Project No. *283496*

transPLANT

Trans-national Infrastructure for Plant Genomic Science

Instrument: **Combination of Collaborative Project and Coordination and Support Action**

Thematic Priority: FP7-INFRASTRUCTURES-2011-2

D10.2
Software for analysis of genome-wide association data

Due date of deliverable: 31.08.2013 (M24)
Actual submission date: 29.08.2013

Start date of project: 1.9.2011 Duration: 48 months

Organisation name of lead contractor for this deliverable: GMI

Project co-funded by the European Commission within the Seventh Framework Programme (2011-2014)		
Dissemination Level		
PU	Public	x
PP	Restricted to other programme participants (including the Commission Services)	
RE	Restricted to a group specified by the consortium (including the Commission Services)	
CO	Confidential, only for members of the consortium (including the Commission Services)	

Contributor
GMI

Introduction

Deliverable reference number:D10.2

The goal of this deliverable is to develop software for analysis of **G**enome **W**ide **A**ssociation **D**ata (GWAS). The deliverable consists of two tasks: 1) develop a web-interface that allows real-time GWAS in A. thaliana; and 2) develop tools for meta-analysis of pleiotropy.

In Task 1 we developed a web-application that allows researchers to easily carry out GWAS in real-time. The user only needs to provide phenotypes as a comma-separated file and choose an appropriate method (mixed model, linear model, non-parametric test) and the analysis should be carried out automatically on the server in real-time. The results of the analysis are presented to the user with interactive Manhattan charts. The main focus of Task 1 is simplicity and usability. The main purpose is to lower the barriers for researchers to run GWAS by removing as many stumbling blocks as possible.

Task 2 builds on top of Task 1 but goes much further. While the primary rationale for the tools described under Task 1 is to enable quick and easy GWAS, Task 2's main goal is to make it possible to compare the results of individual studies with other published results, and thus allow systems-level insights into pleiotropy. Similar to Task 1 the tools (web-application) developed in Task 2 should allow the researcher to upload phenotype data, run different analysis methods and display the results with interactive charts. However, unlike Task 1 all data should be stored in a central database containing both private and public data. The user can change the permissions of the uploaded data (i.e. share it with collaborators) and compare the results to other studies to which the user has permission (public or shared).
This way it should be possible to make phenotypic associations part of the genome annotation. The pleiotropy analysis can range from simply listing phenotypes with which a particular polymorphism appears to be associated, to network displays of correlations between phenotypes. Because the amount of entities (phenotypes, studies, GWAS results) that have to be stored and displayed in the database/web-application can potentially go into the thousands, the main focus of the tools developed in Task 2 is to prepare and display the data in a way that the user can easily navigate, browse and find what he/she is looking for. To accomplish this requirement, a lot of effort was put into good user experience (UX), responsiveness and speed, and intelligent ways to annotate the data, for example by using ontologies.

Methods

Both tools in Task 1 and 2 were developed as a web-application. The choice for a web-application instead of a client-side native application was driven by the goal to minimize the problems related to deployment and availability. Browsers have evolved significantly in recent years and thus provide a mature platform for high performance web-applications. Both web-applications were developed using the Google Web Toolkit (GWT). GWT is a toolkit created and used by Google to create rich web-applications. The big advantage of using a toolkit like GWT is that components (i.e. interactive charts, widgets, etc.) developed for one application can be easily re-used in other applications. For visual representations of the data, interactive charts were either deployed or created/modified when unavailable. These interactive charts were developed using modern web-technologies like HTML5 Canvas and Scalable Vector Graphics (SVG) as self-contained components.
The web-application in Task 2 stores all the data in a PostgreSQL database and uses the Elasticsearch fulltext engine to allow users to search through the data. The analyses (GWAS runs) are distributed using the Celery

distributed task queue and executed on a HPC cluster. This allows us to easily scale with the number of uses and analyses respectively.

This is the overview of technologies that were used in both tasks:

Common to Task 1 & Task 2:

- Google Web Toolkit (GWT) for frontend development [1]
- HDF5 for storing raw results [2]
- HTML5 Canvas and SVG for interactive charts [3]
- JSON data-structure for data exchange protocol
- Python's numpy/scipy and fast Linear Algebra Packages (GotoBlas2, MKL) for fast analysis [4]

Only Task 1:

- CherryPy application server (backend) [5]
- PyTables (HDF5 wrapper) for storing results [6]

Only Task 2:

- Spring as application framework (backend) [7]
- RBAC permission model (Spring security)
- Elasticsearch fulltext search engine [8]
- Python Celery as distributed task queue (analysis) [9]
- PostgreSQL Database based on GDPDM [10]
- OpenID authentication

References:

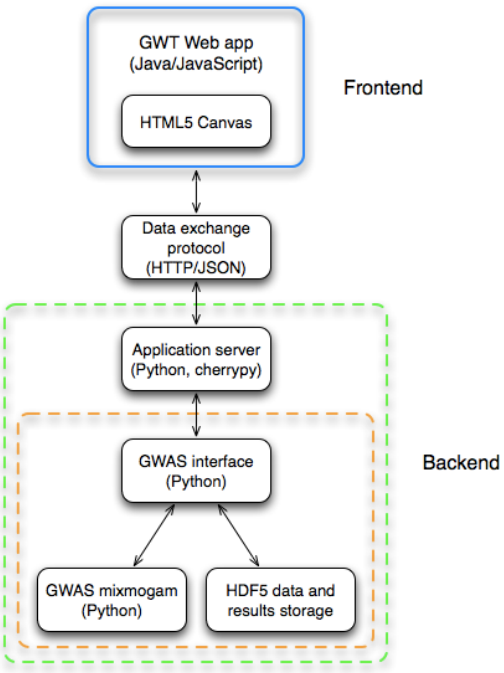
1. http://www.gwtproject.org/
2. The HDF Group (2000-) Hierarchical data format version 5. http://www.hdfgroup.org/HDF5
3. Dygraphs (http://dygraphs.com), Processing.js (http://processingjs.org/) and Google Chart Tools (https://developers.google.com/chart/)
4. Travis E. Oliphant (2007). Python for Scientific Computing. Computing in Science & Engineering 9, 90. (http://www.scipy.org/)
5. http://www.cherrypy.org/
6. Alted F, Vilata I, and others (2002-) PyTables: hierarchical datasets in python. (http://www.pytables.org)
7. http://www.springsource.org/
8. http://www.elasticsearch.org/
9. http://www.celeryproject.org/
10. The Genomic Diversity and Phenotype Data Model (http://www.maizegenetics.net/gdpdm/)

Results (if applicable, interactions with other workpackages)

Taks 1:

The web-application (GWAPP) that was developed in Task 1 has a simple architecture (see Figure 1). The backend is a simple CherryPy python application server that is responsible for running the GWAS analysis, storing the data in HDF5 files and providing the frontend with data. The frontend is a Javascript application that was developed using GWT. GWAPP is accessible to everyone via (http://gwapp.gmi.oeaw.ac.at) and the source-code is open source and available at (https://.github.com/timeu/GWAPP).

Figure 1 – Structure of the web-application



The main goal of GWAPP is to provide researchers with an easy way to run GWAS and to present the data to the user in a useful way.

Here are some examples of the user interface of GWAPP:

Figure 2 – Dataset view

The dataset view allows the user to create a subset for an uploaded phenotype (i.e. only accessions from a specific region).

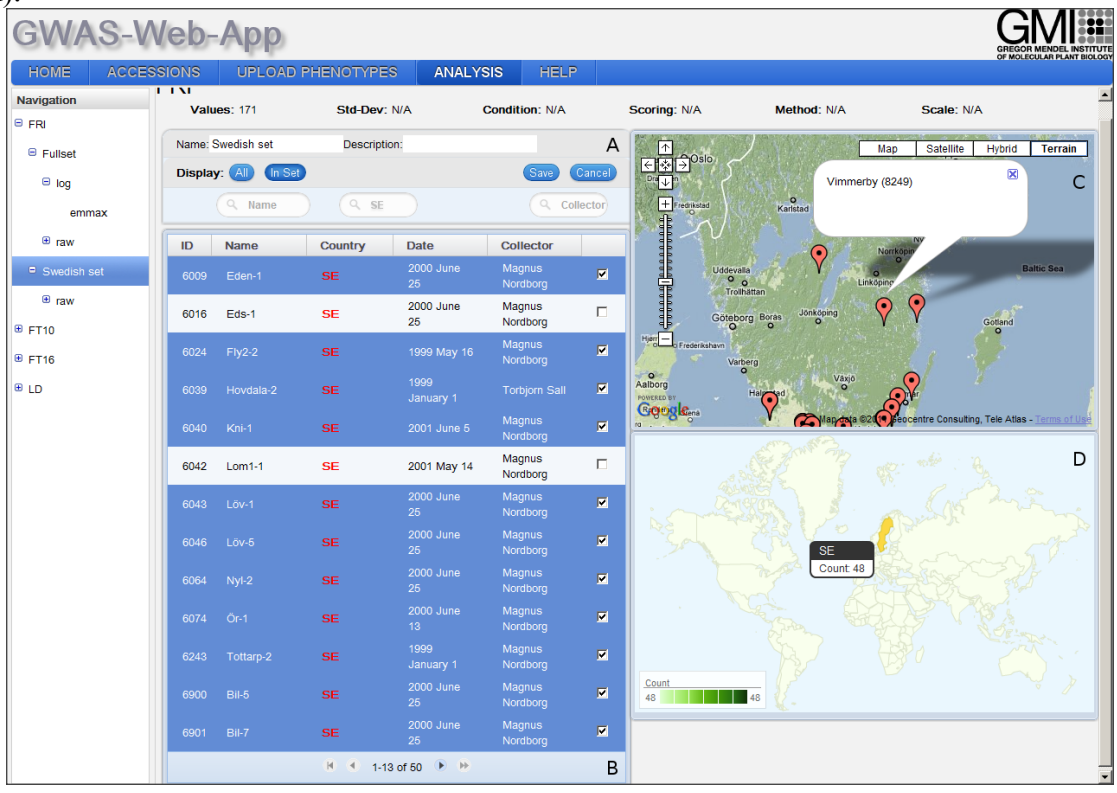


Figure 3 – Transformation view

The transformation view allows the user to choose a transformation from a list of available transformations and view the phenotype values for each accession.

