A Complete Community Genome Annotation System Using GMOD Tools

Scott Cain[1], Ed Lee[2], Carson Holt[3], Stephen Ficklin[4], Meg Staton[5], Lincoln Stein[1] and the GMOD Consortium

[1]Ontario Institute for Cancer Research, Toronto, Ontario, Canada, M5G 0A3
[2]Lawrence Berkeley National Lab, Berkeley, CA 94720
[3]University of Utah, Salt Lake City, UT 84112
[4]Washington State University, Pullman, WA, 99164
[5]Clemson University Genomics Institute, Clemson, SC 29634

The costs associated with DNA sequencing have decreased considerably and the number of organisms being sequenced has increased accordingly. Several organizations, including ParameciumDB, GnpAnnot, AphidBase, and BeeBase, have implemented methods for community annotation using tools from the Generic Model Organism Database (GMOD) project. Here we present an example system based entirely on GMOD tools for community annotation implemented in a VMware virtual machine that could, with little extra effort, be used directly for annotating a nascent genome. The tools include Chado[1] (a genomics database schema), GBrowse[2] (a web based genome browser), Apollo[3] (a genome annotation tool), MAKER[4] (a genome annotation pipeline), and Tripal[5] (a Drupal-based content management system).

Starting with raw sequence...

```
CGACGTGGTGAAATTGTTGCTGTTACTGGCGATGGTGTGAATGACTCTCC
GGCATTAAAAAGAGCCGATATTGGCGTTGCAATGGGTATTTCTGGATCTG
ACGTTTCTAAGCAGGCGGCAGATATGATTCTATTGGATGACAACTTTGCA
TCAATTGTTGTTGGTATTGAAGAGGGGCGGATTATTTTCGATAATCTTAA
AAAGTCCATCGCATATACCTTGACTTCAAATCTTCCTGAAATAGTGCCGT
TTTTATTTTTTGTGATATTTGATATACCTTTAGCTCTGGGCACTATAGCA
ATTCTATGCATCGATATCGGCACTGATATGCTTCCGGCAATATCGCTAGC
TTATGAAAAAGCTGAATCTGATATAATGGCGCGTATGCCAAGAGATCCGT
TTGAAGACCGTTTGGTAAATAAAAAGTTAATTCTAATGGCTTATTTGCAA
ATTGGAGTTATACAAACAGTAGCATGTTTTTTTACTTTTTTTGCTATAAT
GGCAGAACATGGATTCCCCCCCTCCAGACTTAAAGGAATTCGAGAAGACT
GGGACTCAAAAAATGTAGAAGACCTTGAAGATGGCTACGGACAAGAATGG
```

Existing assemblies,
Existing gene predictions,
Expressed sequence data,
NextGen sequencing data,
Protein sequences from related organisms

Raw Data

MAKER identif es repeats, aligns ESTs and proteins to a genome, produces *ab initio* gene predictions and automatically synthesizes these data into gene annotations having evidence-based quality values. MAKER produces as one output GFF3 data that can easily be put into Chado. In fact, MAKER comes with a tool to make loading its data into Chado even easier.

Data Analysis

MAKER

GFF3

Chado is a relational database schema that underlies many GMOD installations. It is capable of representing many of the general classes of data frequently encountered in modern biology such as sequence, sequence comparisons, phenotypes, genotypes, ontologies, publications, and phylogeny. Here is is used as a common data back end for several GMOD applications to allow them all the same views of the sequence

