Exploit your Data

Galaxy

and the

GMOD Tool Suite

University of Tasmania
25 October 2011

Dave Clements, Emory University

http://galaxyproject.org

http://gmod.org

Agenda

Introduction

Galaxy

Worked example
Deployment Options
Microsatellites & SNP calling
Community

GMOD

Software
Community

Goal for this workshop

Give you some idea what these **open source** tools **can do**, and how you might **use them** in your research.

This workshop will not cover
details of how the tools are implemented or
new algorithm designs or
which assembler or mapper or ... is best for you.

The Motivation Slide

Next Generation Genomics: World Map of High-throughput Sequencers
Nick Loman, James Hadfield

http://pathogenomics.bham.ac.uk/hts/

What is GMOD?

- A set of interoperable open-source **software components** for **visualizing, annotating, integrating and querying, and analyzing** biological data.
- An active **community of developers and users** asking diverse questions and facing common challenges with their biological data.

http://gmod.org

Who uses GMOD?

Plus hundreds, if not thousands, of others

GMOD Server Requirements

- GMOD is not a hosted solution
  - *Usually*
- Server
  - Most use Linux or other Unix variant
- GMOD System Administrator
  - Understands Linux package management, scripting, command line interfaces, relational databases, ...
  - Grad/Undergrad, half time when starting up

http://gmod.org/wiki/Computing_Requirements

What is Galaxy?

- **A free (for everyone) web service** integrating a wealth of tools, compute resources, terabytes of reference data and permanent storage
- An **analysis and data integration** tool
- **Open source software** that makes integrating your own tools and data and customizing for your own site simple
- A part of **GMOD**

http://galaxyproject.org

Galaxy URLs to Remember

http://galaxyproject.org

http://usegalaxy.org

http://getgalaxy.org

Agenda

Introduction

Galaxy

Worked example

Deployment Options

Microsatellites & SNP calling

Community

GMOD

Software

Community

Galaxy aims to

Enable
accessible,
reproducible,
and
transparent
computational biomedical research.

Demo: Accessibility

On human chromosome 22,
which coding exons have the most
known SNPs?

http://usegalaxy.org

Galaxy: A Rough Plan

- Get some data
  - Coding exons on chromosome 22
  - SNPs on chromosome 22
- Mess with it
  - Identify which exons have SNPs
  - Count number of SNPs in each of those exons.
  - Identify exons with most SNPs

http://usegalaxy.org/galaxy101

Galaxy aims to

Enable
accessible,
reproducible,
and
transparent
computational biomedical research.

Demo: Reproducibility and Transparency

http://usegalaxy.org

Windshield splatter analysis with the Galaxy metagenomic pipeline

Sergei Kosakovsky Pond[1,2,6,9], Samir Wadhawan[3,6,7], Francesca Chiaromonte[4], Guruprasad Ananda[1,3], Wen-Yu Chung[1,3,8], James Taylor[1,5,9], Anton Nekrutenko[1,3,9] and The Galaxy Team[1]

[+] Author Affiliations

Abstract

How many species inhabit our immediate surroundings? A straightforward collection technique suitable for answering this question is known to anyone who has ever driven a car at highway speeds. The windshield of a moving vehicle is subjected to numerous insect strikes and can be used as a collection device for representative sampling. Unfortunately the analysis of biological material collected in that manner, as with most metagenomic studies, proves to be rather demanding due to the large number of required tools and considerable computational infrastructure. In this study, we use organic matter collected by a

Footnotes

[Supplemental material is available online at http://www.genome.org. All data and tools described in this manuscript can be downloaded or used directly at http://galaxyproject.org. Exact analyses and workflows used in this paper are available at http://usegalaxy.org/u/aun1/p/windshield-splatter.]

Article published online before print. Article and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.094508.109.

Agenda

Introduction

Galaxy

Worked example

Deployment Options

Microsatellites & SNP calling

Community

GMOD

Software

Community

Galaxy main site
http://usegalaxy.org

- Public web site, **anybody can use**
- Hundreds of tools
- **Persistent**
- ~500 new users per month, ~100 TB of user data, ~130,000 analysis jobs per month, every month is our busiest month ever...

But, it's a big world

Main has lots of tools, storage, processor, users, ...

- But **not all tools** - there are thousands and adding new tools is not taken lightly
- But **not infinite storage and processors** - main will continue to be maintained and enhanced, but with use limits and storage quotas

A centralized solution cannot scale to meet data analysis demands of the whole world

Scaling Galaxy

So much data:

- Encourage local Galaxy instances and Galaxy on the cloud. Support increasingly decentralized model and *improve access to existing resources*

So many tools and workflows:

- Focus on building infrastructure to allow community to integrate and share tools, workflows, and best practices

Local Galaxy Instances
http://getgalaxy.org

Galaxy is designed for local installation and customization

- Easily integrate new tools
- Easy to deploy and manage on nearly any (Unix) system
- ***Just download and run, completely self-contained!*** *

* Some assembly required. †
† But not much. ‡
‡ And help is on the way.

Public Galaxy Servers

http://galaxyproject.org/PublicGalaxyServers

Interested in:

ChIP-chip and ChIP-seq?

- ✓ Cistrome

Statistical Analysis?

- ✓ Genomic Hyperbrowser

Sequence and tiling arrays?

- ✓ Oqtans

Text Mining?

- ✓ DBCLS Galaxy

Reasoning with ontologies?

- ✓ GO Galaxy

Internally symmetric protein structures?

- ✓ SymD

Got your own cluster?

- Move tool execution to other systems
- Galaxy works with any DRMAA compliant cluster job scheduler (which is most of them).
- Galaxy is just another client to your scheduler.

