

Managing Genomic Data

Scott Cain
GMOD Project Coordinator
Senior WormBase Developer
Alliance of Genome Resources Developer
Ontario Institute for Cancer Research (OICR)
scott@scottcain.net

Programming for Biology
CSHL
October 2017

One problem people in bioinformatics may face:

Some nice people (possibly you) sequenced one or more organisms or strains, and then built them into assemblies (or if you're really lucky, full chromosomes). What now?

In addition, again if you're really lucky, you may have all sorts of data associated with these sequences, like expression, strains, genotypes, phenotypes, breeding info, physical maps. Seriously, now what?



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?



GMOD Project

- Open Source
- Used to have two full time project staff:
 - Project Coordinator: Scott Cain
 - Help Desk: Amelia Ireland
- 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 least 2 years of support
 - ...



GMOD is Software

Algorithms?

```
INEXACTSEARCH(W,z)
  CALCULATED(W)
  return INEXRECUR(W,[W]-1,z,1,[X]-1)
```

```
CALCULATED(W)
  k ← 1
  I ← [X] - 1
  z ← 0
  for i=0 to |W|-1 do
    k ← C(W[i]) + O'(W[i],k-1) + 1
    I ← C(W[i]) + O'(W[i],k)
    if I > I then
      I ← I
      k ← k + 1
      z ← z + 1
  D(I) ← z
```

Not really.

```
INEXRECUR(W,i,z,k,l)
  if z < D(I) then
    return 0
  if i < 0 then
    return ([k,l])
  I ← 0
  I ← I ∪ INEXRECUR(W,i-1,z-1,k,l)
  for each b ∈ {A,C,G,T} do
    k ← C(b) + O(b,k-1) + 1
    I ← C(b) + O(b,l)
    if k < l then
      I ← I ∪ INEXRECUR(W,i,z-1,k,l)
    if b = W[i] then
      I ← I ∪ INEXRECUR(W,i-1,z,k,l)
    else
      I ← I ∪ INEXRECUR(W,i-1,z-1,k,l)
  return I
```

Plumbing!



GMOD Software

- Configurability and extensibility are central goals of GMOD.
 - GMOD tools are built to be reused
- Emphasize local installs
- Not a hosted solution (with a few exceptions)
- Not monolithic.
 - Most components can stand alone
 - Allows organizations to start slowly



GMOD Software

- Interoperability and data integration are also central goals of GMOD.
- You'll see several mechanisms in this talk
 - GFF3
 - Chado
 - Ontologies



GMOD components can be categorized as

- V** Visualization
- D** Data Management
- A** Annotation and Analysis



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 Galaxy
A Ergatis
A ISGA
A SOBA
A Textpresso
A Apollo
V**A** Table Edit

V GBrowse
V WebGBrowse
V JBrowse
V CMap
V GBrowse_syn
V Sybil
V SynView

D Chado
D Bio::Chado::Schema
D ModWare
A**V** Tripal

D BioMart
D InterMine



A Annotation & Analysis

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.



MAKER

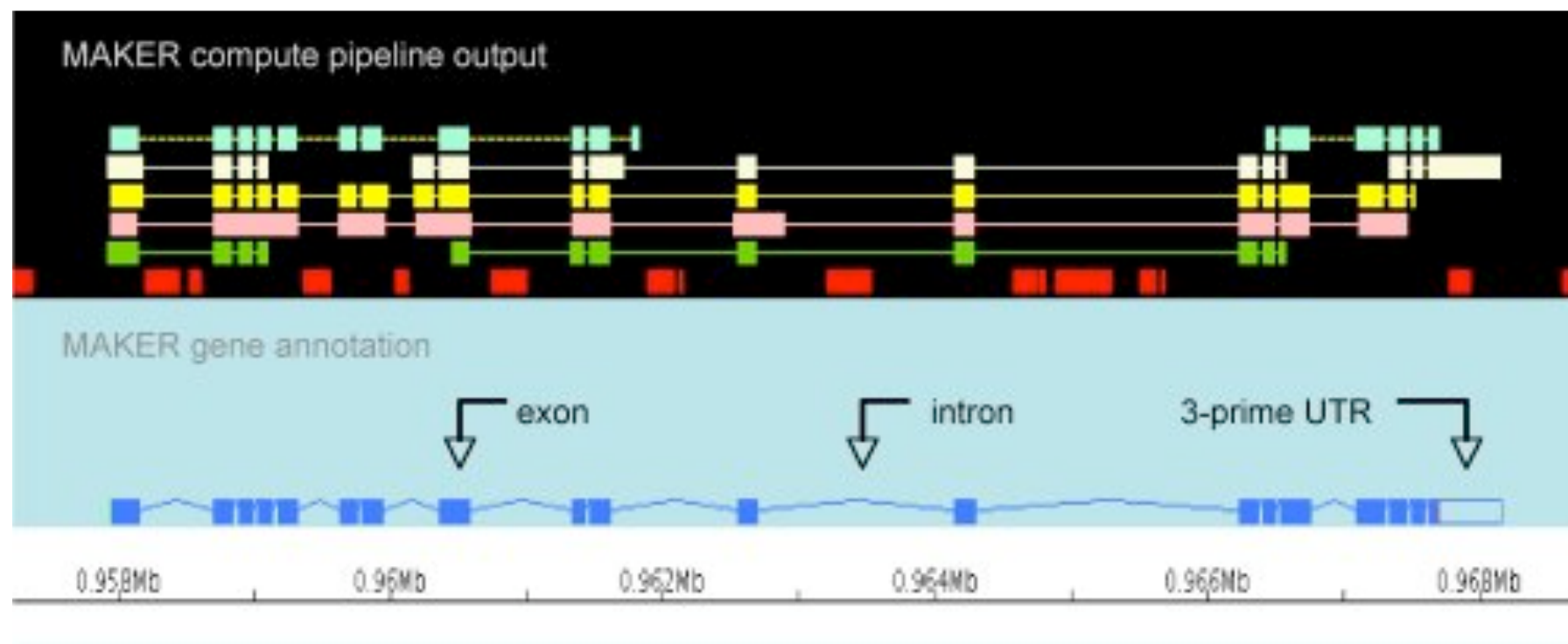


- Genome annotation pipeline for creating gene predictions
- Incorporates
 - SNAP, RepeatMasker, exonerate, BLAST
 - Augustus, FGENESH, GeneMark, MPI
- Other capabilities
 - Map existing annotation onto new assemblies
 - Merge multiple legacy annotation sets into a consensus set
 - Update existing annotations with new evidence
 - Integrate raw InterProScan results
- Maker Web Annotation Service