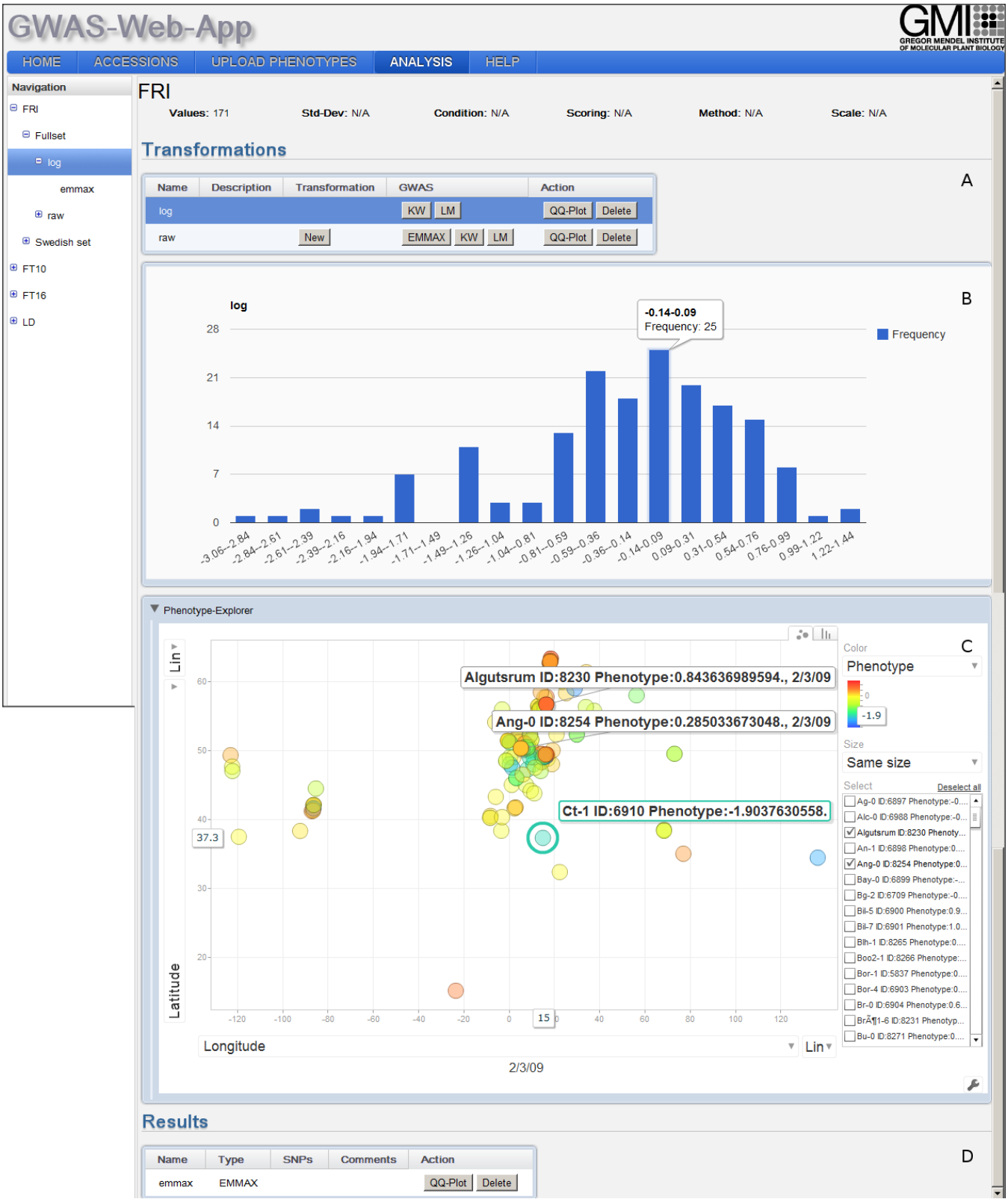


Figure 4 – LD visualization

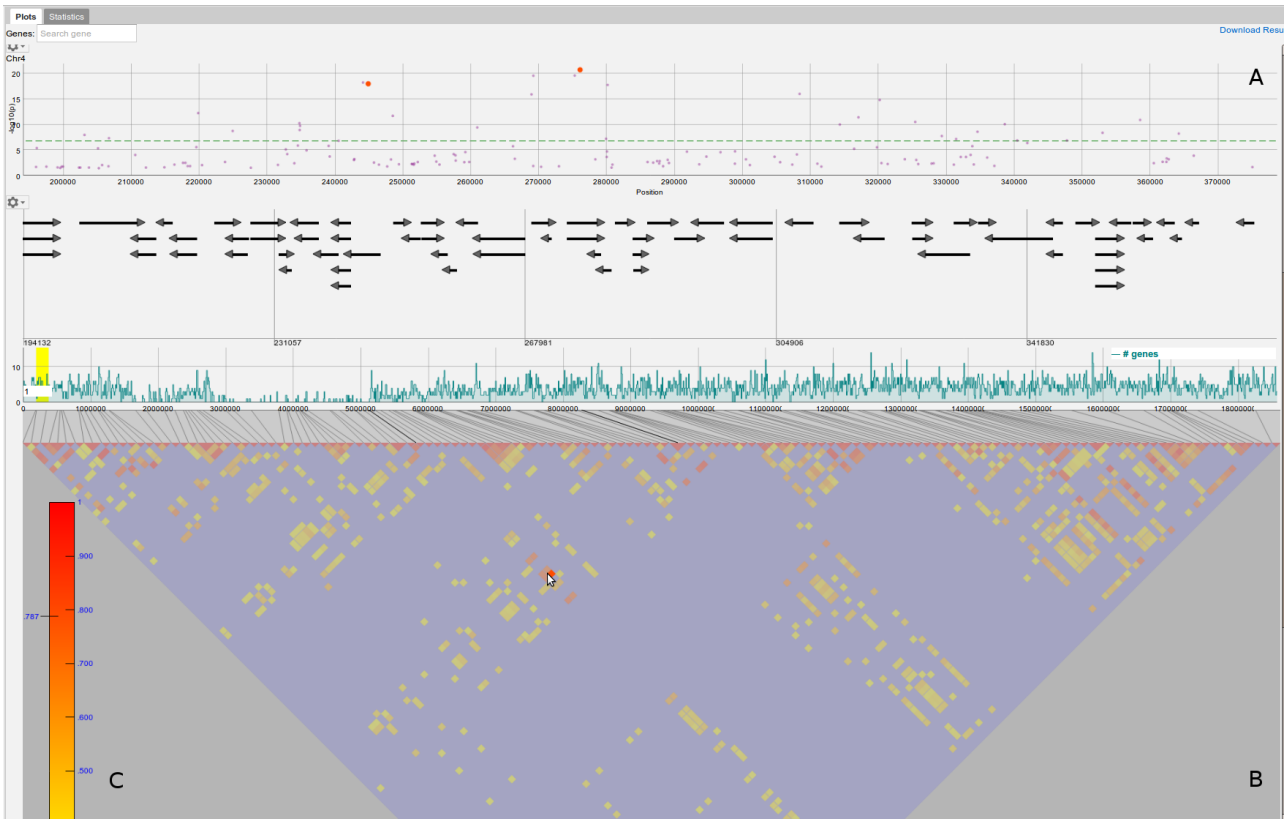
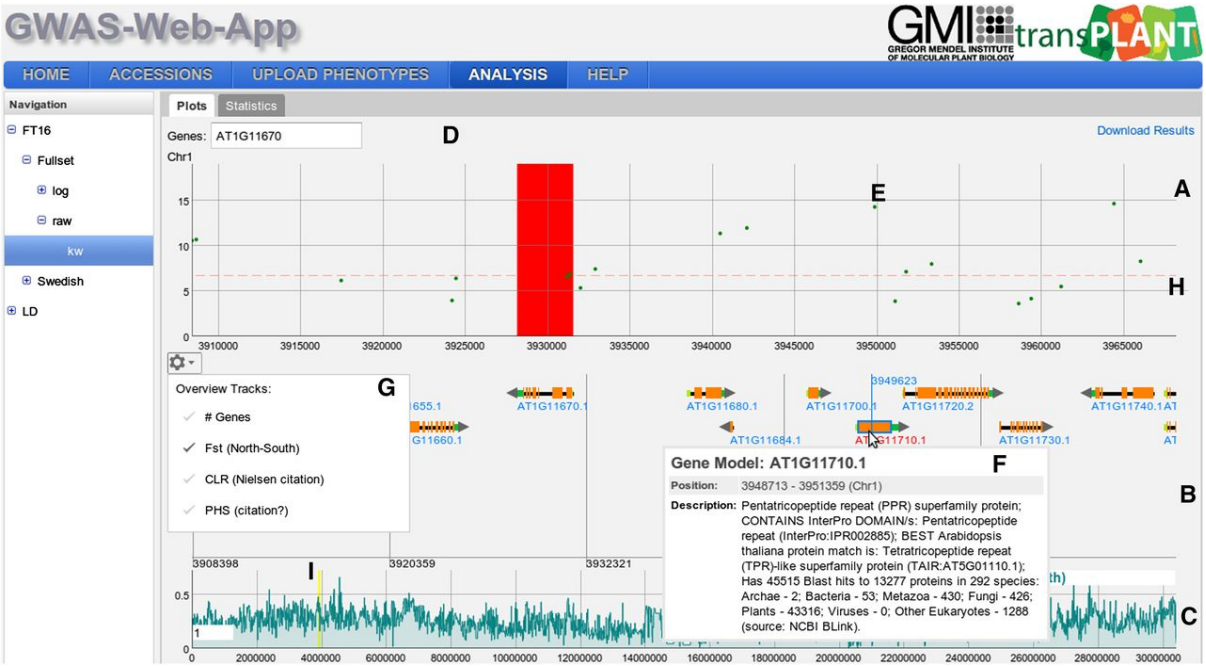