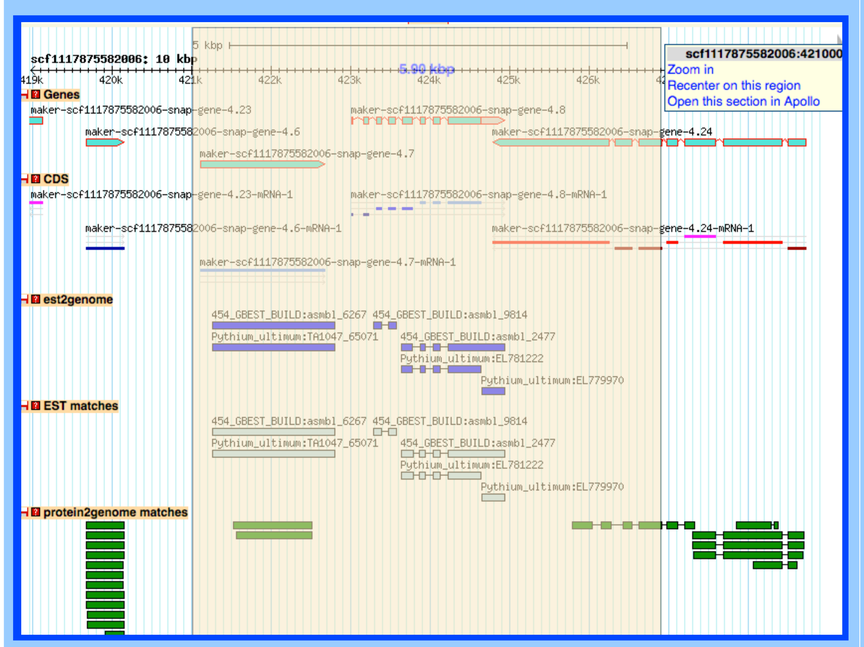
Data Warehouse

Chado

Direct Database Access

Genome Feature Browser

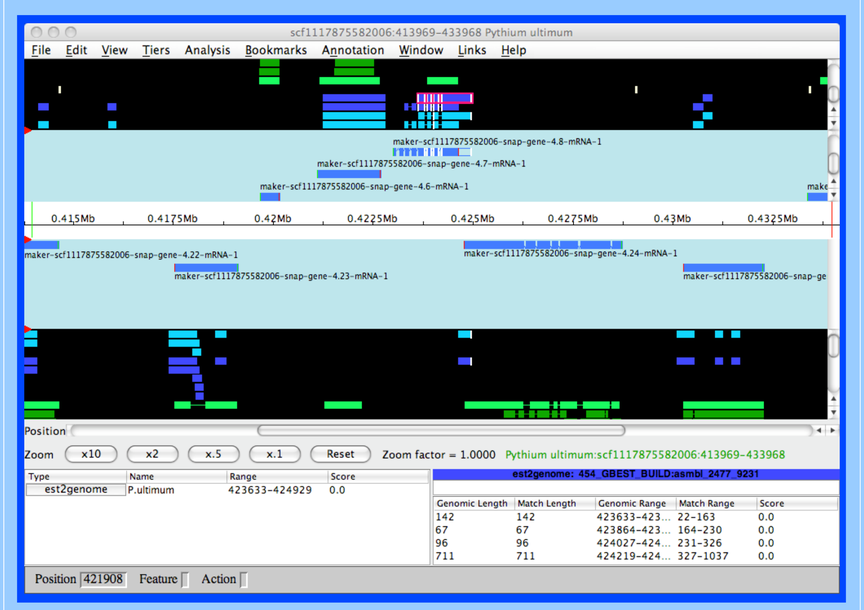
GBrowse



GBrowse2 is a interactive web application for manipulating and displaying annotations on genomes. It can be configured to open a region in Apollo for users to edit annotations.

