Database, Vol. 2011, Article ID bar044, doi:10.1093/database/bar044

Database tool

Tripal: a construction toolkit for online genome databases

Stephen P. Ficklin[1], Lacey-Anne Sanderson[2], Chun-Huai Cheng[1], Margaret E. Staton[3], Taein Lee[1], Il-Hyung Cho[4], Sook Jung[1], Kirstin E. Bett[2] and Doreen Main[1,*]

[1]Department of Horticulture and Landscape Architecture, Washington State University, Pullman, WA, USA 99164, [2]Department of Plant Sciences, University of Saskatchewan, Saskatoon, SK, Canada S7N 5B5, [3]Clemson University Genomics Institute (CUGI), Clemson University, Clemson, SC, USA 29634 and [4]Department of Computer Science, Saginaw Valley State University, University Center, MI, USA 48710

***Corresponding author: Tel: 509 335 9000; Fax: 509 335 8690; Email: dorrie@wsu.edu**

Submitted 1 July 2011; Revised 23 August 2011; Accepted 29 August 2011

As the availability, affordability and magnitude of genomics and genetics research increases so does the need to provide online access to resulting data and analyses. Availability of a tailored online database is the desire for many investigators or research communities; however, managing the Information Technology infrastructure needed to create such a database can be an undesired distraction from primary research or potentially cost prohibitive. Tripal provides simplified site development by merging the power of Drupal, a popular web Content Management System with that of Chado, a community-derived database schema for storage of genomic, genetic and other related biological data. Tripal provides an interface that extends the content management features of Drupal to the data housed in Chado. Furthermore, Tripal provides a web-based Chado installer, genomic data loaders, web-based editing of data for organisms, genomic features, biological libraries, controlled vocabularies and stock collections. Also available are Tripal extensions that support loading and visualizations of NCBI BLAST, InterPro, Kyoto Encyclopedia of Genes and Genomes and Gene Ontology analyses, as well as an extension that provides integration of Tripal with GBrowse, a popular GMOD tool. An Application Programming Interface is available to allow creation of custom extensions by site developers, and the look-and-feel of the site is completely customizable through Drupal-based PHP template files. Addition of non-biological content and user-management is afforded through Drupal. Tripal is an open source and freely available software package found at http://tripal.sourceforge.net

Introduction

Recent improvements in DNA sequencing have vastly increased the quantity and availability of genomic and transcriptomic data for model and non-model species. The availability of open source and free bioinformatics tools for data analysis has also facilitated the adoption of genomics by a larger number of laboratories and investigators. Large entities such as the National Center for Biotechnology Integration (NCBI), the European Molecular Biological Laboratory (EMBL) and the DNA Databank of Japan (DDBJ), continue to serve as repositories for large quantities of biological data, however, model organism and community-specific online databases such as FlyBase (1), WormBase (2), Saccharomyces Genome Database (3), Genome Database for Rosaceae (4, 5), Marine Genomics (6), Gramene (7), MaizeGDB (8, 9), TreeGenes (10), Sol Genomics Network (SGN) (11) and others were developed to address unique needs for targeted communities of scientists. These tailored databases fill an important niche by providing specific analysis tools, data mining capabilities and collaboration resources for their respective communities. Moreover, with increases in quantity, complexity and accessibility of biological data, the need for a manageable solution for constructing such a database is apparent. Non-biological tools and content, such as social networking, forums and outreach are often a necessary component

Database, Vol. 2011, Article ID bar044, doi:10.1093/database/bar044

for the communities that these sites serve. The need is especially great for small community collaborations, newly emerging communities and individual labs that perform the experiments.

Content Management Systems (CMS) were developed to help simplify website installation for site administrators, web development for programmers and content changes for non-technical users. Drupal (http://www.drupal.org) is an open source, popular and well-supported CMS that has been used to construct a wide variety of websites from small to enterprise-level. The Drupal website maintains a repository of thousands of user-contributed extensions (also called modules) that are freely available. These extensions, or modules, are easily obtained and installed thus expanding the functionality of a website within minutes. They include features for blogging, e-commerce and discussion forums to name a few. Hundreds of user-contributed 'themes' are also available for download, which change the look-and-feel of a site within minutes. Books, Drupal community conferences, online documentation and support forums are all available. A well-documented Application Programming Interface (API) is available for programmers to extend the functionality of the software by creating new modules and themes. These custom modules and themes can in turn be shared with the Drupal community of users through a repository on the Drupal website. Drupal reduces the Information Technology (IT) costs of web development by simplifying site construction and management.

To assist with the storage of complex biological data relationships, the Chado database schema was created as an open source, community-supported relational database schema (12, 13). Chado is a product of the Generic Model Organism Database (GMOD) project (http://www.gmod.org) which builds and coordinates construction and distribution of a suite of tools that support storing, mining and visualization of genomic, genetic and other biological data. Chado is designed for storage of complex relationships among data types such as sequence features (e.g. genes, mRNA, chromosome, etc.), controlled vocabularies such as ontologies, phenotypes, genotypes, phylogeny, stock collections, publications, gene expression data and more. Chado tables are grouped into modules, or groups of tables that store data for related data types (http://gmod.org/wiki/Chado_Modules). To avoid confusion with Drupal modules, we will refer to Chado modules as Chado table groups. For example, a sequence table group exists for storing genomic sequence and ancillary data; and a stock table group exists for storing stock collections.

Several popular tools interface with Chado, including GBrowse (14, 15), a commonly used genome browser, Apollo (16), a tool for distributed manual gene annotation, and Ergatis (17), a platform for development and execution of bioinformatics workflows. Additionally, data stored in Chado can be exchanged between collaborative projects using an exchange format called ChadoXML (13). Together, the Chado database and suite of tools provides a standard for storage of biological data and data exchange. These tools reduce IT costs by limiting the need for database design and software development.

Tripal was created to provide a simplified, standards-based, web construction toolkit for display and searching of genomic data using both Chado and Drupal. It is an open source and freely available collection of Drupal modules for management and visualization of data stored within a GMOD Chado database, and is a member of the GMOD family of tools. Tripal version 0.3.1b currently provides support for Chado sequence, library, organism, general, controlled vocabulary and stock table groups. Additionally, several extension modules provide loaders and visualization for common bioinformatic analyses such as unigene assemblies, Kyoto Encylopedia of Genes and Genomes (KEGG) (18), Basic Local Alignment Search Tool (BLAST) (19), InterProScan (20) and Gene Ontology (GO) (21) annotations. A GBrowse extension for Tripal coordinates synchronization of data between Chado and a GBrowse database. Tripal also incorporates features such as a built-in Chado installer (version 1.11); a jobs management subsystem for managing long running tasks; support for materialized views to speed data queries; data loaders for the popular FASTA, GFF3 and Open Biomedical Ontology (OBO) formats; and an API that allows for easy interaction with Chado for the development of custom extension modules. Tripal also integrates with popular and powerful Drupal extensions such as Views and Panels, and draws on the full-text searching mechanism of Drupal to provide default data searching.

Currently, Tripal is in use by the Fagaceae Genome Web (http://www.fagaceae.org), the Cacao Genome Database (http://www.cacaogenomedb.org), the Genome Database for Vaccinium (http://www.vaccinium.org), the Citrus Genome Database (http://www.citrusgenomedb.org), Pulse Crop Genomics & Breeding (http://knowpulse2.usask.ca/portal), the Marine Genomics Project (http://www.marinegenomics.org) and the Cool Season Food Legume Genome Database (http://www.gabcsfl.org). Additionally, the Genome Database for Rosaceae (http://www.rosaceae.org) is currently in the process of incorporating Tripal into its existing framework.

Two other tools exist that provide web front-ends for Chado. These include GMODWeb (22) and ChadoOnRails. GMODWeb is implemented using Turnkey (http://radius.genomics.ctrl.ucla.edu/turnkey/pmwiki.php), a Model-View-Controller (MVC) framework in Perl. GMODWeb provides broad support of Chado tables, and integrates easily with popular Perl-based tools such as BioPerl (23, 24). ChadoOnRails (http://rubyforge.org/projects/chadoonrails/) uses the Ruby on Rails (http://rubyonrails.org/) web application framework to deliver an Object-Relational Mapping

(ORM) to allow other Ruby on Rails applications access to Chado content. Users can use ChadoOnRails to construct a web front-end or other application that requires access to a Chado database. ChadoOnRails serves as an interface to Chado for developers of Ruby on Rails applications and while it can be used for development of a web front-end for Chado, it is not application specific. Tripal differs primarily from these tools through tight coupling with a CMS (i.e. Drupal) which provides for site expandability outside of the realm of genomics, as well as providing authentication, user management and incorporation of non-genomic content.

The Drupal Bioinformatics Server Framework (DBSF) (25) is a Drupal-based set of modules for providing a bioinformatics computational web framework within Drupal. It supports several commonly used bioinformatics tools, but also provides an API to allow the site administrator to easily add additional tools. It supports execution of tools on distributed computational equipment using computational job scheduling tools such as Condor (26). In some cases it uses the Perl-based BioPerl (23, 24) toolkit. DBSF houses data results in database tables similar to Chado, but can later copy that data to a full Chado database once approved by a curator. In comparison, Tripal does not provide a framework for execution of bioinformatics applications, but rather provides a means for dissemination and visualization of data. Tripal does provide an API for interacting with a Chado database, as well as data loaders (e.g. FASTA, GFF3 and OBO loaders) but interacts directly with the Chado database and does not use BioPerl. Because DBSF and Tripal are both Drupal-based solutions it is possible to use them side-by-side to provide a web-based solution for data analysis (DBSF) and public dissemination and visualization (Tripal).

