Database Tools for Biologists

Dave Clements
GMOD Help Desk
US National Evolutionary Synthesis Center (NESCent)
clements@nescent.org

Texas A&M University
Biochemistry and Biophysics
8 December 2009

Sponsored by

Alliance for Bioinformatics, Computational Biology, and Systems Biology

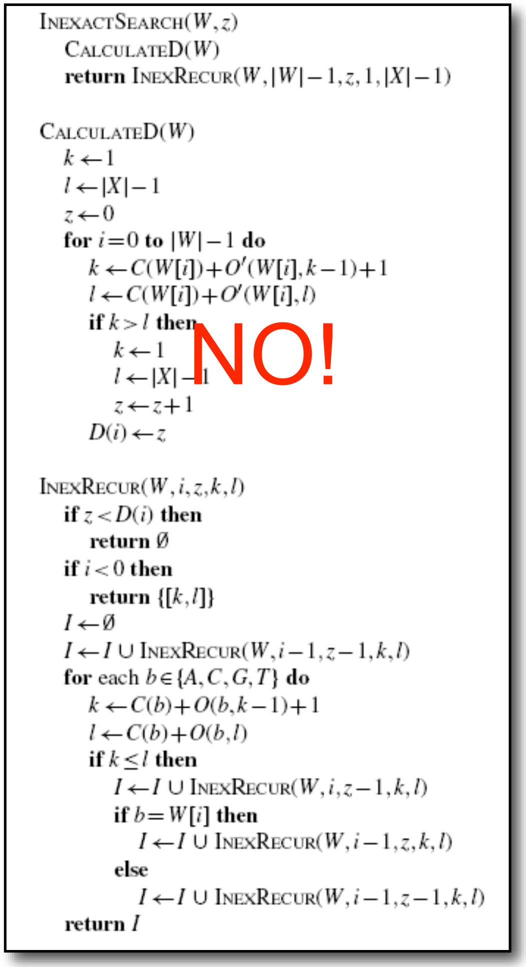
Laboratory for Genome Bioinformatics

Agenda

3:00	Introduction
3:10	Software Visualization GBrowse JBrowse GBrowse_syn Sybil & SynView CMap
4:00	Software, cont. Data Management Chado, Tripal, GMODWeb BioMart and InterMine GFF3 Annotation MAKER & DIYA Apollo Textpresso Community Annotation Pipelines & Workflows
4:40	Community
5:00	Finish

This Talk

Algorithms?



```
InexactSearch$(W, z)$
    CalculateD$(W)$
    return InexRecur$(W, |W|-1, z, 1, |X|-1)$

CalculateD$(W)$
    $k \leftarrow 1$
    $l \leftarrow |X|-1$
    $z \leftarrow 0$
    for $i = 0$ to $|W|-1$ do
        $k \leftarrow C(W[i]) + O'(W[i], k-1) + 1$
        $l \leftarrow C(W[i]) + O'(W[i], l)$
        if $k > l$ then
            $k \leftarrow 1$
            $l \leftarrow |X|-1$
            $z \leftarrow z+1$
        $D(i) \leftarrow z$

InexRecur$(W, i, z, k, l)$
    if $z < D(i)$ then
        return $\emptyset$
    if $i < 0$ then
        return $\{[k, l]\}$
    $I \leftarrow \emptyset$
    $I \leftarrow I \cup$ InexRecur$(W, i-1, z-1, k, l)$
    for each $b \in \{A, C, G, T\}$ do
        $k \leftarrow C(b) + O(b, k-1) + 1$
        $l \leftarrow C(b) + O(b, l)$
        if $k \leq l$ then
            $I \leftarrow I \cup$ InexRecur$(W, i, z-1, k, l)$
            if $b = W[i]$ then
                $I \leftarrow I \cup$ InexRecur$(W, i-1, z, k, l)$
            else
                $I \leftarrow I \cup$ InexRecur$(W, i-1, z-1, k, l)$
    return $I$
```

NO!

Plumbing!

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.

Agenda

3:00	Introduction
3:10	Software Visualization GBrowse JBrowse GBrowse_syn Sybil & SynView CMap
4:00	Software, cont. Data Management Chado, Tripal, GMODWeb BioMart and InterMine GFF3 Annotation MAKER & DIYA Apollo Textpresso Community Annotation Pipelines & Workflows
4:40	Community
5:00	Finish

Software

GMOD components can be categorized as

- V Visualization
- D Data Management
- A Annotation

Software

You have

- Sequence
- Gene models
- Mapping data
- Alternative transcripts
- Expression
- SNP / variation
- Methylation
- GO terms
- Stocks / lines
- Publications / Attribution
- Orthology

GMOD Has

- A MAKER
- A DIYA
- A Galaxy
- A Ergatis

- A Textpresso
- A Apollo
- V A Table Edit

- V GBrowse
- V JBrowse
- V CMap
- V GBrowse_syn
- V Sybil
- V SynView

- D Chado
- A V Tripal
- V GMODWeb

- D BioMart
- D InterMine

A Annotation   D Data Management   V Visualization

GMOD Requirements

- Server
  - – Most use Linux
- GMOD Systems Administrator
  - – Understands Linux package management, a scripting language, command line interfaces, relational databases, …
  - – Grad/Undergrad, half time when starting up.

http://gmod.org/wiki/Computing_Requirements

GBrowse

GMOD's leading
genome browser

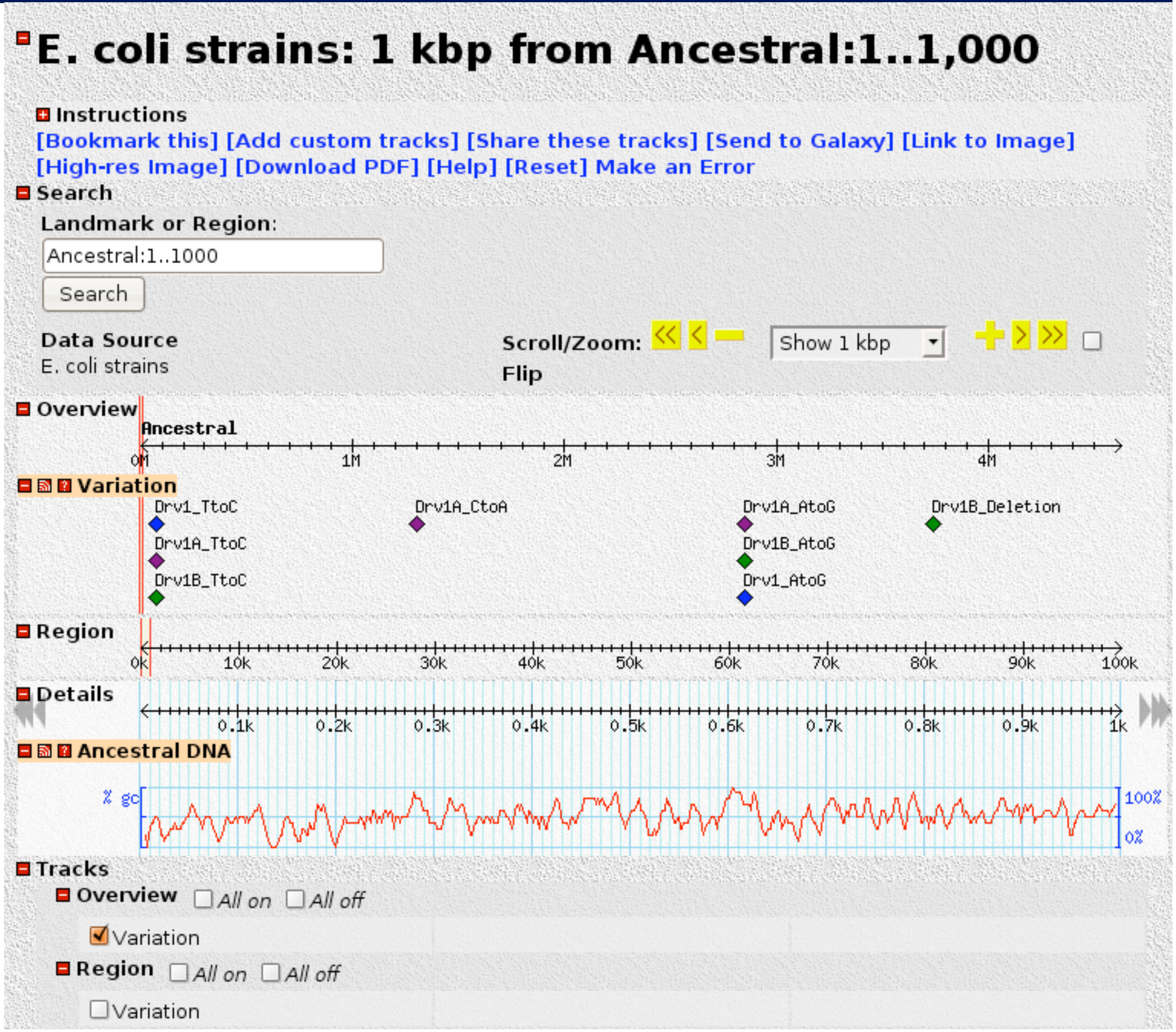
Landing page for *E. coli* example

Overview:
chromosome / contig wide

Region:
intermediate zoom

Details:
current area

Tracks:
current configuration



E. coli strains: 1 kbp from Ancestral:1..1,000

Instructions

[Bookmark this] [Add custom tracks] [Share these tracks] [Send to Galaxy] [Link to Image] [High-res Image] [Download PDF] [Help] [Reset] Make an Error

Search

Landmark or Region: Ancestral:1..1000

Data Source: E. coli strains

Scroll/Zoom: Show 1 kbp

Flip

Tracks

Overview □ All on □ All off

☑ Variation

Region □ All on □ All off

□ Variation

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

GBrowse Example: modENCODE

- Uses GBrowse 2

http://www.modencode.org/gb2/gbrowse/fly/

GBrowse Tutorials

- GBrowse User Tutorial at OpenHelix
  - Flash based, has handouts, very snazzy and thorough
  - Great resource for your users
- GBrowse Admin Tutorial
  - HTML based, written by Lincoln Stein, mostly
  - Excellent way to learn how to configure GBrowse
- GBrowse Admin Tutorial w/ VMware Image
  - From the 2009 GMOD Summer Schools
  - Gives you a system to start with
- NGS in GBrowse and SAMtools Tutorial
  - From Bioinformatics Australia 2009, October
  - Gives you a system to start with

http://gmod.org/wiki/GBrowse_Tutorial

GBrowse Future Plans

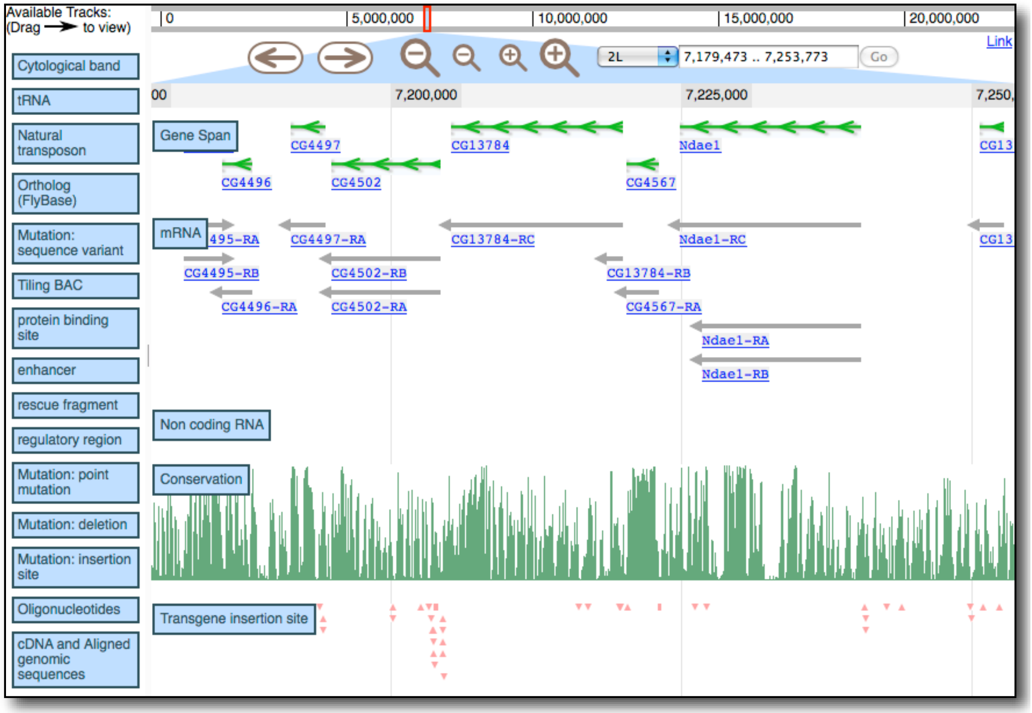
- Circular genome support
  - Work done by Nathan Liles at Texas A&M
- 2.0, Release in 2010
  - Database and rendering multiplexing
  - Asynchronous track loading
  - GBrowse in the cloud
  - User authentication
- 1.x has a few more maintenance releases left.