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

Minimizing “Edit distance”



SNAP *ab-initio* Gene Prediction
EST Alignment - EXONERATE
Protein Alignment - EXONERATE
Protein Alignment - BLASTX

EST Alignment - BLASTN
Repeats
MAKER gene annotation

MAKER Development

■ MAKER

2008. Based on early annotation pipelines developed by Mark Yandell at Celera.

■ MAKER 2

2011. Introduction of MPI parallelization, support for multiple gene predictors, GFF3 pass-through, and quality metrics like AED (Annotation Edit Distance) from the Sequence Ontology consortium.

■ MAKER-P

2015. Support for tRNA and snoRNA annotation. Improved parallelization on large plant genomes.

■ MAKER 3

2016. EVM (Evidence Modeler) support for improved annotation and user defined evidence probability weighting.



MAKER Resources

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



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



+



=

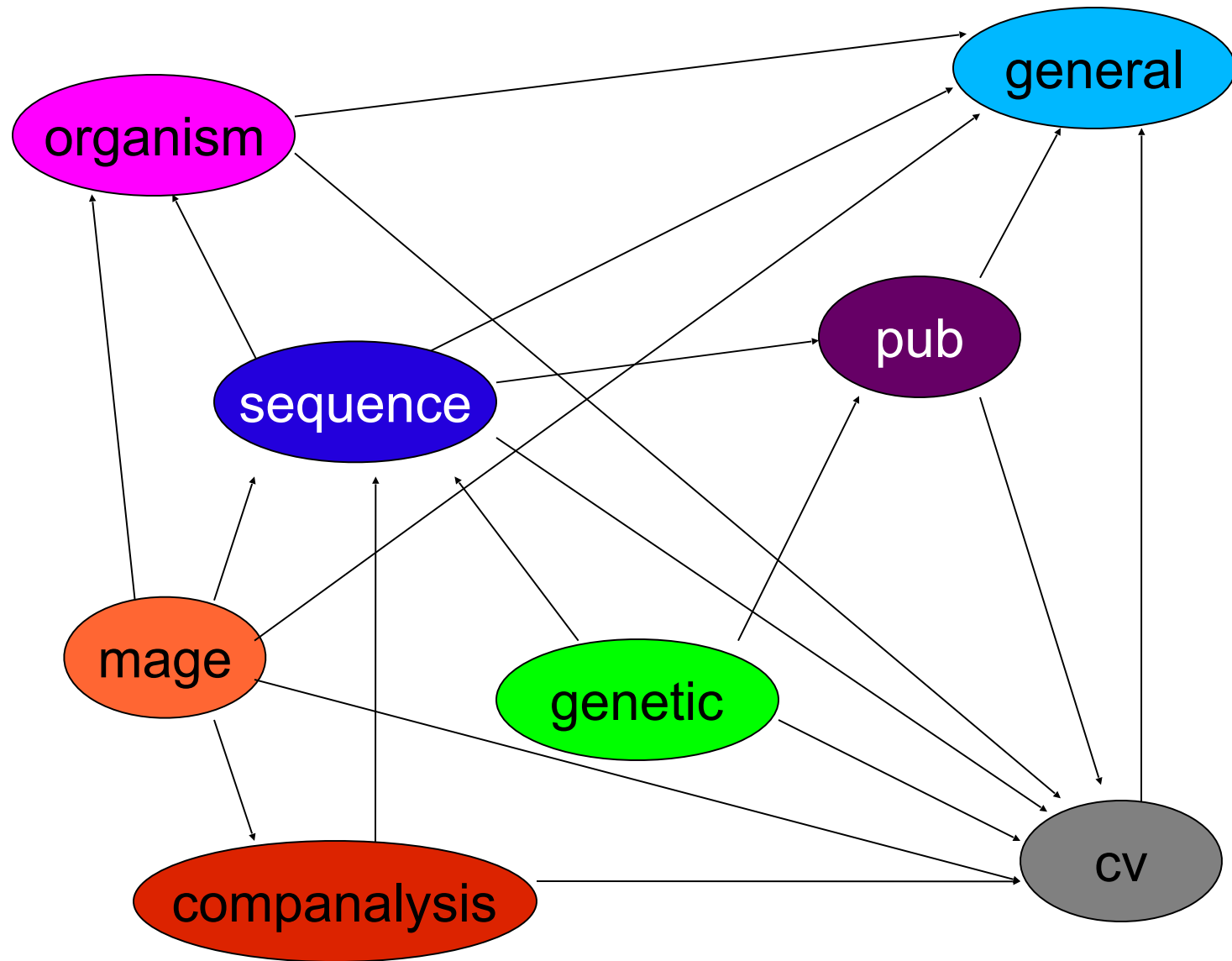


Why use Chado?

