



# 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

The screenshot displays the UCSC Genome Browser interface for Human Mar. 2006 Assembly (hg18). The top navigation bar includes links for Home, Genomes, Blat, Tables, Gene Sorter, PCR, DNA, Convert, Ensembl, NCBI, PDF/PS, Session, and Help. The main title is "UCSC Genome Browser on Human Mar. 2006 Assembly (hg18)". Below the title, there are navigation controls for moving and zooming, and a search bar showing the current position: chrX:112,882,150-112,996,149. A chromosome map shows the current location on chromosome X (q23). The main track area displays various genomic features: RefSeq Genes (UCSC Genes Based on RefSeq, UniProt, GenBank, CCDS and Comparative Genomics), Mammal Cons Multiz Align, SNPs (130), Affy SNP 6.0, Affy SNP 6.0 SV, Illumina 1M-Duo, and RepeatMasker. Below the tracks, there are controls for moving start and end positions, zooming, and track management. A section titled "Mapping and Sequencing Tracks" lists various tracks with dropdown menus to show or hide them.

Home Genomes Blat Tables Gene Sorter PCR DNA Convert Ensembl NCBI PDF/PS Session Help

## UCSC Genome Browser on Human Mar. 2006 Assembly (hg18)

move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x

position/search chrX:112,882,150-112,996,149 jump clear size 114,000 bp. configure

chrX (q23) 22.2 q21.1 22.3 q24 Xq25 Xq28

Scale 50 kb

chrX: 112900000 112950000

UCSC Genes Based on RefSeq, UniProt, GenBank, CCDS and Comparative Genomics RefSeq Genes

RefSeq Genes

Mammal Cons Multiz Align

SNPs (130)

Affy SNP 6.0

Affy SNP 6.0 SV

Illumina 1M-Duo

RepeatMasker

Vertebrate Multiz Alignment & Conservation (44 Species)

Simple Nucleotide Polymorphisms (dbSNP build 138)

SNP Genotyping Arrays

Repeating Elements by RepeatMasker

move start Click on a feature for details. Click or drag in the base position track to zoom in. Click gray/blue bars on left for track options and descriptions. move end

< 2.0 >

default tracks hide all add custom tracks configure reverse refresh

Use drop-down controls below and press refresh to alter tracks displayed. Tracks with lots of items will automatically be displayed in more compact modes.

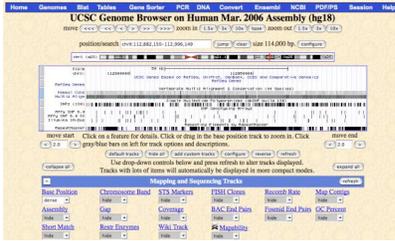
collapse all expand all

### Mapping and Sequencing Tracks

refresh

<a href="#">Base Position</a>	<a href="#">Chromosome Band</a>	<a href="#">STS Markers</a>	<a href="#">FISH Clones</a>	<a href="#">Recomb Rate</a>	<a href="#">Map Contigs</a>
dense ▾	hide ▾	hide ▾	hide ▾	hide ▾	hide ▾
<a href="#">Assembly</a>	<a href="#">Gap</a>	<a href="#">Coverage</a>	<a href="#">BAC End Pairs</a>	<a href="#">Fosmid End Pairs</a>	<a href="#">GC Percent</a>
hide ▾	hide ▾	hide ▾	hide ▾	hide ▾	hide ▾
<a href="#">Short Match</a>	<a href="#">Restr Enzymes</a>	<a href="#">Wiki Track</a>	<a href="#">Mapability</a>		
hide ▾	hide ▾	hide ▾	hide ▾		

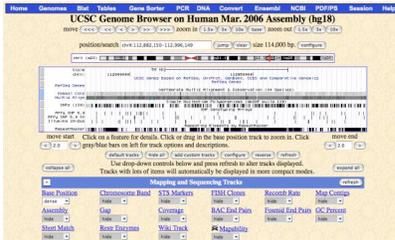
<http://genome.ucsc.edu>



# 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

The screenshot displays the Ensembl Genome Browser interface for Human (GRCh37). The main header shows the Ensembl logo and navigation links: Home > Human [GRCh37], Login / Register, BLAST/BLAT, BioMart, Docs & FAQs, and Mirrors. The current location is specified as 6:133,017,695-133,161,157.

**Location-based displays**

- Whole genome
- Chromosome summary
- Region overview
- Region in detail**
- Comparative Genomics
  - Alignments (image) (5)
  - Alignments (text) (51)
  - Multi-species view (47)
  - Synteny (12)
- Genetic Variation
  - Resequencing (2)
  - Linkage Data
- Markers
- Other genome browsers
  - UCSC
  - NCBI

**Configure this page**

- Manage your data
- Export data
- Bookmark this page

**Chromosome 6: 133,017,695-133,161,157**

Assembly exception... chromosome 6

Assembly exception...  
HSCHR6\_MHC\_AFD  
HSCHR6\_MHC\_COX  
HSCHR6\_MHC\_DBB  
HSCHR6\_MHC\_MANN  
HSCHR6\_MHC\_MCF  
HSCHR6\_MHC\_QBL  
HSCHR6\_MHC\_SSTO

[Export Image](#)

[Region overview](#) **Region in detail** [help](#) [Alignments \(image\)](#)

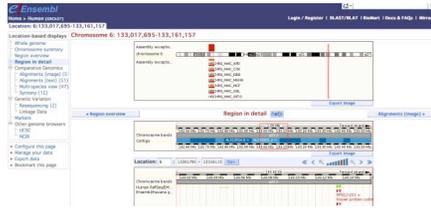
Chromosome bands  
Contigs

Location: 6 : 133017695 - 133161115 [Go>](#)

Chromosome bands  
Human RefSeq/EM...  
Ensembl/Havana g...

RPS12-201 >  
Known protein coding

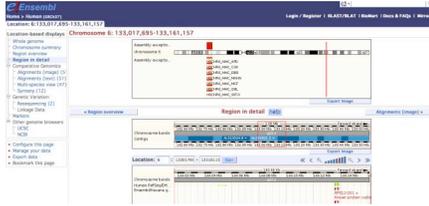
<http://www.ensembl.org>



# 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](http://www.ensembl.org/info/about/ensembl_powered.html)

# October 2009: Ensembl Plants!

Collaboration between EBI and Doreen Ware's group at CSHL

The screenshot shows the Ensembl Plants website interface. At the top left is the logo "e!EnsemblPlants" with a "Home" link below it. On the top right, there is a search bar with a magnifying glass icon and a dropdown menu. Below the logo, navigation links include "Login / Register | BLAST | BioMart | FTP | Docs & FAQs".

The main content area is divided into two columns. The left column features a "Search Ensembl Plants" section with a search bar containing "All species" and a "Go" button. Below the search bar, an example search "e.g. chx28 or Carboxypeptidase" is shown. Underneath is a "Popular genomes" section with a link to "Log in to customize this list". It lists three species: *Arabidopsis thaliana* (TAIR9), *Oryza sativa* (MSU6), and *Sorghum bicolor* (Sbi1), each with a small image icon. Below this is an "All genomes" section with a dropdown menu set to "-- Select a species --" and a link to "View full list of all species".

The right column is titled "Ensembl Genomes" and contains a paragraph explaining the project: "The Ensembl Genomes project produces genome databases for important species from across the taxonomic range, using the Ensembl software system. Five sites are now available: the existing [Ensembl Bacteria](#), [Ensembl Protists](#) and [Ensembl Metazoa](#) sites plus the newly released [Ensembl Plants](#) and [Ensembl Fungi](#) sites. These new sites complement the existing [Ensembl](#) site, with its focus on vertebrate genomes. You can search all Ensembl and Ensembl Genomes databases from the search bar in the top right of this page." Below this paragraph, it states "Ensembl Genomes data is available through many of the same routes as Ensembl data. Data can be accessed via:" followed by a bulleted list of access methods: "this web browser (go to <http://bacteria.ensembl.org>, <http://metazoa.ensembl.org>, etc., or to <http://www.ensemblgenomes.org> for the project homepage)", "through BioMarts (query optimised data warehouses) constructed for each of the Ensembl Genomes sites ( [Bacteria](#) [Metazoa](#) [Protists](#) [Fungi](#) [Plants](#) )", "via FTP (<ftp://ensemblgenomes.org/pub>)", "via the Ensembl Genomes public mysql server ([mysql.ebi.ac.uk:4157:anonymous](mysql://mysql.ebi.ac.uk:4157:anonymous)).", and "using the Ensembl API." Below the list, another paragraph explains: "The API has been modified slightly to support the existence of 'genome collections', i.e. the existence of many small genomes in a single Ensembl database (a model which has been adopted for Ensembl Bacteria). The API makes the use of multi-genome databases transparent to users interested in a single genome, while methods to access a traditional, single-genome database, are unchanged. We aim to keep Ensembl Genomes software in synch with software releases of Ensembl, to ensure that users can access databases from across the taxonomic range using the same software."

At the bottom left of the screenshot, there is a link to "What's in Release 3 (October 2009)".

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

# NCBI Map Viewer

NCBI  NCBI Map Viewer

PubMed Entrez BLAST OMIM Taxonomy Structure

Search  Find Find in This View Advanced Search

[BLAST The Human Genome](#)

**Homo sapiens (human) Build 37.1 (Current)**

Chromosome: [ 1 ] 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X Y MT

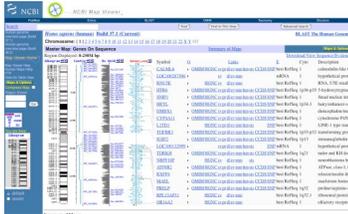
**Master Map: Genes On Sequence** [Summary of Maps](#) [Maps & Options](#)

Region Displayed: 0-250M bp [Download/View Sequence/Evidence](#)

Ideogram	Contig	Hs UniG	Genes_seq	Symbol	Q	Links	E	Cyto	Description
	NT_077402	Hs.517145		<a href="#">CALML6</a>	+	<a href="#">OMIM HGNC sv pr dl ev mm hm sts CCDS SNP</a>	best RefSeq	1	calmodulin-like 6
	NT_077912	Hs.368664		<a href="#">LOC100287506</a>	+	<a href="#">sv dl ev mm</a>	mRNA	1	hypothetical protein
	NT_004356	Hs.370858		<a href="#">RNU5E</a>	+	<a href="#">HGNC sv dl ev mm</a>	best RefSeq	1	RNA, U5E small nu
	NT_021937	Hs.707988		<a href="#">HTR6</a>	+	<a href="#">OMIM HGNC sv pr dl ev mm hm sts CCDS SNP</a>	best RefSeq	1p36-p35	5-hydroxytryptamin
	NT_04610	Hs.151163		<a href="#">SNIP1</a>	+	<a href="#">OMIM HGNC sv pr dl ev mm hm sts CCDS SNP</a>	best RefSeq	1	Smad nuclear intera
	NT_04414	Hs.706890		<a href="#">HEYL</a>	+	<a href="#">OMIM HGNC sv pr dl ev mm hm sts CCDS SNP</a>	best RefSeq	1p34.3	hairy/enhancer-of-sp
	NT_04413	Hs.370581		<a href="#">DMBX1</a>	+	<a href="#">OMIM HGNC sv pr dl ev mm hm sts CCDS SNP</a>	best RefSeq	1	diencephalon/mesen
	NT_04412	Hs.3873		<a href="#">CYP4A11</a>	+	<a href="#">OMIM HGNC sv pr dl ev mm hm sts CCDS SNP</a>	best RefSeq	1	cytochrome P450, fa
	NT_04411	Hs.473583		<a href="#">LITD1</a>	+	<a href="#">HGNC sv pr dl ev mm hm sts SNP</a>	best RefSeq	1	LINE-1 type transpo
	NT_04410	Hs.706748		<a href="#">TGFB3</a>	+	<a href="#">OMIM HGNC sv pr dl ev mm hm sts CCDS SNP</a>	best RefSeq	1p33-p32	transforming growth
	NT_04409	Hs.512676		<a href="#">IGSF2</a>	+	<a href="#">OMIM HGNC sv pr dl ev mm hm sts CCDS SNP</a>	best RefSeq	1p13	immunoglobulin sup
	NT_04408	Hs.180909		<a href="#">LOC100132999</a>	+	<a href="#">sv pr dl ev mm hm sts SNP</a>	mRNA	1	hypothetical protein
	NT_04407	Hs.49727		<a href="#">TDRKH</a>	+	<a href="#">OMIM HGNC sv pr dl ev mm hm sts CCDS SNP</a>	best RefSeq	1q21	tudor and KH doma
	NT_032977	Hs.708014		<a href="#">NBPF18P</a>	+	<a href="#">HGNC sv dl ev mm sts</a>	best RefSeq	1	neuroblastoma break
	NT_032976	Hs.532359		<a href="#">ATP8B2</a>	+	<a href="#">OMIM HGNC sv pr dl ev mm hm sts CCDS SNP</a>	best RefSeq	1	ATPase, class I, type
	NT_077389	Hs.73799		<a href="#">RXFP4</a>	+	<a href="#">OMIM HGNC sv pr dl ev mm hm sts CCDS SNP</a>	best RefSeq	1	relaxin/insulin-like f
	NT_113793	Hs.719862		<a href="#">MAEL</a>	+	<a href="#">OMIM HGNC sv pr dl ev mm hm sts CCDS SNP</a>	best RefSeq	1	maelstrom homolog
	NT_113792	Hs.594444		<a href="#">PRELP</a>	+	<a href="#">OMIM HGNC sv pr dl ev mm hm sts CCDS SNP</a>	best RefSeq	1q32	proline/arginine-rich
	NT_113791	Hs.491494		<a href="#">RPL13AP11</a>	+	<a href="#">HGNC sv dl ev mm</a>	best RefSeq	1q32.1	ribosomal protein L
	NT_079455	Hs.517168		<a href="#">OR14A2</a>	+	<a href="#">HGNC sv pr dl ev mm</a>	best RefSeq	1	olfactory receptor, f

