

OUR DRUPAL/CHADO INTEGRATION EFFORTS

Stephen Ficklin
Clemson University Genomics Institute
Bioinformatics Group

GMOD Annual Meeting, January 15-16 2009.

Who we are and our development partnerships

CUGI: Service & Research

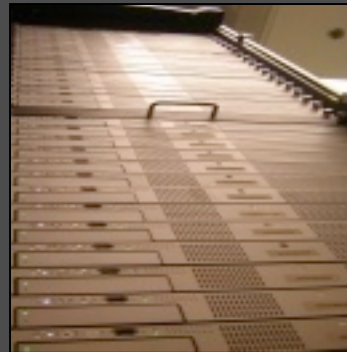
- ◉ Clemson University Genomics Institute
- ◉ <http://www.genome.clemson.edu/>



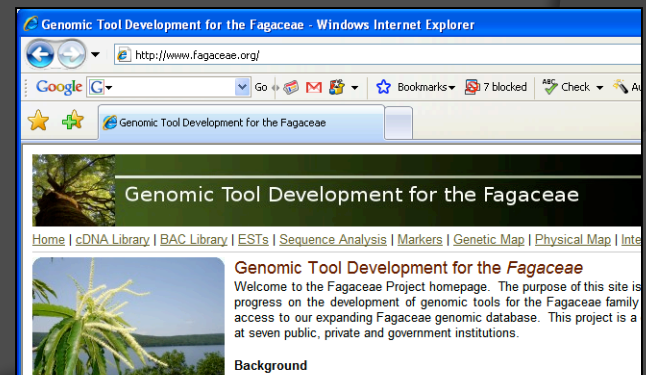
Field Work,
Sample
Collection



Resource
Construction
& Data
Generation



Data Analysis &
Data Management



Data Dissemination
& Tool Development



HML – Marine Genomics Group

- ⦿ <http://www.hml.noaa.gov/>
- ⦿ HML: NOAA Federal Facility
- ⦿ Partnership w/
 - Medical University of South Carolina
 - College of Charleston
 - SC DNR
 - NIST



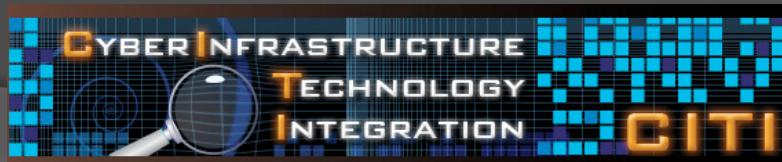
Main Bioinformatics Lab @ WSU

- www.bioinfo.wsu.edu
- Bioinformatics Group
- Computational facilities
- Database development: GDR & original developers of CMD and others....
- Algorithm development
- Comparative genomics
- Genomic annotation



