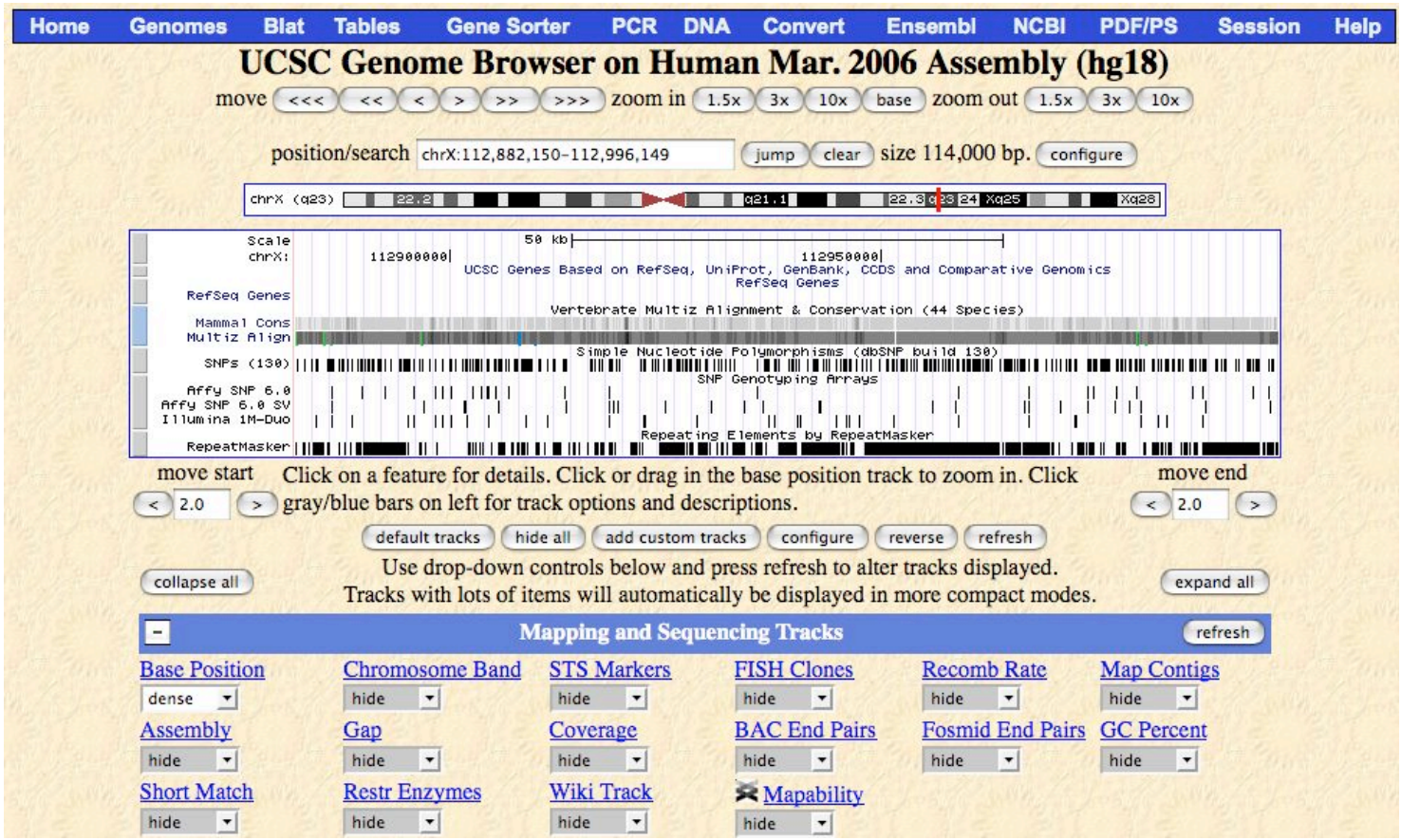
A Survey of Genome and Comparative Genome Browsers

Sheldon McKay, iPlant Collaborative; CSHL
Dec 2, 2009

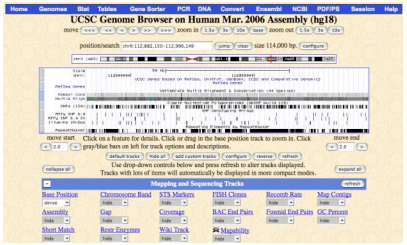
Outline:

- The “big three” centralized genome browsers
- Whole genome browsing
- The Generic Genome Browser and GMOD
- Comparative genome browsing with GBrowse_syn
- Dense Data Browsing

UCSC Genome Browser



http://genome.ucsc.edu



UCSC Genome Browser

- Data sourced from NCBI's RefSeq, the Encyclopedia of DNA Elements (ENCODE) project and UCSC's own genome annotation pipeline

- 48 species represented*.

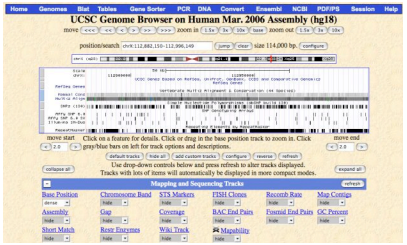
- Features sequence conservation data for 28-way comparative alignment, plus many other tracks

- 178 data tracks in the human genome browser*

- Simple user interface, typical entry point is a home page for each species

- Extensive support for third party data uploads and custom tracks

* As of Sept, 2008



UCSC Genome Browser

• Outbound data sharing via the Distributed Annotation Protocol (DAS), table views and an FTP site.

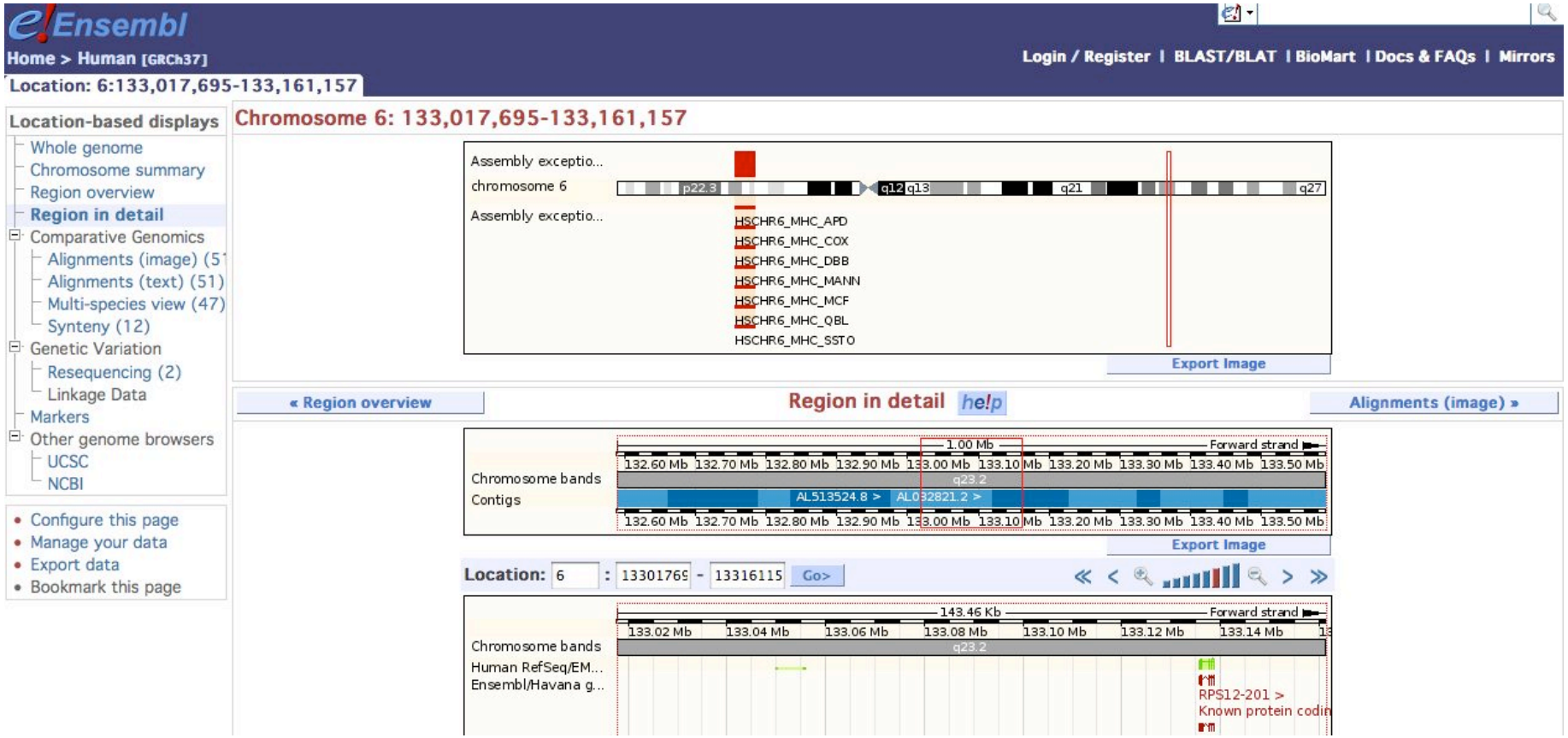
• Written in C, many database optimizations, fast and responsive

• The browser software is open-source for non-commercial users but the code base is complex and not well documented and challenging to deploy.

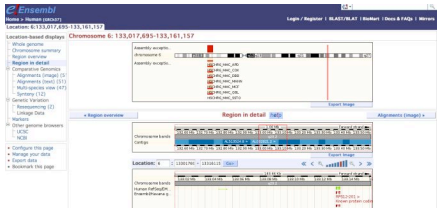
• Three official mirrors: Medical College of Wisconsin, Duke, and Cornell Universities

• Based on web access logs, as many as a dozen unofficial sites mirror UCSC data (H. Clawson, personal communication).

Ensembl Genome Browser



http://www.ensembl.org



Ensembl Genome Browser

- Data sourced from EMBL and other external sources

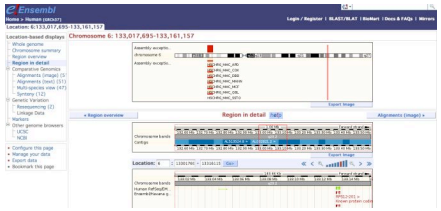
- Ensembl has its own extensive genome annotation pipeline

- Some example recent additions include genome-wide maps of protein-DNA interactions and the regulatory build, an effort to annotate all cis-regulatory sequences.

- As of release 52, 46 species plus three in pre-release, over 200 data tracks for the human genome.

- As of Sept 2008, Ensembl has 250 queries/second on its database (Guilietta Spudich, personal communication)

* As of Sept, 2008



Ensembl Genome Browser

- Ensembl offers data sharing and custom tracks via DAS

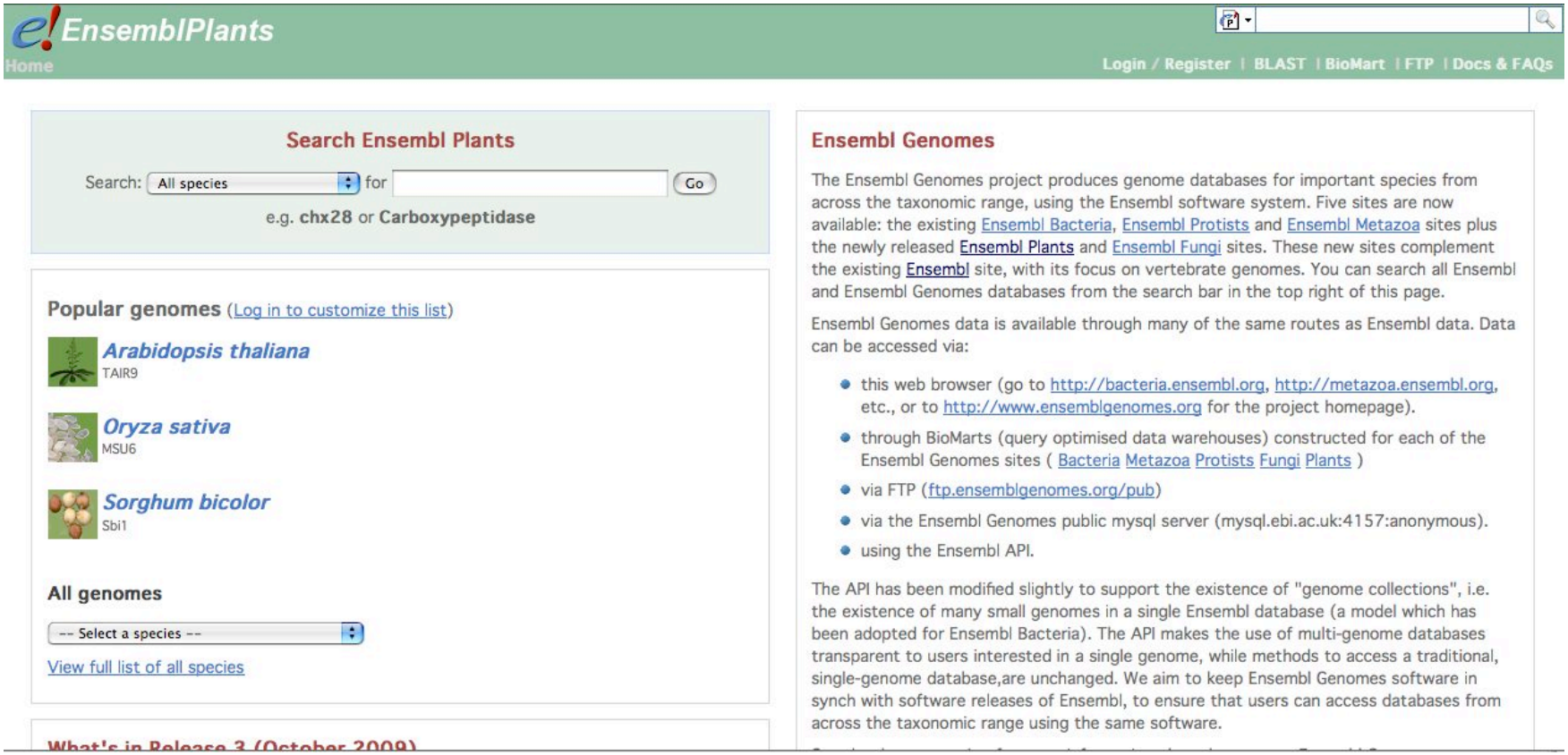
- Data export is also available through an API to its public web server and via BioMart, a GMOD tool that supports data mining for various other major databases, including WormBase, HapMap, and VectorBase.

