NEW TRIPAL MODULES: ELASTICSEARCH AND EXPRESSION

Margaret Staton
Open source content management system (CMS) for biological data

- Specializing in genetic, genomic, breeding, etc.

97 sites report using Tripal!

Benefits:
- Reduces IT costs
- Publishes simple genome sites out-of-the-box
- Provides an API for complete customization
- Uses Chado and community ontologies for standardization
- Allows for sharing of extensions between sites
Extension Module System

• Core modules
  • Developed by core team or vetted by core team
  • Likely to be needed/appreciated by all Tripal sites
  • Well tested

• Extension modules
  • Anyone can contribute extra functionality
  • Take ‘em or leave ‘em - If you don’t need it, it doesn’t clutter your server

Extension modules can eventually be integrated with core.

http://tripal.info

Tripal Extensions
The following is a list of extensions that have been provided by Tripal site developers and which can be imported into another Tripal site. If you have custom extensions you would like to share, please let us know by joining the Tripal mailing list and sending a request. Consider adding your extensions even if only in development state as this encourages collaboration and reduces duplication of effort. Use the fields below to filter.

<table>
<thead>
<tr>
<th>Extension Type</th>
<th>Tripal</th>
<th>Chado</th>
<th>Module Status</th>
<th>Categories</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Any -</td>
<td>$</td>
<td>$</td>
<td></td>
<td>$</td>
</tr>
<tr>
<td>v1.0</td>
<td>v1.1x</td>
<td>v1.2x</td>
<td>No longer supported in Development</td>
<td>- Any -</td>
</tr>
<tr>
<td>v1.1</td>
<td>v1.3x</td>
<td></td>
<td>Needs debugging Released</td>
<td></td>
</tr>
<tr>
<td>v2.0</td>
<td></td>
<td></td>
<td>Needs Review Released</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Ready for Use</td>
<td></td>
</tr>
</tbody>
</table>

ND Genotypes (create marker/stock) Bulk Loader Template

Bulk Loader Template
This template provides a means of loading marker by germplasm genotype matrices (See GenotypeSampleData.txt for an example) into Chado by creating Natural Diversity Experiments for each Genotype (ie: each element in the matrix) and linking it to the corresponding stock and marker. &nbsp;This module requires the sequence and relationship ontologies.

Breeding API Extension Module

Extension Module
Breeding API (https://github.com/plantbreeding) implementation for Tripal. Currently not released and in heavy development. This project has been sponsored by the Bill and Melinda Gates Foundation which funded the breeding API hackathon in June 2015 in Seattle.

http://github.com/tripal
HARDWOOD GENOMICS PROJECT
Welcome to the Hardwood Genomics Project

We house transcriptome and genome resources for hardwood trees.

Interested in contributing data? Please contact us!

This website was originally developed for the NSF grant Comparative Genomics of Environmental Stress Responses in North American Hardwoods, which has now ended. We are expanding to a larger set of tree genomic resources with the support of the NSF grant CIF21 DIBBS: Tripal Gateway, a Platform for Next-Generation Data Analysis and Sharing (PI Stephen Ficklin, Washington State University). Species and resources currently available:

- Chinese Chestnut (draft reference genome, physical map, genetic map)
- American Beech (transcriptome sequencing)
- American Chestnut (transcriptome sequencing)
- Black Cherry (low coverage genome sequence, SSRs)
- Black Walnut (transcriptome sequencing, SSRs, BAC, QTL Map, Reference Populations, BAC Sequencing)
- Blackgum (transcriptome sequencing, SSRs)
- European Chestnut (transcriptome sequencing)
Hardwood Genomics - Data

• Built with a grant for hardwood tree genomics (2010, PI Carlson)

• Seedling stress testing to mimic climate change:
  • Ozone, heat, cold, drought, wounding

• Transcriptome sequencing for 8 species
  • Libraries from diversity of tissue types
  • Libraries from abiotic stress treatments

• Genetic mapping populations for 6 species

• Molecular marker development for 12 species
  • (ranging from in silico only to laboratory confirmed)