Description

In summary, Tripal is a set of Drupal modules that provides data management, visualization and searching capabilities for data stored within a Chado database. The primary objective is to simplify web development for an online genomic database by bridging the strengths of both Drupal and Chado. The modules in Tripal allow the Drupal CMS to interact with Chado data, as well as provide data loaders, display of Chado data and administrative interfaces for data management. Moreover, use of Tripal affords several additional advantages derived from both Drupal and Chado. Drupal provides user-management capabilities, social networking tools and easy creation of non-biological content, such as outreach content. Privileged users can easily update content without knowledge of HTML. Chado provides a common data storage platform, which improves data sharing capabilities, but also provides integration with other tools such as GBrowse (14, 15) and Apollo (16). Users of a Tripal site can customize the look-and-feel without the need for programming. However, to support novel customizations Tripal provides an API that allows for fine-grained changes to the look-and-feel of the site, or for development of extension modules for new functionality—Tripal users need not be bound by the default Tripal offering. Support for Tripal is readily available online at the Tripal website, the GMOD website or through an active email mailing list.

Tripal is released under a GNU Public License (GPL) v2 license. It is a member of the GMOD family of tools, and is available for download from the Tripal website (http://tripal.sourceforge.net) along with installation and upgrade documentation. The GMOD website offers tutorials and a user's mailing list (http://www.gmod.org/tripal). Individuals who develop new extensions for Tripal are welcome to submit their modules for inclusion in the Tripal extension repository. The development version of the Tripal source code is also available in a Subversion (SVN) repository managed by GMOD. Additionally, a pre-installed self-contained virtual machine is available on the Tripal website. Individuals may download this virtual machine, use it on their existing computer or server to bypass the installation steps and begin testing Tripal with their own data.

Tripal can be installed, populated with data and managed by a single bioinformaticist with modest UNIX skills. Projects that require additional customization beyond the default offering of Tripal will need personnel to perform customizations and data analysis dependent on time requirements and scope of the project. A PHP programmer with JavaScript and Cascading Style Sheets (CSS) experience is required for fine-grained customization or construction of new extensions. Familiarity with Chado, or the ability to understand relational databases is also required for customizations.

For small projects, a typical off-the-shelf, high-end desktop computer can be used to install and run a Tripal site. A UNIX operating system (e.g. Linux) is required, as well as, a web server such as Apache (http://httpd.apache.org/) and a PostgreSQL database server (http://postgres.org). For larger projects, specialized server-class hardware may be required. Server costs will vary based on hardware requirements. As with all database applications, sites that anticipate large data sets can experience performance bottlenecks when querying a database if hardware is insufficient. A review of the data sets and required server resources should be performed before implementation of any database application.

Data in Tripal is presented in the form of pages (or nodes in Drupal terminology). Currently Tripal provides pages for organisms, features (genomic sequences, e.g. genes), analyses (e.g. KEGG, InterProScan, BLAST), libraries [e.g. Bacterial Artificial Chromsome (BAC)] and stocks

Database, Vol. 2011, Article ID bar044, doi:10.1093/database/bar044

[e.g. Deoxynucleic acid (DNA) extractions, organism individuals, populations]. To demonstrate these pages, an example site has been created (http://tripal.gmod-dev.oicr.on.ca/0.3.1b/) and screenshots from this site are provided hereafter. The look-and-feel of the example site is controlled by the Sky theme (http://drupal.org/project/sky), one of many freely available themes downloadable from the Drupal website. The look-and-feel for Chado content, however, is provided by Tripal. Within this site is data for a fictitious organism, *Tripalus databasica* (readers are referred to the list of sites indicated in the introduction for examples of actual Tripal-based websites). This example site contains whole genome annotation data, as well as stock information. Here we will provide figures showing an organism page, a genomic feature (gene) page, a stocks page and a demonstration using the Drupal Views extension.

The Organism page (Figure 1) provides details about the organism, such as the common name, genus and species. These data are obtained directly from Chado. The resources sidebar on the right is typical for the default Tripal theme and lists other available data for the organism.

In Figure 1, the surrounding theme, including the header with menu links, is managed by Drupal. The 'About' and 'Contact Us' links are for content added by the site administrator and when clicked will take the user to each respective page. The content on these pages is not stored in Chado. As is typical of the default Tripal installation, there are Tripal-provided menu items in the primary menu of the site (e.g. 'Analyses', 'Organisms', 'Stocks', etc). The links are added by Tripal modules. When clicked, these menu items take the site visitor to either a listing of available resources (e.g. a listing of organisms), or provide a simple search mechanism for finding data. These Tripal added menu items can easily be removed if desired. Users may login to the site using Drupal's user management functionality by using the login box on the bottom left hand side. Figure 2 shows a summary of the GO

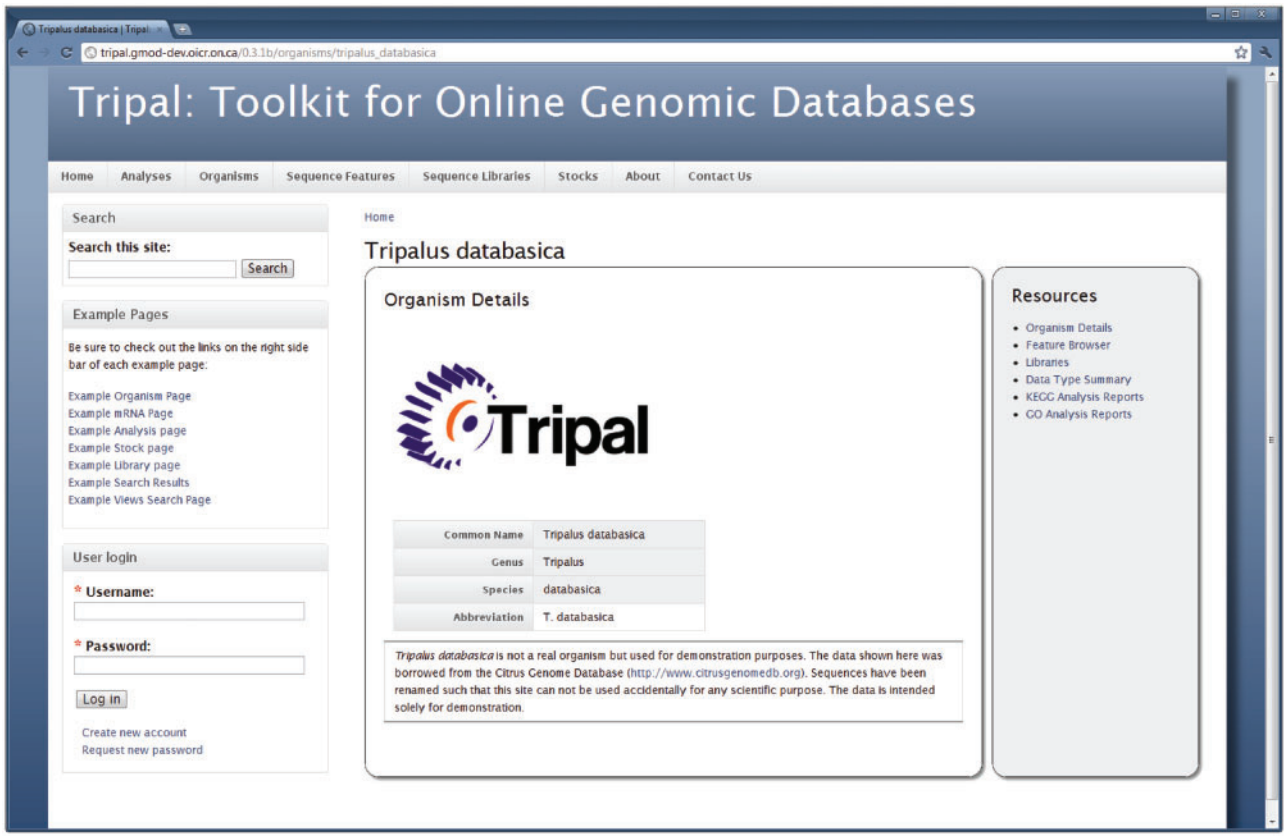

Figure 1. The Organism page. Details about an organism are presented first when visiting the Organism page. The right hand sidebar titled 'Resources' provides additional links for data related to this Organism. Through these links, visitors can use the feature browser to sort through a list of available genomic features (sequences) for this organism, view a list of feature libraries for this organism, and view reports of summarized data for the organism, including GO and KEGG reports. A search box in the top left-hand side is available for full-text searching of the site, and a login box is available for administrative and community logins. By customizing the Tripal template files, site administrators can add any number of links to the resources sidebar.