- Ensembl software and infrastructure are open source and fairly well documented; it is used by the Gramene database, among others. Adopters are tracked at http://www.ensembl.org/info/about/ensembl_powered.html

* As of Sept, 2008

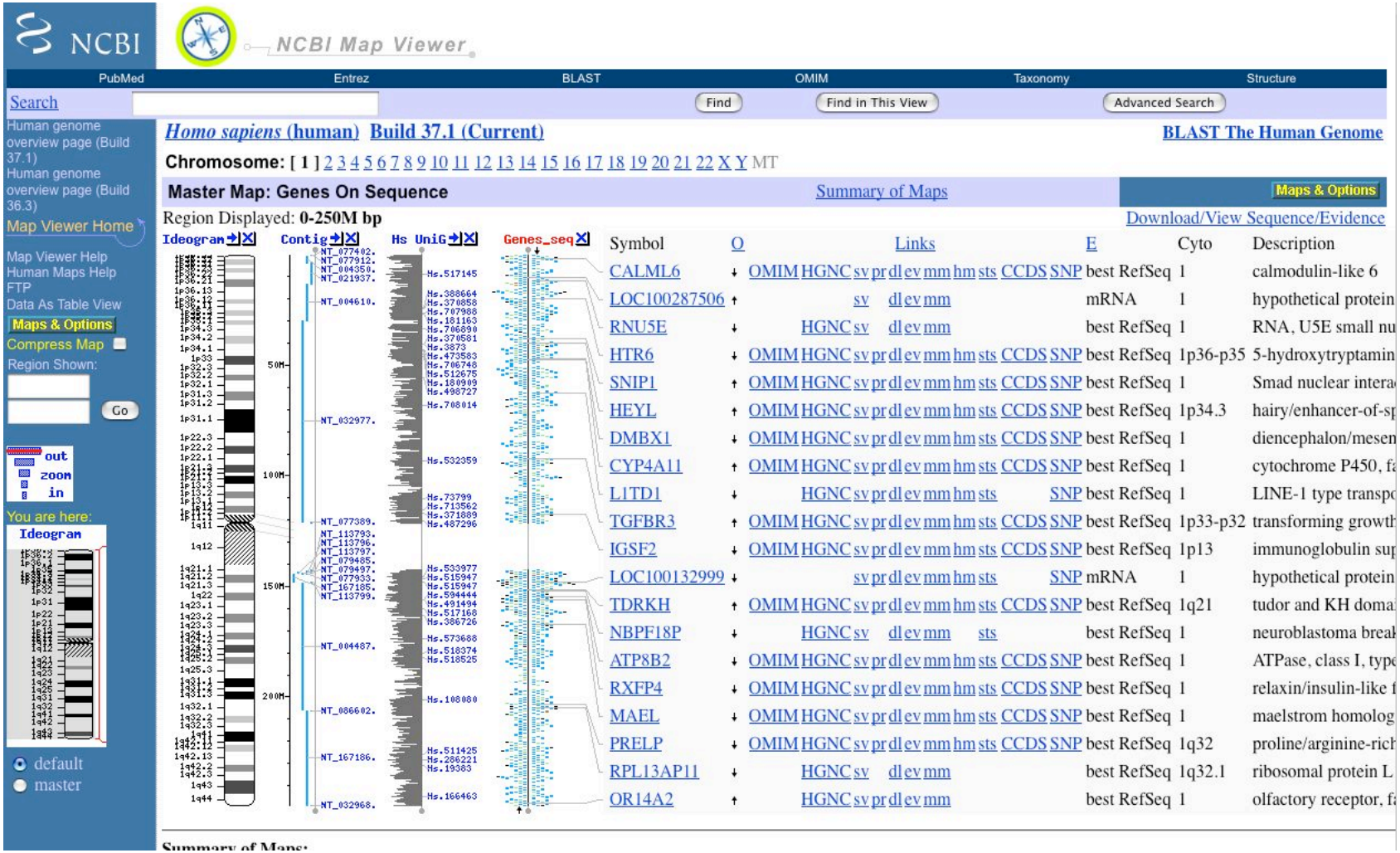
October 2009: Ensembl Plants!

Collaboration between EBI and Doreen Ware's group at CSHL

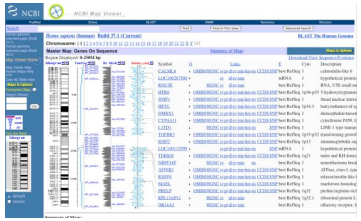


http://plants.ensembl.org/index.html

NCBI Map Viewer



www.ncbi.nih.gov/mapview



NCBI Map Viewer

- National Center for Biotechnology Information best known for GenBank, PubMed, RefSeq, etc.

- Also has a Map Viewer for genome annotations

- Draws from the formidable NCBI toolkit.

- Supports 106 Species*, but a relatively small number of tracks

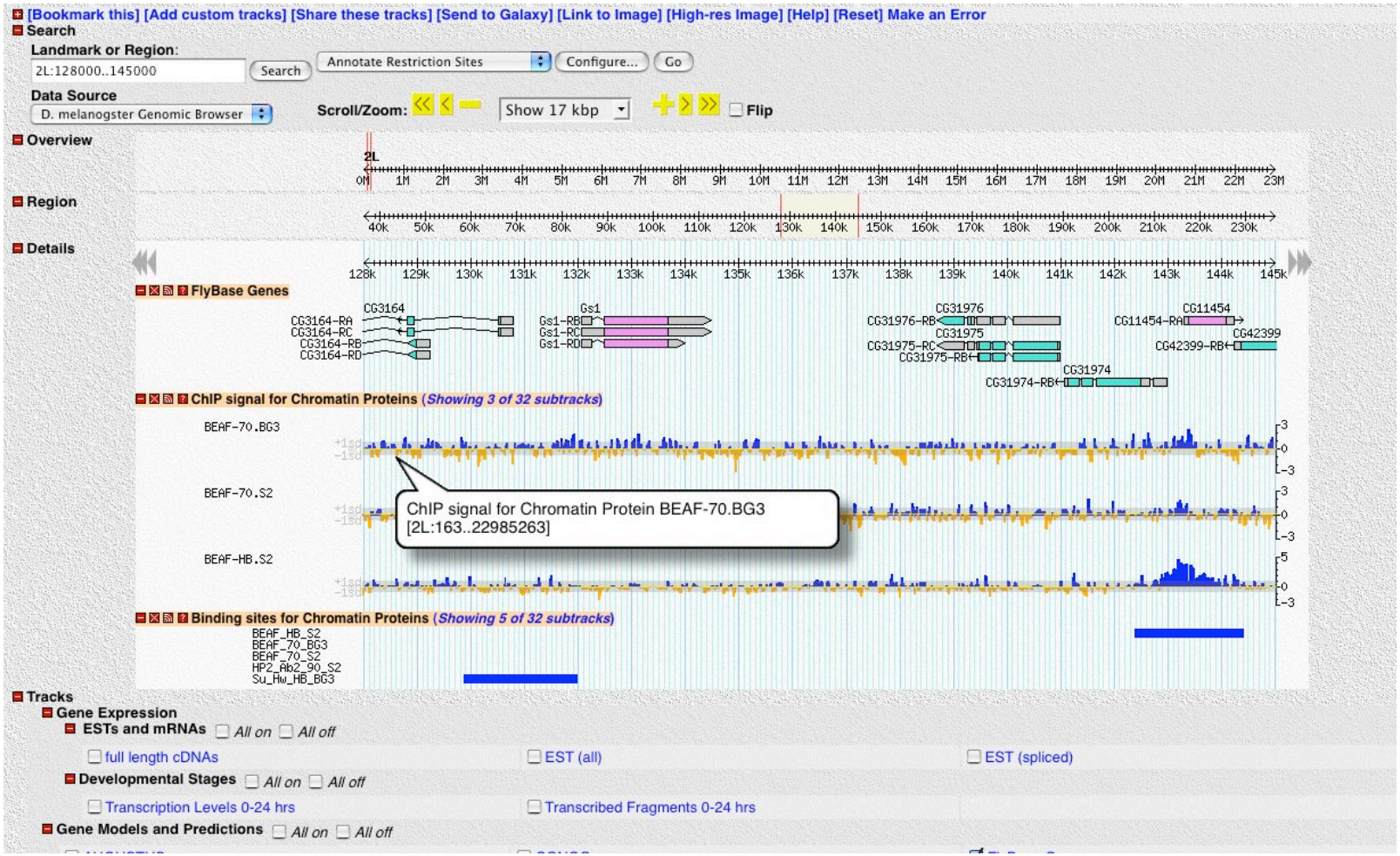
- Navigation features of the interface are somewhat limited

- Underlying data available via ftp or the BioPerl API (DO NOT attempt “screen scraping” scripts)

