Subscribe | Share | Past Issues | Translate

transPLANT Newsletter Spring 2015 Issue 3

View this email in your browser

transPLANT

trans-National Infrastructure for Plant Genomic Science

INSIDE THIS ISSUE

We want to hear from you! Please send any comments or suggestions to us at transplant_help@ebi.ac.uk

1-3 July 2015: 4th transPLANT User Training Workshop on "Mining Plant Variation Data"

We are pleased to announce the 4th transPLANT user Training Workshop in Hinxton, United Kingdom (European Bioinformatics Institute) on **July 01-03, 2015**.

transPLANT is a vibrant consortium of 11 European partners gathered to address the challenges of complex plant genome data integration and analysis. It aims to develop a trans-national infrastructure for plant genomic science. For details, please visit http://transplantdb.eu/user_training_4.

This workshop focuses on current developments in plant data resources at transPLANT partner sites, with a special reference on plant genomic variation data, re-sequencing projects and GWAS analyses. The workshop will provide a basic tutorial on SNP calling and analysis as well as hands-on introductions into partner resources and tools and explain how to obtain, search and use this data.

The workshop is targeted at (experimental) biologists and breeders who have the need to use these resources and concepts in everyday work to interpret own observations and plan new research objectives. No prior (informatics) knowledge or skills are required. Familiarity with using browser- based tools and/or Linux operating systems would be helpful however.

The local organizer is the European Bioinformatics Institute (EBI) in Hinxton, UK.

For details on registration, costs and program please visit
http://www.ebi.ac.uk/training/course/mining-plant-variation-data.

The transPLANT Cloud Infrastructure

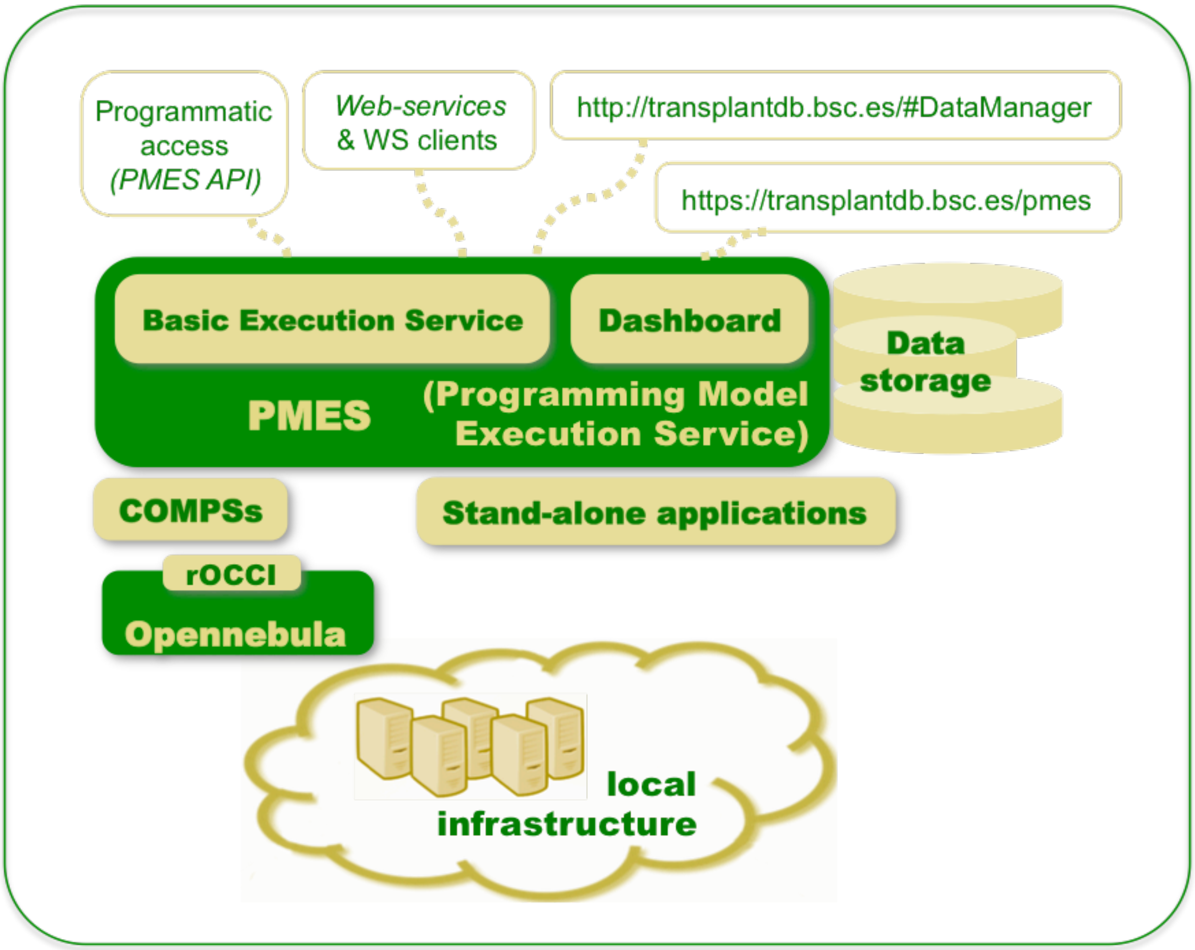
transPLANT offers a cloud-based computational platform that provides interactive and programmatic access to a collection of plant genomic software and analysis pipelines. Applications range from standard bioinformatics tools to specific transPLANT developed applications, as listed in http://transplantdb.bsc.es/transPlantCloud_Apps.htm.

