Comparative Genomics with GBrowse_syn

Sheldon McKay,
Cold Spring Harbor Laboratory

Outline

A brief survey of synteny browsers

A few challenges of rendering comparative data

Comparative genome browsing with GBrowse_syn

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.

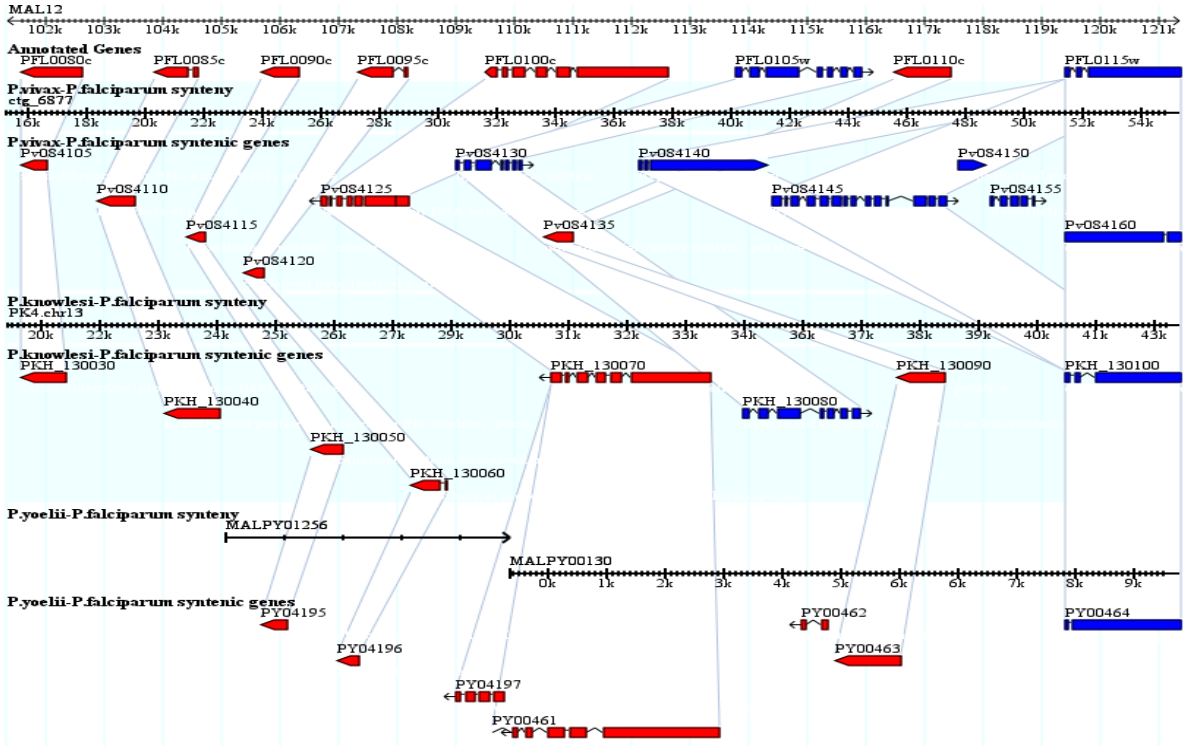
An Embarrassment of Riches*

A Brief Survey of GMOD-friendly Synteny Browsers

*From John Ozell's 1738 translation of a French play, L'Embarras des richesses (1726)