- 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
- Normalized (boring but important for data integrity)



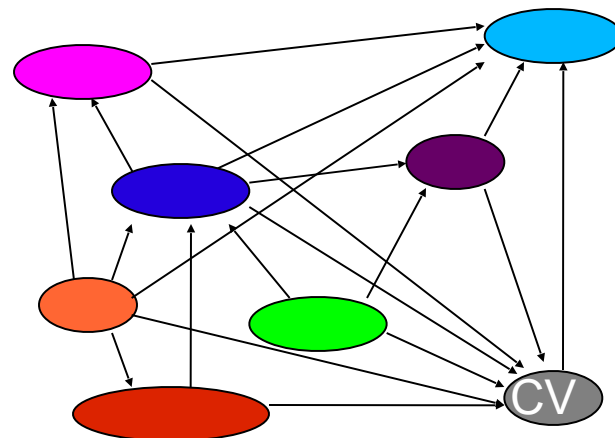
Chado Modules



- muscle 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 junctions

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 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
 - Tripal



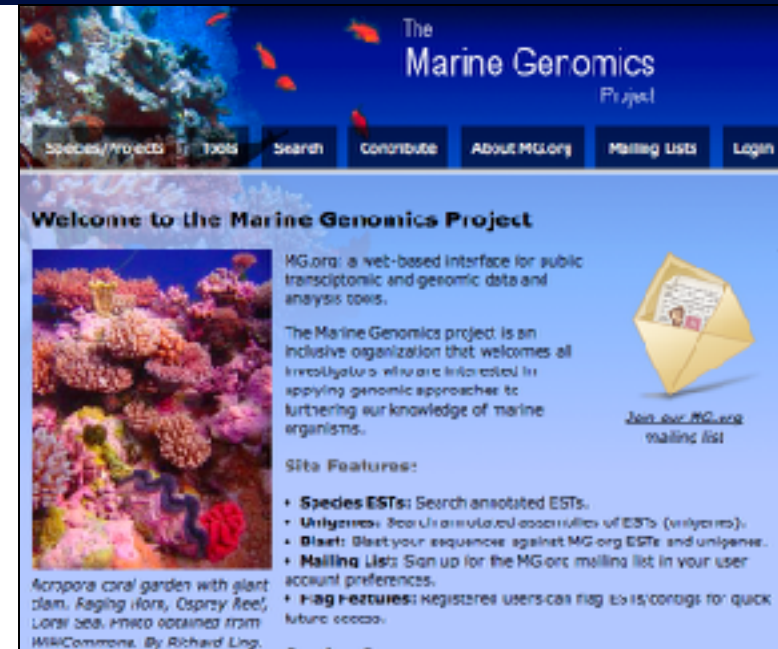
Do it yourself

- GMOD provides some support in form of libraries
- Perl
 - Chado::AutoDBI (phasing out)
 - Modware → Bio::Chado::Schema
- Java
 - At least two projects under development
 - GOBL (the Berkeley/WebApollo people)
 - INRA (France) Hibernate-based
- Drupal / PHP
 - Three projects underway



Tripal

- Set of Drupal modules
 - Feature, Organism, Library, Analysis
 - Modules roughly correspond to Chado modules
 - Easy to create new modules
- Includes user authentication, job management, and data entry support



MarineGenomics.org



Stephen Ficklin, Meg Staton, Chun-Huai Cheng, ...
Washington State University, Clemson University Genomics Institute

Tripal Resources

Home Page	http://gmod.org/wiki/Tripal
Tutorial	http://gmod.org/wiki/Tripal_Tutorial
User Guide	http://tripal.info/tutorials/v3.x
Example	https://www.rosaceae.org
Mailing List	https://lists.sourceforge.net/lists/listinfo/gmod-tripal



Chado Web: DIY or Tripal?

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.

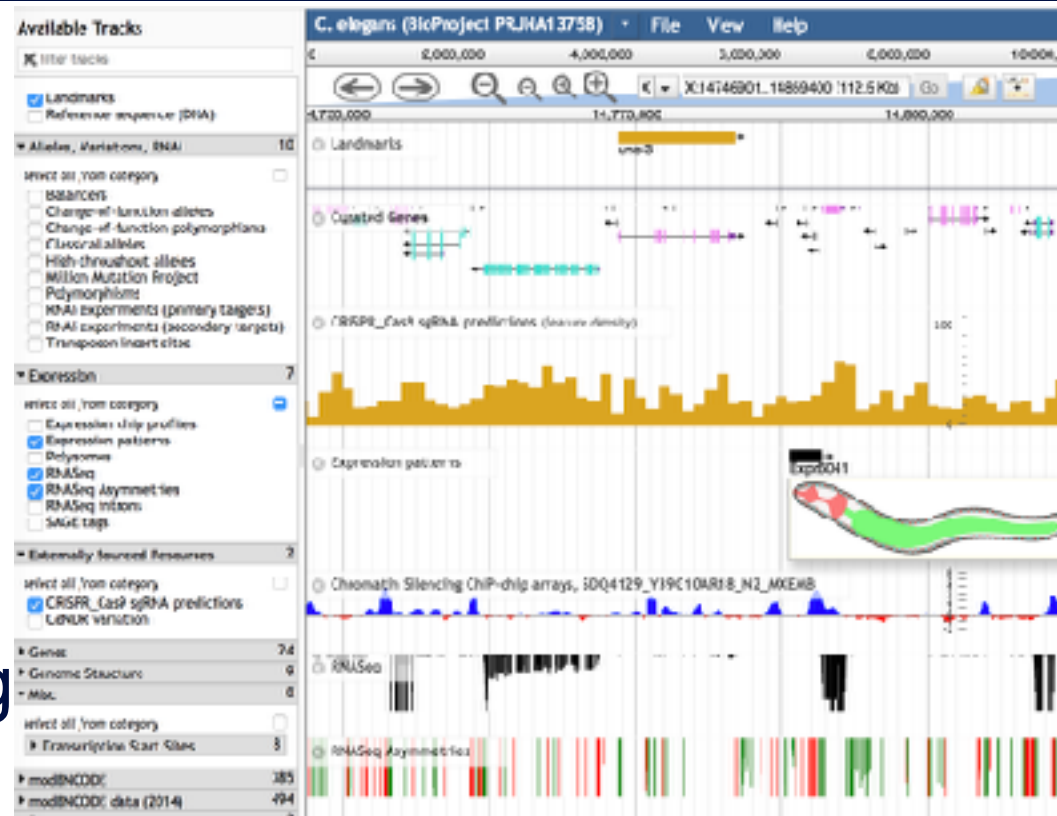
Stephen Ficklin, Lead Tripal Developer



This is not necessarily an either/or choice

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 Future Plans

- 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
- Server side tools for integrated analysis