The applications run on virtualized environments, individual virtual machines that confer high flexibility, modularity and portability to the platform. Hence, the infrastructure can grow together with the constantly evolving tools of the field, and furthermore, it can be installed next to data producer centers, mitigating in this way, data security and transfer issues.

The platform also allows to exploit grid, cloud or HPC distributed architectures, easily enabling the execution of standard single applications into distributed environments by integrating COMPSs, a multi scale programming model that exploits the inherent parallelism of applications at execution time, being no needed a specific HPC software development.

PMES framework (Lordan *et al.*, 2004) is used as the entry point of the whole infrastructure, and enables its programmatic access through standard web service technologies. In addition, a web-based dashboard (https://transplantdb.bsc.es/pmes) provides an interactive way to create, submit and monitor task executions on the transPLANT cloud. Users are offered programmatic and web-based access (http://transplantdb.bsc.es/#DataManager) to 2GB of space on the cloud storage system, where input data and execution results are temporally maintained.

To learn more about the infrastructure, please consult http://transplantdb.bsc.es, or mail transplant@bsc.es.

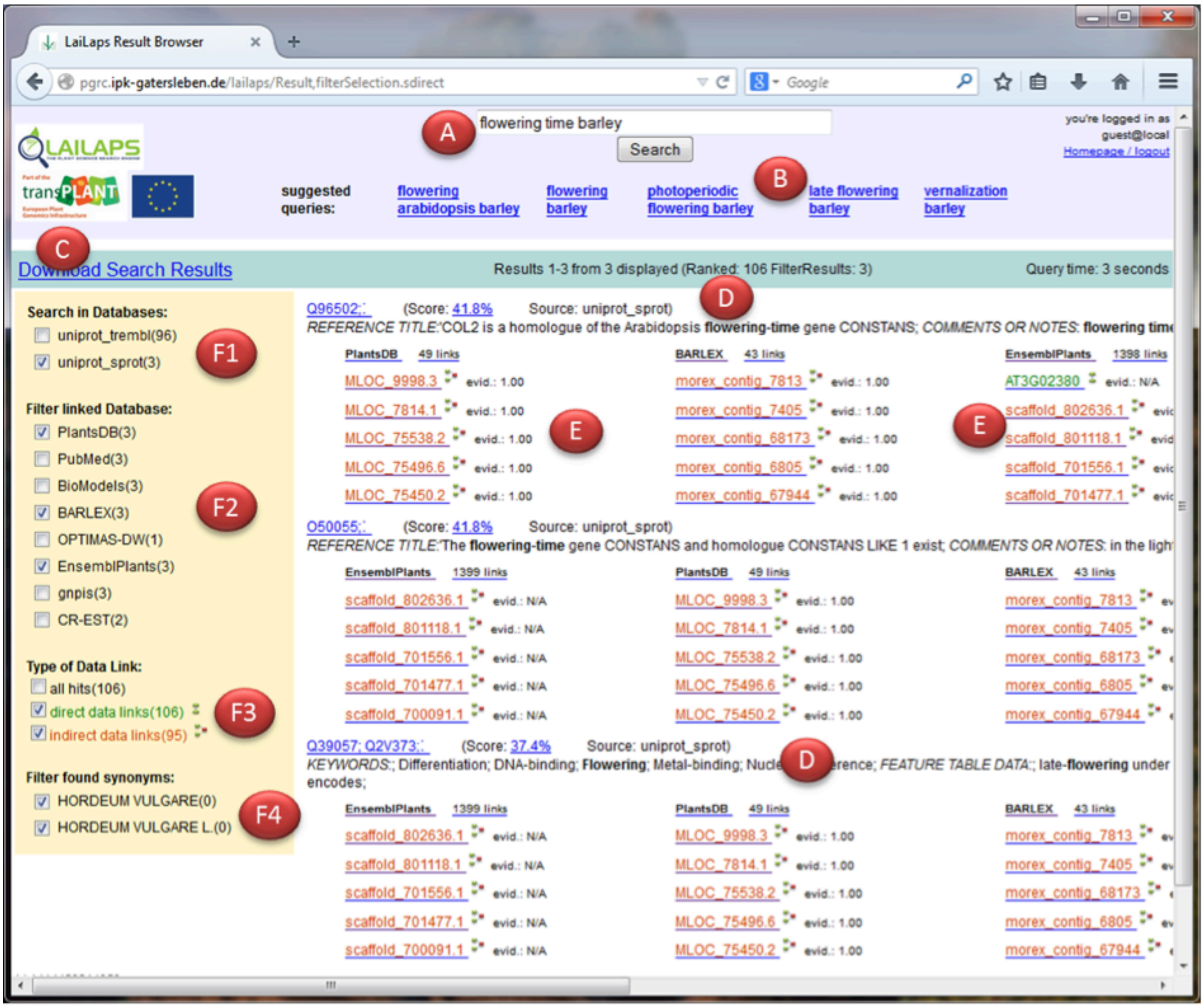

Programmatic access (PMES API)
Web-services & WS clients
http://transplantdb.bsc.es/#DataManager
https://transplantdb.bsc.es/pmes
Basic Execution Service
Dashboard
PMES
(Programming Model Execution Service)