• Genetic mapping for 4 species

• QTL mapping for 2 species
Hardwood Genomics - History

• Incorporates data from the original Fagaceae Genomics Web (built to house data from NSF grant to develop Fagaceae family genomic resources, 2007, PI Sederoff)

• Continuing interest in Chestnut genomics (Forest Health Initiative, The American Chestnut Foundation, USDA)
## Hardwood Genomics - Data

- **Genome**
- **Genetic map**
- **Physical map**

<table>
<thead>
<tr>
<th>Common Name</th>
<th>Transcriptome</th>
<th>RNASeq Biomaterial records</th>
<th>Expression Analysis</th>
<th>Predicted SSRs (transcriptomic)</th>
<th>Predicted SSRs (genomic)</th>
<th>Polymorphic SSRs</th>
<th>Population Description</th>
<th>BAC Libraries</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sugar Maple</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>White Alder</td>
<td>x</td>
<td>o</td>
<td></td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Red Alder</td>
<td>x</td>
<td>o</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Japanese Chestnut</td>
<td>x</td>
<td>o</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>American Chestnut</td>
<td>x</td>
<td>o</td>
<td></td>
<td></td>
<td></td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chinese Chestnut</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>European Chestnut</td>
<td>x</td>
<td>o</td>
<td></td>
<td></td>
<td></td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>American Beech</td>
<td>x</td>
<td>o</td>
<td></td>
<td></td>
<td>x</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>White Ash</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>Green Ash</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td>x</td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>Honeylocust</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td>x</td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>Black Walnut</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>English Walnut</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>American Sweetgum</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td>x</td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>Tulip Poplar</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>Blackgum</td>
<td>x</td>
<td></td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Redbay</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>Black Cherry</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>White Oak</td>
<td>x</td>
<td>o</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>Pedunculate Oak</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Northern Red Oak</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td>x</td>
</tr>
</tbody>
</table>
Hardwood Genomics - Tools

• Jbrowse
• Apollo
• CMap
• BLAST

• Symap – need to replace, any ideas?
Ongoing work - DIBBS

Tripal Gateway Project

NSF Data Information Building Blocks (DIBBs) grant
- Award #1443040
- PI Stephen Ficklin, Washington State University
- 3 years (1.5 years in)

Three components:
- RESTful Web services for Tripal sites
  - Allow sites to exchange data
- Integration with Galaxy
  - Allow sites to provide next-generation sequence analysis tools
- Improve data transfer
  - Big Data Smart Socket Client (BDSS) available
  - Explore Software Defined Networking (SDN)
Upcoming work - PGRP

- Dorrie Main, WSU (PI)
- Ontologies
  - Structure, trait, phenotypic quality and environment
  - Curation of current data
  - Standardize data collection in the future
  - Standardize data submission for users
- Communication between sites
  - Web services
    - Tripal extension module for cross-site querying - enabling a user to collate or view data from multiple Tripal sites
- Better querying and visualization of complex phenotype, genotype, and environment data
- Online educational modules, training courses, and developer/user support for Tripal
ELASTICSEARCH
What problem is being solved?

- Drupal internal search
  - Easy to set up and customize (for normal Drupal data types)
  - No native support for external DBs
  - Slow to index, slow to return results

- Need a solution that will:
  - Access chado database
  - Provide flexible and customizable indexing – index only what is needed, not everything
  - Scales to very large biological data sets
Elasticsearch Software

• distributed, open source search and analytics engine

• Massively distributed – can scale horizontally
• Multitenancy – a search cluster can manage many individual indices that can be queried individually or as a group
• Built on Apache lucene -> autocomplete, fuzzy searching, “did you mean” suggestions
• Document oriented – export database tables as JSON
• RESTful API can be leveraged with JSON over HTTP
• Open source
Elasticsearch Module

Install Elasticsearch

Install Tripal Elasticsearch Module

Index Drupal nodes

Index targeted Chado or Drupal tables

Site-wide search

Customized search

Transcripts

- Any organism
- blast hit description (Ex: [pectinesterase inhibitor])
- transcript unique name (Ex: 3.40-mRNA-1, [contig1187], [comp22850_c0_seq5 OR contig1187])

Search Transcripts
Create a view with the materialized view table.