GBrowse Resources

Home Page	http://gmod.org/wiki/GBrowse
User Tutorial	http://www.openhelix.com/gbrowse
Admin Tutorial	http://gmod.org/wiki/GBrowse_Tutorial
Configuration	http://gmod.org/wiki/GBrowse_Configuration_HOWTO
WebGBrowse	http://webgbrowse.cgb.indiana.edu/
GBrowse.org	http://gbrowse.org
Mailing List	https://lists.sourceforge.net/lists/listinfo/gmod-gbrowse

JBrowse

- GMOD's 2nd generation genome browser
- It's fast
- Completely new
  - Client side rendering
  - Heavily AJAX
  - JSON, Nested Containment Lists



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

JBrowse Demo

http://jbrowse.org

JBrowse Future Plans

- Tools for migrating from GBrowse
- An ecosystem comparable to GBrowse
  - Glyph library, user defined glyphs, callbacks, track sharing, …
- Comparative genomics (more on that later)
- Community Annotation
  - User authentication
  - User uploadable and sharable tracks and annotation

JBrowse Resources

Home Page	http://jbrowse.org
Getting Started	http://jbrowse.org/code/jbrowse-master/docs/tutorial/
Admin Tutorial	http://gmod.org/wiki/JBrowse_Tutorial
Configuration	http://jbrowse.org/code/jbrowse-master/docs/config.html
Demo	http://jbrowse.org/genomes/dmel/
Mailing List	https://lists.sourceforge.net/lists/listinfo/gmod-ajax

GBrowse or JBrowse

GBrowse

Robust ecosystem

Feature rich

Large and growing user base

Track sharing

JBrowse

Very fast

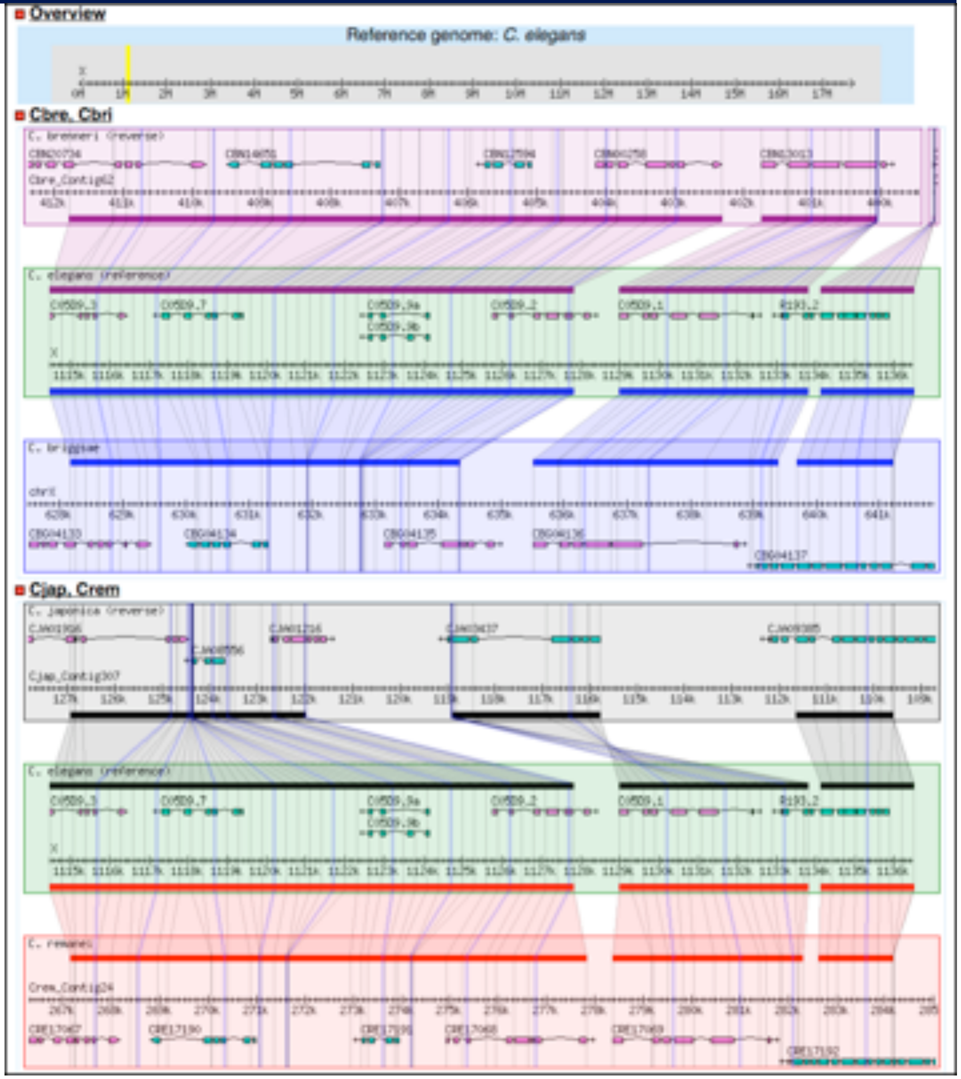
Rapidly growing user base

Lots of future development

Easy to configure

GBrowse_syn

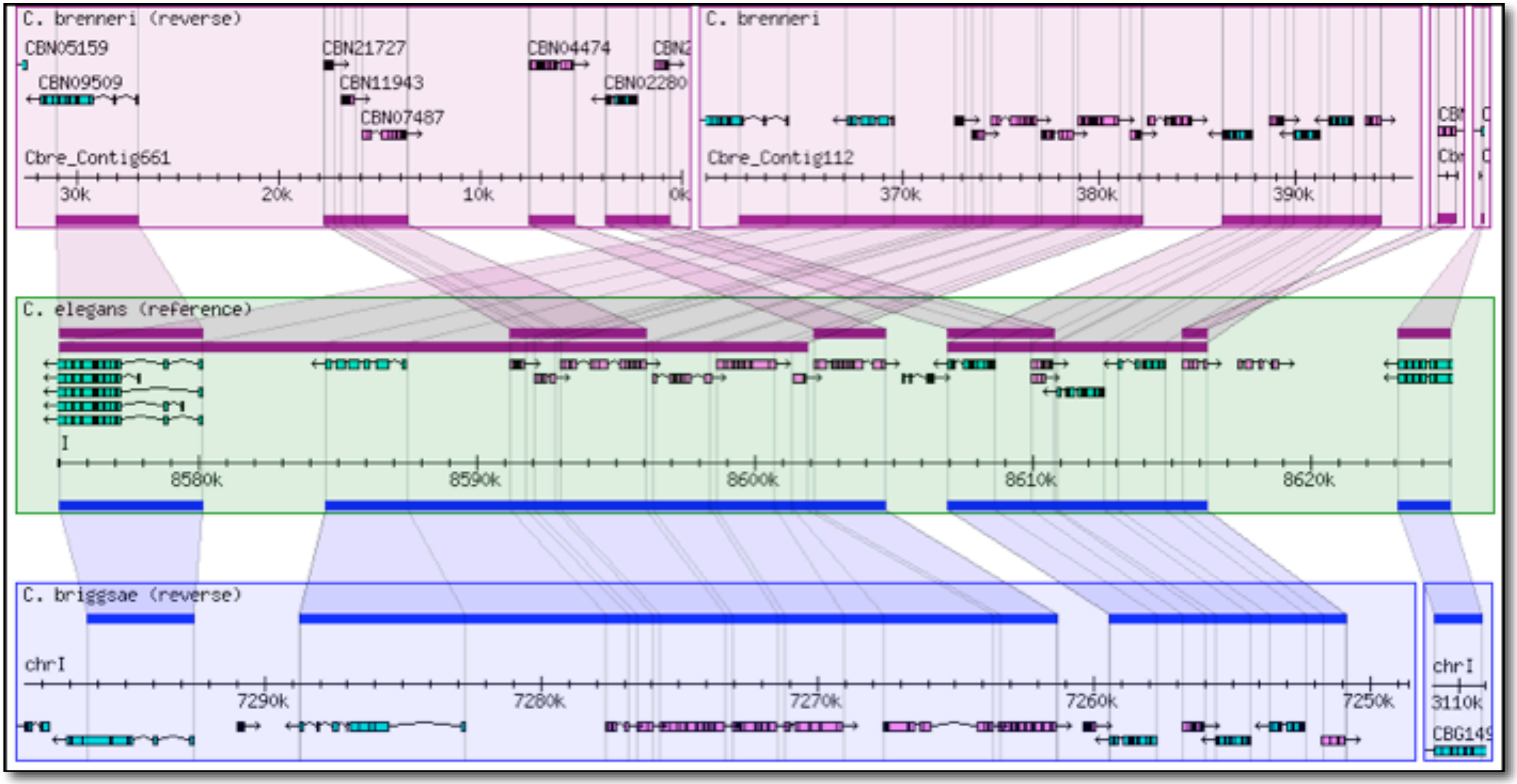
- GBrowse based comparative genomics viewer
- Shows a reference sequence compared to 2 or more others
- Can also show any GBrowse-based annotations



Example comparing *C. elegans* to 4 other species at WormBase

Sheldon McKay, Cold Spring Harbor Laboratory

GBrowse_syn



Syntenic blocks do not have to be colinear

Can also show duplications

Sheldon McKay, Cold Spring Harbor Laboratory

GBrowse_syn Future Work

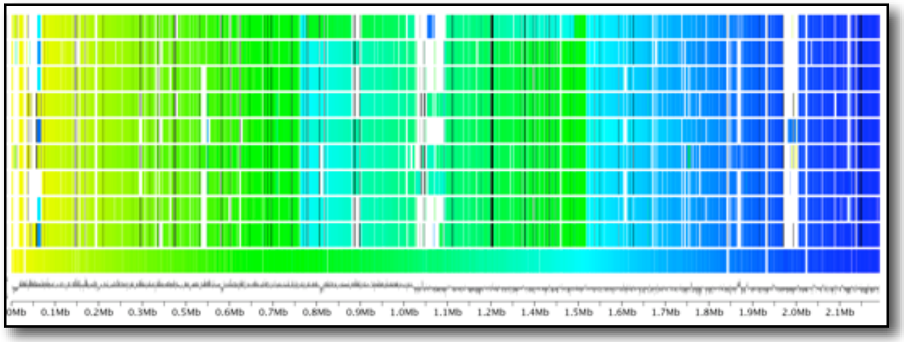
- Integration with GBrowse 2
- High-level graphical overview
- AJAX based user interface and navigation.
  - Submitted grant last month proposing implementing a JBrowse based synteny browser based on GBrowse_syn

GBrowse_syn Resources

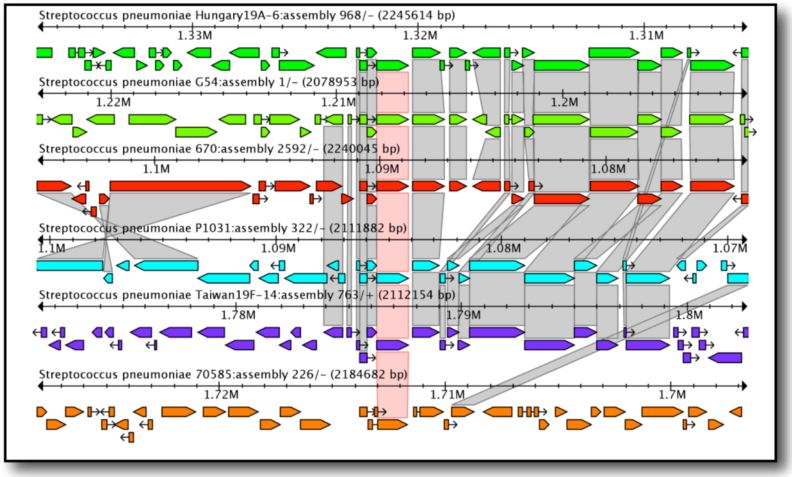
Home Page	http://gmod.org/wiki/GBrowse_syn
Tutorial	http://gmod.org/wiki/GBrowse_syn_Tutorial
User Help	http://gmod.org/wiki/GBrowse_syn_Help
Configuration	http://gmod.org/wiki/GBrowse_syn_Configuration
Example	http://www.wormbase.org/cgi-bin/gbrowse_syn/
Mailing List	https://lists.sourceforge.net/lists/listinfo/gmod-gbrowse

SynView and Sybil

Sybil



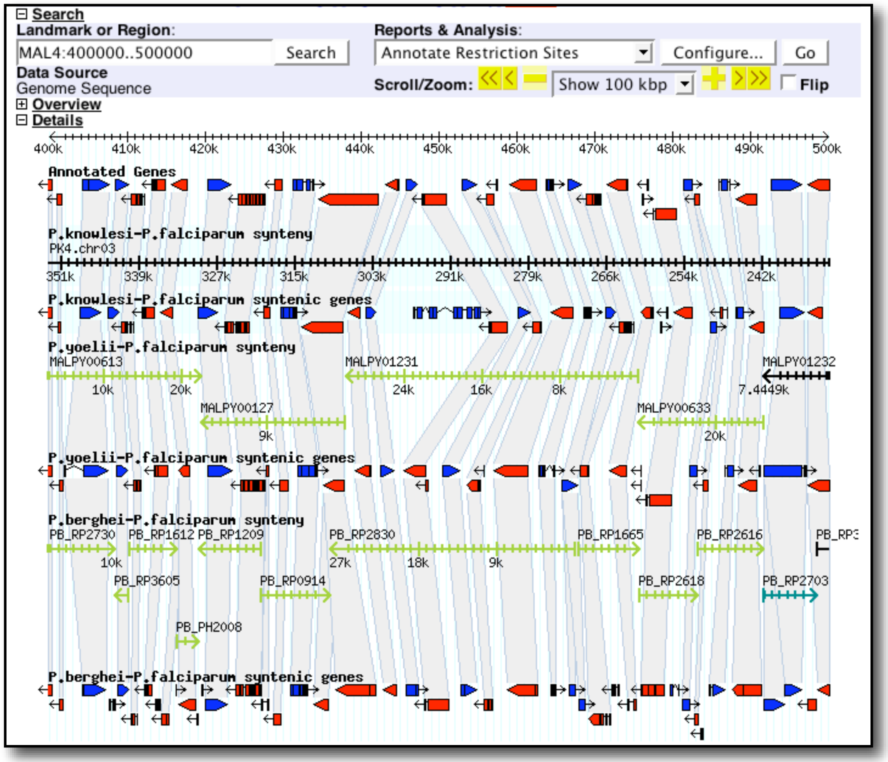
Whole Genome Gradient Display


Streptococcus pneumoniae Hungary19A-6:assembly 968/- (2245614 bp)
1.33M
1.32M
1.31M
Streptococcus pneumoniae G54:assembly 1/- (2078953 bp)
1.22M
1.21M
1.2M
Streptococcus pneumoniae 670:assembly 2592/- (2240045 bp)
1.1M
1.09M
1.08M
Streptococcus pneumoniae P1031:assembly 322/- (2111882 bp)
1.1M
1.09M
1.08M
1.07M
Streptococcus pneumoniae Taiwan19F-14:assembly 763/+ (2112154 bp)
1.78M
1.79M
1.8M
Streptococcus pneumoniae 70585:assembly 226/- (2184682 bp)
1.72M
1.71M
1.7M

