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

D5.1

Updated data warehouses developed for genomic annotation and variation data

Due date of deliverable: 31.8.2013 (M24)

Actual submission date: 19.9.2013

Start date of project: 1.9.2011 Duration: 48 months

Organisation name of lead contractor for this deliverable: INRA

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
INRA, EBI

Introduction

Deliverable reference number: D5.1

The objective of this deliverable (WP5, task 4) is to provide user communities with optimized data mining tools through the deployment of data warehousing technology.

To ensure the readiness and the usability of the technology described in this report, the delivery date was shifted from 15.11.2012 to 31.8.2013 with the agreement of the project officer.

Methods

BioMart is a query-oriented data management system widely used in bioinformatics (http://www.biomart.org). It provides a set of tools for the easy construction of de normalized databases and programmatic interfaces focused on common queries. It is supported by a community of developers and users that use this tool within the biomedical domain. It supports data federation allowing the performance of distributed queries between different marts. The current version is 0.7 but a new version 0.8 has been available since 2010. This new version offers a list of advantages, and has already been deployed in the field of Cancer Genomics (http://dcc.icgc.org/web/)

The provisional working plan was to implement data warehouses and to provide public access to them. transPLANT partners INRA URGI and EBI were already deploying Biomart version 0.7 as a tool to provide access to some of the data available in their resources. The plan was to provide updated warehouses based on the improved infrastructure of the new version of BioMart (0.8).

Results

Update of datasets in BioMart 0.7:

- EBI provides two separate BioMart data schemas, one for gene-centric queries (the gene mart) and one for variant centric queries (the SNP mart). Each contains a number of data sets for query (one per species). EBI have made 9 releases of Ensembl Genomes since the start of the transPLANT project and updated BioMarts are provided with each release; the data, and where appropriate, the schemas, are updated with each release. The latest release (release 19, August 2013) contained 25 gene-centric data sets, of which 16 have been newly provided since the initiation of transPLANT funding, and 10 variant-centric data sets, of which 4 are new since

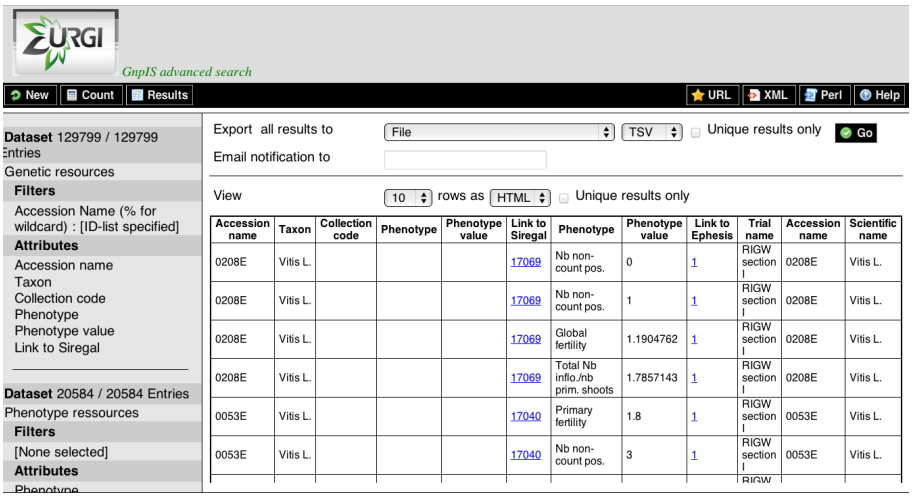
the commencement of transPLANT. Data is available through the Ensembl Plants portal http://plants.ensembl.org.

- INRA partner updated its GnpIS biomart plants datasets. Data are available through its portal: http://urgi.versailles.inra.fr/biomart/martview/

In summary, INRA created a new dataset to explore maize ZmB73 V2 structural and functional genome annotation data. It built a new dataset to explore TAIR V10 genome annotation (Arabidopsis). It improved also its existing datasets that explore variants (SNPs), genomic annotations for GrapeVine species 12X and 8X and for Wheat species. These new tools allow by setting some specific filters, for example to retrieve a list of 'features' that could be genes or SNPs with their genomic coordinates on a chromosome or a set of chromosomes and to have also links to more detailed information contained in GnpIS INRA information system by links based on URLs.

As major event, it created a new dataset dedicated to the exploration of genetic resources (germplasm) and another one to explore phenotypic trials. These two new datasets are also linked together and allow users to do queries that explore the two datasets at the same time based on commons objects (ie accessions in this case). It is indeed for example possible to get all phenotypic values for a set of germplasms and to obtain in a second step, the full description of the accession and the full description of the trial, both contained in GnpIS information system. The **figure 1** below, shows for example the phenotypic results found for the query made on germplasms.