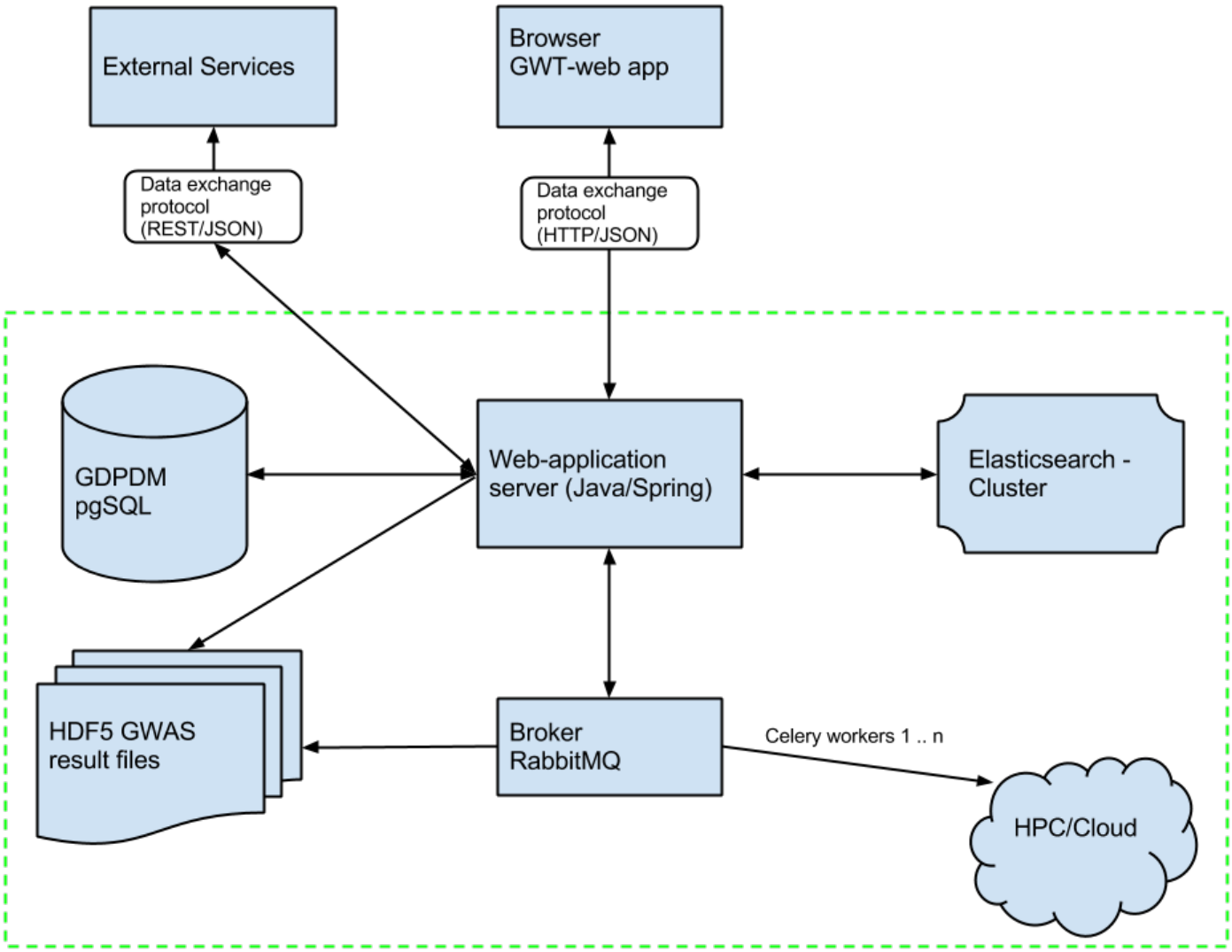
Figure 5 – Interactive Manhattan charts with gene annotation



Taks 2:

The web-application (GWA-portal) developed in Task 2 has a more sophisticated architecture (see Figure 6) as it goes much further than GWAPP's capabilities.

Figure 6 – Architecture of GWA-Portal



Unlike GWAPP, the new web-application stores all the information in a central database. The database structure is based on the **G**enomic **D**iversity and **P**henotype **D**ata **M**odel (GDPDM) used to store maize data, but which also works for other plants.
A fulltext search engine indexes the data and enables the user to search and browse the data.
A distributed task queue is used to run the various analyses. **R**ole **B**ased **A**ccess **L**ists (RBAC) are used to define permissions for the major entities (phenotype, study and analysis). Users can easily authenticate using an OpenID provider (Google, Facebook, Twitter), which frees the user from remembering yet another account. Some effort was put into optimizing the user experience (UX) and user interface so that users can easily browse through potentially huge amounts of information by providing the user with visual representations in the form of charts and graphs, and wizards for adding/modifying information. Additionally, speed and responsiveness, especially when filtering and faceting the data, was an important requirement.
GWA-portal connects and links different information ranging from taxonomy to ontologies and thus provides different views into the stored information.

The following examples of the user interface highlight this aspect:

Home section:
The "Home" section represents the landing page and shows general information and statistics about the information available on GWA-Portal. Users can also take a Tour. The user information popup shows recent notifications (i.e. finished GWAS runs). Users can also use an easy-to-follow wizard to upload phenotypes and create GWAS analysis (Figure 8)

