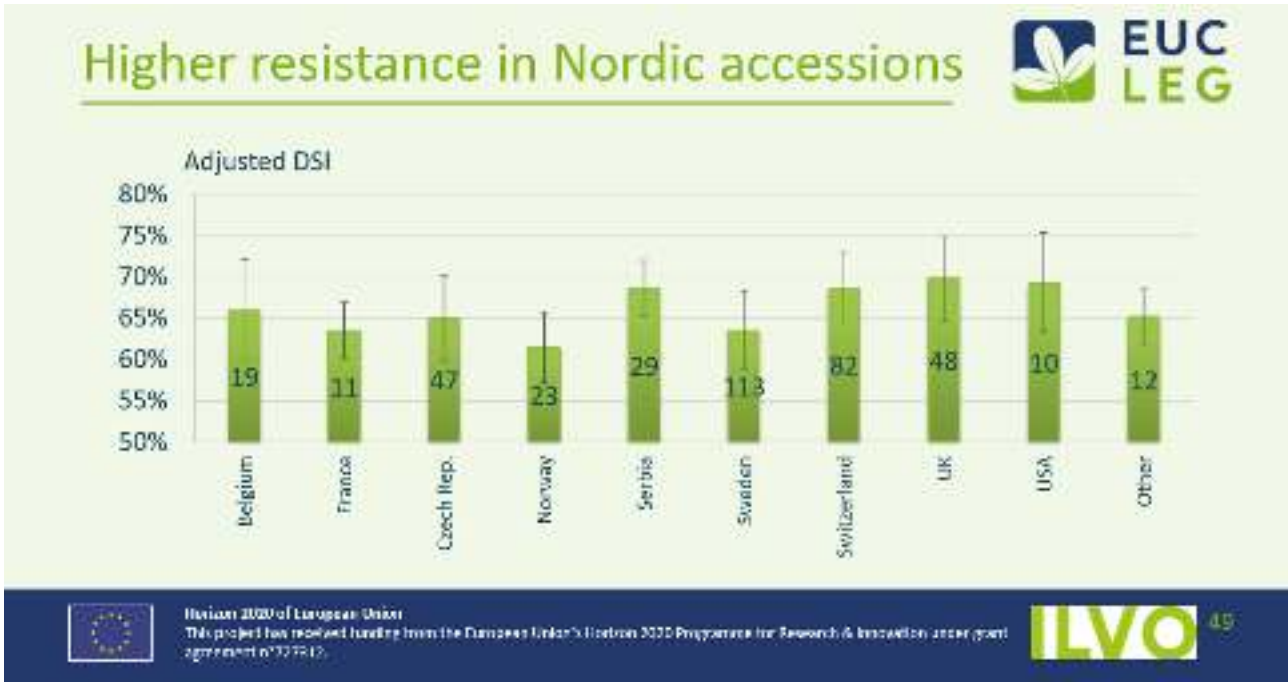


Horizon 2020 of European Union  
This project has received funding from the European Union's Horizon 2020 Programme for Research & Innovation under grant agreement n°727312.

