GMOD

2013
GMOD
COMMUNITY MEETING
CAMBRIDGE, UK
5-6 APRIL 2013

# Other logistical details

Bathrooms

Twitter (#gmod2013 #isb2013)

Lightning talks tomorrow

Where for dinner?

# Introductions

Who you are

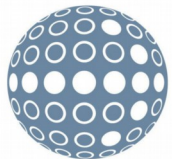Where you're from

What sorts of things you work on

Anything else you might want to add

# The State of GMOD

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

GMOD Meeting
April 5-6, 2013

- 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?



Plus hundreds of others

# Where did GMOD come from?

- Established in 2001 by NIH and USDA-ARS.

- Goal: cutting down on costs for new and existing MODs by making reusable software components.

- Initially there was just an FTP site where you could get software.

- Over the past 10 years, produced interoperable components the fill "all" of the needs of a new MOD.

GMOD components can be categorized as

V Visualization

D Data Management

A Annotation

## GBrowse

JBrowse

GBrowse_syn

CMap

Releases
  1.70
  2.54 (rapid development starting with 2.0 in Jan '10)

Features
  Rubberband region selection
  Drag and drop track ordering
  Collapsible tracks
  Popup balloons
  Allele/gentotype frequency, LD glyphs
  Geolocation popups
  Circular genome support (1.71)
  Asynchronous updates (2.0)
  User authentication
  Multiple server support (2.0)
  SQLite, SAMtools (NGS) adaptors

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

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

Demo: JBrowse Fly

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

# Data Management

## Chado

Tripal

TableEdit

BioMart

InterMine

Chado is the GMOD schema; it is modular and extensible, allowing the addition of new data types "easily." Covered data types in ontologies, organisms, sequence features, genotypes, phenotypes, libraries, stocks, microarrays, with natural diversity recently being rolled into the schema.

1.0 Release solidified the Chado that most people were already using from source.

1.1 Introduced support for GBrowse to use full text searching and "summary statistics" (ie, feature density plots). Version 0.30 of Bio::DB::Das::Chado is needed for these functions.

1.2 Natural diversity and more.

# Data Management

Chado

**Tripal**

TableEdit

BioMart

InterMine

New (2009) 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

Version 1.0 just released:
  – Improved loaders
  – New bulk loader
  – Complete integration with Drupal Views
  – Other good stuff

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

# Annotation

**MAKER**

DIYA

Galaxy

Ergatis

WebApollo

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 Online

Cantarel B. L., et al. MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. Genome Res. 2008 Jan;18(1):188-96

# Annotation

MAKER

DIYA

**Galaxy**

Ergatis

WebApollo

Web interface for creating one off or reproducible analysis pipelines for your genomic data.

MANY applications are already wrapped in.

Clearly defined API for adding more.

Try it out: http://usegalaxy.org/

Install it for yourself: http://getgalaxy.org/

Goecks, J. et al. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. Genome Biol. 2010;11(8):R86. Epub 2010 Aug 25.

# Annotation

MAKER

DIYA

Galaxy

Ergatis

WebApollo

Annotation editor based on JBrowse
Uses servlet engine, berkeley db
Supports BAM/BigWig
Initial release was December 2012