- No support for custom tracks, third-party annotation, DAS

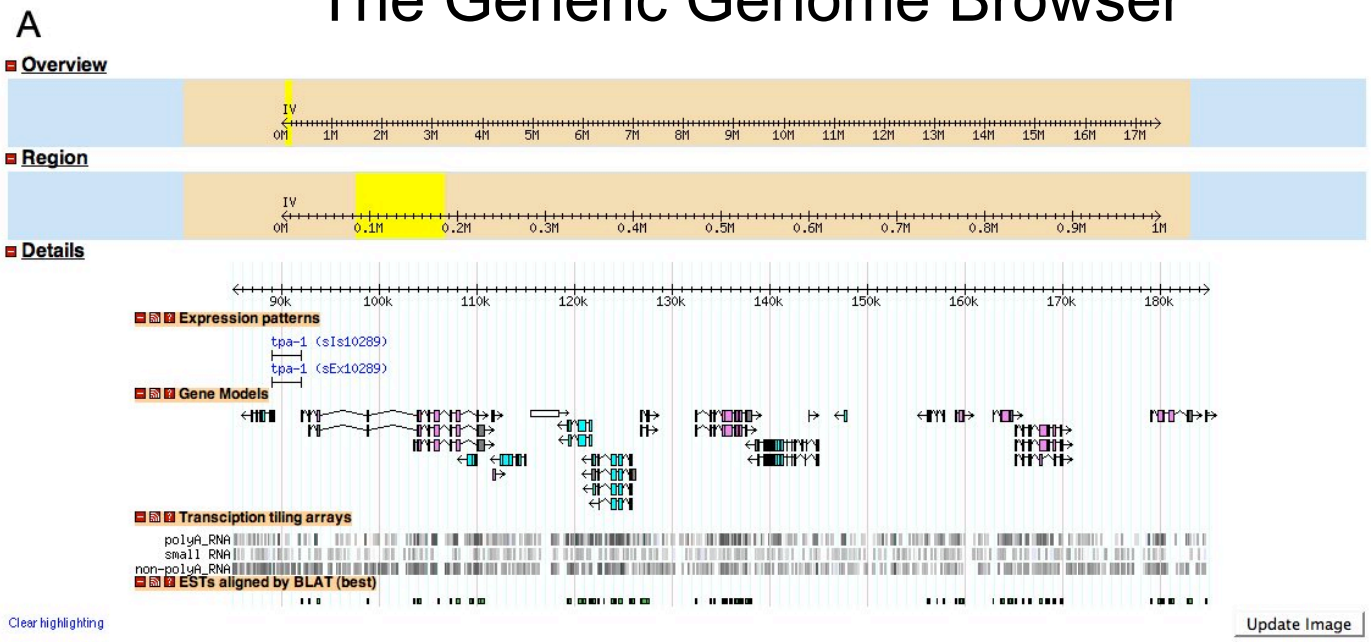
* As of Sept, 2008

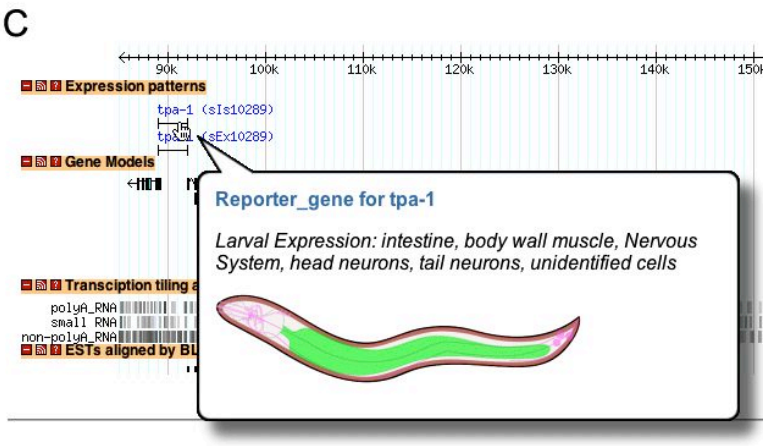
The Generic Genome Browser



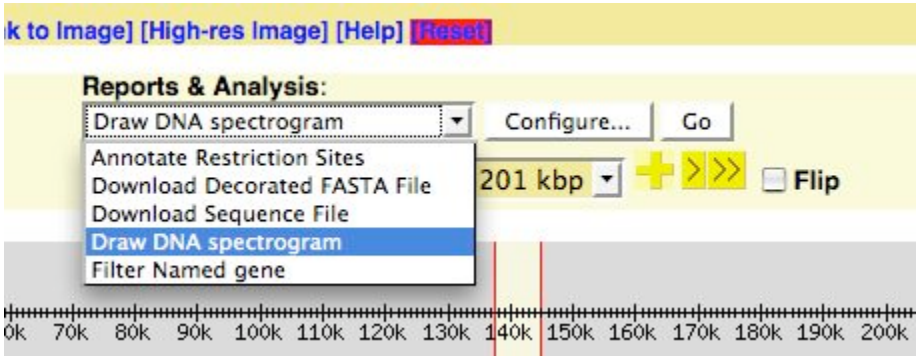
gmod.org

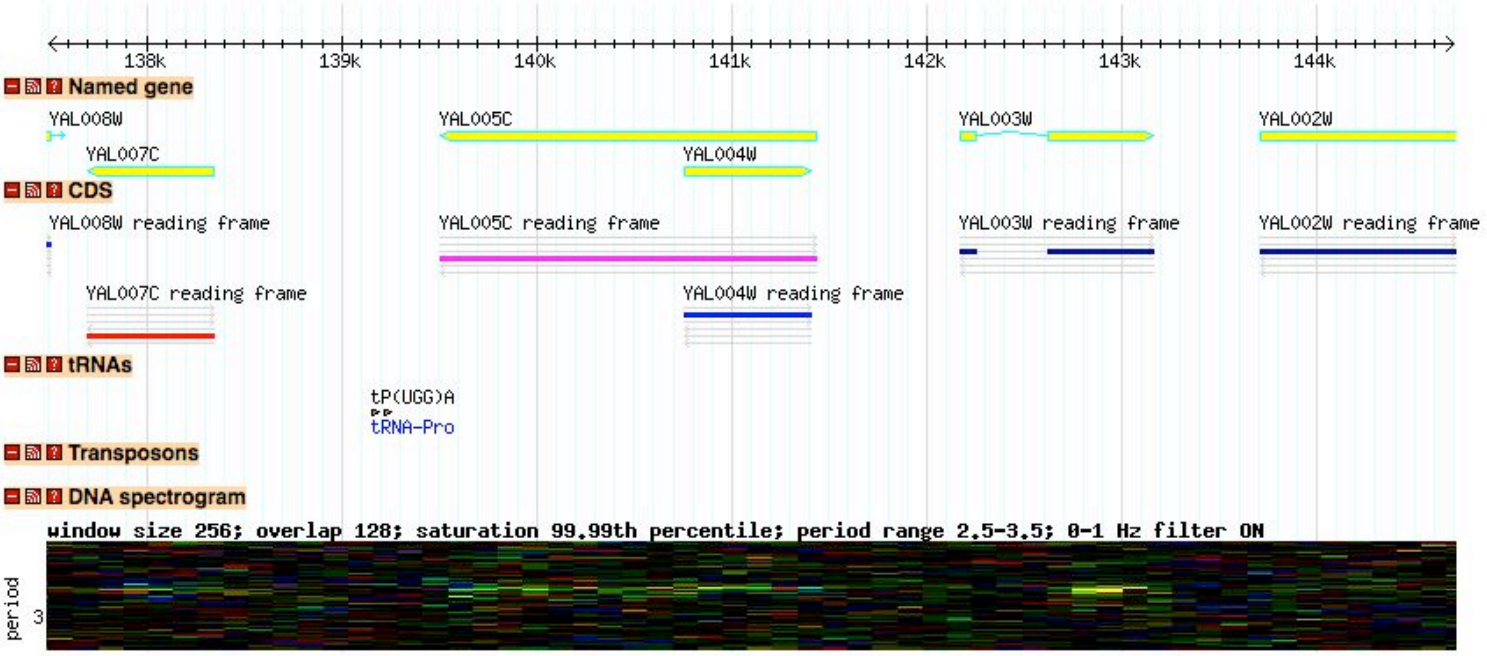
The Generic Genome Browser





Dynamic and Configurable User Interface





Flexible Plugin Architecture

A few words about GMOD

Why GMOD/GBrowse?

Transparent, open source, collaborative development

Decoupled from underlying data sources; portable, configurable, understandable

Interoperability

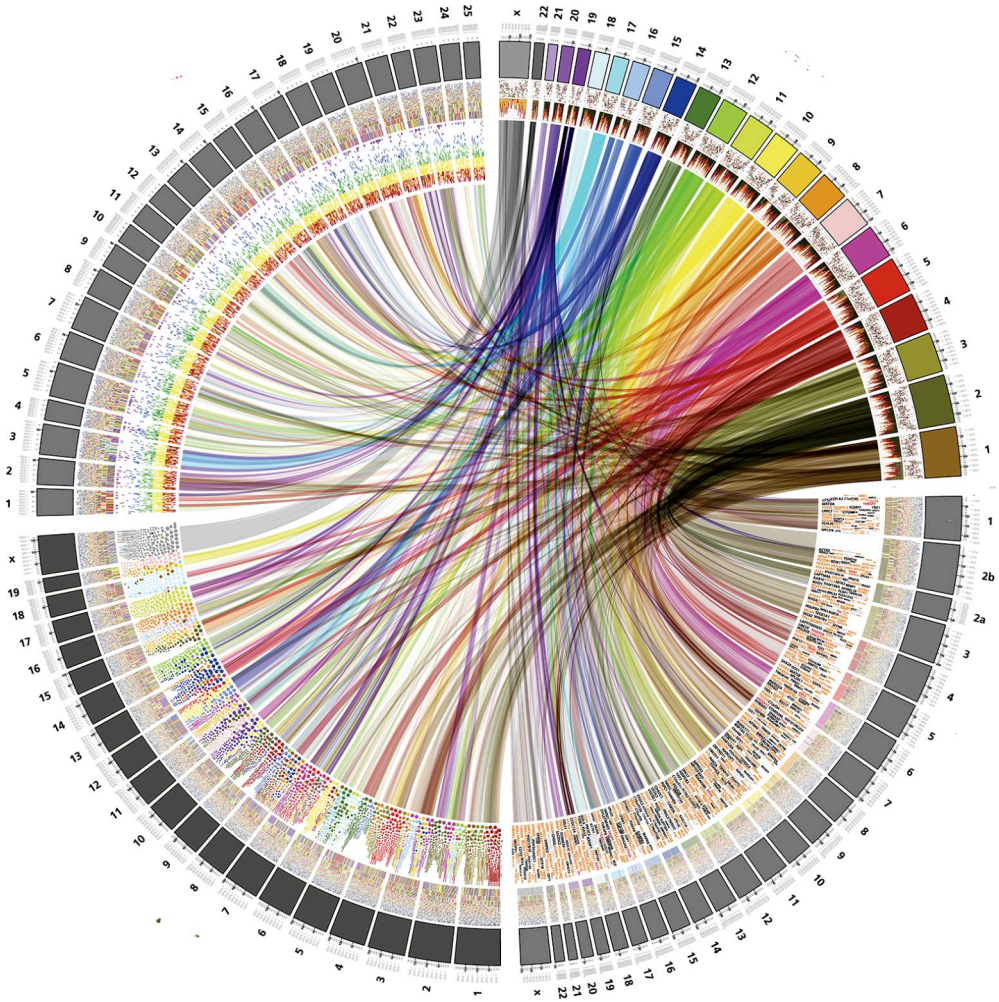
Outreach and training

Free tech support; mailing lists; help-desk

Large and enthusiastic user community (GBrowse is installed on top of hundreds of genome databases, including all major MODS).

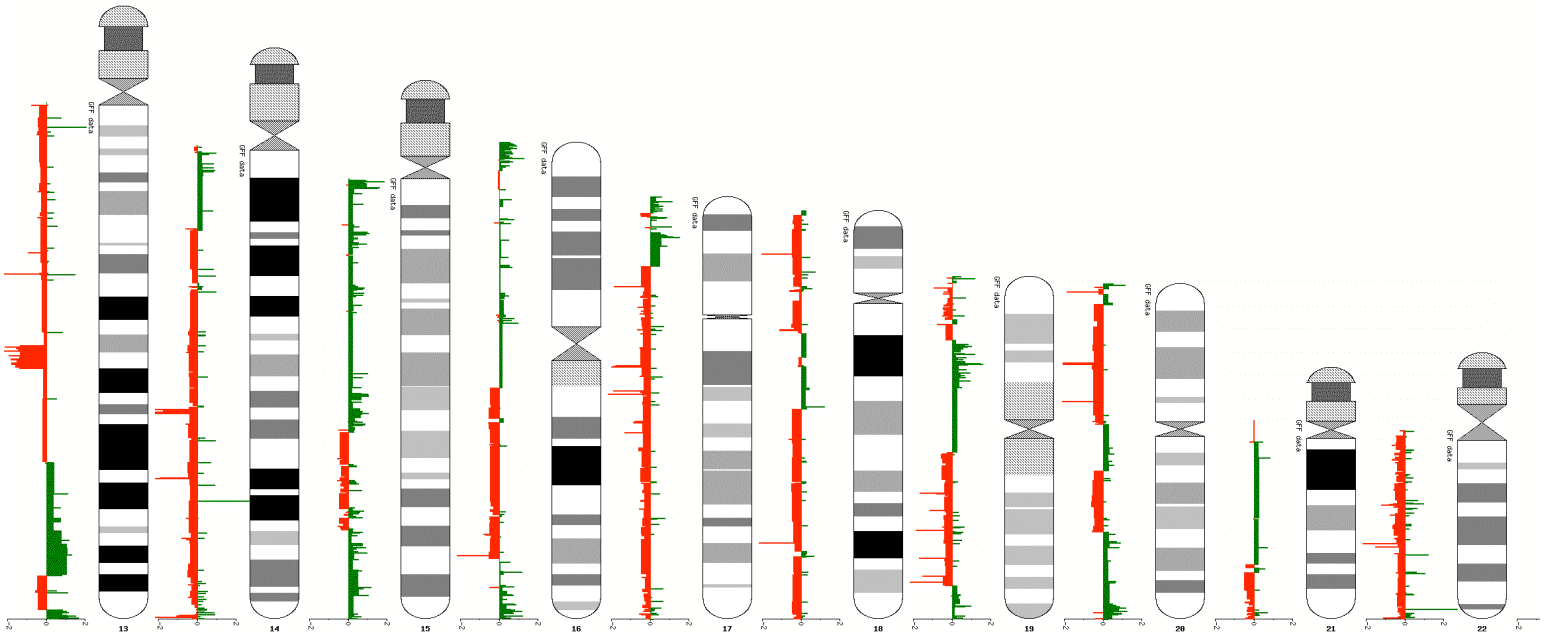
Whole Genome Browsing

CIRCOS



http://mkweb.bcgsc.ca/circos/

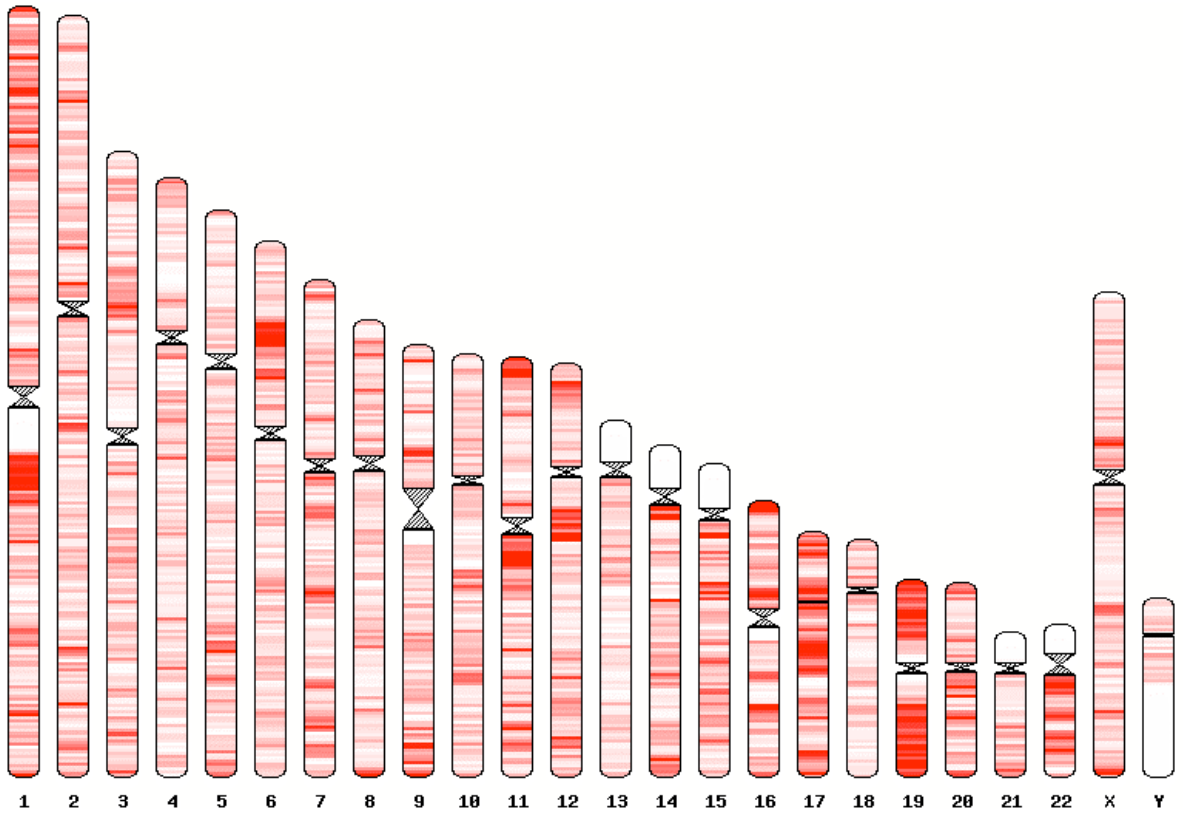
GBrowse_karyotype



CNV data for human chromosomes 13-22

http://gmod.org/GBrowse_karyotype

GBrowse_karyotype



Ensembl gene density plotted on the human karyotype

Synteny Browsers

What is a Synteny Browser?

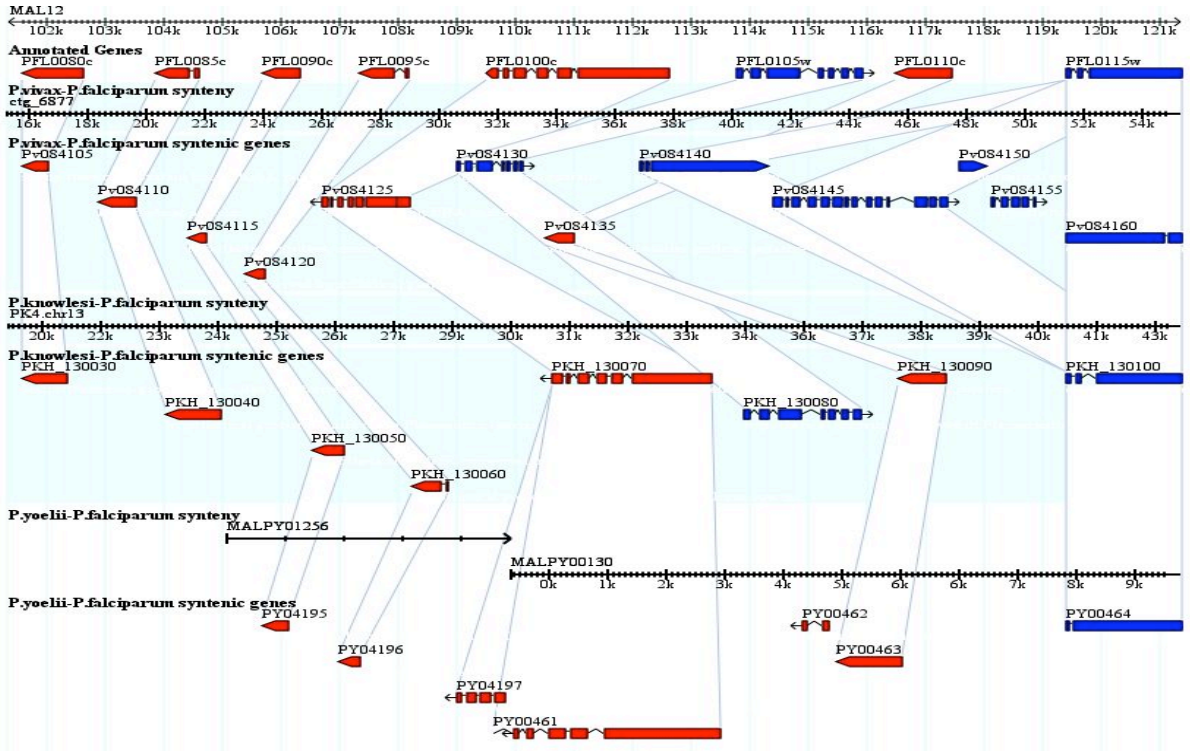
- Has display elements in common with genome browsers

-Uses sequence alignments, orthology or co-linearity data, to highlight different genomes, strains, etc.