Distributed Genome Annotation

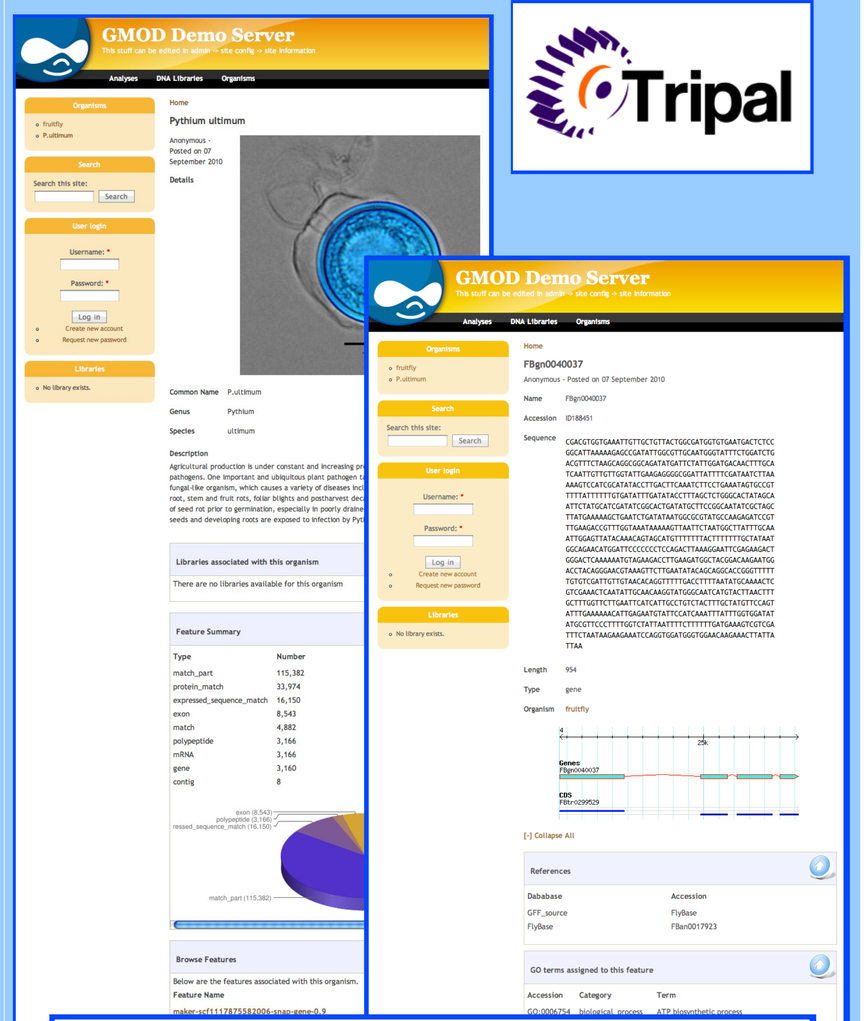
Apollo



The Apollo genome editor is a Java-based application for browsing and annotating genomic sequences. It can be activated (downloaded, installed and launched) from a URL, so a user looking at either a page in Tripal or GBrowse could decide to edit a gene model.

Online Data Visualization

Tripal



Tripal is a Chado database web front end based on a collection of Drupal modules. Drupal (http://drupal.org/) is a widely used, open source content management system. Putting Tripal on top of a Drupal installation takes only minutes, and provides pages for organisms, features (like gene pages), libraries and computational analyses. Pages can be conf gured to link directly to GBrowse and Apollo for browsing and editing.

Other GMOD Software
Other GMOD software could be installed in the VMware image, including JBrowse, Galaxy, BioMart, InterMine,.

Download Now!

The VMWare virtual machine is available for download now from gmod.org. Get the Ubuntu 9.10 image from ftp://ftp.gmod.org/pub/gmod/GMOD_Demo_Server_9_10.tar.gz (2.4 GB).

[1]Mungall, C. J. et al. A Chado case study: an ontology-based modular schema for representing genome-associated biological information. Bioinformatics. 2007 Jul 1;23(13)
[2]Stein, L. D. et al. The generic genome browser: a building block for a model organism system database. Genome Res 12: 1599-610. [PMID: 12368253]
[3]Lewis, S. E. et al. Apollo: a sequence annotation editor. Genome Biol 3(12):RESEARCH0082 [PMID: 12537571]
[4]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 [PMID: 18025269]
[5]http://www.genome.clemson.edu/software/tripal

GMOD is supported by a specific cooperative agreement from the USDA Agricultural Research Service, and by NIH grants co-funded from the National Human Genome Research Institute and the National Institute of General Medical Sciences. GBrowse development is also funded by the Ontario Ministry of Research and Innovation.