Data storage
COMPSs
Stand-alone applications
rOCCI
Opennebula
local infrastructure

transPLANT cloud infrastructure: *The virtual machines executing transPLANT applications are deployed in the local infrastructure, and OpenNebula is here the middle-ware responsible of managing the hardware resources – though OCCI server ensures the system interoperability when using other middle-wares. The Programming Model Execution Service (PMES) connects to the cloud middle-ware through OCCI standards, and deals with deployment, resources provisioning and contextualization operations, in order to meet user job specifications. Such jobs can require the deployment of a single virtual machine, or multiple ones, if COMPSs is used. PMES also manages I/O operations, here centralized in a data storage accessible via FTP within the cloud, but HTTPS or SSH from the outside. The DataManager site enables the web access to such data, and PMES dashboard allow user to submit and monitor batch jobs. If a programmatic access if preferred, PMES implements BES to tackle these operations from a java API, facilitating the use of standard web services.*

LAILAPS: a New Integrative Search Engine for Plant Genomics Data

LAILAPS was recently published as “The Plant Science Search Engine” (Esch *et al.*, 2014). It offers an integrative information retrieval portal over plant genomics resources. Embedded query assistance and an evidence-based annotation system support scientists in efficient exploration of trait genome associations. The adaptive, information potential considering relevance ranking enables a focused exploration over millions of database records.