-Usually displays co-linearity relative to a reference genome.

SynView

A Simple Approach to Visualizing Comparative Genome Data



CryptoDB Cryptosporidium Genome Resource

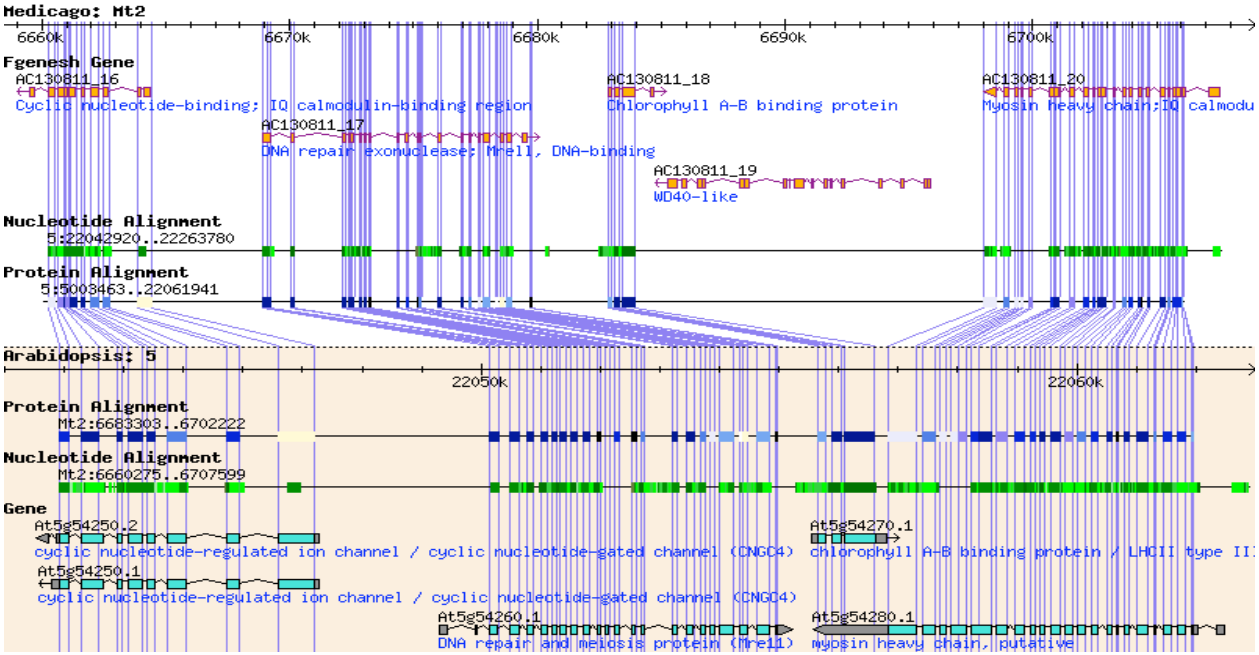
PlasmoDB Plasmodium Genome Resource

PopGenIE Populus Genome Integrative Explorer

Wang H, Su Y, Mackey AJ, Kraemer ET and JC Kissinger . SynView: a GBrowse-compatible approach to visualizing comparative genome data Bioinformatics 2006 22:2308-2309

SynBrowse

...A Synteny Browser for Comparative Sequence Analysis

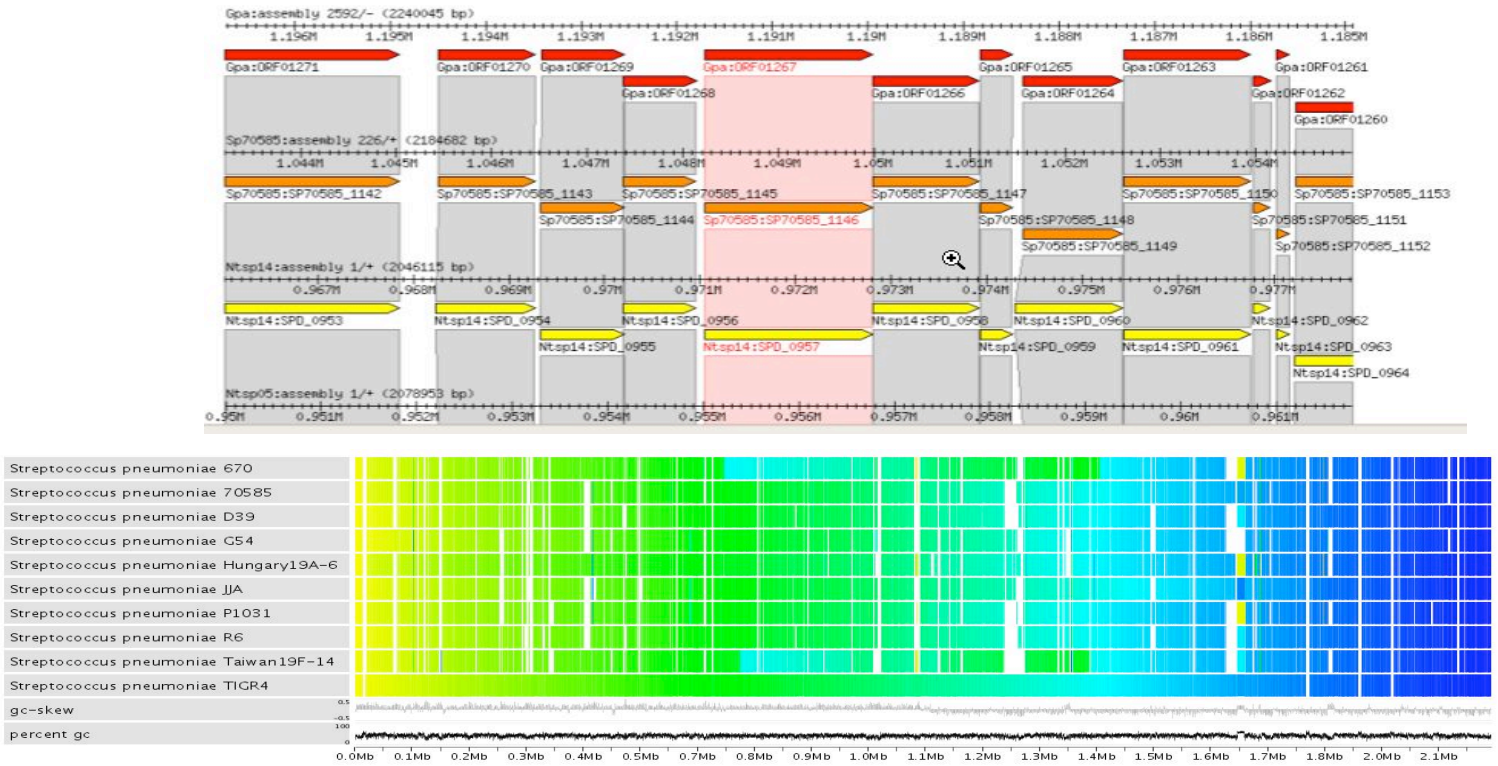


PopGenIE

Populus Genome Integrative Explorer

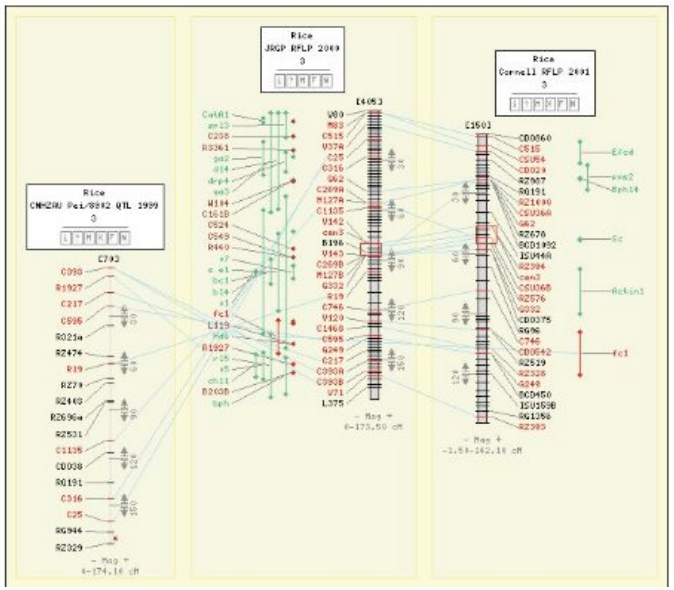
Pan, X., Stein, L. and Brendel, V. 2005. SynBrowse: a Synteny Browser for Comparative Sequence Analysis. Bioinformatics 21: 3461-3468

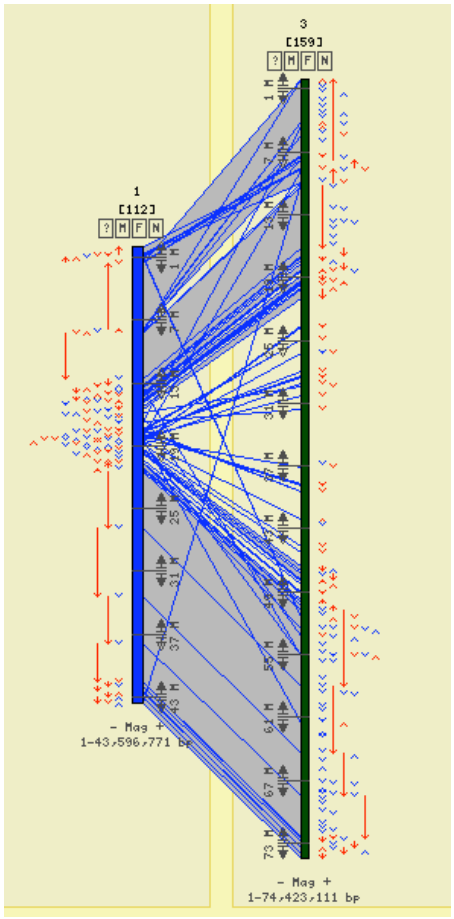
Sybil: Web-based software for comparative genomics



J. Craig Venter™
INSTITUTE

Crabtree, J., Angiuoli, S. V., Wortman, J. R., White, O. R. Sybil: methods and software for multiple genome comparison and visualization Methods Mol Biol. 2007 Jan 01; 408: 93-108.





CMap 2.0

GRAMENE

The Legume Information System

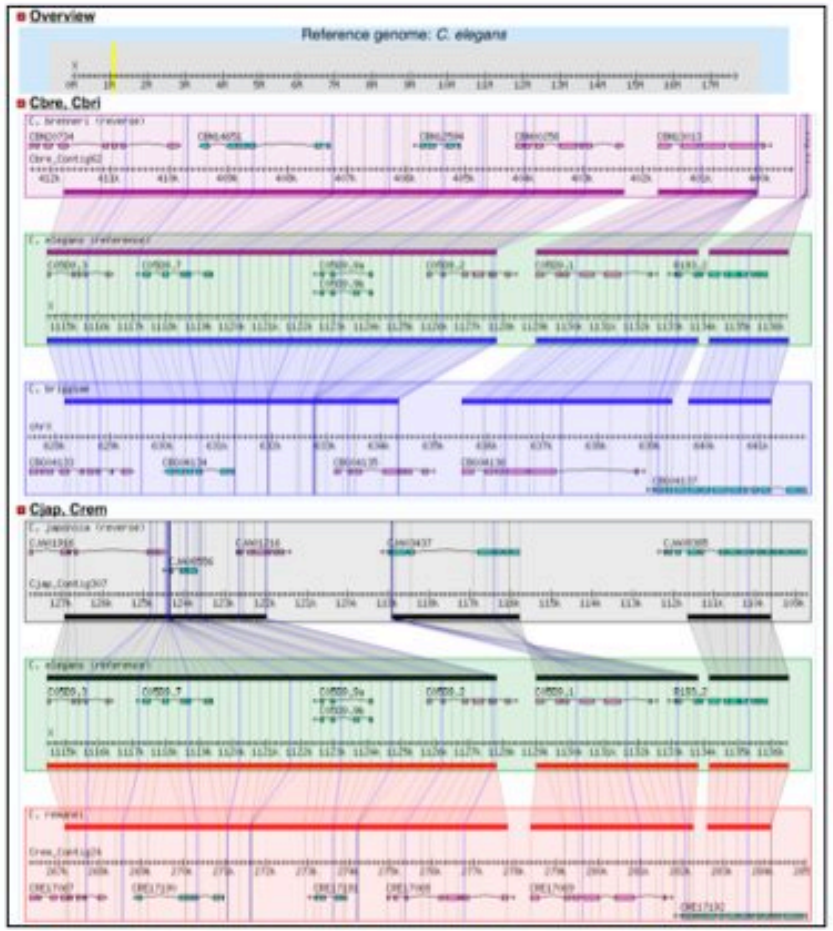
GDR | Genome Database for Rosaceae

+ others...