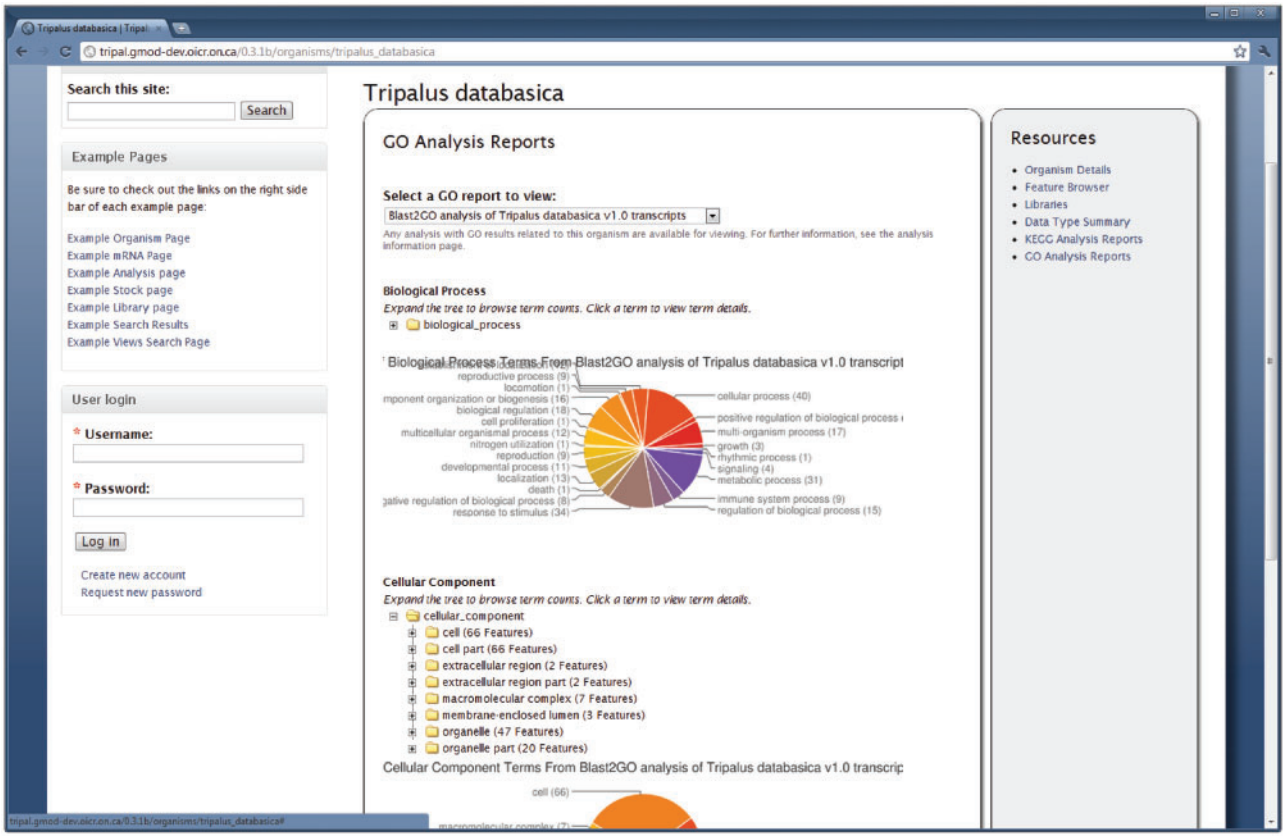

Figure 2. The GO summary on the Organism page. The Tripal Analysis GO module provides summarized reports of GO annotations for an organism. This report is available by clicking the 'Go Analysis Reports' link on the right sidebar of the Organism page. The report shown above is a summarized report from a Blast2GO analysis. The pie charts indicate the percentage of annotations for each of the highest level terms in each of the three branches of the GO. Additionally, expandable trees for each branch are available for browsing all of the assignments. Users can click a GO term in the tree for a description of the term and download a FASTA file of sequences annotated with a particular term.

annotations for an organism. The report includes pie charts for the top-level terms for each branch of the ontology as well as expandable trees for browsing the number of assigned terms. This report, along with a similar KEGG report are generated automatically after data is loaded and prepared by the Tripal GO and KEGG extension modules.

Figure 3 shows a list of biological libraries available for the organism. The visitor clicks the 'Libraries' link in the resources sidebar of the organism page to access this information. This list is automatically added to the organism page when the Tripal Library module is installed and when libraries exist for the organism.

Each genomic feature made public on the site will also have a page. The page layout is similar to that of the organism page with a resources sidebar of data links. Figure 4 shows the feature page when the 'mRNA Colored Sequence' link is clicked. Figure 5 shows the BLAST homology results for the same feature when the 'ExPASy Homology' link is clicked.

As seen in Figure 5, the name for the BLAST homology analysis is provided as a link at the top of the results. When clicked, this takes the site visitor to the Analysis page. Here the layout is similar to the Organism and Feature pages. Figure 6 shows the details for the BLAST homology analysis that generated the results seen in Figure 5. Here, users are presented with information that describes how the analysis was performed.

Each stock made public on the site has a content page similar in layout to both the Feature and Organism pages (Figure 7). Properties, database references, relationships and synonyms for stocks are available for a given stock by clicking the links on the right sidebar. Additionally, content managers can edit the main details of a stock, by clicking the 'Edit' tab near the top of the page. Properties, database references or relationships can be edited by clicking the 'Edit Properties', 'Edit Database References' or 'Edit Relationships' tab respectively. For example, a screenshot of the interface for editing stock properties is shown in Figure 8.

Database, Vol. 2011, Article ID bar044, doi:10.1093/database/bar044

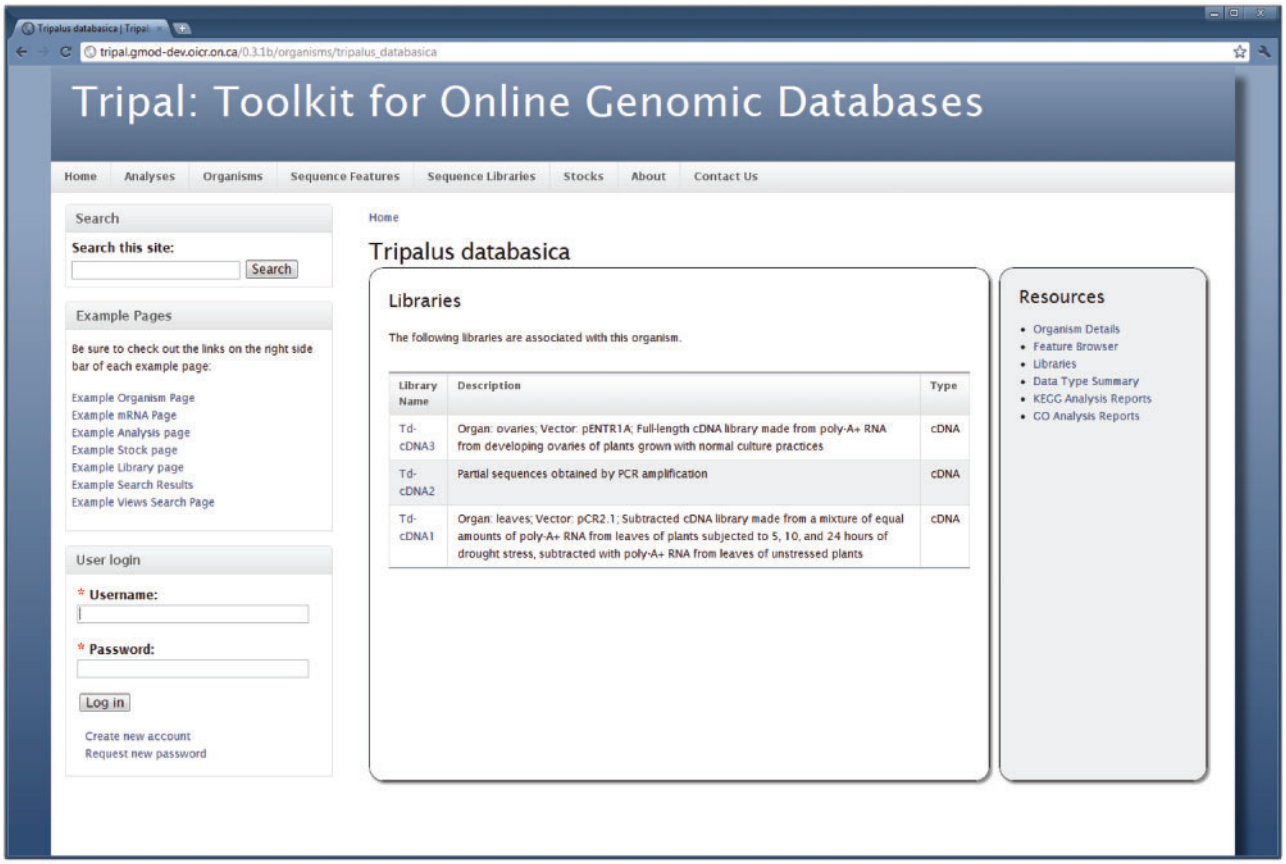

Figure 3. The libraries for an organism. The Tripal Library module will automatically provide a list of libraries associated with an organism. This list is available by clicking the 'Libraries' link on the right sidebar of the Organism page. New libraries will automatically appear in the list as they are added. Visitors can find more information about a library by visiting a library's corresponding content page, which is found by clicking on the library name in the list.

As mentioned previously, the menu links in the site's primary menu are added by Tripal. Tripal comes with several pre-packaged default Views. Views is a Drupal extension module that provides a web interface to the site administrator for creating custom query pages without using SQL (more about Views is discussed in the Software Overview section). The Tripal added menu items link to pages that contain simple lists, such as a list of organisms. However, after Views is enabled these pages become searchable query pages. Figure 9 shows an example page created using Views for the 'Sequence Features' menu item. Here users can easily search for a particular analysis and order the table by a specific column. Using the Views administrative interface (not shown), a site administrator can customize these default Views provided by Tripal or create novel Views without the need for writing SQL.

Finally, Drupal provides a full-text searching tool. This tool will search all pages (including non-Tripal pages) for query words provided by the site visitor. Figure 10 shows an example set of search results when searching for a particular biological process term in the demonstration database. Under the advanced options, site visitors can filter search results using categories such as the organism name or sequence type.

Software overview

Tripal consists of several Drupal modules and a Drupal theme created using the published Drupal API. As shown in Figure 11, Tripal modules are hierarchical, and can be classified into several layers: Core module, Chado modules, Extension modules and Application layers. The Tripal theme and Chado module layers are described in the following paragraphs.