Cluster Report

Sybil: Methods and Software for Multiple Genome Comparison and Visualization. Crabtree, *et al.*; in Gene Function Analysis, ed. by Michael F. Ochs (2007)

SynView



Search

Landmark or Region: MAL4:400000..500000

Data Source: Genome Sequence

Reports & Analysis: Annotate Restriction Sites

Scroll/Zoom: Show 100 kbp

SynView: a GBrowse-compatible approach to visualizing comparative genome data. Haiming Wang, *et al.*; in Bioinformatics 22 (18)

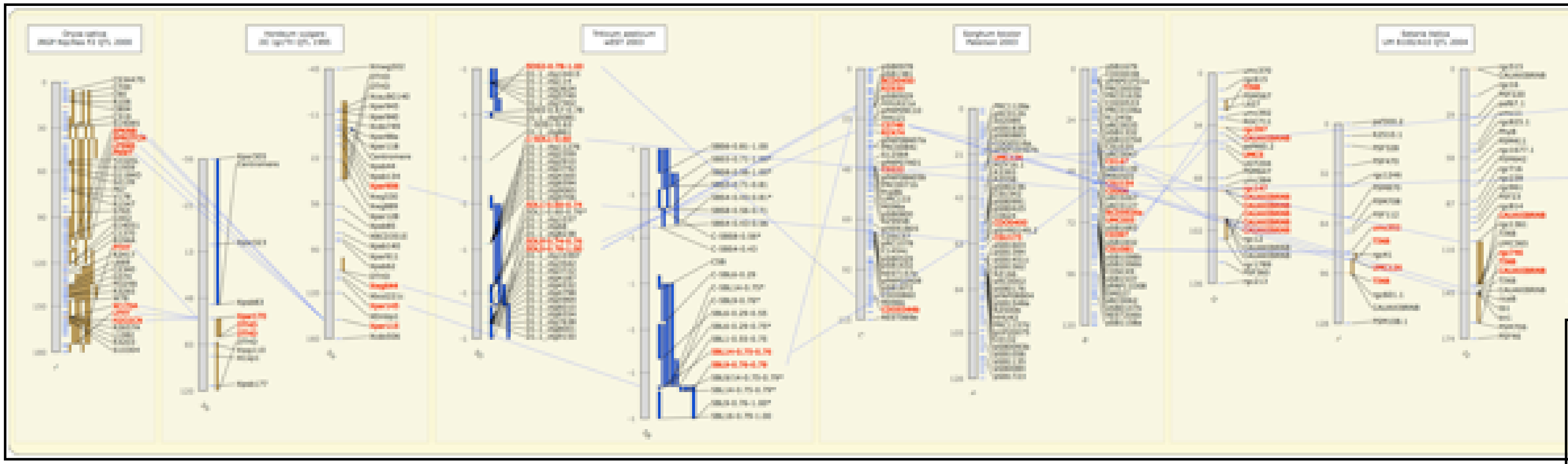
GBrowse_syn or Sybil or SynView?

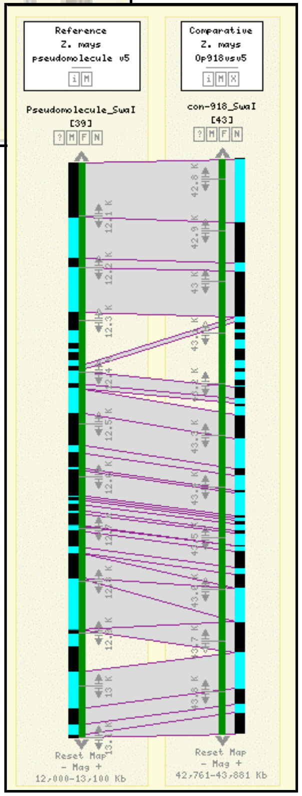
GBrowse_syn
Most actively developed
Scalable
Familiar interface
Extensive documentation
Growing user community

SynView
Scalable
Runs inside GBrowse

Sybil
Scalable
Whole genome and
other unique visualizations
Built on Chado

CMap





Web based comparative map viewer

CMap is data type agnostic: Can link sequence, genetic, physical, QTL, deletion, optical, …

Particularly popular in plant community

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

CMap Future Work

- Streamline the database
- Faster access
- Display in SVG
- Save in Circos / MizBee format
- CMap3D?

CMap Resources

Home Page	http://gmod.org/wiki/CMap
User Tutorial	http://www.gramene.org/tutorials/cmap.html
Admin Guide	http://gmod.svn.sourceforge.net/viewvc/gmod/cmap/trunk/docs/ADMINISTRATION.pod
Example	http://www.gramene.org/cmap/
Mailing List	https://lists.sourceforge.net/lists/listinfo/gmod-cmap

Agenda

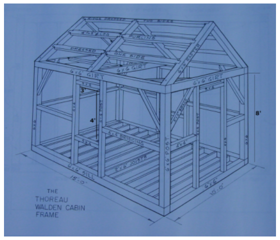
3:00	Introduction
3:10	Software Visualization GBrowse JBrowse GBrowse_syn Sybil & SynView CMap

4:00	Software, cont. Data Management Chado, Tripal, GMODWeb BioMart and InterMine GFF3 Annotation MAKER & DIYA Apollo Textpresso Community Annotation Pipelines & Workflows
4:40	Community
5:00	Finish

Chado: A database schema for biological data

- A *schema* is a database design
  - – Blueprint for a database, a way of organizing data

- Independent of specific data
  - – Chado provides structure
  - – You provide the hard work and data



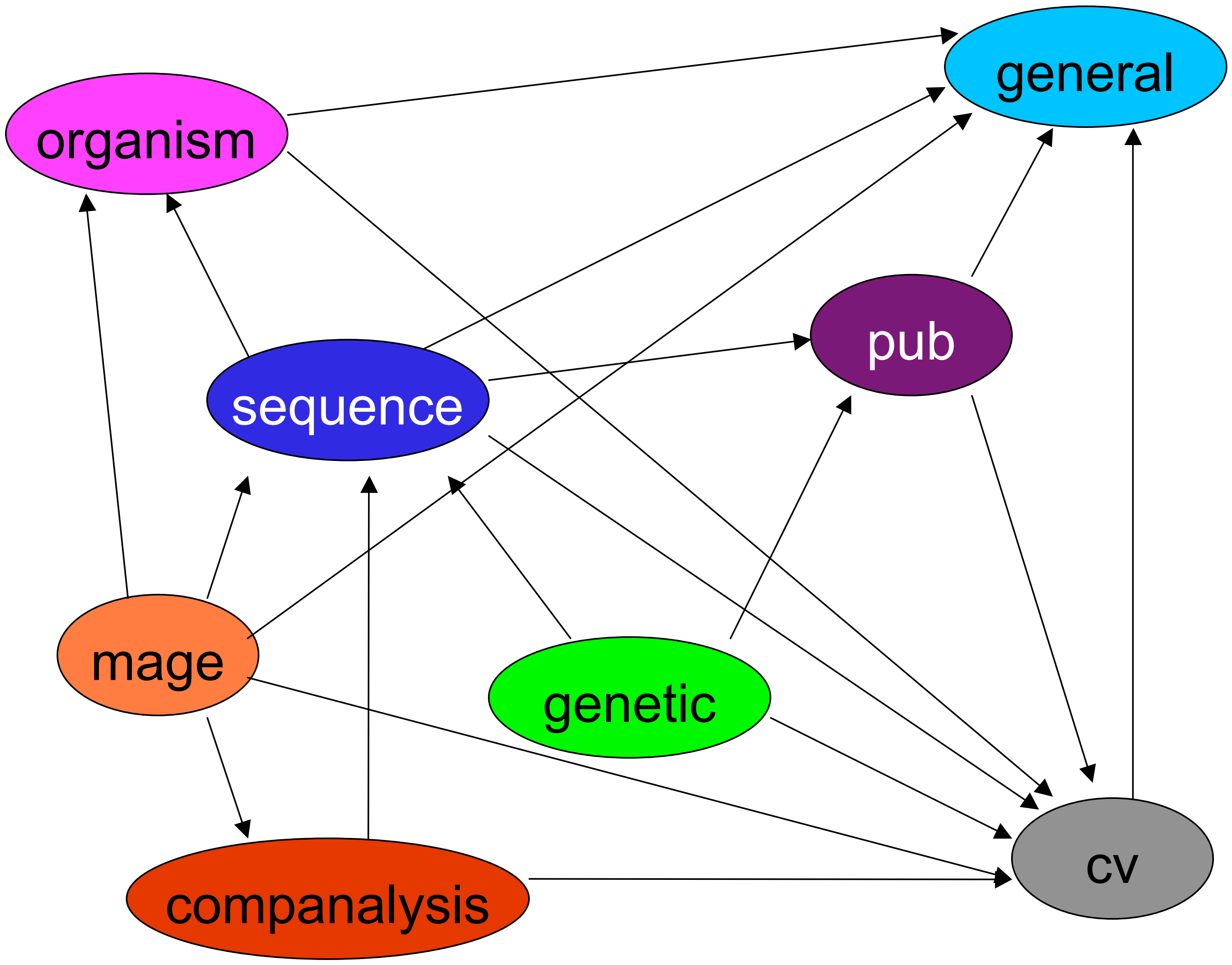
+

=

Why use Chado?

- Very good at genomic data
- Widely used
  - AphidBase, BeetleBase, dictyBase, FlyBase, SGN, SpBase, VectorBase, wFleaBase, …
- Integrates with other GMOD tools
- Community of support
- Modular, flexible and extensible

Chado Modules

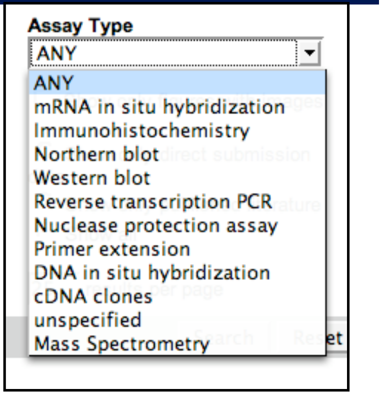


Controlled Vocabularies and Ontologies

Controlled Vocabulary (CV)

List of terms from which a value *must* come

Pull down menus are examples of CVs



Assay Type

ANY

- ANY
- mRNA in situ hybridization
- Immunohistochemistry
- Northern blot
- Western blot
- Reverse transcription PCR
- Nuclease protection assay
- Primer extension
- DNA in situ hybridization
- cDNA clones
- unspecified
- Mass Spectrometry

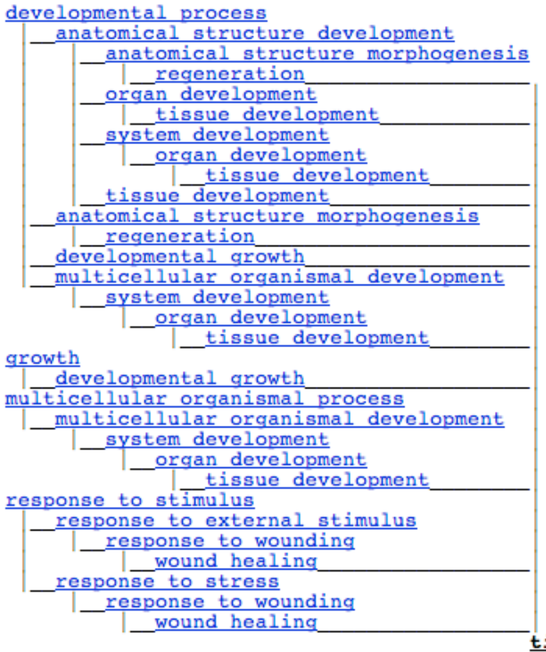
ZFIN assay type CV

Ontology

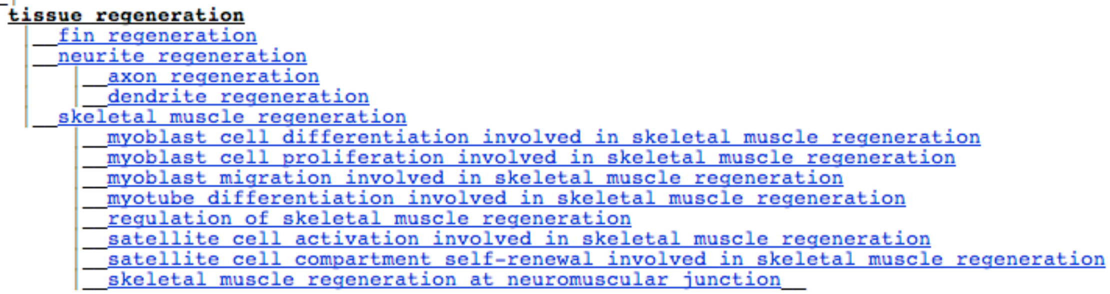
Ontology = CV + rules + relationships between terms