**Summary of Maps:**

www.ncbi.nih.gov/mapview

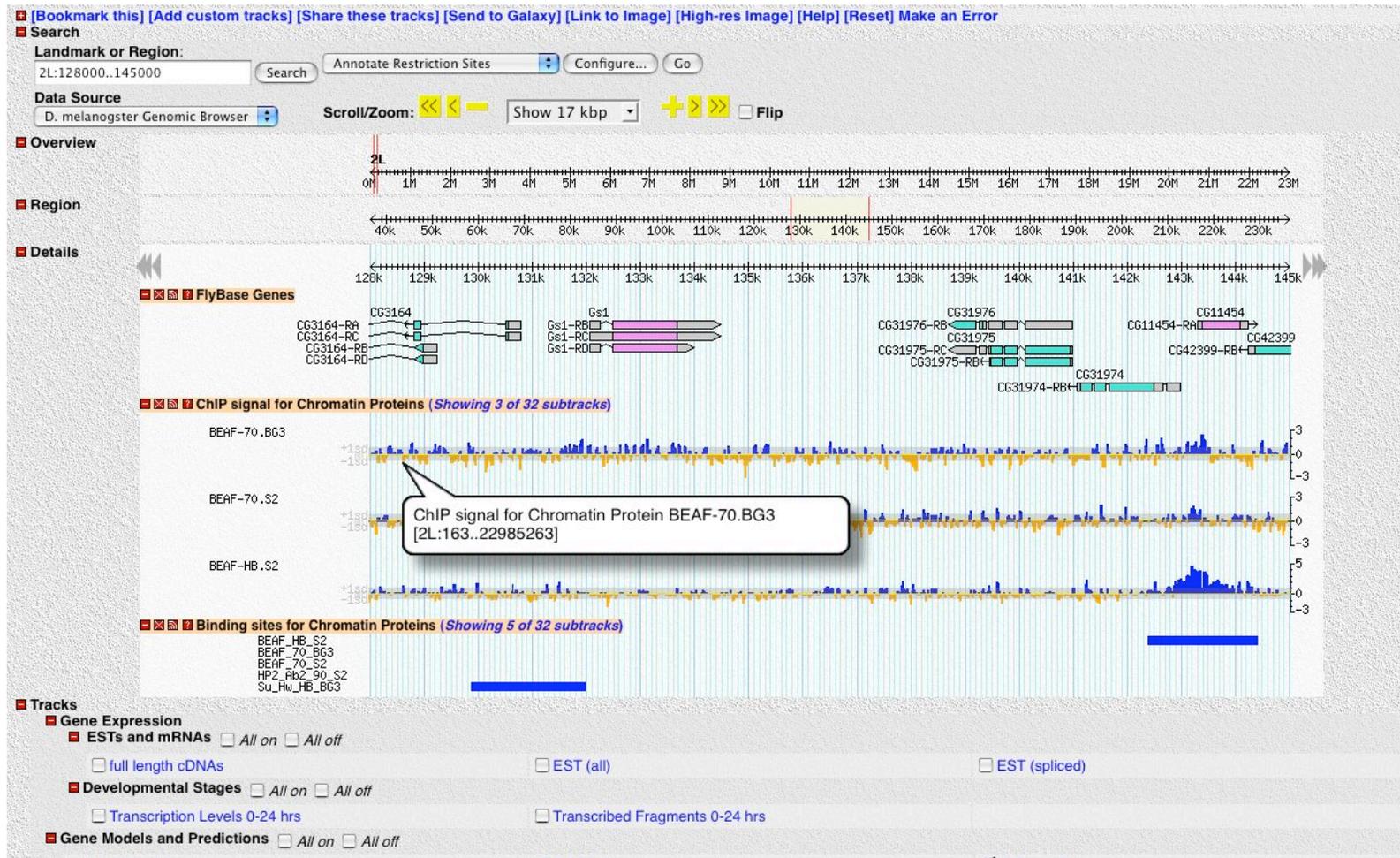


## NCBI Map Viewer

- 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

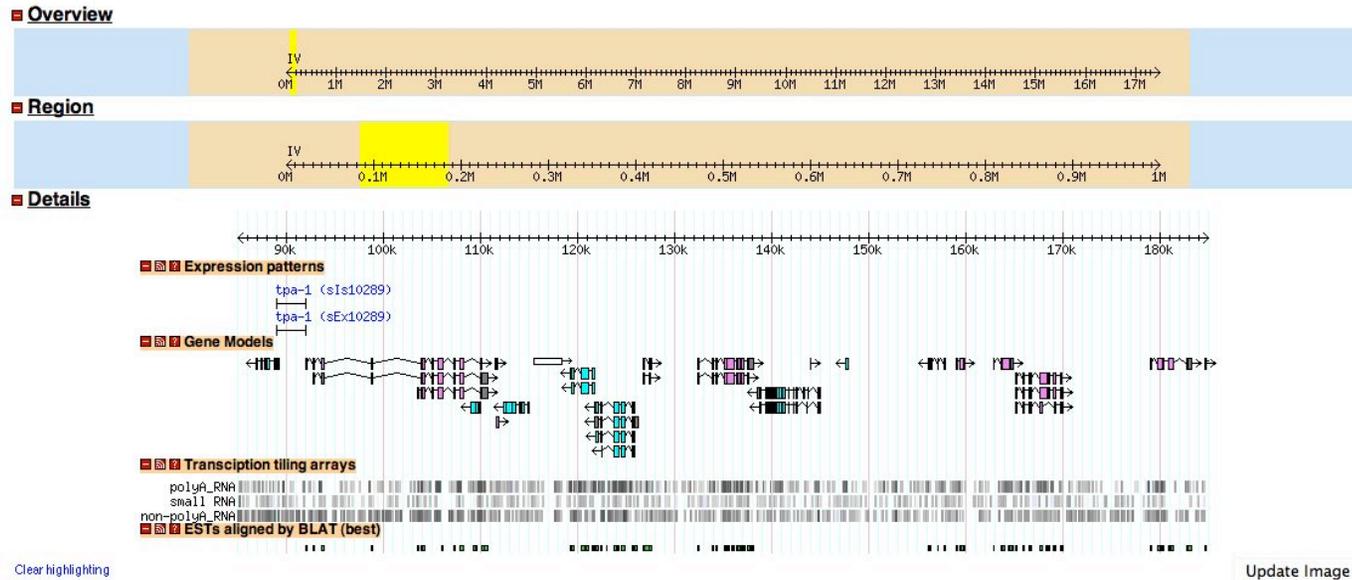
# The Generic Genome Browser



gmod.org

# The Generic Genome Browser

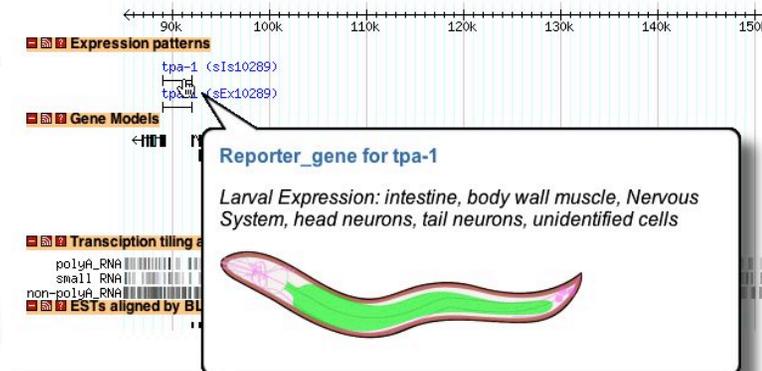
A



B



C



Dynamic and Configurable User Interface

[k to Image](#) [[High-res Image](#)] [[Help](#)] [[Reset](#)]

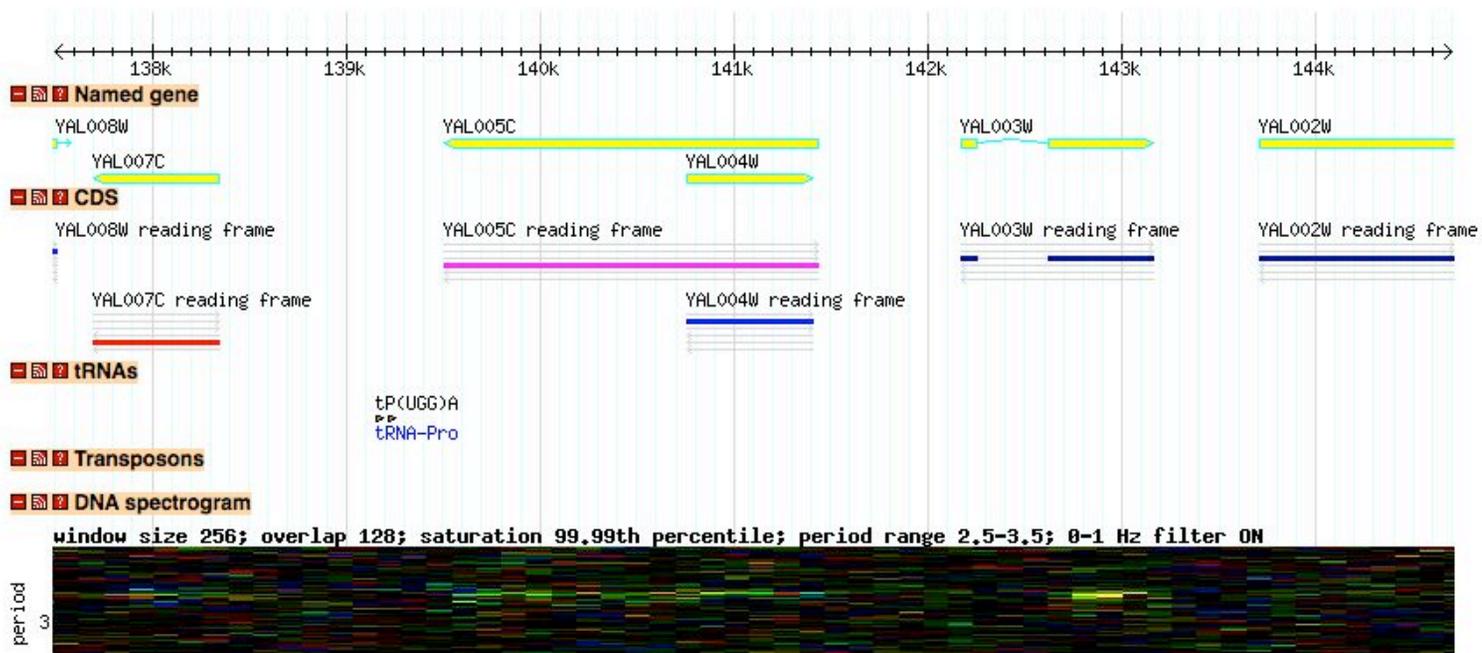
**Reports & Analysis:**

Draw DNA spectrogram

Annotate Restriction Sites    201 kbp     Flip

Download Decorated FASTA File  
 Download Sequence File  
**Draw DNA spectrogram**  
 Filter Named gene

