



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	Х			
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.

	ed search											
New Count Results									👷 URL	🔁 XM	L 🛃 Perl	Help
Dataset 129799 / 129799 Entries Genetic resources	Export all Email noti	results fication	s to to	File			\$	TSV 🛟	🔲 Uniqi	ue resu	Its only	🦻 Go
Filters	View				rows as) – Unique	results only				
Accession Name (% for wildcard) : [ID-list specified]	Accession	Taxon	Collection	Phenotype	Phenotype	Link to	Phenotype	Phenotype	Link to	Trial	Accession	Scientific
Attributes	name		code		value	Sireyai	Nh non-	value	Chilesis	RIGW	name	name
Accession name	0208E	Vitis L.				<u>17069</u>	count pos.	0	1	section I	0208E	Vitis L.
Laxon Collection code Phenotype	0208E	Vitis L.				<u>17069</u>	Nb non- count pos.	1	1	RIGW section I	0208E	Vitis L.
Phenotype value Link to Siregal	0208E	Vitis L.				<u>17069</u>	Global fertility	1.1904762	1	RIGW section I	0208E	Vitis L.
Dataset 20584 / 20584 Entries	0208E	Vitis L.				<u>17069</u>	Total Nb inflo./nb prim. shoots	1.7857143	1	RIGW section	0208E	Vitis L.
Phenotype ressources	0053E	Vitis L.				<u>17040</u>	Primary fertility	1.8	1	RIGW section	0053E	Vitis L.
[None selected]	0053E	Vitis L.				<u>17040</u>	Nb non- count pos.	3	1	RIGW section	0053E	Vitis L.
Attributes										RIGW		<u> </u>

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 (<u>http://central.biomart.org/</u>). 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.

View:			
Ensembl Genomes Plants -			
I. SELECT DATASETS			2. RESTRICT SEARCH
Aegilops tauschii genes (ASM34733v1 (2013-12-BGI))		Chromosome:	Select 💌
Arabidopsis lyrata genes (v.1.0 (2008-12-Araly1.0))		Base Pair - Start (bp):	
Arabidopsis thaliana genes (TAIR10 (2010-09- TAIR10))		babbi an orar (op).	
Brachypodium distachyon genes (v1.0 (2010-02-	P	Base Pair - End (bp):	
Brachy1.2))		ID list limit:	Ensembl Gene ID(s)
Brassica rapa genes (IVFCAASv1 (bra_v1.01_SP2010_01))			
Chlamydomonas reinhardtii genes (v3.0 (2007-11- ENA))			
Cyanidioschyzon merolae genes (ASM9120v1 (2008- 11-ENA))			upload file
Glycine max genes (V1.0 (JGI-Glyma-1.1))		Gene type:	protein_coding
Hordeum vulgare genes (IBSC_1.0 (IBSC_1.0))			
Medicago truncatula genes (MedtrA17_3.5 (2011-11- EnsemblPlants))		Status (gene):	Select 👻
Musa acuminata genes (MA1 (2012-08-Cirad))			
Oryza brachyantha genes (Oryza_brachyantha.v1.4b (OGEv1.4))			
		Go »	
			Powered by bio
			,