Gene Ontology, Sequence Ontology

Many standard ontologies available from OBO



```
developmental process
|__anatomical structure development
|  |__anatomical structure morphogenesis
|  |  |__regeneration
|  |__organ development
|  |  |__tissue development
|  |__system development
|  |  |__organ development
|  |     |__tissue development
|  |__tissue development
|__anatomical structure morphogenesis
|  |__regeneration
|__developmental growth
|__multicellular organismal development
   |__system development
      |__organ development
         |__tissue development
growth
|__developmental growth
multicellular organismal process
|__multicellular organismal development
   |__system development
      |__organ development
         |__tissue development
response to stimulus
|__response to external stimulus
|  |__response to wounding
|     |__wound healing
|__response to stress
   |__response to wounding
      |__wound healing
```

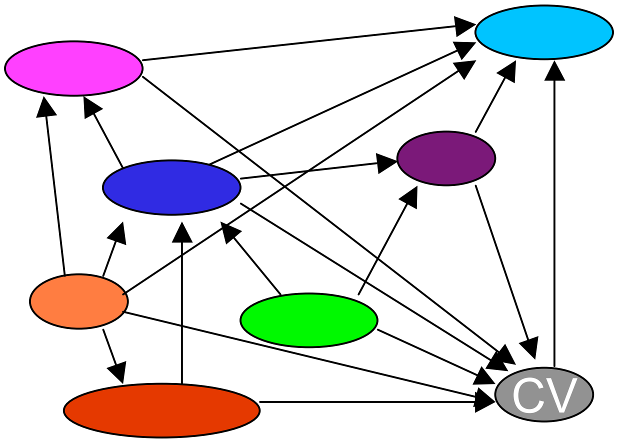


```
tissue regeneration
|__fin regeneration
|__neurite regeneration
|  |__axon regeneration
|  |__dendrite regeneration
|__skeletal muscle regeneration
   |__myoblast cell differentiation involved in skeletal muscle regeneration
   |__myoblast cell proliferation involved in skeletal muscle regeneration
   |__myoblast migration involved in skeletal muscle regeneration
   |__myotube differentiation involved in skeletal muscle regeneration
   |__regulation of skeletal muscle regeneration
   |__satellite cell activation involved in skeletal muscle regeneration
   |__satellite cell compartment self-renewal involved in skeletal muscle regeneration
   |__skeletal muscle regeneration at neuromuscular junction
```

FlyBase CV Term Viewer:
GO term "tissue regeneration"

CVs and Ontologies in Chado

- Controlled vocabularies and ontologies are key in Chado
- Maximally used for
  - – Integrity
  - – Interoperability



- Can create your own, *but* …
  - – Please use standard ontologies when they exist
  - – See OBO: http://www.obofoundry.org/

Chado Future Developments

Flexibility means core schema changes *slowly*

That's a feature.

- Natural Diversity module
  - Better support for phenotypes, crosses, individuals, geolocation, …
  - Based on GDPDM from Cornell University, Terry Casstevens, *et al.* (http://www.maizegenetics.net/gdpdm/)
- Expression / Anatomy / Cell Fate Atlas support
  - Aniseed (http://aniseed-ibdm.univ-mrs.fr/) converting to Chado and extending it to better support atlases
  - Will have a web front end for atlases
-

Chado Resources

Home Page	http://gmod.org/wiki/Chado
Tutorial	http://gmod.org/wiki/Chado_Tutorial
Introduction	http://gmod.org/wiki/Introduction_to_Chado
Manual	http://gmod.org/wiki/Chado_Manual
Modules	http://gmod.org/wiki/GBrowse_Modules
Mailing List	https://lists.sourceforge.net/lists/listinfo/gmod-schema

Chado Web Front Ends

- Chado is a schema, a server side technology
- It is not a web front end or a desktop client

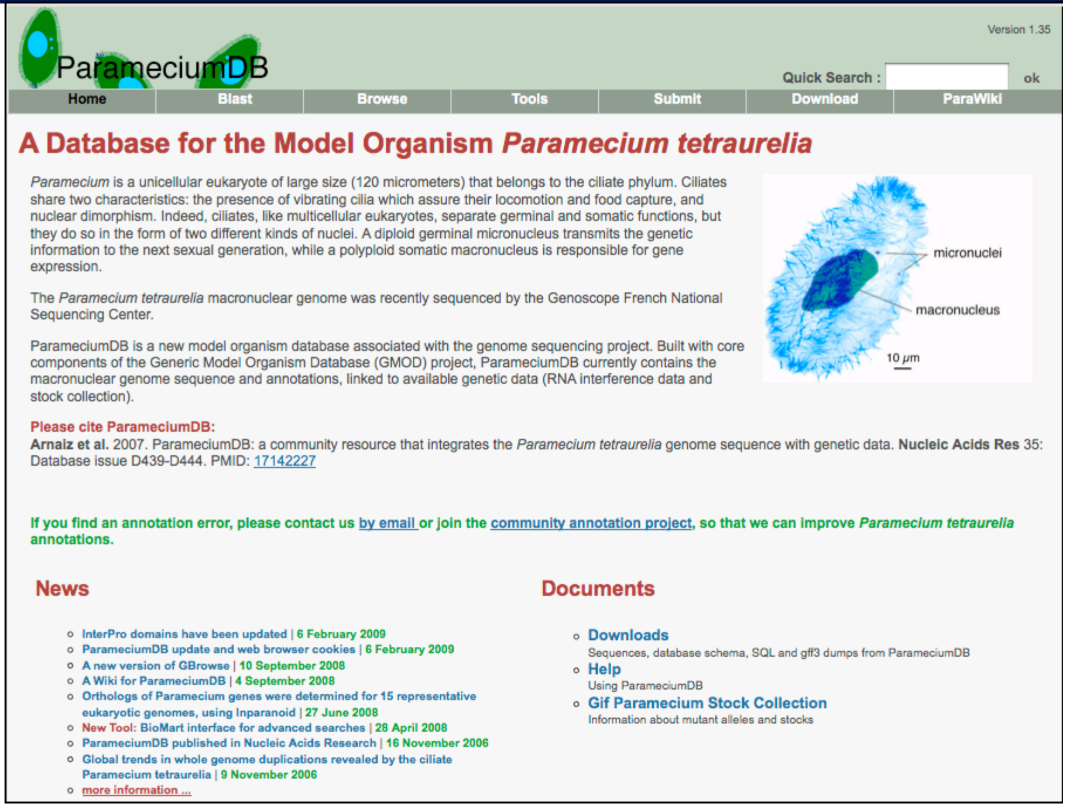
- Options for Chado web front ends:
  - Do it yourself
  - GMODWeb
  - Tripal

Do it yourself

- GMOD provides some support in form of libraries
- Perl
  - Chado::AutoDBI
  - Modware → Bio:Chado:Schema
- Java
  - Two projects under development

GMODWeb

- A Chado specific set of templates for the generic Turnkey web site generation system
- Written in Perl
- Lots of Perl module dependencies



A Database for the Model Organism *Paramecium tetraurelia*

Paramecium is a unicellular eukaryote of large size (120 micrometers) that belongs to the ciliate phylum. Ciliates share two characteristics: the presence of vibrating cilia which assure their locomotion and food capture, and nuclear dimorphism. Indeed, ciliates, like multicellular eukaryotes, separate germinal and somatic functions, but they do so in the form of two different kinds of nuclei. A diploid germinal micronucleus transmits the genetic information to the next sexual generation, while a polyploid somatic macronucleus is responsible for gene expression.

The *Paramecium tetraurelia* macronuclear genome was recently sequenced by the Genoscope French National Sequencing Center.

ParameciumDB is a new model organism database associated with the genome sequencing project. Built with core components of the Generic Model Organism Database (GMOD) project, ParameciumDB currently contains the macronuclear genome sequence and annotations, linked to available genetic data (RNA interference data and stock collection).

Please cite ParameciumDB:
Arnaiz et al. 2007. ParameciumDB: a community resource that integrates the *Paramecium tetraurelia* genome sequence with genetic data. **Nucleic Acids Res** 35: Database issue D439-D444. PMID: 17142227

If you find an annotation error, please contact us by email or join the community annotation project, so that we can improve *Paramecium tetraurelia* annotations.

News

- InterPro domains have been updated | 6 February 2009
- ParameciumDB update and web browser cookies | 6 February 2009
- A new version of GBrowse | 10 September 2008
- A Wiki for ParameciumDB | 4 September 2008
- Orthologs of Paramecium genes were determined for 15 representative eukaryotic genomes, using Inparanoid | 27 June 2008
- New Tool: BioMart interface for advanced searches | 28 April 2008
- ParameciumDB published in Nucleic Acids Research | 16 November 2006
- Global trends in whole genome duplications revealed by the ciliate Paramecium tetraurelia | 9 November 2006
- more information ...

Documents

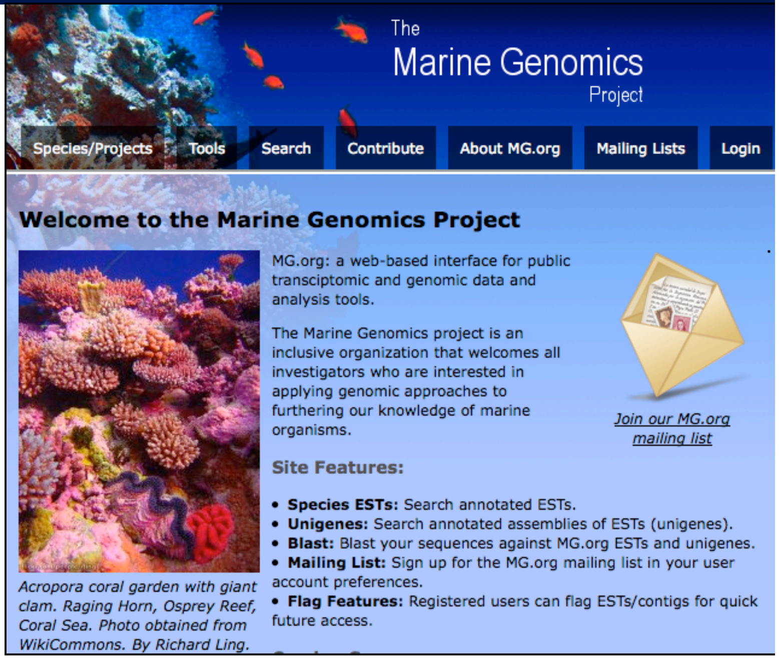
- **Downloads**
  Sequences, database schema, SQL and gff3 dumps from ParameciumDB
- **Help**
  Using ParameciumDB
- **Gif Paramecium Stock Collection**
  Information about mutant alleles and stocks

ParameciumDB, a website built with GMODWeb
http://paramecium.cgm.cnrs-gif.fr/

GMODWeb: a web framework for the generic model organism database, O'Connor *et al*., Genome Biology 2008, 9:R102.

Tripal

- Added to GMOD this year
- Set of Drupal modules
  - Feature, Organism, Library, Analysis
  - Modules roughly correspond to Chado modules
  - Easy to create new modules



Welcome to the Marine Genomics Project

MG.org: a web-based interface for public transciptomic 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.

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.

Acropora coral garden with giant clam. Raging Horn, Osprey Reef, Coral Sea. Photo obtained from WikiCommons. By Richard Ling.

MarineGenomics.org

- Includes user authentication, job management, and data entry support
- Developed by Clemson University Genomics Institute

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

Tripal Resources

Home Page	http://gmod.org/wiki/Tripal
Tutorial	http://gmod.org/wiki/Tripal_Tutorial
User Guide	http://gmod.org/wiki/Media:TripalUsersGuideJune2009.pdf
Example	http://marinegenomics.org
Mailing List	https://lists.sourceforge.net/lists/listinfo/gmod-tripal

Chado Web: DIY or GMODWeb or Tripal?

GMODWeb

Complete

Requires some tuning

Perl

Do It Yourself

More work

Get exactly what you want

Tripal

User authentication

Data entry

Actively developed

Well documented

Easy to extend

Drupal

> What really made us decide to switch over to Drupal was that we needed authentication mechanisms, customized data entry mechanisms, and the ability to add social networking features and other non-biological components to our sites. Drupal supported all of this and was widely used, well documented, and well supported.
>
> Stephen Ficklin, CUGI

BioMart and InterMine

- Chado well-suited for setting up organism databases that have
  - Easy to use query interface to support common types of questions
  - Unified, coherent presentation of information
- BioMart and InterMine
  - Allow users to ask complex queries on all data
  - At the expense of having to do more work

GFF3

- The common file format of GMOD for genomic annotation
- Supported by Chado, GBrowse, JBrowse, CMap, Apollo, ….