INRA released several versions of GnpIS since the beginning of transPLANT project, 3 in 2012 and 2 in 2013. The new marts developed in the frame of transPLANT are released also in GnpIS information system, they are also online on INRA web site. The full information system was recently published, D. Steinbach & al. Database, Vol. 2013, Article ID bat058, doi:10.1093/database/bat058.

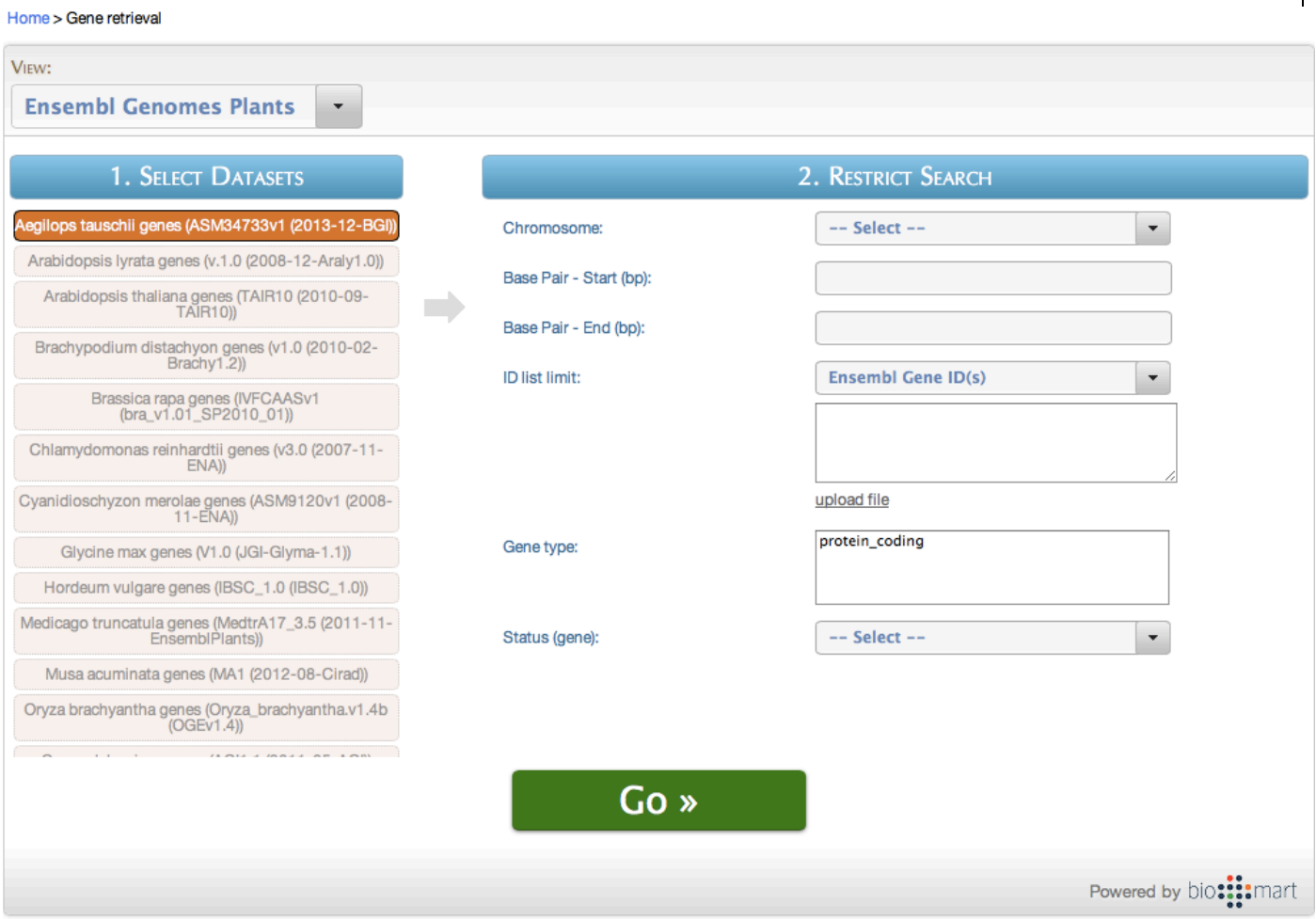


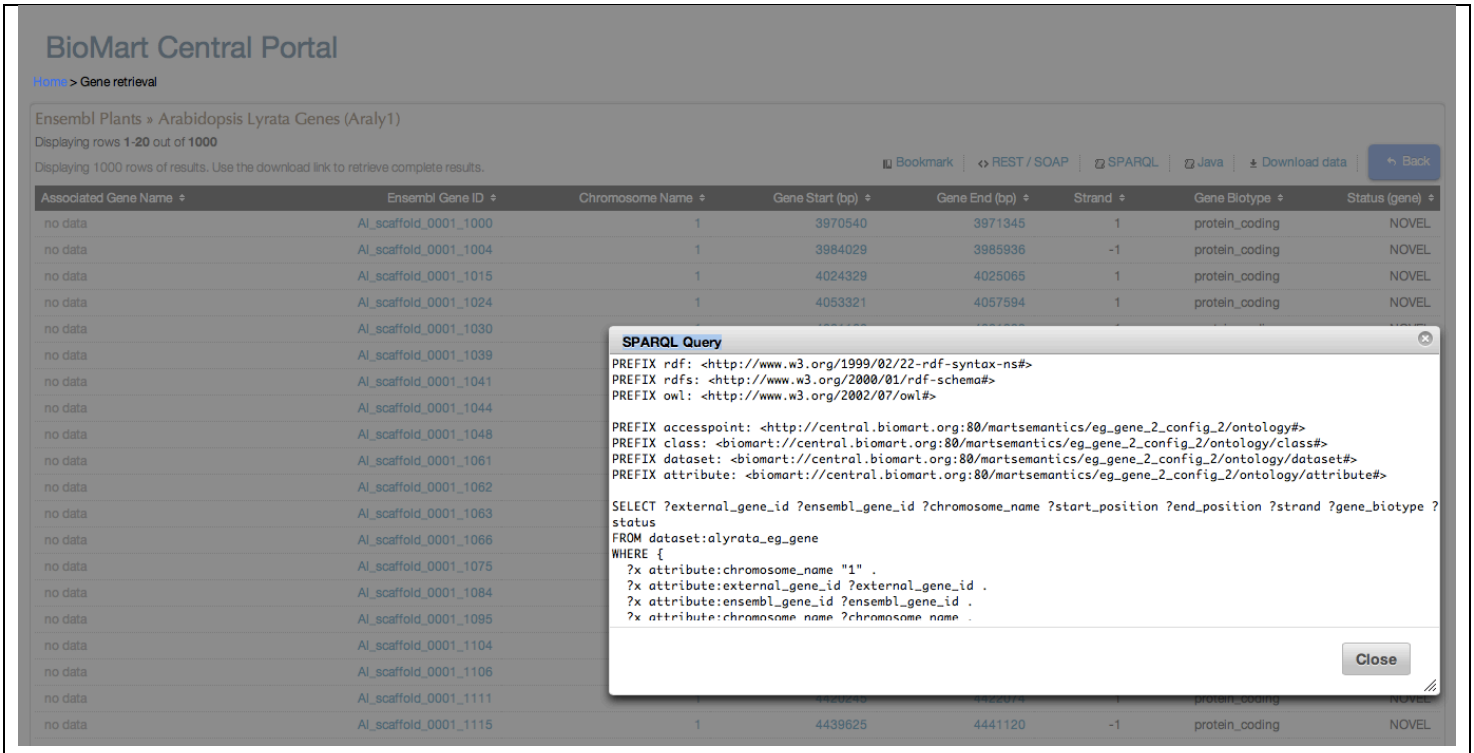
Accession name	Taxon	Collection code	Phenotype	Phenotype value	Link to Siregal	Phenotype	Phenotype value	Link to Ephesis	Trial name	Accession name	Scientific name
0208E	Vitis L.				17069	Nb non-count pos.	0	1	RIGW section I	0208E	Vitis L.
0208E	Vitis L.				17069	Nb non-count pos.	1	1	RIGW section I	0208E	Vitis L.
0208E	Vitis L.				17069	Global fertility	1.1904762	1	RIGW section I	0208E	Vitis L.
0208E	Vitis L.				17069	Total Nb inflo./nb prim. shoots	1.7857143	1	RIGW section I	0208E	Vitis L.
0053E	Vitis L.				17040	Primary fertility	1.8	1	RIGW section I	0053E	Vitis L.
0053E	Vitis L.				17040	Nb non-count pos.	3	1	RIGW section I	0053E	Vitis L.

