GMOD Project Update

Scott Cain
GMOD Project Coordinator
Ontario Institute for Cancer Research
scott@scottcain.net

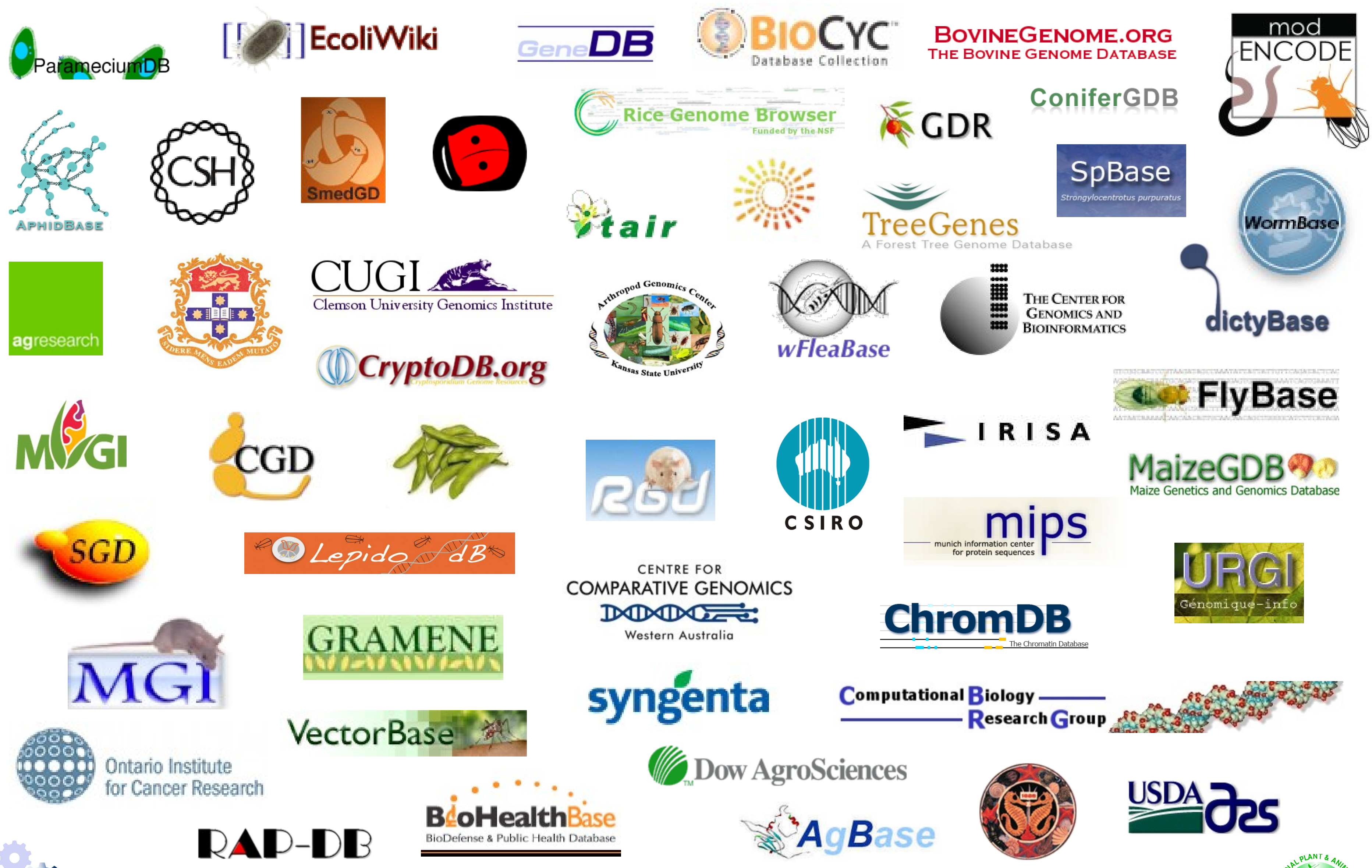
GMOD Meeting
San Diego, CA
January 14-15, 2010

Introduction: GMOD is …

- A set of interoperable open-source **software** components for visualizing, annotating, and managing biological data.

- An active **community** of developers and users asking diverse questions, and facing common challenges, with their biological data.

Who uses GMOD?


ParameciumDB
EcoliWiki
GeneDB
BioCyc Database Collection
BovineGenome.org
The Bovine Genome Database
mod ENCODE
Rice Genome Browser
Funded by the NSF
GDR
ConiferGDB
AphidBase
CSH
SmedGD
tair
TreeGenes
A Forest Tree Genome Database
SpBase
Strongylocentrotus purpuratus
WormBase
agresearch
SIDERE MENS EADEM MUTATO
CUGI
Clemson University Genomics Institute
CryptoDB.org
Cryptosporidium Genome Resources
Arthropod Genomics Center
Kansas State University
wFleaBase
The Center for Genomics and Bioinformatics
dictyBase
FlyBase
MGI
CGD
RGD
CSIRO
IRISA
MaizeGDB
Maize Genetics and Genomics Database
mips
munich information center for protein sequences
SGD
Lepido dB
CENTRE FOR COMPARATIVE GENOMICS
Western Australia
ChromDB
The Chromatin Database
URGI
Génomique-info
MGI
GRAMENE
syngenta
Computational Biology Research Group
VectorBase
Ontario Institute for Cancer Research
Dow AgroSciences
RAP-DB
BioHealthBase
BioDefense & Public Health Database
AgBase
USDA ARS

