

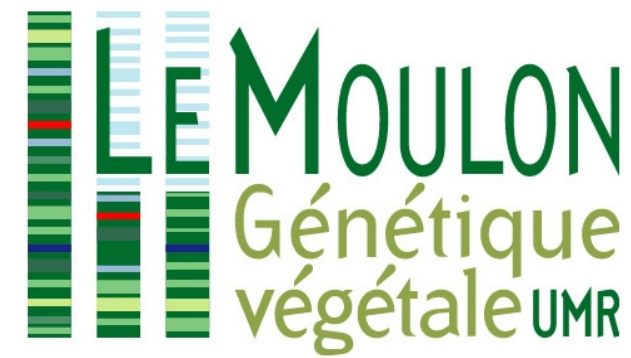
# Thaliadb

## A database dedicated to association genetics in plants

Yannick De Oliveira<sup>1</sup>, Julien Cornouiller<sup>1</sup>, Guy-Ross Assoumou-Ella<sup>1</sup>, Johann Joets<sup>1</sup> and Alain Charcosset<sup>1</sup>

<sup>1</sup> UMR Génétique Végétale, INRA – Université Paris-Sud - CNRS, Ferme du Moulon, 91190, Gif-sur-Yvette, France

Contact : ydeoliveira@moulon.inra.fr



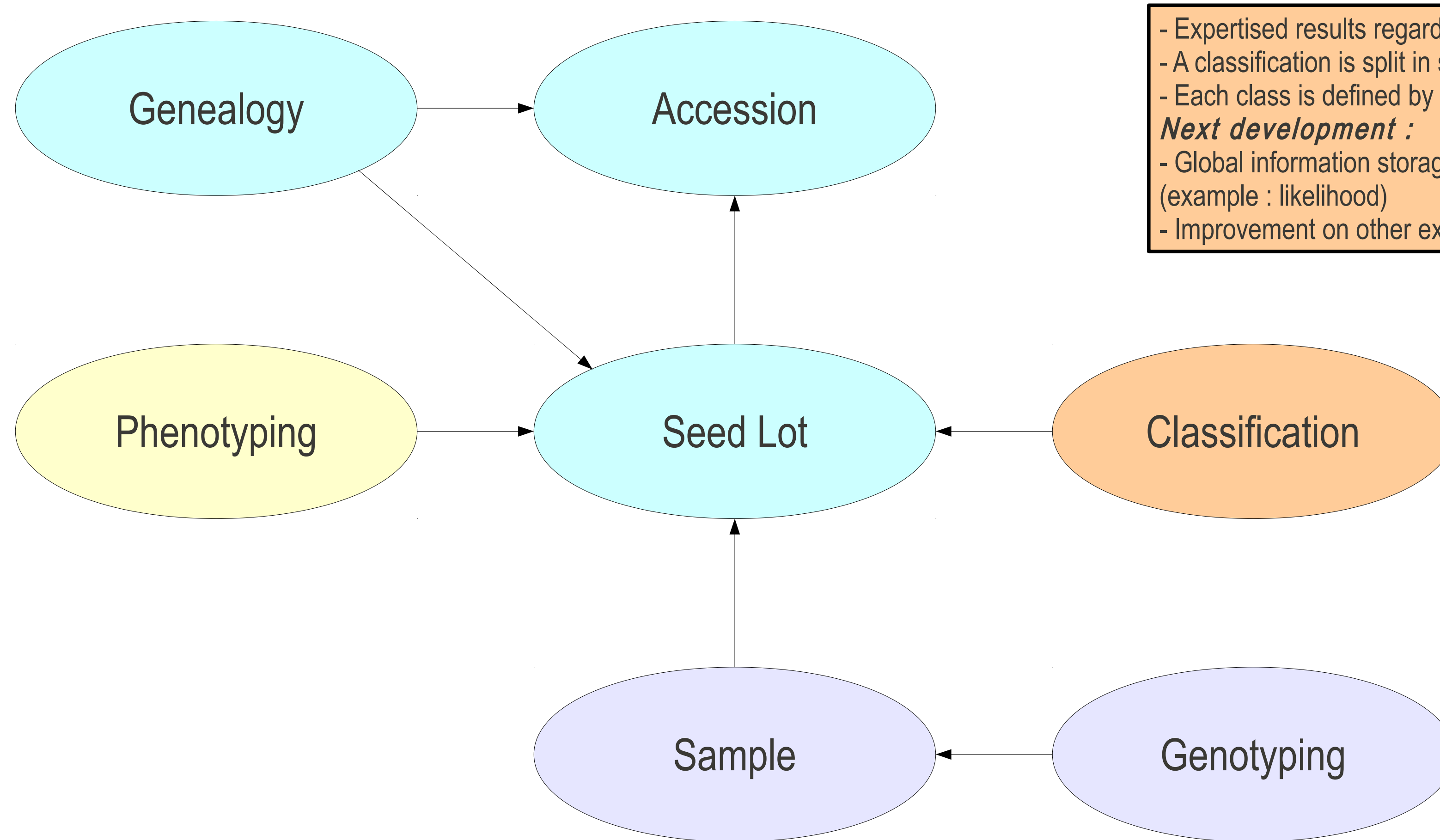
### Introduction

Diversity and association genetics studies lead to manipulate a large number of individual, lines, clones and/or populations. Moreover, emergence of high-throughput technologies for both genotyping and phenotyping generates a large amount of data. These need to be stored and managed in order to perform requests and organize datasets to conduct association genetics studies.

The Thaliadb database has been developed with this aim. It manages genetic resources, phenotyping and genotyping data, and also population structure information. Thaliadb enables data extraction in formats used by genetic association softwares.

### Data Structure

- Dynamic description of types
- URL to pictures can be stored
- Genealogy management
- Next development :**
- Improvement of genealogy functionalities
- Add a link to Sirégal database [1] (taxonomy informations)