CITI - Cyberinfrastructure

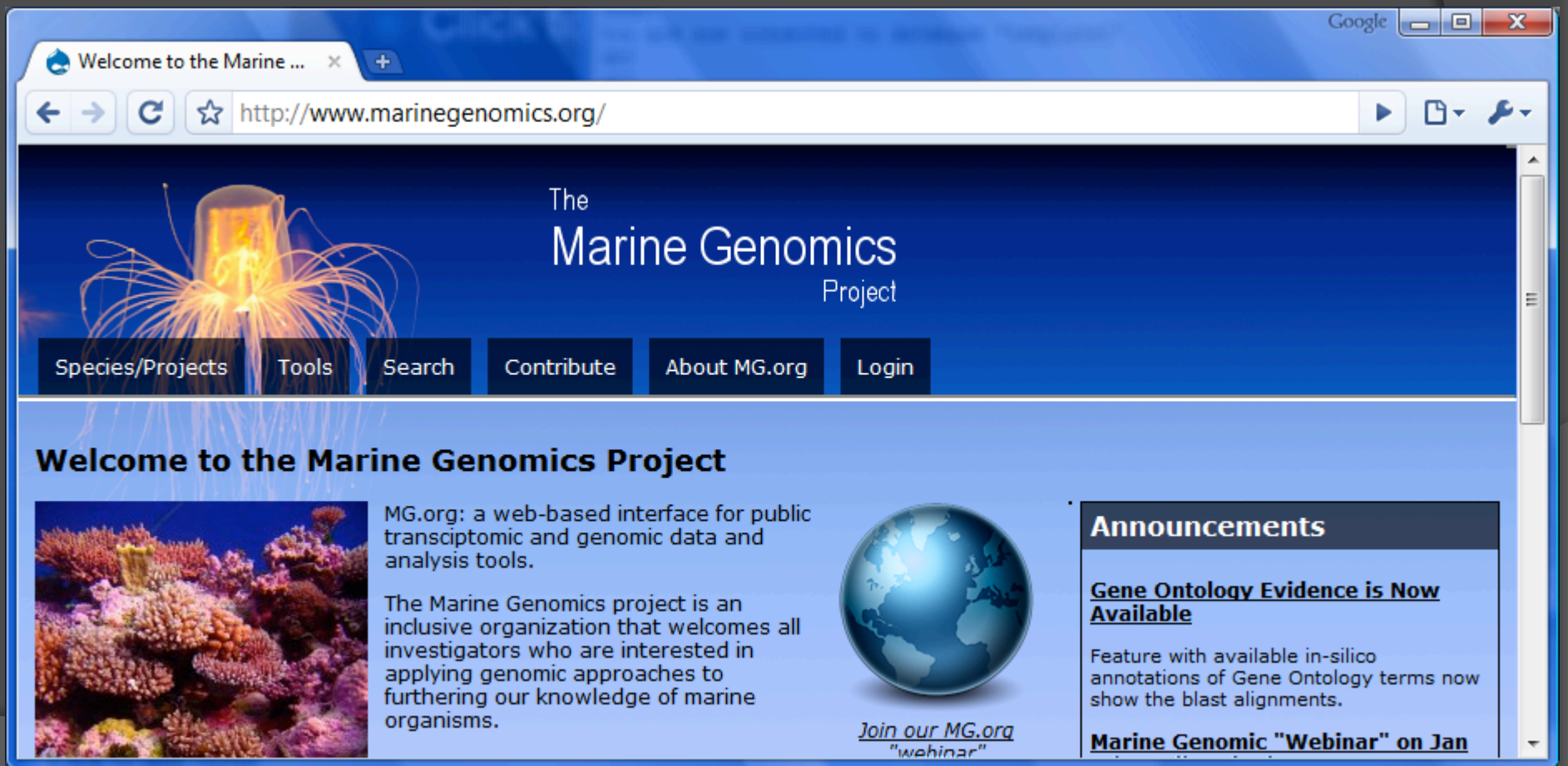
- <http://citi.clemson.edu/>
- CITI: Clemson Cyberinfrastructure and Technology integration group
- Research arm of Clemson IT group
- “Palmetto” cluster: ranked 60th in the world
- State-of-the-art data center
 - Redundant cooling, networking, power
 - Offsite duplicated data center
 - Two routes for every
- 364/24/7 monitoring
- Staff dedicated assistance



Our Collective Database Projects

Marine Genomics Project

- Developed by CUGI, CITI & HML
- Drupal front-end
- Chado backend



The screenshot shows a web browser window with the URL <http://www.marinegenomics.org/>. The page features a blue header with a glowing jellyfish image on the left and the text "The Marine Genomics Project" on the right. Below the header is a navigation menu with buttons for "Species/Projects", "Tools", "Search", "Contribute", "About MG.org", and "Login". The main content area is titled "Welcome to the Marine Genomics Project" and includes a coral reef image, a globe, and several text blocks. The text blocks describe the project's purpose and provide information about Gene Ontology evidence and a webinar.


Welcome to the Marine ... x +

← → ↻ ☆ <http://www.marinegenomics.org/> ▶ ⌵ ⚙


The
Marine Genomics
Project

Species/Projects Tools Search Contribute About MG.org Login

Welcome to the Marine Genomics Project

 MG.org: a web-based interface for public transcriptomic and genomic data and analysis tools.

The Marine Genomics project is an inclusive organization that welcomes all investigators who are interested in applying genomic approaches to furthering our knowledge of marine organisms.



*Join our MG.org
"webinar"*

Announcements

Gene Ontology Evidence is Now Available

Feature with available in-silico annotations of Gene Ontology terms now show the blast alignments.

Marine Genomic "Webinar" on Jan

Fagaceae Genomics Web

- Developed by CUGI
- NSF funded, collaborations with Dendrome Project
- Currently using GMODWeb, switching to chado/drupal



The screenshot shows a web browser window with the address bar displaying <http://www.fagaceae.org/web/db/index>. The page features a header with a green leaf image and the title "Fagaceae Genomics Web" with the subtitle "genomic tools for chestnut, oak, beech, and other trees." Below the header is a navigation menu with links: [Homepage](#), [Search](#), [cDNA Libraries](#), [BAC Libraries](#), [ESTs](#), [Markers](#), [Genetic Maps](#), [Physical Maps](#), and [Integrated Maps](#). The main content area is divided into two columns. The left column, titled "Project Information", contains a list of links: [Project Phases and Objectives](#), [Outreach](#), [Progress Reports](#), [Publications](#), [Photo Gallery](#), [Project Team](#), and [Links](#). The right column, titled "Project Description", contains a welcome message: "Welcome to the Fagaceae Project homepage. The purpose of this site is to provide background and up progress on the development of genomic tools for the Fagaceae family of trees, and to provide centra access to our expanding Fagaceae genomic database. This project is a collaborative effort among scie at seven public, private and government institutions." Below this text is a section titled "Background".