GMOD

Plus hundreds of others

INTERNATIONAL PLANT & ANIMAL GENOME CONFERENCE

Software

GMOD components can be categorized as

- V Visualization
- D Data Management
- A Annotation

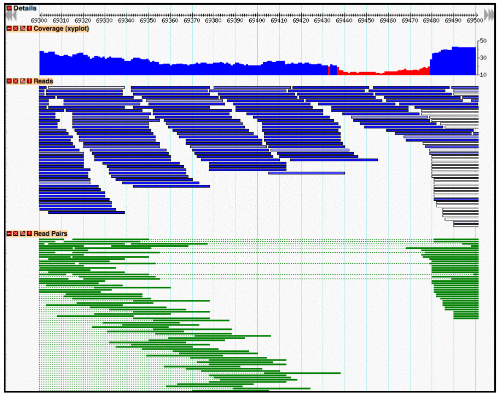
Visualization: GBrowse

GBrowse

JBrowse

GBrowse_syn

CMap



Releases

- 1.70 released
- 2.0, 1.71 in the pipe

AJAX/Interface:

- Rubberband region selection, drag and drop track ordering, collapsible tracks, popup balloons, asynchronous rendering (2.0)

Biology:

- Allele/gentotype frequency, LD glyphs, geolocation popups, circular genome support (1.71)

Infrastructure

- User logins, server multiplexing (2.0), SQLite and SAMtools (NGS) adaptors

modENCODE Fly:

- http://modencode.oicr.on.ca/cgi-bin/gb2/gbrowse/fly/

The generic genome browser: a building block for a model organism system database. Stein LD et al. (2002) *Genome Res* 12: 1599-610

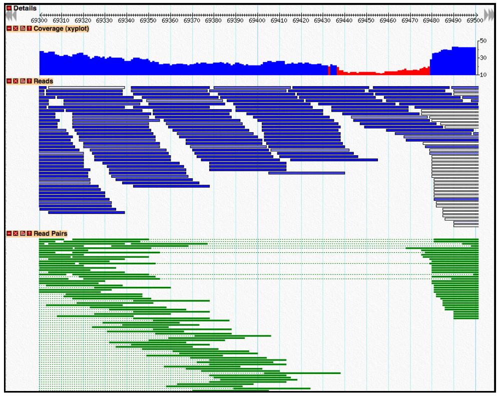
Visualization

GBrowse

JBrowse

GBrowse_syn

CMap



Resources

Tutorials (http://gmod.org/wiki/GBrowse_Tutorial):

- GBrowse User Tutorial at OpenHelix.com
- GBrowse Admin Tutorial
- NGS in GBrowse and SAMtools Tutorial

Web Sites:

GMOD	http://gmod.org/wiki/GBrowse
WebGBrowse	http://webgbrowse.cgb.indiana.edu/
GBrowse.org	http://gbrowse.org

Mailing List:

https://lists.sourceforge.net/lists/listinfo/gmod-gbrowse

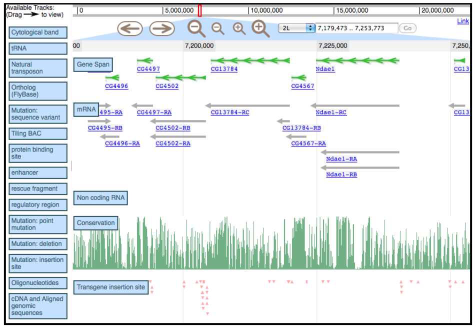
Visualization

GBrowse

JBrowse

GBrowse_syn

CMap



GMOD's 2nd Generation Genome Browser
It's *fast*

Completely new genome browser implementation:

Client side rendering

Heavy use of AJAX

Uses JSON and Nested Containment Lists

JBrowse Fly:

http://jbrowse.org/genomes/dmel/

Web Sites:

GMOD	http://gmod.org/wiki/JBrowse
JBrowse	http://jbrowse.org

Mailing List:

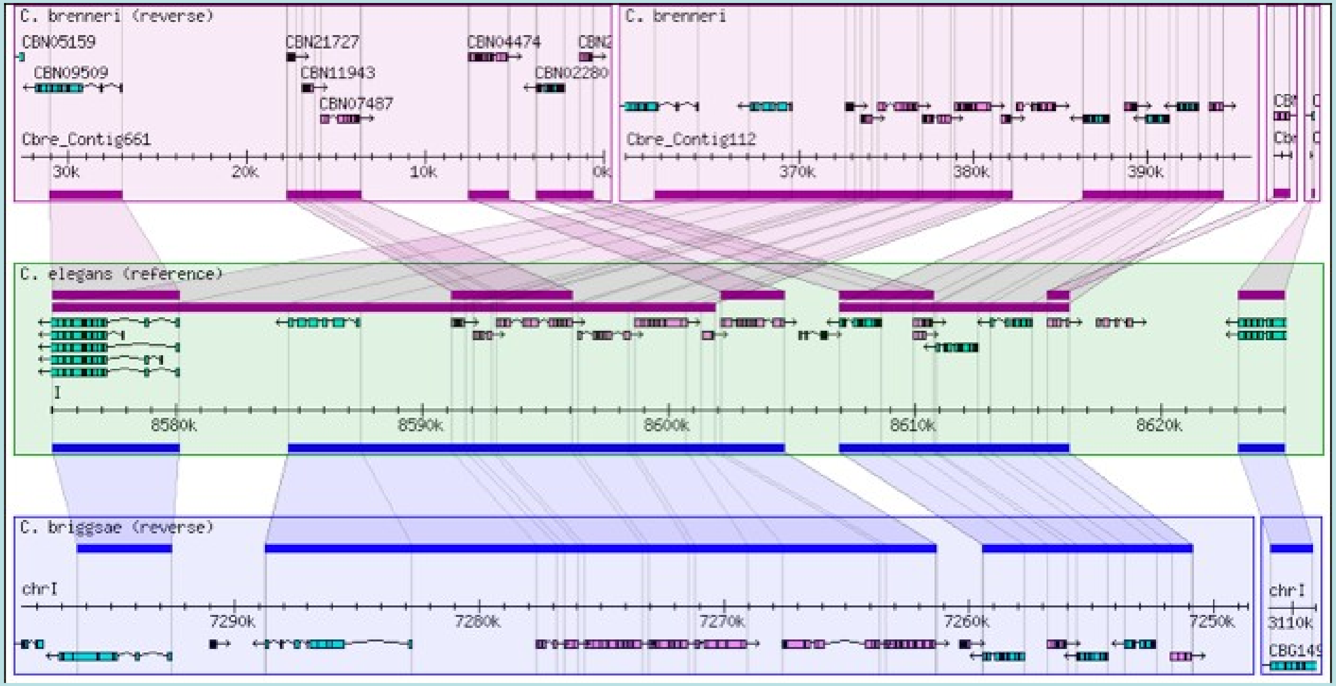
https://lists.sourceforge.net/lists/listinfo/gmod-ajax

JBrowse: A next-generation genome browser, Mitchell E. Skinner, Andrew V. Uzilov, Lincoln D. Stein, Christopher J. Mungall and Ian H. Holmes, *Genome Res.* 2009. 19: 1630-1638

Visualization

- GBrowse
- JBrowse
- **GBrowse_syn**
- CMap

GBrowse based comparative genomics viewer
Shows a reference sequence compared to 2+ others
Can also show any GBrowse-based annotations



wormbase.org

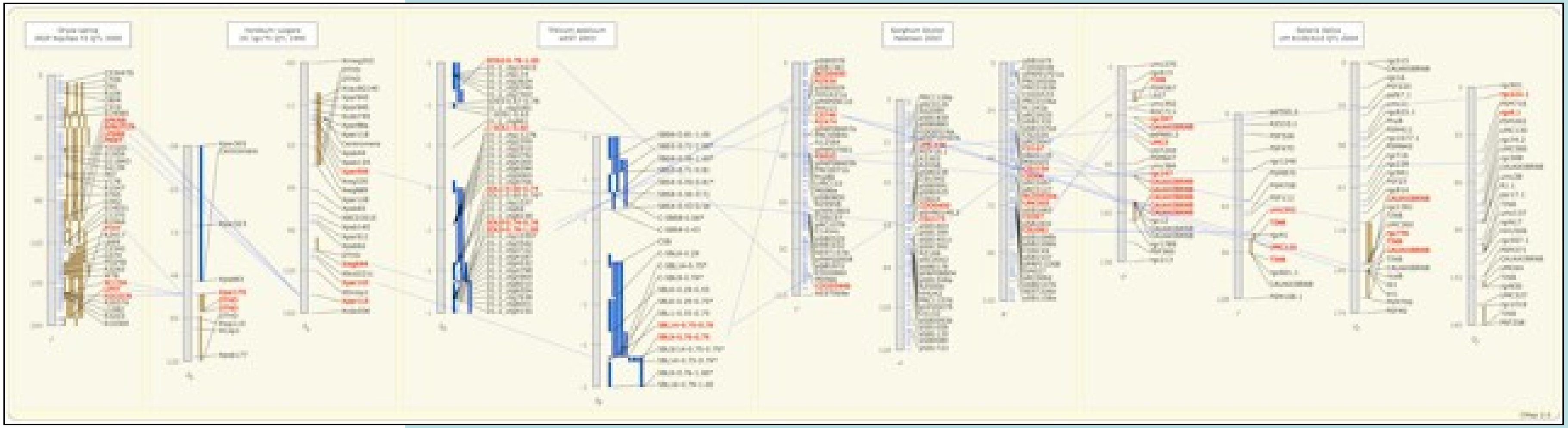
Syntenic blocks do not have to be colinear
Can also show duplications

http://gmod.org/wiki/GBrowse_syn

Visualization

- GBrowse
- JBrowse
- GBrowse_syn
- **CMap**

- Web based comparative map viewer
- CMap is data type agnostic:
  - Can link sequence, genetic, physical, QTL, deletion, optical, …
- CMap 2.0 coming
  - Faster, internals cleanup
  - Circos export



CMap 1.01: A comparative mapping application for the Internet, Ken Youens-Clark, Ben Faga, Immanuel V. Yap, Lincoln Stein and Doreen Ware, *Bioinformatics*, doi:10.1093/bioinformatics/btp458