0k 70k 80k 90k 100k 110k 120k 130k 140k 150k 160k 170k 180k 190k 200k



## 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

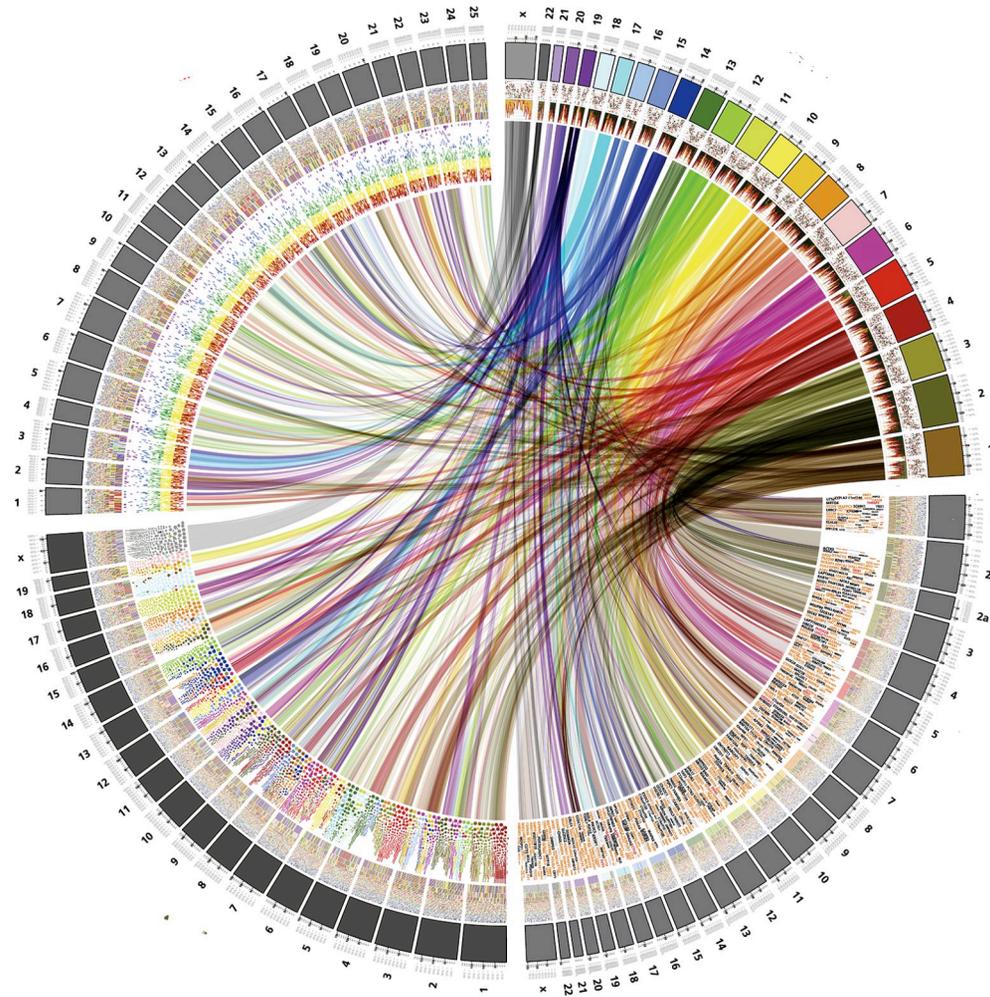
Outreach and training

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).

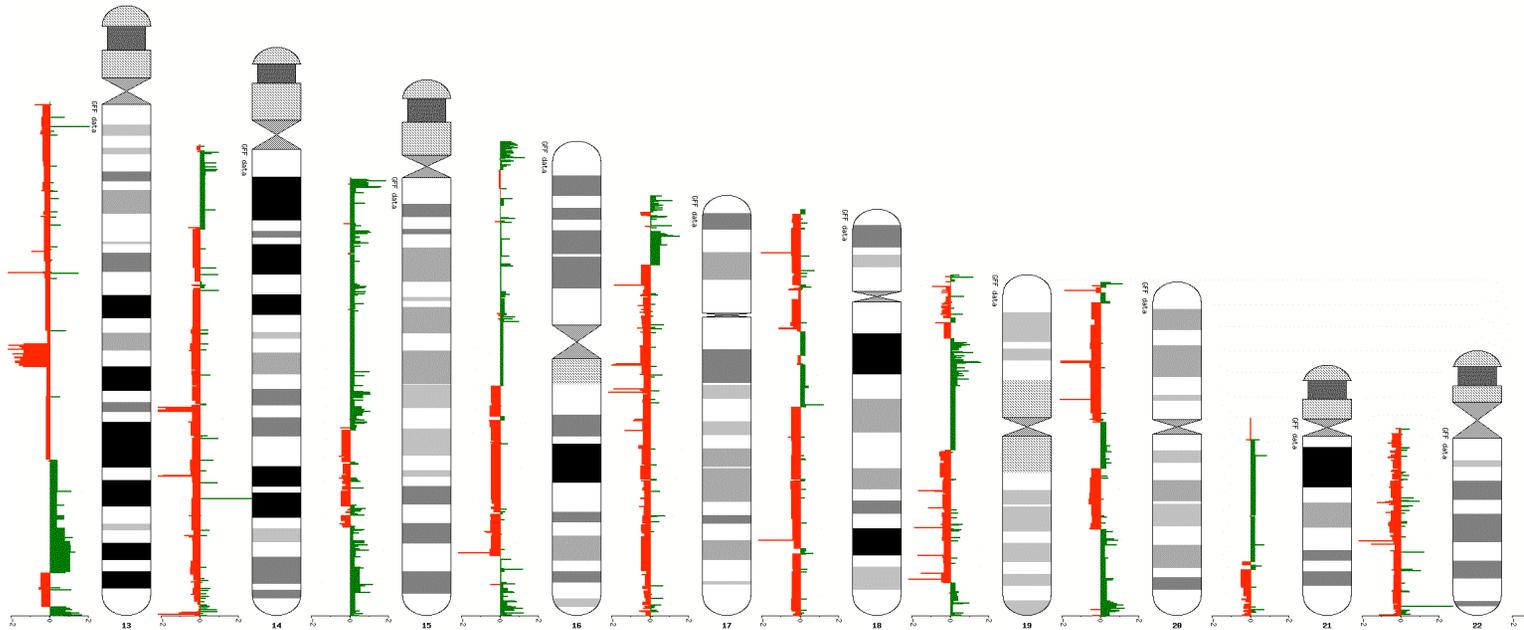
# Whole Genome Browsing

# CIRCOS



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

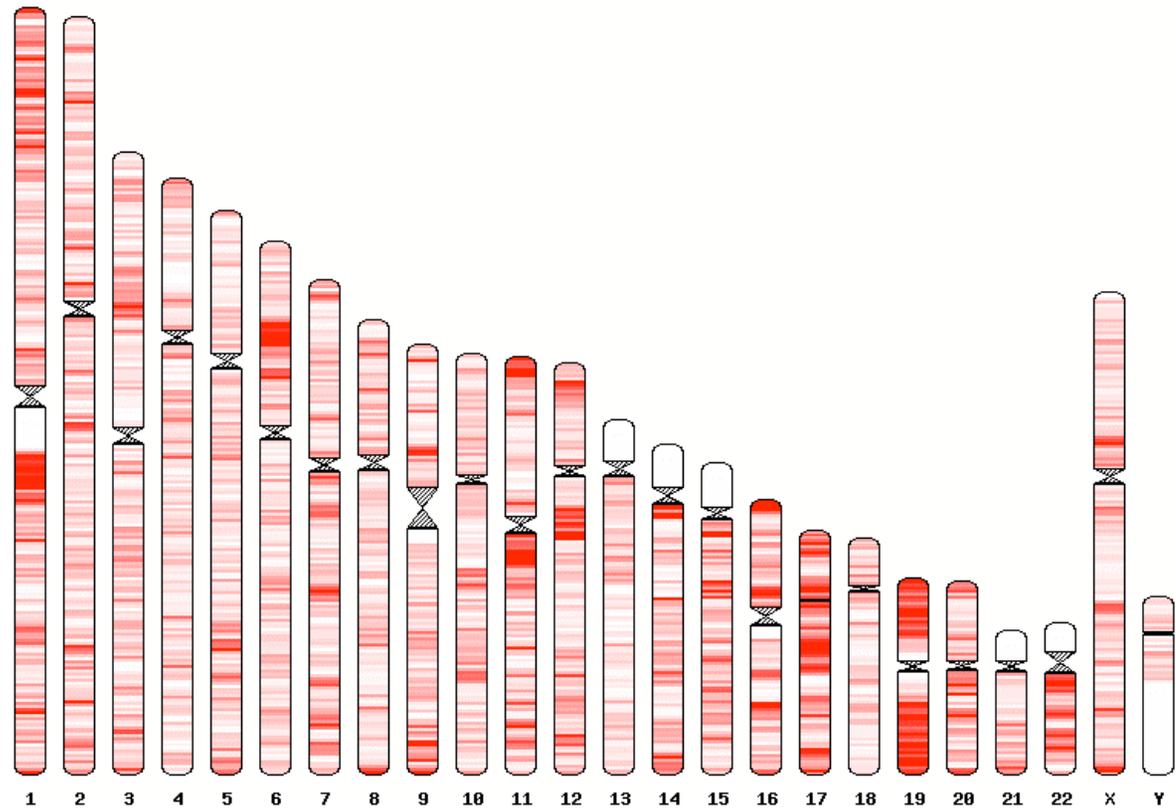
# GBrowse\_karyotype



CNV data for human chromosomes 13-22

[http://gmod.org/GBrowse\\_karyotype](http://gmod.org/GBrowse_karyotype)