Figure 7 – Landing page

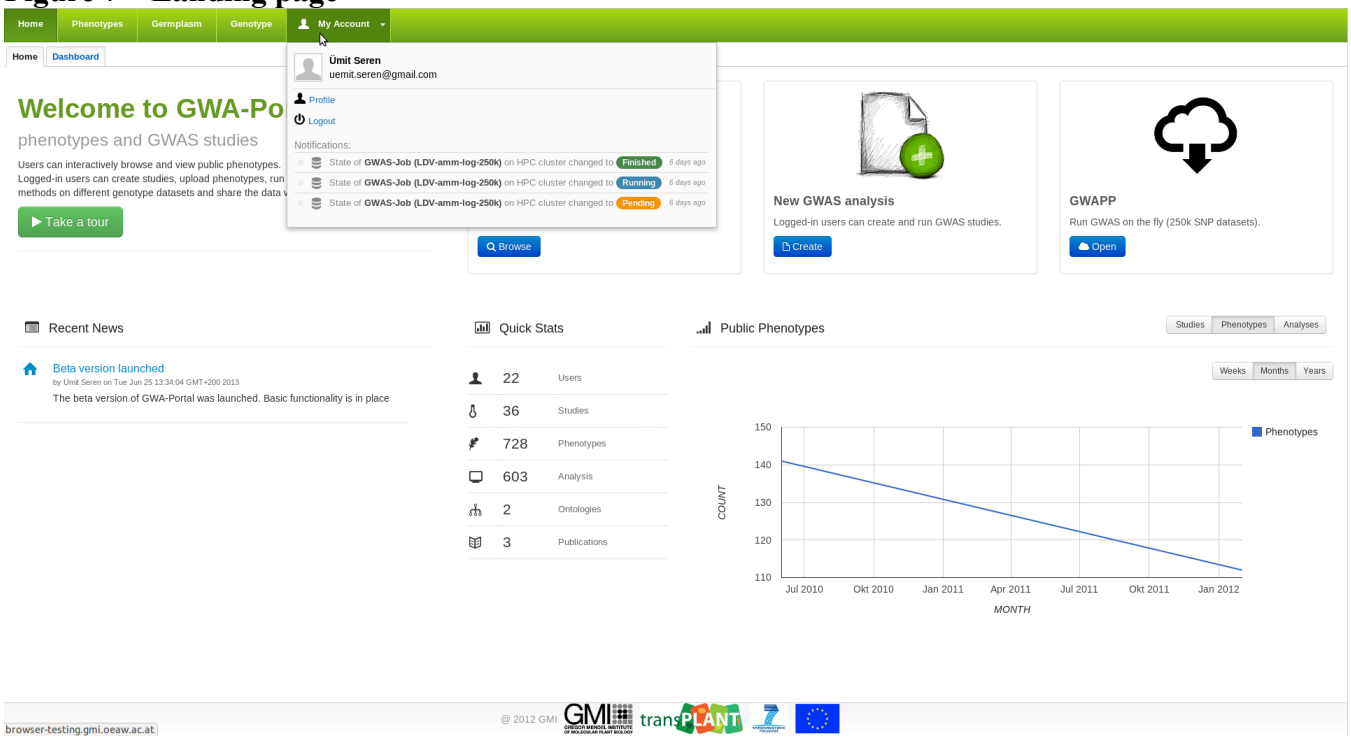
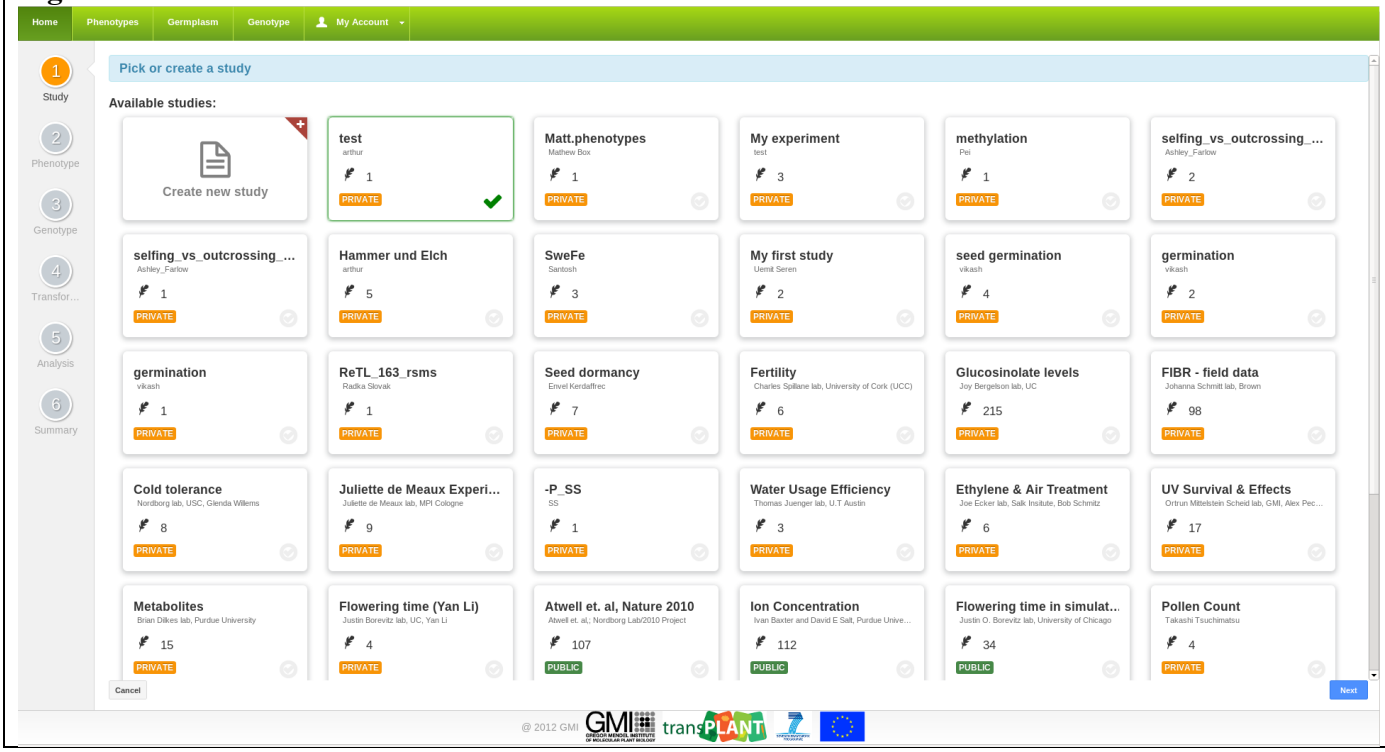
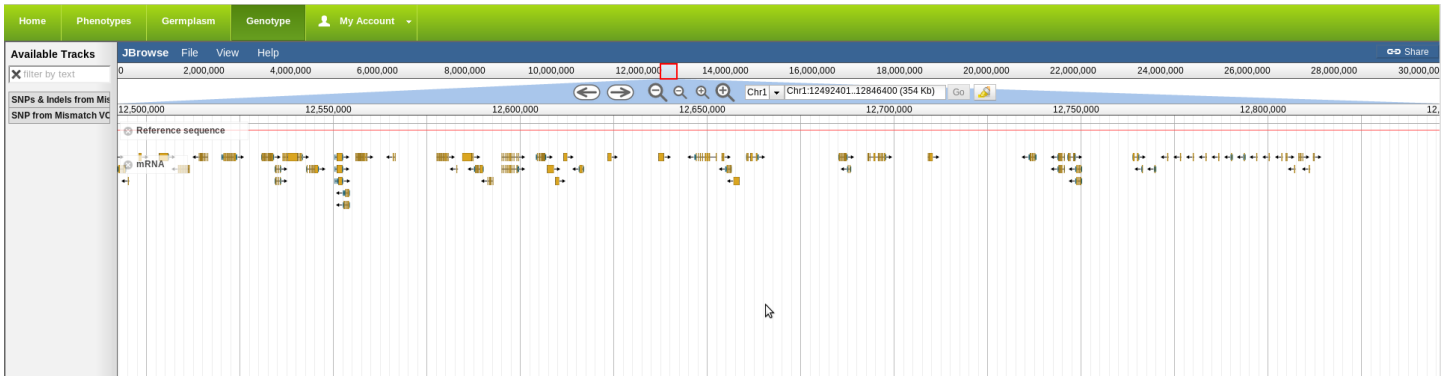


Figure 8 – Wizard to create new data



Genotype section:
The genotype section shows a genome browser (JBrowse) with some standard tracks.

Figure 9 – Genotype section



Germplasm section:
The Germplasm section contains general information about taxonomies, passports and stocks stored in the database. Users can view the available taxonomies (Figure 10), detailed information and statistics (Figure 11). Furthermore, they can view all the passports (Figure 12), passport information and stock information (Figure 13).