Data Management

- Chado
- Tripal
- TableEdit
- BioMart
- InterMine

A extensible, modular database schema for storing biological data

1.0 release:

- Stable schema
- Tools for data in/out

1.1 release (soon):

- Stable schema (minor, nondestructive changes)
- Improvements to data loading scripts
- Additional modules: cell line, natural diversity
- Tool for managing materialized views
- Tool for creating ontology-based views

Data Management

Chado

Tripal

TableEdit

BioMart

InterMine

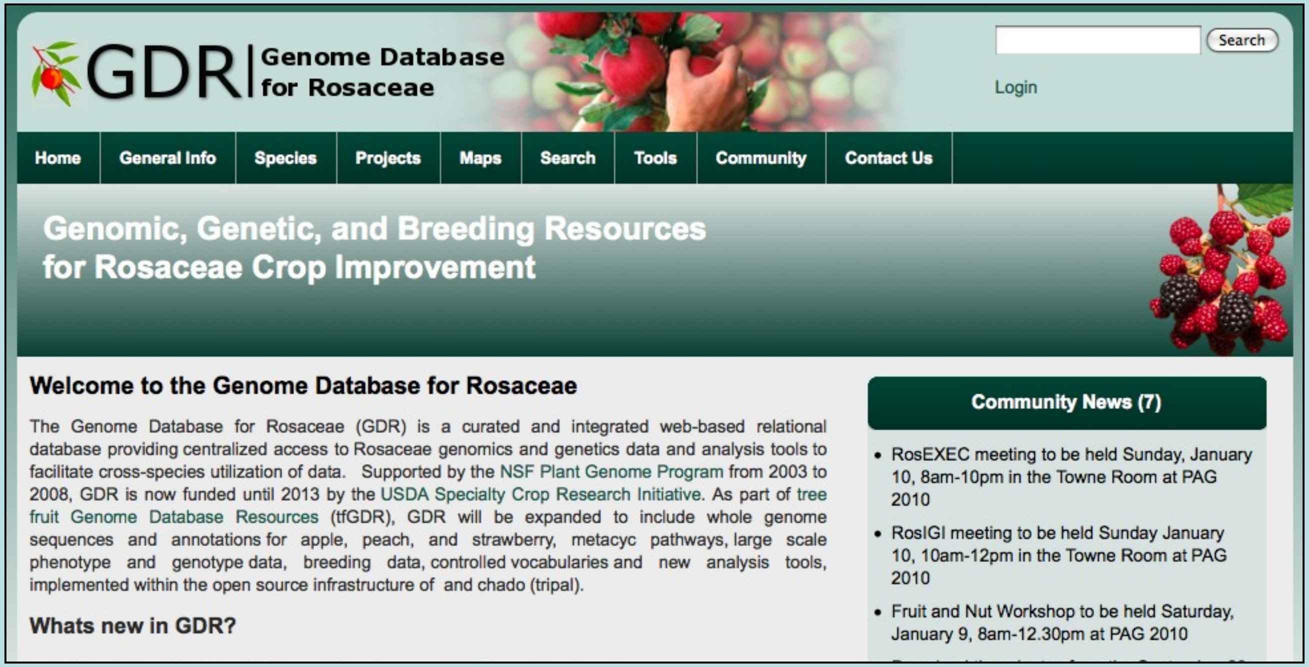
New web front end for Chado databases

Set of Drupal modules

Modules approximately correspond to Chado modules

Easy to create new modules

Includes user authentication, job management, curation support



Genomic, Genetic, and Breeding Resources for Rosaceae Crop Improvement

Welcome to the Genome Database for Rosaceae

The Genome Database for Rosaceae (GDR) is a curated and integrated web-based relational database providing centralized access to Rosaceae genomics and genetics data and analysis tools to facilitate cross-species utilization of data. Supported by the NSF Plant Genome Program from 2003 to 2008, GDR is now funded until 2013 by the USDA Specialty Crop Research Initiative. As part of tree fruit Genome Database Resources (tfGDR), GDR will be expanded to include whole genome sequences and annotations for apple, peach, and strawberry, metacyc pathways, large scale phenotype and genotype data, breeding data, controlled vocabularies and new analysis tools, implemented within the open source infrastructure of and chado (tripal).

Whats new in GDR?

Community News (7)

- RosEXEC meeting to be held Sunday, January 10, 8am-10pm in the Towne Room at PAG 2010
- RosIGI meeting to be held Sunday January 10, 10am-12pm in the Towne Room at PAG 2010
- Fruit and Nut Workshop to be held Saturday, January 9, 8am-12.30pm at PAG 2010

Stephen Ficklin, Meg Staton, Chun-Huai Cheng, …
Clemson University Genomics Institute