Agenda

3:00	Introduction
3:10	Software Visualization GBrowse JBrowse GBrowse_syn Sybil & SynView CMap

4:00	Software, cont. Data Management Chado, Tripal, GMODWeb BioMart and InterMine GFF3 Annotation MAKER & DIYA Apollo Textpresso Community Annotation Pipelines & Workflows
4:40	Community
5:00	Finish

MAKER

- Genome annotation pipeline for creating gene predictions
- 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 Online in beta

MAKER: An easy-to-use annotation pipeline designed for emerging model organism genomes, Brandi L. Cantarel, *et al.*, *Genome Res*. 2008. 18: 188-196

MAKER Resources

Home Page	http://www.yandell-lab.org/software/maker.html
Tutorial	http://gmod.org/wiki/MAKER_Tutorial
PAG Workshop	http://gmod.org/wiki/MAKER_PAG_2010_Workshop
Mailing List	http://yandell-lab.org/mailman/listinfo/maker-devel_yandell-lab.org

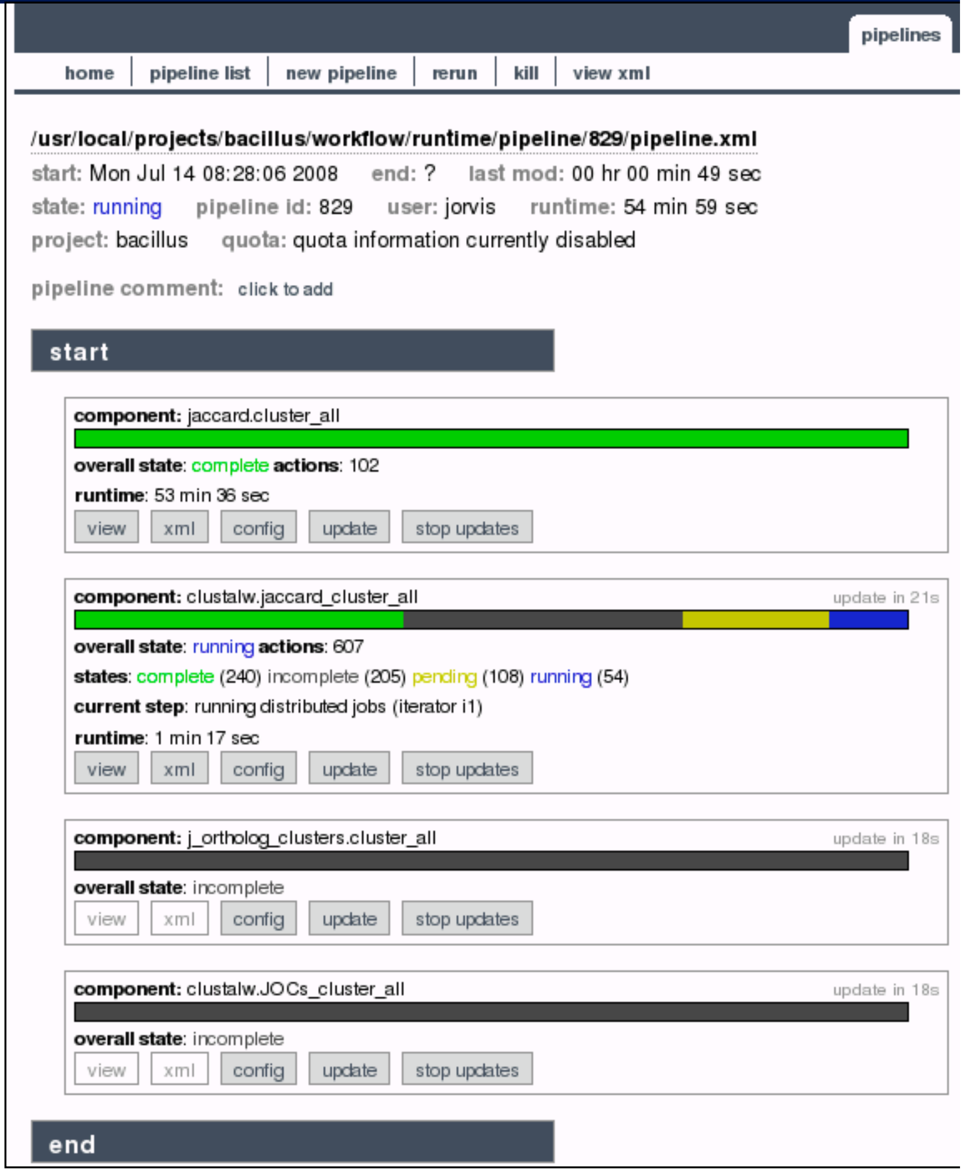
DIYA

- Lightweight, modular, and configurable Perl-based pipeline framework.
- Initial application is gene prediction pipeline 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

Ergatis

- Web interface to the TIGR-Workflow engine
- Create, run and monitor reusable computational analysis pipelines
- Manage compute clusters or single machines
- Comes with several pre-configured pipelines



/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

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

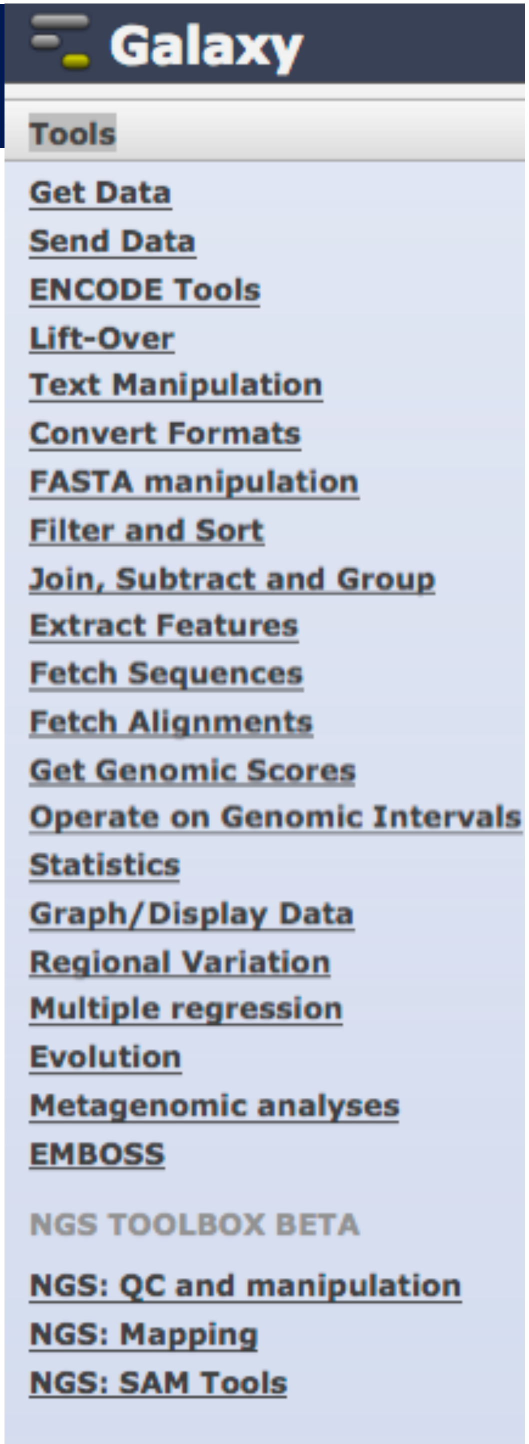
overall state: incomplete

component: clustalw.JOCs_cluster_all update in 18s

overall state: incomplete

Galaxy

- Web portal
  - Search remote resources, combine data from independent queries and visualize results
- Queries / pipelines can be saved and referenced in papers or rerun later.
- Supports set-theory operations on results
- Links to outside tools, including GBrowse
- Can use central server or install locally



Galaxy

Tools

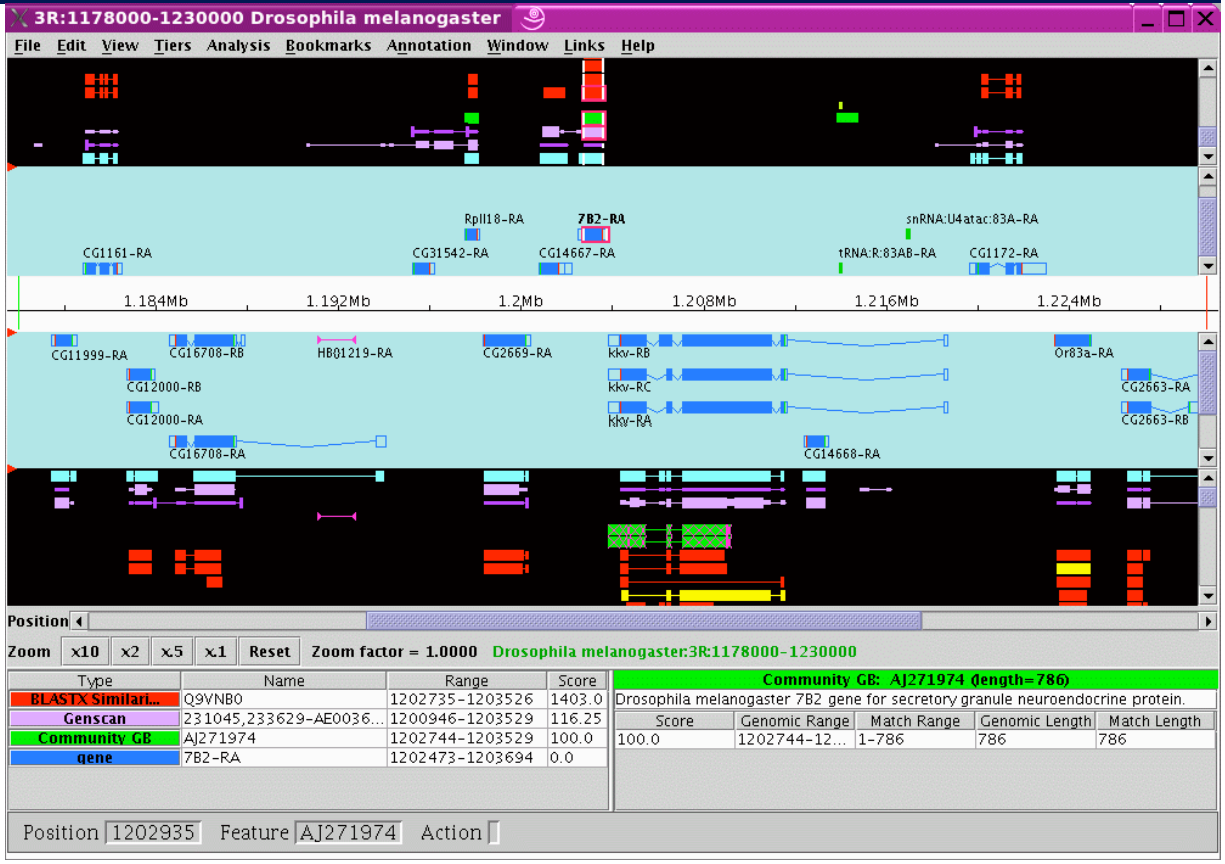
Get Data
Send Data
ENCODE Tools
Lift-Over
Text Manipulation
Convert Formats
FASTA manipulation
Filter and Sort
Join, Subtract and Group
Extract Features
Fetch Sequences
Fetch Alignments
Get Genomic Scores
Operate on Genomic Intervals
Statistics
Graph/Display Data
Regional Variation
Multiple regression
Evolution
Metagenomic analyses
EMBOSS

NGS TOOLBOX BETA

NGS: QC and manipulation
NGS: Mapping
NGS: SAM Tools

Apollo

- GMOD's genome annotation editor
- Add and refine annotations.
- Java desktop client
- Widely used
- Read/write in multiple formats
- Keep track of evidence, curator
- Used in several community annotation efforts



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

Apollo Future Work

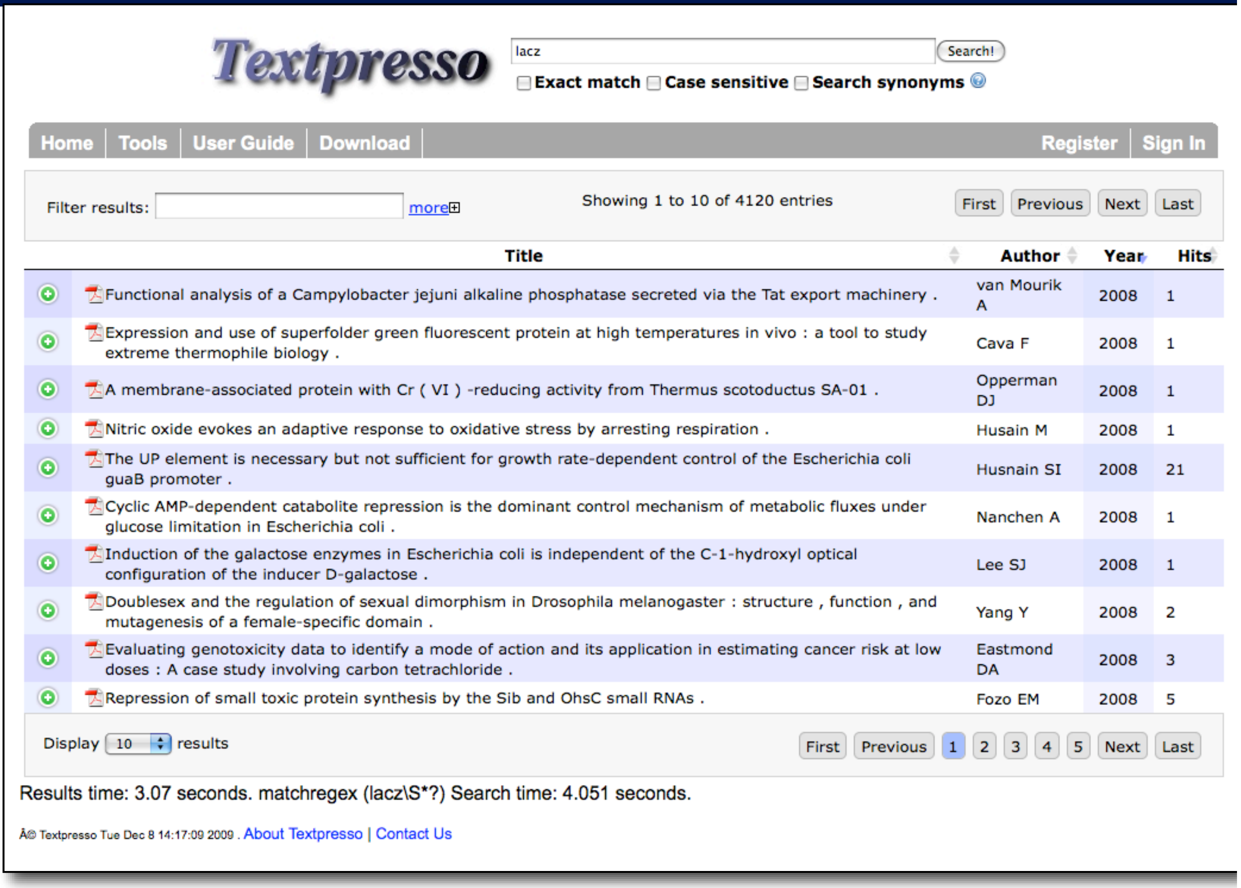
- Berkeley Bioinformatics Open-source Projects (BBOP)
  - Current developers of Apollo
  - Submitted a grant proposal for
    - Apollo on the web
    - Using same underlying tools as JBrowse
- Meanwhile, CCG/ABF
  - Is using Apollo (and Chado) for genome annotation
  - ABF is exploring the possibility of developing a web-based application to complement Apollo
  - NCRIS 5.1 funding for a 6 month project
- These two groups are talking to each other