Figure 10 – Available taxonomies

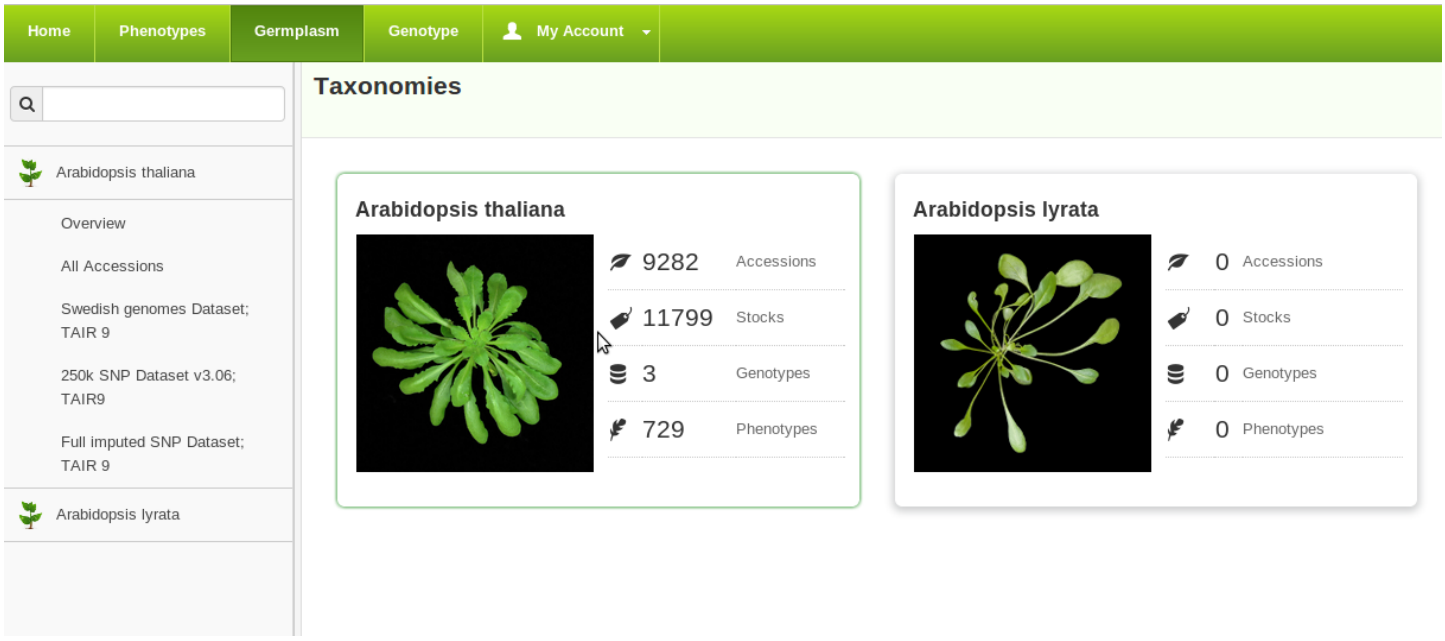


Figure 11 – Taxonomy information

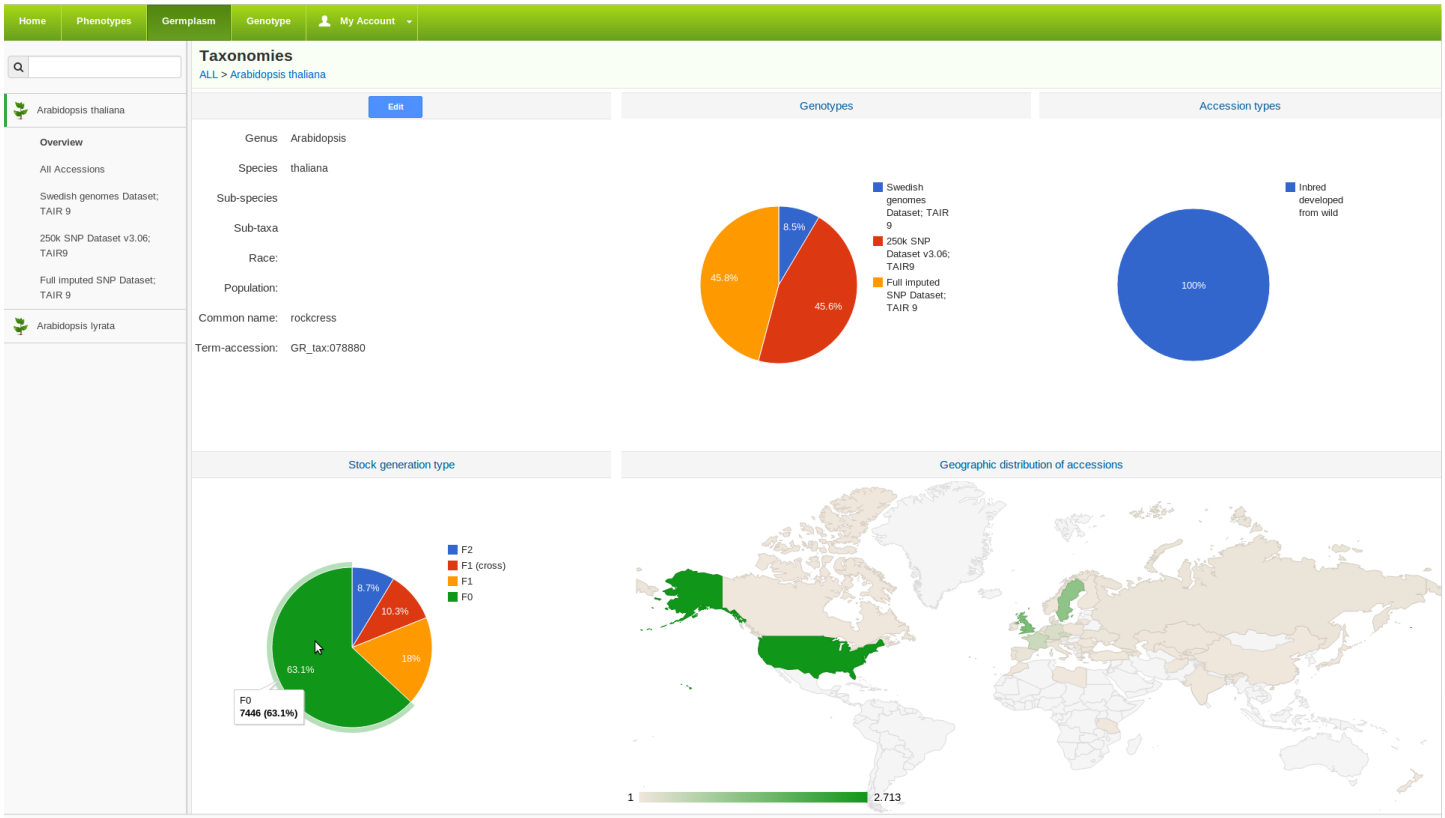


Figure 12 – List of passports

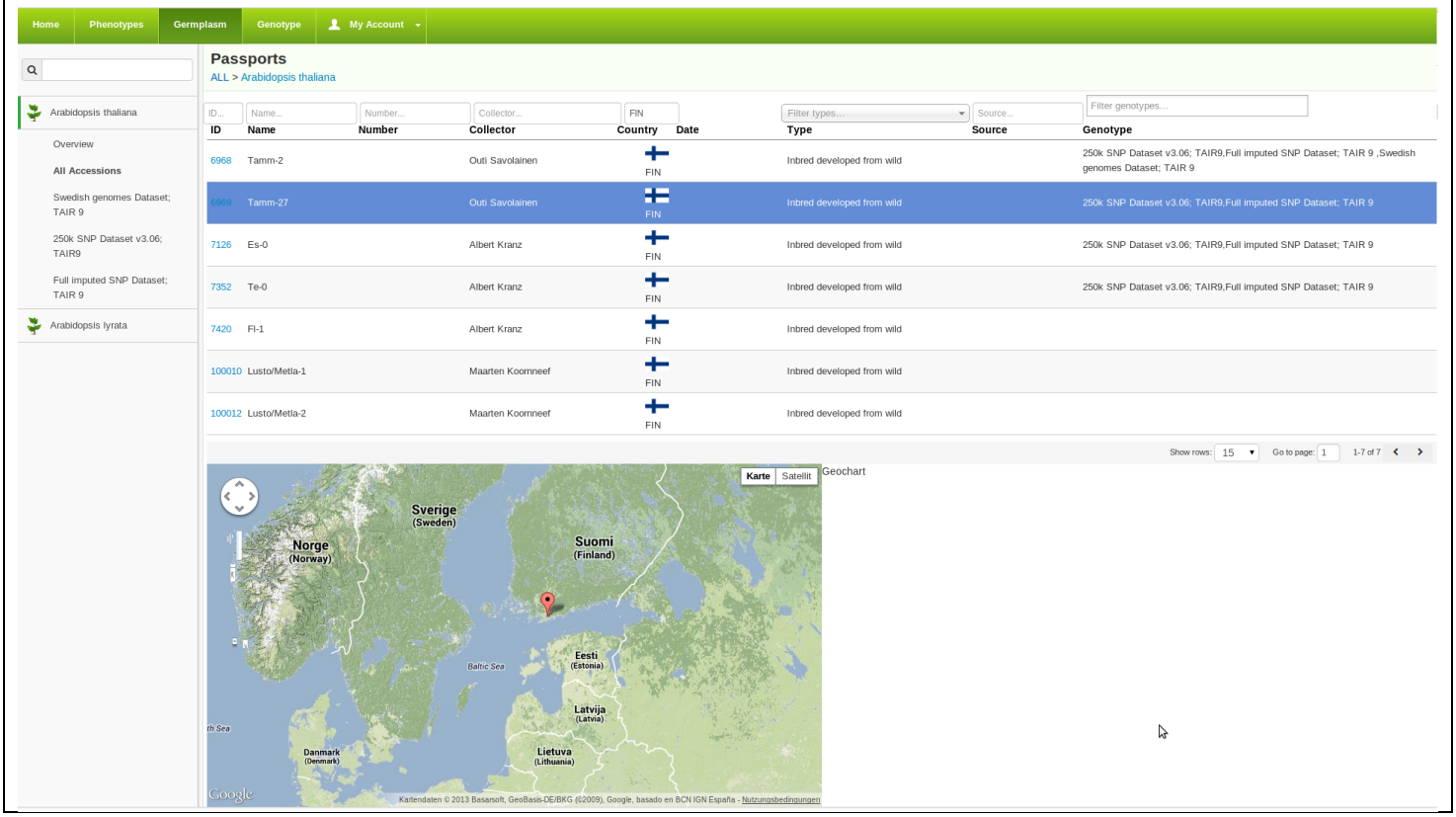


Figure 13 – Passport information

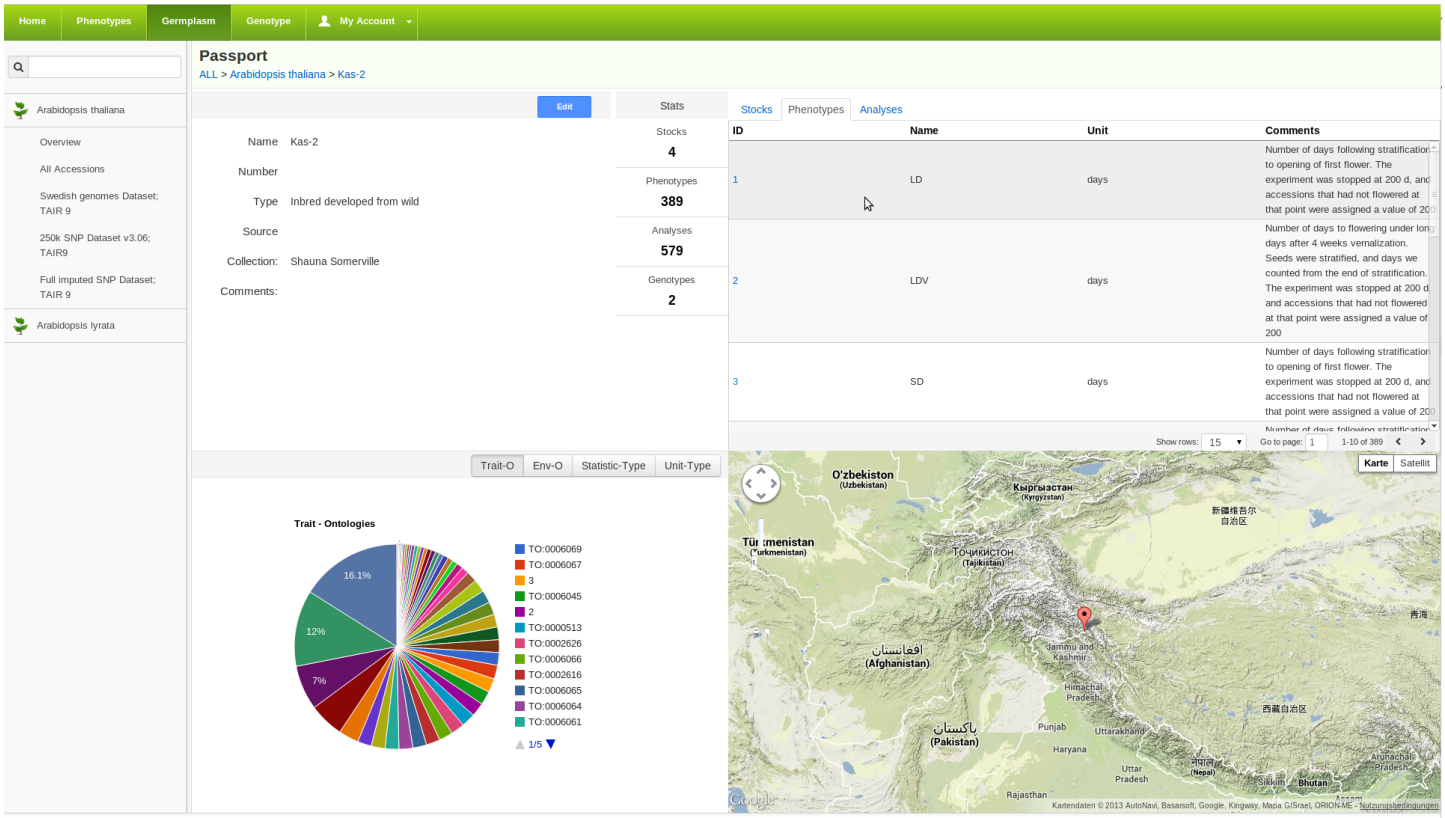
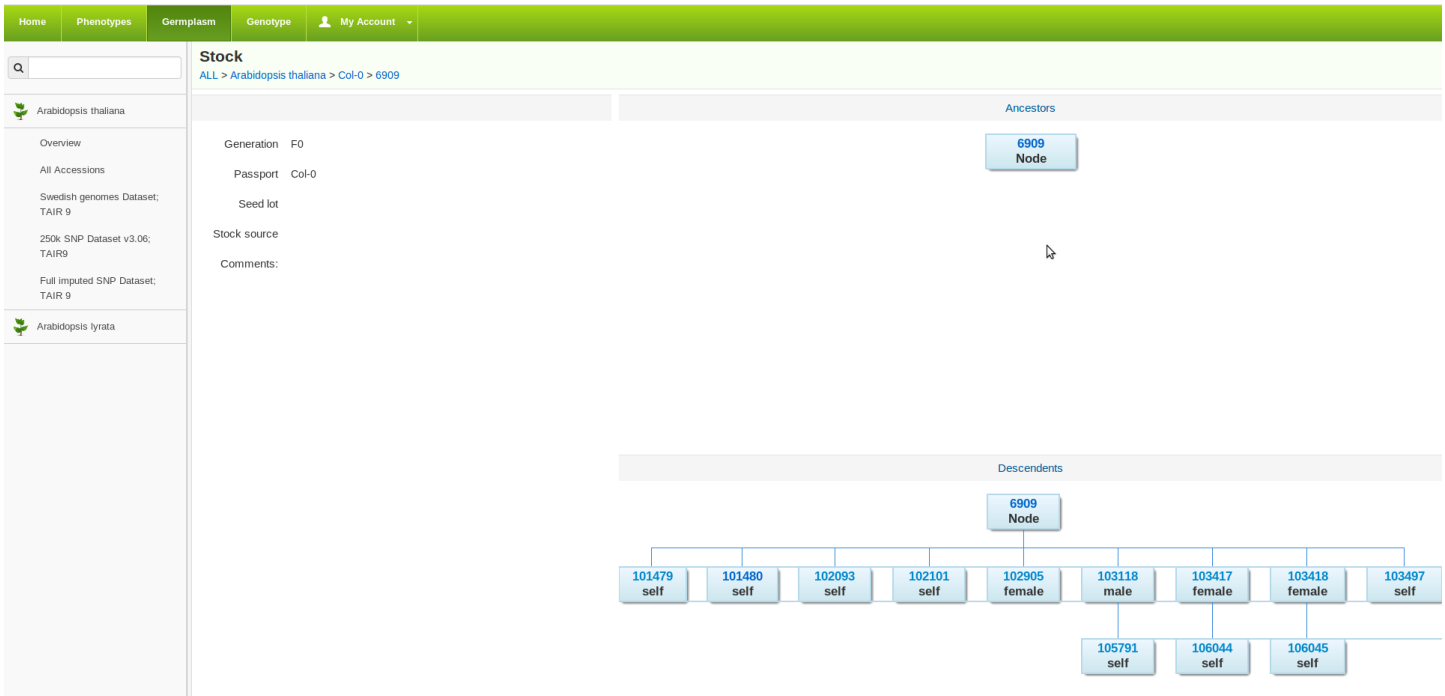


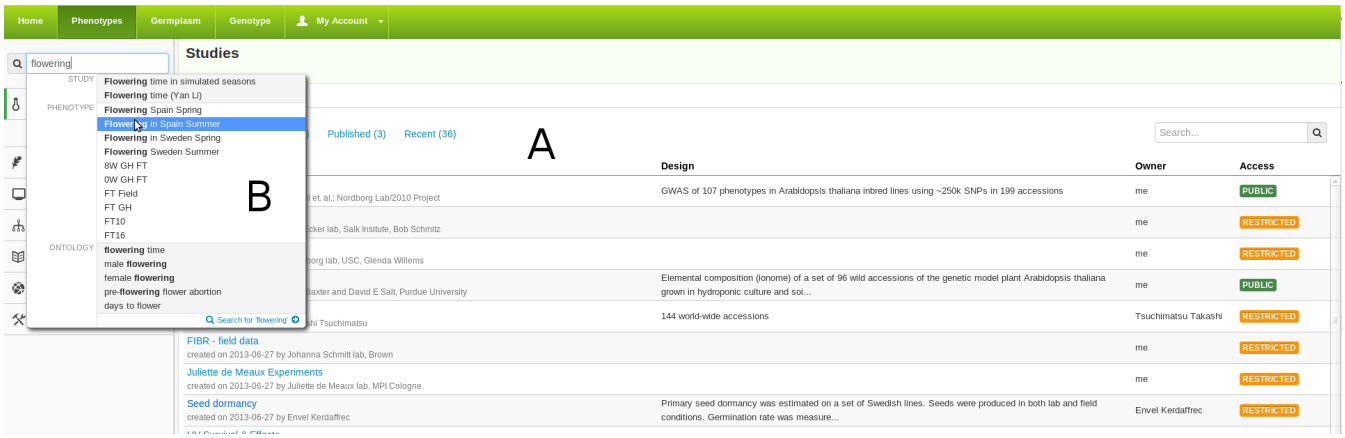
Figure 14 – Stock information



Phenotype section:

The phenotype section contains information about phenotypes, studies and analyses. The data is structured in a hierarchical way (study > phenotype > analysis). There are different ways to navigate through this section (Figure 15): A.) Users can browse and filter/search entity-lists; or B) use the global searchbox to search across all categories.

Figure 15 – A) use lists to browse/search corresponding category or B) search in global searchbox.



For each entity (phenotype, study, analysis) there is a detailed view available that shows additional information (Figure 16). Permissions can be set for each of these entities (Figure 17).