Data Management

- Chado
- Tripal
- **TableEdit**
- BioMart
- InterMine

MediaWiki extension

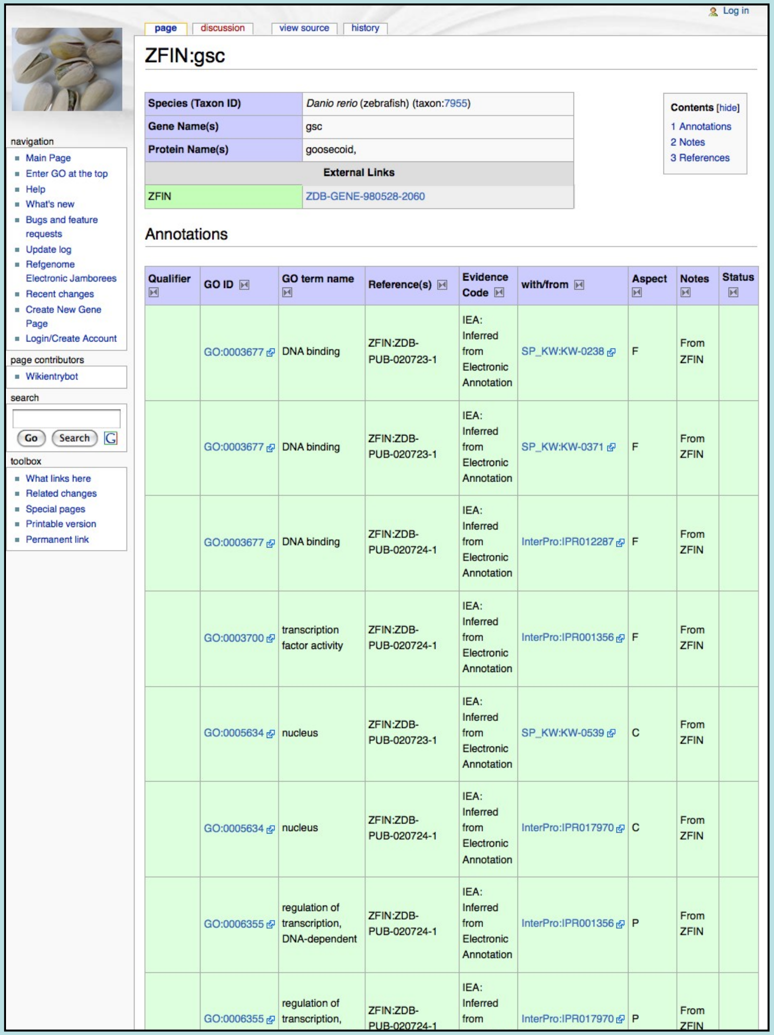
- MediaWiki software used at Wikipedia, GMOD.org, …

GUI to wiki tables

Also a GUI to database tables

Work in progress to use this with Chado

Potential to give wiki access to Chado databases



ZFIN:gsc

Species (Taxon ID)	*Danio rerio* (zebrafish) (taxon:7955)
Gene Name(s)	gsc
Protein Name(s)	goosecoid,
External Links	
ZFIN	ZDB-GENE-980528-2060

Annotations

Qualifier	GO ID	GO term name	Reference(s)	Evidence Code	with/from	Aspect	Notes	Status
	GO:0003677	DNA binding	ZFIN:ZDB-PUB-020723-1	IEA: Inferred from Electronic Annotation	SP_KW:KW-0238	F	From ZFIN	
	GO:0003677	DNA binding	ZFIN:ZDB-PUB-020723-1	IEA: Inferred from Electronic Annotation	SP_KW:KW-0371	F	From ZFIN	
	GO:0003677	DNA binding	ZFIN:ZDB-PUB-020724-1	IEA: Inferred from Electronic Annotation	InterPro:IPR012287	F	From ZFIN	
	GO:0003700	transcription factor activity	ZFIN:ZDB-PUB-020724-1	IEA: Inferred from Electronic Annotation	InterPro:IPR001356	F	From ZFIN	
	GO:0005634	nucleus	ZFIN:ZDB-PUB-020723-1	IEA: Inferred from Electronic Annotation	SP_KW:KW-0539	C	From ZFIN	
	GO:0005634	nucleus	ZFIN:ZDB-PUB-020724-1	IEA: Inferred from Electronic Annotation	InterPro:IPR017970	C	From ZFIN	
	GO:0006355	regulation of transcription, DNA-dependent	ZFIN:ZDB-PUB-020724-1	IEA: Inferred from Electronic Annotation	InterPro:IPR001356	P	From ZFIN	
	GO:0006355	regulation of transcription,	ZFIN:ZDB-PUB-020724-1	IEA: Inferred from	InterPro:IPR017970	P	From ZFIN	

Example: GONUTS (http://gowiki.tamu.edu/)

Jim Hu, Daniel Renfro, *et al*., Texas A&M

Data Management

- Chado
- Tripal
- TableEdit
- **BioMart**
- InterMine

??

New GUIs - more configurable and easier to use

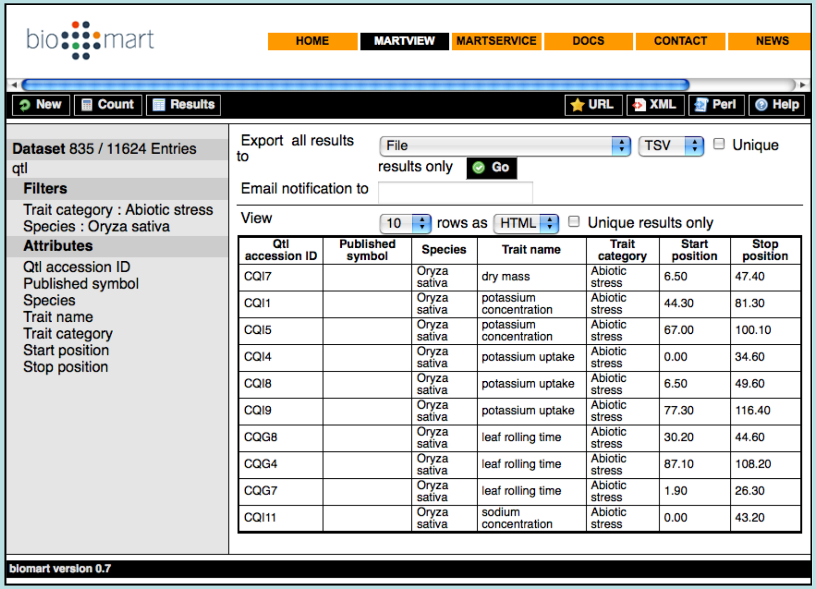
Virtual Marts - marts running off source schema without materializing