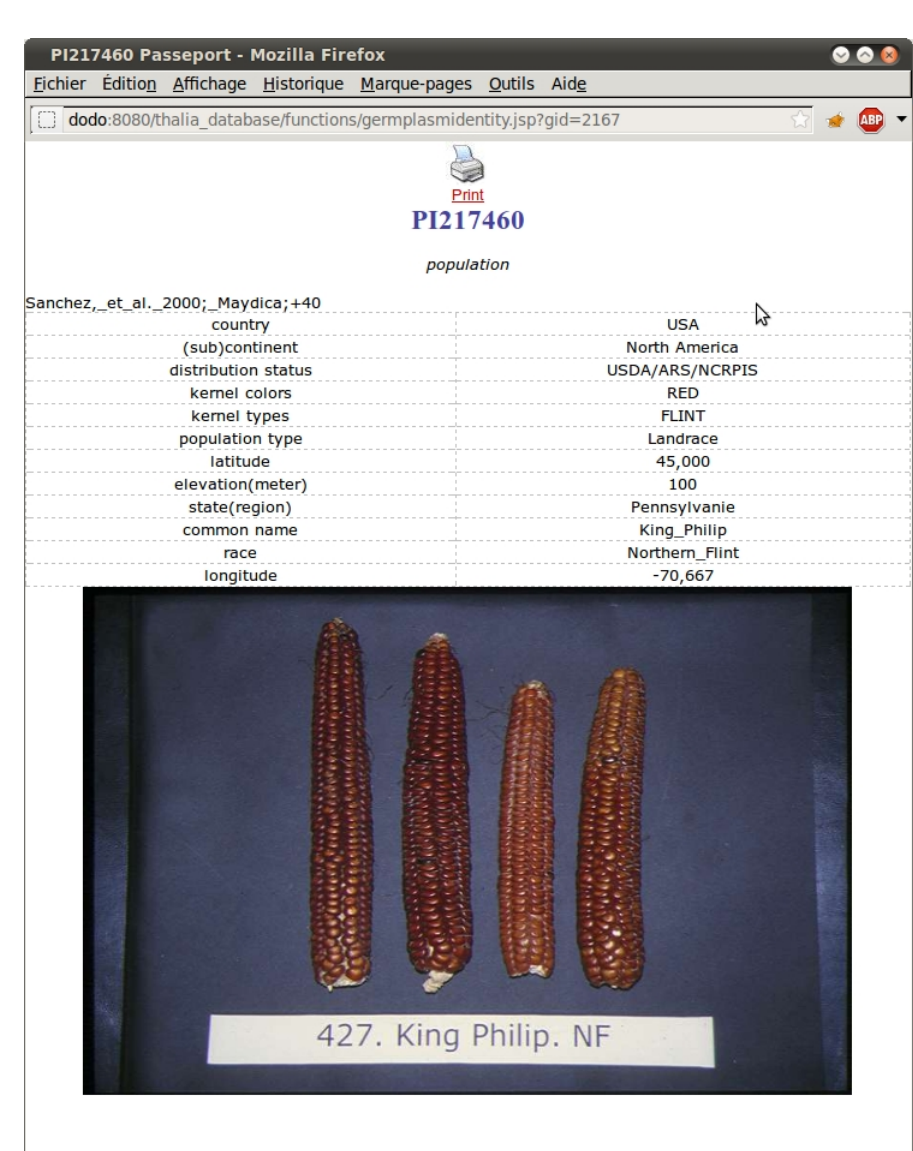


- Expertised results regarding population structure
- A classification is split in several classes
- Each class is defined by a value and linked to a seed lot
- Next development :**
- Global information storage concerning statistical analysis (example : likelihood)
- Improvement on other expertised data type (PCA)

- Average values stored
- Data are linked to seed lot and defined by an environment, a trait, and a statistical method
- Next development :**
- Use a standard ontology for traits definition [2]
- Homogeneity with Ephesis [3]

- Dynamic description of marker types (SNP, RFLP, SSR etc.)
- A (DNA) sample is characterized for a locus in a given experiment (conducted by a person in an institute, both documented in the database).
- One or more alleles may be observed with given frequencies.
- Alleles are described following a referential used for an experiment.
- A correspondence option allows to bind heterogeneous data (observed with different referentials).
- Next development :**
- Improve performance and visualisation strategy to submit NGS data

### Data Extraction and Visualization



Seed lot	Vgt1 MITE A	Vgt1 MITE B	Days to pollen	Northern flint	Andes	Caribbean	Corn belt dent	Europe	Mexican	Italian
PPS48	0,00	1,00	57,9	0,9140	0,0104	0,0088	0,0092	0,0162	0,0106	0,0308
PI217460	0,21	0,79	72,3	0,9156	0,0056	0,0088	0,0412	0,0084	0,0096	0,0110
PI218143	0,52	0,48	85,8	0,0852	0,0178	0,0164	0,6516	0,0178	0,2016	0,0088
PI213719	0,80	0,20	81,8	0,0810	0,0136	0,0268	0,8024	0,0150	0,0204	0,0410
PPS774	1,00	0,00	77,2	0,0090	0,0664	0,2494	0,0174	0,0410	0,0898	0,5274

Thaliadb enables data extraction and visualization at multiple level :