Youens-Clark K, Faga B, Yap IV, Stein LD, Ware, D. 2009.
CMap 1.01: A comparative mapping application for the Internet. doi:10.1093

GBrowse_syn

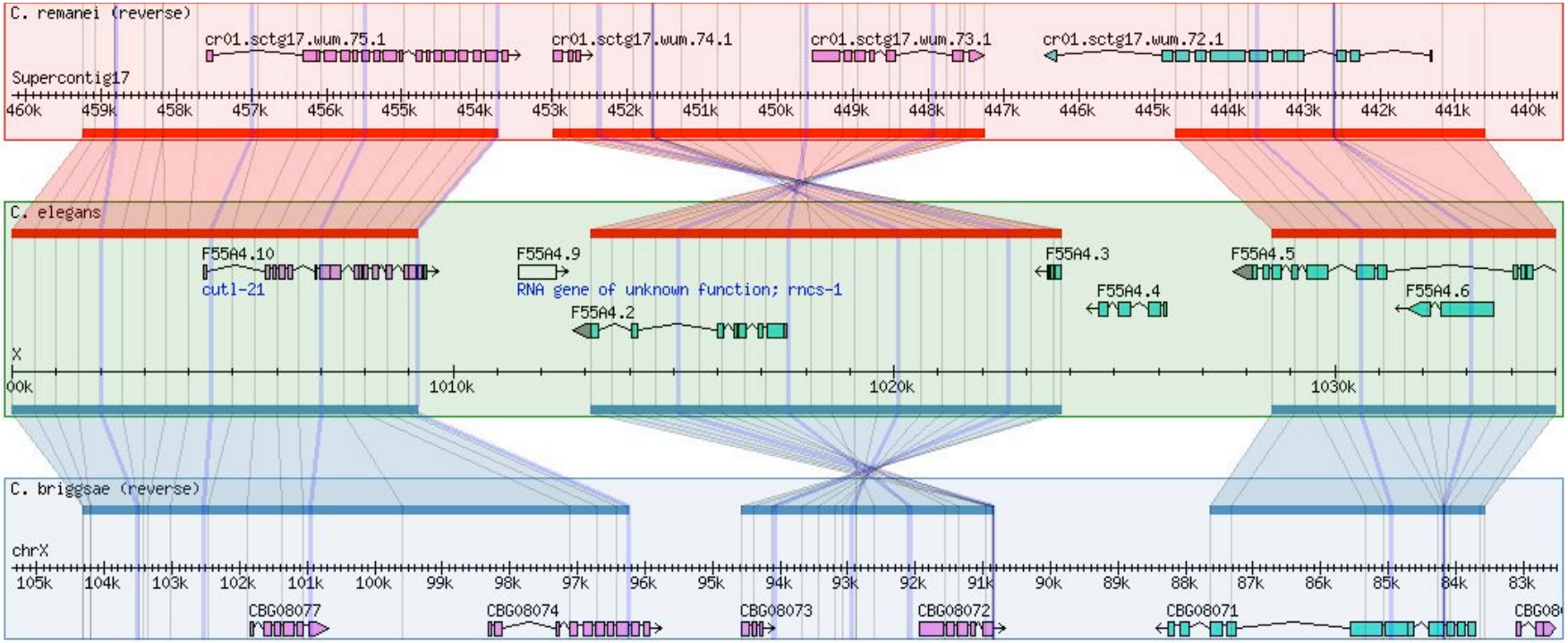
- GBrowse based comparative genomics viewer
- Shows a reference sequence compared to 2 or more others
- Can also show any GBrowse-based annotations



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

Sheldon McKay, Cold Spring Harbor Laboratory

GBrowse_syn



WormBase

tair

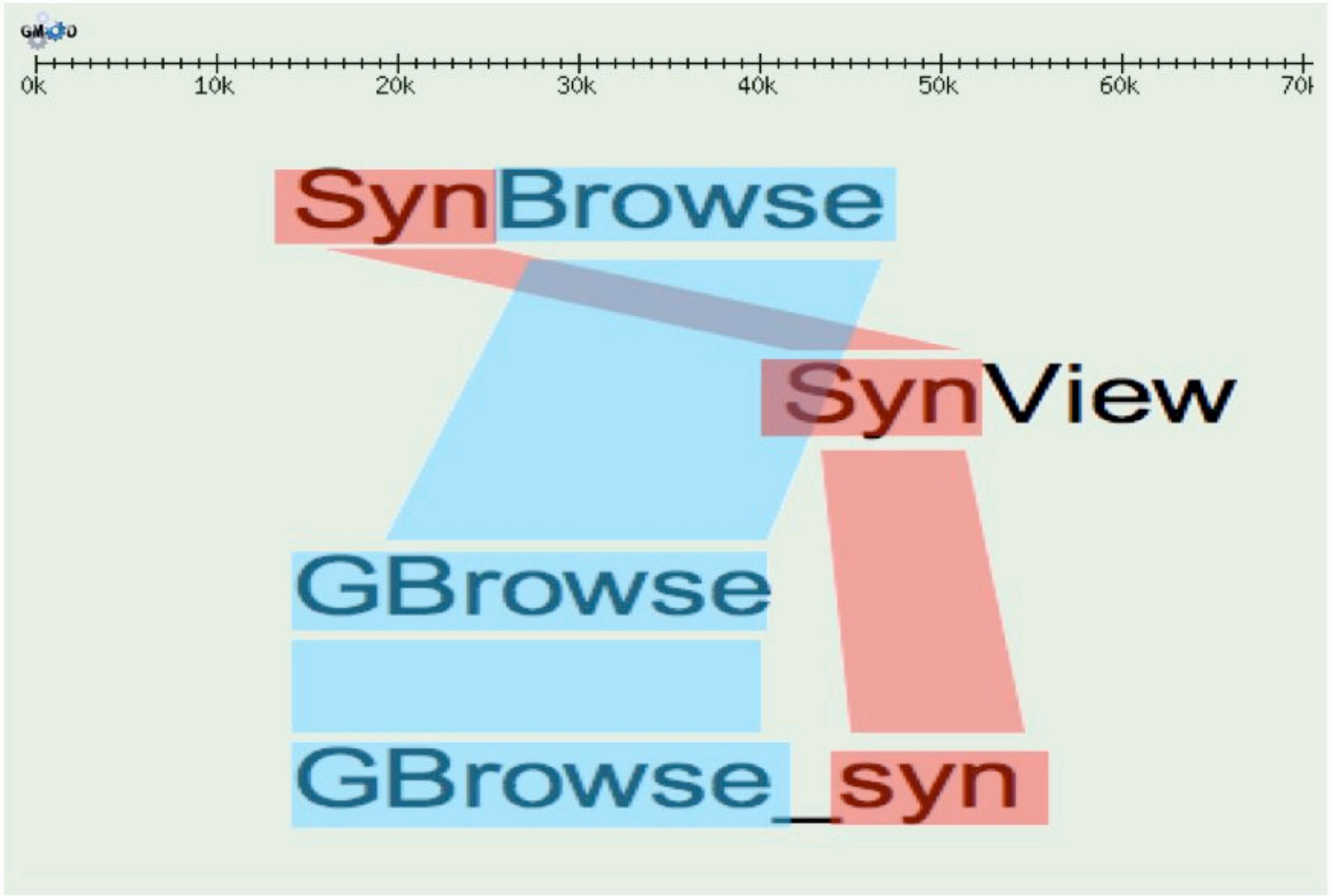
mod ENCODE

+others…

Pseudomonas Genome Database V2

Branding ideas...

GMOD Browser branding/nomenclature issues…



SynView:
- Add-on to native GBrowse package
- Uses GFF3 or DAS1 compliant data adapters
- GFF requires special tags (allowed in spec.)
- Reference panel on top

SynBrowse:
- Uses same core libraries as Gbrowse
- Uses GFF database adapter
- GFF2 uses standard 'Target' syntax
- Currently only supports two species
- Central reference panel?

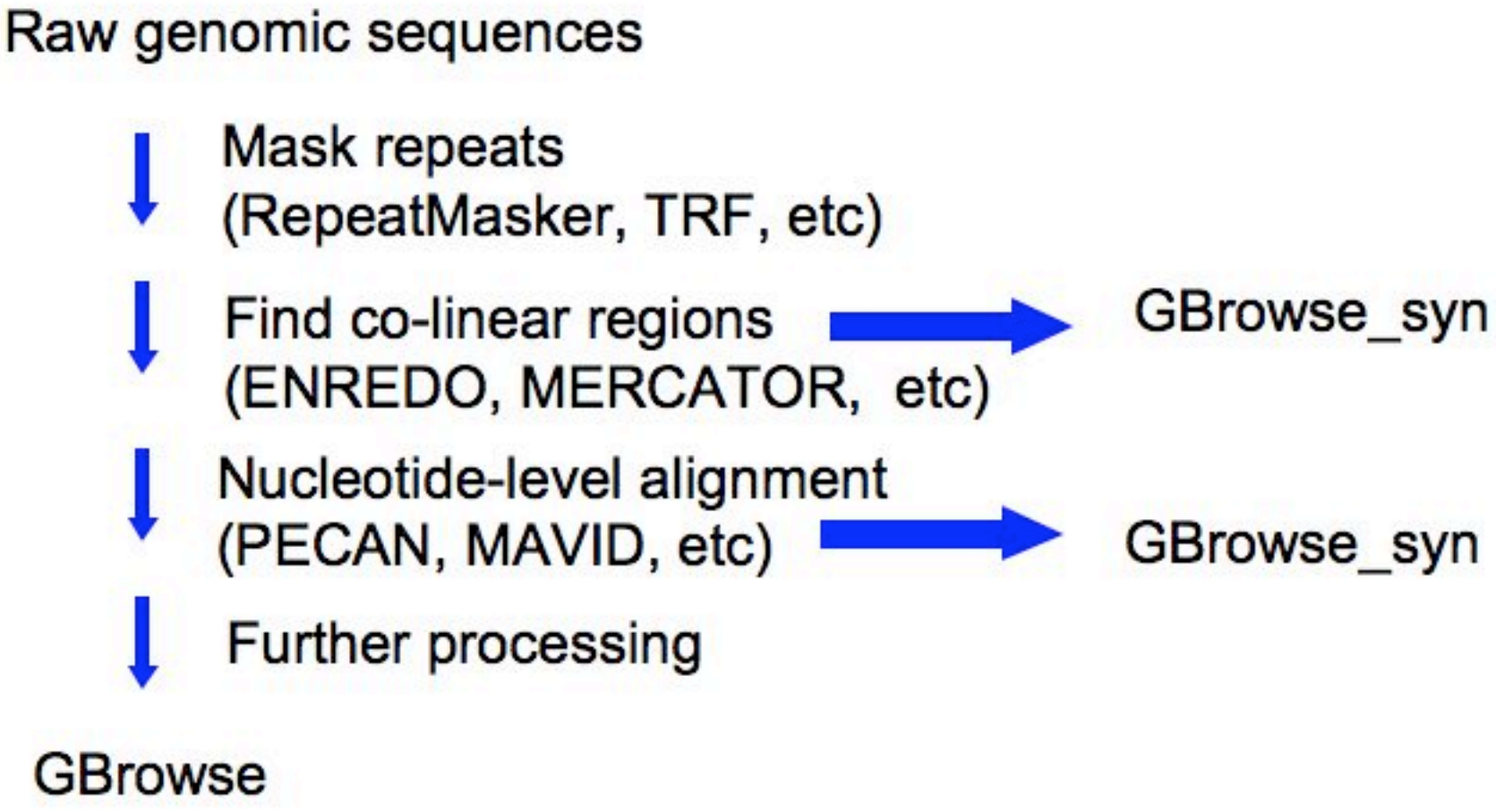
Sybil:
- Not GBrowse-based
- Uses chado database
- Whole genome and detailed views

GBrowse_syn:
- Part of GBrowse distribution
- Uses native GFF2/3 or chado adapters for species' data
- Synteny data are stored in a separate joining database

How is GBrowse_syn different?

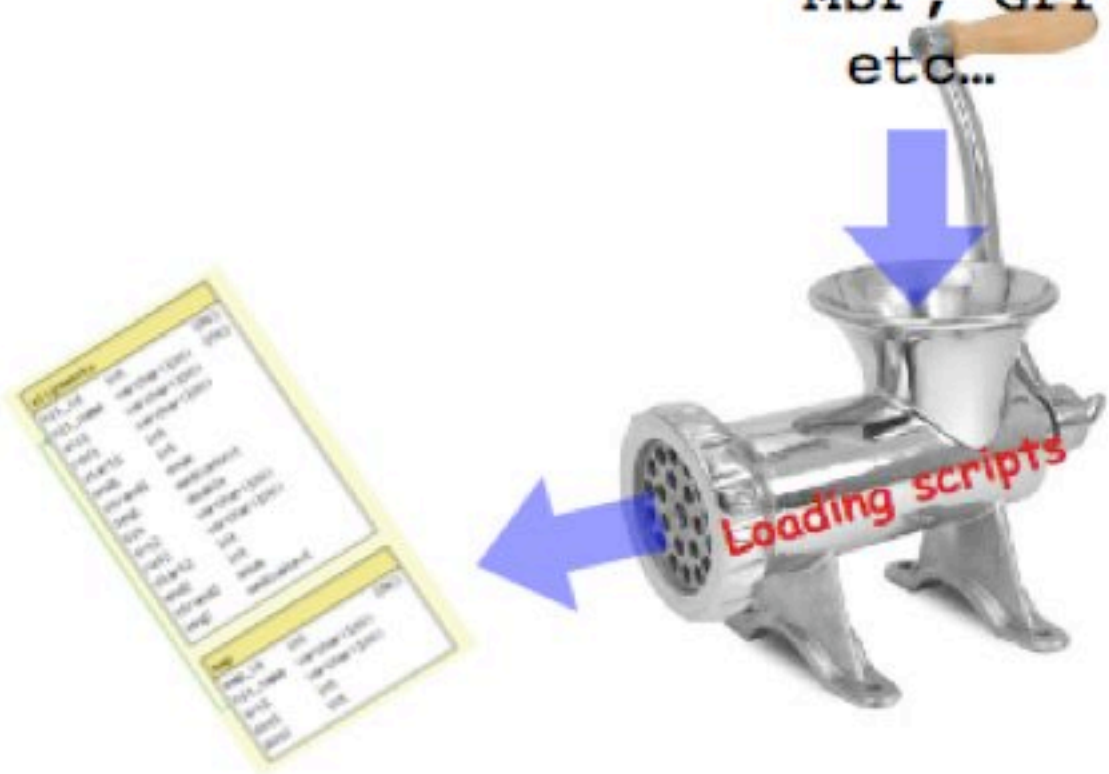
- Does not rely on perfect co-linearity across the entire displayed region (no orphan alignments)
- Offers on the fly alignment chaining
- No upward limit on the number of species
- Used grid lines to trace fine-scale sequence gain/loss
- Seamless integration with GBrowse data sources
- Ongoing support and development
- Some people think it looks nice