Galaxy Cloud

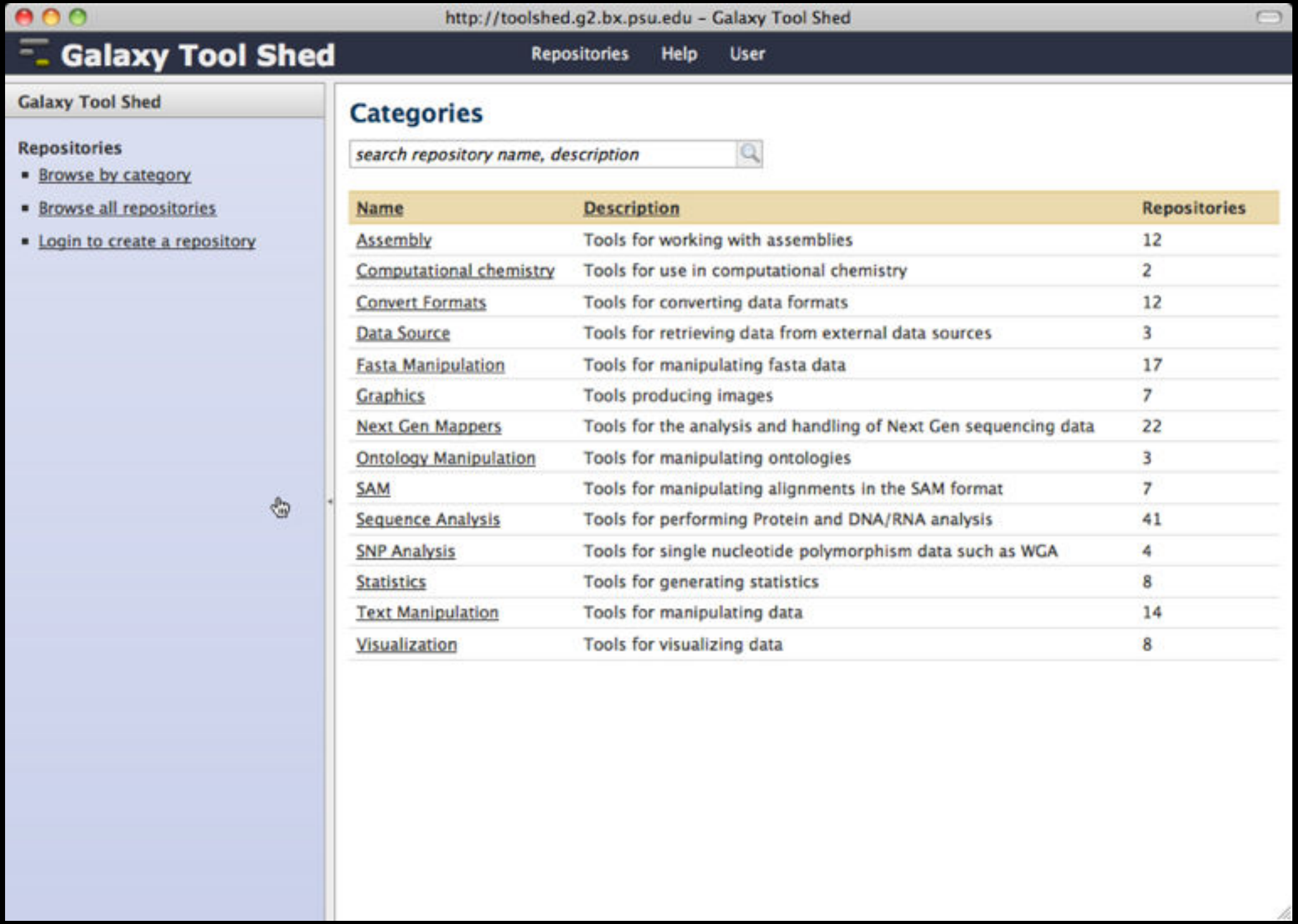
http://usegalaxy.org/cloud

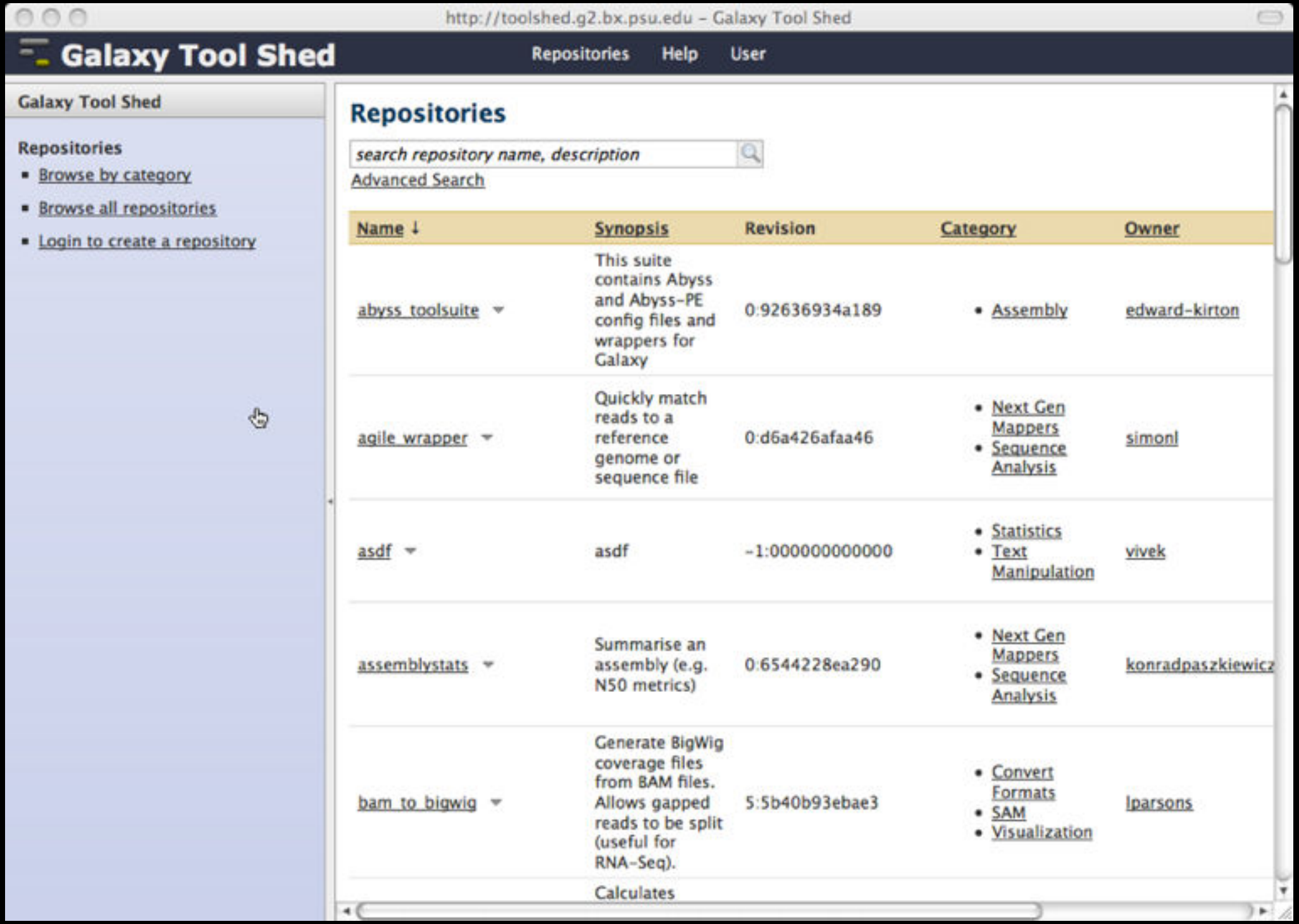
- Start with a fully configured and populated (tools and data) Galaxy instance.
- Allows you to scale up and down your compute assets as needed.
- Someone else manages the data center.

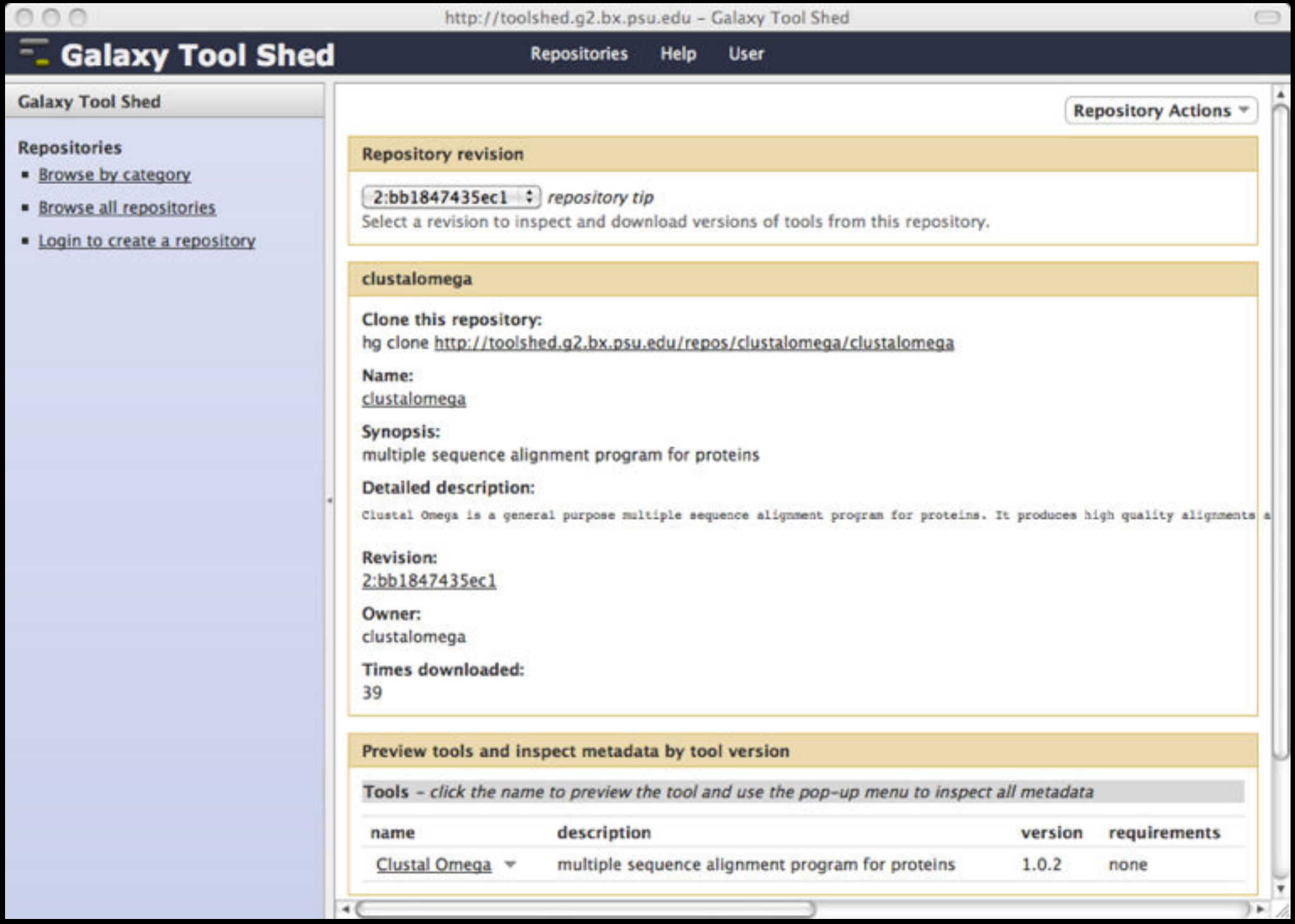
Galaxy Tool Shed

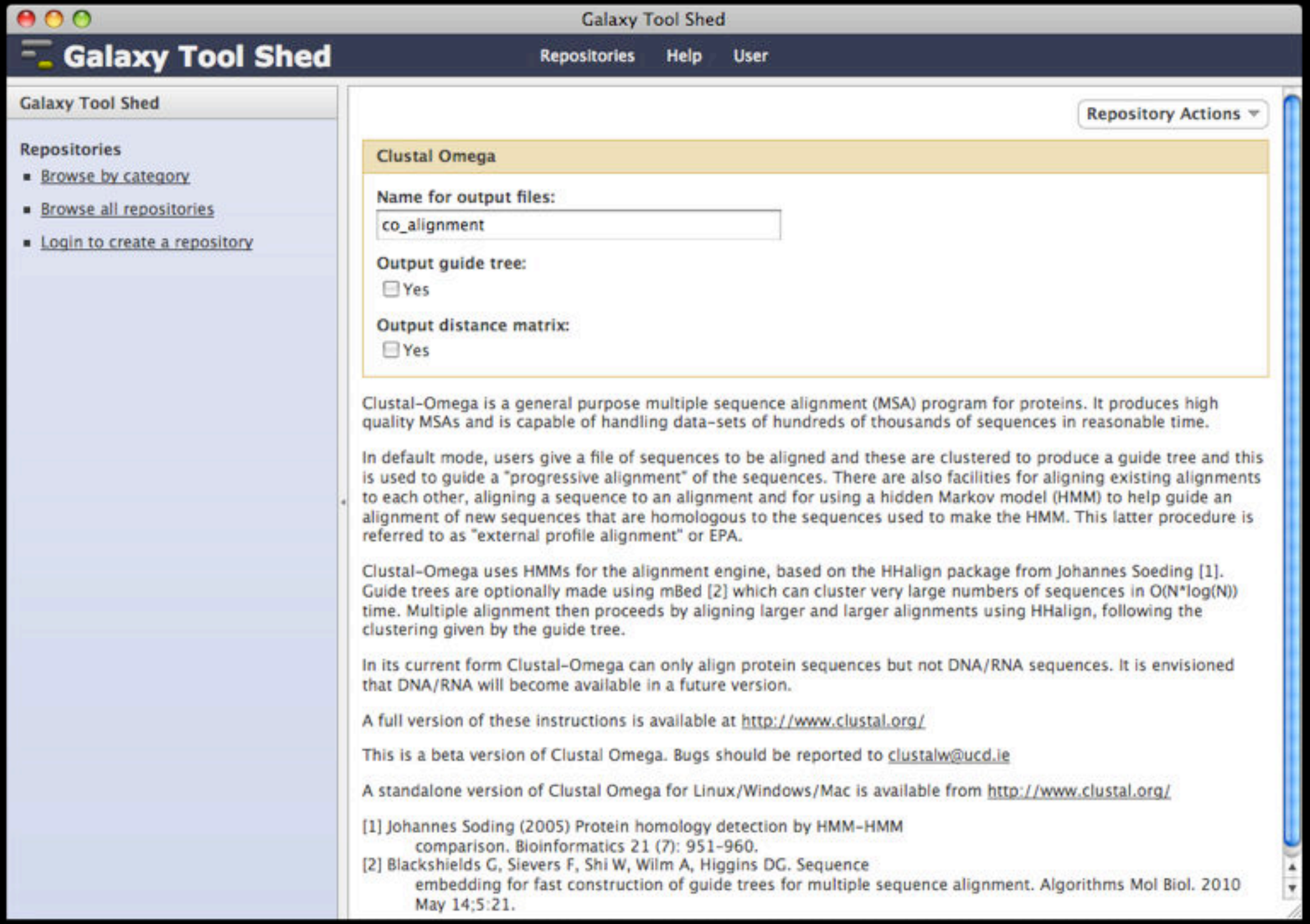
- Allow users to share "suites" containing tools, datatypes, workflows, sample data, and automated installation scripts for tool dependencies
- Integration with Galaxy instances to automate tool installation and updates