Apollo Resources

Home Page	http://apollo.berkeleybop.org/
Tutorial	http://gmod.org/wiki/Apollo_Tutorial
User Guide	http://apollo.berkeleybop.org/current/userguide.html
Mailing List	http://mail.fruitfly.org/mailman/listinfo/apollo

Textpresso

- Text mining system for scientific papers
- Analyzes full article text
- Indexes articles by keywords and by category tags
- Stand alone search engine w/ web interface
- Curation tool


Textpresso
Search!
Home | Tools | User Guide | Download
Register | Sign In
more⊞
First | Previous | Next | Last
First | Previous | 1 | 2 | 3 | 4 | 5 | Next | Last

lacz

□ Exact match □ Case sensitive □ Search synonyms

Filter results:

Showing 1 to 10 of 4120 entries

Title	Author	Year	Hits
Functional analysis of a Campylobacter jejuni alkaline phosphatase secreted via the Tat export machinery .	van Mourik A	2008	1
Expression and use of superfolder green fluorescent protein at high temperatures in vivo : a tool to study extreme thermophile biology .	Cava F	2008	1
A membrane-associated protein with Cr (VI) -reducing activity from Thermus scotoductus SA-01 .	Opperman DJ	2008	1
Nitric oxide evokes an adaptive response to oxidative stress by arresting respiration .	Husain M	2008	1
The UP element is necessary but not sufficient for growth rate-dependent control of the Escherichia coli guaB promoter .	Husnain SI	2008	21
Cyclic AMP-dependent catabolite repression is the dominant control mechanism of metabolic fluxes under glucose limitation in Escherichia coli .	Nanchen A	2008	1
Induction of the galactose enzymes in Escherichia coli is independent of the C-1-hydroxyl optical configuration of the inducer D-galactose .	Lee SJ	2008	1
Doublesex and the regulation of sexual dimorphism in Drosophila melanogaster : structure , function , and mutagenesis of a female-specific domain .	Yang Y	2008	2
Evaluating genotoxicity data to identify a mode of action and its application in estimating cancer risk at low doses : A case study involving carbon tetrachloride .	Eastmond DA	2008	3
Repression of small toxic protein synthesis by the Sib and OhsC small RNAs .	Fozo EM	2008	5

Display 10 results

Results time: 3.07 seconds. matchregex (lacz\S*?) Search time: 4.051 seconds.

Â© Textpresso Tue Dec 8 14:17:09 2009 . About Textpresso | Contact Us

Textpresso for *E. coli*, extensions by Nathan Liles, Hu Lab

Textpresso: an ontology-based information retrieval and extraction system for biological literature, Muller HM, Kenny EE, Sternberg PW, *PLoS Biol.* 2004 Nov;2(11):e309

Community Annotation

- How do you get others to contribute?

- Social:
  - Sticks
    - Work well if your database is already the authority on your topic/organism and you have curators and a huge community
  - Carrots
    - Give people credit
    - Give people ownership
    - Seek mutually beneficial relationships
  - Comfort Level
    - Recent popularity of social computing

Community Annotation

- Technological
  - – Make it easier to fix something then it is to be irritated by its error or absence.
    - The Wikipedia model.
  - – Make it relatively easy for people who really care to contribute significant content
    - Also the Wikipedia model.

Community Annotation: GMOD Technology

- Apollo
  - Several projects use Apollo to distribute genome annotation efforts
  - Apollo infrastructure supports:
    - Read from Chado → Save to XML → Review → Upload to Chado
  - But
    - Java application; Infrequent Apollo users forget a lot.
    - Web Apollo will help some, maybe a lot

- Tripal
  - Supports update interfaces for data in Chado databases.
  - Has access to all of Drupal's social networking.

Community Annotation: GMOD Technology

- Table Edit
  - A MediaWiki extension that provides a GUI interface to updating MediaWiki tables.
  - MediaWiki software used at Wikipedia
  - Has been extended to update and render database tables through a MediaWiki interface.
  - Work is in progress to apply it to Chado.
  - See http://ecoliwiki.net
  - Has potential to turn Chado into a wiki.

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

Agenda

3:00	Introduction
3:10	Software Visualization GBrowse JBrowse GBrowse_syn Sybil & SynView CMap

4:00	Software, cont. Data Management Chado, Tripal, GMODWeb BioMart and InterMine GFF3 Annotation MAKER & DIYA Apollo Textpresso Community Annotation Pipelines & Workflows
4:40	Community
5:00	Finish

Who uses GMOD?

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

GMOD

Plus hundreds of others

NESCent

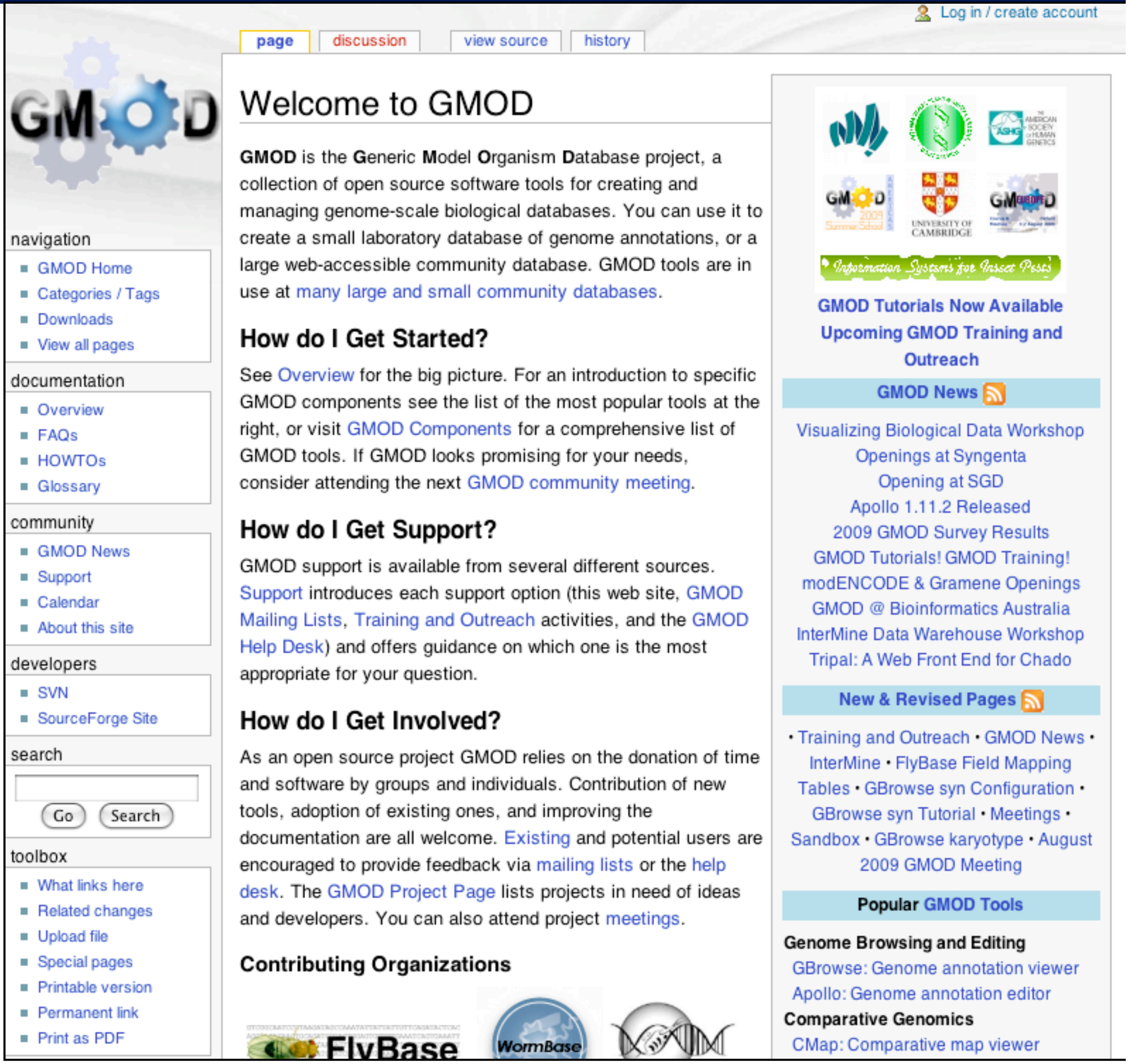
GMOD Project

- Open Source
- Two full time project staff:
  - Project Coordinator: Scott Cain
  - Help Desk: Dave Clements
- Components
  - Some have dedicated funding
  - Others are contributed
  - New components must have:
    - An open source license
    - Interoperability with other GMOD components
    - A good faith commitment of at lest 2 years of support

GMOD.org

A wiki, of course. GMOD.org is the hub for all things related to the project:

- Documentation
- News
- Links
- Calendar
- Tutorials
- HOWTOs
- Glossary
- Overview
- …



Log in / create account

page | discussion | view source | history

navigation

- GMOD Home
- Categories / Tags
- Downloads
- View all pages

documentation

- Overview
- FAQs
- HOWTOs
- Glossary

community

- GMOD News
- Support
- Calendar
- About this site

developers

- SVN
- SourceForge Site

search

Go | Search

toolbox

- What links here
- Related changes
- Upload file
- Special pages
- Printable version
- Permanent link
- Print as PDF

Welcome to GMOD

GMOD is the **G**eneric **M**odel **O**rganism **D**atabase project, a collection of open source software tools for creating and managing genome-scale biological databases. You can use it to create a small laboratory database of genome annotations, or a large web-accessible community database. GMOD tools are in use at many large and small community databases.

How do I Get Started?

See Overview for the big picture. For an introduction to specific GMOD components see the list of the most popular tools at the right, or visit GMOD Components for a comprehensive list of GMOD tools. If GMOD looks promising for your needs, consider attending the next GMOD community meeting.

How do I Get Support?

GMOD support is available from several different sources. Support introduces each support option (this web site, GMOD Mailing Lists, Training and Outreach activities, and the GMOD Help Desk) and offers guidance on which one is the most appropriate for your question.

How do I Get Involved?

As an open source project GMOD relies on the donation of time and software by groups and individuals. Contribution of new tools, adoption of existing ones, and improving the documentation are all welcome. Existing and potential users are encouraged to provide feedback via mailing lists or the help desk. The GMOD Project Page lists projects in need of ideas and developers. You can also attend project meetings.

Contributing Organizations

FlyBase WormBase

Information Systems for Insect Pests

GMOD Tutorials Now Available

Upcoming GMOD Training and Outreach

GMOD News

Visualizing Biological Data Workshop
Openings at Syngenta
Opening at SGD
Apollo 1.11.2 Released
2009 GMOD Survey Results
GMOD Tutorials! GMOD Training!
modENCODE & Gramene Openings
GMOD @ Bioinformatics Australia
InterMine Data Warehouse Workshop
Tripal: A Web Front End for Chado

New & Revised Pages

• Training and Outreach • GMOD News • InterMine • FlyBase Field Mapping Tables • GBrowse syn Configuration • GBrowse syn Tutorial • Meetings • Sandbox • GBrowse karyotype • August 2009 GMOD Meeting

Popular GMOD Tools

Genome Browsing and Editing

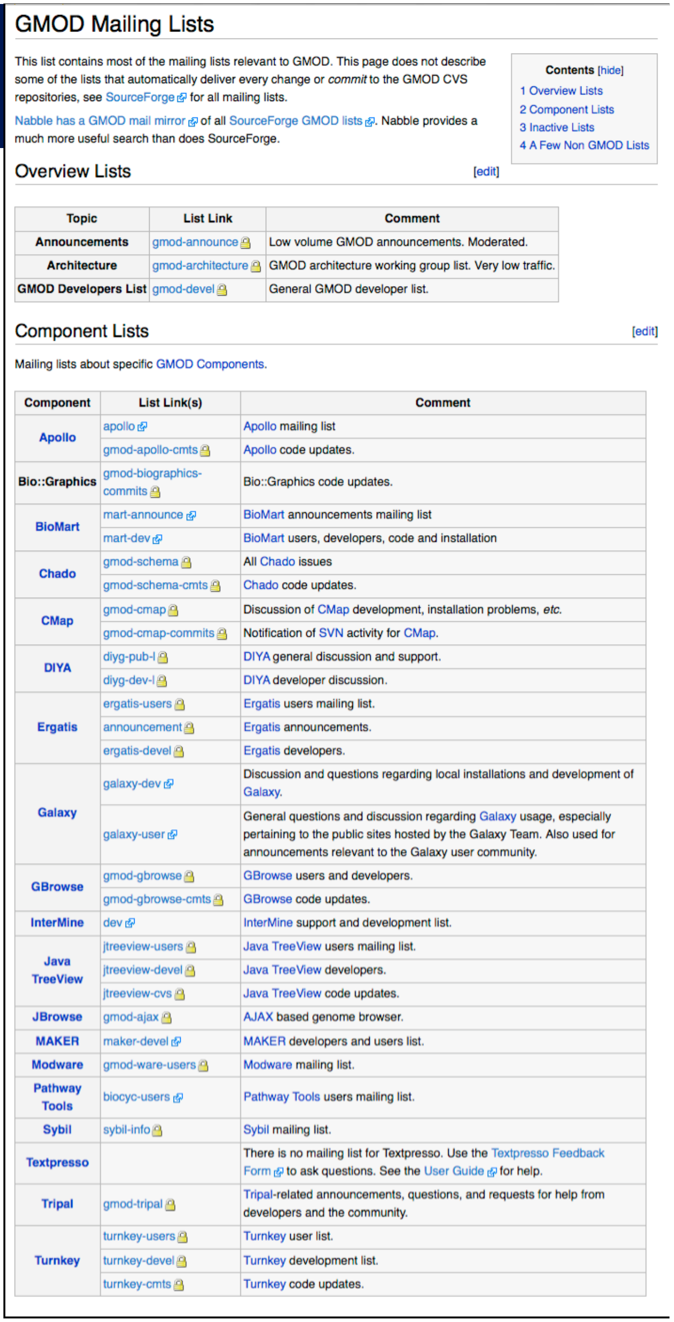
GBrowse: Genome annotation viewer
Apollo: Genome annotation editor

Comparative Genomics

CMap: Comparative map viewer

Mailing Lists

- Several project lists
- Many component-specific lists
- 3100 messages in last 12 months on the 7 lists managed by GMOD staff
- Up 69% from previous year
- Mailing lists are very active



GMOD Mailing Lists

This list contains most of the mailing lists relevant to GMOD. This page does not describe some of the lists that automatically deliver every change or *commit* to the GMOD CVS repositories, see SourceForge for all mailing lists.