Data Warehouses For Plant Data in BioMart 0.8:

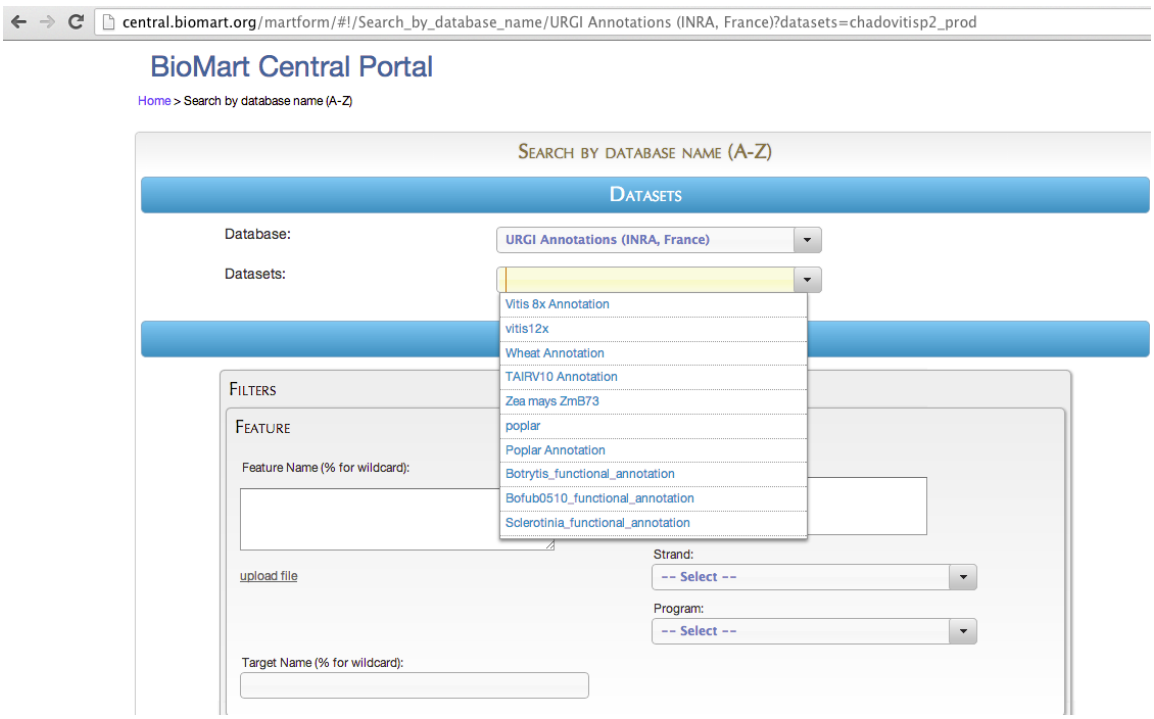
The Ensembl Plants BioMarts can now be accessed in the BioMart 0.8 user interface through the BioMart central portal (http://central.biomart.org/). The interface provides more powerful features to sort data, and provides Java and SPARQL endpoints in addition to the support already offered for Perl, SOAP and REST in BioMart 0.7. Updated databases are supplied to the operators of the BioMart Central portal and made available with each release of Ensembl Genomes.

Figure 2: A. Selecting databases and B. visualising results from the Ensembl Plans BioMarts in the BioMart central portal. In the second image, the user has performed a search for genes in *Arabidopsis lyrata*, and selected an option to generate code to retrieve the data set generated using the SPARQL query language.