Improved scalability

Security and access control

Improved federation

New configuration tool



Dataset 835 / 11624 Entries

qtl

Filters

Trait category : Abiotic stress
Species : Oryza sativa

Attributes

Qtl accession ID
Published symbol
Species
Trait name
Trait category
Start position
Stop position

Export all results to: File, TSV, □ Unique results only, Go

Email notification to

View: 10 rows as HTML □ Unique results only

Qtl accession ID	Published symbol	Species	Trait name	Trait category	Start position	Stop position
CQI7		Oryza sativa	dry mass	Abiotic stress	6.50	47.40
CQI1		Oryza sativa	potassium concentration	Abiotic stress	44.30	81.30
CQI5		Oryza sativa	potassium concentration	Abiotic stress	67.00	100.10
CQI4		Oryza sativa	potassium uptake	Abiotic stress	0.00	34.60
CQI8		Oryza sativa	potassium uptake	Abiotic stress	6.50	49.60
CQI9		Oryza sativa	potassium uptake	Abiotic stress	77.30	116.40
CQG8		Oryza sativa	leaf rolling time	Abiotic stress	30.20	44.60
CQG4		Oryza sativa	leaf rolling time	Abiotic stress	87.10	108.20
CQG7		Oryza sativa	leaf rolling time	Abiotic stress	1.90	26.30
CQI11		Oryza sativa	sodium concentration	Abiotic stress	0.00	43.20

biomart version 0.7

Data Management

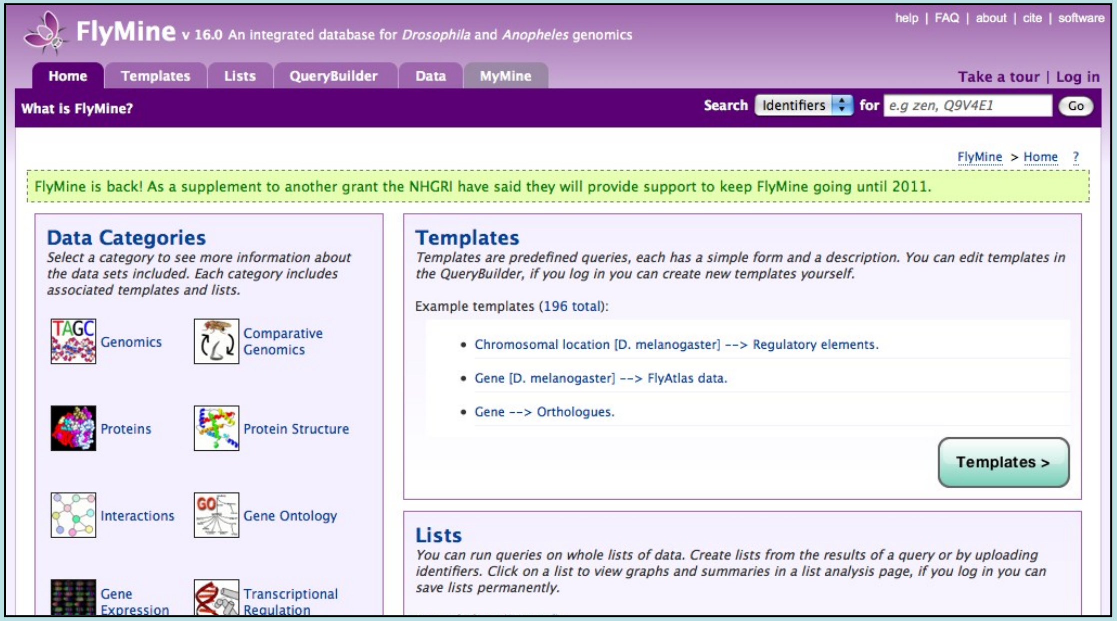
- Chado
- Tripal
- TableEdit
- BioMart
- **InterMine**

Data integration and web-based query package

Now supports ~20 common data formats:
GFF3, Chado, GO annotation, biopax, BioGrid, TreeFam, PubMed, Ensembl, …

Interfaces: RESTful web service, Java & Perl APIs

Upload & analyse gene lists with graphical and statistical widgets



FlyMine v 16.0 An integrated database for *Drosophila* and *Anopheles* genomics

FlyMine is back! As a supplement to another grant the NHGRI have said they will provide support to keep FlyMine going until 2011.

Data Categories

Select a category to see more information about the data sets included. Each category includes associated templates and lists.