Figure 16 – Detailed phenotype information

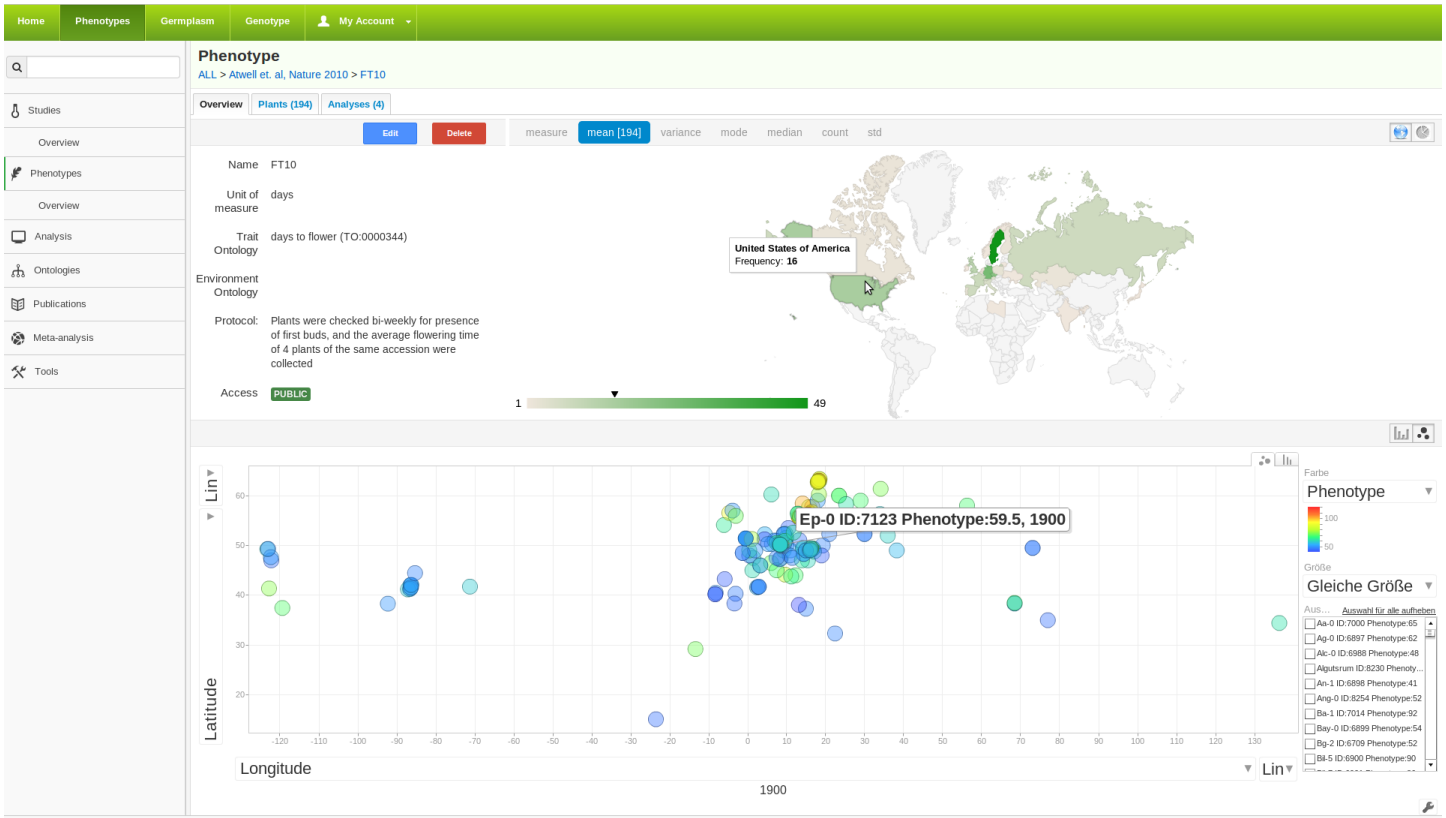
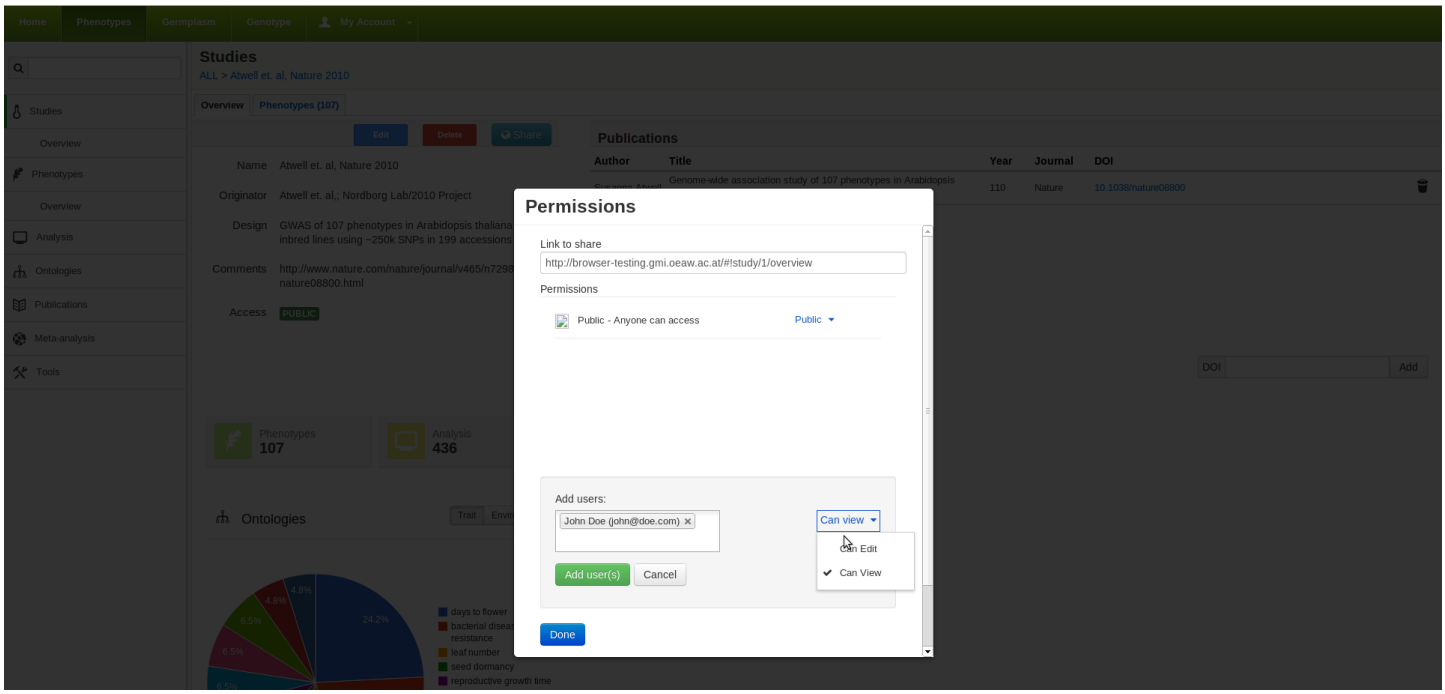
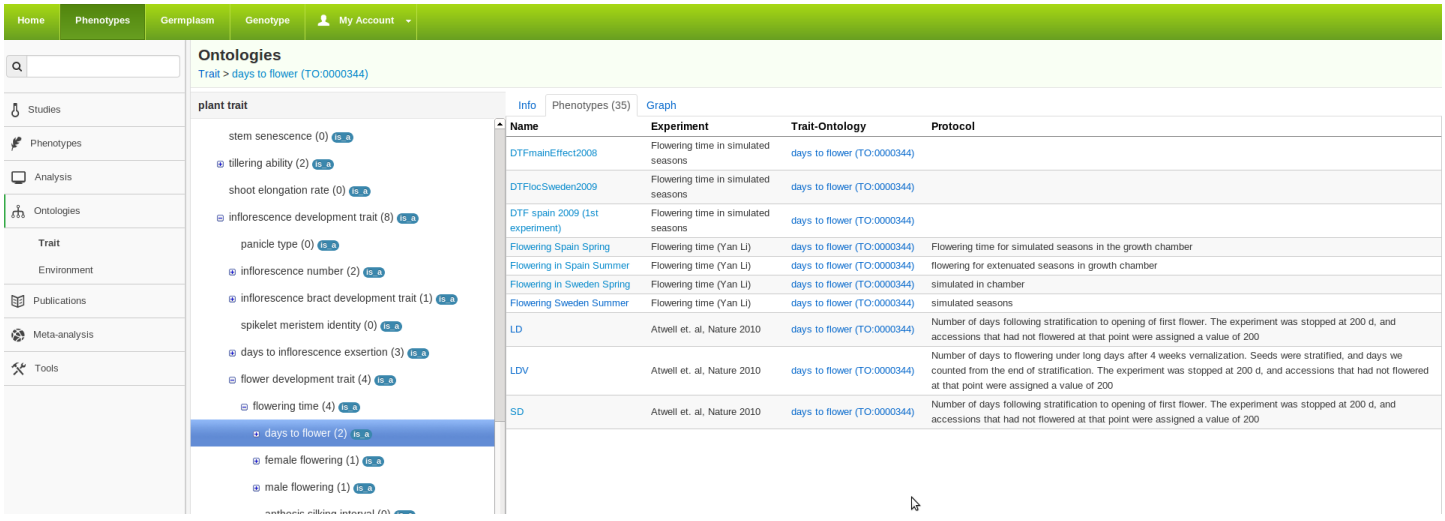