# Google Summer of Code

"Genome Informatics" Project:
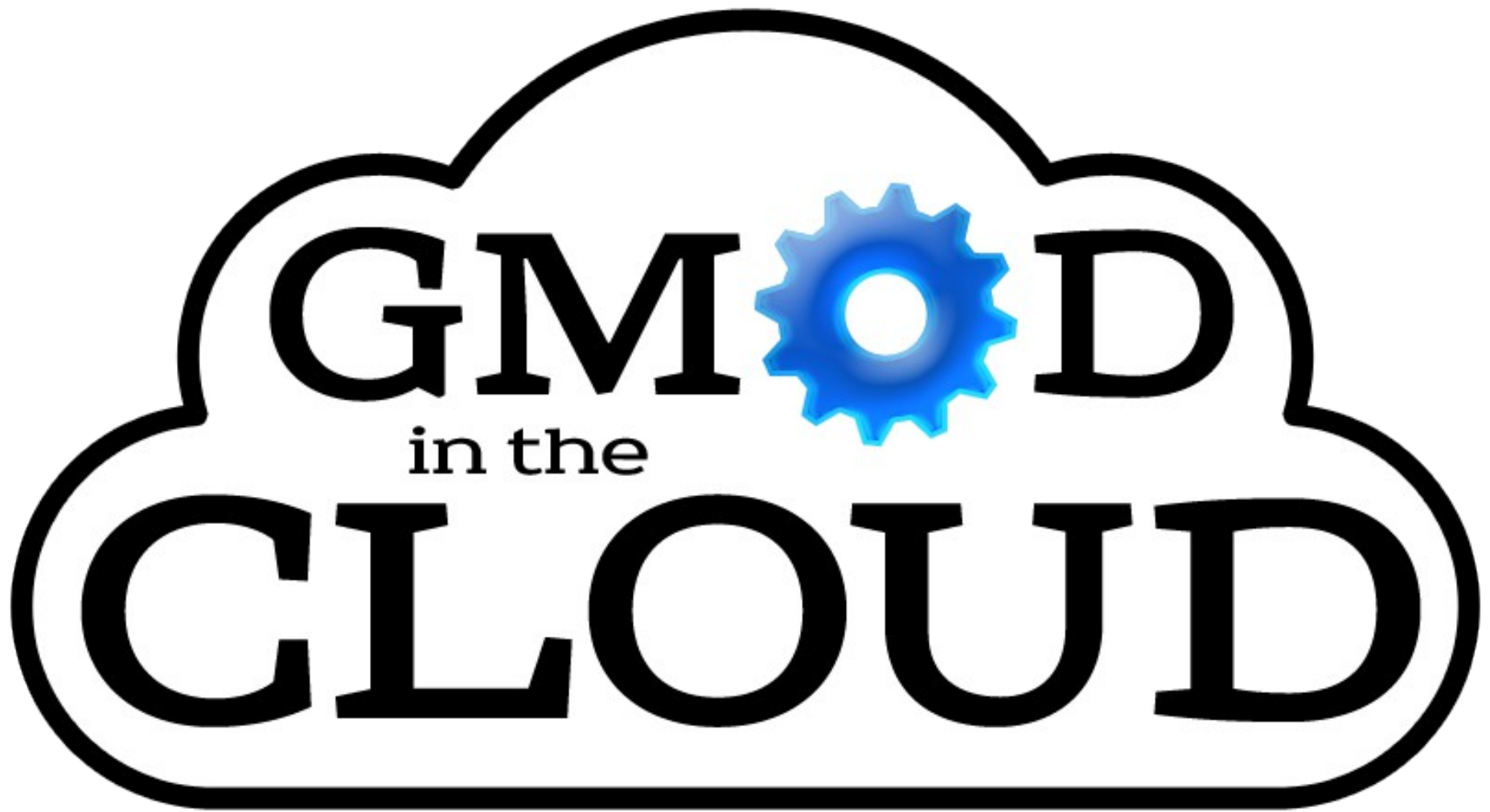    Reactome
    GMOD/JBrowse/GBrowse
    Galaxy
    WormBase

Initial proposals due SOON!
Decisions by mid April
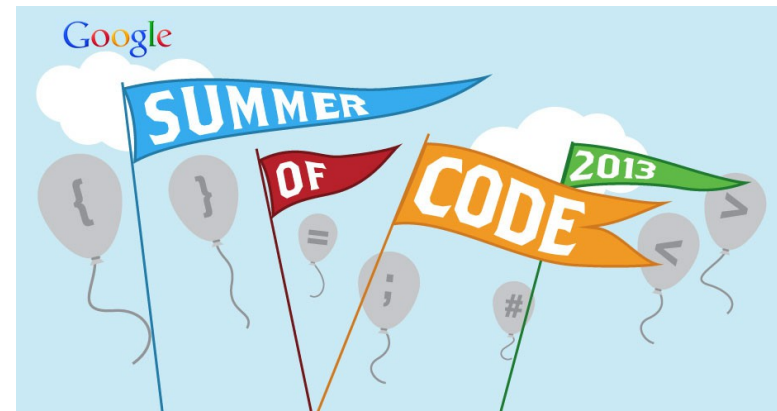
# GMOD in the Cloud

## Two cloud efforts:

### GBrowse2

- Goal to allow easy addition of data, rapid scaling of rendering servers
- Provide servers with "trickier" prereqs installed (eg, Kent's BigWig libraries)

### GMOD

- Goal to have a "reference implementation" for several GMOD tools that people can "try out" and continue to use if desired.
- Could use for an annotation jamboree.