SEVENTH FRAMEWOF



EVENTH FRAMEWORK

BioMart Central Portal

Ensembl Plants » Arabidopsis Lyr	ata Genes (Araly1)						
Displaying rows 1-20 out of 1000							
Displaying 1000 rows of results. Use the do	wnload link to retrieve complete results.		in E	Bookmark <> REST / SC	DAP 2 SPARQL	🛛 Java 🛛 🛓 Download	d data
Associated Gene Name +	Ensembl Gene ID +	Chromosome Name +	Gene Start (bp) ÷	Gene End (bp) 🗧	Strand +	Gene Biotype +	Status (gene)
	Al_scaffold_0001_1000				1	protein_coding	NOVE
	Al_scaffold_0001_1004				-1	protein_coding	NOVE
	Al_scaffold_0001_1015		4024329	4025065	1	protein_coding	NOVE
	Al_scaffold_0001_1024		4053321	4057594	1	protein_coding	NOVE
	Al_scaffold_0001_1030		1001100	1001000			6
	Al_scaffold_0001_1039	PREETX rdf: chttp	://www.w3.org/1999/02/	22-rdf-syntax-ns#>			
	Al_scaffold_0001_1041	PREFIX rdfs: <http: 01="" 2000="" rdf-schema#="" www.w3.org=""></http:>					
	Al_scaffold_0001_1044	PREFIX owl: <http: 07="" 2002="" owl#="" www.w3.org=""></http:>					
	Al_scaffold_0001_1048	PREFIX accesspoint	t: <http: central.bio<="" td=""><td>mart.org:80/martsema t.org:80/martsemanti</td><td>ntics/eg_gene_2</td><td>_config_2/ontology#></td><td>#~</td></http:>	mart.org:80/martsema t.org:80/martsemanti	ntics/eg_gene_2	_config_2/ontology#>	#~
	Al_scaffold_0001_1061	PREFIX dataset: <	biomart://central.biom	art.org:80/martseman	tics/eg_gene_2_	config_2/ontology/dat	aset#>
	Al_scaffold_0001_1062	PREFIX attribute:	<biomart: central.bi<="" td=""><td>omart.org:80/martsem</td><td>antics/eg_gene_</td><td>2_config_2/ontology/a</td><td>ttribute#></td></biomart:>	omart.org:80/martsem	antics/eg_gene_	2_config_2/ontology/a	ttribute#>
	Al_scaffold_0001_1063	SELECT ?external_s	gene_id ?ensembl_gene_	id ?chromosome_name	?start_position	?end_position ?stran	d ?gene_biotype
	Al_scaffold_0001_1066	FROM dataset:alyr	ata_eg_gene				
		WHERE {					
	Al_scaffold_0001_1075	?x attribute:ch	romosome_name "1" .				
	Al_scaffold_0001_1075 Al_scaffold_0001_1084	?x attribute:ch ?x attribute:ext	romosome_name "1" . ternal_gene_id ?extern sembl sene id ?ensembl	al_gene_id .			
	Al_scaffold_0001_1075 Al_scaffold_0001_1084 Al_scaffold_0001_1095	?x attribute:ch ?x attribute:ex ?x attribute:en ?x attribute:ch	romosome_name "1" . ternal_gene_id ?extern sembl_gene_id ?ensembl romosome_name ?chromos	al_gene_id . _gene_id . ome name .			
	Al_scaffold_0001_1075 Al_scaffold_0001_1084 Al_scaffold_0001_1095 Al_scaffold_0001_1104	?x attribute:ch ?x attribute:ex ?x attribute:ch ?x attribute:ch	romosome_name "1" . ternal_gene_id ?extern sembl_gene_id ?ensembl. romosome_name ?chromos	al_gene_id . _gene_id . ome name .			
	Al_scaffold_0001_1075 Al_scaffold_0001_1084 Al_scaffold_0001_1095 Al_scaffold_0001_1104 Al_scaffold_0001_1106	?x attribute:ch ?x attribute:ex ?x attribute:ch ?x attribute:ch	romosome_name "1" . ternal_gene_id ?extern sembl_gene_id ?ensembl, romosome name ?chromos	al_gene_id . _gene_id . ome name .			Close
	Al_scaffold_0001_1075 Al_scaffold_0001_1084 Al_scaffold_0001_1095 Al_scaffold_0001_1104 Al_scaffold_0001_1106 Al_scaffold_0001_1111	?x attribute:ch ?x attribute:en ?x attribute:en ?x attribute:ch	romosome_name "1" . ternal_gene_id ?extern sembl_gene_id ?ensembl romosome name ?chromoso	al_gene_id . _gene_id . ome name . 1922074		protein_cooling	Close

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:

→ C 🗋 centr	ral.biomart.org/martform/#!/Search_by_dat	tabase_name/URGI Annotations (INRA, France)?datasets=chadovitisp2_prod
Но	BioMart Central Portal me > Search by database name (A-Z)	
		Search by database name (A-Z)
		Datasets
	Database:	URGI Annotations (INRA, France)
	Datasets:	Vitis 8x Annotation
		vttist2x What Annotation
	Filters	TAIRV10 Annotation Zea mays 2mB73
	Feature	poplar Poplar Annotation
	Feature Name (% for wildcard):	Botrytis_functional_annotation
		Scierotinia_functional_annotation
	upload file	Strand:
		Program: Select
	Target Name (% for wildcard):	