Hierarchical Genome Alignment Strategy



Interoperability

CLUSTALW, *ad hoc* tab-delimited
PECAN, STOCKHOLM,
MSF, GFF3,
etc…



Problem : How to use Insertions/Deletion data

A

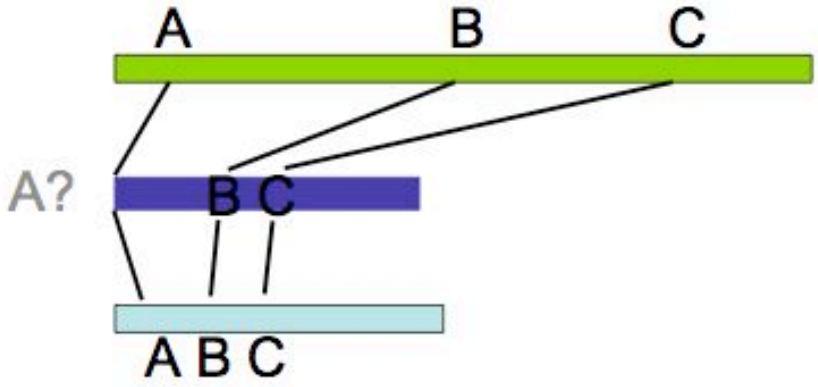
```
Ce-CHROMOSOME_I(+)/5195-16585 TGGCAAAAATATTTTGCATTTGCCGTTTTTCCCGTTTGCCGAAAAGTCTAATTTCGGTAA
Cb-chrI(-)/4091935-4097143    ------------------------------------------------------------
Cr-Contig8(+)/571990-577344   TTCGAAAC---------------------------------------------------
```

B

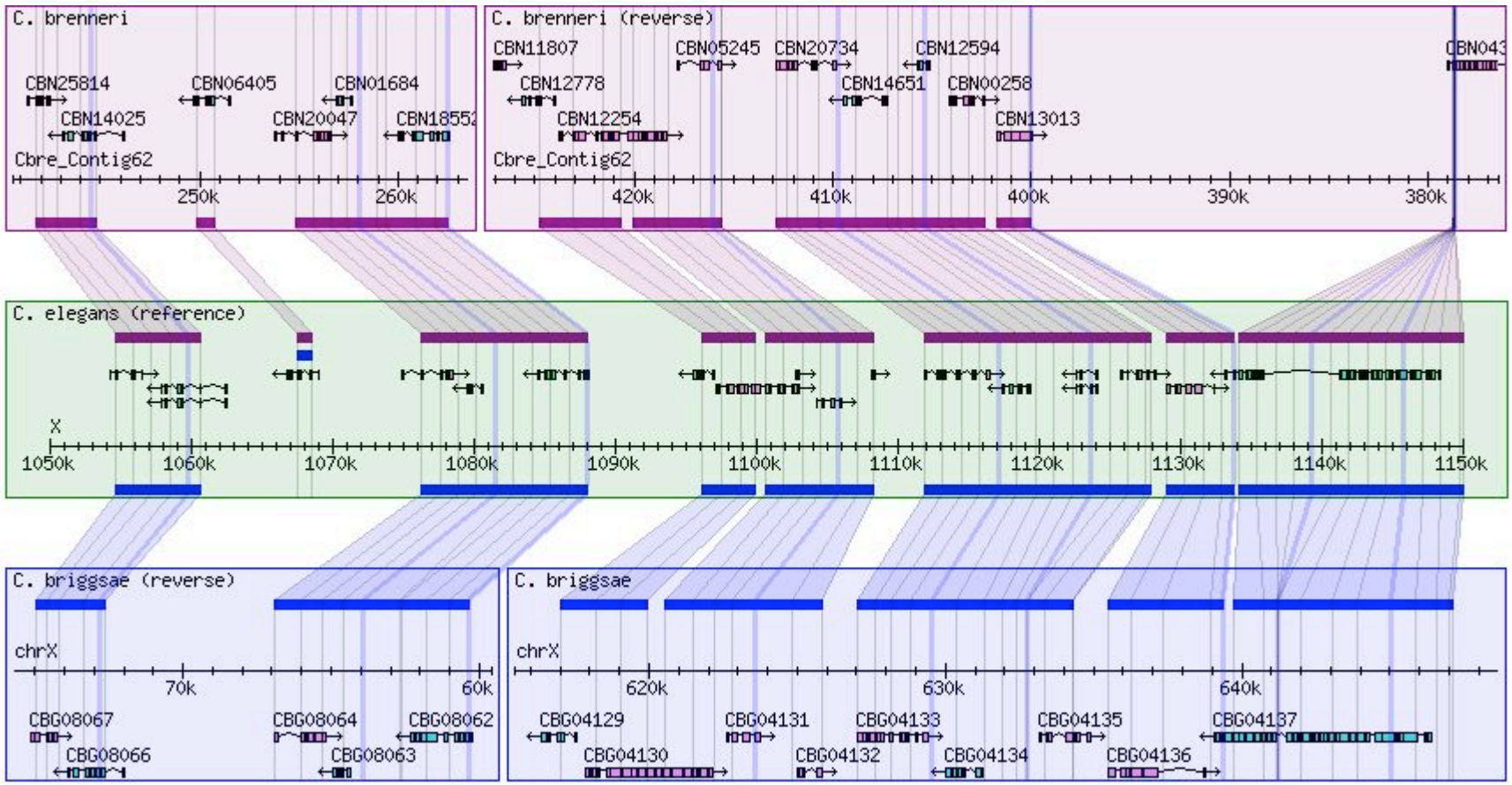
```
Ce-CHROMOSOME_I(+)/5195-16585 TTGGGCCATTTTTCGAAATTTTGAGCCACATAAAAAACTTTGAACCATTTTTGAGAAGTA
Cb-chrI(-)/4091935-4097143    ----------------------AGAAGAATGTGAAGATCTTCA------------------
Cr-Contig8(+)/571990-577344   ----------------------CAGAGAAACAGAAACAATTTTA------------------
                                                     **   *    ** *  **
```

C

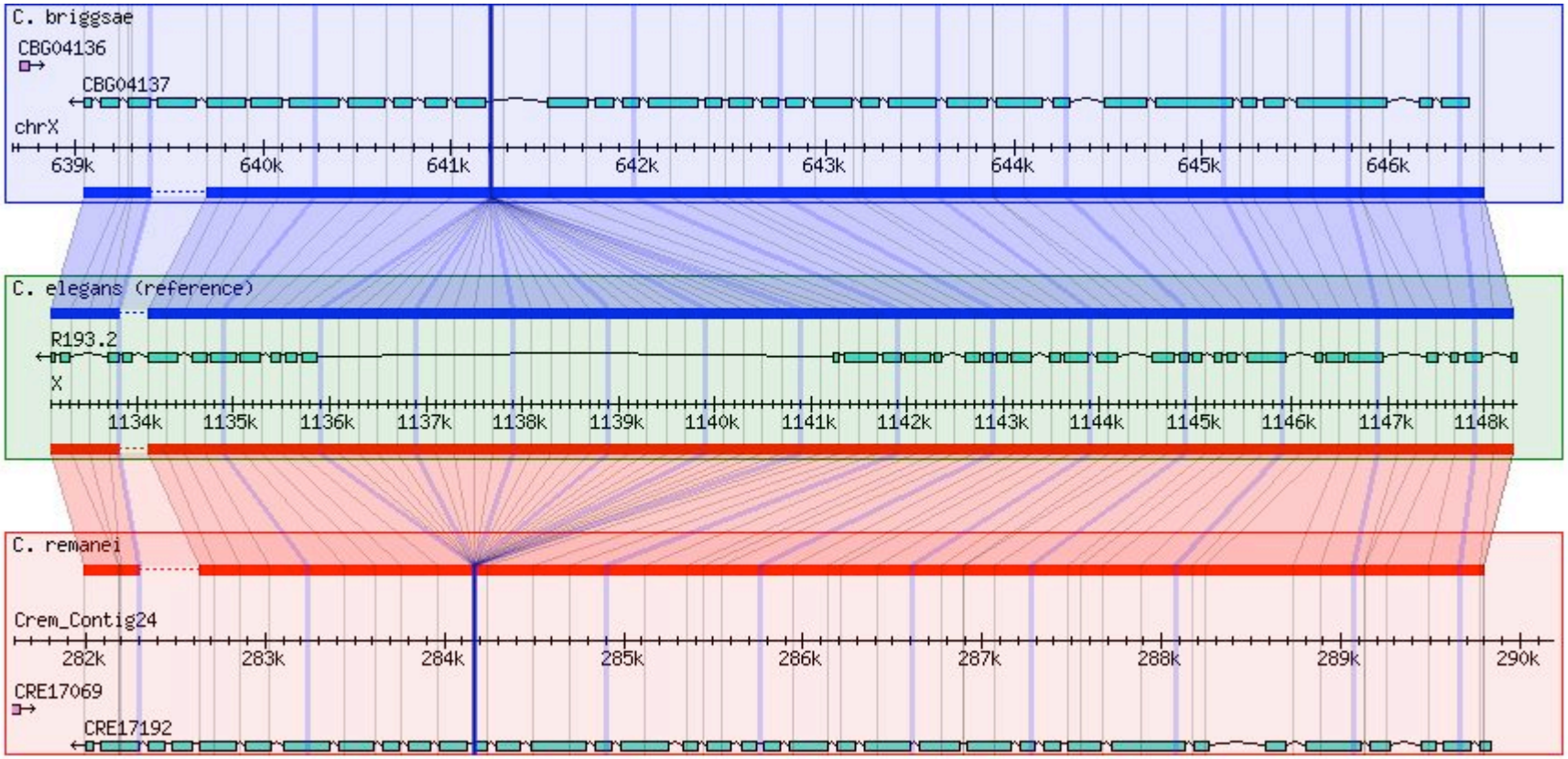
```
Ce-CHROMOSOME_I(+)/5195-16585 TTATTACGACATTCGTTTATTTGAGCACAATTTGGGCCTATACTTTCAAAATCGGGGTTT
Cb-chrI(-)/4091935-4097143    --TTCATGTCAA-----------------------------------------------TCAT
Cr-Contig8(+)/571990-577344   --TTTCTGAAAACAGGTAGTATTATGGTTCCGAGGGTGTAGGGTTCGAAACCGGGCCTAG
                                 *   *  *
```