Nabble has a GMOD mail mirror of all SourceForge GMOD lists. Nabble provides a much more useful search than does SourceForge.

Contents [hide]

Overview Lists

[edit]

Topic	List Link	Comment
Announcements	gmod-announce	Low volume GMOD announcements. Moderated.
Architecture	gmod-architecture	GMOD architecture working group list. Very low traffic.
GMOD Developers List	gmod-devel	General GMOD developer list.

Component Lists

[edit]

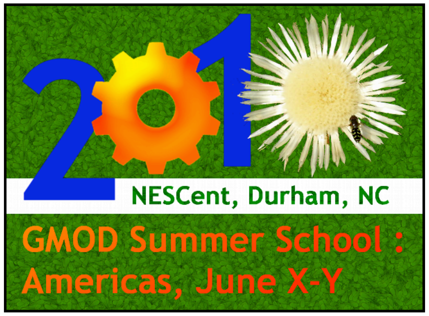
Mailing lists about specific GMOD Components.

Component	List Link(s)	Comment
Apollo	apollo	Apollo mailing list
	gmod-apollo-cmts	Apollo code updates.
Bio::Graphics	gmod-biographics-commits	Bio::Graphics code updates.
BioMart	mart-announce	BioMart announcements mailing list
	mart-dev	BioMart users, developers, code and installation
Chado	gmod-schema	All Chado issues
	gmod-schema-cmts	Chado code updates.
CMap	gmod-cmap	Discussion of CMap development, installation problems, *etc.*
	gmod-cmap-commits	Notification of SVN activity for CMap.
DIYA	diyg-pub-l	DIYA general discussion and support.
	diyg-dev-l	DIYA developer discussion.
Ergatis	ergatis-users	Ergatis users mailing list.
	announcement	Ergatis announcements.
	ergatis-devel	Ergatis developers.
Galaxy	galaxy-dev	Discussion and questions regarding local installations and development of Galaxy.
	galaxy-user	General questions and discussion regarding Galaxy usage, especially pertaining to the public sites hosted by the Galaxy Team. Also used for announcements relevant to the Galaxy user community.
GBrowse	gmod-gbrowse	GBrowse users and developers.
	gmod-gbrowse-cmts	GBrowse code updates.
InterMine	dev	InterMine support and development list.
Java TreeView	jtreeview-users	Java TreeView users mailing list.
	jtreeview-devel	Java TreeView developers.
	jtreeview-cvs	Java TreeView code updates.
JBrowse	gmod-ajax	AJAX based genome browser.
MAKER	maker-devel	MAKER developers and users list.
Modware	gmod-ware-users	Modware mailing list.
Pathway Tools	biocyc-users	Pathway Tools users mailing list.
Sybil	sybil-info	Sybil mailing list.
Textpresso		There is no mailing list for Textpresso. Use the Textpresso Feedback Form to ask questions. See the User Guide for help.
Tripal	gmod-tripal	Tripal-related announcements, questions, and requests for help from developers and the community.
Turnkey	turnkey-users	Turnkey user list.
	turnkey-devel	Turnkey development list.
	turnkey-cmts	Turnkey code updates.

http://gmod.org/wiki/GMOD_Mailing_Lists

Meetings, Training and Outreach

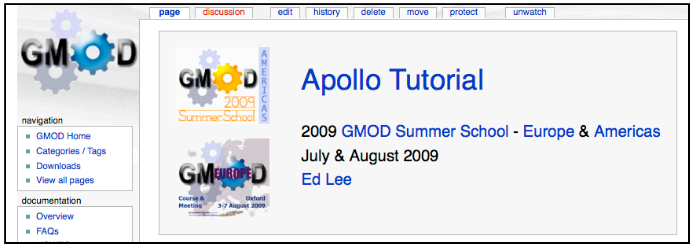
- Semi-annual community meetings
  - Next Meeting:
    - January 2010, San Diego, after PAG
- GMOD Summer Schools
  - 2009
    - July, NESCent, North Carolina, US
    - August, Oxford, UK
  - 2010
    - ??, NESCent, North Carolina, US
    - ??, Asia / Pacific, maybe
- Outreach
  - BA, SMBE, PAG, Arthropod Genomics, …



http://gmod.org/wiki/Training_and_Outreach

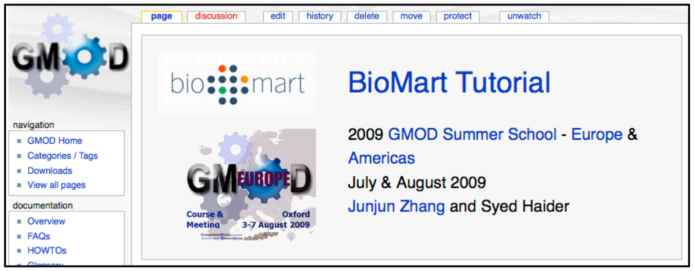
Tutorials

- Summer school sessions become online tutorials with
  - Starting VMware images
  - Step by step instructions
  - Example datasets
  - Ending VMware images
- Topics:
  - Apollo, Artemis-Chado Integration, BioMart, Chado, CMap, GBrowse, GBrowse_syn, JBrowse, MAKER, Tripal, GBrowse NGS



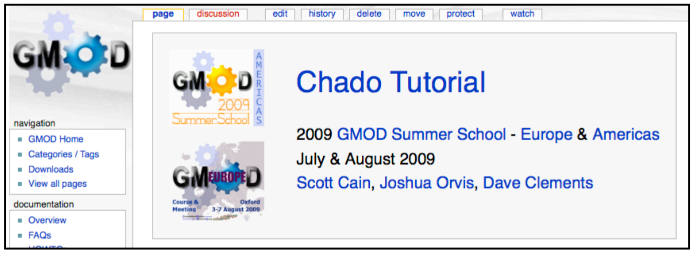
Apollo Tutorial

2009 GMOD Summer School - Europe & Americas
July & August 2009
Ed Lee



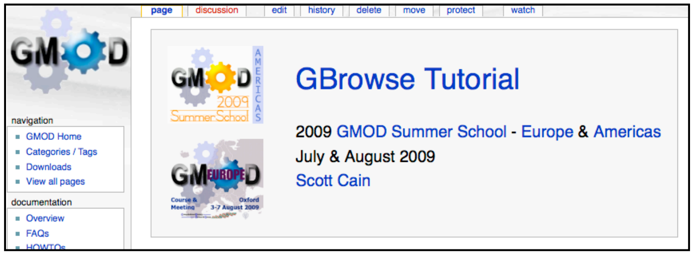
BioMart Tutorial

2009 GMOD Summer School - Europe & Americas
July & August 2009
Junjun Zhang and Syed Haider



Chado Tutorial

2009 GMOD Summer School - Europe & Americas
July & August 2009
Scott Cain, Joshua Orvis, Dave Clements



GBrowse Tutorial

2009 GMOD Summer School - Europe & Americas
July & August 2009
Scott Cain

⋮

http://gmod.org/wiki/Training_and_Outreach#Online_Tutorials

Agenda

3:00	Introduction
3:10	Software Visualization GBrowse JBrowse GBrowse_syn Sybil & SynView CMap

4:00	Software, cont. Data Management Chado, Tripal, GMODWeb BioMart and InterMine GFF3 Annotation MAKER & DIYA Apollo Textpresso Community Annotation Pipelines & Workflows
4:40	Community
5:00	Finish

Acknowledgements

NESCent
Todd Vision
Hilmar Lapp

BBOP
Ed Lee

OICR
Scott Cain
Lincoln Stein

Oregon
Patrick Phillips
Phillips Lab

Texas A&M
Jim Hu
Daniel Renfro
Nathan Liles
Brenley McIntosh

CBRG, Oxford
Simon McGowan

Broad
Heng Li

CUGI
Stephen Ficklin

CSHL
Sheldon McKay
Ken Youens-Clark

Thank You!

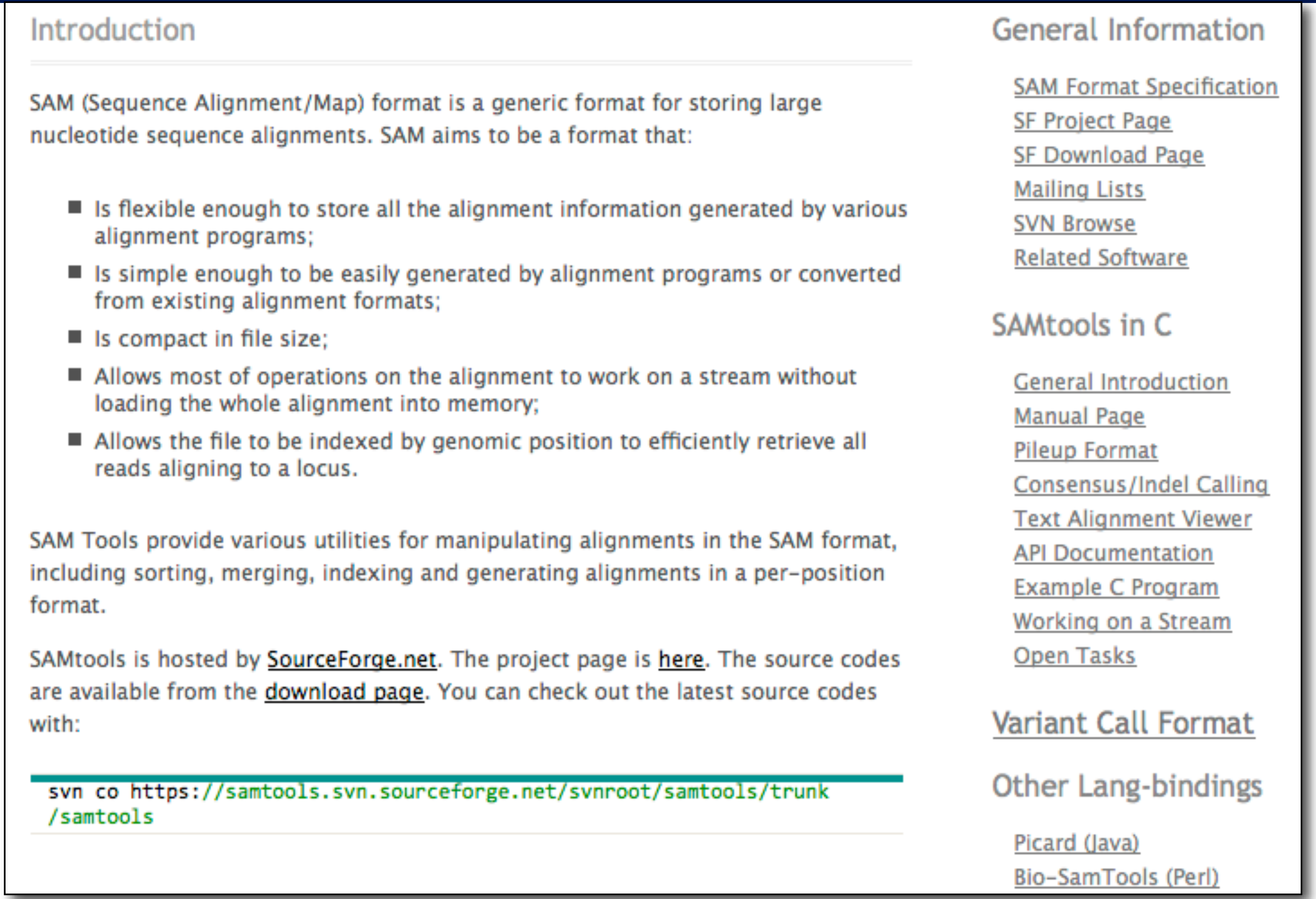
Dave Clements
GMOD Help Desk

US National Evolutionary
Synthesis Center
http://nescent.org

clements@nescent.org
help@gmod.org

http://gmod.org/wiki/GMOD_Help_Desk

SAMtools



Introduction

SAM (Sequence Alignment/Map) format is a generic format for storing large nucleotide sequence alignments. SAM aims to be a format that:

- Is flexible enough to store all the alignment information generated by various alignment programs;
- Is simple enough to be easily generated by alignment programs or converted from existing alignment formats;
- Is compact in file size;
- Allows most of operations on the alignment to work on a stream without loading the whole alignment into memory;
- Allows the file to be indexed by genomic position to efficiently retrieve all reads aligning to a locus.

SAM Tools provide various utilities for manipulating alignments in the SAM format, including sorting, merging, indexing and generating alignments in a per-position format.