SynView

A Simple Approach to Visualizing Comparative Genome Data



CryptoDB Cryptosporidium Genome Resource

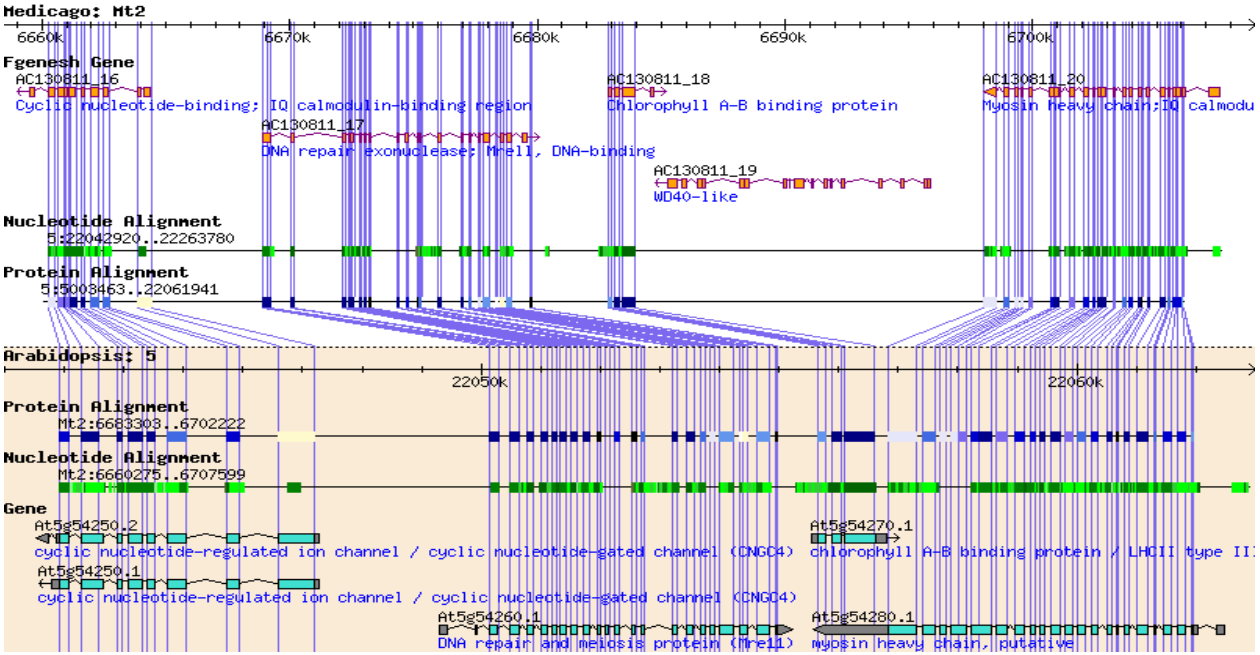
PlasmoDB Plasmodium Genome Resource

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

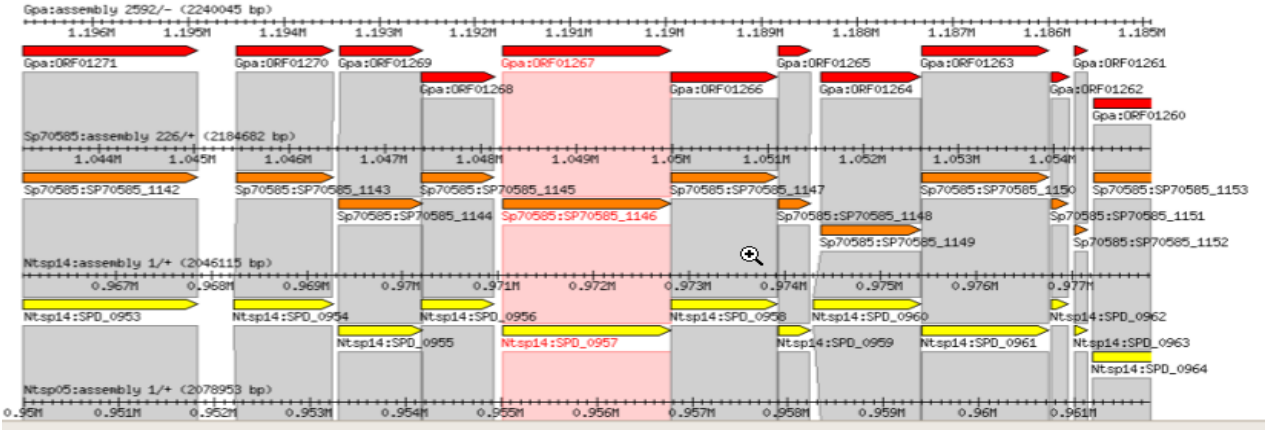
...A Synteny Browser for Comparative Sequence Analysis