# GBrowse\_karyotype



Ensembl gene density plotted on the human karyotype

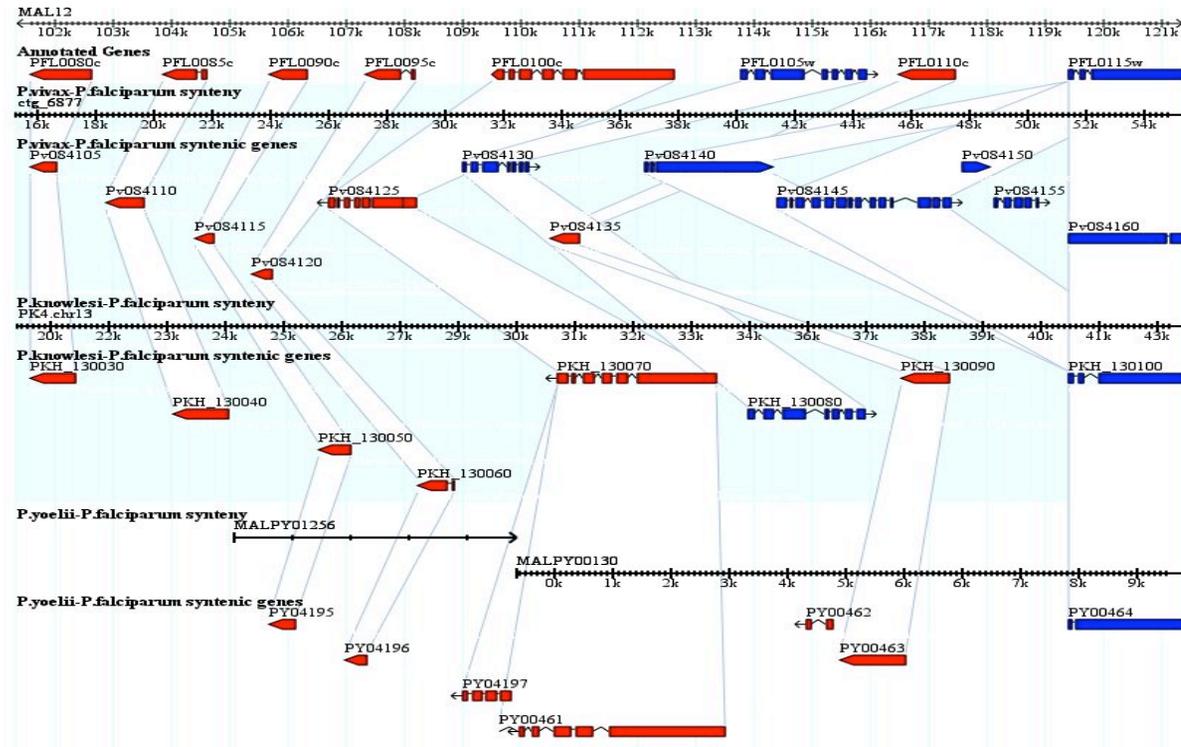
# Synteny Browsers

## **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

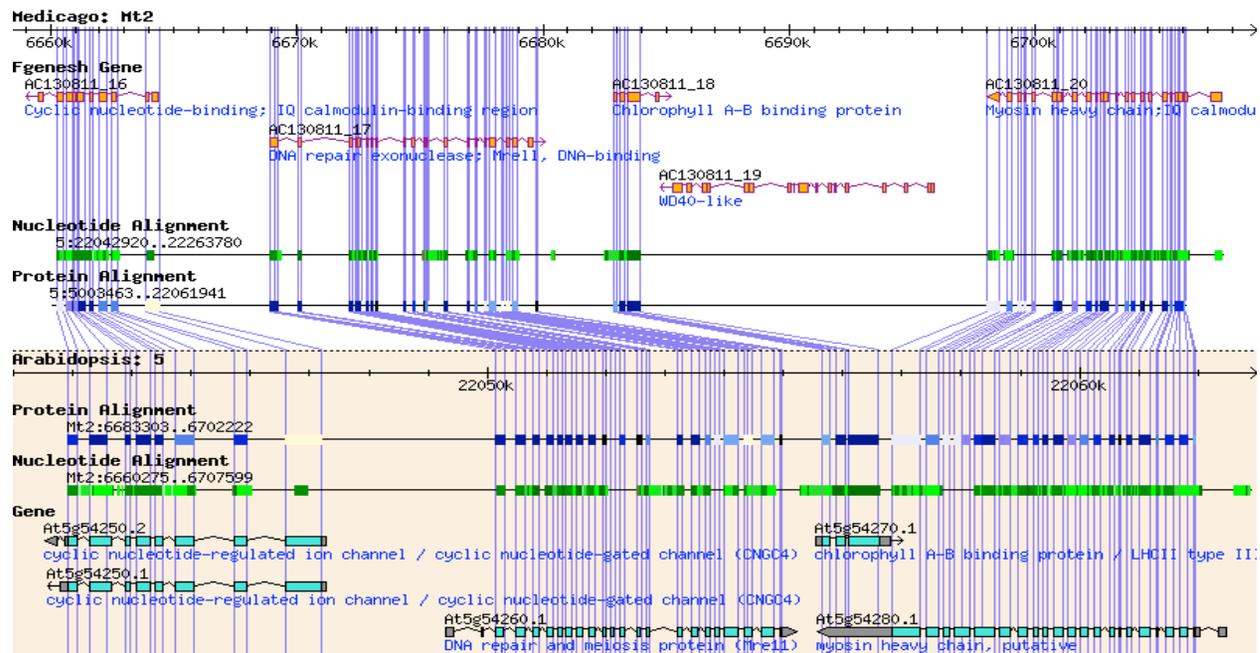
A Simple Approach to Visualizing Comparative Genome Data



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

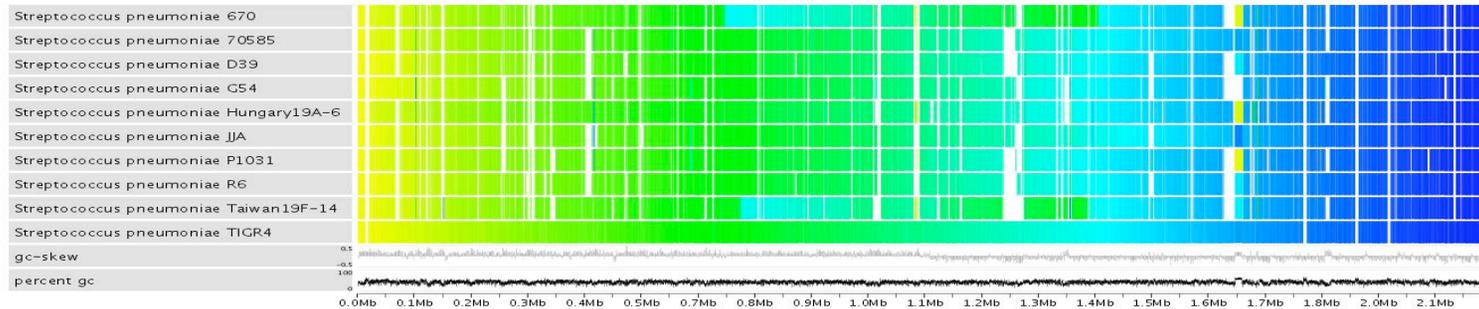
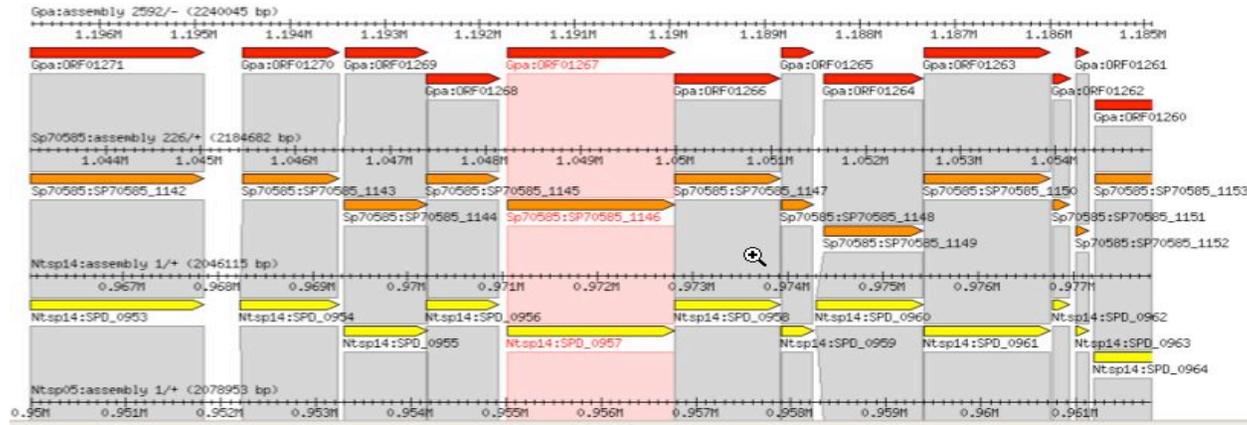
# SynBrowse

...A Synteny Browser for Comparative Sequence Analysis



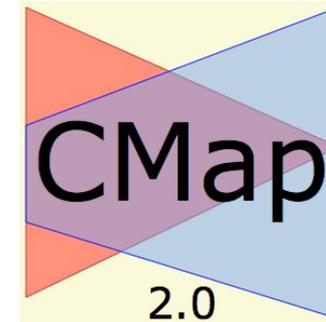
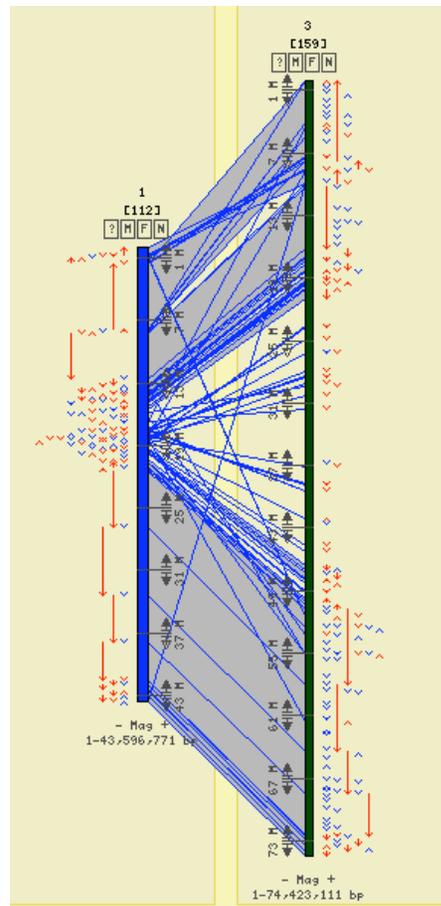
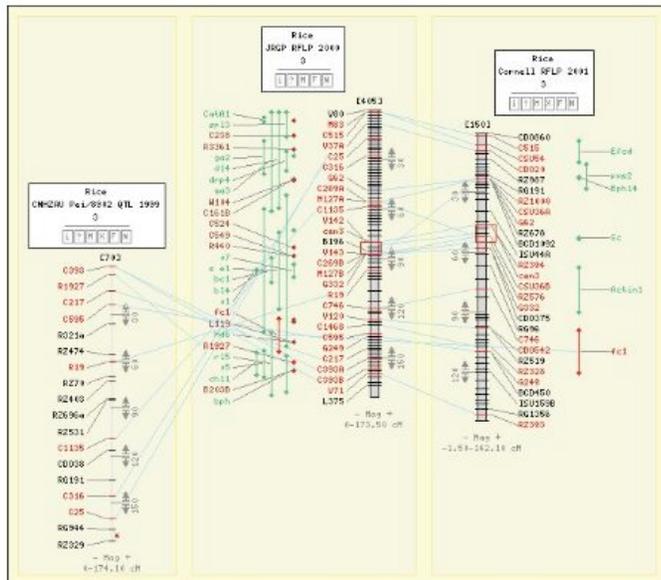
Pan, X., Stein, L. and Brendel, V. 2005. SynBrowse: a Synteny Browser for Comparative Sequence Analysis. *Bioinformatics* 21: 3461-3468

# 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.

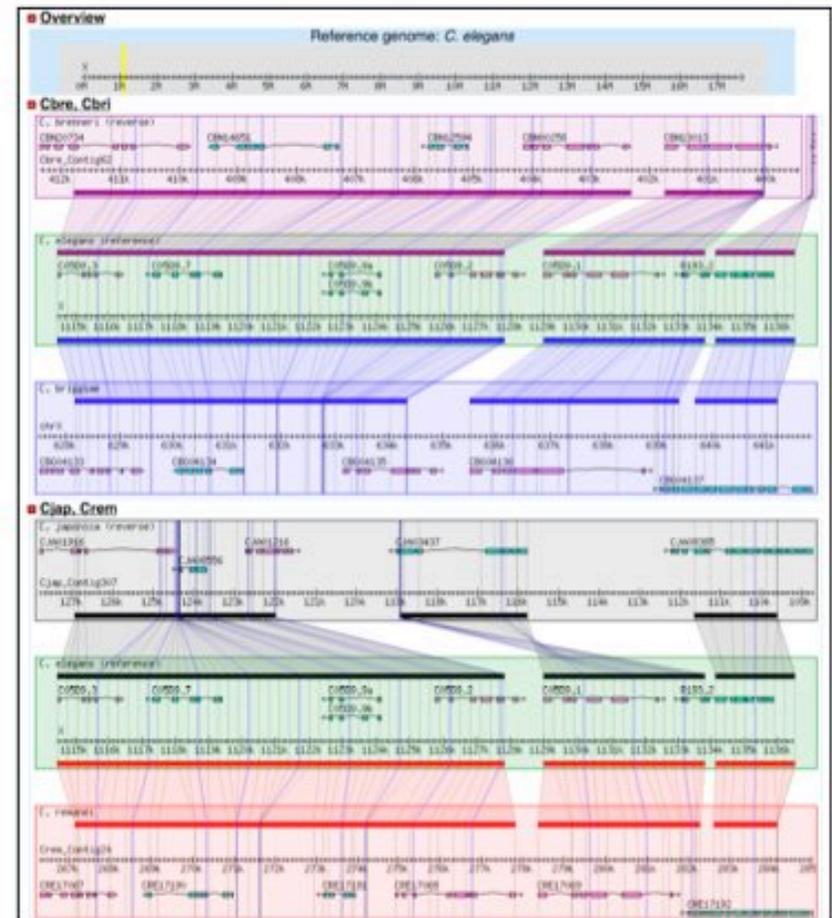


+ others...