The Tripal base theme

As mentioned previously, Drupal allows for changes to the look-and-feel of the site by means of custom theming. Many customizations can be implemented by a site administrator without the need for programming. The look-and-feel of a site can be altered almost instantly by

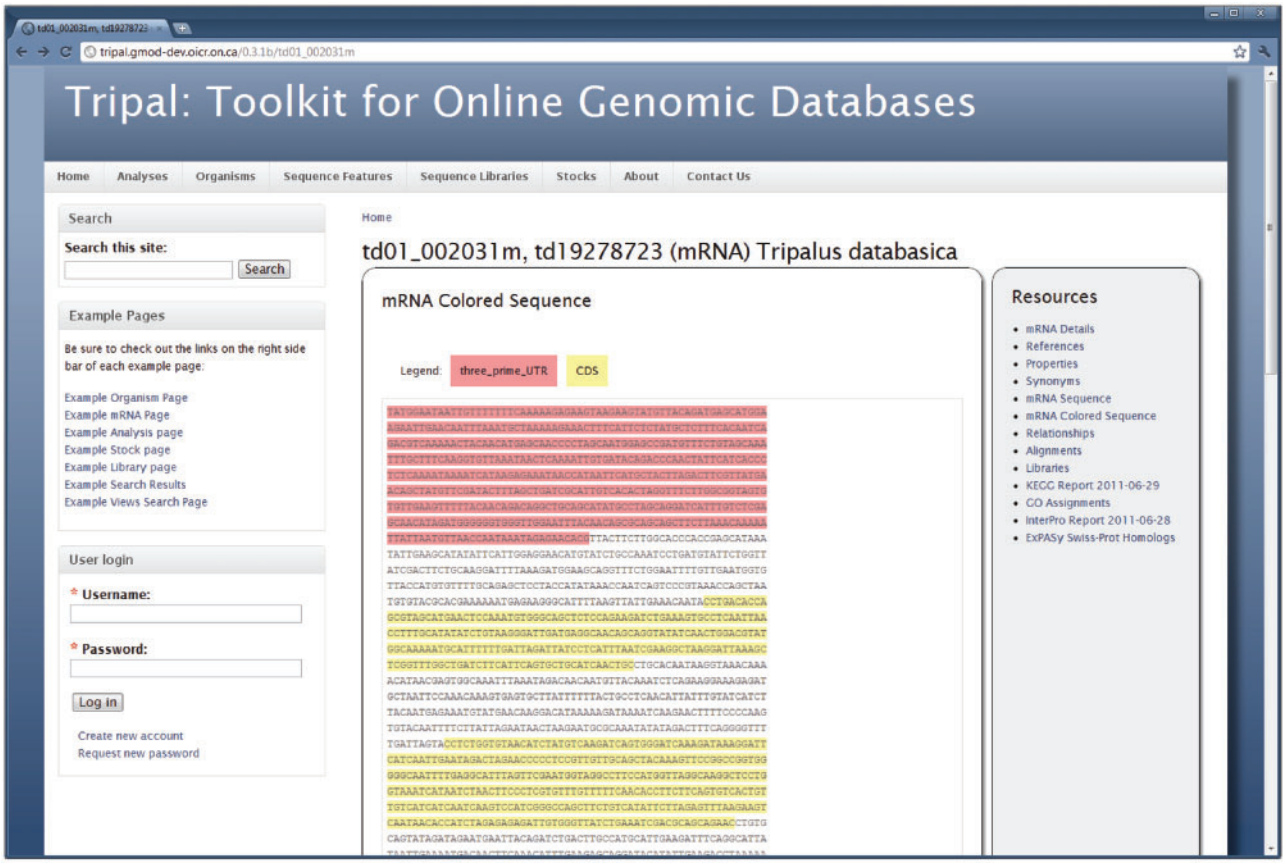

Figure 4. The feature page. Similar to the organism page, the right sidebar of the feature page provides a list of available data for a genomic sequence. Here the genomic sequence for an mRNA feature is shown. This is viewable by clicking the 'mRNA Colored Sequence' link. Visitors can also view synonyms, properties, alignments and external references for the feature. Library information will be viewable if the feature is derived from a library. Analyses results are also available on this page. In this example, BLAST homology results against the ExPASy Swiss-Prot database, InterProScan and KEGG/KASS results are available for viewing. Site administrators may customize which links are shown and can change the wording on many of the data views to better fit their community by editing the Tripal template files. Here the default look-and-feel is shown.

downloading and installing a new theme from the Drupal repository. Powerful tools such as Views and Panels provide custom presentation of content without the need for programming. If needed, however, fine-grained customizations can be made by editing PHP template files that accompany a Drupal theme.

Tripal includes a base theme that provides a default look-and-feel for only Chado-derived data. This allows for complete customization of the remaining portion of the site using any publicly available Drupal theme or in-house custom theme. Moreover, the Tripal theme is itself customizable. Template files are made available within this theme and provide fine-grained control for presentation of Chado data. Site developers familiar with PHP, JavaScript, HTML and CSS can easily edit these templates to change or customize the presentation. Chado content is automatically made available to these templates by way of PHP variables, which the developer can access and expand. The structure of these template variables mimics the Chado tables and makes accessible any record in the data set. Thus, site developers can customize templates in new ways not provided in the default theme.

Tripal modules

The Tripal Core module provides base functionality and an API for which all other Tripal modules are dependent. This API includes functions for selecting, inserting, updating and deleting records in Chado. Also present in the API are functions for generating the variables used in the template files of the Tripal theme. The Core also provides a jobs management subsystem and a mechanism for managing materialized views. The jobs management subsystem allows site administrators to execute long running tasks, such as loading of very large data files. A jobs history is preserved and administrators can cancel a pending job, reschedule a

Database, Vol. 2011, Article ID bar044, doi:10.1093/database/bar044

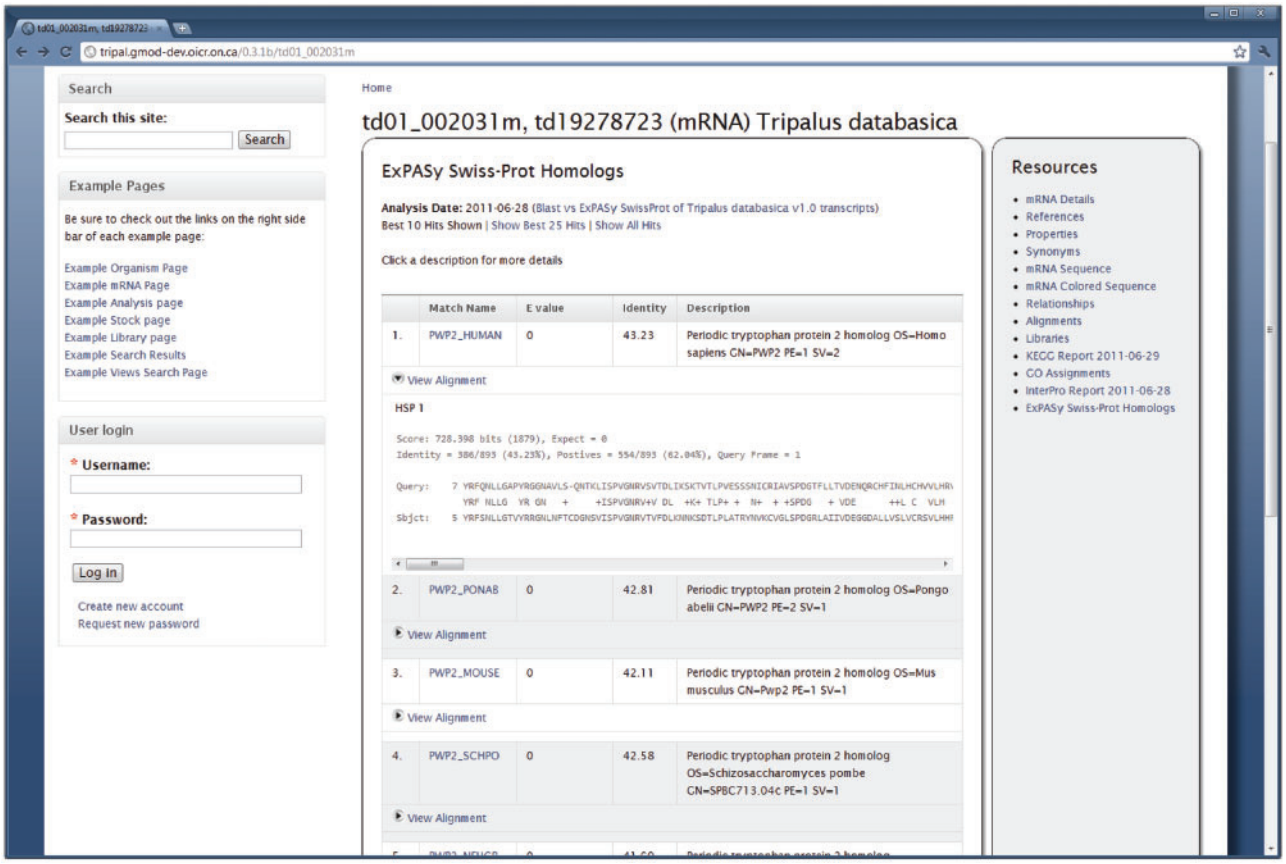

Figure 5. BLAST results for a feature. The Tripal Analysis BLAST module provides a summary of BLAST homology results on the Feature page. This example is for a feature that was compared to the ExPASy Swiss-Prot protein database. At first glance visitors view the match, e-value, percent identity and a description of the matching sequence. The actual alignment can be seen by clicking the 'View Alignment' link under each match. By default only the top 10 results are shown, however, visitors can choose to see the top 25 or all of the matches by clicking the appropriate links at the top of the report. The name given to the BLAST analysis is also visible at the top of the report. Visitors can find more information about this blast job by visiting the BLAST Analysis page by clicking the analysis name.

previously run job or view a job's details. Currently, the jobs management tool is intended primarily for data loading and content management, and is not suited for data analysis. Performing data analysis or processing of analytical workflows is better performed using tools such as Taverna (27), Ergatis (17) or Galaxy (28).

Materialized views were introduced in Chado as a means for speeding long running queries. Because Chado is highly normalized some complex queries can take more time than is acceptable by web site visitors. A materialized view is a database table pre-populated with results from an otherwise long-running query, which can then be queried simply and quickly. Tripal supports the use of materialized views and provides an administrative interface, as well as API functions for creating and updating these views.

In the layer above the Tripal Core are the Tripal Chado modules. In most cases these modules correlate directly with Chado table groups—one Tripal module per Chado table group. Currently, there are seven Tripal Chado modules. These include: the CV module for managing controlled vocabularies; the DB module for external database references; the Feature module for managing genomic features; the Library module for molecular libraries; the Organism module for managing organisms; the Stock module for biological stock collections; and the Analysis module, a generic module providing support for computational analyses. These modules provide a web-based mechanism for adding and updating new data such as a new organism, stock or analysis. They also provide an API of their own, such as for accessing data properties (e.g. stock properties). Most Chado modules also provide pages for viewing of Chado content. For example, Tripal provides pages for organisms, genomic features (i.e. sequences), stocks, libraries and analyses. Screenshots for some of these content pages were shown in the 'Description' section. Both the Tripal Core module and the Tripal Chado modules combine to form the base Tripal package.