Coral Microbes Project

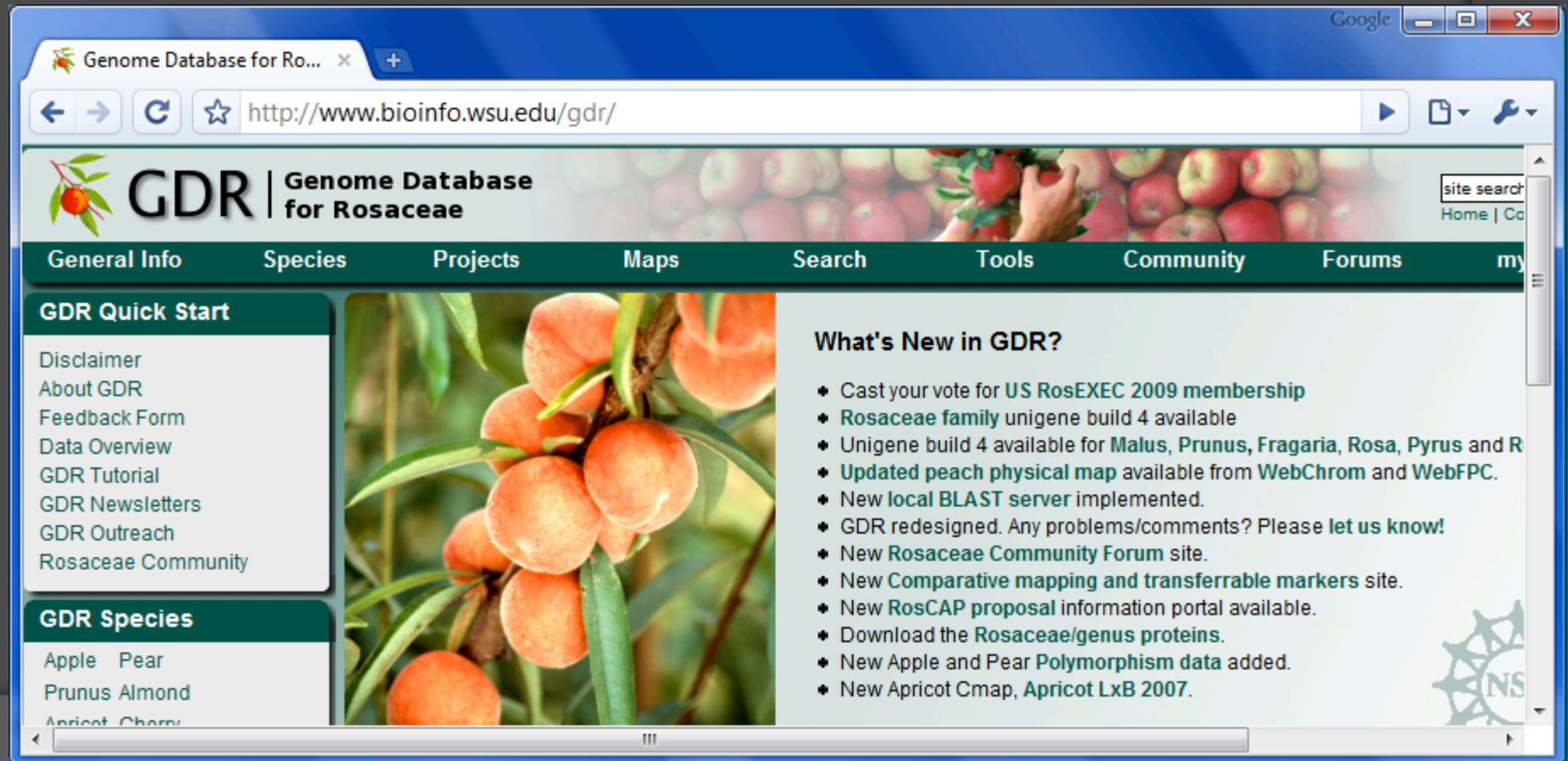
- Developed by Morris Lab @ MUSC/HML – CUGI/CITI consultation
- Drupal front-end
- Custom back-end for 16S rRNA data
- Collaborates with Ribosomal Database Project



The screenshot shows a web browser window with two tabs: 'Welcome | www.coralmi...' and 'Ribosomal Database Proj...'. The address bar displays 'http://www.coralmicrobes.org/'. The website header features a grid of 14 microscopic images of various coral-associated microbes, with the text 'CoralMicrobes.org' overlaid in a large, white, 3D-style font. Below the grid is a navigation menu with links for 'Home', 'Data Location Map', 'About', 'Contact Us', and 'Login'. The main content area begins with a 'Welcome' heading, followed by a bullet point: 'Bacterial communities associated with corals vary among coral species and along environmental gradients and are further influenced by ecological stressors, including increased ocean temperature and disease. A global perspective of the temporal and spatial'. To the right of this text is a small image of a coral branch against a blue sky.