Figure 17 – Permissions



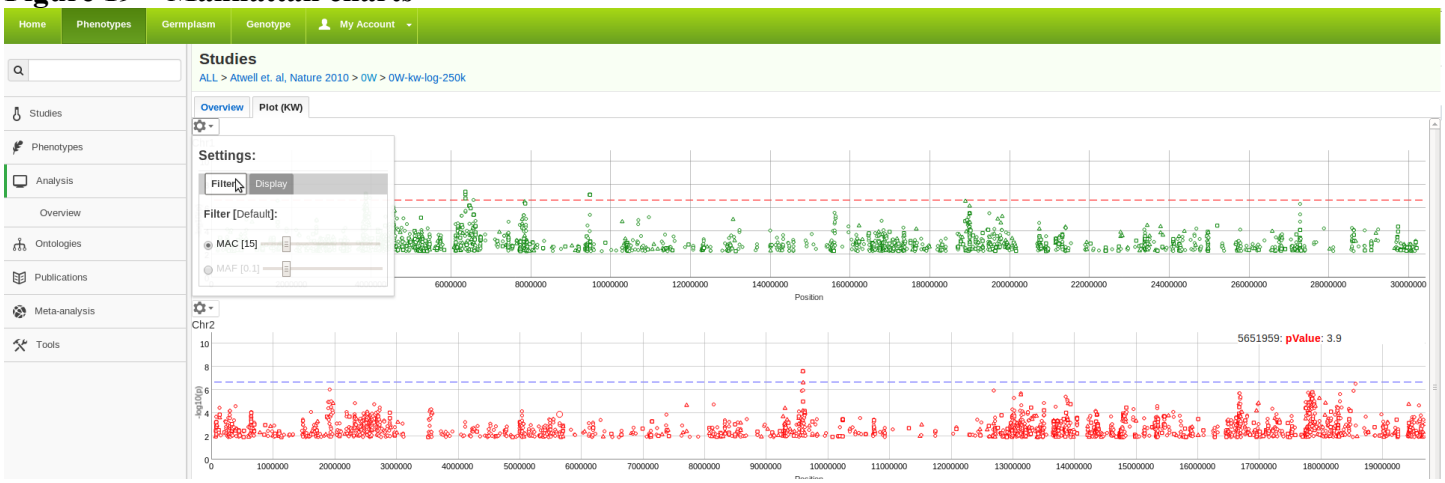
The phenotype section also contains information about ontologies (trait & environment) and annotated phenotypes (Figure 18). The ontologies where imported from Gramene (http://gramene.org).

Figure 18 – Ontology view



GWAS results are displayed similar to GWAPP with interactive Manhattan charts (Figure 19). These interactive Manhattan charts now support filtering based on MAC/MAF and displaying SNP annotation using the shape of datapoint (i.e. triangle, rectangle, etc).

Figure 19 – Manhattan charts



The meta-analysis section allows the user to view all associations in a specific genomic region (Figure 20) or show the top associations across all available information (Figure 21).

Figure 20 – Association in genomic region

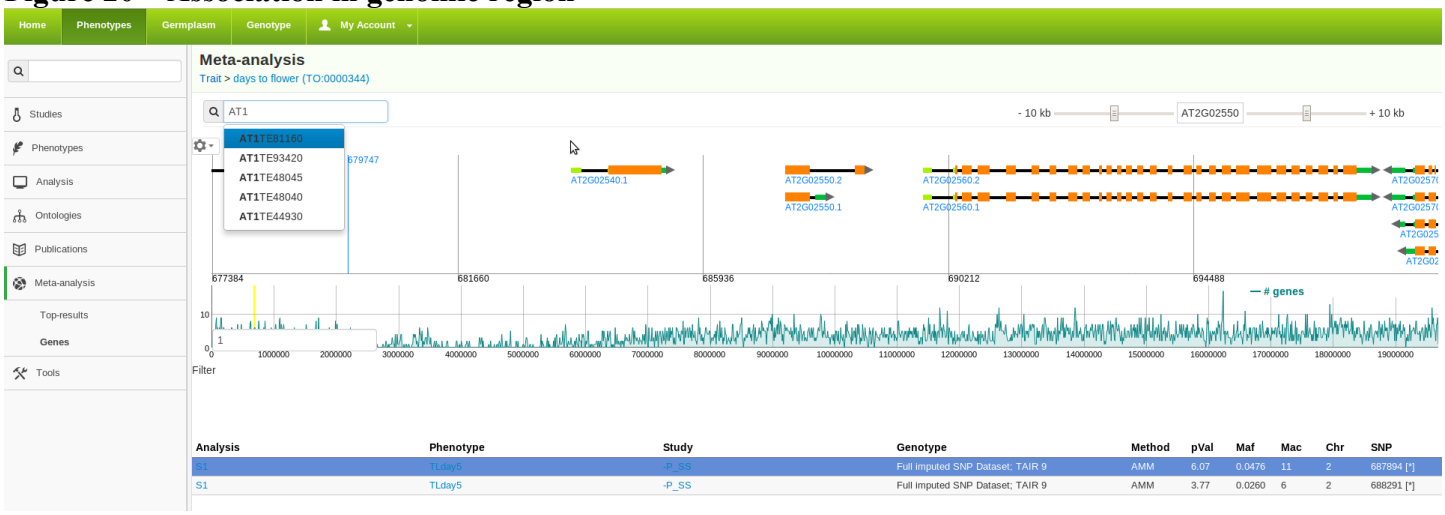
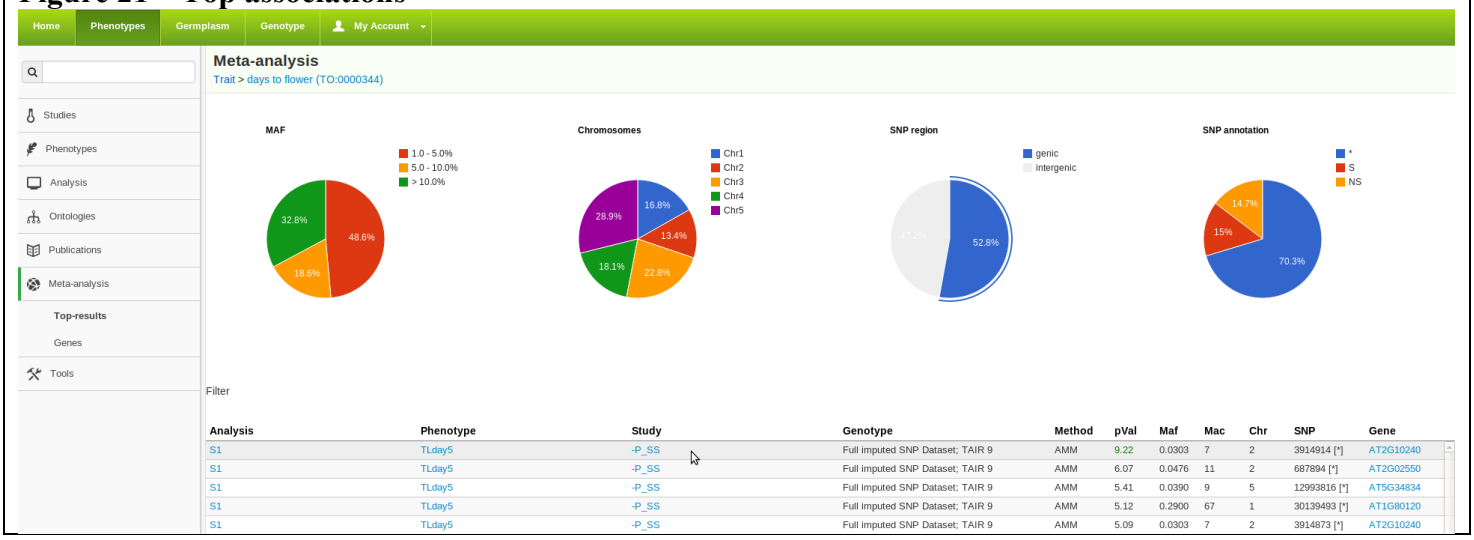


Figure 21 – Top associations



GWA-Portal is open source (https://github.com/timeu/geno-phen-browser) and public access will be soon available using following URL: http://gwas.gmi.oeaw.ac.at

Interaction with other work-packages:
There are a some areas in GWA-Portal that allow for interaction and integration with tasks and deliverables of other work-packages:

- Currently the format for uploading phenotypes is quite simple. However there are future plans to support the phenotype format and ontologies that are being developed in WP 3 (D3.2).
- The GWAS analyses are put on a distributed task queue and later processed by our HPC. Because of the modular setup it should be quite easy to incorporate cloud computing as they are developed in the course of WP5.
- Public data (phenotypes, studies and analyses) can be made available to the information retrieval system (WP11) and integrated search of the transplant portal (WP6).

Publications

Seren, et. al. (2012) GWAPP: A Web Application for Genome-Wide Association Mapping in Arabidopsis thaliana. The Plant Cell. vol. 24 (12): 4793-4805.