http://usegalaxy.org/community


http://toolshed.g2.bx.psu.edu - Galaxy Tool Shed
Galaxy Tool Shed
Repositories
Help
User
Galaxy Tool Shed
Repositories
Browse by category
Browse all repositories
Login to create a repository
Categories
search repository name, description
Name
Description
Repositories
Assembly
Tools for working with assemblies
12
Computational chemistry
Tools for use in computational chemistry
2
Convert Formats
Tools for converting data formats
12
Data Source
Tools for retrieving data from external data sources
3
Fasta Manipulation
Tools for manipulating fasta data
17
Graphics
Tools producing images
7
Next Gen Mappers
Tools for the analysis and handling of Next Gen sequencing data
22
Ontology Manipulation
Tools for manipulating ontologies
3
SAM
Tools for manipulating alignments in the SAM format
7
Sequence Analysis
Tools for performing Protein and DNA/RNA analysis
41
SNP Analysis
Tools for single nucleotide polymorphism data such as WGA
4
Statistics
Tools for generating statistics
8
Text Manipulation
Tools for manipulating data
14
Visualization
Tools for visualizing data
8


http://toolshed.g2.bx.psu.edu - Galaxy Tool Shed
Galaxy Tool Shed
Repositories
Help
User
Galaxy Tool Shed
Repositories
Browse by category
Browse all repositories
Login to create a repository
Repositories
search repository name, description
Advanced Search
Name ↓
Synopsis
Revision
Category
Owner
abyss_toolsuite
This suite contains Abyss and Abyss-PE config files and wrappers for Galaxy
0:92636934a189
Assembly
edward-kirton
agile_wrapper
Quickly match reads to a reference genome or sequence file
0:d6a426afaa46
Next Gen Mappers
Sequence Analysis
simonl
asdf
asdf
-1:000000000000
Statistics
Text Manipulation
vivek
assemblystats
Summarise an assembly (e.g. N50 metrics)
0:6544228ea290
Next Gen Mappers
Sequence Analysis
konradpaszkiewicz
bam_to_bigwig
Generate BigWig coverage files from BAM files. Allows gapped reads to be split (useful for RNA-Seq).
5:5b40b93ebae3
Convert Formats
SAM
Visualization
lparsons
Calculates


http://toolshed.g2.bx.psu.edu - Galaxy Tool Shed
Galaxy Tool Shed
Repositories
Help
User
Galaxy Tool Shed
Repositories
Browse by category
Browse all repositories
Login to create a repository
Repository Actions
Repository revision
2:bb1847435ec1
repository tip
Select a revision to inspect and download versions of tools from this repository.
clustalomega
Clone this repository:
hg clone http://toolshed.g2.bx.psu.edu/repos/clustalomega/clustalomega
Name:
clustalomega
Synopsis:
multiple sequence alignment program for proteins
Detailed description:
Clustal Omega is a general purpose multiple sequence alignment program for proteins. It produces high quality alignments a
Revision:
2:bb1847435ec1
Owner:
clustalomega
Times downloaded:
39
Preview tools and inspect metadata by tool version
Tools - click the name to preview the tool and use the pop-up menu to inspect all metadata
name
description
version
requirements
Clustal Omega
multiple sequence alignment program for proteins
1.0.2
none


Galaxy Tool Shed
Galaxy Tool Shed
Repositories
Help
User
Galaxy Tool Shed
Repositories
Browse by category
Browse all repositories
Login to create a repository
Repository Actions
Clustal Omega
Name for output files:
co_alignment
Output guide tree:
Yes
Output distance matrix:
Yes
Clustal-Omega is a general purpose multiple sequence alignment (MSA) program for proteins. It produces high quality MSAs and is capable of handling data-sets of hundreds of thousands of sequences in reasonable time.
In default mode, users give a file of sequences to be aligned and these are clustered to produce a guide tree and this is used to guide a "progressive alignment" of the sequences. There are also facilities for aligning existing alignments to each other, aligning a sequence to an alignment and for using a hidden Markov model (HMM) to help guide an alignment of new sequences that are homologous to the sequences used to make the HMM. This latter procedure is referred to as "external profile alignment" or EPA.
Clustal-Omega uses HMMs for the alignment engine, based on the HHalign package from Johannes Soeding [1]. Guide trees are optionally made using mBed [2] which can cluster very large numbers of sequences in O(N*log(N)) time. Multiple alignment then proceeds by aligning larger and larger alignments using HHalign, following the clustering given by the guide tree.
In its current form Clustal-Omega can only align protein sequences but not DNA/RNA sequences. It is envisioned that DNA/RNA will become available in a future version.
A full version of these instructions is available at http://www.clustal.org/
This is a beta version of Clustal Omega. Bugs should be reported to clustalw@ucd.ie
A standalone version of Clustal Omega for Linux/Windows/Mac is available from http://www.clustal.org/
[1] Johannes Soding (2005) Protein homology detection by HMM-HMM comparison. Bioinformatics 21 (7): 951-960.
[2] Blackshields G, Sievers F, Shi W, Wilm A, Higgins DG. Sequence embedding for fast construction of guide trees for multiple sequence alignment. Algorithms Mol Biol. 2010 May 14;5:21.

Agenda

Introduction

Galaxy

Worked example

Deployment Options

Microsatellites & SNP calling

Community

GMOD

Software

Community

Microsatellites

Research

A matter of life or death: How microsatellites emerge in and vanish from the human genome

Yogeshwar D. Kelkar,[1,2,6] Kristin A. Eckert,[2,3] Francesca Chiaromonte,[2,4,5,7] and Kateryna D. Makova[1,2,5,7]

Galaxy tools

Our completely reproducible computational pipeline is available in the "Regional variation" toolset of Galaxy test site (usegalaxy.org), and can be used to investigate births/deaths for in genome-wide alignments. The tool "Extract Microsatellite information" generates a summary of orthologous microsatellites, while the tool "Extract Microsatellite births and deaths" identifies births and deaths.

Genome Research, publ. online 12 Oct 2011, doi:10.1101/gr.122937.111

Microsatellites

Note different URL from paper:
http://test.g2.bx.psu.edu

- Tool list from the Test instance
- Most also available on Main
- Four of these mention “microsatellites” in their name

SNP Calling

- Two days ago, all I knew about SNP Calling was that it is hard.
- So, I turned to Galaxy:
  - Searched mailing list archive for “SNP calling”
  - Searched Tools, shared Histories, Workflows, and Pages for “SNP”
  - Search Tool Shed’s “SNP Analysis” category

SNP Calling

- Discovered
  - Tools → NGS: SAM Tools → Generate pileup
  - Tools → NGS: SAM Tools → Filter pileup
  - Tools → SNP/WGA RGenetics tool set
  - Published Histories → human mt snp discovery
  - Tool Shed → mummer_toolsuite
  - Tool Shed → ssr_marker_design
- I still don’t know how to do SNP calling, but I have an idea where to start

Agenda

Introduction

Galaxy

Worked example

Deployment Options

Microsatellites & SNP calling

Community

GMOD

Software

Community

Galaxy Community

- Tool Shed
- Local Public Installs
- Mailing Lists (very active)
- Screencasts
- Events Calendar, News Feed
- Community Wiki
- Annual Community Meeting
  - Summer 2012 in Chicago

http://galaxyproject.org/wiki/Get%20Involved

Try it now:
http://UseGalaxy.org

Develop and deploy:
http://GetGalaxy.org

EMORY

Enis Afgan

Dannon Baker

Jeremy Goecks

Dave Clements

Kanwei Li

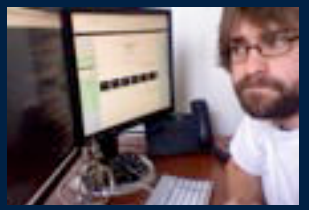
James Taylor

PENNSTATE

1855

Guru Ananda

Dan Blankenberg



Nate Coraor

Jennifer Jackson

Greg von Kuster

Anton Nekrutenko

Supported by the **NHGRI** (HG005542, HG004909, HG005133), **NSF** (DBI-0850103), Penn State University, Emory University, and the Pennsylvania Department of Public Health

http://GalaxyProject.org

Agenda

Introduction

Galaxy

Worked example

Deployment Options

Microsatellites & SNP calling

Community

GMOD

Software

Community

GMOD Software

Visualization

- **GBrowse** & **JBrowse** genomic browsers
- **CMap** & **GBrowse_syn** for comparative genomics

Annotation

- **MAKER** & **Apollo**

Data Integration and Querying

- **BioMart, InterMine, Chado**

Analysis

- **Galaxy, Ergatis, ISGA**

http://gmod.org/

Visualization: GBrowse

- Visualize features, SNPs, quantitative data, NGS data (uses SAMtools)
- Custom tracks, shared tracks, DAS
- Highly customisable by both admin and end-user
- Huge community