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

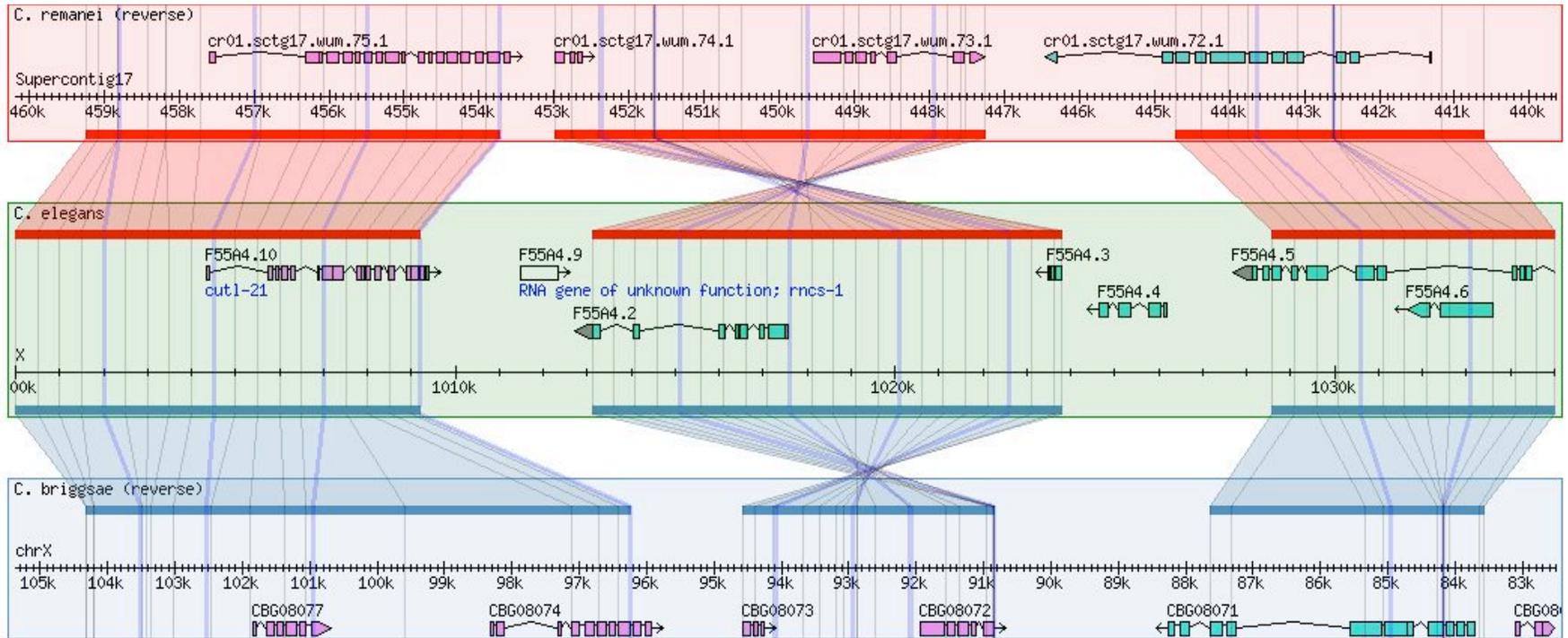
# GBrowse\_syn

- 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

# GBrowse\_syn



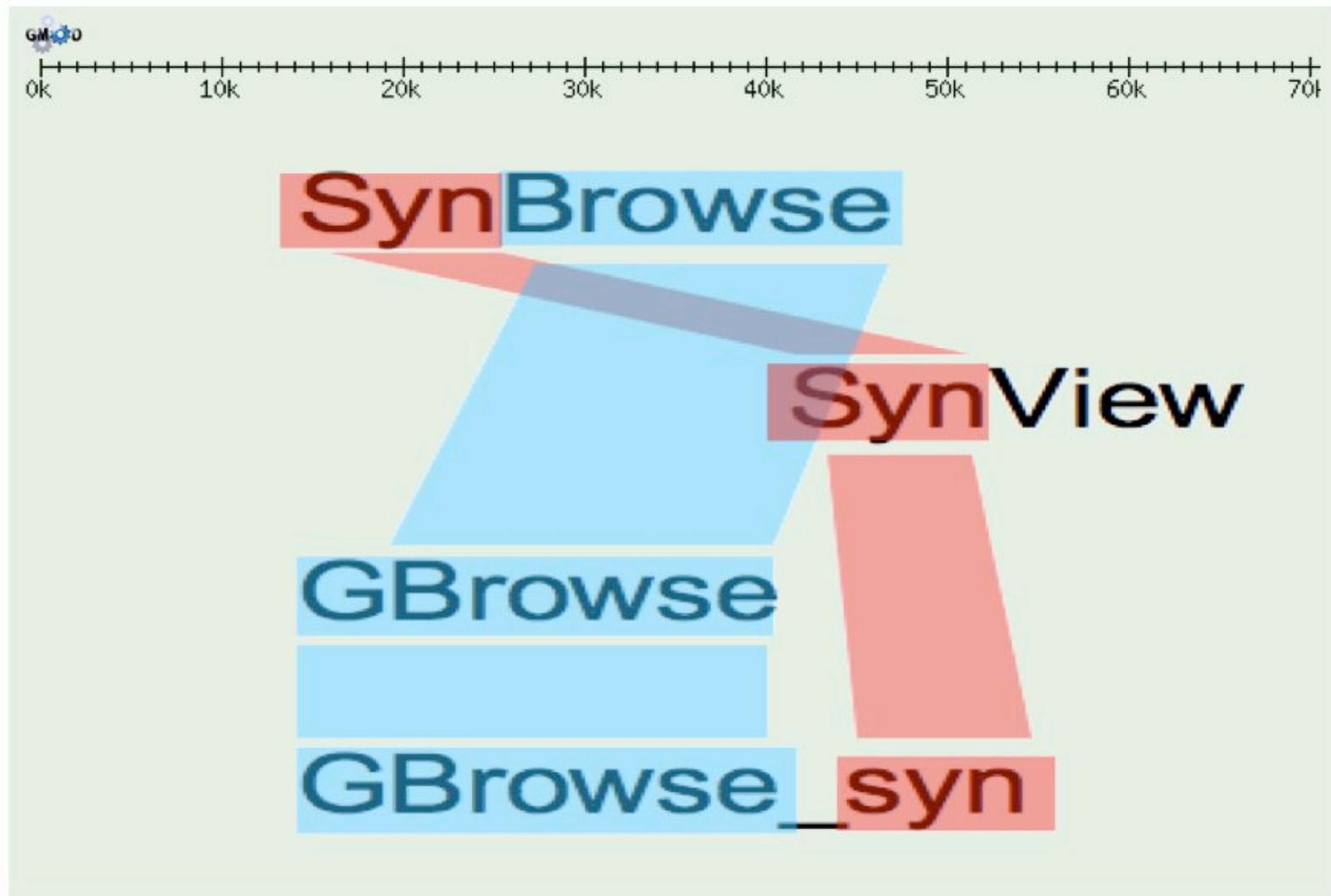
+others...

**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:**

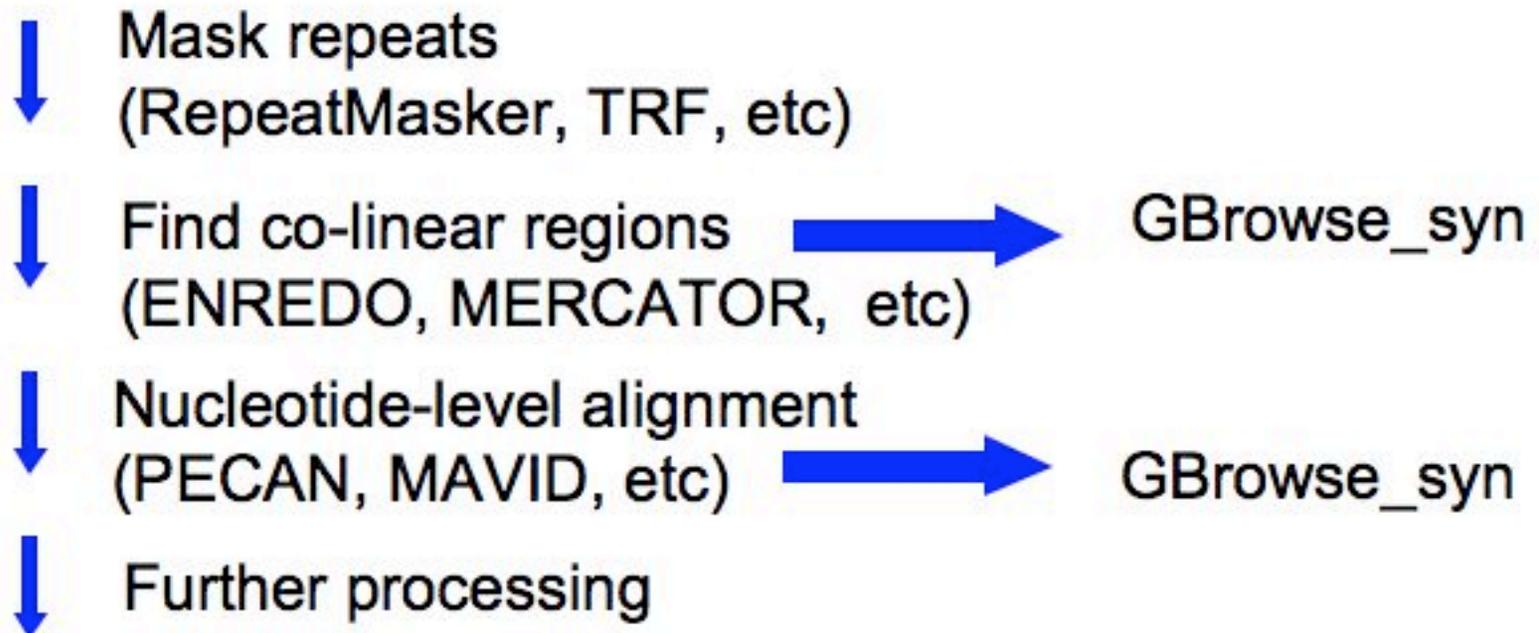
- Part of GBrowse distribution
- Uses native GFF2/3 or chado adapters for species' data
- Synteny data are stored in a separate joining database