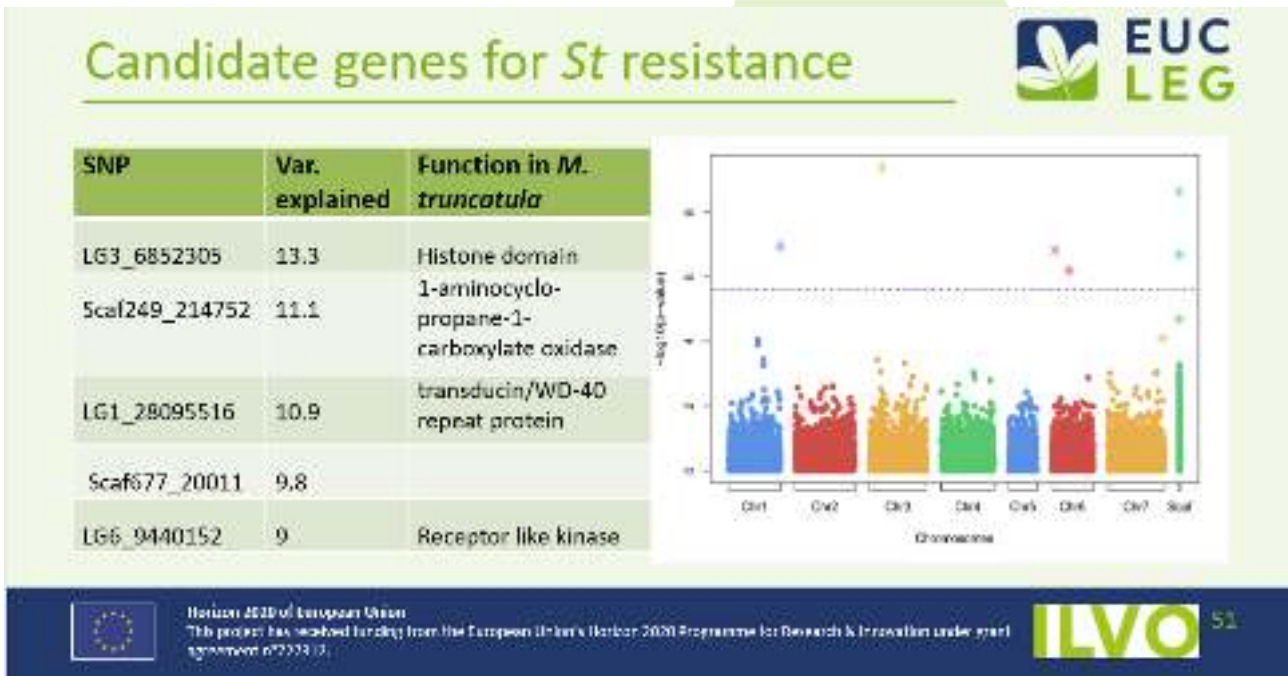

48

Basically a similar greenhouse experiment with artificial inoculation using one single spore isolate of *Sclerotinia trifoliorum* was performed and the survival rate was determined.





Then the disease severity index was calculated, and the higher this index is, the more susceptible the plants were in the greenhouse. Again, there was substantial susceptibility in these accessions, but this time the Nordic accessions seem perform better.



Again GWAS analysis identified a few interesting candidates, which we will now use for further characterization. For example one on chromosome 3 and also some explaining a very high amount of variability on the not mapped scaffolds.

## Freezing tolerance

Pregrowth



Cold acclimation



Freezing



Regrowth



- 393 accessions subjected to different freezing temperatures in the growth chamber
- LT50 – the temperature at which 50% of the plants perish



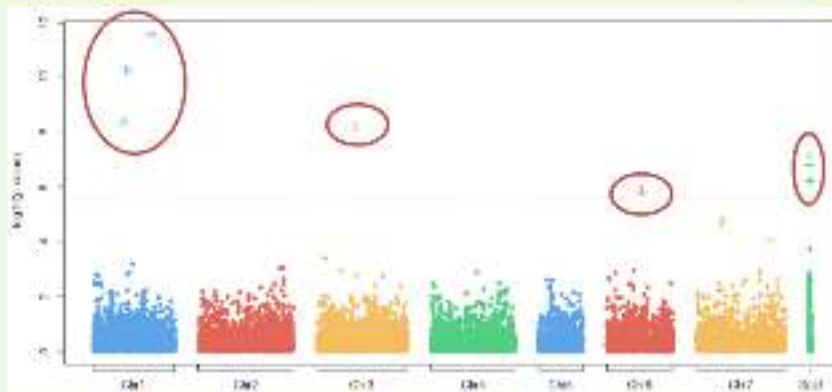
Horizon 2020 of European Union  
This project has received funding from the European Union's Horizon 2020 Programme for Research & Innovation under grant agreement n°727312



54

Another very important trait, particularly in northern countries, is freezing tolerance, which is known to be associated with general persistence. This is work conducted in Norway where the freezing tolerance was assessed under controlled conditions, by using 393 accessions subjected to different freezing temperatures in the growth chamber. LT50 was determined as the temperature where 50% of the plants perished.

## GWAS for Freezing Tolerance



**8 significant SNPs were detected explaining ~45% of the phenotypic variation**



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



56

GWAS analysis identified a set of eight significant SNPs, which explained around 45% of the phenotypic variation.