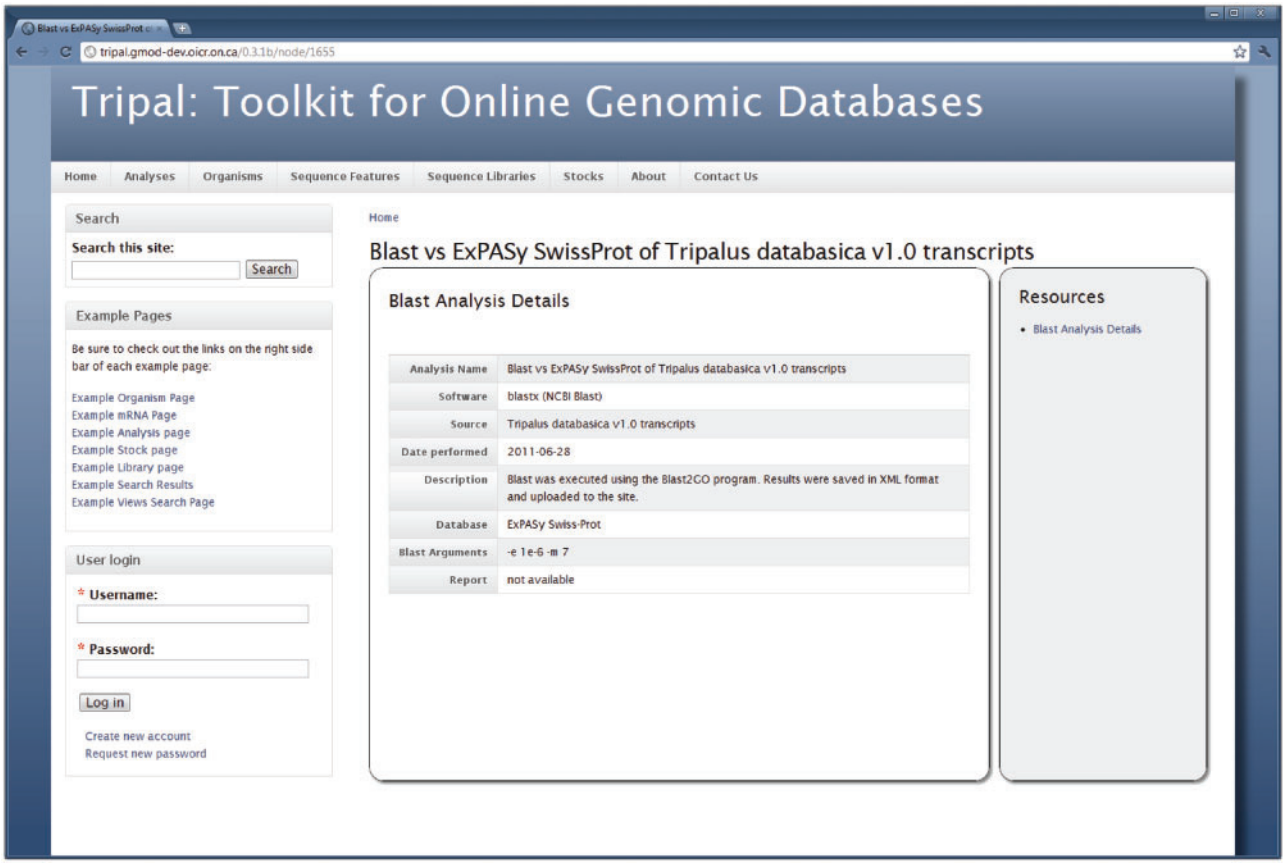

Figure 6. The Analysis page. When adding results from an analysis, content managers may provide specific details about the analysis including the software used, the date the analysis was performed, parameters used and the data source used. This example shows information for a BLAST homology analysis, specifically the BLAST homology results shown in Figure 5.

A layer above the Tripal Core and Chado modules are extension modules. These are modules that do not correlate directly with Chado table groups but provide additional functionality. For example, Chado can store analysis results for any type of computational tool. The Tripal Analysis module provides a generic framework and API for inserting and modifying any type of analysis in Chado. However, different analyses will require different methods for loading and visualization, therefore, the analysis module is designed to be extended. Examples of Tripal Analysis module extensions include the Analysis BLAST homology module that loads and displays XML results from the NCBI BLAST (19) tool; the Analysis InterPro module, that loads and displays XML results from the InterProScan (29) tool; the Analysis KEGG module, that loads and displays annotations from the KEGG Automated Annotation Service (KAAS) (18, 30) tool; and the Analysis GO module for displaying trees and charts for GO mappings (21) derived from other analyses [e.g. InterProScan or Blast2GO (31)]. In each of these cases different data parsers and methods for displaying results are required. Extension modules may provide additional functionality. For example, a Tripal GBrowse Management module is available that provides interaction between Tripal and GBrowse (14, 15, 32), including synchronization of data from Chado to a non-Chado GBrowse database. Extensions of Tripal created using the Tripal API from both the Core and the Chado modules can in turn be shared with others through a module repository at the Tripal website. Developers who create new Tripal extensions are welcome to submit their modules for inclusion in the repository. The extension modules are not included in the base Tripal package but can be downloaded separately from the extensions page of the Tripal website (http://tripal.sourceforge.net/?q=extensions). A list of Tripal modules, including several extension modules, with brief descriptions and their dependencies can be seen in Table 1. Any module that uses the Tripal API may be submitted for inclusion in the repository. Instructions for submitting new modules are available on the extensions page.

Tripal provides a mechanism for visualization of data within Chado and is intended to serve as a tool for community genomic databases. However, Tripal can also be used

Database, Vol. 2011, Article ID bar044, doi:10.1093/database/bar044

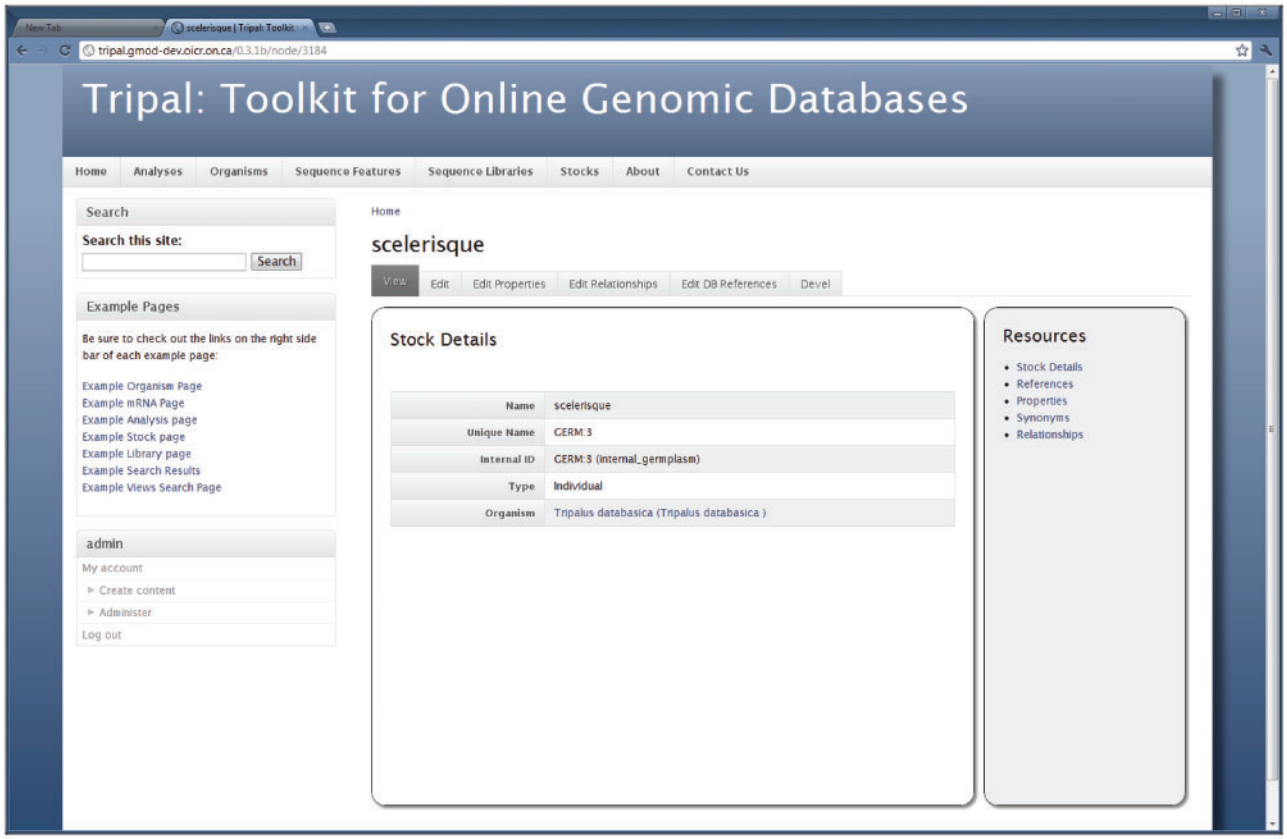

Figure 7. The Stock page. As shown, the Stock page is similar to the Organism and Feature pages in that a visitor first views basic details describing the stock. The right sidebar provides information for the given stock including references to the stock in other databases, additional properties for the stock (such as comments), synonyms and relationships with other stocks. The tabs above the stock details allow privileged users to edit the basic details and add, update or delete properties, references or relationships for the current stock. These tabs are not present for non-privileged visitors. As with other pages, the site administrator may tailor the look-and-feel of this page by editing the appropriate template files. In this example, the site administrator is logged on. A navigation box appears on the bottom-right hand side that provides administrative access to the site.

to build web-based applications. The application layer of the Tripal hierarchy is meant to categorize collections of modules that provide a specific application beyond just management and visualization of a Chado database. These applications could include not just the Tripal Core, Tripal Chado and other extension modules, but also other publicly available Drupal modules and custom built modules as well. A Tripal-based application would bundle all of these modules together to provide a tool or resource for a community database. An example of such an application could be a bundle of Tripal Chado modules, Tripal Extensions and Drupal Modules that together provide functionality for managing a breeding program or a sequencing center.