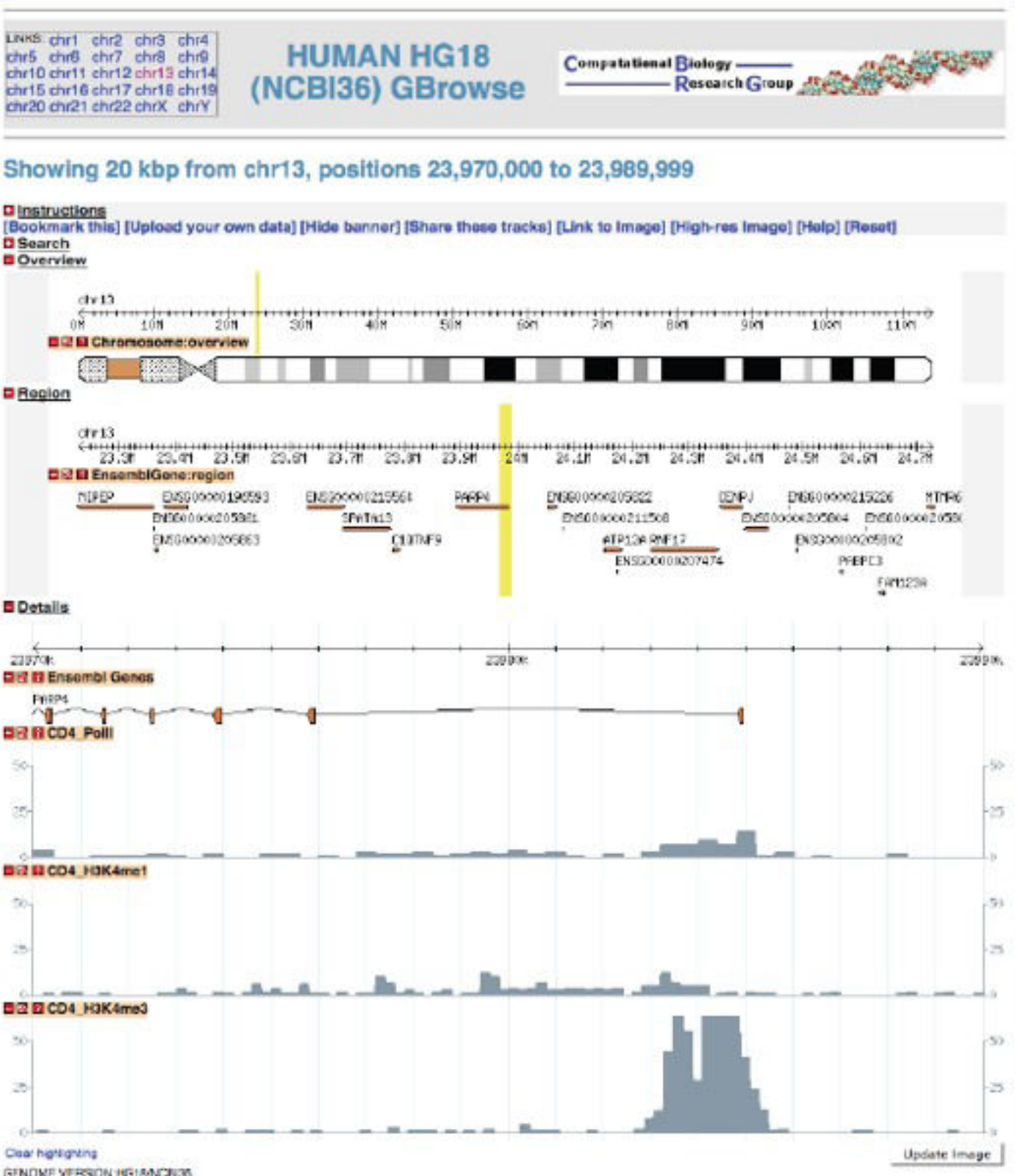
Stein, *et al.*, The Generic Genome Browser: A Building Block for a Model Organism System Database, *Genome Res.* 2002. 12: 1599-1610

GBrowse:

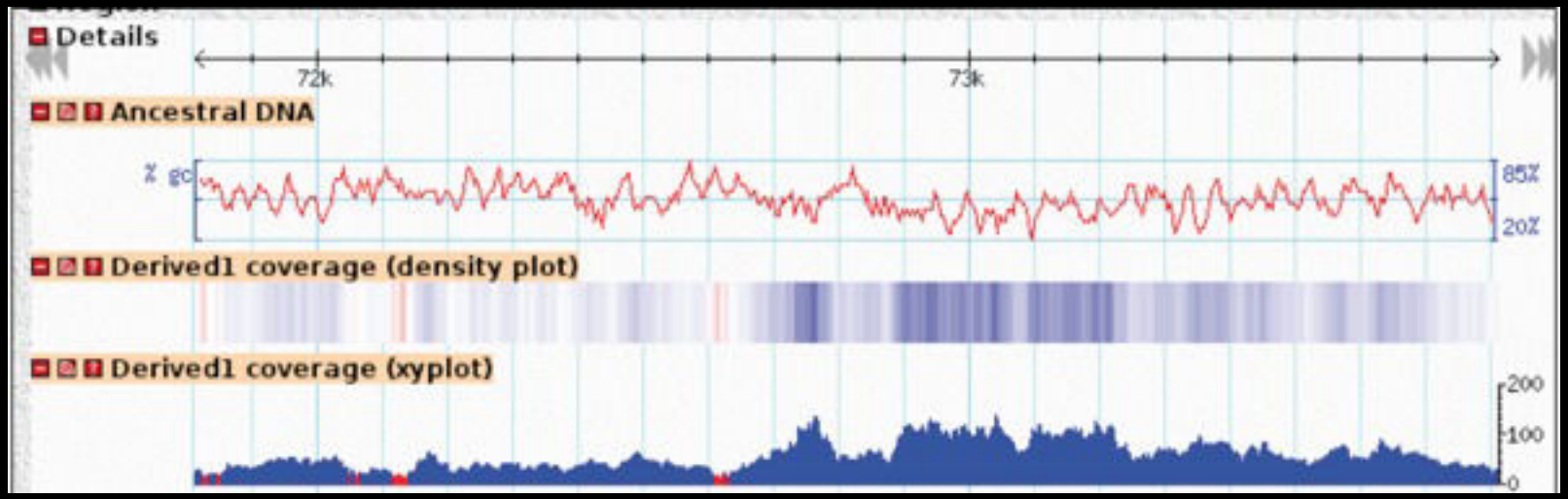
Genome Browser

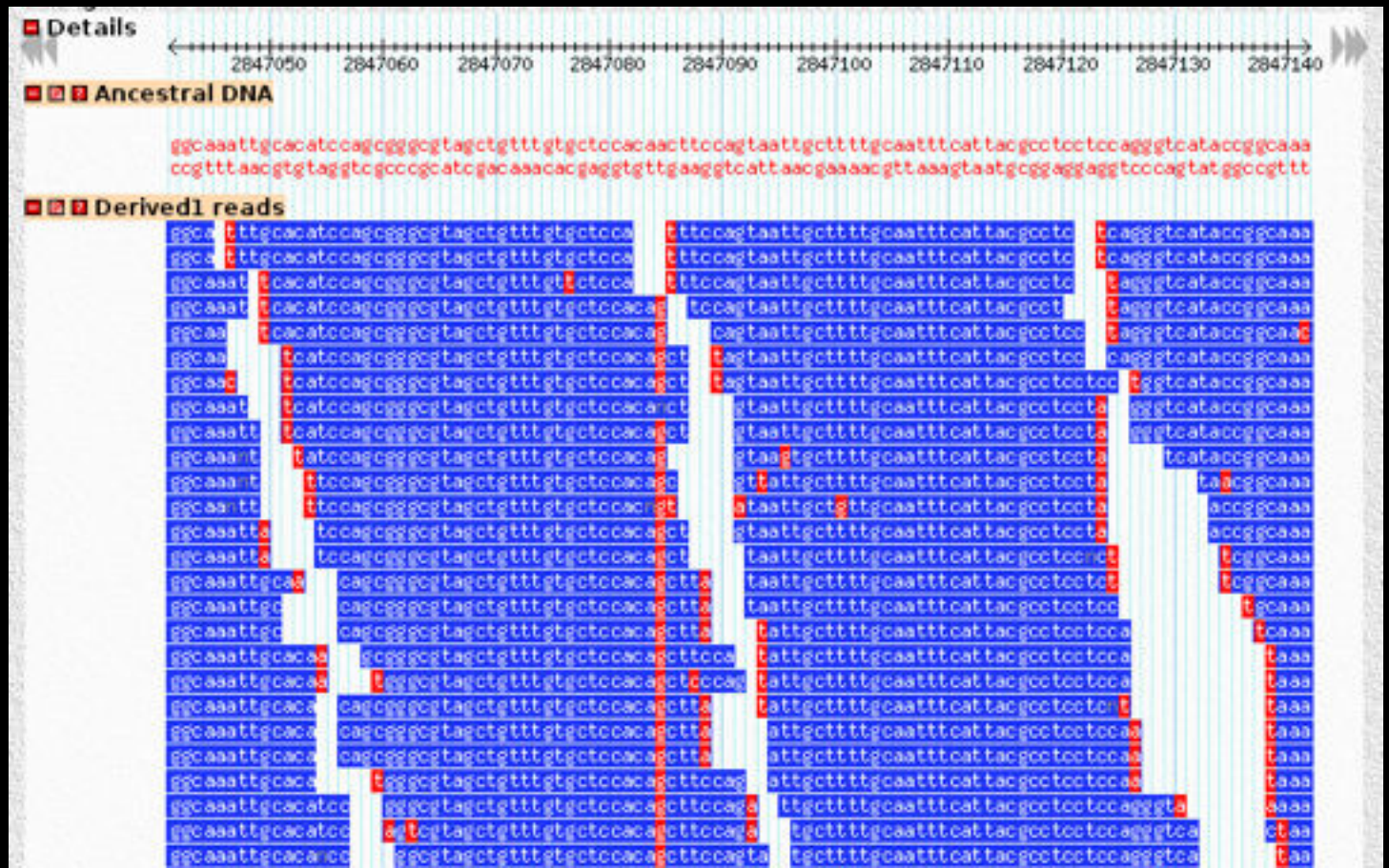
ChIP-Seq data

Visualization by Computational Biology Research Group @ Oxford.