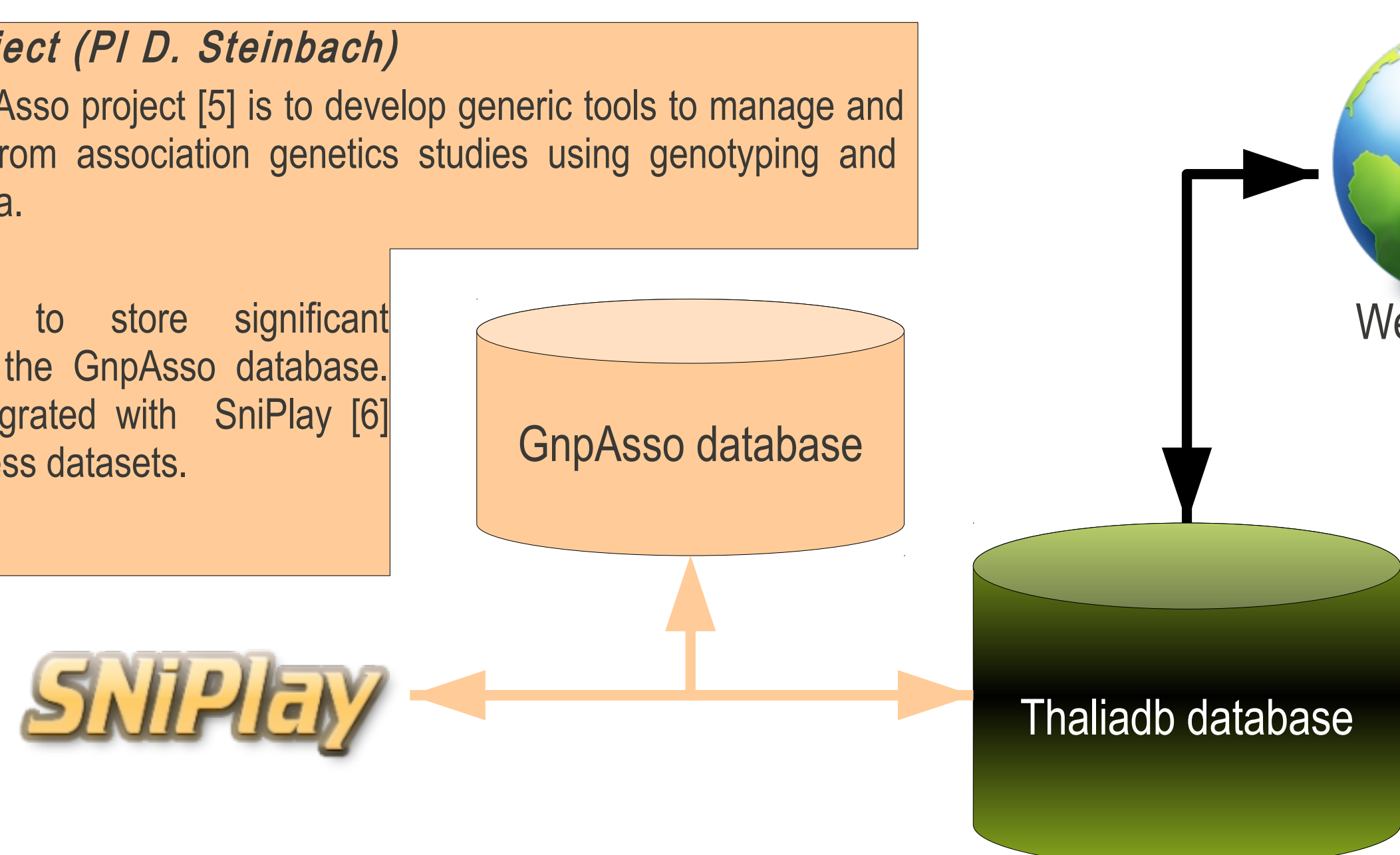
- A view of a single entity (population, loci,...), picture on the left) with all data related to this entity
- A table view (on the top) describing several entities. Thaliadb allows user to request and extract data in format compatible with association genetic software.
- On the right, a map view (Google map) with the geographic repartition of accession genotypes. This example shows the frequencies of the allele B of VGT1-MITE locus for accessions located in Europe.



### Main projects using Thaliadb

**GnpAsso project (PI D. Steinbach)**  
The aim of GnpAsso project [5] is to develop generic tools to manage and exploit results from association genetic studies using genotyping and phenotyping data.

The goal is to store significant associations in the GnpAsso database. Thaliadb is integrated with SniPlay [6] pipeline to process datasets.



**Seed lot history management project**  
This project aim to develop a database to store and manage seed lot history data (relations between seed lots, stock etc.)  
The web service we are developing will allow request on phenotyping and genotyping data.