Data loaders

Previously, users were required to download and install the Chado Perl package, as well as, dependencies, such as BioPerl (23, 24), to install Chado into a PostgreSQL database (the recommended database for Chado). To simplify this process, a Chado loader is introduced into version 0.3.1b of Tripal. An administrative interface allows for installation of Chado within a few clicks, and populates the tables with basic data typically installed by the Perl-based installer. The Chado database is installed within the Drupal database, but within a separate schema. Installation of Chado within the Drupal database allows for integration with powerful Drupal features such as Views. However, Tripal also supports Chado when installed in a database external to Drupal. Such will be the case if Chado was installed using the Perl-based installers or existed prior to installation of Tripal. In this case, Tripal will experience some limitations.

Controlled vocabularies must be added to Chado before most other data. Using the Tripal ontology loader, users can specify and import a file in OBO format (http://www.geneontology.org/GO.format.obo-1_2.shtml), which can reside on the same server as Tripal, the local server, or on

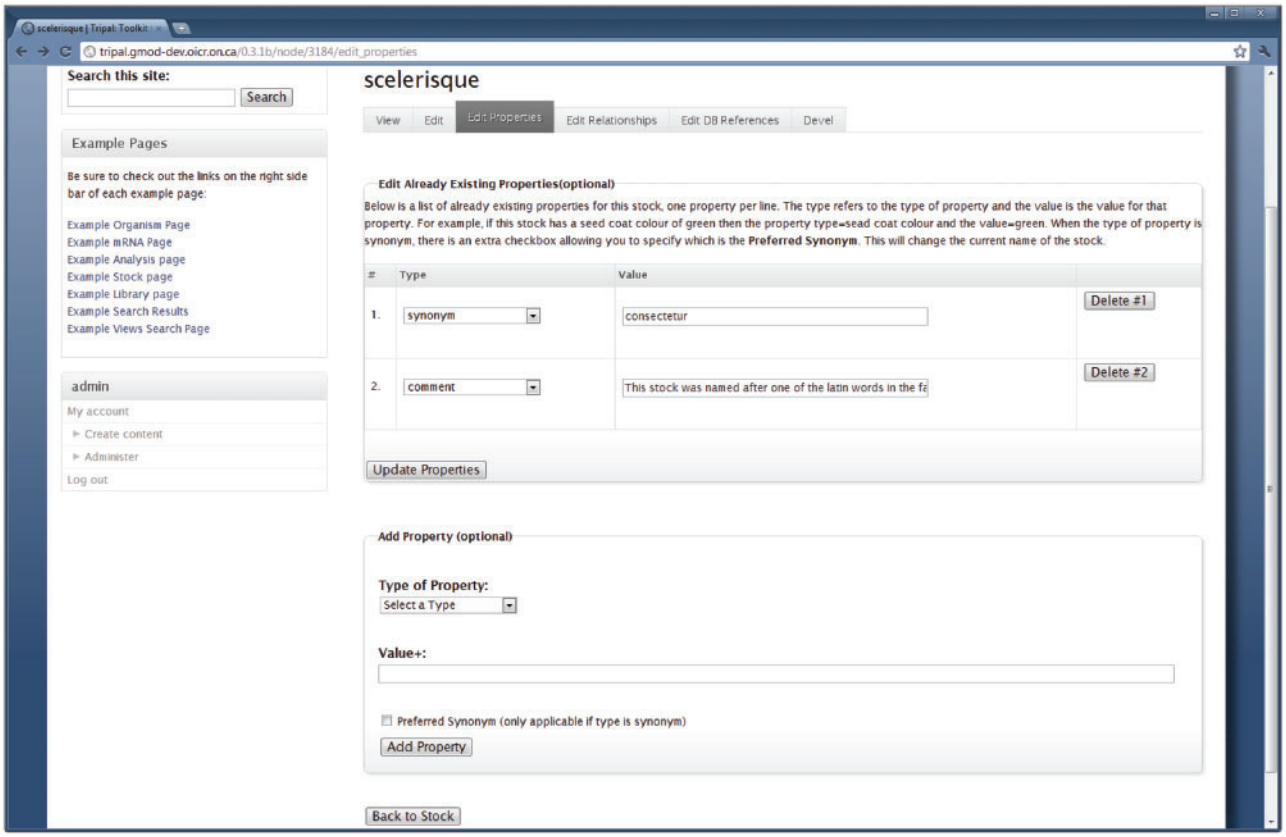

Figure 8. The interface for editing stock properties. Content managers with proper privilege may use this form to edit the properties of a stock. This example is for editing the stock shown in Figure 7, and is available by clicking the 'Edit Properties' tab at the top of the Stock page. The first fieldset lists all current properties for the stock and allows the content manager to change the property type or edit the text for the property. The content manager can also delete a property by clicking the 'Delete' button to the right side of the property. The second fieldset allows for adding new properties to the current stock. Editing relationships and references is done in a similar manner.

a remote server. The file is downloaded (if stored remotely), parsed and then loaded into the Chado database. The site administrator may return to update ontologies as regularly as desired. The OBO loader comes pre-populated with several common biological ontologies but any additional ontology added by the administrator will be saved for future updates.

Once populated with the necessary controlled vocabularies, administrators may then upload genomic features (sequences) using the FASTA and GFF3 loaders. The site administrator specifies the location of the file on the local server. The GFF3 loader will import genomic features, their synonyms (aliases), locations relative to others (alignments), external database cross-references, GO assignments and parent–child relationships. The residues for these genomic features can then be imported using the FASTA loader.

The various analysis extension modules also provide loaders for their respective data. As mentioned previously, the Tripal BLAST homology, KEGG and InterPro modules all provide an interface for loading results. The Tripal Analysis GO module provides a basic GO Annotation File (GAF) parser (http://www.geneontology.org/GO.format.gaf-2_0.shtml) for loading of GO terms associated with genomic features. The loader is not fully compliant with the standard, but will import GO assignments and can be useful for loading assignments derived from the Blast2GO (33) tool.

Integration with Drupal Views and Panels

As mentioned previously, Drupal maintains a repository of hundreds of user contributed extensions that anyone may download and use to add functionality to a Drupal website. Two popular extensions specifically supported by Tripal are Views (http://www.drupal.org/projects/views) and Panels (http://www.drupal.org/projects/panels). The Views module gives site administrators the ability to query tables, using a web-based graphical interface without

Database, Vol. 2011, Article ID bar044, doi:10.1093/database/bar044

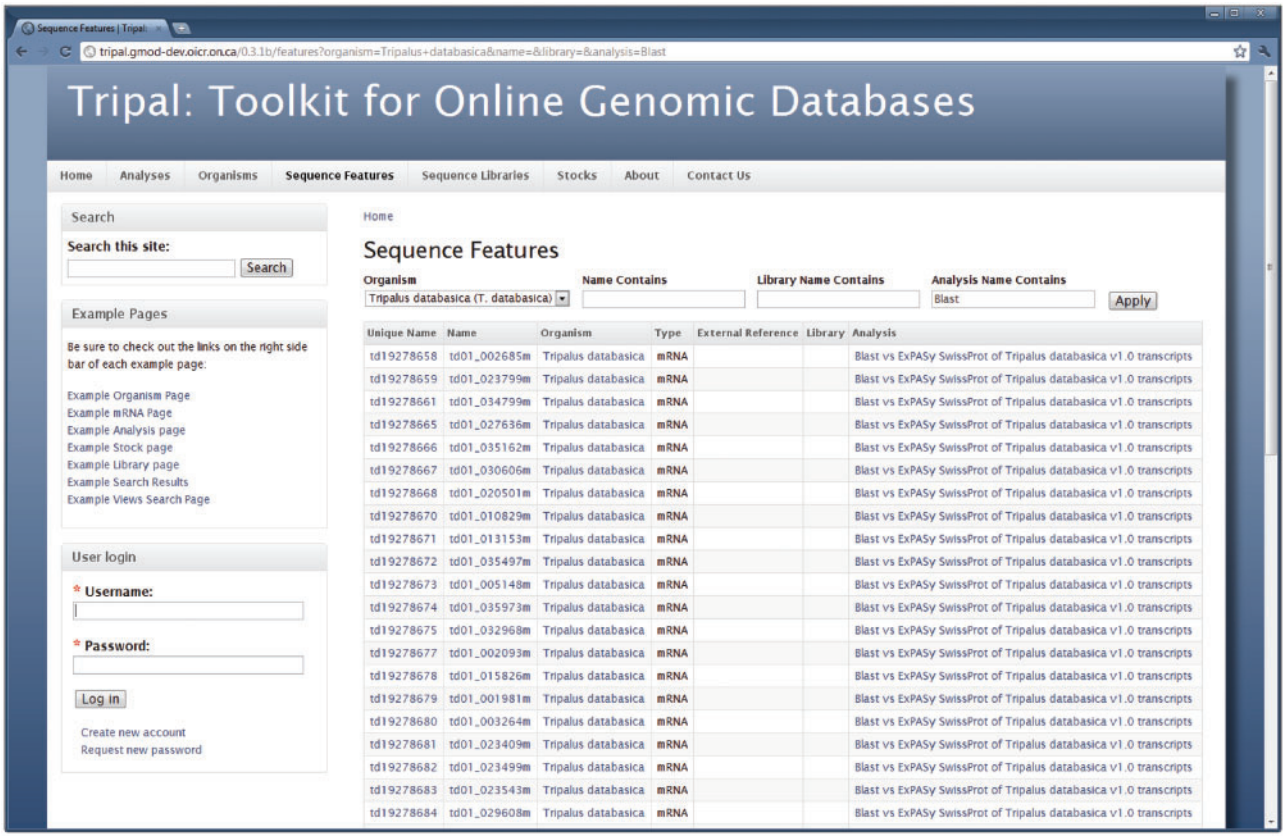

Figure 9. A Drupal Views query page. Tripal exposes the Chado database to the Drupal Views extension module. This allows site administrators to create custom lists and search pages without the need for SQL or programming. When the Views module is installed Tripal will provide several pre-existing Views. This example shows the 'Sequence Features' View, which is available by clicking the 'Sequence Features' link in the primary menu. This View allows site visitors to search for features (sequences) within the database by filtering results by a specific organism or feature name. Visitors may also limit results to features that belong to a specific library or analysis. Here features for our example organism that are associated with a blast analysis are shown.

knowing SQL, the programming language used to query a relational database. Tripal exposes the Chado database to the Views module. Unique lists, tables and even search pages can be created using the Views web interface allowing for custom filtering, sorting and displaying of Chado data. Thus site administrators can use Views to customize their site without scripting or programming. If the Tripal Chado loader is used to create the Chado database then these queries can be restricted to include only Chado data published on the site and for linking results to Tripal generated content pages. If Chado is installed separately then Views will function properly but will not allow filtering of non-published data or linking to Tripal pages.