## 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

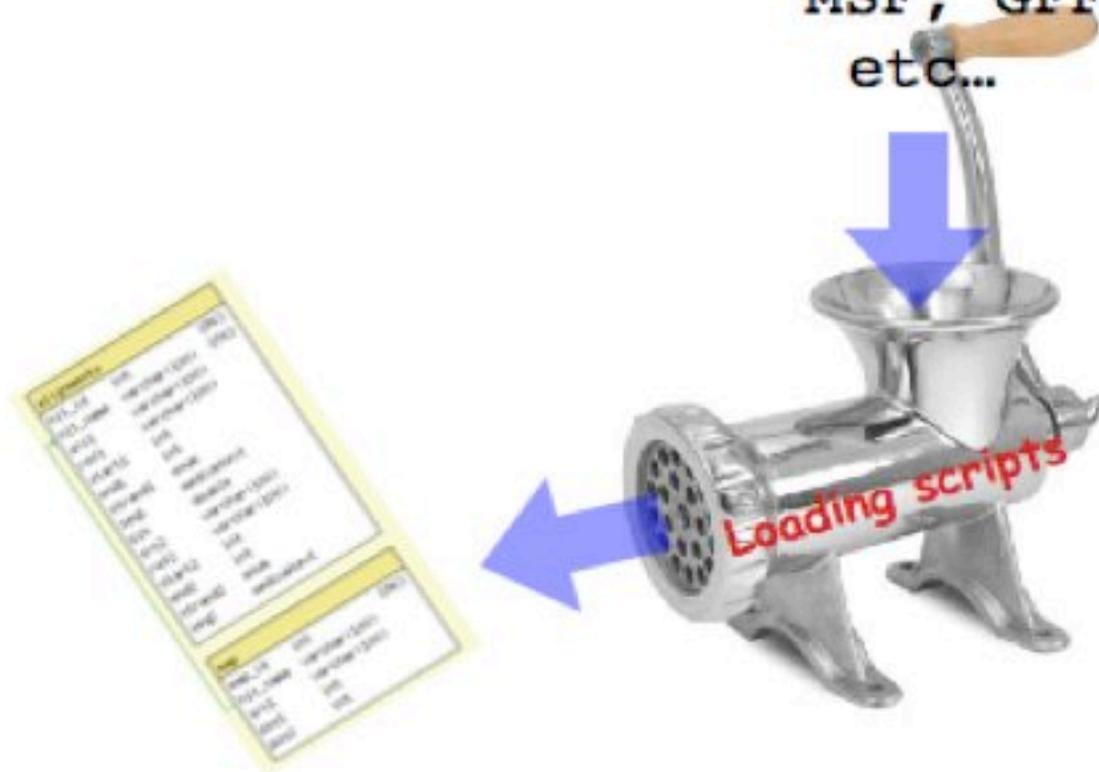
Raw genomic sequences



GBrowse

# Interoperability

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



# Problem : How to use Insertions/Deletion data

**A**

```

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

**B**

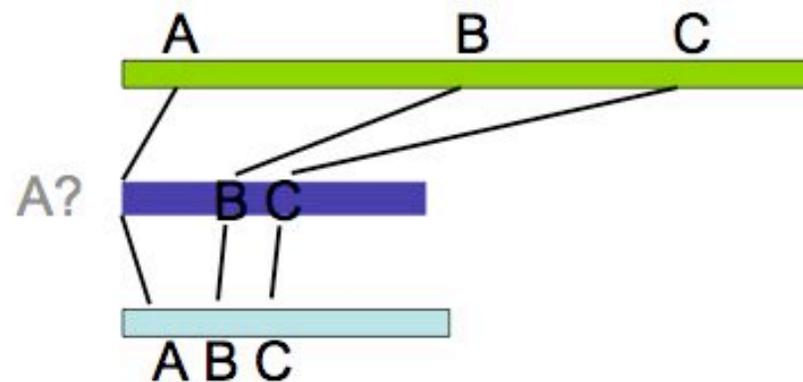
```

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

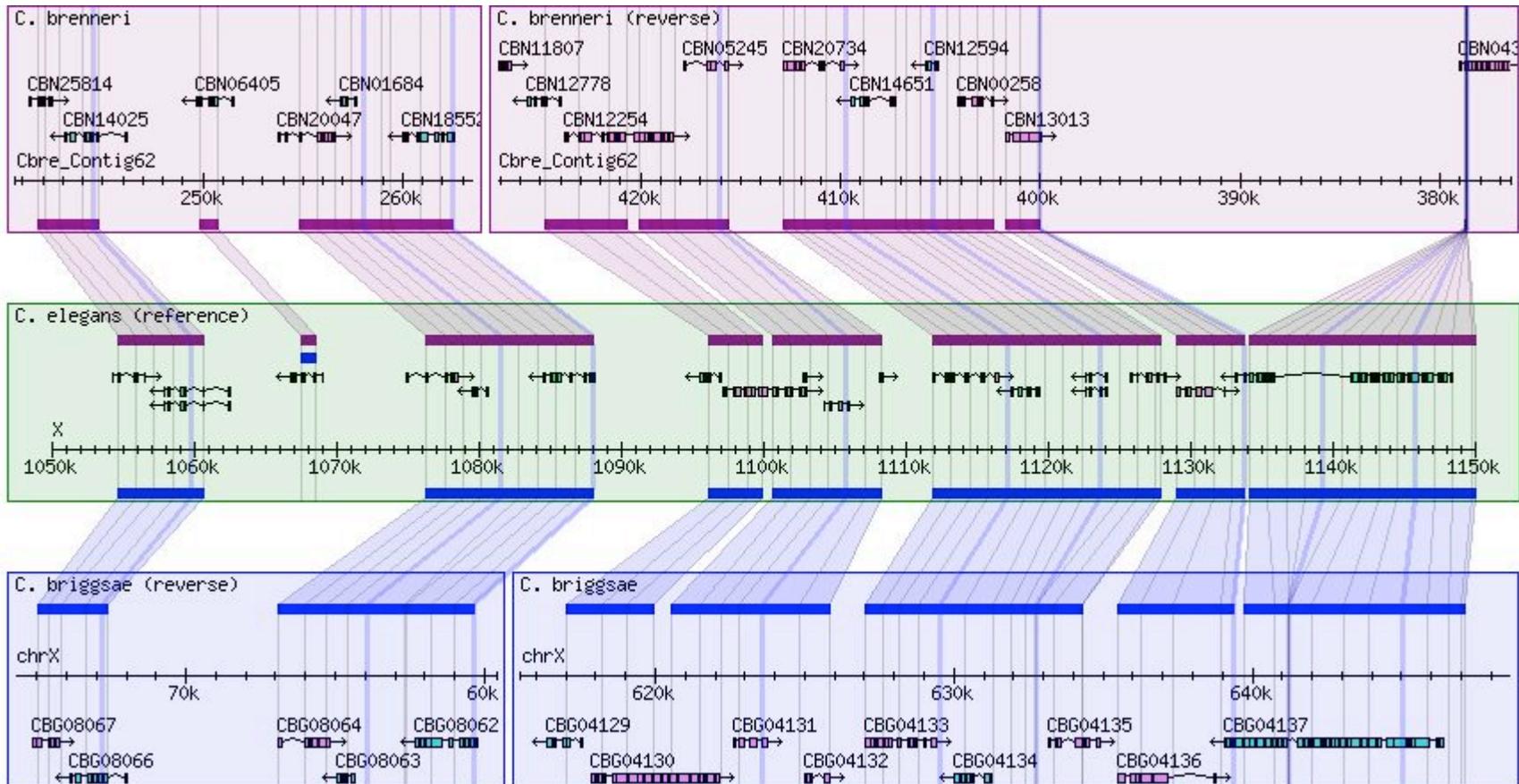
**C**

```

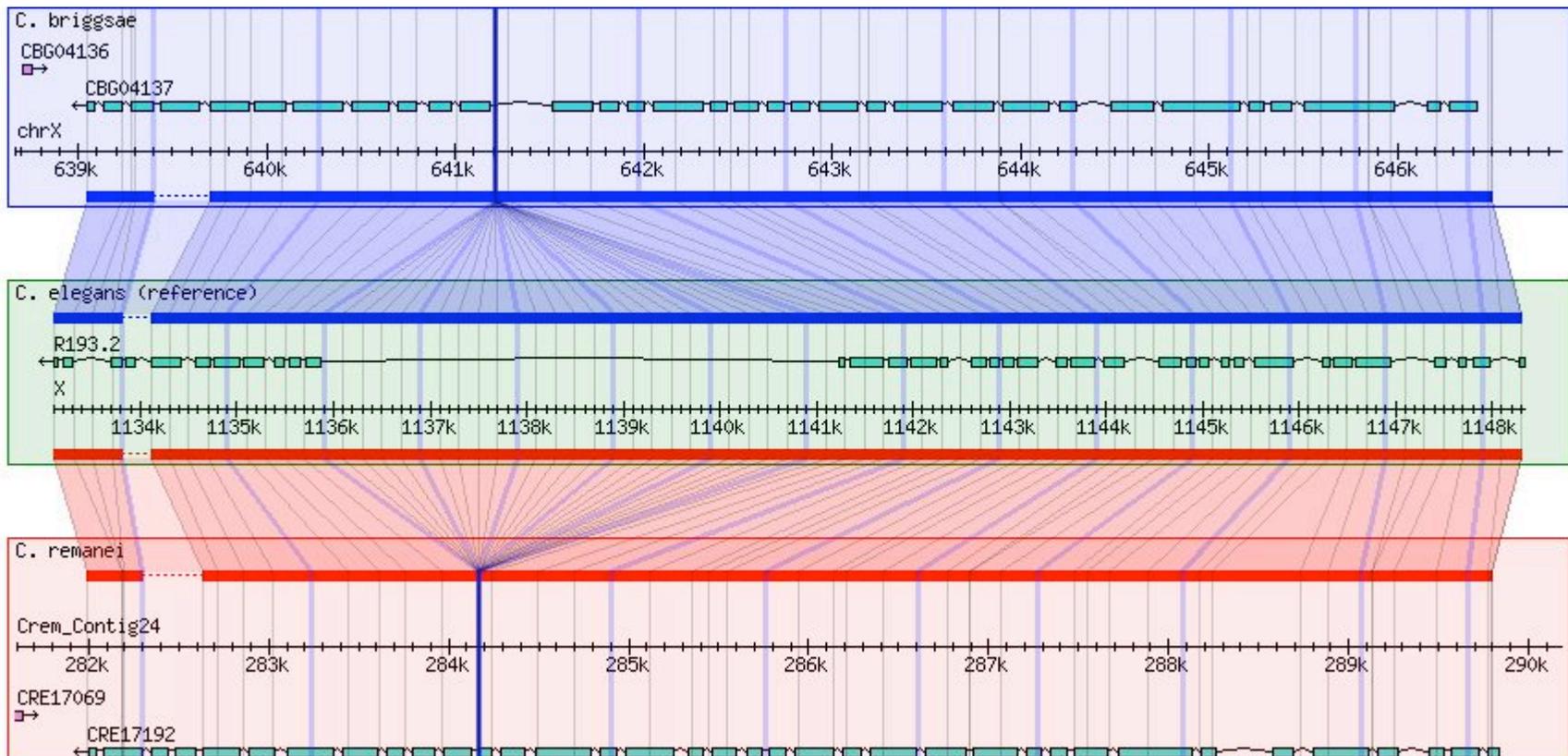
Ce-CHROMOSOME_I(+)/5195-16585  TTATTACGACATTCGTTTTATTTGAGCACAATTTGGGCCTATACTTTCAAAATCGGGGTTT
Cb-chrI(-)/4091935-4097143      --TTCATGTCAA-----TCAT
Cr-Contig8(+)/571990-577344    --TTTCTGAAAACAGGTAGTATTATGGTTCCGAGGGTGTAGGGTTTCGAAACCGGCCTAG
                                   * * *
    
```