GBrowse: Short Reads

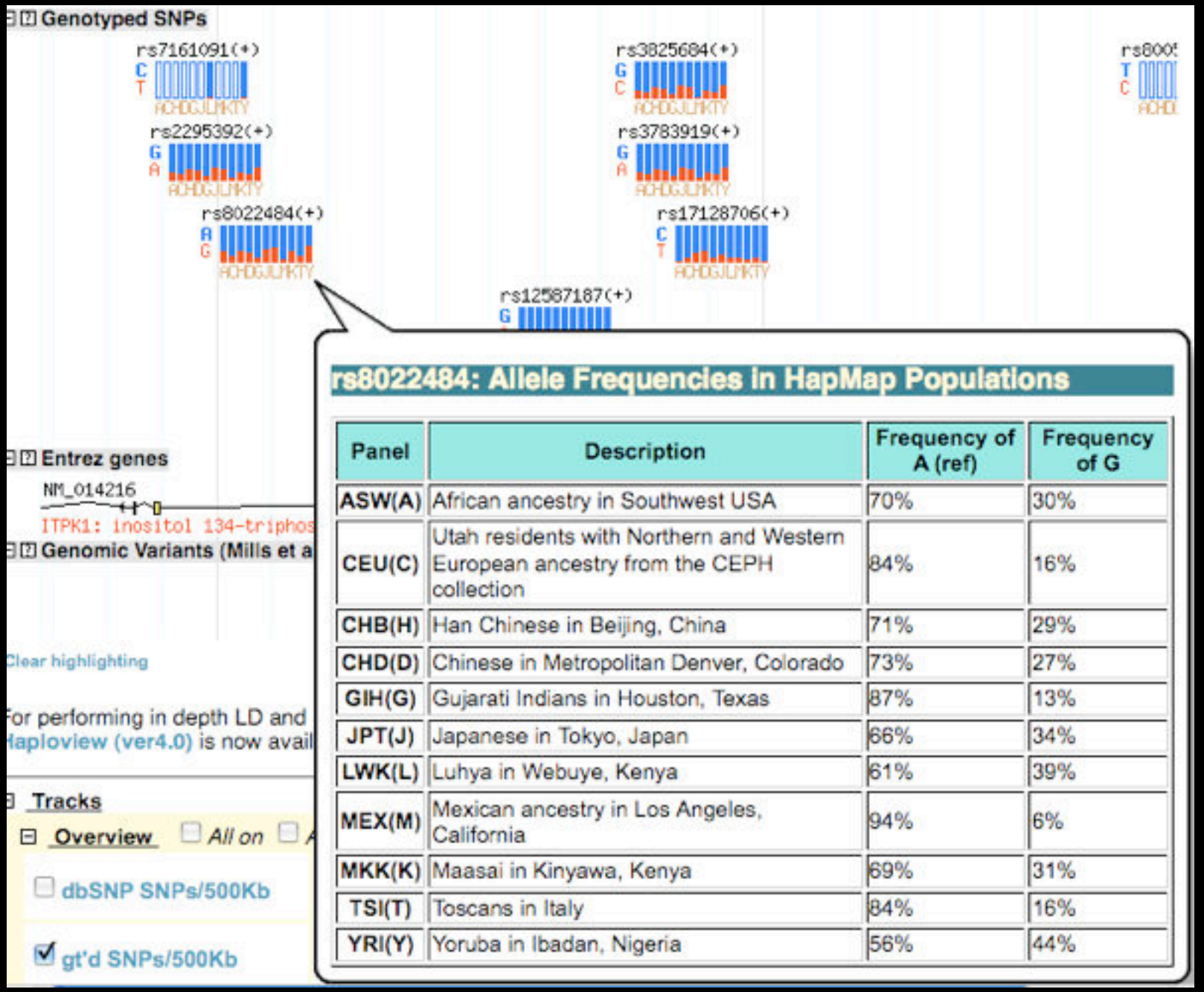




GBrowse

HapMap Allele Frequencies

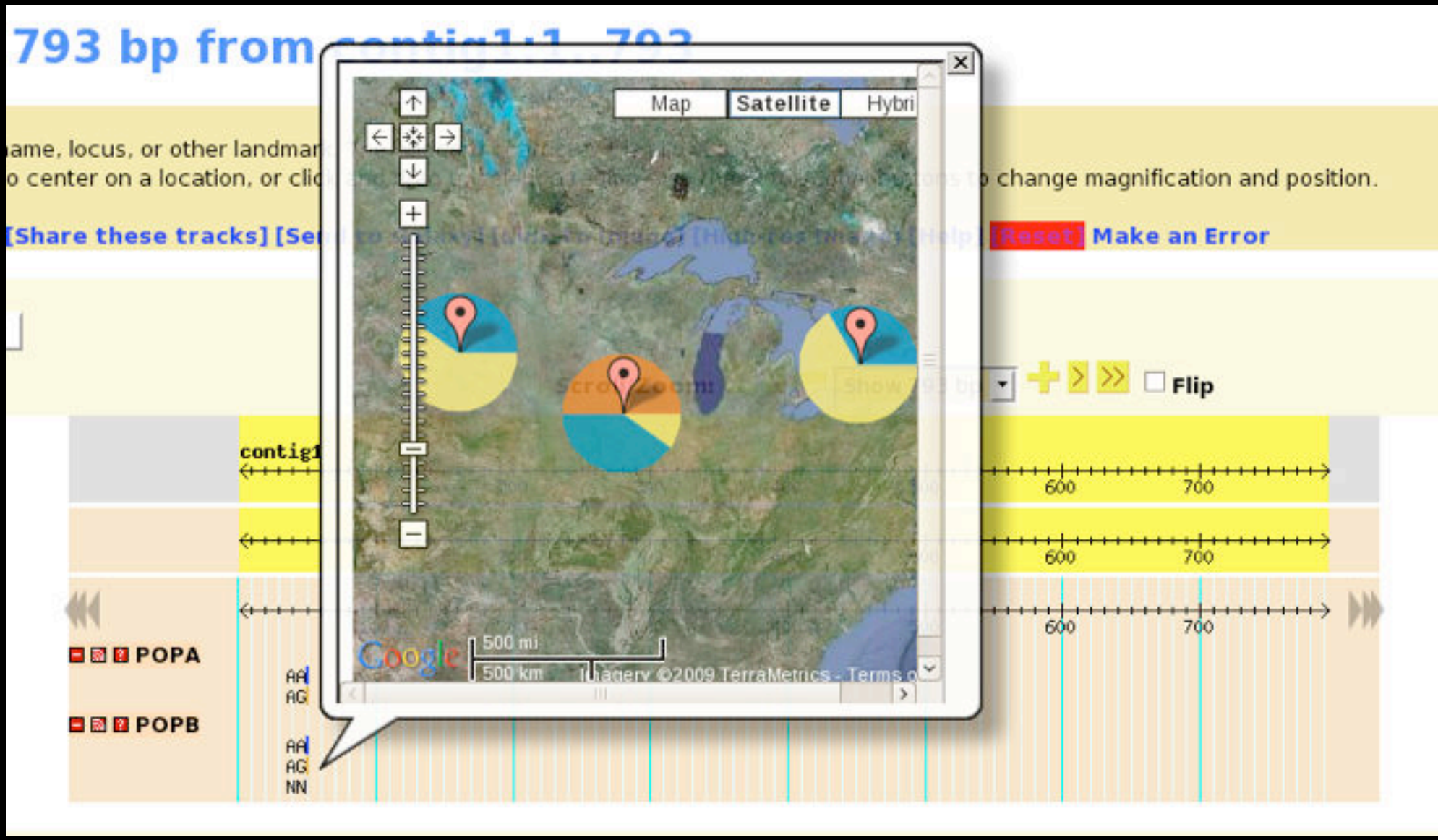
hapmap.org



rs8022484: Allele Frequencies in HapMap Populations

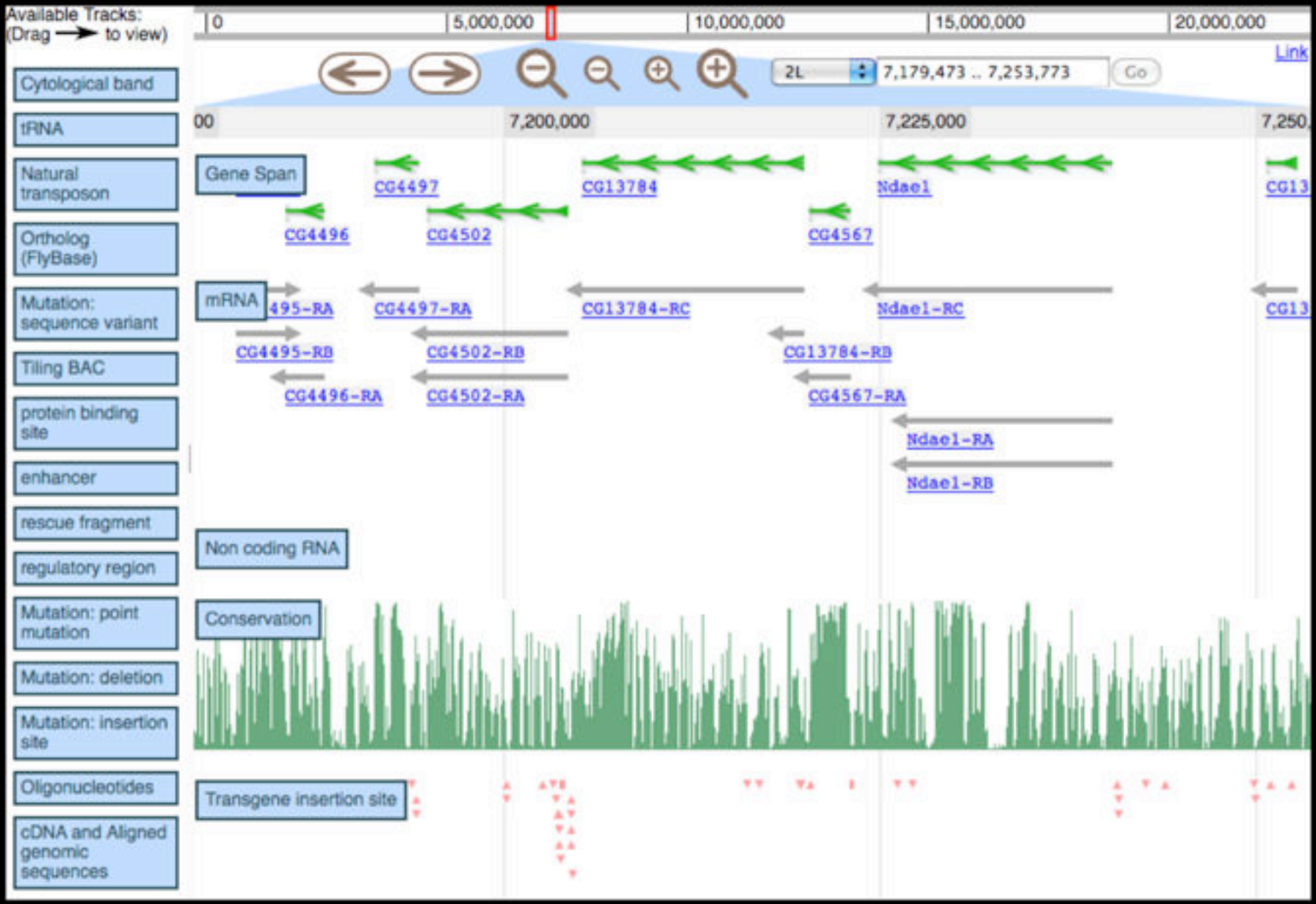
Panel	Description	Frequency of A (ref)	Frequency of G
ASW(A)	African ancestry in Southwest USA	70%	30%
CEU(C)	Utah residents with Northern and Western European ancestry from the CEPH collection	84%	16%
CHB(H)	Han Chinese in Beijing, China	71%	29%
CHD(D)	Chinese in Metropolitan Denver, Colorado	73%	27%
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%

GBrowse



Work by Yi-Hsin Erica Tsai & Ben Faga, using PhyloGeoViz

JBrowse



Skinner, *et al.*, JBrowse: A next-generation genome browser, *Genome Res*. 2009. 19: 1630-1638

http://jbrowse.org/

GBrowse or JBrowse?