Similarly, the Panels module allows site administrators to customize the layout of components on a page. Tripal exposes the various components of the organism, genomic feature, analysis, library and stock content pages for use by the Panels module. For example, if a layout is desired which is different from the default Tripal genomic feature page layout, a site administrator may use the Panels web interface to reorganize or exclude components on the feature page. The same is true for organism, analysis, library and stock pages. This can be useful if the user does not desire to manually edit the default Tripal template theme files to perform customizations. However, Panels cannot provide as fine-grained control as editing templates.

Searching capabilities

Currently, Tripal uses Drupal's full-text searching method to allow site visitors to find data of interest. Genomic features, organisms, stocks, libraries and analysis results are indexed by Drupal's searching mechanism. Words on these pages are extracted and ranked. Visitors who would like to find, for example, all genomic features for a specific biological process need only provide the name of the process in the search box. All pages in the Drupal site that contain the words specified will be presented as search results. This allows for an easy, all-in-one solution for searching terms that might be present in a BLAST homology match name,

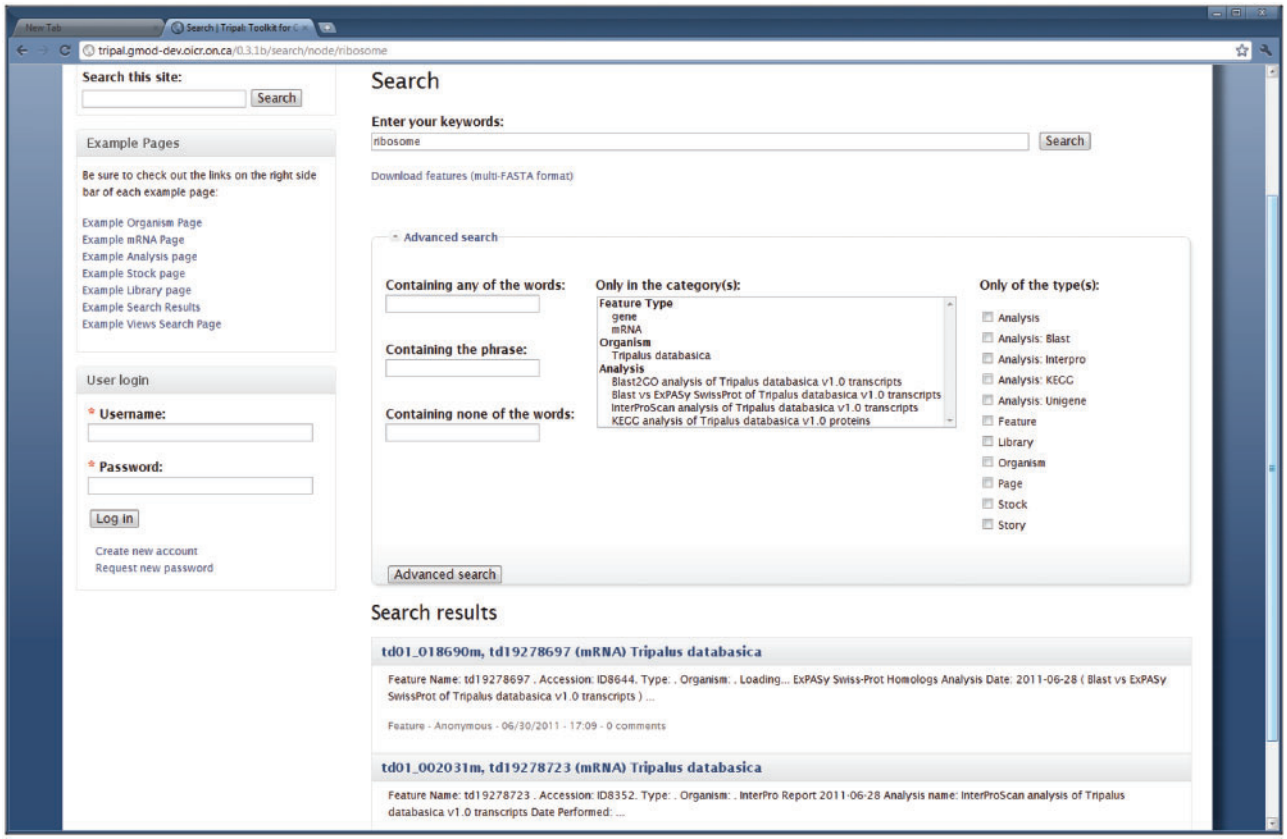

Figure 10. Full-text search page. Drupal provides a full-text searching mechanism which Tripal supports for searching Chado content. The search page can be found by using the search box in the top right corner on the example website. In this example a search for the term 'ribosome' yielded two mRNA results. An advanced search field set is also available and shown expanded. This allows visitors to filter searches by a specific organism, feature type, analysis or Drupal content type (shown in the list along the right hand side). Users can download a FASTA file of search results by clicking the 'Download features' link near the top of the search page.

InterPro domain, GO term, etc. An administrative option exists in Tripal that allows pages to be categorized by organism, feature type (SO term), library or analysis. This form of 'tagging' allows for advanced searching where site visitors can limit their searches to specific organisms, sequence ontology terms, analyses or libraries. Search results for features can be downloaded in FASTA format. One limitation for Drupal searching is that for large amounts of data, indexing of all pages can take several days to weeks. However, once indexing is completed, searching is reasonably fast.

Alternatively, with Tripal version 0.3.1b and the Views module, custom search pages for Chado data can be created. Drupal Views allows administrators to create custom displays of data, but to also create filters, which can be used to restrict the displayed results. These filters can be exposed to the site visitor, which will provide a web form for custom searching of the data. For communities requiring more fine-grained control of search parameters and display, the Chado database can be queried directly using a custom search module that could be built with the aid of the Tripal API. Such is the case for sites like the Cacao Genome Database and the Citrus Genome Database.

Future work

Currently, Tripal supports a subset of the Chado table groups. Only those described in this text are supported. Work is currently underway to broaden availability to other Chado table groups. Additionally, the current data loaders (GFF3, FASTA, OBO and GAF) are not sufficient to load all types of data into Chado. Under active development is a Chado bulk loader that will allow a site administrator to create templates for loading of any type of tab delimited or Excel file by authorized site users. In the future, it may be possible to use Tripal as a web front-end for data analysis workflow engines (such as Galaxy, Taverna or Ergatis) but currently no active work is underway to support analytical workflows. Additionally, an extension

Database, Vol. 2011, Article ID bar044, doi:10.1093/database/bar044

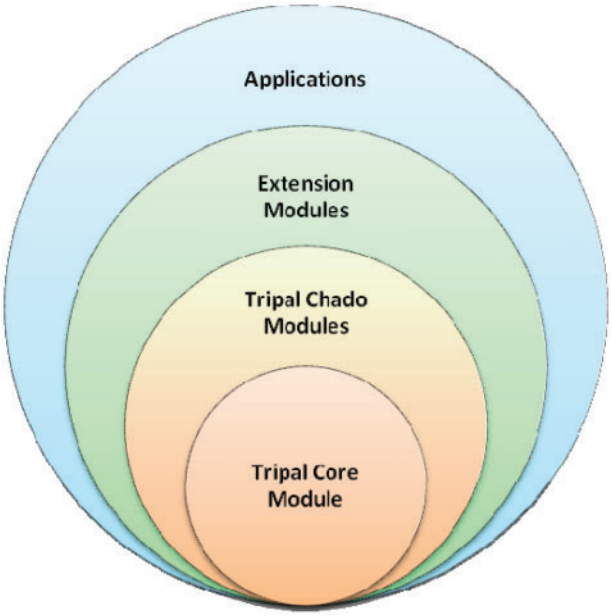

Figure 11. Tripal is designed in hierarchical layers. The baseTripal package consists of the Tripal core and Tripal Chado modules. Extension modules use the Tripal Core and Tripal Chado Modules and provide additional functionality not provided by the base package. Applications use a combination of the Core, Chado and extension modules, as well as, other Drupal modules.

module is planned for support of JBrowse (34), a javascript based genome browser. Tripal may be offered through the Drupal modules repository in the future, but is currently housed on a SourceForge project page.

The default Drupal full-text searching will return results where the search term appears anywhere on the page. This may be too inclusive, as in the case where a user wants to see only coding sequences annotated with a specific GO term, and not where the term may appear anywhere on the page (such as in a blast result). Work is currently underway to provide a set of search tools using the Drupal Views interface to be included as part of the Tripal core package. Once complete these views will allow for custom direct searching of Chado tables and content.