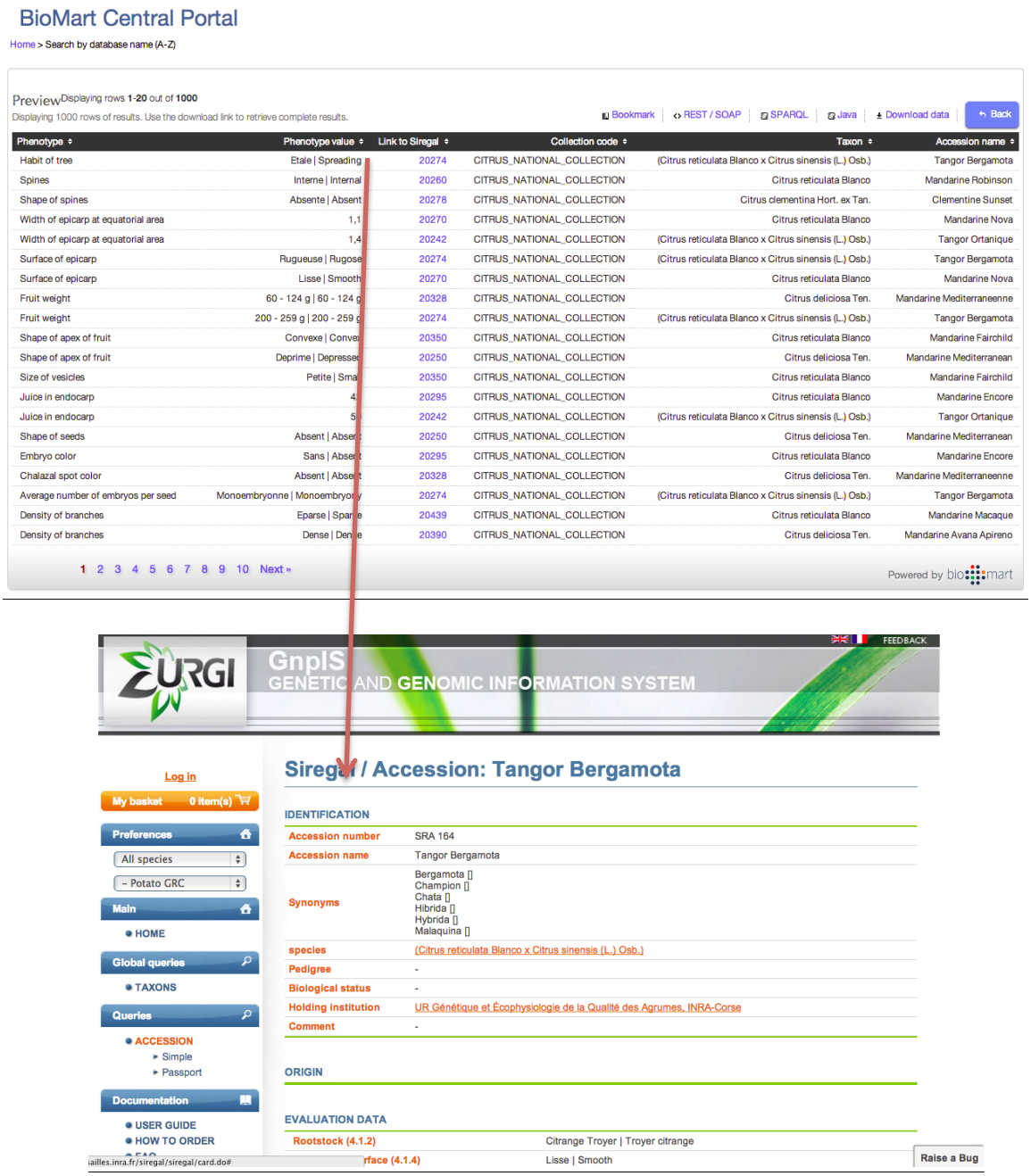
In august 2013, INRA collaborated with Biomart central team to make available through the BioMart central portal its databases and its corresponding datasets. As result, the genome annotation databases and its 10 datasets are not online: See **Figure 3** below:



The second INRA biomart database was also added. It is more focussed on genetic data and it contains 4 datasets dedicated to genetics maps, one on NGS variant, one on genetic resources and one on phenotypic results. For example we can retrieve from this server and from the phenotypic dataset, all phenotypic values from a list of genetic resources, here citrus accessions and to switch to GnpIS information system for getting

more information in the global information system. Results are also available in a format exportable in table sheets, in SPARQL or in JAVA.

Figure 4: the 2 screenshots below, show all phenotypic evaluation results retrieved from a query done on biomart central server on a citrus genetic resources collection (3A) and gives after all information retrieved from GnpIS for the selected germplasm (3B)
http://urgi.versailles.inra.fr/siregal/siregal/card.do?className=genres.accession.AccessionImpl&dbName=common&id=20274



Options for the Longer Term : Local Installations of BioMart 0.8

Before a more comprehensive commitment was made to local deployment of the new BioMart software, an

analysis was made of its suitability for further developments. The new version of the software was installed at INRA and migration of all existing version 0.7 datasets to version 0.8 was attempted. Unfortunately, several bugs were found. While some of the existing datasets could be migrated, for example, the (relatively simple) genomic annotation datasets for Arabidopsis and maize, more complex datasets (in particular those making interoperability between genomics and genetics i.e. genetic markers, genomic annotations and genetic resources) failed to migrate. These root causes was diagnosed due to new constraint that appeared in version 0.8 and which had not been present in version 0.7. It was indeed not possible to migrate both 0.7 database and the 0.7 query system into 0.8. A solution could be to keep 0.7 and 0.8 databases on line and to have a unique query interface in 0.8 that is plugged on the two types of databases. This solution is indeed heavier to maintain. Additionally, the rate of recent BioMart development has been slow.

We therefore decided to evaluate an alternative platform to fulfil the future needs for warehousing plant data.