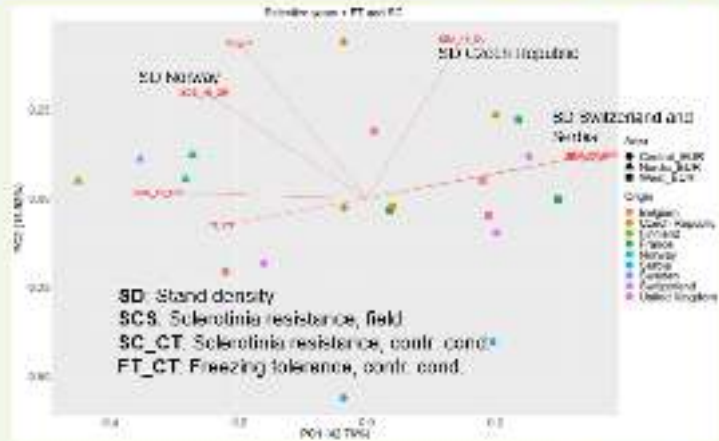


This project has received funding from the European Union's Horizon 2020 Programme for Research & Innovation under grant agreement n°727312.

**EUCLEG.eu**

## Persistence at four locations

- Persistence of 20 accessions measured as stand density
- Nordic varieties (triangles), one Belgian and one Swiss variety had
  - Better persistence in Norwegian location
  - Lower persistence in Czech, Swiss and Serbian test locations
  - Better *Sclerotinia* resistance
  - Better freezing tolerance



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



58

Also, we looked at the persistence, particularly in the field. The arrows indicate the different explanatory variables that were used. You can see that there were some accessions that were clearly performing better in Norway. They showed a better persistence in the Norwegian location and these were mainly the Nordic varieties and one Belgium and one Swiss variety. These varieties also showed better *Sclerotinia* resistance and better freezing tolerance. So these are traits also associated to better persistence. On the other hand you have the other accessions performing better on the Czech, Swiss and Serbian conditions.

## Persistence at Arneberg (GRAMINOR)

- 393 accessions
- Nordic material (triangles) performed better at Arneberg
  - Higher yield and persistency.
  - Better tolerance to *Sclerotinia*
  - Better freezing tolerance
  - Later flowering
  - Lower stands in the autumn after establishment



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



63


If we just look at the performance in the Nordic location at Arneberg at Graminor, the Nordic accessions are clearly separated by principle components associated to winter survival, stand density in the years 2000

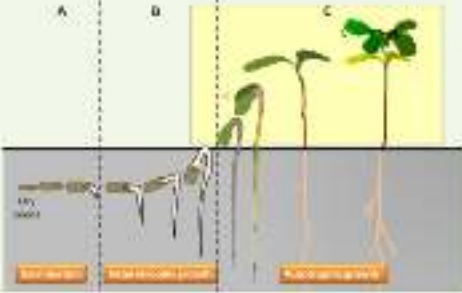


This project has received funding from the European Union's Horizon 2020 Programme for Research & Innovation under grant agreement n°727312.


and 2021, but also *Sclerotinia* resistance, and freezing tolerance under the controlled conditions from the remaining other accessions. On the other hand the other accessions were characterised by a better growth and indicated here by a higher stand height in the first year, in autumn 2018, before they had to undergo the severe winter conditions of Norway.


## Seedling emergence






trait	unit
Germination percentage	%
Germination speed	Time (h)
Germination heterogeneity (T80-T20)	Time (h)
Root length (RL)	cm
Hypocotyl length (HL)	cm
Emergence (%EM)	%
Speed of emergence	AHC






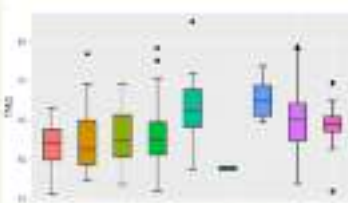
Horizon 2020 of European Union  
This project has received funding from the European Union's Horizon 2020 Programme for Research & Innovation under grant agreement n°727312.


ES

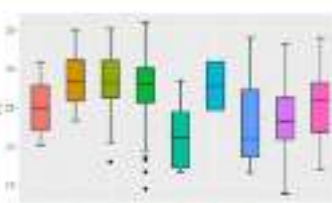
Seedling emergence, an important trait concerning the establishment rapid establishment of red clover swards, was investigated in France at INRAE.