Genome Database for Rosaceae

- Developed by Main Lab @ WSU
- PGRP proposal, partnership with Main Lab & CUGI
- To be converted to Chado/Drupal



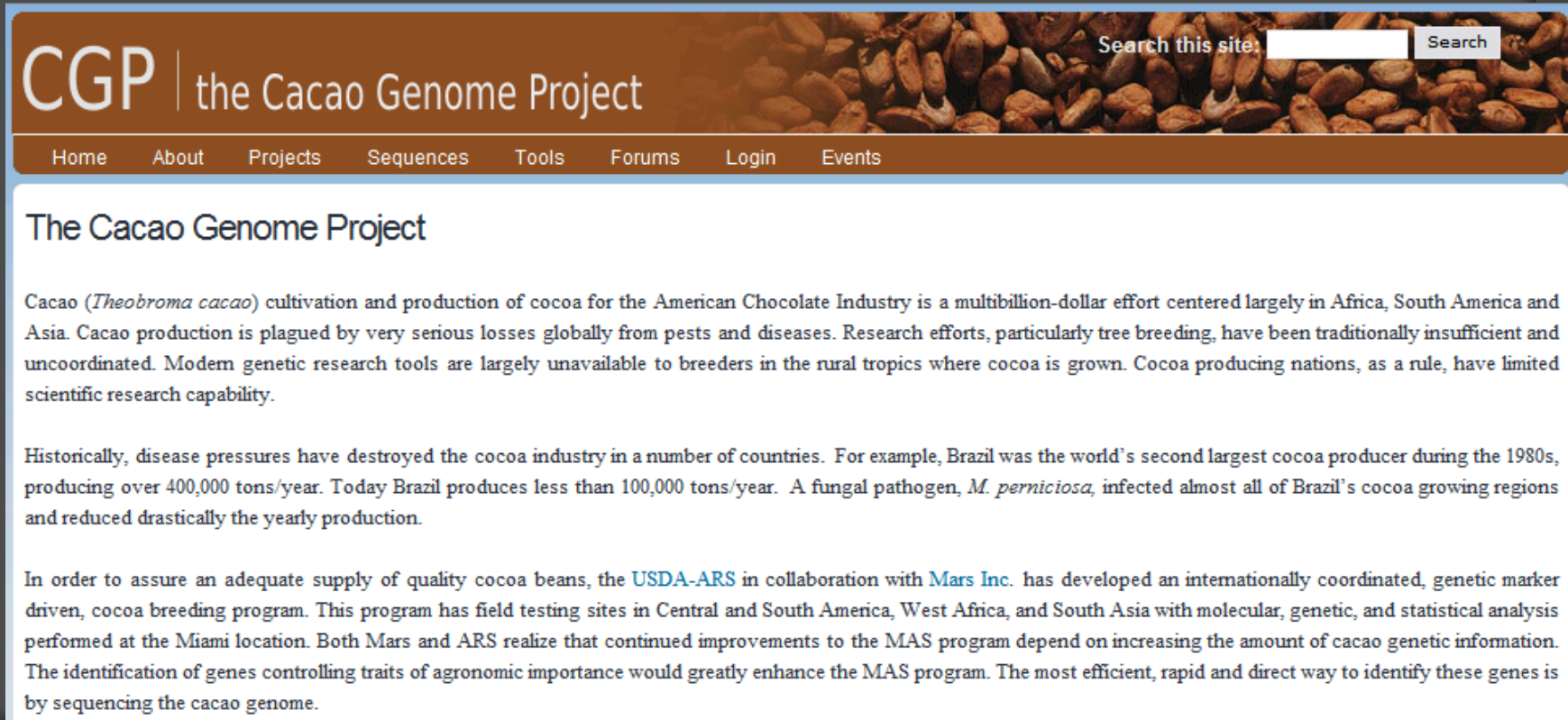
The screenshot shows a web browser window displaying the Genome Database for Rosaceae (GDR) website. The browser's address bar shows the URL <http://www.bioinfo.wsu.edu/gdr/>. The website header features the GDR logo and the text "Genome Database for Rosaceae". A navigation menu includes links for General Info, Species, Projects, Maps, Search, Tools, Community, and Forums. The main content area is divided into several sections:

- GDR Quick Start**: A list of links including Disclaimer, About GDR, Feedback Form, Data Overview, GDR Tutorial, GDR Newsletters, GDR Outreach, and Rosaceae Community.
- GDR Species**: A list of species including Apple, Pear, Prunus, Almond, Apricot, and Cherry.
- What's New in GDR?**: A list of recent updates and news items, such as "Cast your vote for US RosEXEC 2009 membership", "Rosaceae family unigene build 4 available", and "New local BLAST server implemented".

The website also features a search bar and a sidebar with a "site search" box and "Home | Co" links. The background of the website is a collage of various fruits, including apples and peaches.

Cacao Genome Project

- Main Lab @ WSU
- Currently funded: development just starting
- Will use Chado/Drupal

The image is a screenshot of the Cacao Genome Project website. At the top, there is a dark blue header with the text "CGP | the Cacao Genome Project" in white. To the right of the text is a search bar with the placeholder "Search this site:" and a "Search" button. Below the header is a navigation menu with links for "Home", "About", "Projects", "Sequences", "Tools", "Forums", "Login", and "Events". The main content area has a white background and a blue border. It features a title "The Cacao Genome Project" followed by three paragraphs of text. The first paragraph discusses the scale of cacao cultivation and the challenges of pest and disease management. The second paragraph describes historical losses in Brazil due to a fungal pathogen. The third paragraph explains the current breeding program involving Mars Inc. and USDA-ARS, highlighting the need for genetic information and the role of genome sequencing.

CGP | the Cacao Genome Project

Search this site: Search

Home About Projects Sequences Tools Forums Login Events

The Cacao Genome Project

Cacao (*Theobroma cacao*) cultivation and production of cocoa for the American Chocolate Industry is a multibillion-dollar effort centered largely in Africa, South America and Asia. Cacao production is plagued by very serious losses globally from pests and diseases. Research efforts, particularly tree breeding, have been traditionally insufficient and uncoordinated. Modern genetic research tools are largely unavailable to breeders in the rural tropics where cocoa is grown. Cocoa producing nations, as a rule, have limited scientific research capability.

Historically, disease pressures have destroyed the cocoa industry in a number of countries. For example, Brazil was the world's second largest cocoa producer during the 1980s, producing over 400,000 tons/year. Today Brazil produces less than 100,000 tons/year. A fungal pathogen, *M. perniciosa*, infected almost all of Brazil's cocoa growing regions and reduced drastically the yearly production.

In order to assure an adequate supply of quality cocoa beans, the [USDA-ARS](#) in collaboration with [Mars Inc.](#) has developed an internationally coordinated, genetic marker driven, cocoa breeding program. This program has field testing sites in Central and South America, West Africa, and South Asia with molecular, genetic, and statistical analysis performed at the Miami location. Both Mars and ARS realize that continued improvements to the MAS program depend on increasing the amount of cacao genetic information. The identification of genes controlling traits of agronomic importance would greatly enhance the MAS program. The most efficient, rapid and direct way to identify these genes is by sequencing the cacao genome.

Our Drupal/Chado Integration

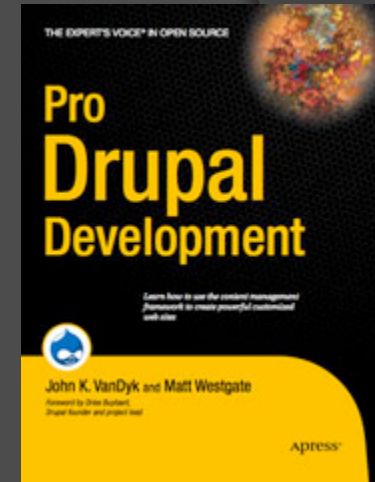
Why we chose Drupal

- ⦿ Advantages of a content management system:
 - Quicker development time
 - Easier for the end user
- ⦿ Advantages with Drupal
 - Very well documented
 - Large user-community
 - Large repository of plug-in modules
 - Themes provide highly customizable sites & easy to develop
- ⦿ Social-networking
- ⦿ RSS Feeds
- ⦿ Mash-ups



Drupal Resources

- ⦿ Drupal 6 book
 - Pro Drupal Development
- ⦿ Online API: <http://api.drupal.org/>
- ⦿ Tutorials
- ⦿ How-To's
- ⦿ Developer Forums
- ⦿ Contributed Modules



Our Database infrastructure



“Palmetto” computational cluster:

- 512 Dell PowerEdge (8 cores, 12 GB RAM)
- 244 SUN, AMD

Shibboleth User Authentication



Postgres Database Server:

Separate Schemas:

- Drupal schema
- Chado schema

Web Server:

<http://www.marinegenomcs.org>

- Drupal front-end
- User's upload data for processing
- Launches analysis jobs
- Monitors jobs
- Updates database
- Notifies User

Drupal Interface (Themes)

- Menus
- Nodes
- Blocks

- PHP
- CSS
- JQuery
- Ajax

Primary menu Page node Block

The screenshot shows a web browser displaying the homepage of the Marine Genomics Project. Three red arrows point to specific elements: one to the top navigation menu (labeled 'Primary menu'), one to the main content area (labeled 'Page node'), and one to a sidebar section titled 'Announcements' (labeled 'Block').

Welcome to the Marine Genomics Project

MG.org: a web-based interface for public transcriptomic and genomic data and analysis tools.

The Marine Genomics project is an inclusive organization that welcomes all investigators who are interested in applying genomic approaches to furthering our knowledge of marine organisms.

[Join our MG.org "webinar"](#)

Site Features:

- ◆ **Species ESTs:** Search annotated ESTs.
- ◆ **Unigenes:** Search annotated assemblies of ESTs (unigenes).
- ◆ **Blast:** Blast your sequences against MG.org ESTs and unigenes.
- ◆ **Mailing List:** Sign up for the MG.org mailing list in your user account preferences.
- ◆ **Flag Features:** Registered users can flag ESTs/contigs for quick future access.

Coming Soon:

Announcements

Gene Ontology Evidence is Now Available

Feature with available in-silico annotations of Gene Ontology terms now show the blast alignments.

Marine Genomic "Webinar" on Jan 7th -- All Invited

On Jan 7th we will be holding an online "Webinar" to demonstrate the updated features of MG.org, anyone is invited to attend.

MG.org to be Presented at PAG 2009

MG.org will be presented at this years aquaculture workshop at PAG.

Drupal/Chado modules

- ⦿ We have the following modules
 - Feature
 - Organism
 - Library
- ⦿ Did not use CCK. Wrote custom modules

Drupal Database Setup

- ⦿ We kept database schemas separate for Drupal and Chado
- ⦿ Drupal must know about both schemas:
- ⦿ Setting in `./sites/default/settings.php`

```
$db_url = array(  
  'default' => 'pgsql://<drupal_dbuser>:<password>@<hostname>/<drupal_schema_name>',  
  'chado' => 'pgsql://<chado_dbuser>:<password>@<hostname>/<chado_schema_name>',  
);
```

Retrieving Chado Data

A Chado SQL statement



```
// We'll use the following SQL statement to get the feature
$sql = "SELECT F.feature_id, F.name, F.uniquename, O.genus, ".
      "      O.species, CVT.name as cvname, F.residues, F.organism_id ".
      "FROM FEATURE F ".
      "  INNER JOIN Cvterm CVT ON F.type_id = CVT.cvterm_id ".
      "  INNER JOIN Organism O ON F.organism_id = O.organism_id ".
      "WHERE F.feature_id = %d";
```

Switch over to the chado database

```
$previous_db = db_set_active('chado');
$feature = db_fetch_object(db_query($sql,$feature_id));
db_set_active($previous_db);
```

Execute the Chado query

Switch back to the Drupal database

Feature Module

- Creates a Drupal 'chado_feature' node & Drupal table:

Column	Type	Not Null	Default	Actions		Comment
vid	int_unsigned	NOT NULL	0	Alter	Drop	
nid	int_unsigned	NOT NULL	0	Alter	Drop	
feature_id	integer	NOT NULL	0	Alter	Drop	

- Drupal insert_hook and update_hook interact with these chado tables:
 - feature
 - featureprop
 - dbxref
 - feature_dbxref
 - synonym
 - feature_synonym

Unique Feature Name: *

Argopecten_irradians-Contig_164-v1

Enter a unique name for this feature

▼ **Flags**

Flag this feature for easy access
Flag this feature for easy access

Feature Type: *

contig ▼

Choose the feature type.

Organism: *

Argopecten irradians (Atlantic Bay Scallop) ▼

Choose the organism with which this feature is associated

Genbank Accession:

Enter the ID assigned by genbank for this feature

Synonyms:

Argopecten_irradians-Contig_164-v1

Enter alternate names (synonyms) for this feature to help in searching and identification. You may enter as many alternate names as needed separated by spaces or on different lines.