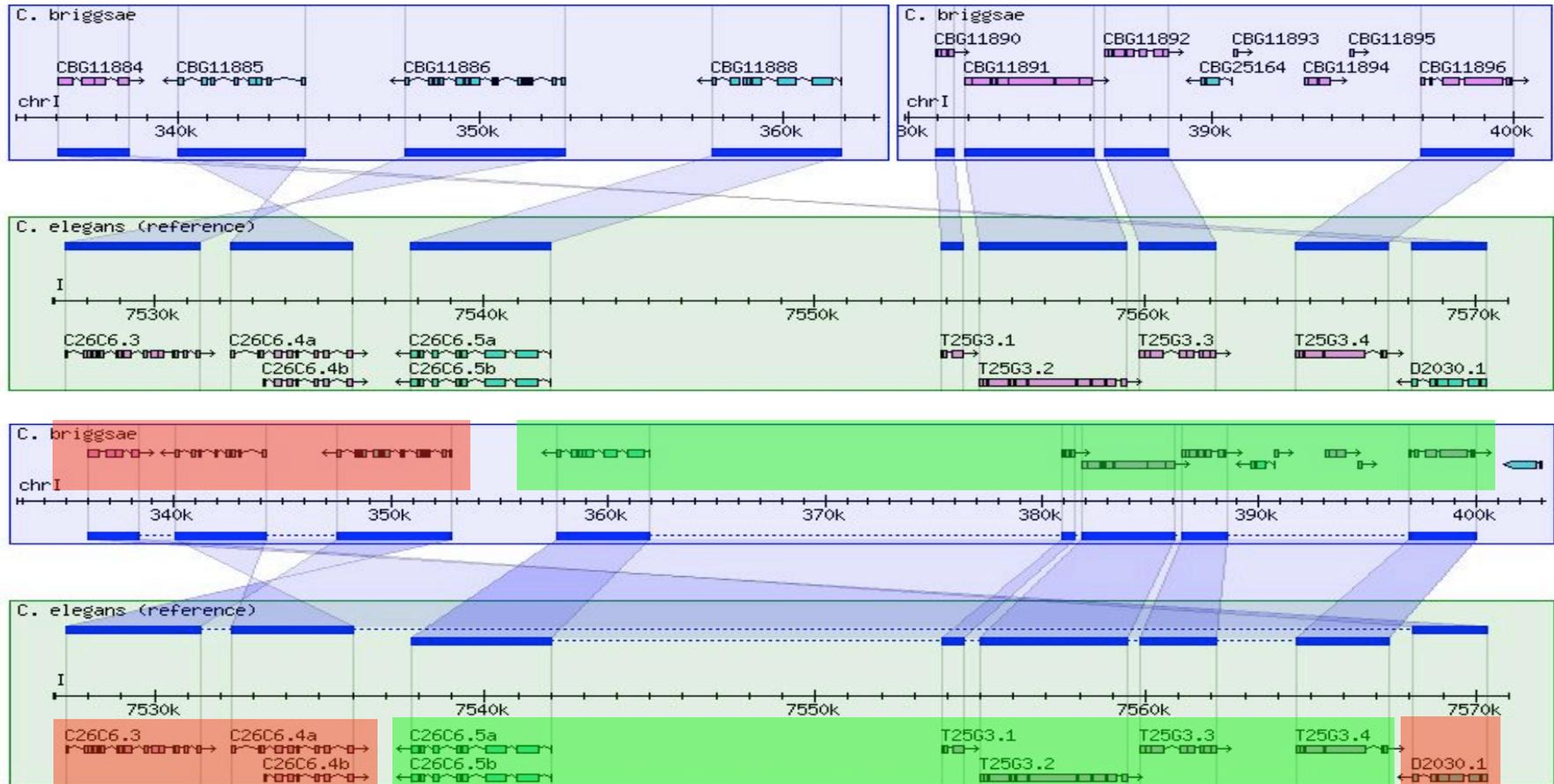


# Tracking Indels with grid lines

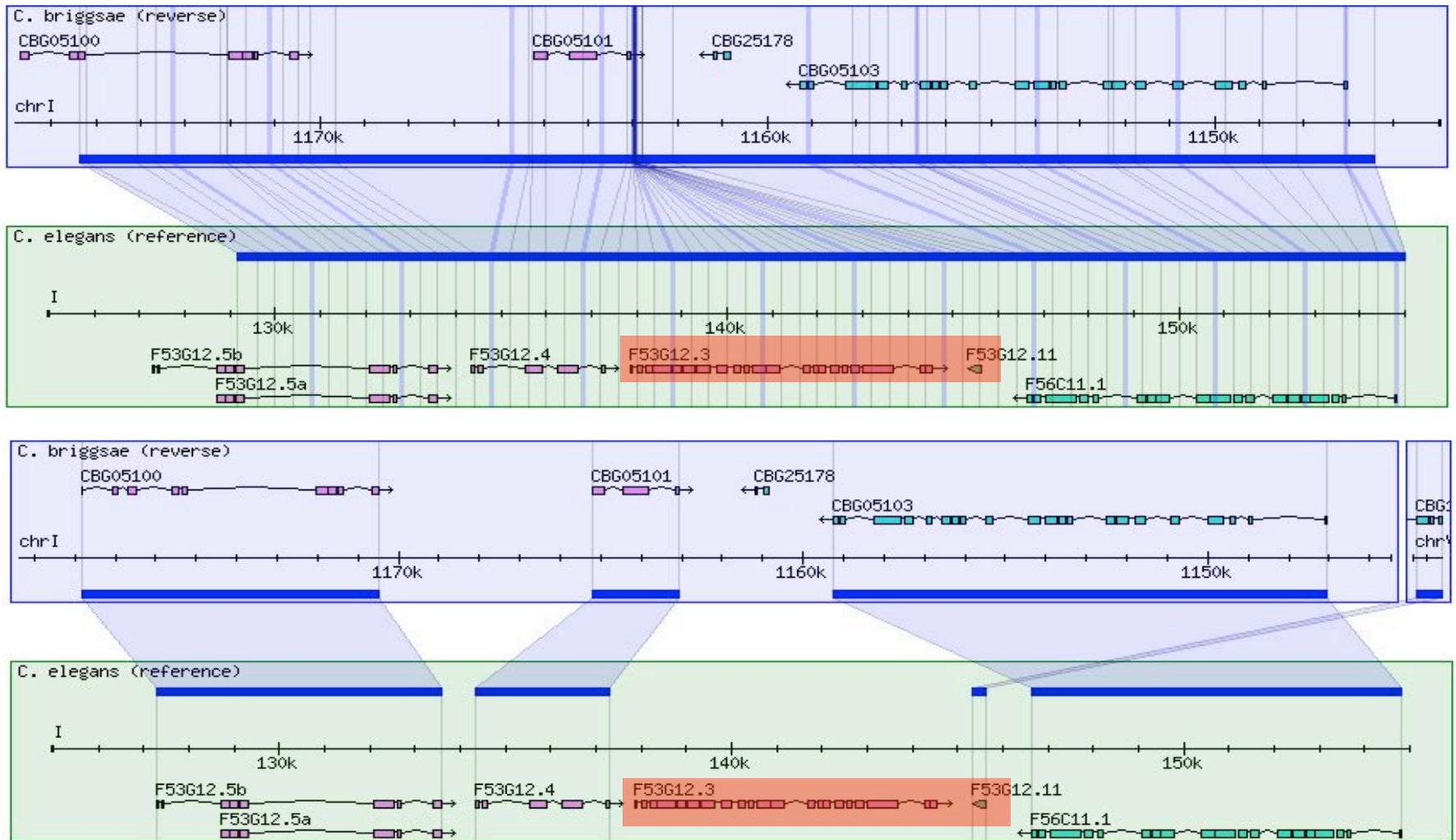


# Evolution of Gene Structure





**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*.

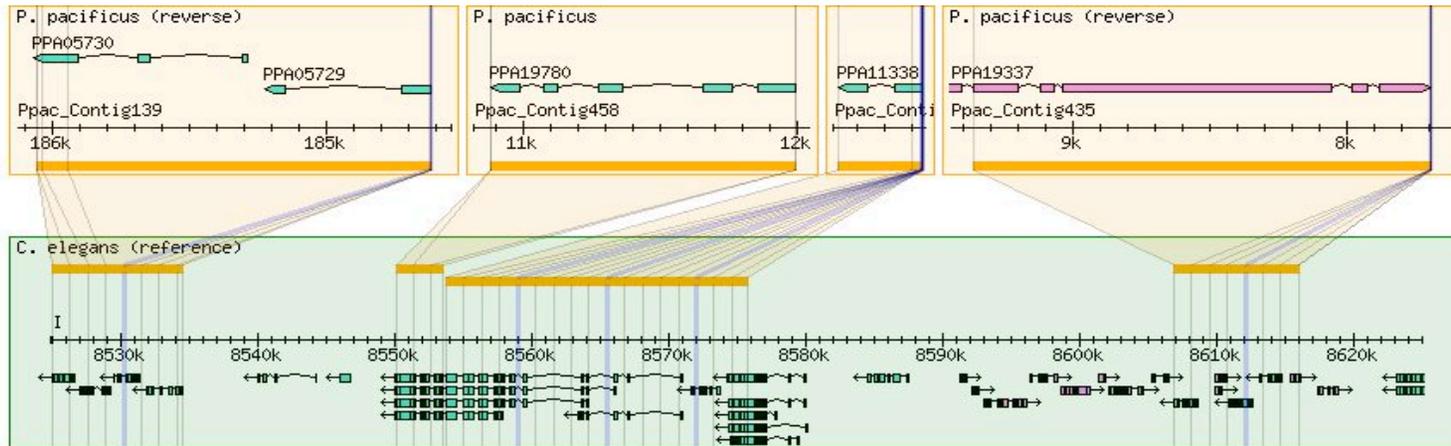
What if the aligned DNA sequences are too distant?



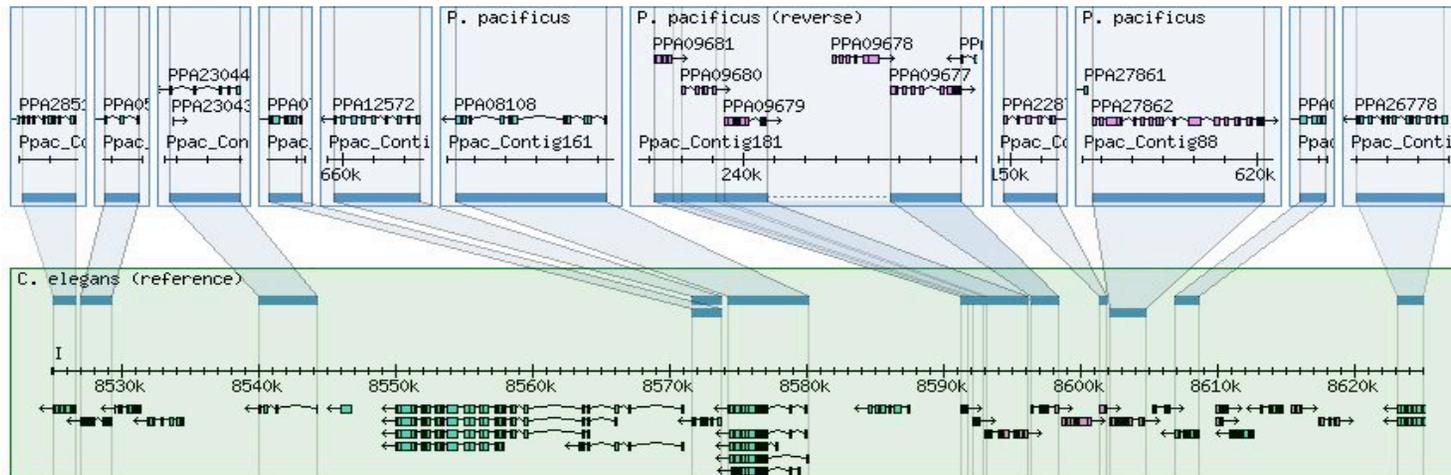
!  
=



# 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

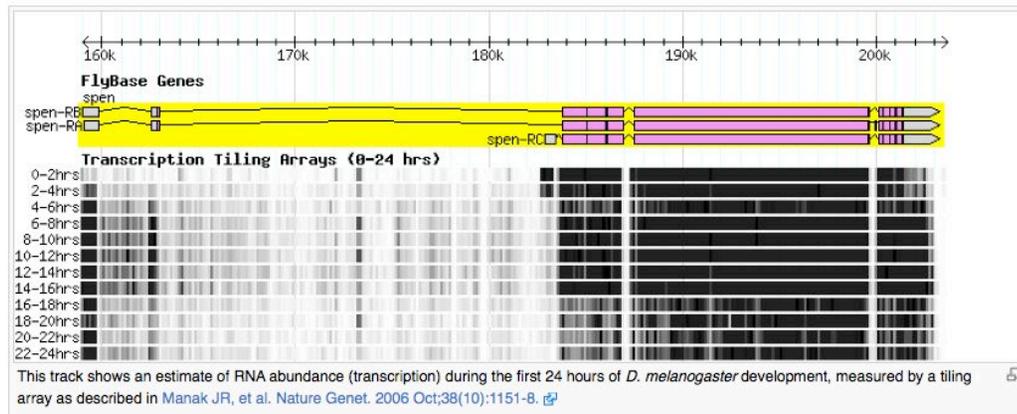
## GBrowse\_syn Resources

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

High Density Data

# Dealing with very dense data

- Microarrays
- Next-gen Sequencing



```
g tgaactgggtgtggaattgcagcaacggtttgatgatgtcaggcgttgatccacat ttggccgggaagggtgtgcagccacaccaccaga
g tgaactgggtgtggaattgcagcaacggtttgatgatgtcaggcgttgatccacat ttggccgggaagggtgtgcagccacaccaccaga
aactgggtgtggaattgcagcaacggtttgatgatgtcaggcgttgatccacat ttggccgggtgcagccacaccaccaga
gt acgggtgtggaattgcagcaacggtttgatgatgtcaggcgttgatccacat ttggccgggtgcagccacaccaccaga
gt aatgggtgtggaattgcagcaacggtttgatgatgtcaggcgttgatccacat ttggccggggagaa gcagccacaccaccaga
gt aatgggtgtggaattgcagcaacggtttgatgatgtcaggcgttgatccacat ttggccggggaga gcagccacaccaccaga
ggtt ctgggtgtggaattgcagcaacggtttgatgagcgggtcaggcgttgatccacat ttggccggggaga gcagccacaccaccaga
ggtga tgggtgtggaattgcagcaacggtttgatgagcgtcaggcgttgatccacat ttggccggggaga agccacaccaccaga
ggtga tgggtgtggaattgcagcaacggtttgatgagctcaggcgttgatccacat ttggccggggagaa agccacaccaccaga
ggtga tgggtgtggaattgcagcaacggtttgatgagcaggcgttgatccacat ttggccggggagaa agccacaccaccaga
ggtga tgggtgtggaattgcagcaacggtttgatgagcaggcgttgatccacat ttggccggggagaa agccacaccaccaga
ggtga tgggtgtggaattgcagcaacggtttgatgagcaggcgttgatccacat ttggccggggagaa agccacaccaccaga
```