## Variation for seedling emergence







**Time to mean germination**  
(time in h for fifty percent of the seed to germinate)




**Root length**  
(length in mm reached under dark conditions)



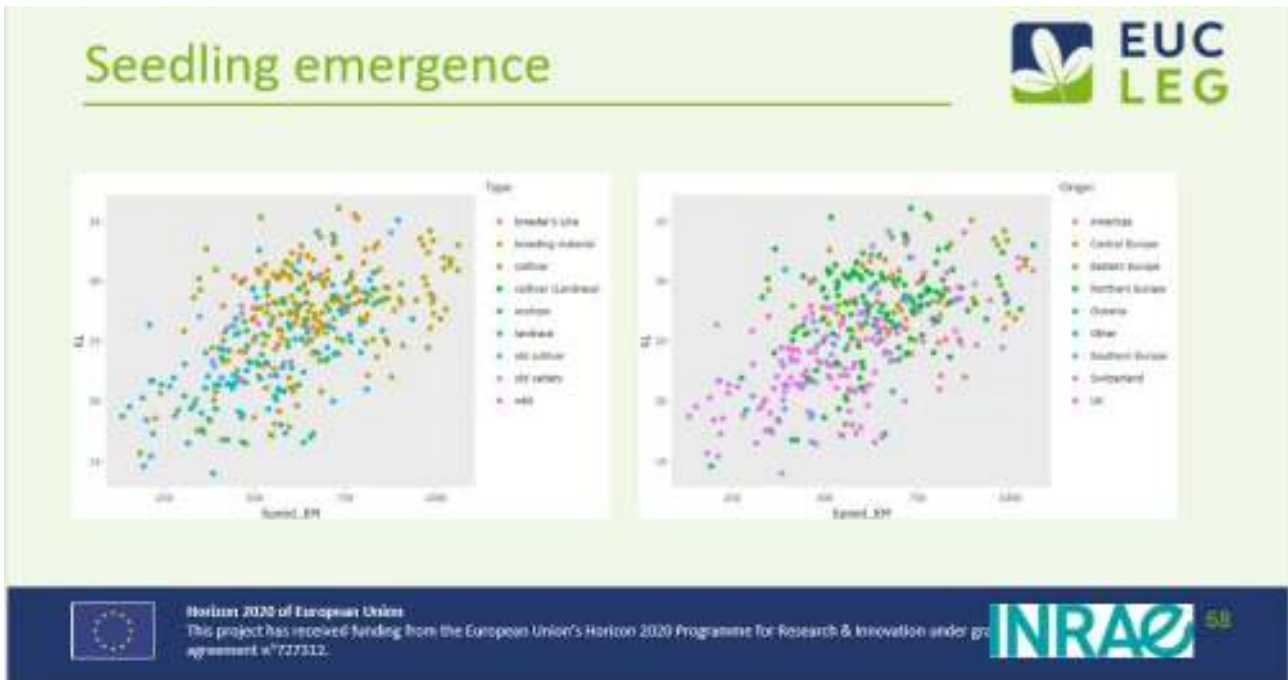
**Speed of emergence**  
(time in h for fifty percent of the seed to show emergence of cotyledons above soil surface)



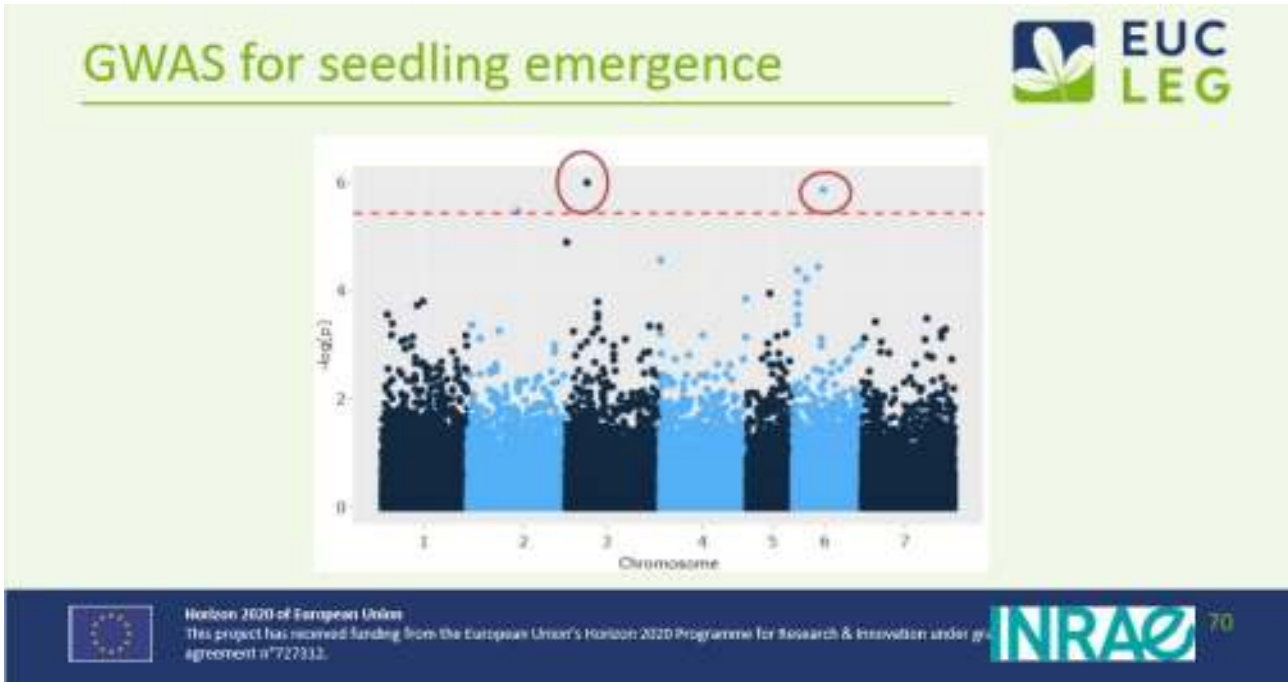
Horizon 2020 of European Union  
This project has received funding from the European Union's Horizon 2020 Programme for Research & Innovation under grant agreement n°727312.


