

# Lawrence Berkeley National Laboratory

## LBL Publications

### Title

InterMine Webservices for Phytozome

### Permalink

<https://escholarship.org/uc/item/0j39j94m>

### Authors

Carlson, Joseph  
Hayes, David  
Goodstein, David  
[et al.](#)

### Publication Date

2014-01-10

# InterMine Webservices for Phytozome

Authors: Joseph Carlson<sup>1</sup>, Richard Hayes<sup>1</sup>, David Goodstein<sup>1</sup>, Daniel Rokhsar<sup>1</sup>

<sup>1</sup> U.S. Department of Energy Joint Genome Institute // LBNL - Walnut Creek, CA

*\* To whom correspondence may be addressed. Joseph Carlson, Joint Genome Institute, 2800 Mitchell Drive, Walnut Creek, CA, 94598, USA - JWCarlson@lbl.gov@lbl.gov*

January 10, 2014

## **ACKNOWLEDGMENTS:**

Work by the U.S. Department of Energy Joint Genome Institute is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.

## **DISCLAIMER:**

This document was prepared as an account of work sponsored by the United States Government. While this document is believed to contain correct information, neither the United States Government nor any agency thereof, nor The Regents of the University of California, nor any of their employees, makes any warranty, express or implied, or assumes any legal responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by its trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof, or The Regents of the University of California. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof or The Regents of the University of California.

# InterMine Webservices for Phytozome

Joseph W Carlson, Richard D Hayes, David M Goodstein and Daniel S Rokhsar

DOE Joint Genome Institute, Walnut Creek, CA



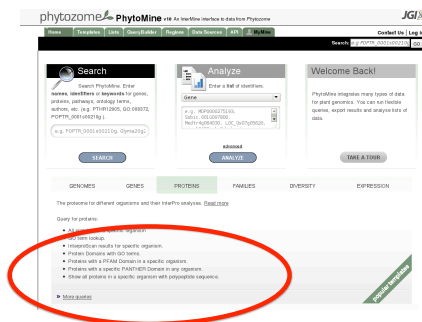
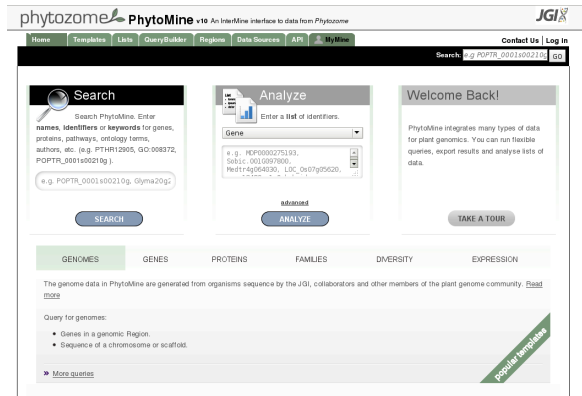
Visit [www.phytozome.net](http://www.phytozome.net) for current status and deployment URL

A data warehousing framework for biological information provides a useful infrastructure for providers and users of genomic data. For providers, the infrastructure give them a consistent mechanism for extracting raw data. While for the users, the web services supported by the software allows them to make either simple and common, or complex and unique, queries of the data.

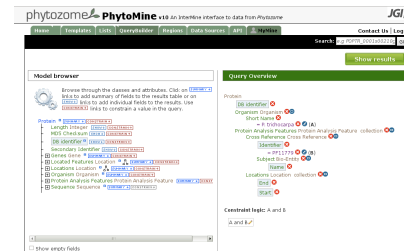
Previously, phytozome.net used BioMart to provide the infrastructure. As the complexity, scale and diversity of the dataset as grown, we decided to implement an InterMine web service ([www.intermine.org](http://www.intermine.org)) on our servers. This change was largely motivated by the ability to have a more complex table structure and richer web reporting mechanism than BioMart.

Our implementation of InterMine is used to serve data for the next version of the web app on phytozome.net through their API. But it is also available for general use for information beyond what is available on our web pages.

For InterMine to achieve its more complex database schema it requires an XML description of the data and an appropriate loader. Unlimited one-to-many and many-to-many relationship between the tables can be enabled in the schema. Within this framework, complex SQL queries are generated which can return results in a highly optimized manner, with performance exceeding what we saw with BioMart.



One interface to phytomine relies on templated queries for generating results. A set of templates is linked from the start page. These can be used directly, or used as starting points for more complex queries by editing the templates.

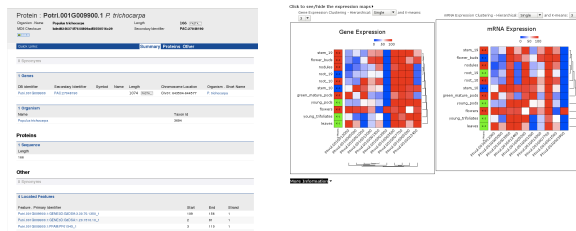


The QueryBuilder interface of InterMine enables browsing the database schema and graphical selection of fields extracted in a query. It also allows constraints on the query as well. This example shows the edits possible for the query that extract all proteins in poplar that contain the PFAM domain PF11779.

Track Query: SNPs Near an Annotated Gene.

Location On chr	SNP Identifier	Start	End	Reference	Strand	Allele	Distance From Gene	Gene Name	Type	GeneSource	Subcellular
Dv01	36811	G	A	BE00108	DOWNSTREAM	NOISE					
Dv01	36812	G	A	BE00108	DOWNSTREAM	NOISE					
Dv01	36813	G	A	BE00108	DOWNSTREAM	NOISE					
Dv01	36814	G	A	BE00108	DOWNSTREAM	NOISE					
Dv01	36815	G	A	BE00108	DOWNSTREAM	NOISE					
Dv01	36816	G	A	BE00108	DOWNSTREAM	NOISE					
Dv01	36817	G	A	BE00108	DOWNSTREAM	NOISE					
Dv01	36818	G	A	BE00108	DOWNSTREAM	NOISE					
Dv01	36819	G	A	BE00108	DOWNSTREAM	NOISE					
Dv01	36820	G	A	BE00108	DOWNSTREAM	NOISE					
Dv01	36889	C	A	BE00108	DOWNSTREAM	NOISE					

The output of a query is normally a list. The example shows a list of SNPs near an annotated gene in poplar. The output of the list may be downloaded, or, by selecting one of the displayed values, further information can be extracted.



Full reports of individual items displays information about the item and other items related to it. The example in the upper left shows a protein report and includes the associated gene, transcript and Interproscan results. Alternatively, as shown on the right, a set of genes can be used to extract a heat map of gene or mRNA expression.

We have implemented a database for:

1. Genomes and annotations for the data in Phytozome. This set is the 45 organisms currently stored in a back end CHADO datastore. The data loaders are modified versions of the CHADO data adapters from Flymine.
2. Interproscan results from all proteins in the Phytozome database.
3. Clusters of proteins into a grouped hierarchically by similarity.
4. Cufflinks results from tissue-specific RNA-Seq data of Phytozome organisms.
5. Diversity data (GATK and SnpEFF results) from a set of individual organism.