Thaliadb : A database dedicated to association genetics in plants

Guy-Ross Assoumou-Ella, Yannick de Oliveira, Johann Joets, Alain Charcosset

Thaliadb : A database dedicated to association genetics in plants

Guy-Ross Assoumou-Ella[1], Yannick De Oliveira[1], Johann Joets[1] and Alain Charcosset[1]

[1] INRA, UMR Génétique Quantitative et Evolution - Le moulon, F-91190 Gif-sur-Yvette, France

JOBIM Clermont-Ferrand - 6-9 juillet 2015

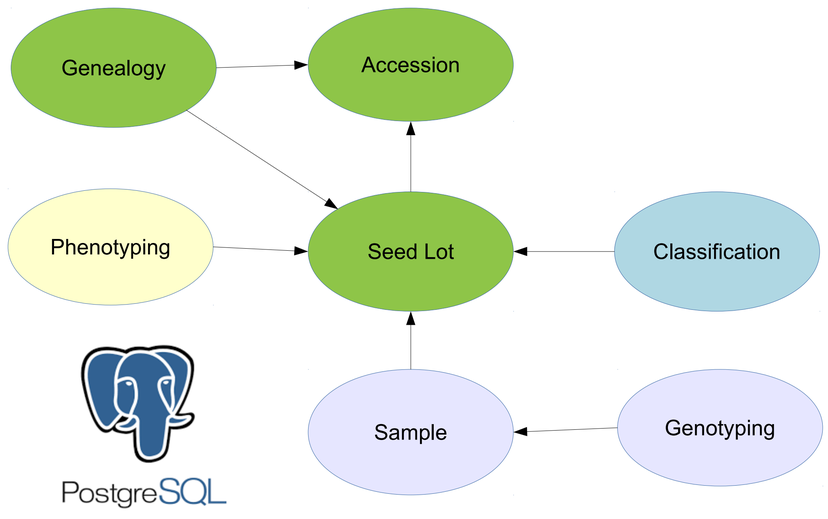
Introduction

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 software.

Data Structure

- Dynamic description of types
- URL to pictures can be stored
- Genealogy management

- Dynamic description of types
- URL to pictures can be stored
- Genealogy management

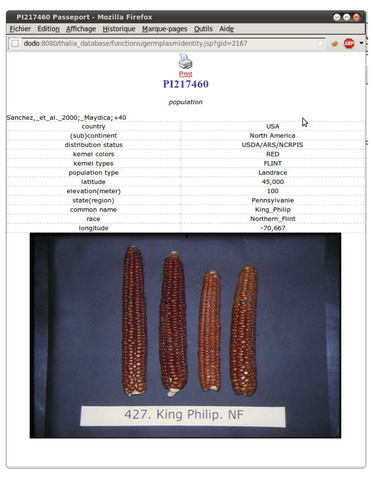


Thaliadb has been developed using the PostgreSQL database management system. Recently we migrate the Genotyping part of the database in the MongoDB system to manage the data produced by high-throughput technologies.

- Expertised results regarding population structure
- A classification is split in several classes
- Each class is defined by a value and liked to a seed lot

- Dynamic description of marker types (SNP, RFLP, SSR, etc.)
- A (DNA) sample is characterized for a locus in given experience (conducted by a person in an institute, both documented in a 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).

Data Extraction and visualization



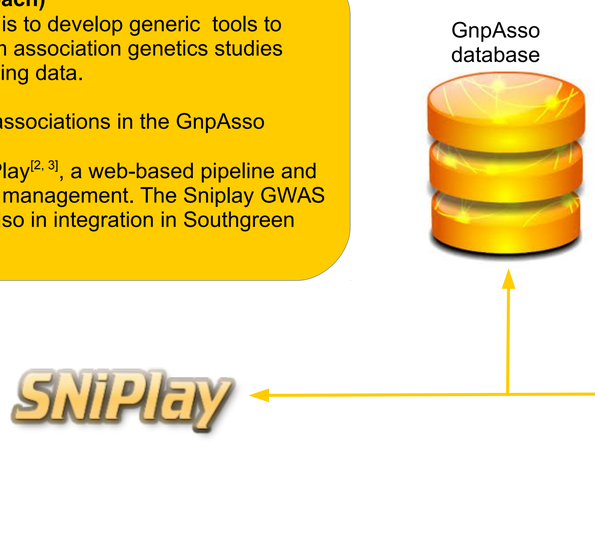
Seed lot	Vgt1 MITE A	Vgt1 MITE B	Days to pollen	Northern flint	Andes	Caribbean	Corn belt dent	Europe	Mexican	Italian
PPS48	0,00	1,00	57,9	0,9140	0,0104	0,0088	0,0092	0,0162	0,0106	0,0308
PI217460	0,21	0,79	72,3	0,9156	0,0056	0,0088	0,0412	0,0084	0,0096	0,0110
PI218143	0,52	0,48	85,8	0,0852	0,0178	0,0164	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[1, 4] is to develop generic tools to manage and exploit results from association genetics studies using genotyping and phenotyping data.

The goal is to store significant associations in the GnpAsso database.
Thaliadb is integrated with SniPlay[2, 3], a web-based pipeline and database for SNP analysis and management. The Sniplay GWAS analysis workflow is currently also in integration in Southgreen and URGI Galaxy servers.



***Amaizing project*[5]**
Amaizing is an ANR project of 8 years. The aim of this project is to develop knowledge, selection methods and cultural practices to elaborate variety with a high yield, and an improved environmental value.
In this project Thaliadb is used to manage and store the genotyping and phenotyping data

***Interoperability with SHiNeMaS database*[6]**
SHiNeMaS is a database that store and manage seed lot history data (relations between seed lot, cultural practices data, stock etc...).
The aim is to establish interoperability between these two tools.

References

[1] http://www.agence-nationale-recherche.fr/projet-anr/?tx_lwmsuivibilan_pi2[CODE]=ANR-10-GENM-0006
[2] A. Dereeper, F. Homa, G. Andres, G. Sempere, G. Sarah, Y. Hueber, J.F. Dufayard and M. Ruiz, SNiPlay3: a web-based application for exploration and large scale analyses of genomic variations *Nucleic Acids Research* [Epub ahead of print], 2015.
[3] http://sniplay.southgreen.fr/cgi-bin/home.cgi
[4] https://urgi.versailles.inra.fr/association/
[5] http://www.amaizing.fr/
[6] http://moulon.inra.fr/index.php/en/tranverse-team/atelier-de-bioinformatique/projects/181