**DROPS project**  
DROPS [4] aims at developing novel methods and strategies for genetic yield improvement under dry environments and for enhanced plant water-use efficiency. In this project, we are developing a web service to allow user request Thaliadb through DROPS application.

### References

[1] <http://urgi.versailles.inra.fr/siregal/siregal/welcome.do>  
 [2] P. Jaiswal, D. Ware, J. Ni, K. Chang, W. Zhao, S. Schmidt, X. Pan, K. Clark, L. Teytelman, S. Cartinhour, L. Stein, S. McCouch, Gramene: development and integration of trait and gene ontologies for rice. *Comp Funct Genom* 3: 132–136, 2002.  
 [3] <http://urgi.versailles.inra.fr/Projects/URGI-software/Ephesis>  
 [4] <http://www.drops-project.eu/>  
 [5] [http://www.agence-nationale-recherche.fr/projet-anr/?tx\\_lwmsuivibilan\\_pi2\[CODE\]=ANR-10-GENM-0006](http://www.agence-nationale-recherche.fr/projet-anr/?tx_lwmsuivibilan_pi2[CODE]=ANR-10-GENM-0006)  
 [6] A. Dereeper, S. Nicolas, L. Le Cunff, R. Bacilieri, A. Doligez, J-P Peros, M Ruiz, P This, SniPlay: a web-based tool for detection, management and analysis of SNPs. Application to grapevine diversity projects. *BMC Bioinformatics* 12:134, 2011.

Funding : Thaliadb is developed with funding from ANR project ANR-10-GENM-0006 (PI D. Steinbach)