The current LAILAPS release comprises about 91 million indexed database records of trait knowledge within 13 major life science data collections and more than 60 million associations to –omics data sets. The ergonometric Web-frontend is featured by encompassing query assistance. A guided result filtering and the suggestion of alternative, semantic similar search queries are intended to guide scientist through the plant genomics big data.



flowering time barley

you're logged in as guest@local

suggested queries: flowering arabidopsis barley | flowering barley | photoperiodic flowering barley | late flowering barley | vernalization barley

Results 1-3 from 3 displayed (Ranked: 106 FilterResults: 3)

Query time: 3 seconds

Search in Databases:
- ☐ uniprot_trembl(96)
- ☑ uniprot_sprot(3)

Filter linked Database:
- ☑ PlantsDB(3)
- ☐ PubMed(3)
- ☐ BioModels(3)
- ☑ BARLEX(3)
- ☐ OPTIMAS-DW(1)
- ☑ EnsemblPlants(3)
- ☐ gnpis(3)
- ☐ CR-EST(2)

Type of Data Link:
- ☐ all hits(106)
- ☑ direct data links(106)
- ☑ indirect data links(95)

Filter found synonyms:
- ☑ HORDEUM VULGARE(0)
- ☑ HORDEUM VULGARE L.(0)

Q96502;: (Score: 41.8%) Source: uniprot_sprot)
REFERENCE TITLE:'COL2 is a homologue of the Arabidopsis **flowering-time** gene CONSTANS; *COMMENTS OR NOTES:* **flowering time**

PlantsDB 49 links		BARLEX 43 links		EnsemblPlants 1398 links	
MLOC_9998.3	evid.: 1.00	morex_contig_7813	evid.: 1.00	AT3G02380	evid.: N/A
MLOC_7814.1	evid.: 1.00	morex_contig_7405	evid.: 1.00	scaffold_802636.1	evic
MLOC_75538.2	evid.: 1.00	morex_contig_68173	evid.: 1.00	scaffold_801118.1	evid
MLOC_75496.6	evid.: 1.00	morex_contig_6805	evid.: 1.00	scaffold_701556.1	evic
MLOC_75450.2	evid.: 1.00	morex_contig_67944	evid.: 1.00	scaffold_701477.1	evic

O50055;: (Score: 41.8%) Source: uniprot_sprot)
REFERENCE TITLE:'The **flowering-time** gene CONSTANS and homologue CONSTANS LIKE 1 exist; *COMMENTS OR NOTES:* in the ligh

EnsemblPlants 1399 links		PlantsDB 49 links		BARLEX 43 links	
scaffold_802636.1	evid.: N/A	MLOC_9998.3	evid.: 1.00	morex_contig_7813	ev
scaffold_801118.1	evid.: N/A	MLOC_7814.1	evid.: 1.00	morex_contig_7405	ev
scaffold_701556.1	evid.: N/A	MLOC_75538.2	evid.: 1.00	morex_contig_68173	
scaffold_701477.1	evid.: N/A	MLOC_75496.6	evid.: 1.00	morex_contig_6805	ev
scaffold_700091.1	evid.: N/A	MLOC_75450.2	evid.: 1.00	morex_contig_67944	

Q39057; Q2V373;: (Score: 37.4%) Source: uniprot_sprot)
KEYWORDS:; Differentiation; DNA-binding; **Flowering**; Metal-binding; Nucl erence; *FEATURE TABLE DATA:*; late-**flowering** under encodes;

EnsemblPlants 1399 links		PlantsDB 49 links		BARLEX 43 links	
scaffold_802636.1	evid.: N/A	MLOC_9998.3	evid.: 1.00	morex_contig_7813	ev
scaffold_801118.1	evid.: N/A	MLOC_7814.1	evid.: 1.00	morex_contig_7405	ev
scaffold_701556.1	evid.: N/A	MLOC_75538.2	evid.: 1.00	morex_contig_68173	
scaffold_701477.1	evid.: N/A	MLOC_75496.6	evid.: 1.00	morex_contig_6805	ev
scaffold_700091.1	evid.: N/A	MLOC_75450.2	evid.: 1.00	morex_contig_67944	

Screenshots of the LAILAPS web interface: *It illustrates the search and results page. The text box (A) enables an interactive and spelling corrected keyword query submission. After query was executed successfully, a list of semantically related phrases is provided for query refinement (B). The query results can be either downloaded as a Microsoft Excel sheet (C) or interactively explored. For this, all relevant hits are displayed as short excerpts in the result* *panel (D). Connected to each hit is a list of links to associated genomic data (E). Those links can be either refer to genome data directly (green) or reflect an indirect, transitive relationship (red). The left hand filter panel enables to restrict the results by fact databases (F1), linked genome databases (F2), direct or indirect linked gene annotations (F3) or synonyms (F4).*

LAILAPS is available at http://www.transplantdb.eu/lailaps or http://lailaps.ipk-gatersleben.de.