JBrowse Resources

Home Page	http://jbrowse.org
Admin Tutorial	http://gmod.org/wiki/JBrowse_Tutorial_PAG_2017
Configuration	http://jbrowse.org/code/jbrowse-master/docs/config.html
Example site	http://staging.wormbase.org/tools/genome/jbrowse/
Mailing List	https://lists.sourceforge.net/lists/listinfo/gmod-ajax



GBrowse or JBrowse

GBrowse

Robust ecosystem

Feature rich

Large user base

Track sharing

JBrowse

Very fast

Rapidly growing user base

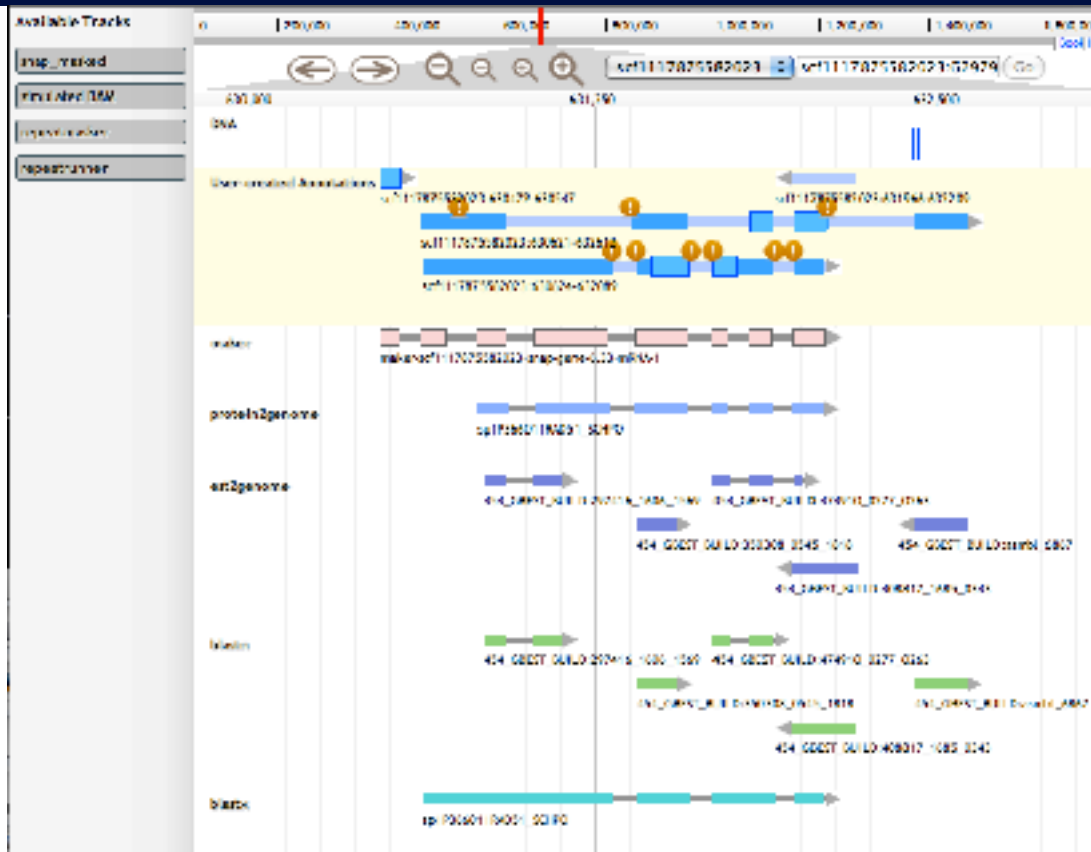
Lots of future development

Easy to configure

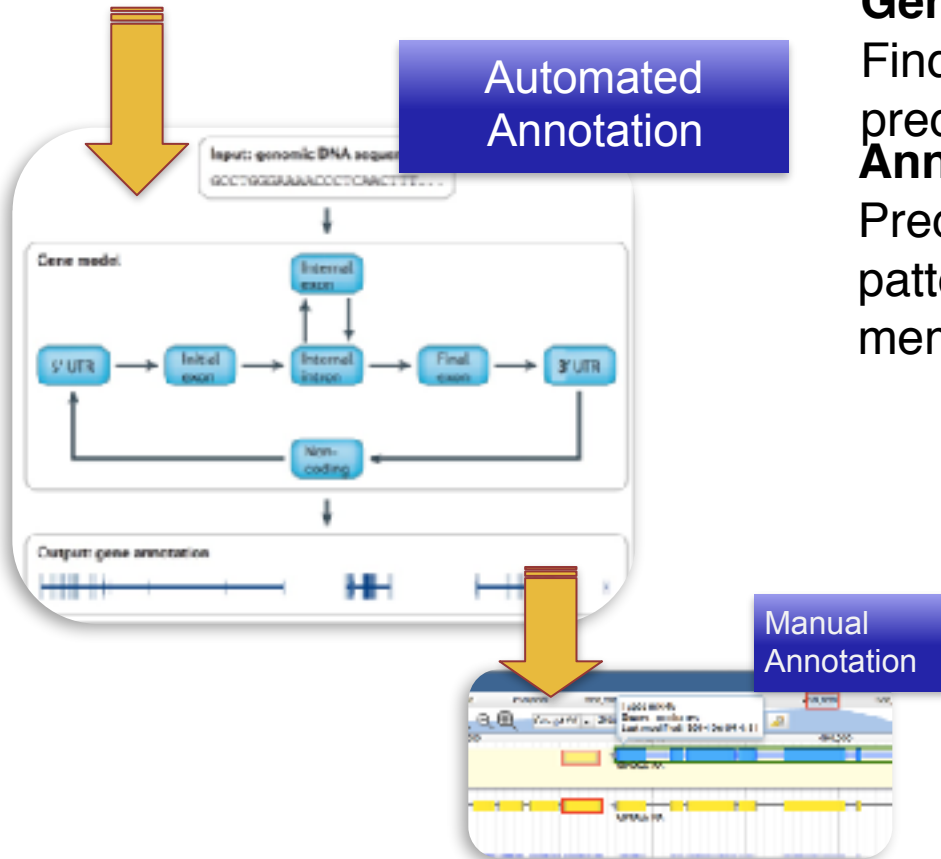


Apollo (used to be called WebApollo)