**View Name** *

search_features_all

Please enter the name for this materialized view.

**MView Description**

This view joins feature uniquenames to BLAST hit information (description, e-value, and hit score) and organism information (genus, species, common_name).

Optional. Please provide a description of the purpose for this materialized view.
Elasticsearch Module - Example

Create a view with the materialized view table.

```
Query *

hit_best_eval, b.hit_best_score AS hit_best_score, o.common_name AS common_name, o.genus AS genus, o.species AS species
FROM
chado.feature f
INNER JOIN chado.blast_hit_data b ON b.feature_id = f.feature_id
INNER JOIN chado.organism o ON f.organism_id = o.organism_id
```

Please enter the SQL statement used to populate the table.
Elasticsearch Module - Example

```json
Schema Array
array {
  'description' => 'This view joins feature uniqueness to BLAST hit information (description, e-value, and hit score) and organism information (genus, species, common_name).',
  'table' => 'search_features_all',
  'fields' => array {
    'unique_name' => array {
      'type' => 'text',
      'not null' => true,
    },
    'hit_description' => array {
      'type' => 'text',
      'not null' => true,
    },
    'hit_best_eval' => array {
      'type' => 'text',
      'not null' => true,
    },
    'hit_best_score' => array {
      'type' => 'text',
      'not null' => true,
    },
    'common_name' => array {
      'type' => 'text',
      'not null' => true,
    },
    'genus' => array {
      'type' => 'text',
      'not null' => true,
    },
    'species' => array {
      'type' => 'text',
      'not null' => true,
    },
  },
}
```

After describing, populate.
Elasticsearch Module - Example

Now to ElasticSearch admin.

Select the materialized view to index (or any other table).

Select the fields.
Elasticsearch Module - Example

Queue UI and ultimate cron are dependencies. You can check on cron jobs and run them in parallel. This is convenient if you have many processors and are on a dev server.
Elasticsearch Module - Example

Move to demo...
Elasticsearch Module

Future Development

• Edit the search form – change field labels, type of search field (dropdown, checkboxes), order of fields
• Paths - Are all fields easily accessed by URLs? Automate discovery of URL links for datatypes?
• Fasta file download for feature table (include script)
• Multisite installs – use the flexibility of elasticsearch
• Scale with bigger data and different types of data
• Port to Tripal 3.0 and compare to new internal searching
EXPRESSION MODULE
What problem is being solved?

Need a better way to store and visualize RNASeq differential gene expression experiments.
Expression Module

Module to display expression data collected from RNASeq

We have left open the possibility for microarray expression data sources as well, currently untested.

Chado Tables/Modules used:

- MAGE
- Organism
- Contact
- Sequence
- Companalysis modules.
Content Types

- Tripal content types are created for these tables:
  - Biomaterial
    - Similar to NCBI BioSample and SRA
    - We do not differentiate between samples and libraries
  - Array design
    - Can be used for microarray data, but not used for NGS projects
  - Protocol
    - Define protocols for the experimental analysis
- New Chado analysis content type:
  - Analysis: Expression.
Loading Data

- Import biomaterial
  - BioSample data downloaded from NCBI (xml)
  - Flat file format (based on NCBI biomaterial bulk load form)

- Import expression values
  - (assumed to be normalized, features must already exist)
  - Individual file per sample
  - Tab delimited file with gene rows, sample columns
Visualization

- Demo…
Future Work on Expression Module

• Biomaterials
  • Upload SRA records from NCBI automatically via web services
  • Link the properties to ontologies
  • Link to individual analyses (currently only displays as associated with an organism)
    • IE – A transcriptome is built from a subset of biomaterials

• Expression
  • Allow user to provide a list of genes (cart system) and generate heatmap for all
  • Add significance/p-values from differential gene expression test results
    • Important functional data
    • Aid searching – limit results only to genes that respond to cold stress
Acknowledgements

University of Tennessee
- Ming Chen
- Nathan Henry

Washington State University
- Stephen Ficklin

University of Saskatchewan
- Lacey Anne Sanderson

All the developers of Tripal