- **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

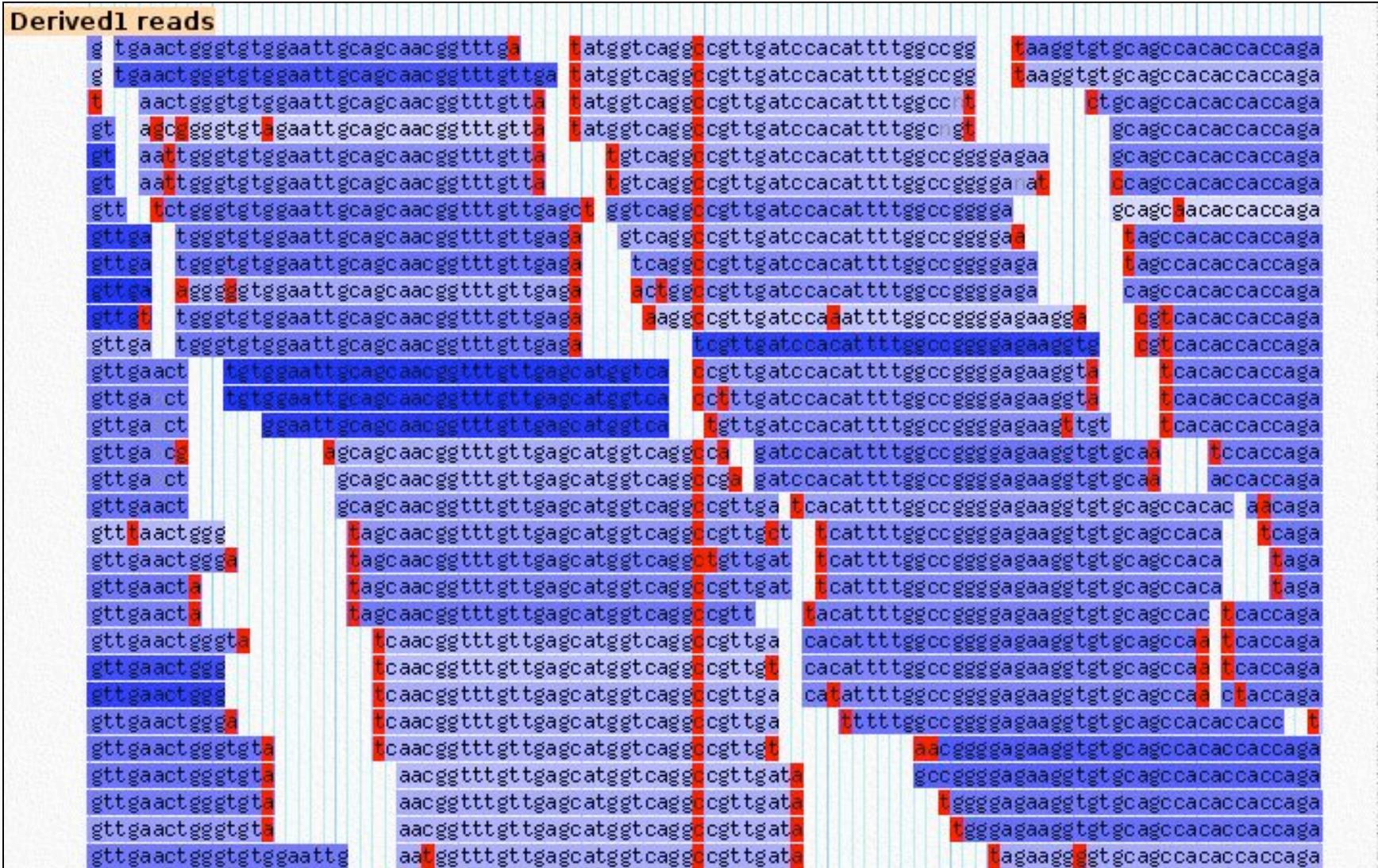
<http://genome.ucsc.edu/goldenPath/help/wiggle.html>

[http://gmod.org/wiki/GBrowse/Uploading\\_Wiggle\\_Tracks](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/>



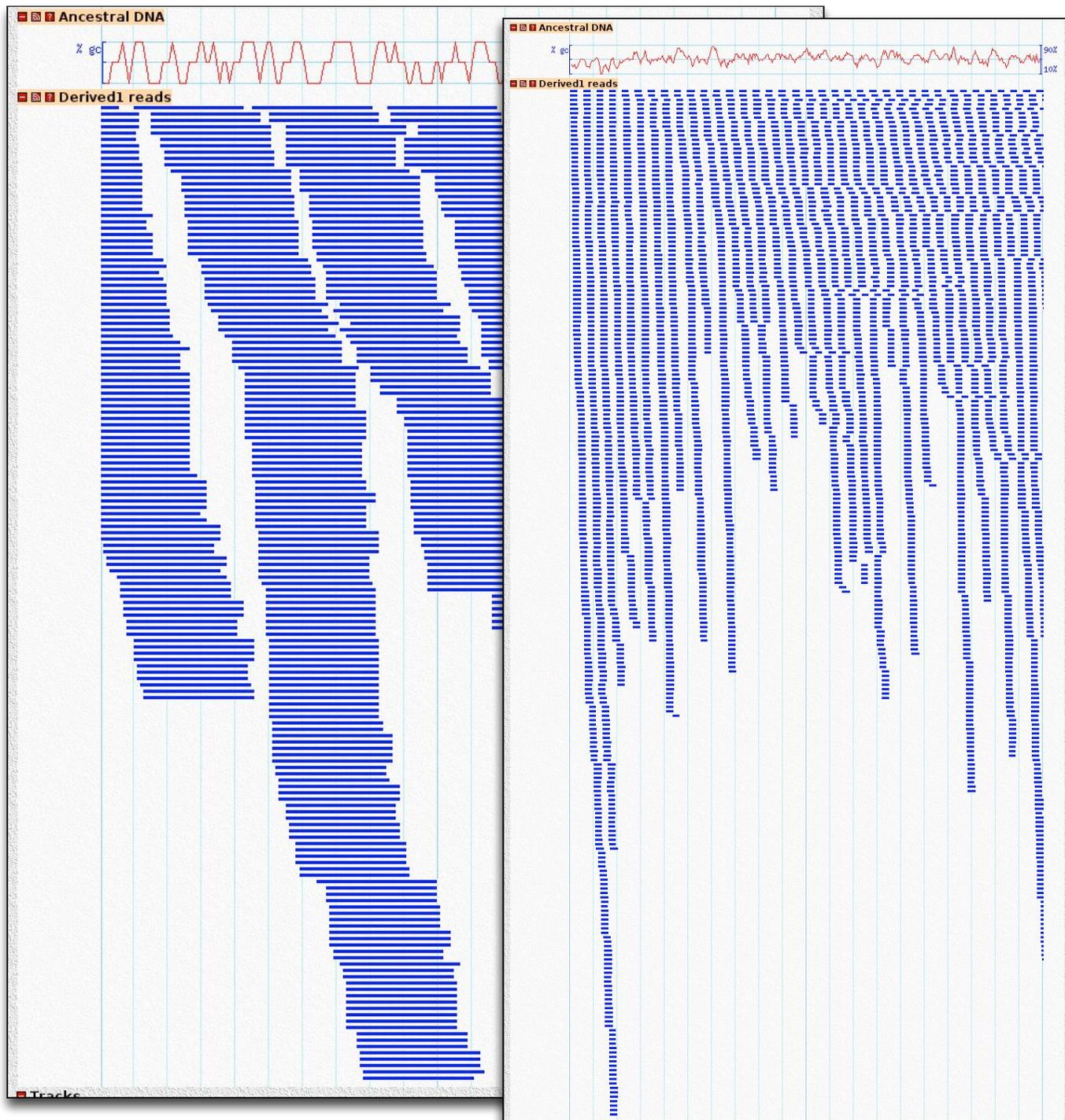


Anything in SAM format is accessible.

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

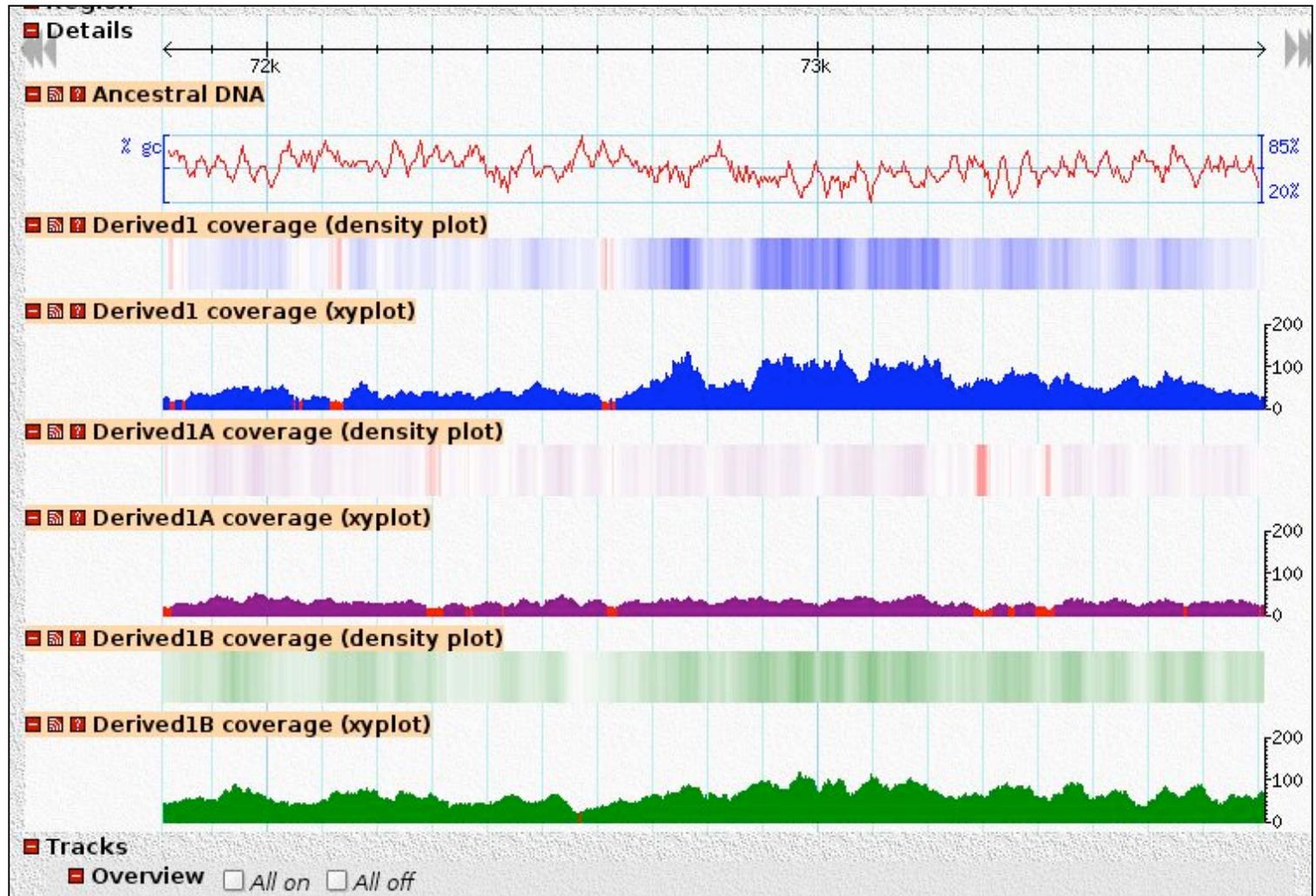


# GBrowse as an Alignment Viewer

Summarize!

SAMtools and  
GBrowse help  
here

SAMtools can  
summarize  
values from  
the data.



# Acknowledgments

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