Residues:

```
GCATAGACCAGCTAAACACACAAAGTCATCAACTCATCTCGGAACAGCAGCAAATCACACAAAAACAAACATGTCGGCCGAACCAGAATACAAGAAACAGC
CAGTTGCTGACTTGTGGGCAAAACCTTGATGGAACATAAAGAATGTAAATCTTTATTGAAAAAGCATCTTACCAAGGAAAGATATGATGCTTTAAAGGATTTGA
AAACAAGCATTGGCGGTGACCTTGGGGATTGTATTGAGTCAGGTTGTATGAACTTGGACAGTGGTGTGGTATTTATGCTTGTGACCCTGAAGGATACGATA
CATTGCTCCCCTTGGATGAAGTCATTAAAGATTACCATAAGGTCGAGAAGGTTGAACACCCTGAACCAAACCTTTGGAGATTTGGAAAACCTTGAATCTTC
CCGATCTTGACCCTAACAAATGAGATGATTGTCAGCACCCGAGTACGTGTTGGAAGGAGTCATGTAGGCTTTCCATTCCCACCTGCTGCCACCAAAGAGCAA
```

Enter the nucleotide sequences for this feature

Is Obsolete

Organism Module

- Creates a Drupal 'chado_organism' node & Drupal table:

Column	Type	Not Null	Default	Actions	Comment
vid	int_unsigned	NOT NULL	0	Alter Drop	
nid	int_unsigned	NOT NULL	0	Alter Drop	
organism_id	integer	NOT NULL	0	Alter Drop	

- Drupal insert_hook and update_hook interact with these chado tables:
 - organism

phpPgAdmin x Karenia brevis | MarineGe... x

http://www.marinegenomics.org/node/27110


The Marine Genomics Project

Species/Projects Tools Search Contribute About MG.org Login

[Home](#)

Karenia brevis

View EST Assembly GO Analysis




Overview

Common Name	Karenia brevis (toxic dinoflagellate)
Number of Contigs	893
Number of ESTs	6,991

Download ESTs: [Karenia brevis.EST.fasta](#)

Background:
 Karenia brevis is a dinoflagellate whose expressed genome is of significant interest because of its role in producing harmful algal blooms (HABs) or "red tides" that occur annually in the Gulf of Mexico. Dinoflagellates are microscopic, unicellular, flagellated, often photosynthetic protists, commonly regarded as "algae" (Division Dinoflagellata). K. brevis "red tides" cause extensive marine animal mortalities and human illness through the production of highly potent neurotoxins known as brevetoxins. Although K. brevis has come to be known as the Florida red tide organism, it has been implicated in blooms in Louisiana, Texas, Mississippi, Mexico, and the Carolinas.

Disclaimer
 email: info@marinegenomics.org
 Copyright (c) 2005-2008



Library Module

- Creates a Drupal 'chado_library' node & Drupal table:

Column	Type	Not Null	Default	Actions		Comment
vid	int_unsigned	NOT NULL	0	Alter	Drop	
nid	int_unsigned	NOT NULL	0	Alter	Drop	
library_id	integer	NOT NULL	0	Alter	Drop	

- Drupal insert_hook and update_hook interact with these chado tables:
 - library
 - libraryprop

Searching

- Currently using Drupal full-text search
- All node content is indexed for searching
 - Feature name, synonyms and all other content

```
CATTGATGCGGACTCGGGGGGCGTTGCCATACTTTGTGTTTACAACAGG
GCGTCTGGCCAAAGTCCACCCCACTTTTTCCAATAANTCCTCCTTTTGGG
TGAAGAAGGCNAAGTTGCAAAAATGAATT
```

Length 629
Type EST
Organism [North Atlantic Right Whale](#)

References

Genbank accession numbers searchable

Dababase	Accession
Genbank dbEST	ES556888

ExpASY Swissprot Blast Hits

Blast hit descriptions searchable

Best 10 Hits Shown | [Show Best 25 Hits](#) | [Show All Hits](#)
Note: Click a description for more details.

Match Name	E value	Description
BCKD_BOVIN	4.60471e-82	[3-methyl-2-oxobutanoate dehydrogenase [lipoamide]] kinase, mitochondrial precursor - Bos taurus (Bovine)
BCKD_RAT	1.33976e-81	[3-methyl-2-oxobutanoate dehydrogenase [lipoamide]] kinase, mitochondrial precursor - Rattus norvegicus (Rat)

Drupal “Taxonomy” (categories)

- Nodes can be assigned “taxonomy” terms
- Searching can be filtered by categories
- We assign cvterm and organism as taxonomy to “feature” nodes.