Templates

Templates are predefined queries, each has a simple form and a description. You can edit templates in the QueryBuilder, if you log in you can create new templates yourself.

Example templates (196 total):

- Chromosomal location [D. melanogaster] --> Regulatory elements.
- Gene [D. melanogaster] --> FlyAtlas data.
- Gene --> Orthologues.

Lists

You can run queries on whole lists of data. Create lists from the results of a query or by uploading identifiers. Click on a list to view graphs and summaries in a list analysis page, if you log in you can save lists permanently.

FlyMine: an integrated database for *Drosophila* and *Anopheles* genomics, Rachel Lyne, *et al.*, *Genome Biol.* 2007; 8(7): R129.

Annotation

- MAKER
- DIYA
- Galaxy
- Ergatis
- Apollo

MAKER
Annotate this!

Genome annotation pipeline for creating gene models

Output can be loaded into GBrowse, Apollo, Chado, …

Incorporates

- SNAP, RepeatMasker, exonerate, BLAST,
- Augustus, FGENESH, GeneMark, MPI

Other capabilities

- Map existing annotation onto new assemblies
- Merge multiple legacy annotation sets into a consensus set
- Update existing annotations with new evidence
- Integrate raw InterProScan results

MAKER Web Annotation Service - MAKER online

MAKER: An Easy-to-use Annotation Pipeline Designed for Emerging Model Organism Genomes. Cantarel B, *et al., Genome Res*. Jan;18(1):188-96

Annotation

MAKER

DIYA

Galaxy

Ergatis

Apollo

Lightweight, modular, and configurable Perl-based pipeline framework

Initial application is gene prediction for prokaryotes

Working on integration of Amos assembly tools

DIYA: a bacterial annotation pipeline for any genomics lab, Andrew C. Stewart, Brian Osborne and Timothy D. Read, *Bioinformatics* 2009 25(7):962-963

Annotation

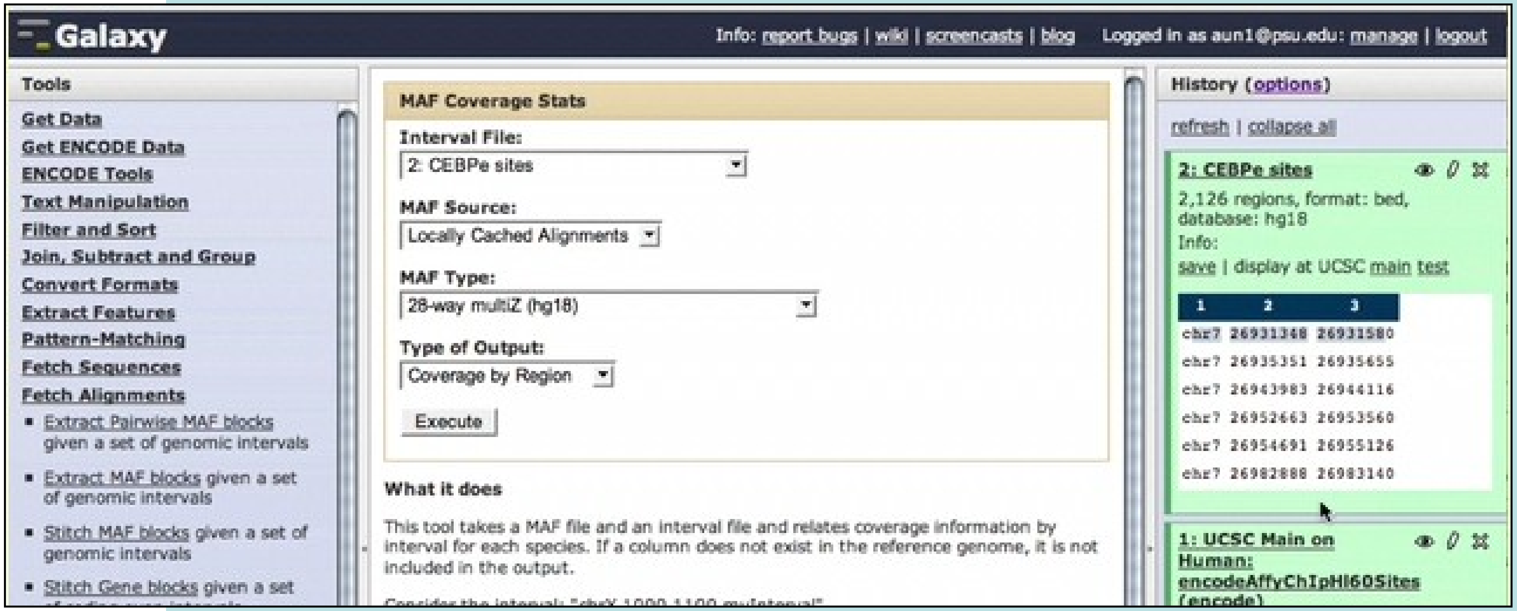
- MAKER
- DIYA
- Galaxy
- Ergatis
- Apollo

NGS tools support

- QC and Manipulation: FASTQ, 454, SOLiD support
- Mapping: Bowtie or BWA, Megablast
- SAMtools: Web interface to SAMtools scripts

LIMS system in beta.