GBrowse

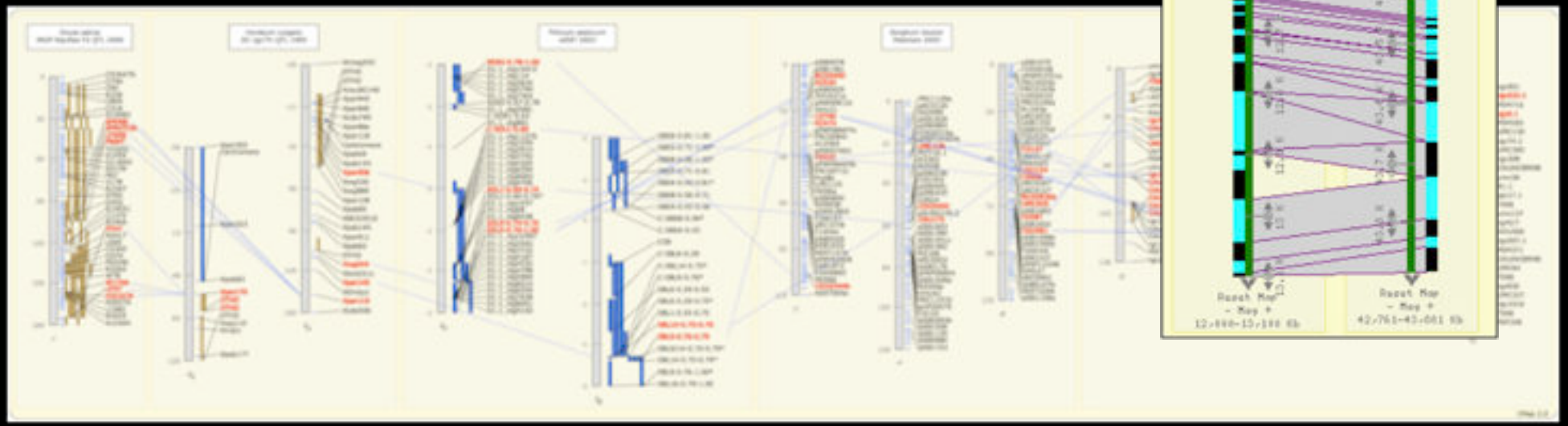
- Widely used
- Robust features & ecosystem
- Familiar interface
- Configuration is trial and error (see WebGBrowse)
- Requires more server
- Slower

JBrowse

- Limited features
- Unfamiliar
- Lots of future development
- Configuration is simpler
- Requires less server
- Much Faster

CMap

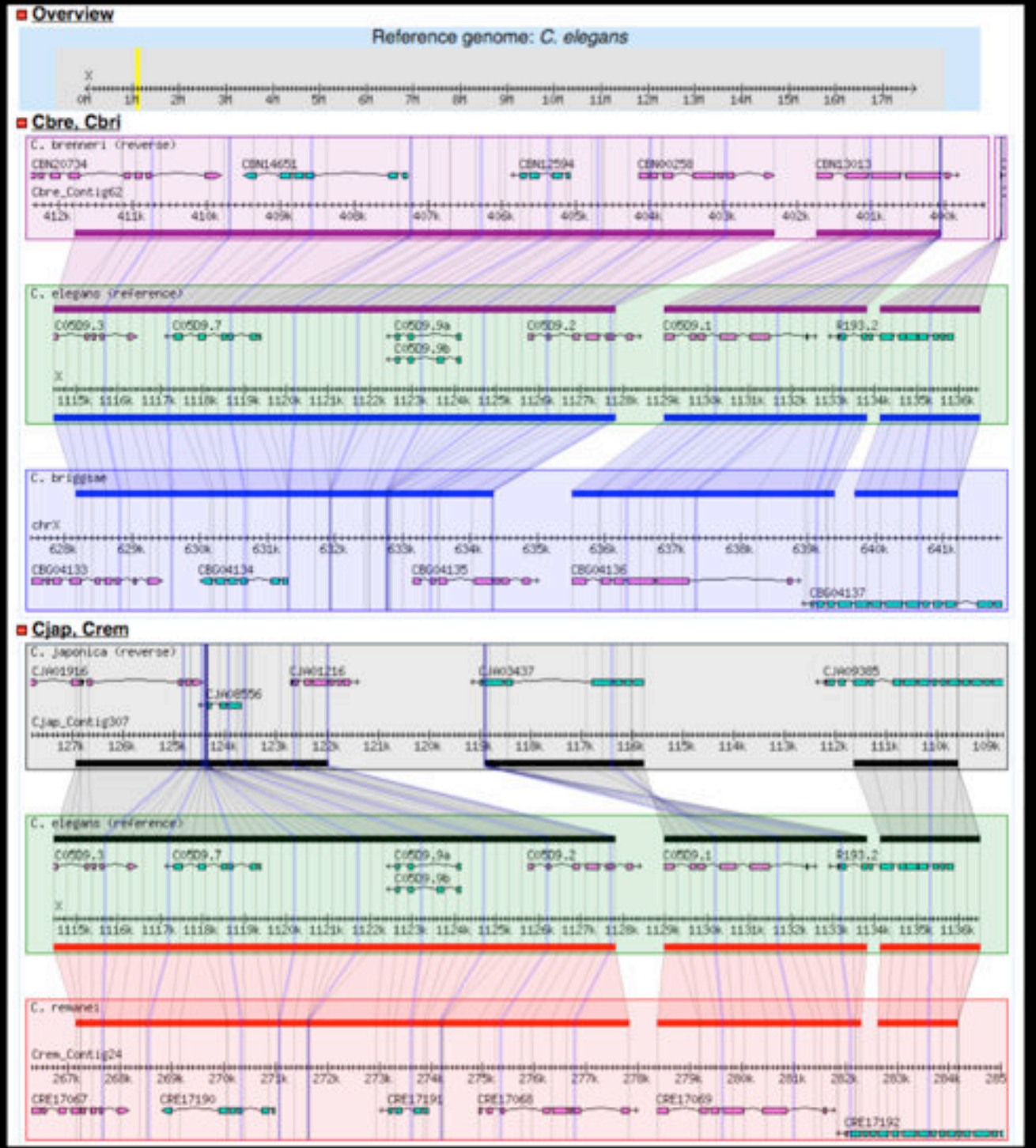
- Comparative map viewer
- Data type agnostic. Can show linkage, physical, deletion, QTL, ...
- Anything that is points or regions on a line



Youens-Clark, *et al*., CMap 1.01: a comparative mapping application for the Internet, *Bioinformatics* (2009) 25 (22): 3040-3042.

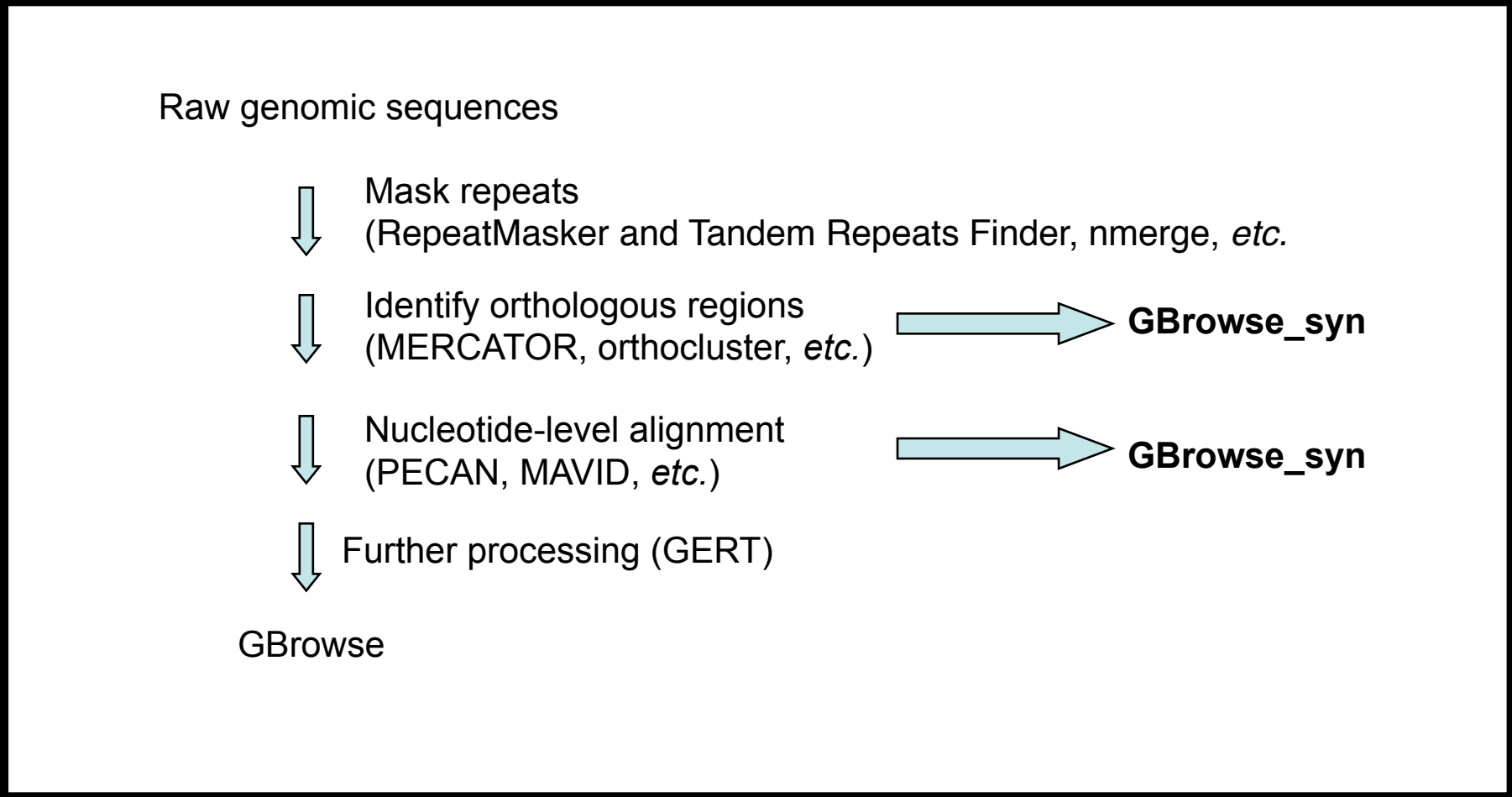
GBrowse_syn

- Comparative genomics browser
- Shows a “reference” compared to 1 or more others
- Built on GBrowse and can show any GBrowse-based annotations