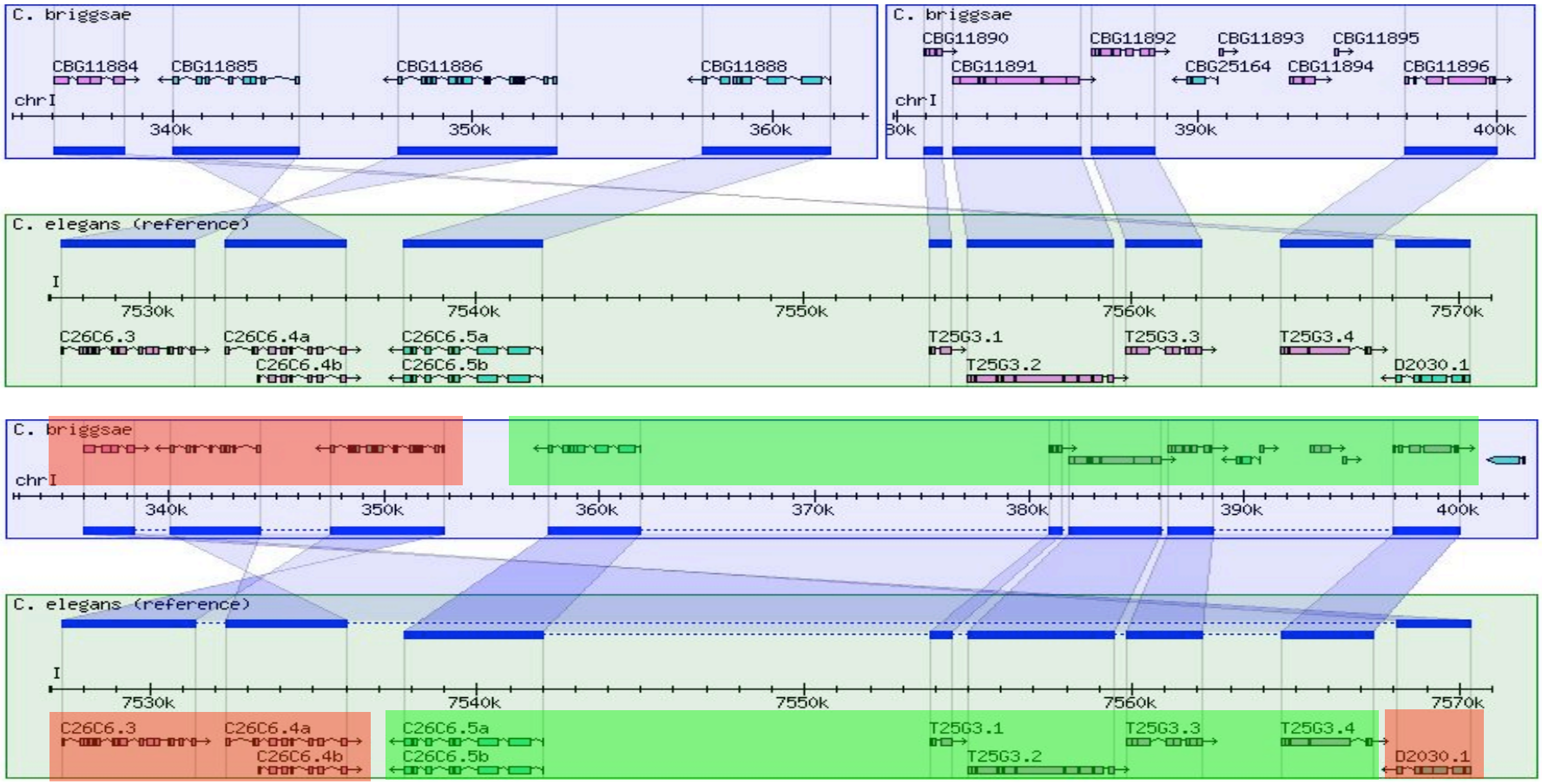


Tracking Indels with grid lines

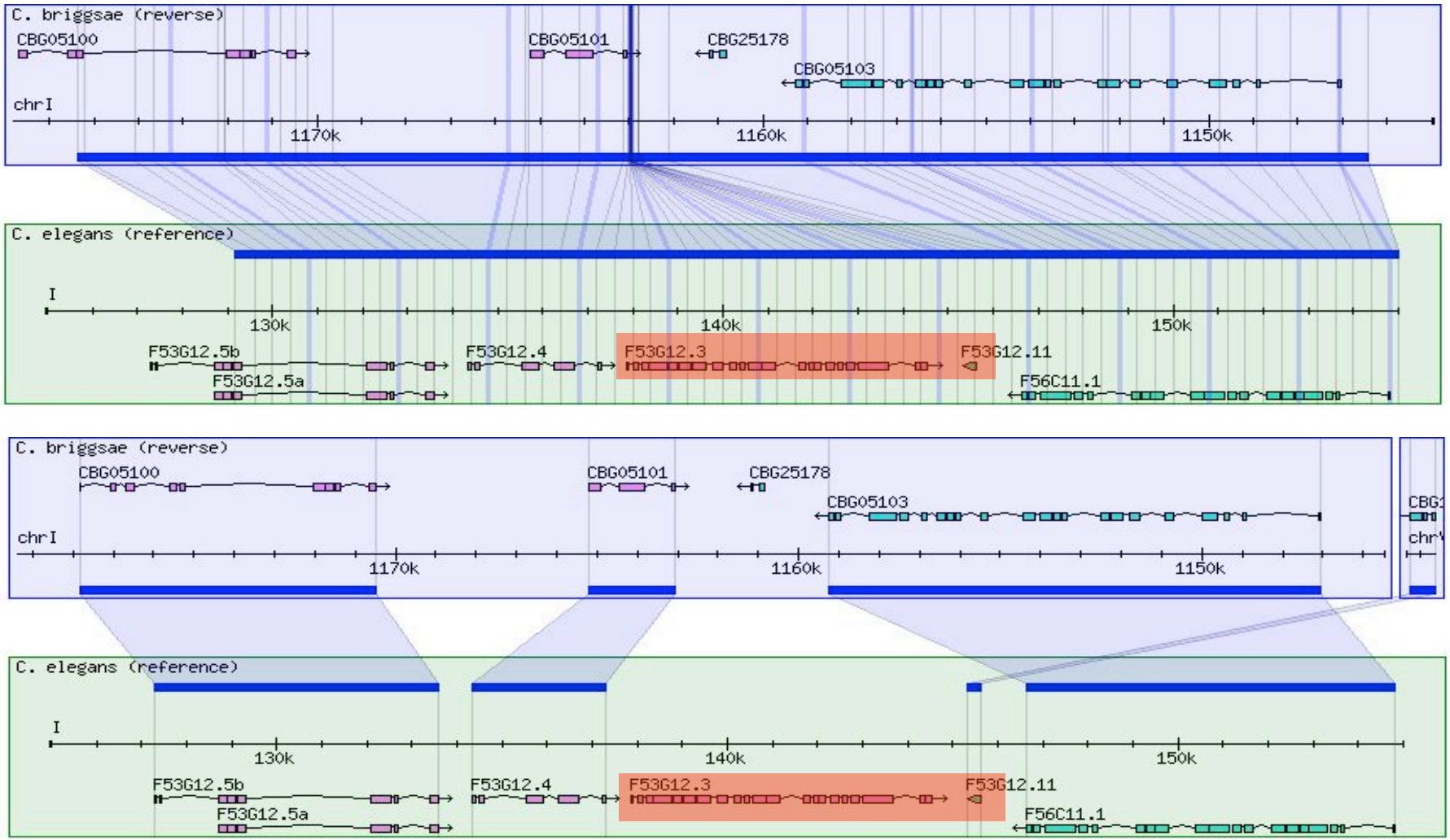


Evolution of Gene Structure





Alignment chaining. GBrowse_syn provides on-board functionality to "chain" alignments together if they are co-linear, in the same orientation and have monotonically increasing coordinates. This is sometimes helpful to visualize higher order chromosome rearrangements. In this example, chaining the alignments (lower panel) helps to visualize a possible model where an inversion affecting the genes highlighted in red was followed by a nested insertion of the block of genes highlighted in green.

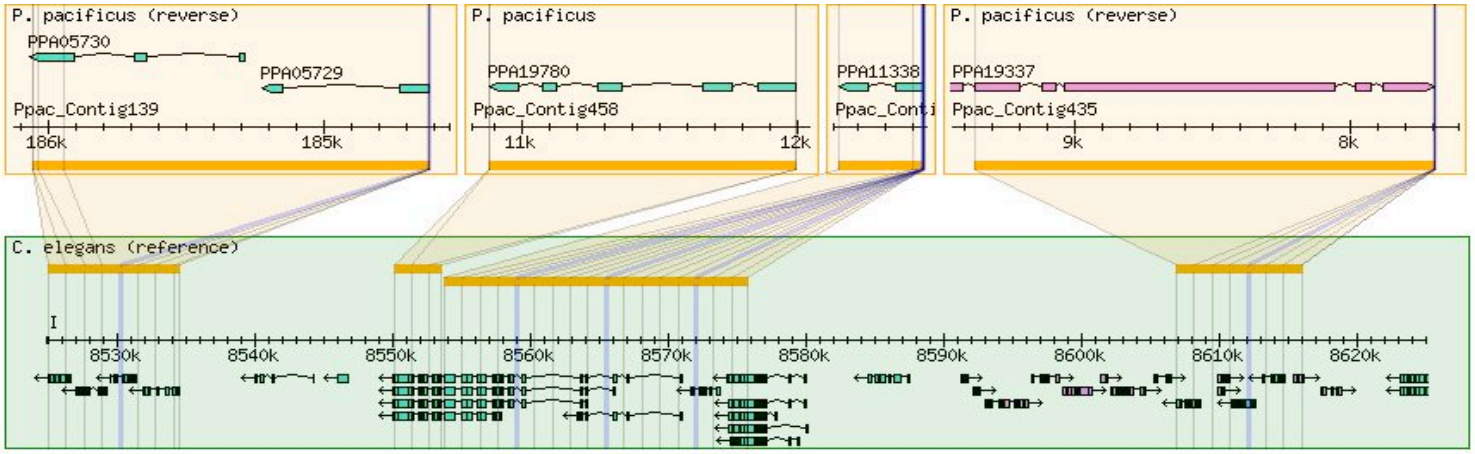


Gene loss. A portion of the *C. elegans* and *C. briggsae* genomes from WormBase. The top view shows DNA sequence alignment data. Grid-lines indicate the relative coordinates in the two sequence. Smaller and larger spaces indicate gaps or insertions relative to the reference sequence, respectively. There is a large gap in the *C. briggsae* chromosome sequence that affects the genes highlighted in red. Independent orthology data (shown in the lower panel) are consistent with a translocation of the small gene and a complete loss of the larger gene in *C. briggsae*.

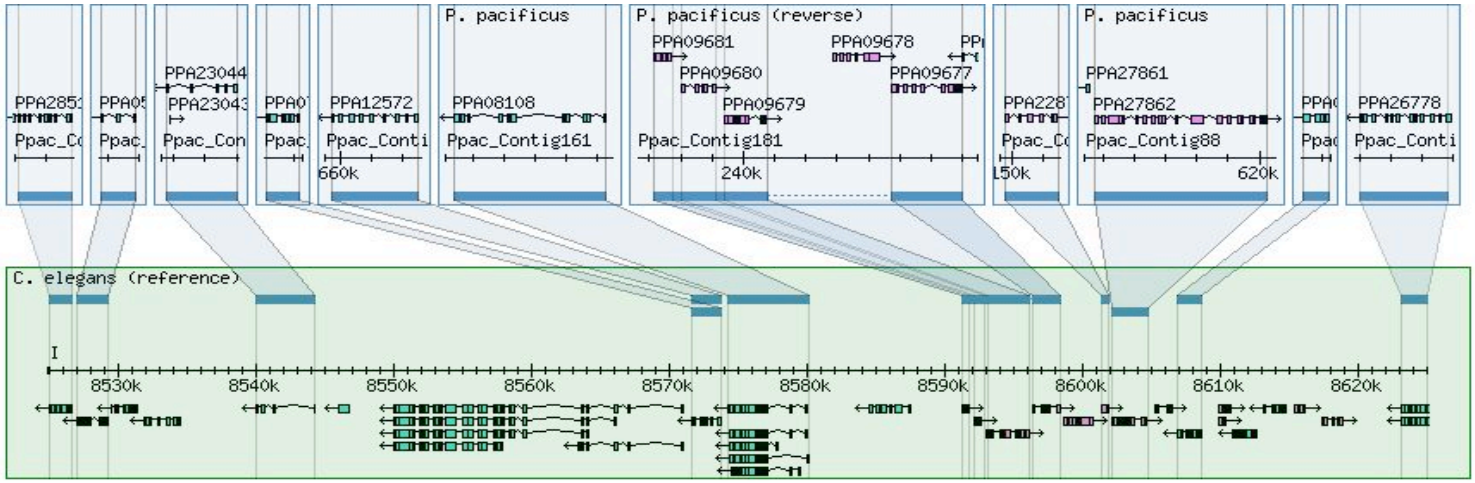
What if the aligned DNA sequences are too distant?

!=

Pecan alignments



Protein orthology based Synteny blocks



GBrowse_syn or Sybil or SynView?

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

SynView
Scalable
Familiar interface

Sybil
Whole genome and
other unique visualizations
Unfamiliar interface

GBrowse_syn Future Work

- Integration with GBrowse 2
- "On the fly" sequence alignment view
- High-level graphical overview

- AJAX based user interface and navigation.
  - Submitting grant next week proposing implementing a JBrowse based synteny browser

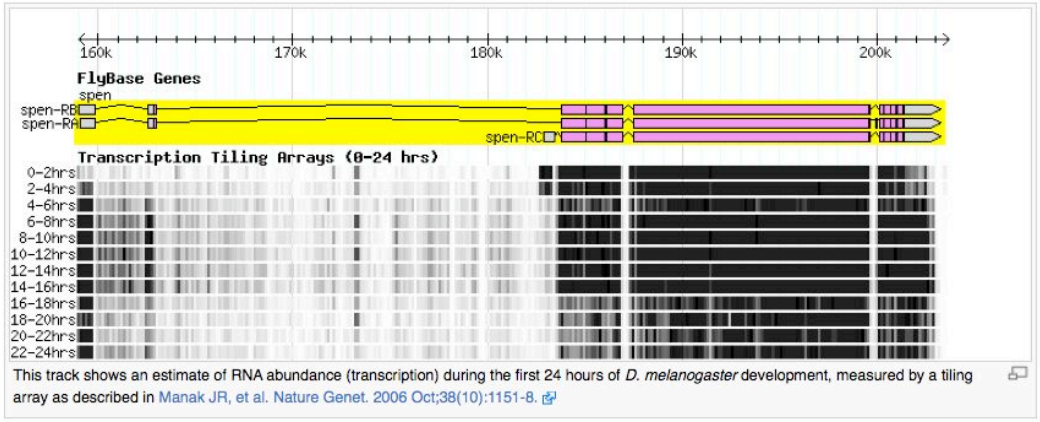
GBrowse_syn Resources

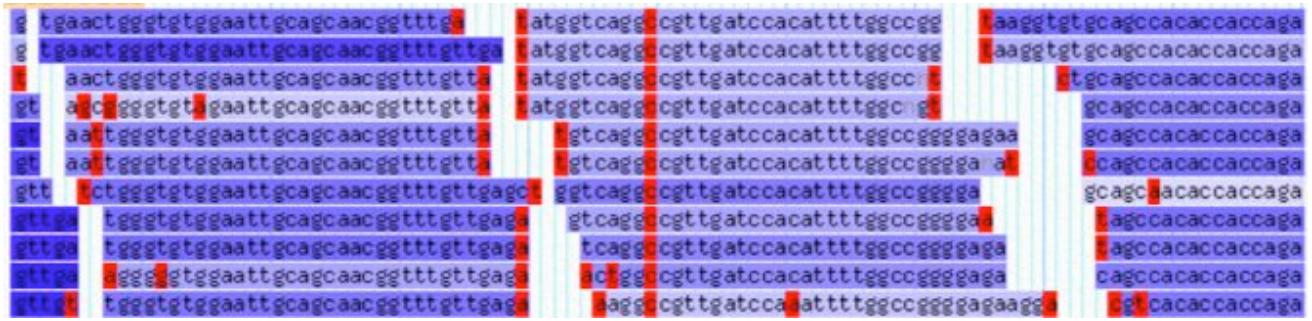
Home Page	http://gmod.org/wiki/GBrowse_syn
Tutorial	http://gmod.org/wiki/GBrowse_syn_Tutorial
User Help	http://gmod.org/wiki/GBrowse_syn_Help
Configuration	http://gmod.org/wiki/GBrowse_syn_Configuration
Example	http://www.wormbase.org/cgi-bin/gbrowse_syn/
Mailing List	https://lists.sourceforge.net/lists/listinfo/gmod-gbrowse

High Density Data

Dealing with very dense data

- Microarrays
- Next-gen Sequencing





- Wiggle
  - Large amounts of scored data with genomic coordinates
  - Too many table rows for a relational database
  - Solution is a hybrid database/serialized data approach

WIG is a format specification introduced by the UCSC Genome Browser and also adopted by GBrowse

1) The WIG file is converted to a query-optimized binary file
2) A pointer to the binary file is stored in the database
3) An external adapter queries the binary file