- GMOD's genome annotation editor
- Add and refine annotations.
- Based on JBrowse
- Multiple simultaneous users
- Keep track of evidence, curator
- Used in several community annotation efforts



Automated Identification is not Perfect



Generation of Gene Models

Find ORFs, multiple rounds of gene prediction

Annotation of Gene Models

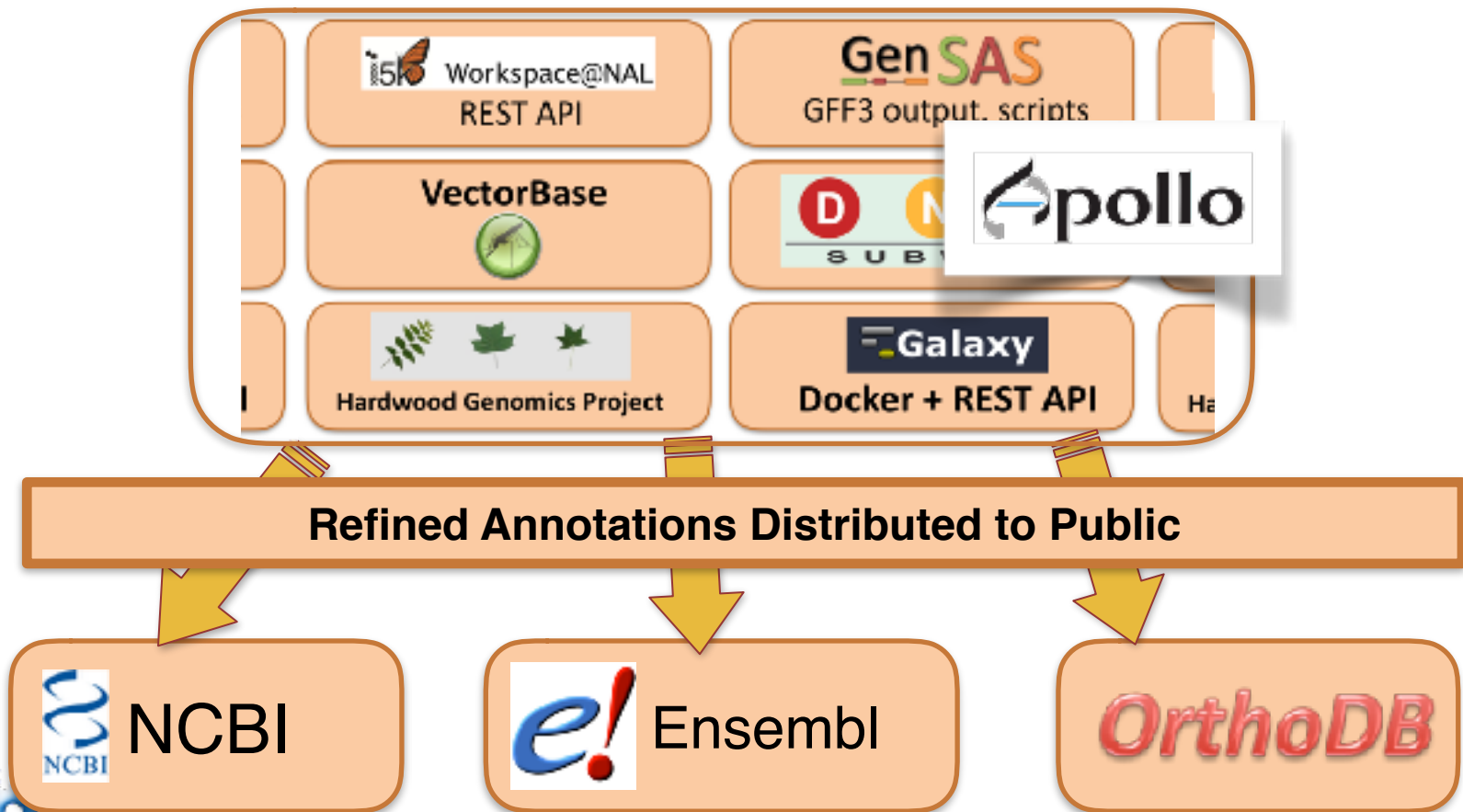
Predicting function, expression patterns, metabolic network memberships



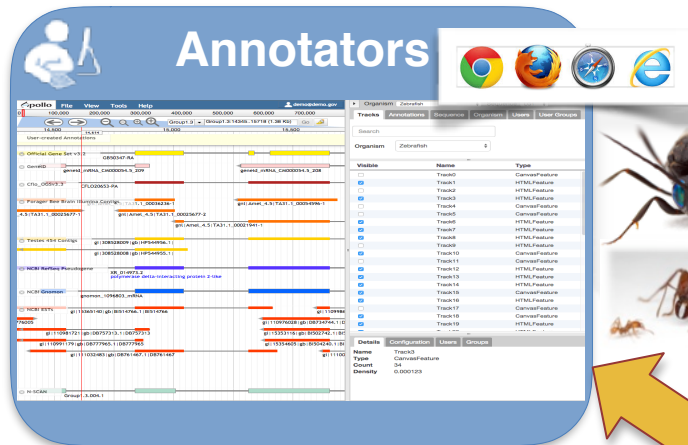
- Assembly errors can cause fragmented annotations
- Limited coverage makes precise identification difficult


Apollo: Used to Produce High Quality Annotations

- Over 100 organizations use Apollo
- Multiple genomes and labs per server



Apollo is a Tool for Collaborative Annotation



- Web-based Editor
- Visual feedback
- Real-time collaborative
-  genomic browser