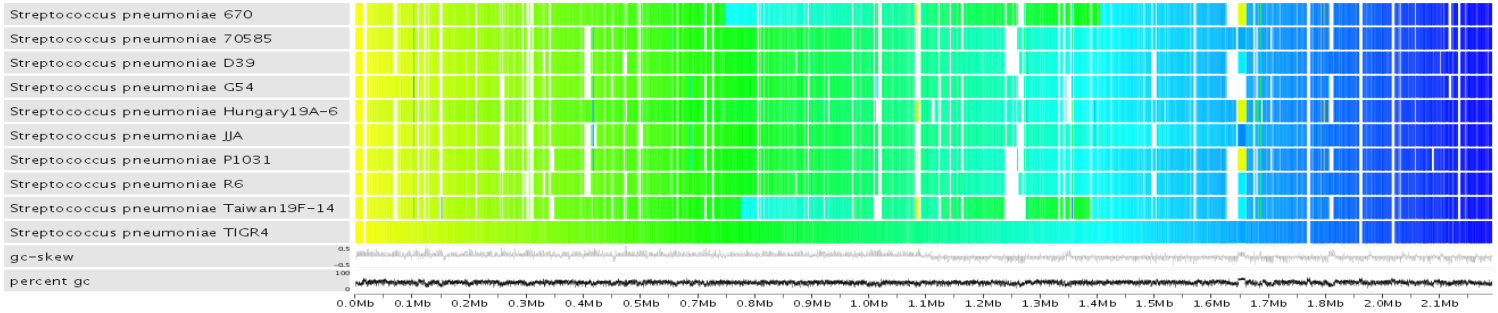


PopGenIE
Populus Genome Integrative Explorer

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

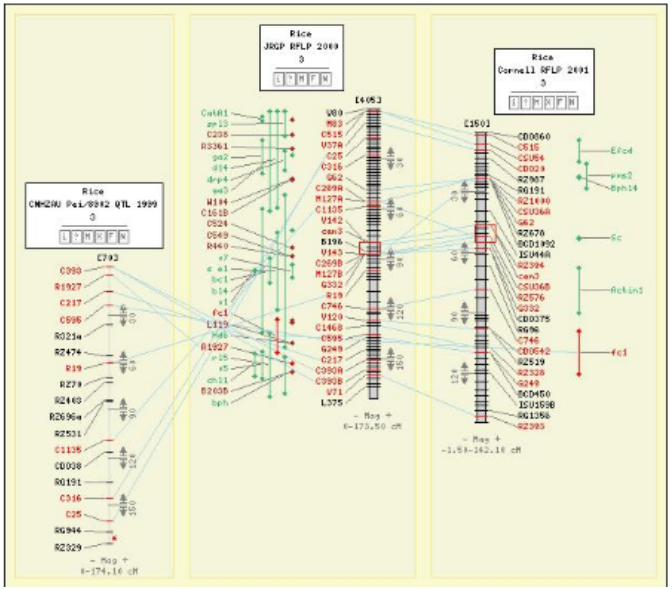
Sybil: Web-based software for comparative genomics