Evaluaton of InterMine :

InterMine is an open source data warehouse built specifically for the integration and the analysis of complex biological resource. It is developed by Micklem laboratory at the University of Cambridge. As with BioMart, the system supports the generation of data warehouses and associated tools for web-based query tools. Parsers are provided for many common biological data sources and formats, and automatically integrate data within a relational database. Queries are optimised by the generation of pre-computed tables. The tool was originally developed to meet the needs of model organism databases and the current parsers and query structures are focussed mainly on genes. However, the tool provides a framework for supporting additional data types. The query interface can also be customised for specific needs and an API is available to allow programmatic access to data.

InterMine was installed locally at INRA.

Datatypes :

An existing genomic dataset (already available through BioMart datasets) was entered into InterMine, comprising grapevine genomic structural and functional annotation. Further datasets were extracted from the GnpIS data warehouse (used internally at INRA) to test the capacity of the system to accommodate further data types that are not natively supported in intermine tool. Data was converted into GFF3 standard format as a solutin for import into InterMine. INRA succeeded in loading indeed genetic marker data (with their position in cM), QTLs and SNP markers into the same instance of InterMine already holding the structural and functional annotation. INRA continued the integration work by adding also genetic resources data (passport data) and phenotyping data (trial data).

External links and data federation :

INRA also tested the set up of links between this mine and external tools. It was able:

- To construct links with the GnpIS URGI portal to have access to more detailed information not contained in the mine and also with Gbrowse/Gmod tool.

- To construct links to external mines. INRA tested the functionality with for example a link with FlyMine tool (even if there no plant data in FlyMine) and the link with another local mine at INRA was also tested.

- To link to BioMart datasets contained in GnpIS, but also accessible at EBI Ensembl Plants. For that we chose an example present in both information system, based on GrapeVine (12X) data.

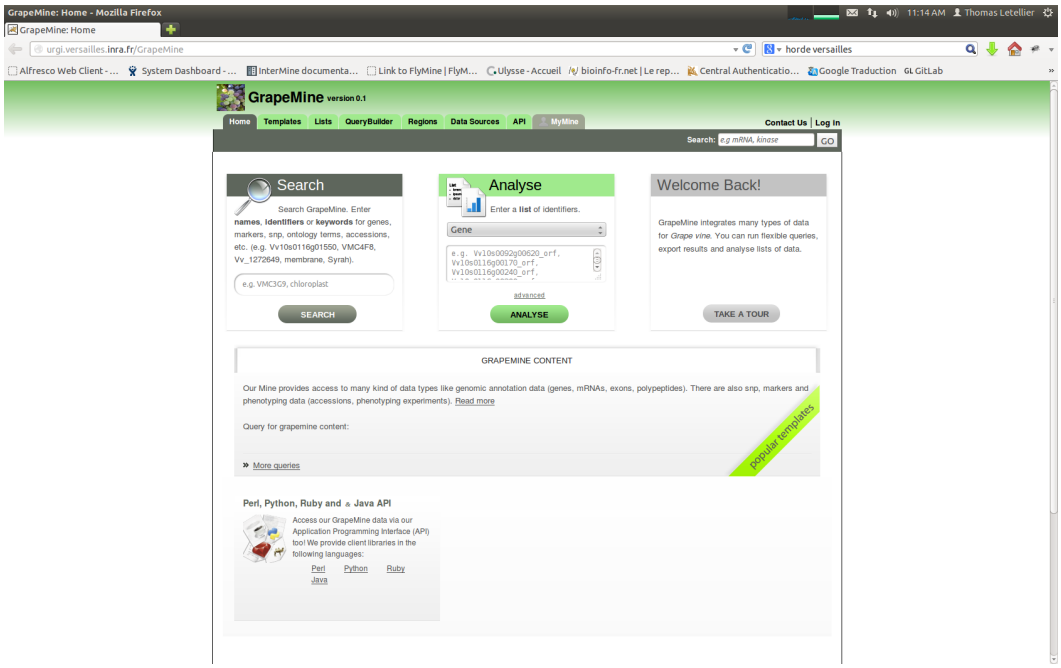
Query interface:

Modifications were necessary to be done to draw the object types in the InterMine query interface. It required to make changes in Java parsers and also XML edition because intermine is not adapted for those new datatypes.

Data is now accessible through the QuickSearch and Query Builders tool and through the use of the Region tab.

To query the tool: http://urgi.versailles.inra.fr/grapemine.

See the **figure 5** below:



This tool was presented for getting feedbacks at INRA and to the coordinator of structural grape genomics. They found the tool very useful for grape communities and encouraged us to continue the work and to build also other marts on other species.

Perspectives:

We intend in the next future to build a new mart dedicated to wheat genomics and genetics.