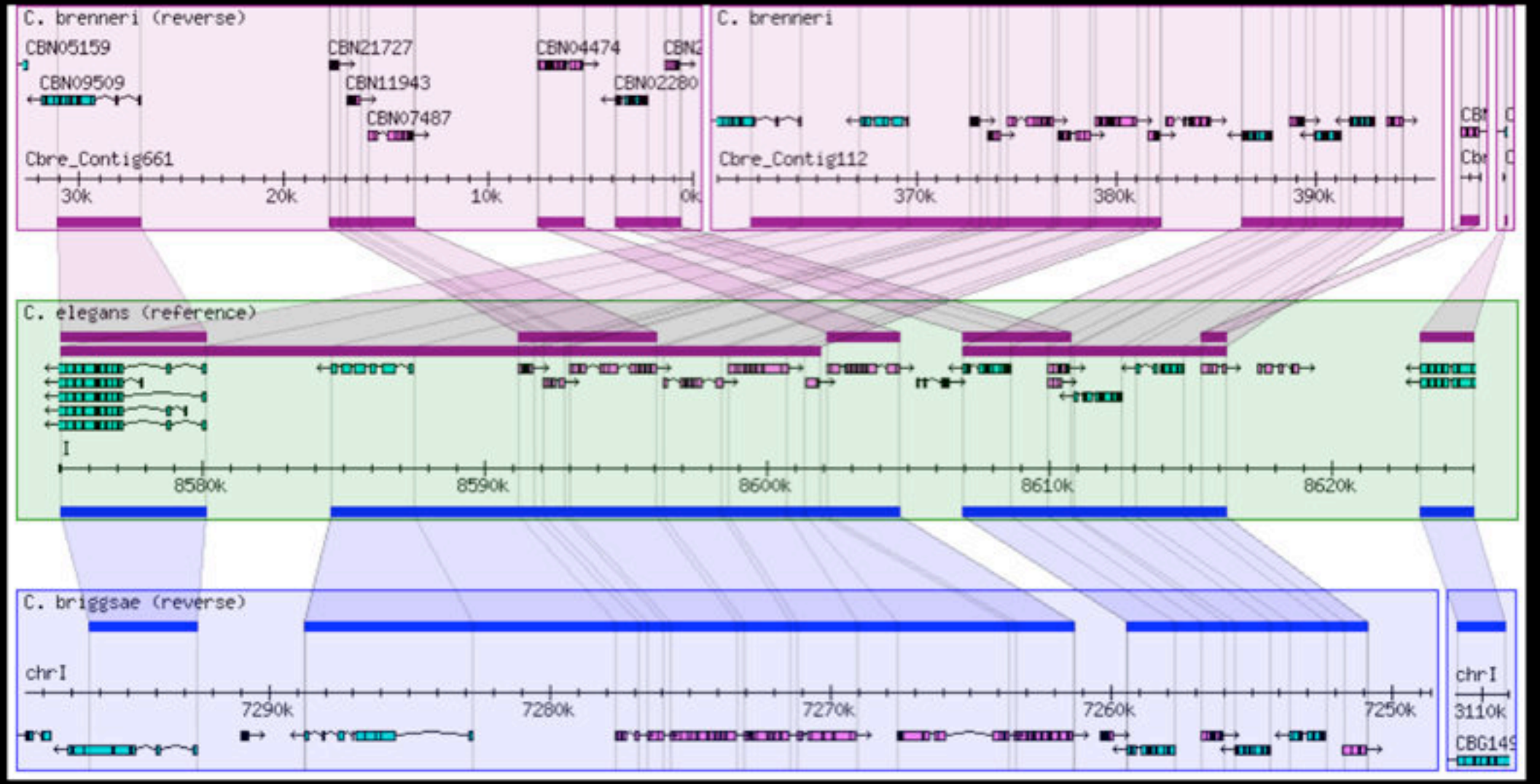


McKay, *et al.*, Using the Generic Synteny Browser (GBrowse_syn), *Current Protocols in Bioinformatics*, Unit 9.12, September 2010

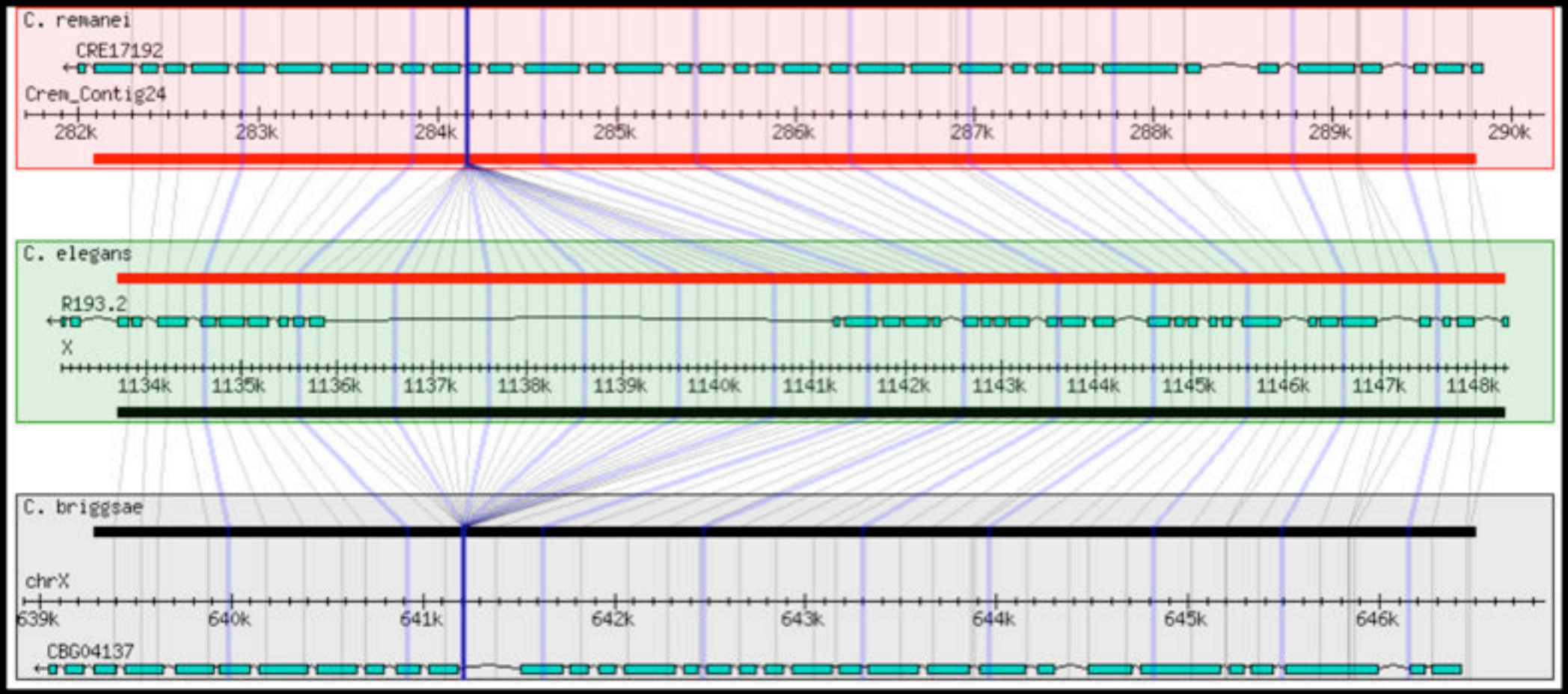
GBrowse_syn: Big Picture



GBrowse_syn: Duplications



GBrowse_syn: Base level alignments

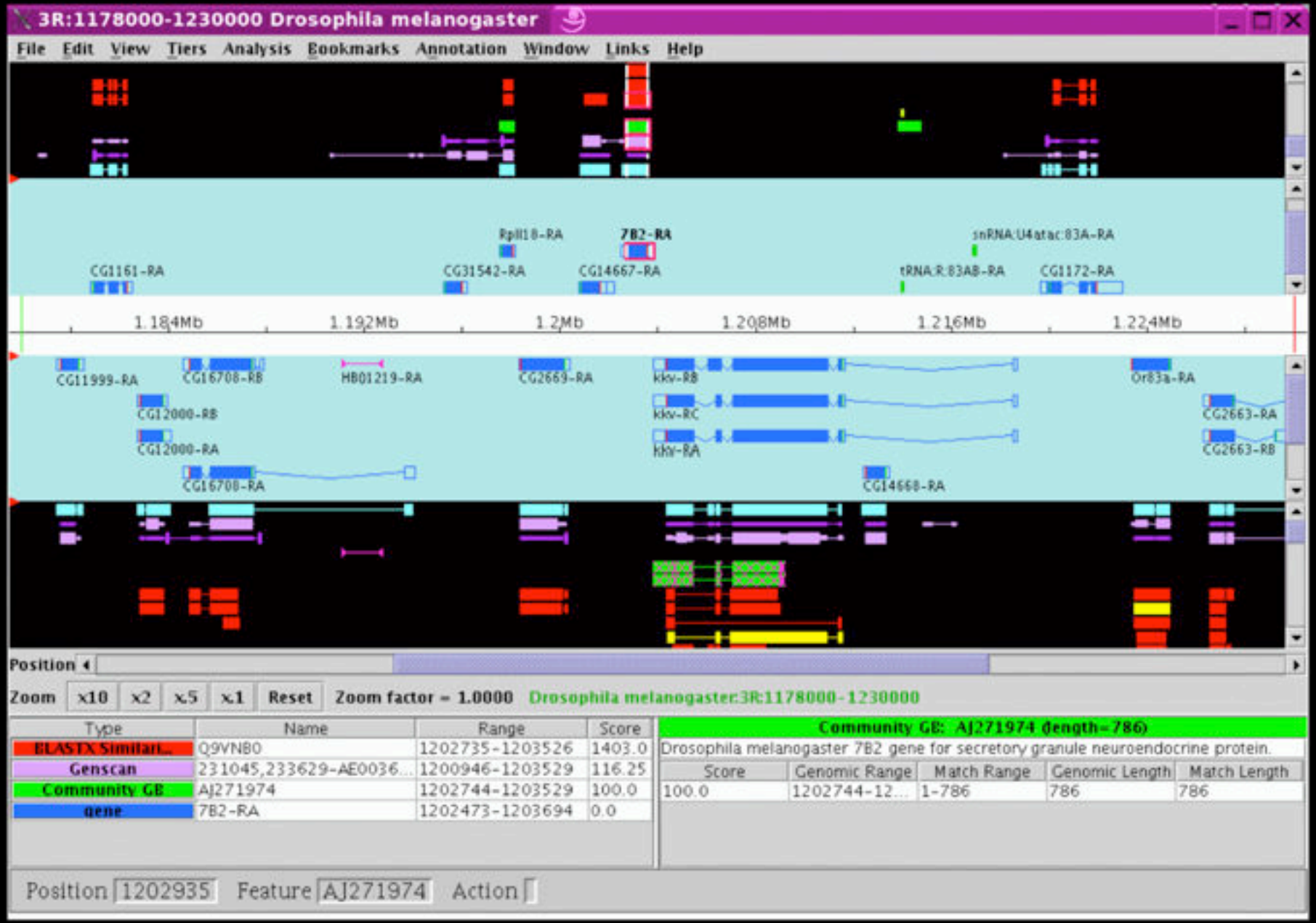


Annotation: MAKER

- Genome annotation pipeline for creating gene models
- Produces GFF3 and FASTA which can be loaded into GBrowse, JBrowse, Apollo, Chado, Galaxy, BioMart, InterMine, ...
- Incorporates SNAP, RepeatMasker, exonerate, BLAST, Augsustus, FGENESH, GeneMark, MPI
- Can also map existing annotation on to new assemblies and update existing annotations with new evidence
- Also available as a hosted service.