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=com mon&id=20274

BioMart Central Portal

Preview Displaying rows 1-20 out of 1	1000		n Rookm	ark OREST/SOAP SPAROL Slava	Download data
splaying 1000 rows of results. Use the	download link to retrieve complete results.			and Oneshiroon gorande goard	
Henotype =	Etele Sereeding =	Ink to Siregal ÷		(Citrue ratioulate Blance v Citrue electrole (L.) Och.)	Accession nan
Habit of tree	Etale Spreading	20274	CITRUS_NATIONAL_COLLECTION	(Citrus reticulata Blanco x Citrus sinensis (L.) OSD.)	Tangor Bergan
Spines	Interna Interna	20260	CITRUS_NATIONAL_COLLECTION	Citrus reticulata Blanco	Clementine Robin
Snape of spines	Absente Absent	20278	CITRUS_NATIONAL_COLLECTION	Citrus clementina Hort, ex Tan.	Clementine Sur
width of epicarp at equatorial area	1,1	20270	CITHUS_NATIONAL_COLLECTION	Citrus reticulata Blanco	Mandarine N
Surface of epicarp at equatorial area	1,4	20242		(Citrus reticulata Blanco x Citrus sinensis (L.) Osb.)	Tangor Ortani
Surface of epicarp	Hugueuse Hugose	20274	CITHUS_NATIONAL_COLLECTION	(Citrus reticulata Blanco x Citrus sinensis (L.) USD.)	I angor Bergan
Surface of epicarp	Lisse Smooth	20270	CITHUS_NATIONAL_COLLECTION	Citrus reticulata Blanco	Mandarine N
Fruit weight	60 - 124 g 60 - 124 g	20328	CITHUS_NATIONAL_COLLECTION	Citrus deliciosa Ten.	Mandarine Mediterranee
Fruit weight	200 - 259 g 200 - 259 g	20274	CITRUS_NATIONAL_COLLECTION	(Citrus reticulata Blanco x Citrus sinensis (L.) Osb.)	Tangor Bergan
Shape of apex of fruit	Convexe Convex	20350	CITRUS_NATIONAL_COLLECTION	Citrus reticulata Blanco	Mandarine Fairc
Shape of apex of fruit	Deprime Depresser	20250	CITRUS_NATIONAL_COLLECTION	Citrus deliciosa Ten.	Mandarine Mediterran
Size of vesicles	Petite Sma	20350	CITRUS_NATIONAL_COLLECTION	Citrus reticulata Blanco	Mandarine Fairc
Juice in endocarp	4:	20295	CITRUS_NATIONAL_COLLECTION	Citrus reticulata Blanco	Mandarine En
Juice in endocarp	5	20242	CITRUS_NATIONAL_COLLECTION	(Citrus reticulata Blanco x Citrus sinensis (L.) Osb.)	Tangor Ortani
Shape of seeds	Absent Abser	20250	CITRUS_NATIONAL_COLLECTION	Citrus deliciosa Ten.	Mandarine Mediterran
Embryo color	Sans Absert	20295	CITRUS_NATIONAL_COLLECTION	Citrus reticulata Blanco	Mandarine En
Chalazal spot color	Absent Absert	20328	CITRUS_NATIONAL_COLLECTION	Citrus deliciosa Ten.	Mandarine Mediterranee
Average number of embryos per seed	Monoembryonne Monoembryor y	20274	CITRUS_NATIONAL_COLLECTION	(Citrus reticulata Blanco x Citrus sinensis (L.) Osb.)	Tangor Bergan
Density of branches	Eparse Spar e	20439	CITRUS_NATIONAL_COLLECTION	Citrus reticulata Blanco	Mandarine Maca
1 2 3 4 5 6	7 8 9 10 Next»		_	XI	Powered by bio
	7 8 9 10 Next»	O GENON	NIC INFORMATION SY	/STEM	Powered by blo
1 2 3 4 5 6	7 8 9 10 Next*	O GENON	MIC INFORMATION SY	/STEM	Powered by blot
1 2 3 4 5 6	7 8 9 10 Next*	O GENOR ACCESSIC	NIC INFORMATION SY	/STEM	
1 2 3 4 5 6	7 8 9 10 Next*	O GENOR ACCESSIC SRA 164 Tangor Br	AIC INFORMATION SY	/STEM	
1 2 3 4 5 6	7 8 9 10 Next*	SRA 164 SRA 164 Tangor Ba Bergamot Chata [] Hibrida [] Hybrida [] Malaguia]	AIC INFORMATION SY	/STEM	
1 2 3 4 5 6	7 8 9 10 Next*	SRA 164 Tangor Ba Bergamot Chata [] Hibrida [] Hybrida [] Malaquia (Citrus rel	AIC INFORMATION SY	/STEM	
1 2 3 4 5 6	7 8 9 10 Next*	SRA 164 Tangor Ba Bergamot Champior Hibrida [] Hibrida [] Malaquina Cirtus refi	AIC INFORMATION SY	rSTEM	
1 2 3 4 5 6	7 8 9 10 Next* RGI GnpIS Genetic And Siregi / A Dentification Accession number Accession number	SRA 164 SRA 164 Tangor Bu Bergan Champior Chata [] Hybrida [] Malaquina (Citrus rel -	AIC INFORMATION SY DDN: Tangor Bergan argamota a [] a [] a [] a [] b tculata Blanco x Citrus sinensis (L.) Osb.	rstem	
1 2 3 4 5 6	7 8 9 10 Next*	SRA 164 SRA 164 Tangor B Bergamot Chata [] Hibrida [] Hibrida [] Hibrida [] Citrus ref - - - -	AIC INFORMATION SY	/STEM	
1 2 3 4 5 6	7 8 9 10 Next*	SRA 164 Tangor Ba Bergamot Chata IJ Hibrida (I Hybrida (I) Hybrida (I) Malaquina (Citrus ref - - - - - - - - -	AIC INFORMATION SY DDN: Tangor Bergan ergamota a () 1) culture at Ecophysiologie de la Qualité des	rSTEM	
1 2 3 4 5 6	7 8 9 10 Next*	SRA 164 Tangor Ba Bergamot Champior Chata j Hibrida j Hibrida j Malaquina (Citrus ref - - UR Généti -	AIC INFORMATION SY DDN: Tangor Bergan ergamota a [] 10 a [] ticulata Blanco x Citrus sinensis (L.) Osb. ticulata Blanco x Citrus sinensis (L.) Osb.	/STEM	
1 2 3 4 5 6	7 8 9 10 Next*	SRA 164 Tangor Br Bergamoto Chata [] Hibrida	AIC INFORMATION SY	/STEM	
1 2 3 4 5 6	7 8 9 10 Next*	SRA 164 SRA 164 Tangor Bt Bergan Champior Chata [] Hibrida [] Malaquina (Citrus rel - UR Génét -	AIC INFORMATION SY DDD: Tangor Bergan argamota a [] b 1 2 1 1 1 1 1 1 1 1 1 1 1 1 1	INRA-Corse	
1 2 3 4 5 6	7 8 9 10 Next*	SRA 164 Tangor Br Bergamoto Chenaging Hibrida [] Hibrida [] Hibrid	AIC INFORMATION SY	rSTEM	