Acknowledgements

Special thanks are extended to Scott Cain and Dave Clements of the GMOD project team for logistical support and consultation, as well as, to the Marine Genomics Group at the Hollings Marine Lab and Clemson's Cyberinfrastructure and Technology Integration (CITI) Group who provided assistance with early developments of Tripal for the Marine Genomics Project.

Table 1. Tripal modules

Tripal module name	Chado table group[a]	Purpose	Tripal dependencies[b]
Tripal main package			
Core	NA	Provides generic support and API for all other modules.	None
DB	General	Provides an interface for management of external databases where data may also be referenced. Ontologies, features, libraries and analyses and more all may have external references.	Core
CV	Controlled vocabulary	Provides an interface for management of controlled vocabularies, including editors and bulk loaders for ontologies.	Core, DB
Organism	Organism	Exports data about organisms from the Chado database, as well as, a generic organism page for viewing of content.	Core
Feature	Sequence	Exports data about features, feature properties, feature locations on reference features, feature synonyms (aliases) and a generic feature page.	Analysis, organism
Library	Library	Exports data about molecular libraries, their properties features associated with these libraries, and a generic library page.	Organism, feature
Stock	Stock	Exports data about biological stocks, stock properties including synonyms, external database references, relationships and a generic stock page.	Organism
Analysis	Companalysis	Provides generic support and API for all analysis modules.	Organism

(Continued)

Table 1. Continued

Tripal module name	Chado table group[a]	Purpose	Tripal dependencies[b]
Currently available extension modules			
Analysis blast	NA	Exports blast results for features. Results appear on feature pages.	Analysis, feature
Analysis KEGG	NA	Exports KEGG results imported from a KAAS results hier file. Results appear on feature pages and organism pages.	Analysis, organism, feature
Analysis GO	NA	Provides GO reports on organism pages for any analyses that track GO assignments (e.g. the tripal_analysis_interpro module).	Analysis, organism, feature,
Analysis InterPro	NA	Exports and formats InterProScan XML or HTML results on the feature pages and organism page.	Analysis, feature
GBrowse Management	NA	Provides an interface for managing GMOD GBrowse instances including creating and deleting instances, loading feature data stored in Chado into a non-Chado GBrowse backend and a Drupal GBrowse page including display of a GBrowse instance within Tripal.	Analysis, feature, library

[a]Tables in Chado are organized into groups of related tables called modules. To reduce confusion we refer to these Chado modules as Chado table groups.
[b]Dependencies are inherited, for example, analysis modules are dependent on the Tripal Analysis module, but also on its dependencies. NA, Not applicable.

Funding

This work was supported by the informal collaboration between various institutions with an overlapping mission to create an easy to manage Drupal-based biological website. No single funding source is responsible. However, individual who have contributed to Tripal received funding from various sources including the Clemson University Genomics Institute (CUGI), the National Science Foundation (NSF) funded 'Tool Development for the Fagaceae' project (Grant number #0605135); Clemson's Cyberinfrastructure and Technology Integration group (CITI); a United States Department of Agriculture (USDA) Specialty Crops Research Initiative grant: 'Tree Fruit GDR: Translating Genomics into Advances in Horticulture' (2009–13) (Grant number #2009-51181-06036); an Agriculture and Agri-Food Canada (AAFC) Grant: 'Purenet' under the Agricultural Bioproducts Innovation Program (ABIP) (2009–11); a grant from the Saskatchewan Pulse Growers Association: 'iMAP' (2010–13) (Grant number BRE1010); a USDA Specialty Crops Research Initiative grant: 'RosBREED' (2009–13) (Grant number #2009-51181-05808); and a USDA Agriculture Research Service (USDA-ARS) specific cooperative agreement award: 'Development of a Cacao Genome Database' (2009–13) (Grant number #5866319200).

References

1. Drysdale,R. (2008) FlyBase : a database for the Drosophila research community. *Methods Mol. Biol.*, **420**, 45–59.
2. Harris,T.W., Antoshechkin,I., Bieri,T. *et al*. (2010) WormBase: a comprehensive resource for nematode research. *Nucleic Acids Res.*, **38**, D463–D467.
3. Hirschman,J.E., Engel,S., Hong,E. *et al*. (2007) The Saccharomyces Genome Database provides comprehensive information about the biology of S-cerevisiae and tools for studies in comparative genomics. *FASEB J.*, **21**, A264–A264.
4. Jung,S., Staton,M., Lee,T. *et al*. (2008) GDR (Genome Database for Rosaceae): integrated web-database for Rosaceae genomics and genetics data. *Nucleic Acids Res.*, **36**, D1034–D1040.
5. Jung,S., Jesudurai,C., Staton,M. *et al*. (2004) GDR (Genome Database for Rosaceae): integrated web resources for Rosaceae genomics and genetics research. *BMC Bioinformatics*, **5**, 130.
6. McKillen,D.J., Chen,Y.A., Chen,C. *et al*. (2005) Marine genomics: a clearing-house for genomic and transcriptomic data of marine organisms. *BMC Genomics*, **6**, 34.
7. Jaiswal,P. (2011) Gramene database: a hub for comparative plant genomics. *Methods Mol. Biol.*, **678**, 247–275.
8. Lawrence,C.J., Dong,Q., Polacco,M.L. *et al*. (2004) MaizeGDB, the community database for maize genetics and genomics. *Nucleic Acids Res.*, **32**, D393–D397.
9. Lawrence,C.J., Harper,L.C., Schaeffer,M.L. *et al*. (2008) MaizeGDB: the maize model organism database for basic, translational, and applied research. *Int. J. Plant Genomics*, **2008**, 496957.
10. Wegrzyn,J.L., Lee,J.M., Tearse,B.R. *et al*. (2008) TreeGenes: a forest tree genome database. *Int. J. Plant Genomics*, **2008**, 412875.
11. Bombarely,A., Menda,N., Tecle,I.Y. *et al*. (2011) The Sol Genomics Network (solgenomics.net): growing tomatoes using Perl. *Nucleic Acids Res.*, **39**, D1149–D1155.
12. Mungall,C.J. and Emmert,D.B. (2007) A Chado case study: an ontology-based modular schema for representing genome-associated biological information. *Bioinformatics*, **23**, i337–i346.

Database, Vol. 2011, Article ID bar044, doi:10.1093/database/bar044

13. Zhou,P., Emmert,D. and Zhang,P. (2006) Using Chado to store genome annotation data. *Curr. Protoc. Bioinformatics*, **Chapter 9**, Unit 9.6.
14. Donlin,M.J. (2007) Using the Generic Genome Browser (GBrowse). *Curr. Protoc. Bioinformatics*, **Chapter 9**, Unit 9.9.
15. Stein,L.D., Mungall,C., Shu,S.Q. *et al*. (2002) The Generic Genome Browser: A building block for a model organism system database. *Genome Res.*, **12**, 1599–1610.
16. Lewis,S.E., Searle,S.M., Harris,N. *et al*. (2002) Apollo: a sequence annotation editor. *Genome Biol.*, **3**, RESEARCH0082.
17. Orvis,J., Crabtree,J., Galens,K. *et al*. (2010) Ergatis: a web interface and scalable software system for bioinformatics workflows. *Bioinformatics*, **26**, 1488–1492.
18. Kanehisa,M., Araki,M., Goto,S. *et al*. (2008) KEGG for linking genomes to life and the environment. *Nucleic Acids Res.*, **36**, D480–D484.
19. Altschul,S.F., Madden,T.L., Schaffer,A.A. *et al*. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
20. Apweiler,R., Attwood,T.K., Bairoch,A. *et al*. (2001) The InterPro database, an integrated documentation resource for protein families, domains and functional sites. *Nucleic Acids Res.*, **29**, 37–40.
21. Ashburner,M., Ball,C.A., Blake,J.A. *et al*. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.
22. O'Connor,B.D., Day,A., Cain,S. *et al*. (2008) GMODWeb: a web framework for the Generic Model Organism Database. *Genome Biol.*, **9**, R102.
23. Stajich,J.E., Block,D., Boulez,K. *et al*. (2002) The Bioperl toolkit: Perl modules for the life sciences. *Genome Res.*, **12**, 1611–1618.
24. Stajich,J.E. (2007) An introduction to BioPerl. *Methods Mol. Biol.*, **406**, 535–548.
25. Papanicolaou,A. and Heckel,D.G. (2010) The GMOD Drupal bioinformatic server framework. *Bioinformatics*, **26**, 3119–3124.
26. Thain,D., Tannenbaum,T. and Livny,M. (2005) Distributed computing in practice: the Condor experience. *Concurr. Comput. Pract. Exp. Grid Perform.*, **17**, 232–356.
27. Oinn,T., Addis,M., Ferris,J. *et al*. (2004) Taverna: a tool for the composition and enactment of bioinformatics workflows. *Bioinformatics*, **20**, 3045–3054.
28. Goecks,J., Nekrutenko,A. and Taylor,J. (2010) Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol.*, **11**, R86.
29. Zdobnov,E.M. and Apweiler,R. (2001) InterProScan–an integration platform for the signature-recognition methods in InterPro. *Bioinformatics*, **17**, 847–848.
30. Moriya,Y., Itoh,M., Okuda,S. *et al*. (2007) KAAS: an automatic genome annotation and pathway reconstruction server. *Nucleic Acids Res.*, **35**, W182–W185.
31. Conesa,A., Gotz,S., Garcia-Gomez,J.M. *et al*. (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics*, **21**, 3674–3676.
32. Donlin,M.J. (2009) Using the Generic Genome Browser (GBrowse). *Curr. Protoc. Bioinformatics*, **Chapter 9**, Unit 9.9.
33. Conesa,A. and Gotz,S. (2008) Blast2GO: a comprehensive suite for functional analysis in plant genomics. *Int. J. Plant Genomics*, **2008**, 619832.
34. Skinner,M.E., Uzilov,A.V., Stein,L.D. *et al*. (2009) JBrowse: a next-generation genome browser. *Genome Res.*, **19**, 1630–1638.