ES

The results of three important traits are highlighted. One is the time to mean germination on the left, so the time in hours for 50% of the seeds to germinate. The next is the root length, which was reached under dark conditions during germination and emergence and finally the speed of emergence, which is characterised as the time for 50% of the seed to show emergence of the cotyledons above the soil surface. Depending on the origin of the accessions, we see quite a large variability in this germplasm and some accessions can actually have a higher time to mean germination, but still be faster in emergence.




Root length and speed of emergence were highly correlated, but more interestingly we saw a clear grouping of the land races or in this case the Swiss landraces, which seemed to be faster in emergence than most of the other accessions.





GWAS identified interesting candidates on two chromosomes which need to be investigated in more detail in the future.

## Conclusions



- Extensive set of well characterised red clover accessions
- Phenotypic and genotypic data for genomics assisted breeding strategies
- Basis for
  - elucidating the genetic control of key traits
  - further development of genomic selection models
- An important step towards improved red clover breeding

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

 71

So we have quite a nice set of well characterised red clover accessions and we generated a very nice data set of phenotypic and genotypic data, which will provide the basis for genomics assisted breeding strategies.



**(Q1) You refer to year one, was that the year of sowing or the year after?**

The yield data I showed was after sowing. The sowing year was considered as year zero and then we have year one which will be the first winter after sowing.

**(Q2) The next question is referring to the use of GBS on pooled populations samples. What is the relationship between the precision of the allele frequency estimates and the trait prediction accuracy?**

As far as I understand the question we would have to genotype individual plants and then determine the allele frequency on the basis of that and of course that is more accurate than pooling the individual genotypes, from populations, but this would be quite an expensive genotyping exercise. I am not sure it would actually improve a lot, because the phenotyping is done on the populations as well and not on single plants.

**(Q3) A specific question on the seedling emergence. Was any comparison made for performance of emergence in the field?**

That is actually a very interesting point. No, I don't think we have looked at that. We have the juvenile establishment in the field, so that's just a one year observation, but it would be very interesting to compare this with the data from the controlled experiments.

**(Q4) There was a question concerning the PCA analysis of the phenotypic data. Looking at the structure of the population, was it based on all the recorded traits or just a subset of the traits?**

I can't answer this question 100%, I know that a lot of traits were used, but probably not all of them, but most of the traits were used.

## About the author

**Dr Roland Kölliker** is a senior scientist in the Molecular Plant Breeding group of ETH Zurich with research focus in molecular genetics and genomics of grassland species and the development of tools for plant breeding and plant ecology. In particular, he is interested in the analysis of the complex interactions of forage grasses and legumes with their fungal and bacterial pathogens using genome sequencing, transcriptomics and applied statistics. As the species expert for red clover in EUCLEG, Roland Kölliker supervised and coordinated research activities on this important forage legume.

**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 30<sup>th</sup> September and 1<sup>st</sup> October 2021**

Recording link to the presentation: <https://youtu.be/J7l-eXJlme0>