New approaches for accessing large variation data sets

Variation data can be very large, comprising polymorphisms from across entire genomes from thousands of individuals, stocks or populations. Common use cases including searching for variants near particular genes, associated with particular phenotypes, present/absent in particular individuals, or variant loci with different values at particular loci. Variant data stored in Ensembl Plants is currently available through the variation warehouses produced in the BioMart system. These, however, can be slow to construct, can freeze when confronted with large or poorly formed queries, and can only support a limited number of query types. By way of explanation, the current variation BioMart for *Arabidopsis thaliana* contains 14 million variant features from 1609 individuals, and that for maize contains 51 million variant features for 103 individuals, giving over 100

million data points each. By contrast, the gene-centric BioMarts for both species are centred on gene sets containing 34 thousand and 110 thousand genes respectively. Moreover, much additional variation data is expected through the development of the transPLANT variation archive in work package 9. Because variation warehouses are currently the biggest cause for concern, we set about prototyping a new warehouse for querying variation data.

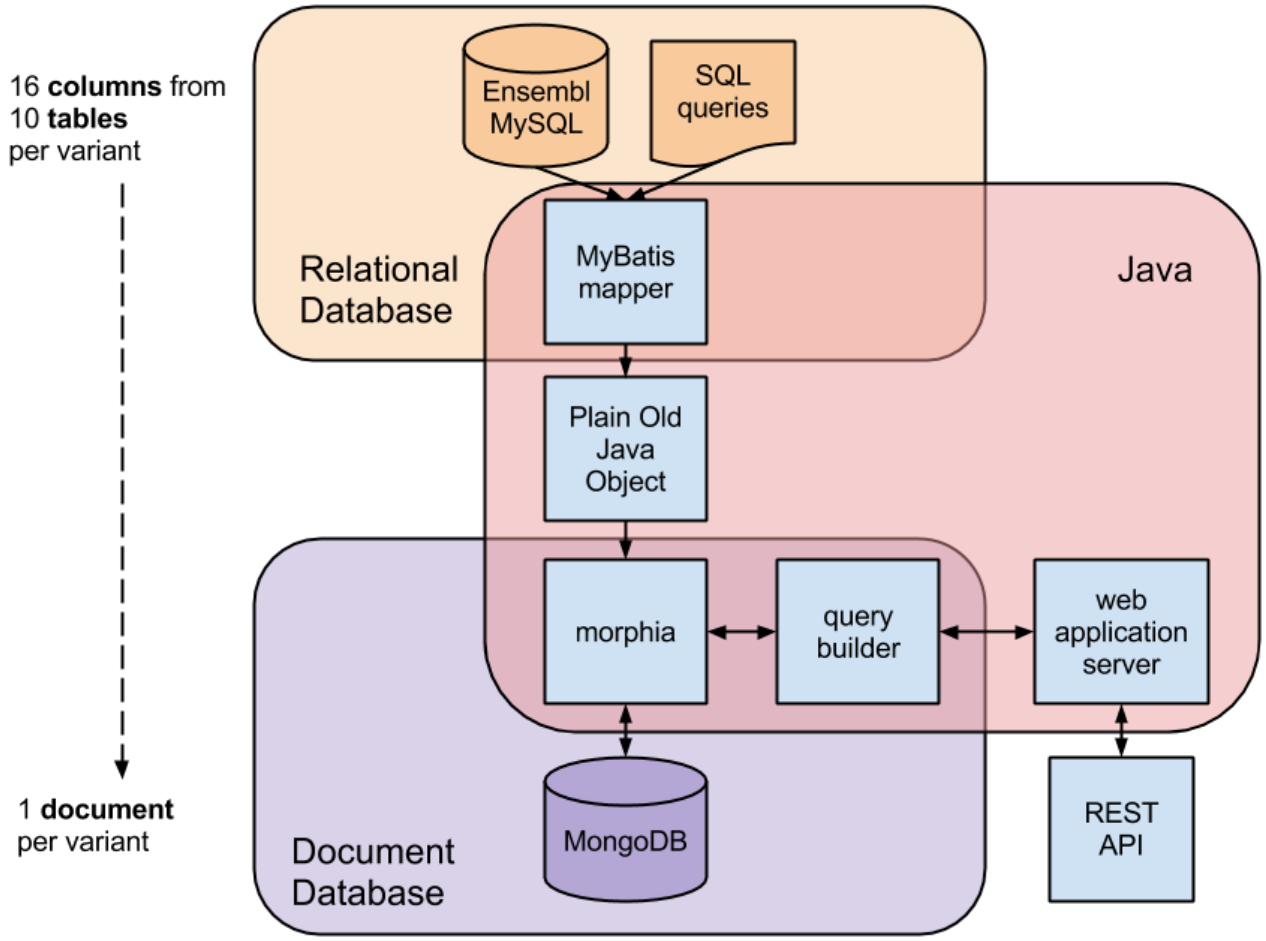
After discussion with users and collaborators, a list of common use cases was developed for the variation data warehouse. A sample of these is listed in table 1.