“Taxonomy” terms

Search

Enter your keywords:

▼ Advanced search

Containing any of the words: <input type="text"/>	Only in the category(s): Feature Type EST contig Organism/Project Acropora palmata Anas platyrhynchos Argopecten irradians Calanus finmarchicus Callinectes sapidus Crassostrea gigas	Only of the type(s): <input type="checkbox"/> Announcements <input type="checkbox"/> Book page <input type="checkbox"/> EST Batch Upload <input type="checkbox"/> Feature <input type="checkbox"/> Organism/Project <input type="checkbox"/> Page <input type="checkbox"/> Story
---	--	--

Syncing Drupal and Chado

- ⦿ Data added to Chado first:
 - Use chado GFF upload scripts to add data
 - Sync scripts to run on command-line which:
 - Generate Drupal nodes for specific feature types
 - Add taxonomy (cvterm, organism)
 - Index the node for searching
- ⦿ Data added to Drupal first:
 - In the case of our MG.org EST pipeline.
 - Curators review data before inclusion into chado
 - Drupal module adds features to data once approved.
 - Add to feature, featureprop, feature_dbxref, synonyms

Integration of independent tools

- Drupal page node + theme + HTML Iframes
- Perl-based GBrowse:

The screenshot displays a web browser window with the URL <http://www.marinegenomics.org/node/288318/gbrowse>. The page title is "Litopenaeus_vannamei-Contig_5089-v1". Below the title, there are two tabs: "View" and "GBrowse". The main content area shows "Showing 780 bp from litopenaeus_vannamei-contig_5089-v1, positions 1 to 780".

On the left side, there are several sections:

- Instructions:** Includes links for [Hide banner], [Bookmark this], [Link to Image], [Help], and a [Reset] button.
- Search:** A search bar with the text "Litopenaeus_vannamei-Contig_" and a "Search" button.
- Data Source:** "MarineGenomics.org Public Species Data".
- Overview:** A horizontal scale from 0 to 700+ bp with a red box highlighting the current view range.
- Details:** A section titled "CAP3 Alignments" showing multiple horizontal green bars representing sequence alignments.

On the right side, there are controls for "Reports & Analysis" (Highlight Selected Properties, Configure..., Go) and "Scroll/Zoom" (navigation arrows, Show 780 bp, Flip).

Large Putative Data Sets

- ◎ Blast results are not stored in chado:
 - too large
 - only putative
 - but needed for manual curation!!!
 - we need them searchable
- XML formatted blast results for each feature
- Stored in file system directory structure based off of feature_id and db_id
 - e.g: Eubalaena_glacialis-Contig_49-v1
 - feature_id == 154351
 - db_id (for Uniprot) == 105
 - Blast results found here:
 - /<blast repository path>/154/154351/105.xml
 - <http://www.marinegenomics.org/MGID154351>

Where we're going!

“Tripal”

- ◉ Our package of Drupal modules
- ◉ Chado refers to a Japanese tea ceremony
- ◉ Tripel is a strong belgian beer.
- ◉ Change the ‘e’ to an ‘a’ to mimic drupal: Tripal
- ◉ Drupal + Chado ~ Tripal

Release of our Tripal modules

- Spring 2009 release of Tripal for general use:
 - Feature module
 - Organism module
 - Library module
- Available in Drupal module repository and CUGI website
- Continued development of new modules:
 - Currently funded through WSU's cacao genome database
 - Additional funding requested through WSU's GDR PGRP proposal and future WSU/Clemson/Cornell Fruit & Nut GDR SCRI submission (04/09)
 - Continued development at CUGI for our funded research projects.

Upgrades to Existing Sites

⦿ Data Types:

- Our current “TriPal” databases have:
 - ESTs
 - Marker data
- Expand to whole genome, physical mapping, and ancillary annotation data
- Breeding and phenotypic data.

Cyberinfrastructure

- “Leadership in cyberinfrastructure may well become the determinant in measuring pre-eminence in higher education among nations”. -- Dr. Arden Bement, Director of NSF, 2007.
- NSF’s Cyberinfrastructure Vision for 21st century:
http://www.nsf.gov/od/oci/ci_v5.pdf
 - Computing systems
 - Data storage systems
 - Advanced instruments and data repositories
 - Visualization environments
 - People
 - Linked by high speed networks.
 - **Enable scholarly innovation and discoveries not otherwise possible**
- Clemson’s CITI group is working with Purdue and the HubZero folks

A Current Project

- Cyberinfrastructure Project
 - Involves marinegenomics.org
- South Carolina Marine Genomics Consortium
 - <http://www.genome.clemson.edu/activities/projects/marineGenomics/>
- Mission Statement:

“Using cyber-enabled genomic and bioinformatic approaches to predict and address the impact of climate change and environmental stressors on ecosystems.”

SOUTH CAROLINA
Marine Genomics Consortium



Thank You