SAMtools is hosted by SourceForge.net. The project page is here. The source codes are available from the download page. You can check out the latest source codes with:

```
svn co https://samtools.svn.sourceforge.net/svnroot/samtools/trunk
/samtools
```

General Information

- SAM Format Specification
- SF Project Page
- SF Download Page
- Mailing Lists
- SVN Browse
- Related Software

SAMtools in C

- General Introduction
- Manual Page
- Pileup Format
- Consensus/Indel Calling
- Text Alignment Viewer
- API Documentation
- Example C Program
- Working on a Stream
- Open Tasks

Variant Call Format

Other Lang-bindings

- Picard (Java)
- Bio-SamTools (Perl)

Platform neutral set of programs and file formats specifically for short reads.

Heng Li, *et al.*, The Sequence Alignment/Map format and SAMtools, *Bioinformatics*. 2009 Aug 15;25(16):2078-9.

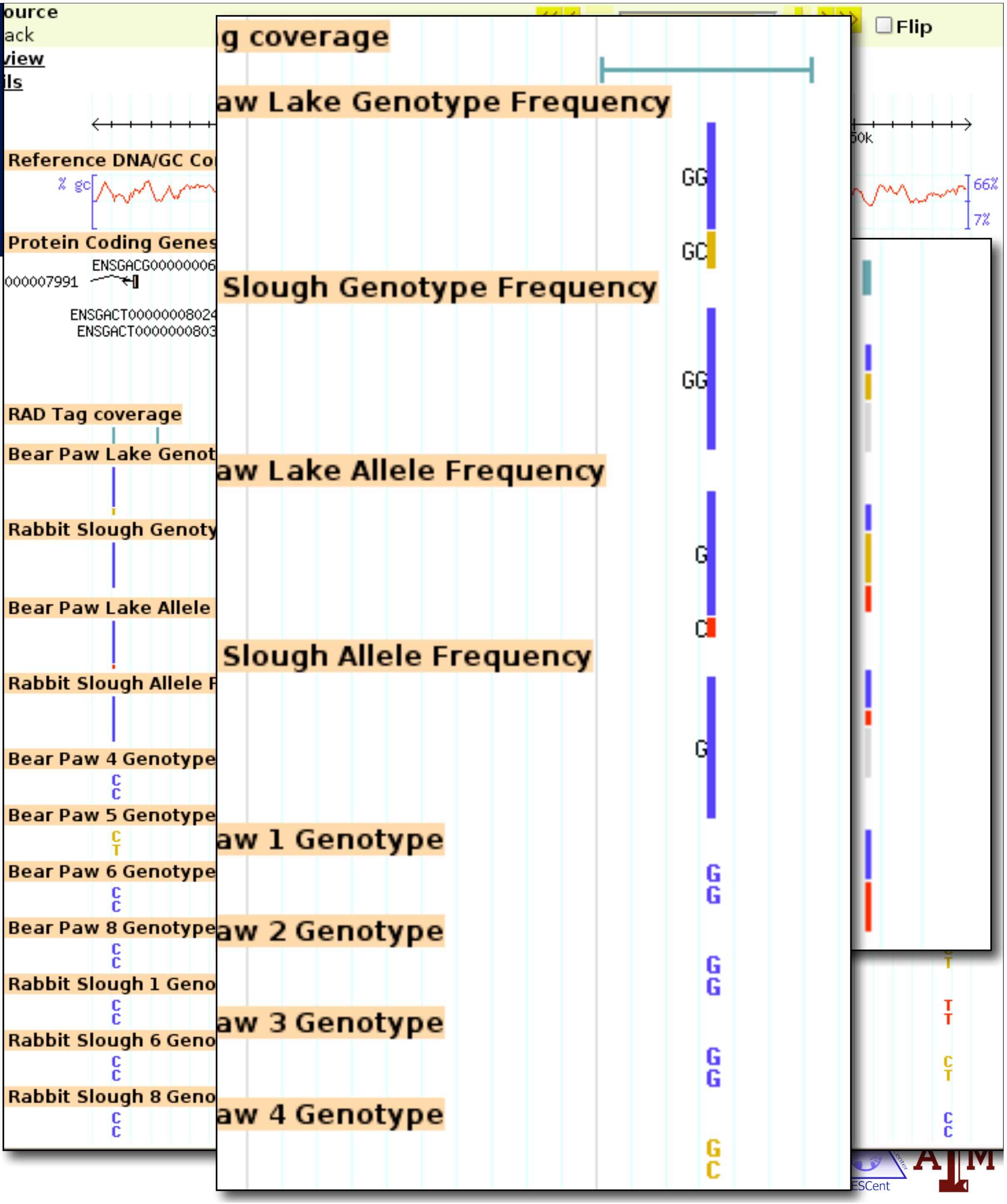
GBrowse for Population Genetics

Shows

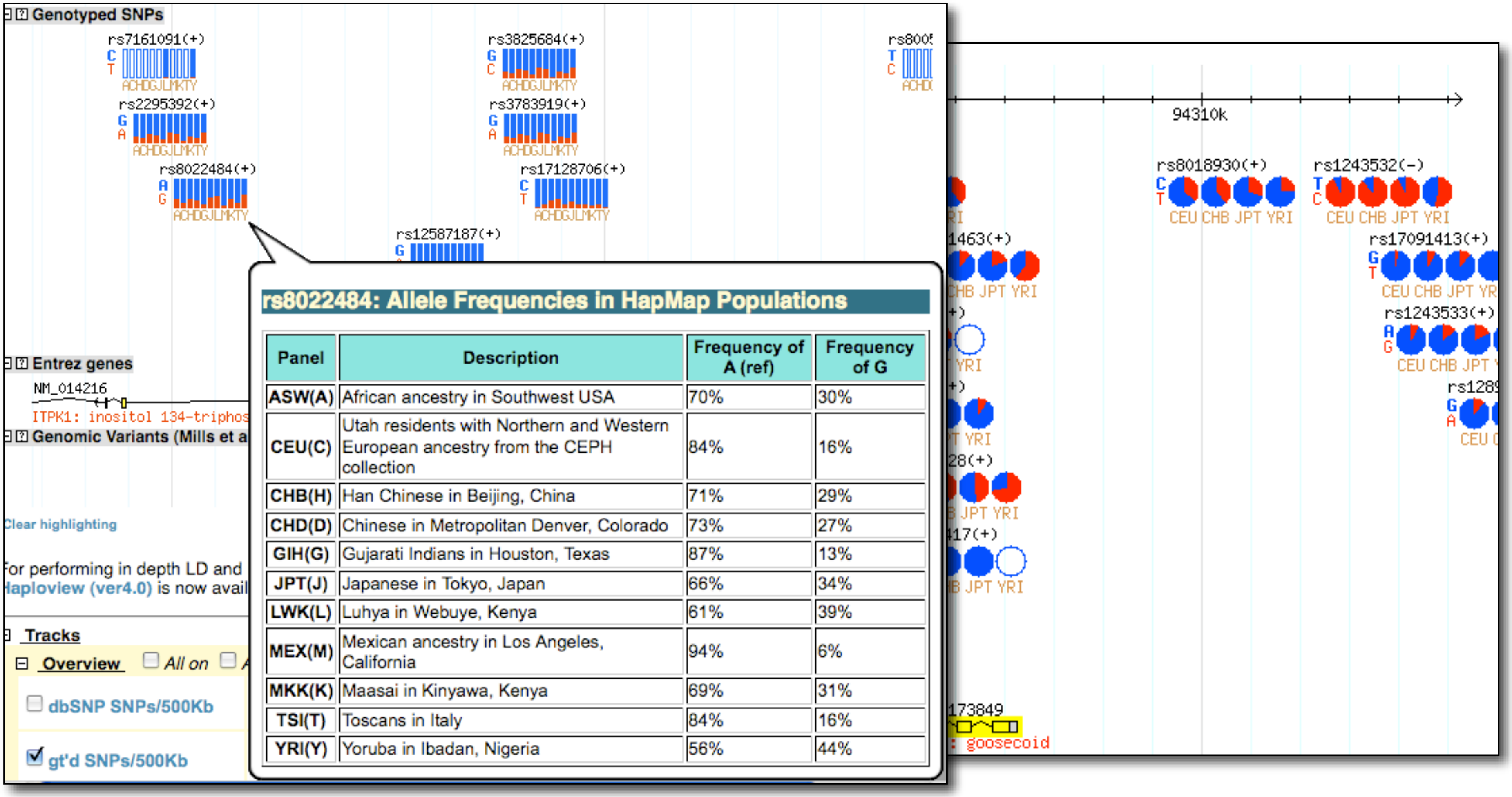
- Where we looked
- Allele & genotype frequencies
  - By population
  - Individual genotypes

Could also show:

- Frequency by phenotype or any other characteristic
- Sliding window stats



HapMap Allele Frequencies



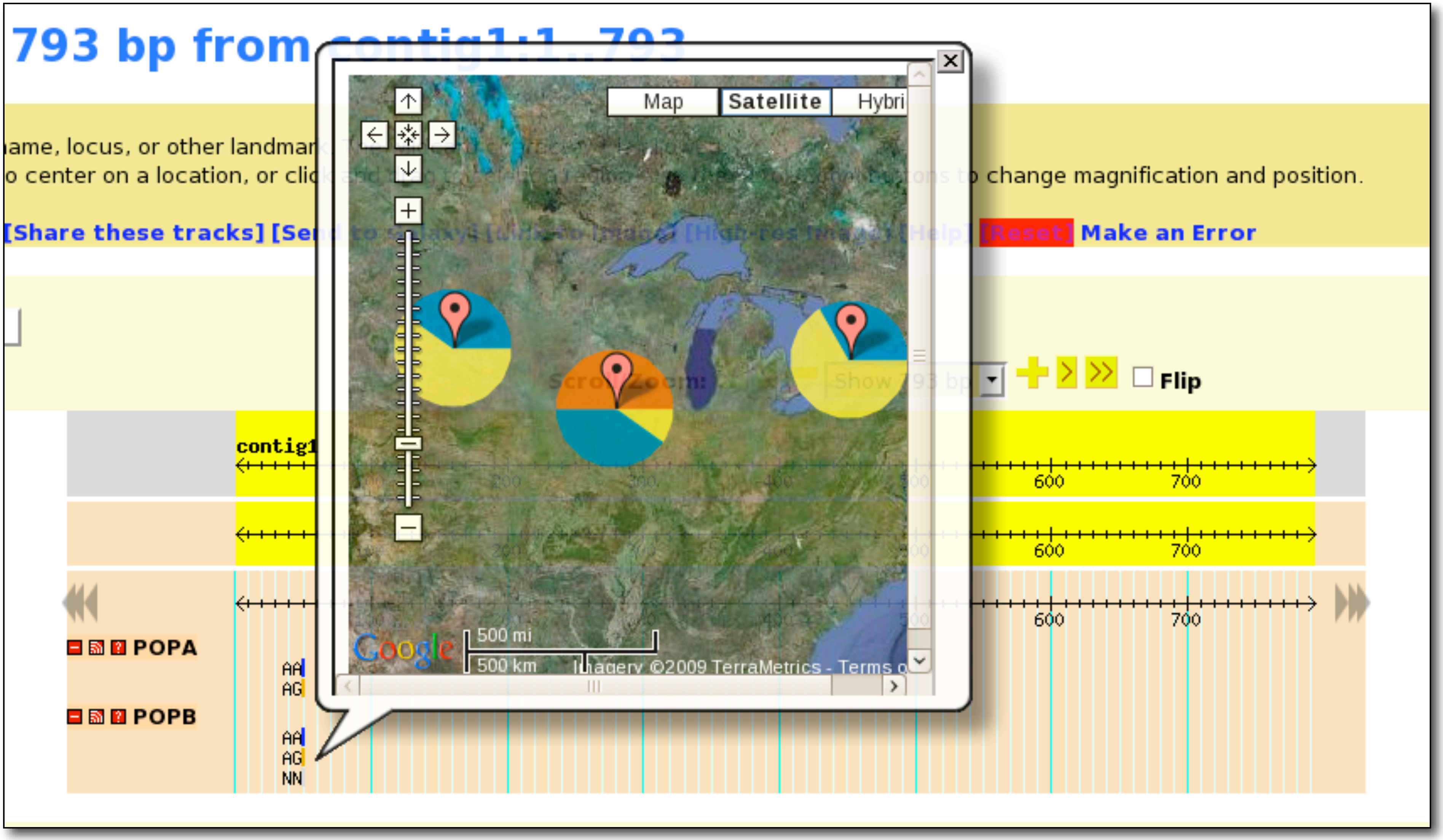
for performing in depth LD and
Haploview (ver4.0) is now avail

rs8022484: Allele Frequencies in HapMap Populations

Panel	Description	Frequency of A (ref)	Frequency of G
ASW(A)	African ancestry in Southwest USA	70%	30%
CEU(C)	Utah residents with Northern and Western European ancestry from the CEPH collection	84%	16%
CHB(H)	Han Chinese in Beijing, China	71%	29%
CHD(D)	Chinese in Metropolitan Denver, Colorado	73%	27%
GIH(G)	Gujarati Indians in Houston, Texas	87%	13%
JPT(J)	Japanese in Tokyo, Japan	66%	34%
LWK(L)	Luhya in Webuye, Kenya	61%	39%
MEX(M)	Mexican ancestry in Los Angeles, California	94%	6%
MKK(K)	Maasai in Kinyawa, Kenya	69%	31%
TSI(T)	Toscans in Italy	84%	16%
YRI(Y)	Yoruba in Ibadan, Nigeria	56%	44%

http://hapmap.org

Geolocation data



793 bp from contig1:1..793

ame, locus, or other landmar

o center on a location, or clic ... change magnification and position.

[Share these tracks] [Se ... [Reset] Make an Error

Yi-Hsin Erica Tsai Ben Faga