Table 1. A selection of use cases for the variation data warehouse. Essentially, the queries envisage a matrix of variant alleles against sequenced individuals; additionally, both individuals and variant loci may themselves be linked to query able metadata (gene names/functions, geographical location, plant phenotype or gene-phenotype association, etc.)

- Find me variants:
  - in a particular specified coordinate range
  - within a defined distance range of a gene on the supplied list
  - in a given sequence (primer checking etc.)
  - present in a given strain
  - common (locus) among two strains
  - different between two strains
  - for a phenotype with a given P-value
- Find me genomes
  - with these variant alleles

An initial warehouse has been implemented in MongoDB, a noSQL database backend also being used for the development of the variation archive (see work package 9). MongoDB is fast and scalable (promising linear performance with increased data size providing a linear increase in compute resource is available). Data from Ensembl MySQL databases is loaded into Java objects using mybatis (https://code.google.com/p/mybatis/), a persistence framework for custom SQL, and stored in MongoDB using Morphia, (https://github.com/mongodb/morphia) a type-safe Java library for accessing MongoDB. The overall infrastructure is specified in figure 1.

Figure 6. The infrastructure stack supporting the new variation data warehouse.



Results

A prototype warehouse was constructed for variation data from A. thaliana. This data set comprises 350 million alleles from 1610 individuals and has a total size (when stored in a standard Ensembl relational database schema) of 18.5 GB. Table 1 shows a comparison of build and query time with the existing use of BioMart. Queries were run on 5 shard servers on the EBI cluster (8GB memory, data stored on NFS).

	BioMart	**MongoDB**
Build time	<= 1 week	7 hours
Disk space	18.5 GB	35 GB
Q1 – first ten results		
Q1 – full results	6.7 min	2.8 min
Q2 – first ten results	8.5 sec	0.1. sec
Q2 – all results	40 sec	0.8 sec

Q1: all variants that are common between 2 strains
Q2: all variants in the ranges of 20 genes

Next steps

A basic user interface has been provided at http://tertic.transplantdb.eu to enable test usage and to obtain feedback from collaborators. The longer term plan is to embed a new UI within the Ensembl Plants site for bulk data retrieval and download.

Summary

New data warehouses have been made containing the latest data from Ensembl Plants (9 data releases) and INRA (12.4 data releases) in the BioMart data warehousing system, version 0.7.

Ensembl Plants data can be queried using the BioMart 0.7 user interface at the BioMart central portal. INRA marts can be queried at the BioMart central portal using BioMart 0.8 for all INRA marts (those built in 0,7 and those 0,8 database schema) and can also be queried on INRA web portal by using Biomart 0.7.

BioMart version 0.8 has been evaluated.

Due to certain limitations with the BioMart 0.8 software, an alternative system, InterMine, has been tested by INRA. A datawarehouse for the grapevine has been developed using InterMine and made available to public.

A new warehousing system is in development at EBI for large variation data sets.

Future Directions

We will continue to develop InterMine and the new system for variation data mining over the remainder of the project. Regular updates of our BioMart 0.7 databases will be produced until replacement systems are in place and fully functional. For some data types, BioMart 0.7 may prove serviceable for the medium term.

One of the great advantages of BioMart was that it defined a standard for database interoperability, i.e. it is relatively easy for different instances of BioMart, located in different locations, to communicate with each other and to be used for inter-warehouse queries. Another useful feature is that it is compliant with Galaxy tool which is able to get Data from Biomart natively. In future, however, it seems less likely that a single back-end solution will work for all data sets due to differences in data type and volume. INRA and EBI are presently exploring the possibility of establishing a lightweight data structure, similar to that in development for the search facilities in the transPLANT site (WP6), to be recognised by newly developed warehouses to maintain this possibility for interoperability.

Publications

Delphine Steinbach, Michael Alaux, Joelle Amselem, Nathalie Choisne, Sophie Durand, Raphaêl Flores, Aminah-Olivia Keliet, Erik Kimmel, Nicolas Lapalu, Isabelle Luyten, Célia Michotey, Nacer Mohellibi, Cyril Pommier, Sébastien Reboux, Dorothée Valdenaire, Daphné Verdelet and Hadi Quesneville : **GnpIS: an information system to integrate genetic and genomic data from plants and fungi. Database, Vol. 2013**, Article ID bat058, doi:10.1093/database/bat058