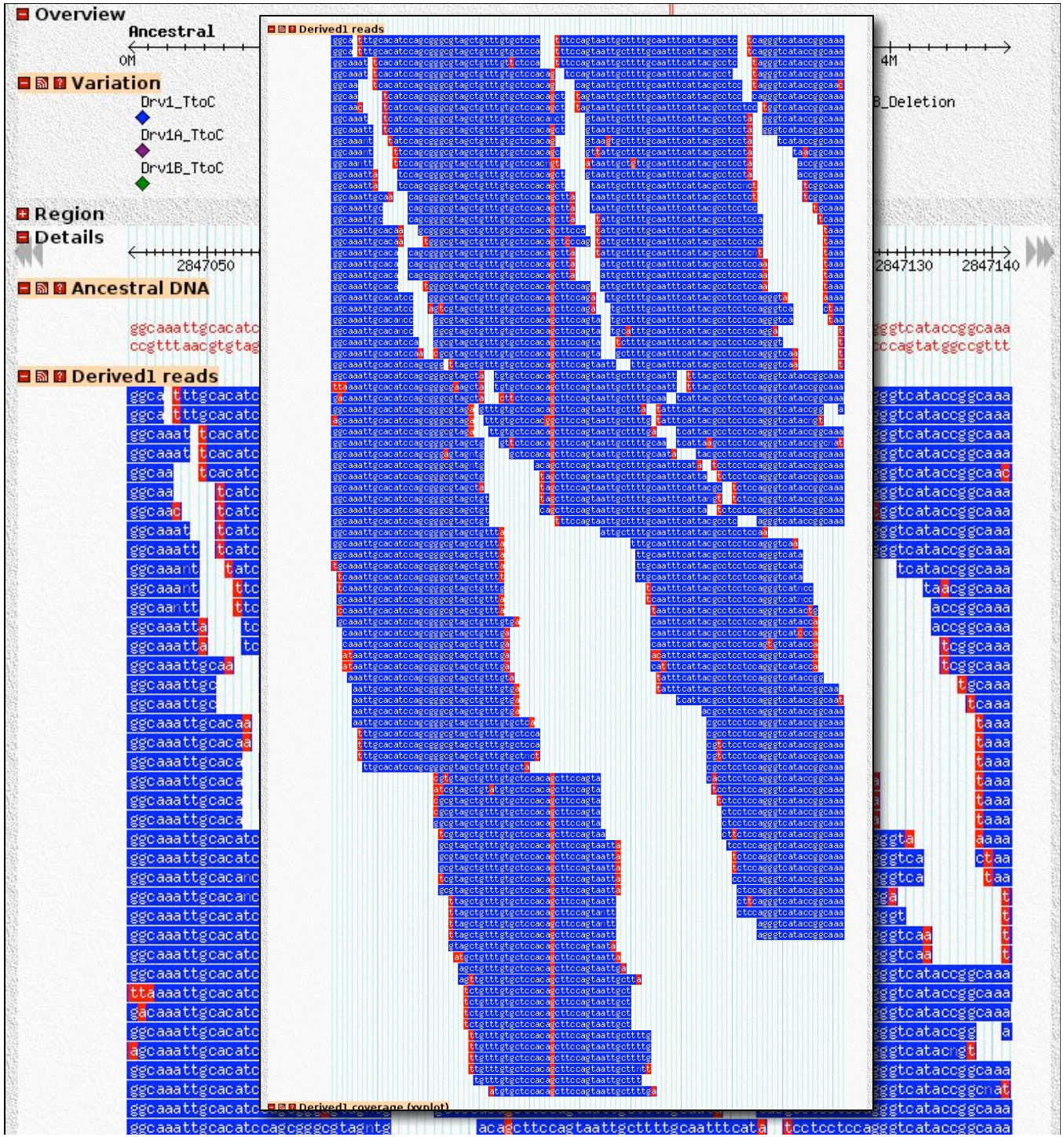
http://genome.ucsc.edu/goldenPath/help/wiggle.html
http://gmod.org/wiki/GBrowse/Uploading_Wiggle_Tracks

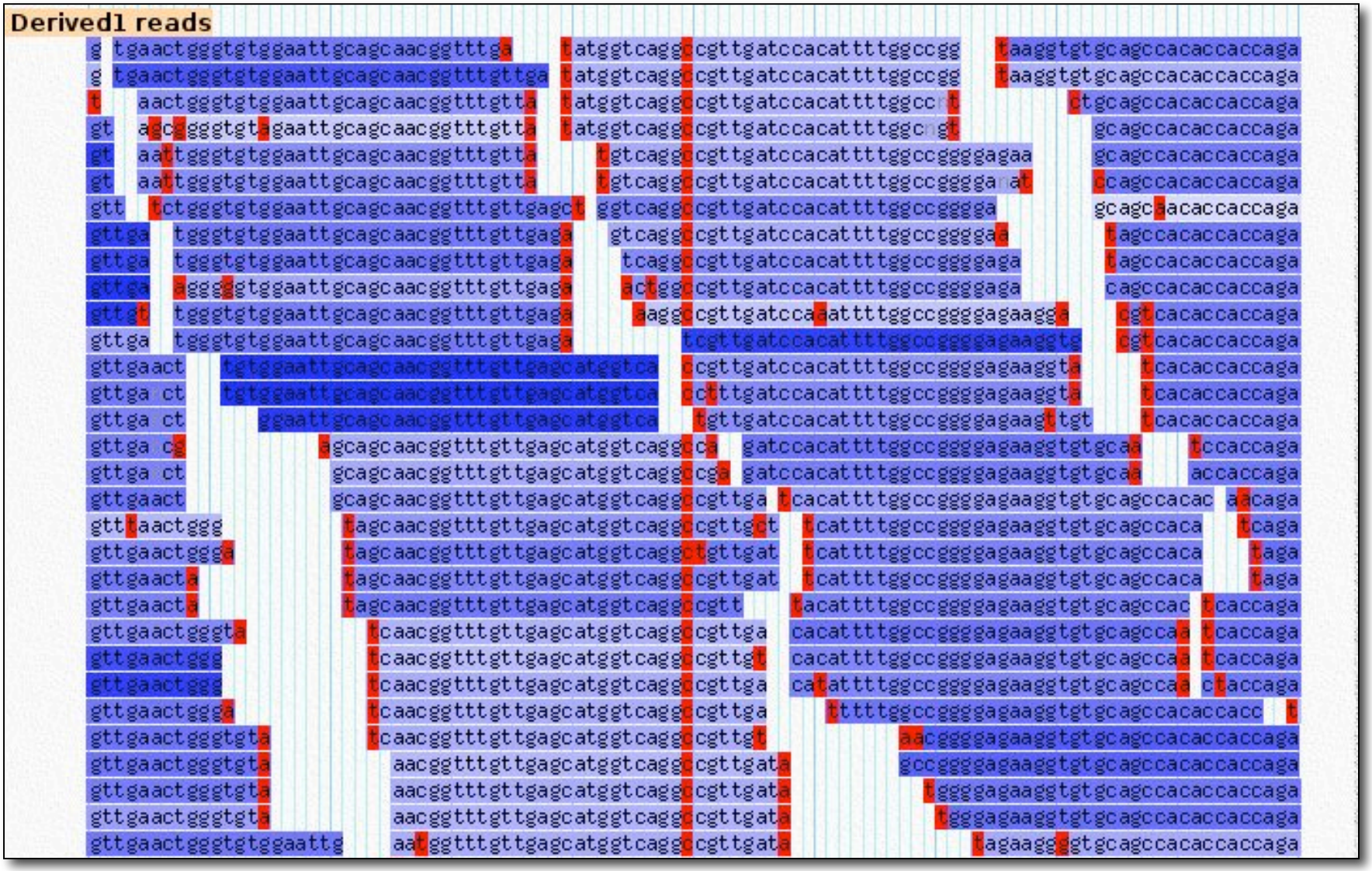
- SAM/BAM (Sequence Alignment/Map)
  - NGS data generates huge numbers of aligned reads
  - The SAM specification allows efficient storage of read alignments against reference sequences
  - BAM is a highly efficient, compressed binary version of SAM
  - The SAMTools package provides utilities for handling the alignment data.
  - Third party implementers are starting to support SAM/BAM, for example Bio::DB::SAM/GBrowse

http://samtools.sourceforge.net/

High magnification view: 100bp

Uses GBrowse 2 (Beta) and the Bio::DB::Sam GBrowse database adaptor (Alpha), and SAMtools



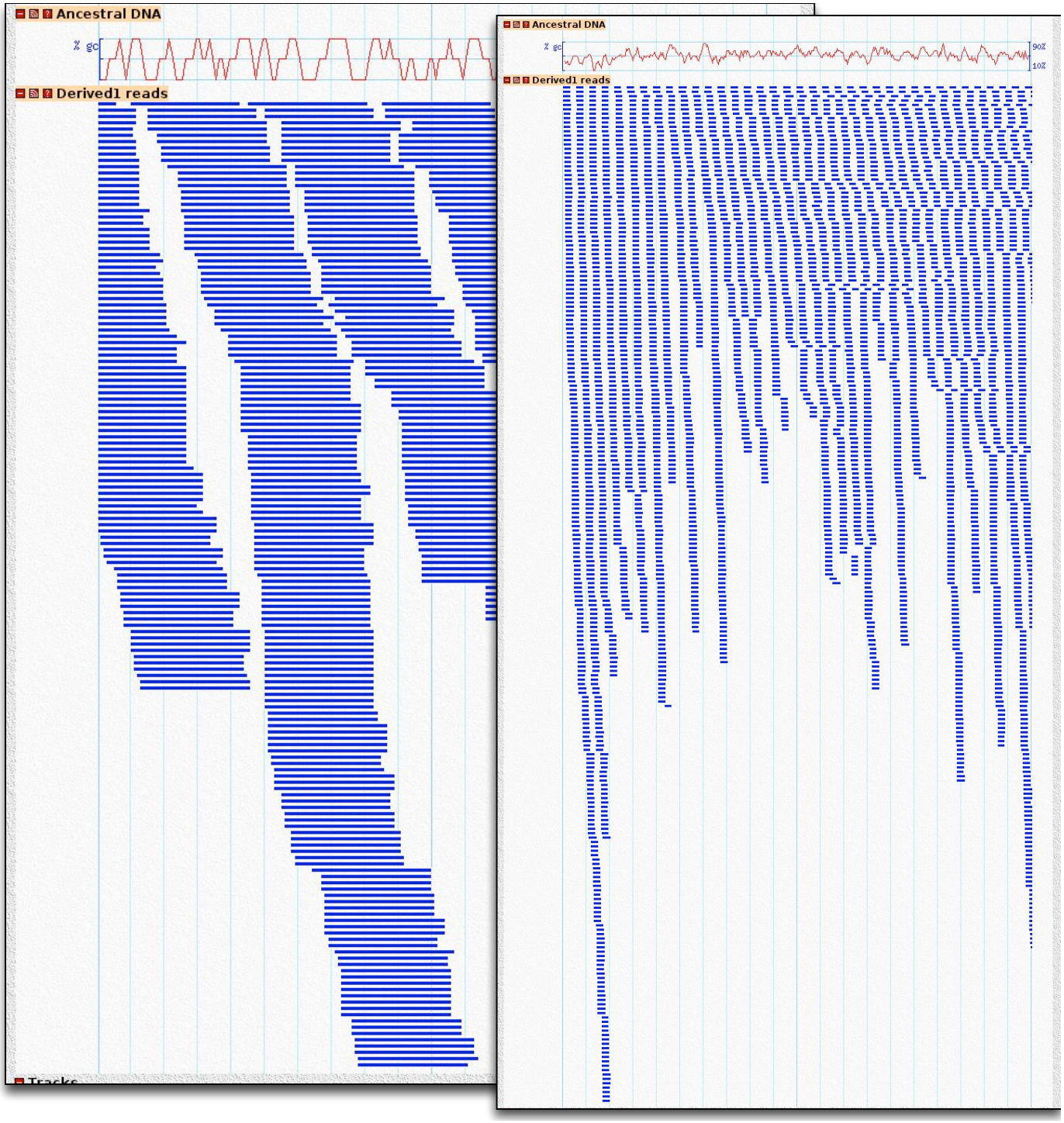


Anything in SAM format is accessible.

As you zoom out to 200bp you lose letters.

As you zoom out to 2000bp the view becomes much less useful.

SAMtools, GBrowse 2, & Bio::DB::Sam adaptor make this volume of data computationally tractable

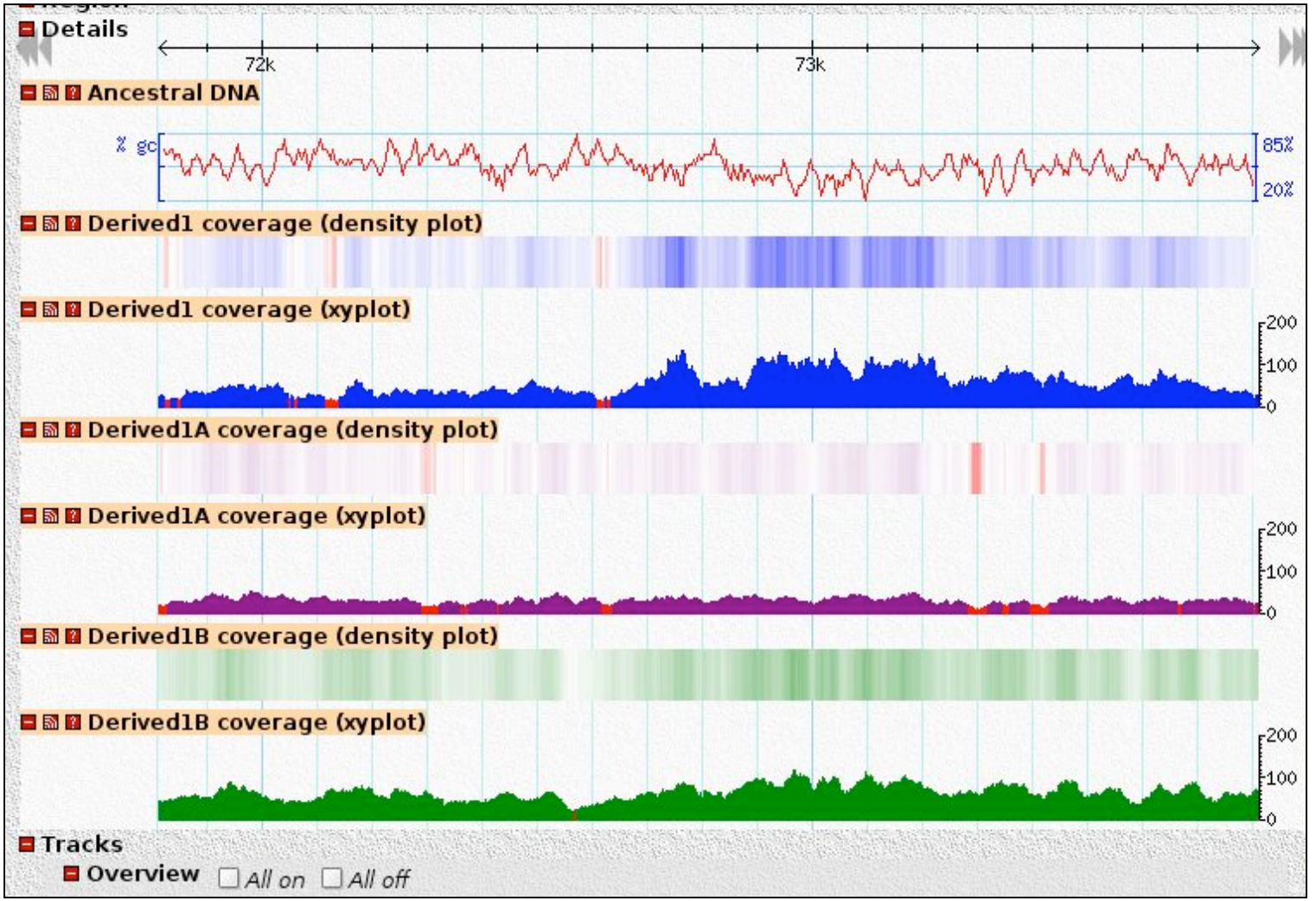


GBrowse as an Alignment Viewer

Summarize!

SAMtools and GBrowse help here

SAMtools can summarize values from the data.



Acknowledgments

- GMOD
- iPlant Collaborative
- NESCent
- TAIR
- WormBase
- ModENCODE