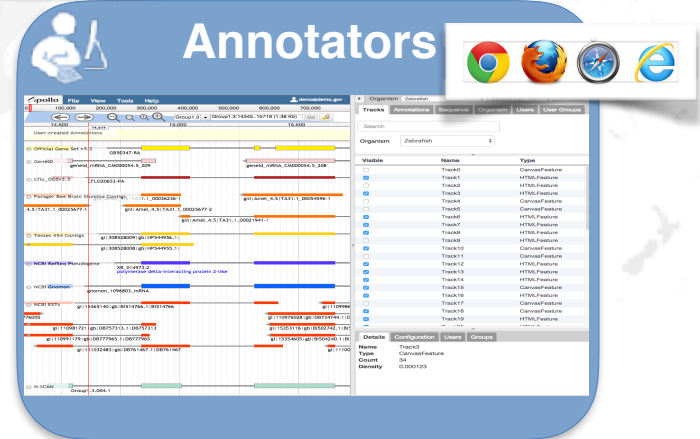
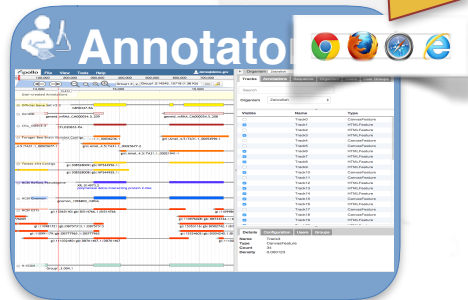


Photo Credits: i5K; Alex Wild at <http://www.alexanderwild.com/>: leaf cutter ant, ensign wasp; Leo Bukeboom: *Nasonia vitripennis* jewel wasp; Wikimedia Commons: *Apis mellifera* honey bee; Mike MacNeil USDA/ARS Fort Keogh LARRL: *Bos taurus* cow.

Apollo Resources

Home Page	http://apollo.berkeleybop.org/
Documentation	http://genomearchitect.github.io/documentation/
Demo	http://genomearchitect.github.io/demo/
Mailing List	apollo@lists.lbl.gov



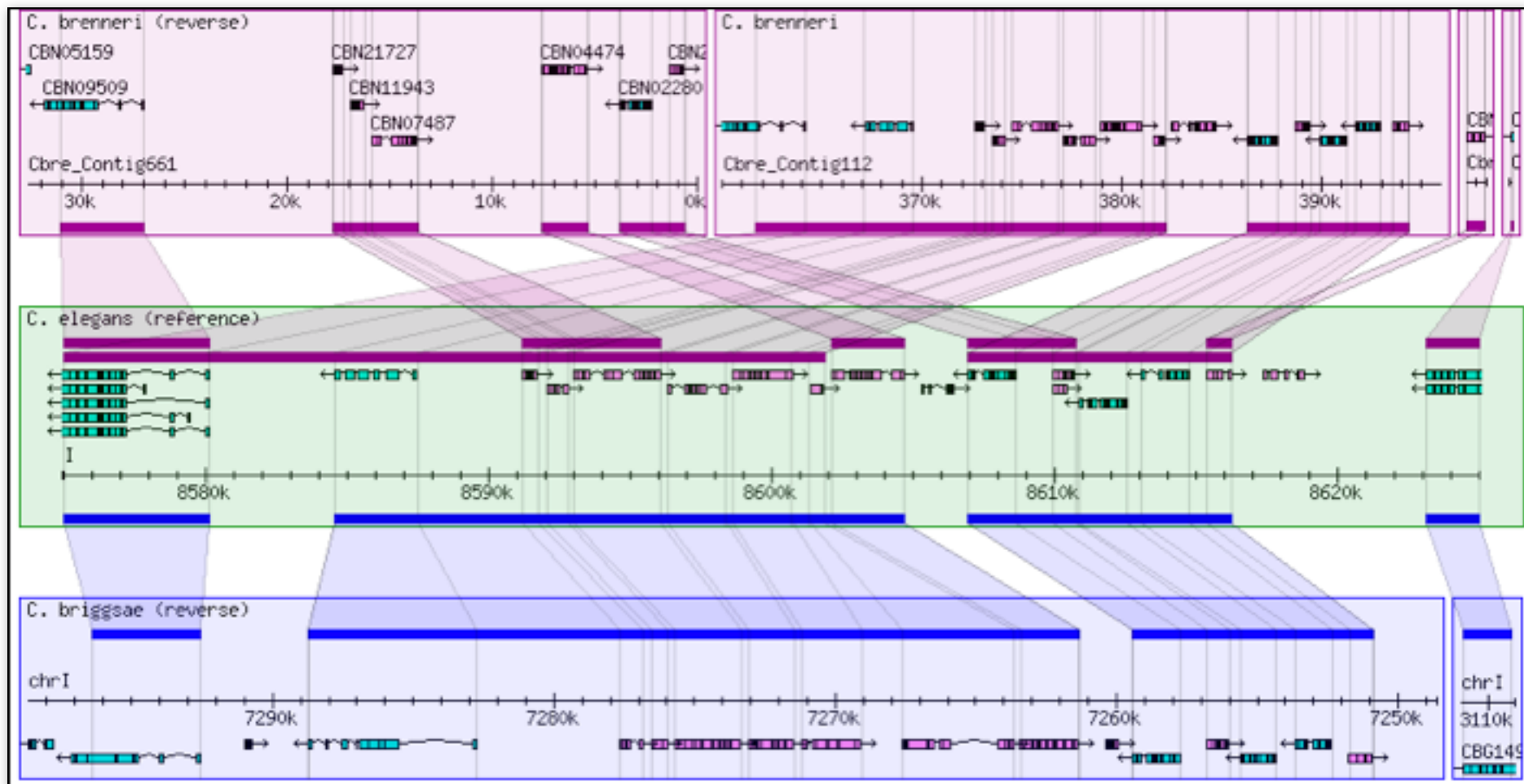
GBrowse_syn

- 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

GBrowse_syn



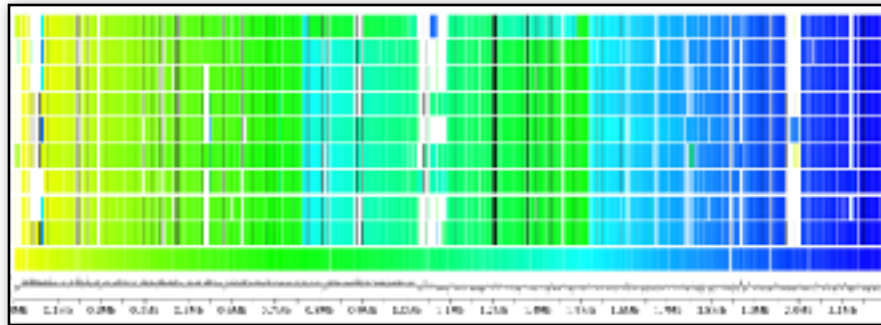
Syntenic blocks do not have to be colinear
Can also show duplications