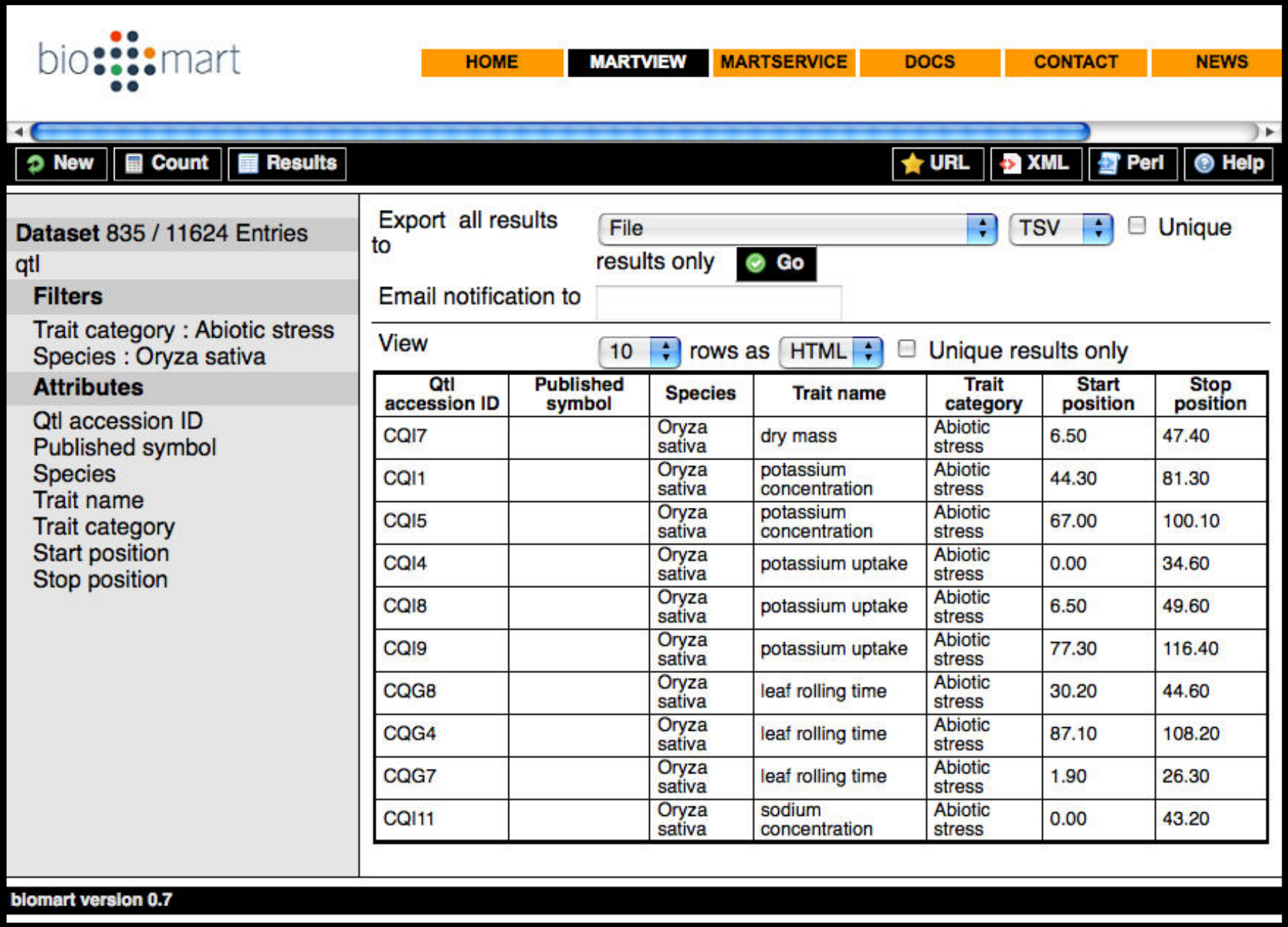
Cantarel, *et al.,* MAKER: An Easy-to-use Annotation Pipeline Designed for Emerging Model Organism Genomes,. *Genome Research* 2008 18(1) 188-96.

Apollo: Manual genome annotation editor



Lewis *et al.*, Apollo: a sequence annotation editor.,
Genome Biology 2002, 3(12)

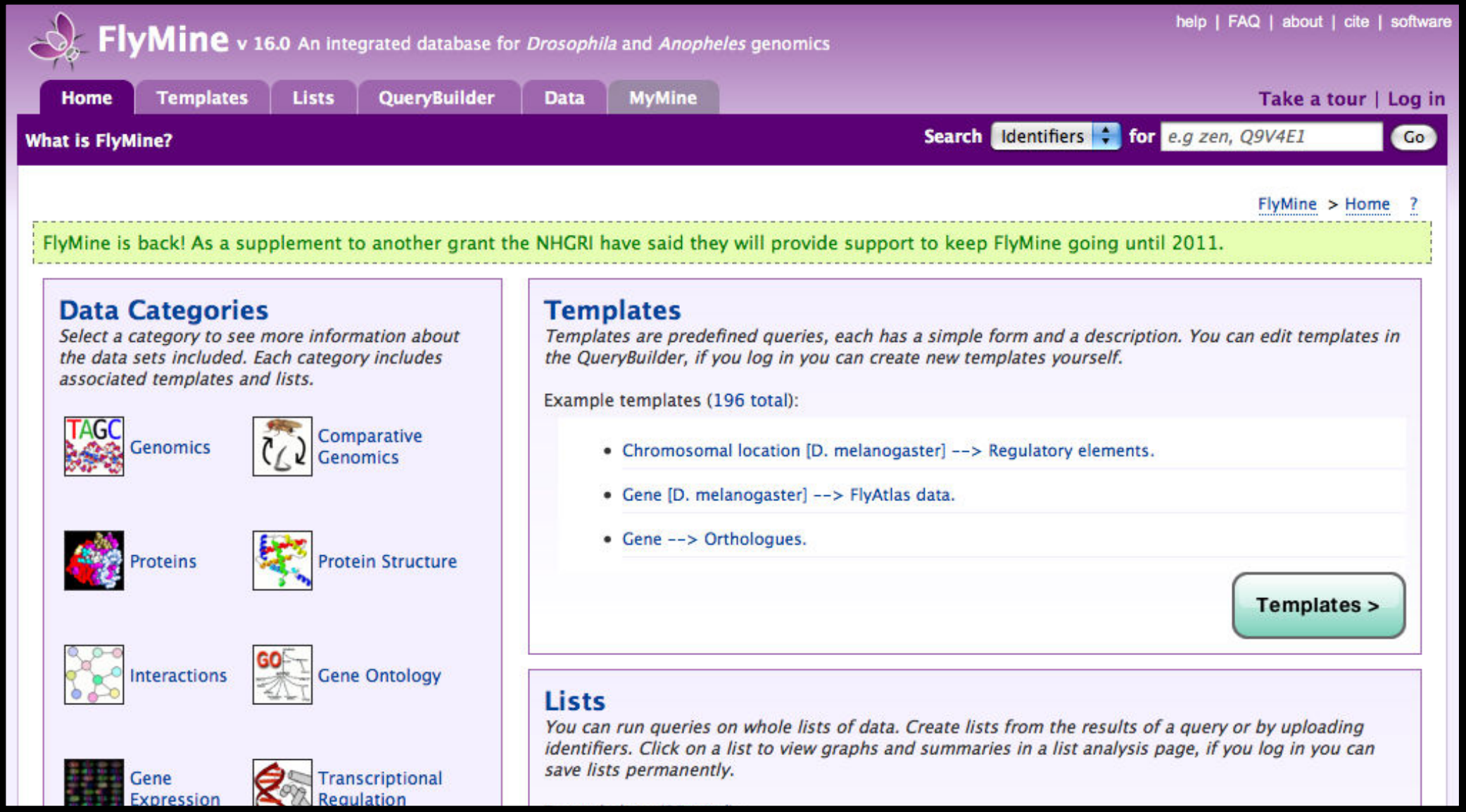
Data Integration & Querying: BioMart



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

Smedley, *et al.*, BioMart – biological queries made easy, *BMC Genomics* 2009, 10:22

Data Integration & Querying: InterMine



Agenda

Introduction

Galaxy

Worked example
Deployment Options
Microsatellites & SNP calling
Community

GMOD

Software
Community

GMOD Community

- Mailing lists (very active)
- Annual week long course
- Tutorials
- Events Calendar, News Feed
- Community Wiki
- Semi-Annual Community Meetings
  - April 2012, Washington, DC, co-located with Biocuration 2012

http://gmod.org/

GMOD Contributors

Scott Cain
Carson Holt
Brian Osborne
Ed Lee
Stephen Ficklin
Mitch Skinner
Ian Holmes
Gos Micklem
Barry Moore
Dave Emmert
Alex Kalderimis
Steve Taylor
Dorie Main
Brian O'Connor
Sook Jung
…

Sheldon McKay
Lincoln Stein
Jim Hu
Daniel Renfro
Todd Vision
Nathan Liles
Chris Hemmerich
James Taylor
LBCC Adv. Design
Seth Redmond
Chris Fields
Ben Faga
Joshua Orvis
Junjun Zhang
Naama Menda
…

Hilmar Lapp
Dan Blankenberg
John Aikman
Robert Buels
Richard Smith
Ken Youens-Clark
Nicole Washington
Alexie Papanicolaou
Josh Goodman
Yuri Bendana
Jason Stajich
Suzi Lewis
Mark Yandell
Don Gilbert
Meg Staton
…

Plus hundreds, *if not thousands*, of others

Galaxy & GMOD: Some Australian Resources

Dr Ross Lazarus @ Baker IDI, Melbourne,
Core Galaxy Team Member

CSIRO

CSIRO looking at deploying Galaxy and GBrowse across CSIRO life sciences

vlsci
LIFE SCIENCES COMPUTATION

VLSCI doing pilot project to deploy Galaxy in the Australian Research Cloud infrastructure

Thanks

UTAS

Dr Adam Smolenski

Dr Rene Vaillancourt

CSIRO

Dr David Lovell

Dr Jen Taylor

Dr Annette McGrath

Dr Alexie Papanicolaou

Dr Warwick Irwin