A New Method For QTL Candidate Gene Prioritization

In the context of transPLANT, Plant Research International, part of Wageningen University and Resarch Centre, recently published a paper on Quantitative Trait Locus candidate gene prioritization (Bargsten *et al.*, 2014). Quantitative Trait Loci (QTLs) indicate genome regions influencing complex traits, such as yield. A bottleneck in application of these data is that QTL regions are relatively large, containing tens to hundreds of genes. Selecting potential causal genes often is done by hand, which is time consuming. Our bioinformatics method prioritizes causal genes in an automatic fashion, for many different traits. Such candidate gene prioritization is useful for marker development, further use of genes, and for better understanding of the trait.

The method consists of two steps: (1) Prediction of biological processes that genes are involved in. This step takes as input a genome sequence with structural gene annotation, and expression data in different conditions. Its output is for each gene, a list of predicted gene functions (biological processes). (2) Integration of predicted gene functions with QTL data. This step takes as input QTL regions for traits of interest, mapped to the genome sequence. Its output consists of lists of gene functions that are overrepresented in the different QTL regions for each trait. On the basis of these lists, prioritization of genes in the QTL regions is performed. This method was applied to a rice QTL-compendium with around 150 different traits; see also http://www.ab.wur.nl/bmrftrait. We compared our predictions with experimentally validated causal genes underlying those QTLs; this indicated very significant performance (p<0.001). Currently we are working on further improvement of the method, as well as application to other crops and other types of data such as eQTLs and mQTLs.

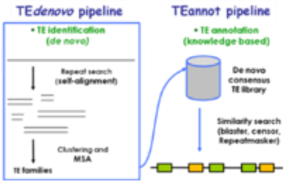
REPET: A Software for the Analysis of Repeats

The recent successes of new sequencing technologies allow today to sequence increasingly large genomes at reduced costs. Transposable elements (TEs) constitute the most structurally dynamic components and the largest portion of nuclear sequences of these large genomes, e.g. 85% of the maize genome (Schnable *et al.*, 2009), and 88% of the wheat genome (Choulet *et al.*, 2014). Therefore, TEs annotation should be considered as a major task in these genome projects. However, this still remains a major computational challenge, this crucial step is now a bottleneck for many genome analyses. We scaled-up a repeat detection and annotation package called REPET (Flutre *et al.*, 2011), now at its v2.2 release (http://urgi.versailles.inra.fr/Tools/REPET).

Improvements to the latest version include:

1. A structural TE detection is now implemented. LTRharvest (Ellinghaus *et al.*, 2008) is used to search for LTR retrotransposons, using structural features of this TE category. Potential detected TEs are then classified to remove false positives.
2. Classification has also been improved with the development of PASTEC (Hoede *et al.*, 2014). It tests all TE classifications, each result being weighted according to the evidences found. In addition to similarities to known TEs in Repbase Update and the search for repeated structures, it also uses HMM profiles, which are interesting to classify TEs and to detect host genes.
3. We propose a new pipelines based on Tallymer (Kurtz *et al.*, 2008), called TallymerPipe, as pre-processing tool for a fast repeated region detection.
4. We also propose SegDup, a pipeline to detect segmental duplications, taking care of TEs, based on our previous work (Fiston-Lavier *et al.*, 2007).

Using these pipelines and the tools from the REPET package, we applied a new strategy, to cope with very large genomes such as the wheat (Choulet *et al.*, 2014, Daron *et al.*, 2014).


TEdenovo pipeline
• TE identification (de novo)
Repeat search (self-alignment)
Clustering and MSA
TE families
TEannot pipeline
• TE annotation (knowledge based)
De novo consensus TE library
Similarity search (blaster, censor, Repeatmasker)

Back to top

The transPLANT project is funded by the **European Commission** within its 7th Framework Programme, under the thematic area "iInfrastructure", **contract number 283496.**

If you wish to unsubscribe from this mailing list, please email:
transplant_help@ebi.ac.uk