GBrowse_syn Resources

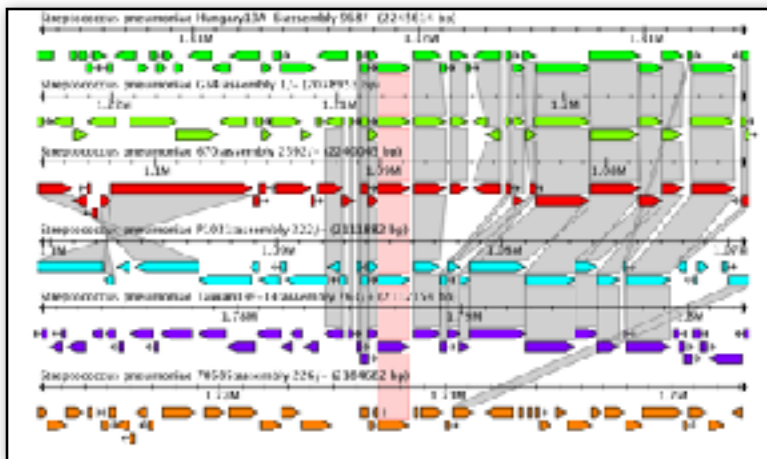
Home Page	http://gmod.org/wiki/GBrowse_syn
Tutorial	http://gmod.org/wiki/GBrowse_syn_Tutorial
User Help	http://gmod.org/wiki/GBrowse_syn_Help
Configuration	http://gmod.org/wiki/GBrowse_syn_Configuration
Example	http://www.wormbase.org/cgi-bin/gbrowse_syn/
Mailing List	https://lists.sourceforge.net/lists/listinfo/gmod-gbrowse



Sybil



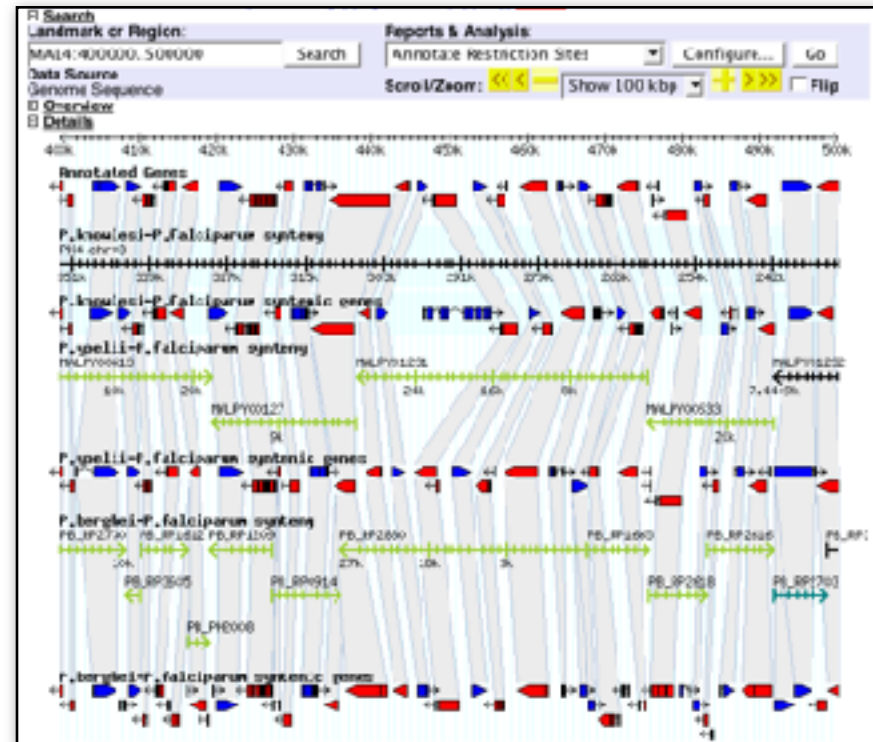
Whole Genome Gradient Display



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



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

Scalable (sort of)
Familiar interface
Extensive documentation
Growing user community