Installed and configured:

- Chado 1.23

- GBrowse 2.54 (with SAMtools, BigWig)

- JBrowse 1.8.1

- Tripal 1.0

- WebApollo

- Sample data

  - S. cerevisiae

  - P. ultimum

# apt-get GMOD

Thanks to the efforts of debian-med (Olivier Sallou in particular), you can now do this:

```
sudo apt-get install gbrowse
sudo apt-get install libchado-perl
```

Gets all prereqs (libgd, BioPerl, etc).

Chado not perfect yet; still working with Olivier to get it right.

# NCBI GFF3

NCBI now provides GFF3 for some reference genomes (more on the way I think). From the NCBI Facebook page:

NCBI now offers Reference Sequence (RefSeq) genome annotation files in the latest Generic Feature Format (GFF3) specification (1.20). RefSeq genome data can be downloaded from the genomes area of the NCBI FTP site (http://bit.ly/gers8b). GFF3 files are in the GFF directory within each organism directory. Currently GFF3 files are available for the NCBI annotations of the latest assemblies for human (http://bit.ly/GEoisT), cow (http://bit.ly/GIs96D), dog (http://bit.ly/GE0yBu), chicken (http://bit.ly/GFd6WA), and many others.

Lots of bug fixes

Added gbrowse_import_ucsc_db.pl:
    create a GBrowse instance from UCSC
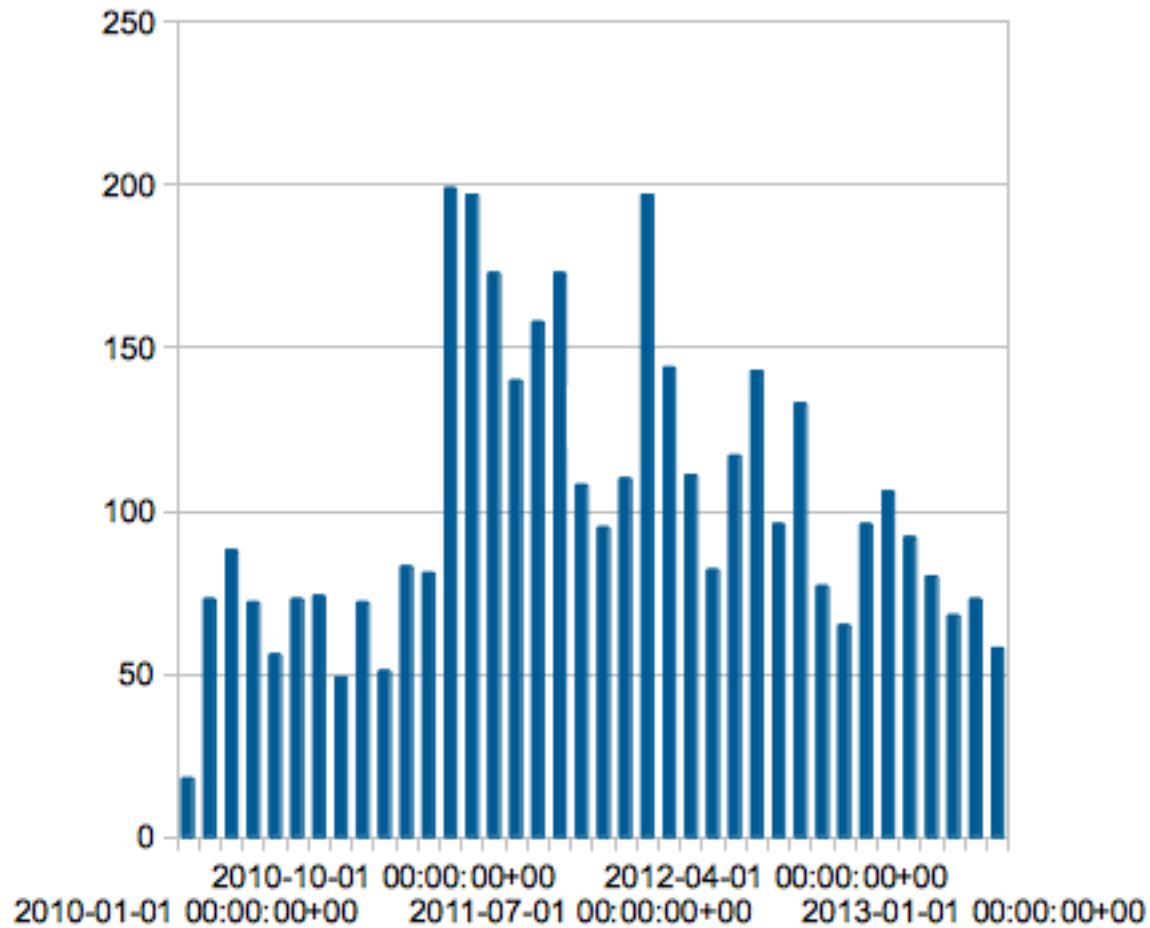
VirtualBox and EC2 support