- Import data from your sequencer into Galaxy


Galaxy
Info: report bugs | wiki | screencasts | blog
Logged in as aun1@psu.edu: manage | logout
Tools
Get Data
Get ENCODE Data
ENCODE Tools
Text Manipulation
Filter and Sort
Join, Subtract and Group
Convert Formats
Extract Features
Pattern-Matching
Fetch Sequences
Fetch Alignments
Extract Pairwise MAF blocks given a set of genomic intervals
Extract MAF blocks given a set of genomic intervals
Stitch MAF blocks given a set of genomic intervals
Stitch Gene blocks given a set
History (options)
refresh | collapse all
Execute

MAF Coverage Stats

Interval File: 2: CEBPe sites

MAF Source: Locally Cached Alignments

MAF Type: 28-way multiZ (hg18)

Type of Output: Coverage by Region

What it does

This tool takes a MAF file and an interval file and relates coverage information by interval for each species. If a column does not exist in the reference genome, it is not included in the output.

2: CEBPe sites

2,126 regions, format: bed, database: hg18

Info:

save | display at UCSC main test

1	2	3
chr7	26931348	26931580
chr7	26935351	26935655
chr7	26943983	26944116
chr7	26952663	26953560
chr7	26954691	26955126
chr7	26982888	26983140

1: UCSC Main on Human: encodeAffyChIpHl60Sites (encode)

DIYA: a bacterial annotation pipeline for any genomics lab, Andrew C. Stewart, Brian Osborne and Timothy D. Read, *Bioinformatics* 2009 25(7):962-963

Annotation

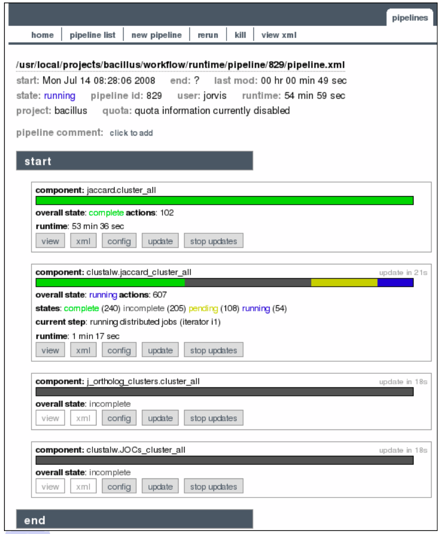
MAKER

DIYA

Galaxy

Ergatis

Apollo



/usr/local/projects/bacillus/workflow/runtime/pipeline/829/pipeline.xml

start: Mon Jul 14 08:28:06 2008 end: ? last mod: 00 hr 00 min 49 sec

state: running pipeline id: 829 user: jorvis runtime: 54 min 59 sec

project: bacillus quota: quota information currently disabled

pipeline comment: click to add

start

component: jaccard.cluster_all

overall state: complete actions: 102

runtime: 53 min 36 sec

component: clustalw.jaccard_cluster_all update in 21s

overall state: running actions: 607

states: complete (240) incomplete (205) pending (108) running (54)

current step: running distributed jobs (iterator i1)

runtime: 1 min 17 sec

component: j_ortholog_clusters.cluster_all update in 18s

overall state: incomplete

component: clustalw.JOCs_cluster_all update in 18s

overall state: incomplete

end

- Currently up to 162 analysis tool / components for use in pipelines
- Updated prokaryotic annotation pipeline template
- Updated comparative annotation pipeline template
- Lots of work for use on Amazon EC2
- Now the engine behind the new CloVR cloud computing project (http://clovr.igs.umaryland.edu/)

http://ergatis.sourceforge.net/

Annotation

- MAKER
- DIYA
- Galaxy
- Ergatis
- Apollo

Better Chado support

  including DBMS independent support)

GFF3 support

GUI based configurations

Multiple alignment transcript viewer and editor

Continuous data display

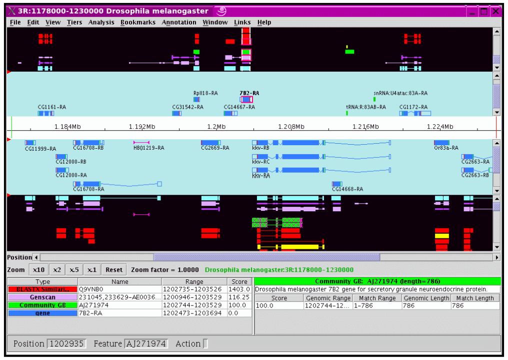
  sgr, wiggle

Remote analysis to NCBI services

  BLAST, Primer-BLAST

Undo support

More robust Java Web Start support



Zoom factor = 1.0000 Drosophila melanogaster:3R:1178000-1230000

Type	Name	Range	Score
BLASTX Similari...	Q9VNB0	1202735-1203526	1403.0
Genscan	231045,233629-AE0036...	1200946-1203529	116.25
Community GB	AJ271974	1202744-1203529	100.0
gene	7B2-RA	1202473-1203694	0.0

Community GB: AJ271974 (length=786)

Drosophila melanogaster 7B2 gene for secretory granule neuroendocrine protein.

Score	Genomic Range	Match Range	Genomic Length	Match Length
100.0	1202744-12...	1-786	786	786

Position 1202935 Feature AJ271974 Action

DIYA: a bacterial annotation pipeline for any genomics lab, Andrew C. Stewart, Brian Osborne and Timothy D. Read, *Bioinformatics* 2009 25(7):962-963