SynView

Scalable
Runs inside GBrowse 1

Sybil

Scalable
Whole genome and
other unique visualizations
Built on Chado



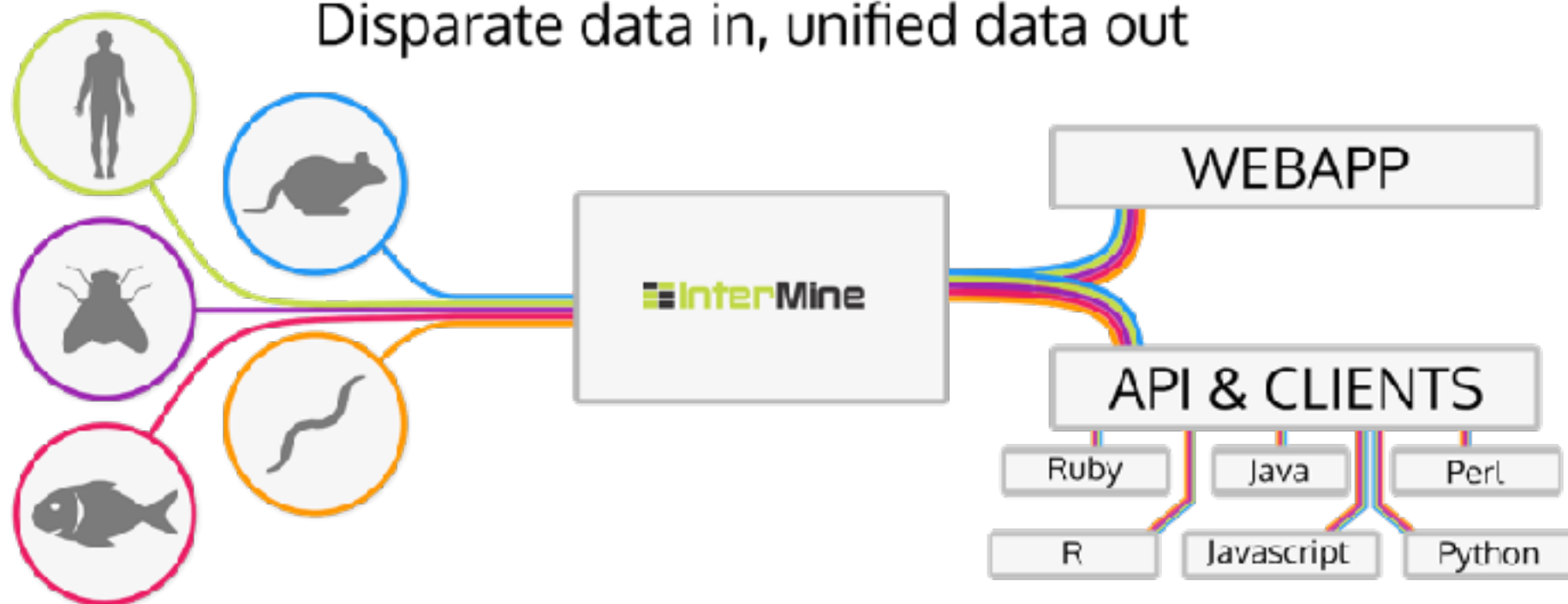
BioMart and InterMine

- 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



InterMine: What is it?

Disparate data in, unified data out



Model organism images Designed by freepik and distributed by Flaticon

Who uses InterMine

Over 30 installations, mostly at MODs, like FlyMine, YeastMine, WormMine, Wheat3BMine.

Others as well though, like TargetMine (drug discovery) and Shaare (gene candidate prioritization).

What do they use it for:

1. Genomic
2. Pathways
3. Interactions (complex and binary)
4. Protein structures
5. GO and other ontologies (Mammalian phenotype, etc)
6. Protein domains
7. Variation
8. Expression and regulation
9. Lots more ...



What can you do with InterMine

1. Keyword search
2. Query results
 1. Column summaries
3. List Analysis
 1. Visualizations - heat maps, expression.
 2. Enrichment
4. Template queries - quick forms to make for searches that your users do often
5. My account
6. All data is integrated, fast performance



GFF3

- The common file format of GMOD for genomic annotation
- Tab delimited, 9 column format
- Supported by Chado, GBrowse, JBrowse, CMap, Apollo, InterMine, BioMart, Galaxy,



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
- Talks/Posters
- Mailing Lists
- ...



[page](#) [discussion](#) [view source](#) [history](#)

The March 2011 GMOD Meeting starts this weekend. Register now.

Welcome to GMOD

GMOD is the Generic Model Organism Database 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 (including [GMOD Schools](#)), 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. The next meeting will be held March 5-6, 2011 at NESCent in Durham, North Carolina, as a part of [GMOD Americas 2011](#).

Contributing Organizations



Start this weekend!
Register Now
(Satellites are free)



Abstracts due Feb 28

[GMOD news](#) [RSS](#)

Planned downtime for gmod.org
InterMine 0.96 Release
GMOD Helpdesk Position Open
Galaxy Conf. Abstracts Due Feb 28
GMOD Wiki Migration Complete
March 2011 GMOD Meeting
Openings @ Xenbase
GMOD @ PAC 2011
GMOD Roadshow in San Diego
New GMOD.org Beta site

[New & Revised Pages](#) [RSS](#)

• March 2011 GMOD Meeting • GSoC •
• Browse 2.8 Prerequisites • GMOD
Schools • GMOD in the Sequencing
Center • News/Planned downtime for
gmod.org • News/InterMine 0.96 Release
• Browse sys • Galaxy • News/GMOD
Helpdesk Position Open

Popular GMOD Tools

Genome Browsing and Editing
[Gbrowse](#): Genome annotation viewer
[Azollo](#): Genome annotation editor

Comparative Genomics
[CMap](#): Comparative map viewer
[GBrowse_syn](#): Synteny viewer

Database Tools
[Chado](#): Biological database schema
[BioMart](#): Data mining system
[GMODTools](#): Chado to Fasta, GFF, ...
[InterMine](#): Data warehousing

Analysis and Annotation
[Galaxy](#): Data analysis & integration
[MAKER2](#): Genome annotation pipeline

Biological Pathways
[Pathway Tools](#): Metabolic, regulatory

Publication Curation
[Textpresso](#): text mining

Mailing Lists

- Several project and topic based lists
- Many component-specific lists
- Mailing lists are very active
- Nabble archive of all lists

Topic	URL Link	Accessed	Author(s)
Introduction to the	www.khanacademy.org	Viewed on 08/05/2015 at 10:00 AM	Dr. Michael J. Smith (2015)
Calculus I and II	www.khanacademy.org	Viewed on 08/05/2015 at 10:00 AM	Dr. Michael J. Smith (2015)

Component Lists

Learning Objectives: Identify the different types of business organizations.

[illegible]

Topic Headed Link:

I HAVE MANY OTHERS WHO COULD SIGN THIS OF THEIR OWN FREE WILL, BUT I AM NOT A PERSON WHO WOULD DO SO.

Name	Life Link	Notes	Availability
Co-Cu-Pb	great example	Cluster of <i>complanata</i> , <i>transfusa</i> , <i>perlegrus</i> , and related taxa.	Habitat: South Africa
Indus-Pan	great example	Cluster of <i>Stenophis</i> , <i>Nautil</i> , <i>Stenophis</i> , <i>Stenophis</i> , and related taxa.	Habitat: South Africa



<http://gmod.org/wiki/GMOD> Mailing Lists

Acknowledgements

Literally, too many to count (OK, that was figurative, but it's really a lot!)

Each of these projects has a group of people that contribute code or testing or feedback. Actually counting them up would be too hard.

Thanks, and now lets look at some of the websites I referred to earlier...