TrueType font support (!)

## Some usage statistics

– When the user runs ./Build install, it prompts the user to send a registration email to me.

– Asks for email address, site, organism, organization, submits to a CGI that collects IP address too.

GM☼D

# GBrowse 2

Some usage statistics

- 8047 registrations

- 2868 distinct IP address

- 1233 distinct "organisms" (though this is free text, and has entries like "labuser", "Scientists" and "Not Specified"

- 1424 distinct organizations (though again free text—every continent represented (I think))

# GBrowse 2

Distinct IP addresses per month since Jan, 2010

# GBrowse

## Faceted track management (soon!)

# Chado 1.20 release

Fixes to several scripts:

> gmod_fasta2gff.pl

> gmod_bulk_load_gff3.pl

> load_ncbi_taxonomies.pl

Fixes so that GFF3 dumping from GBrowse works as expected

Added the ability for the database to "introspect" (chadoprop)

# Chado 1.21-1.23 releases

1.21 (October 2011)

- Added rudimentary support for non-public schemas (to better support Tripal)
- Bug fixes related to schema versioning

1.22 (November 2011)

- Fixes to make automated schema version migration easier

1.23 (December 2012)

- Made perl scripts more "modern" (circa 2004)
- Bug fix to GFF3 preprocessor (splitter/sorter)

# Chado versioning

Addition to the schema: chadoprop table

Additional cv: chado_properties

SQL diffs (ie, CREATE and ALTER statements)

Makefile target: "make update"

New scripts:

- gmod_chado_properties.pl to gather info about schema (specifically version)

- gmod_update_chado.pl to update from an older version of Chado (1.0+) to the current schema

# Chado 1.24 or 2.0 release

New interaction module (provided by FlyBase)

- In use at FlyBase now to model binary protein-protein interactions, but is general enough to model many to many or "self" interaction

Stored procedures to facilitate common tasks (eg, loading GFF)

- Requires coordination with Chado core, Tripal, WebApollo at least.

Funding for my position is drying up in 2014.

Need a five year plan to go grant hunting:

- Integration

- Expanding Cloud efforts (iPlant/iAnimal?)

- What else?

# Outreach

GMOD Help desk (Yay Amelia!)

Presence at many meetings/conferences

  Genome Informatics, ISMB, PAG, Biology of Genomes

Courses/Workshops

  GMOD summer school, Comparative Genomics Workshop, Progamming for Biologists, PAG

Wiki, mailing lists

# Acknowlegements

- You! (Remember that part about GMOD being a community?)
- Gos Micklem's group, notably Shelley Lawson
- There are literally too many people associated with GMOD to thank them all, but Lincoln Stein and Dave Clements are a big part of why GMOD is successful.  Amelia did a lot of work setting up this meeting.
- Eventbrite
- Funding agencies: NIH and USDA ARS