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: <u>http://urgi.versailles.inra.fr/grapemine</u>.

See the **figure 5** below:

GrapeMine: Home - Mozilla Firefox			allowed as	🖾 👣 🗤) 11:14 AM 👤	. Thomas	Letelli	er⊰‡⊱
👉 🕘 urgi.versailles. inra.fr /GrapeMine			🔻 🥙 🔡 🔻 horde versaille	rs	۹ 🦊		# v
🗌 Alfresco Web Client 💡 System Dashboa	ırd 🔢 InterMine documenta 📋 Link to FlyM	ine FlyM 🕒 Ulysse - Accueil 🥀 bioinfo-fr.ne	et Le rep K Central Authenticatio 🛽 🔊 Googl	Traduction GL GitLab			39
	GrapeMine version 0.1						
	Home Templates Lists QueryBuilder Regio	ns Data Sources API 🔬 MyMine	Contact Us Log In				
			Swaren: e.g minut, kinase GO				
	Search	Analyse	Welcome Back!				
	Search GrapeMine. Erter names, ledniffare or kaywords for genes, markers, snp. ontology terms, accessions, etc. (e.g., VMC051050, VMC4F6, VV. 1272464, membrane, Synah).	Enter a list of identifiers.	GrapeMine Integrates many types of data for Grape vine. You can run flexible queries, export results and analyse lists of data.				
	SEARCH	advanced ANALYSE	TAKE A TOUR				
		GRAPEMINE CONTENT					
	Our Mine provides access to many kind of data typ phenotyping data (accessions, phenotyping experim	pes like genomic annotation data (genes, mRNAs, exons ments). <u>Read more</u>	, polypeplides). There are also snp, markers and				
	Query for grapemine content:		us endo				
	» More queries		Pope				
	Perl, Python, Ruby and & Java API						
	Access or dicplation data via our dicatication Programming Instruction (AP) to We provide client Entraries in the body the provide client Entraries in the discovery language.						0
							*

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