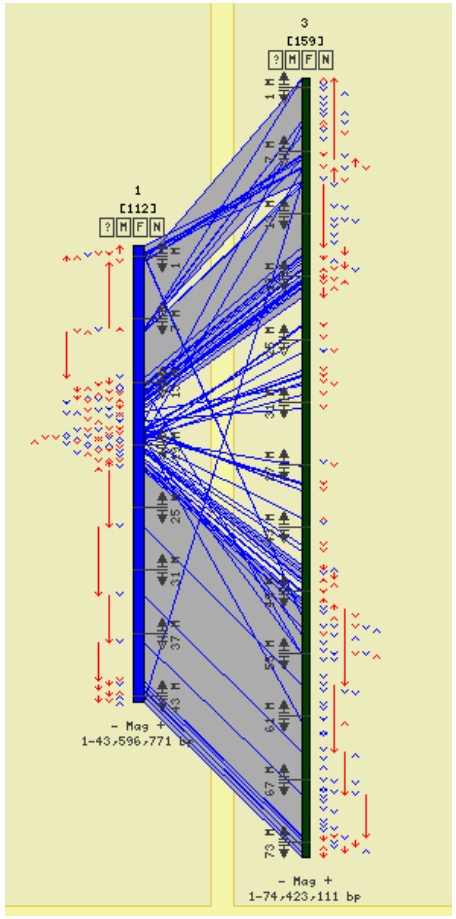




J. Craig Venter™ INSTITUTE

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

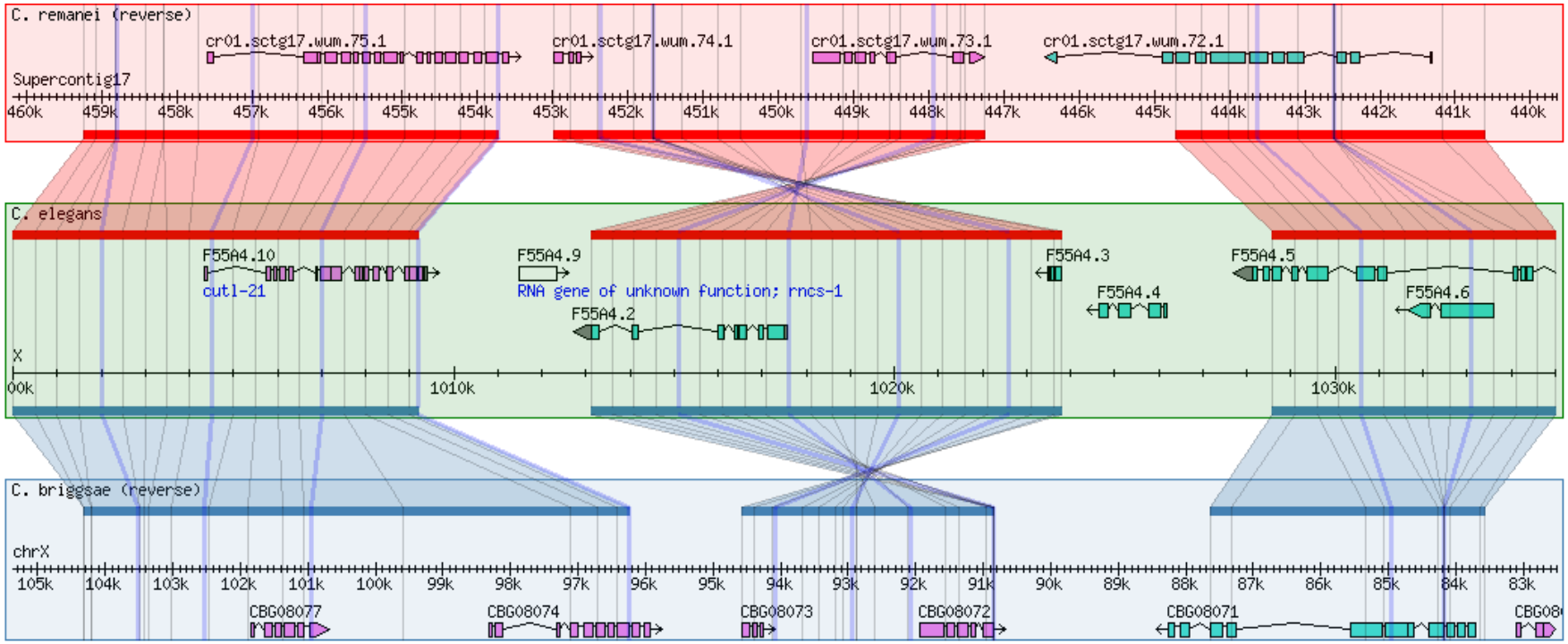
The Legume Information System

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



WormBase

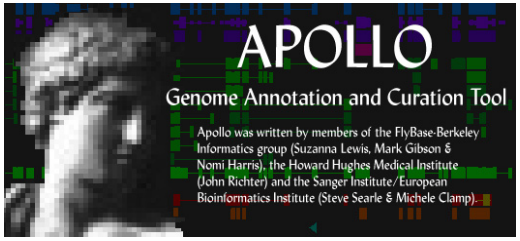
tair

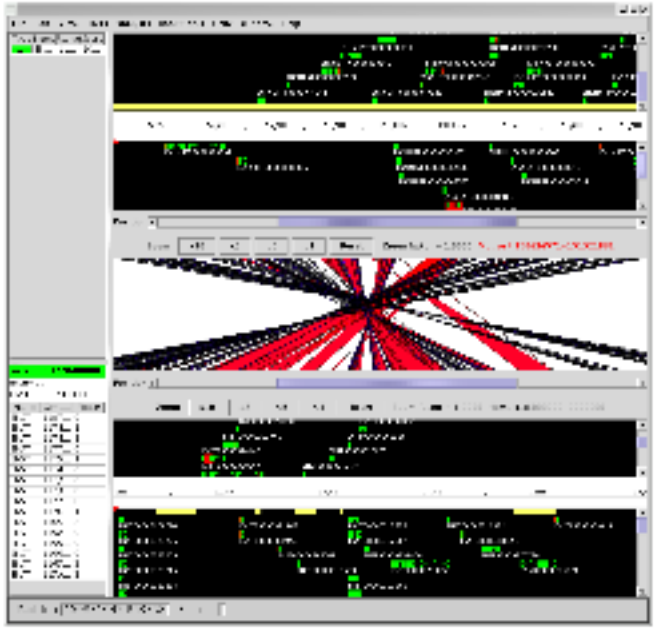
mod ENCODE

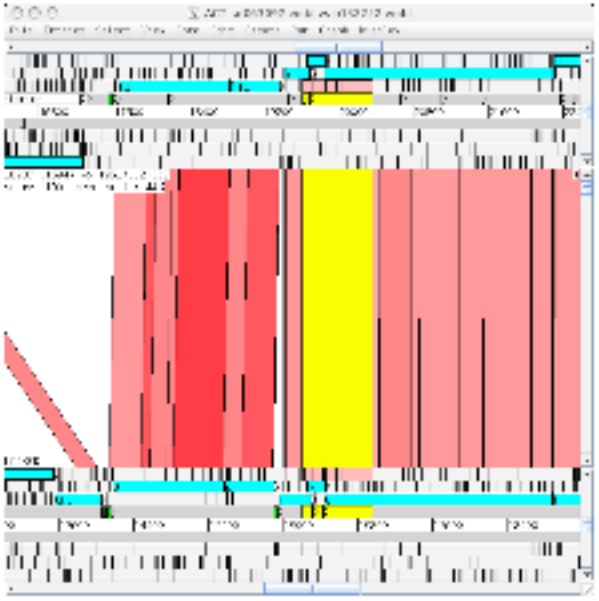
+others…

Pseudomonas Genome Database V2

Branding ideas...





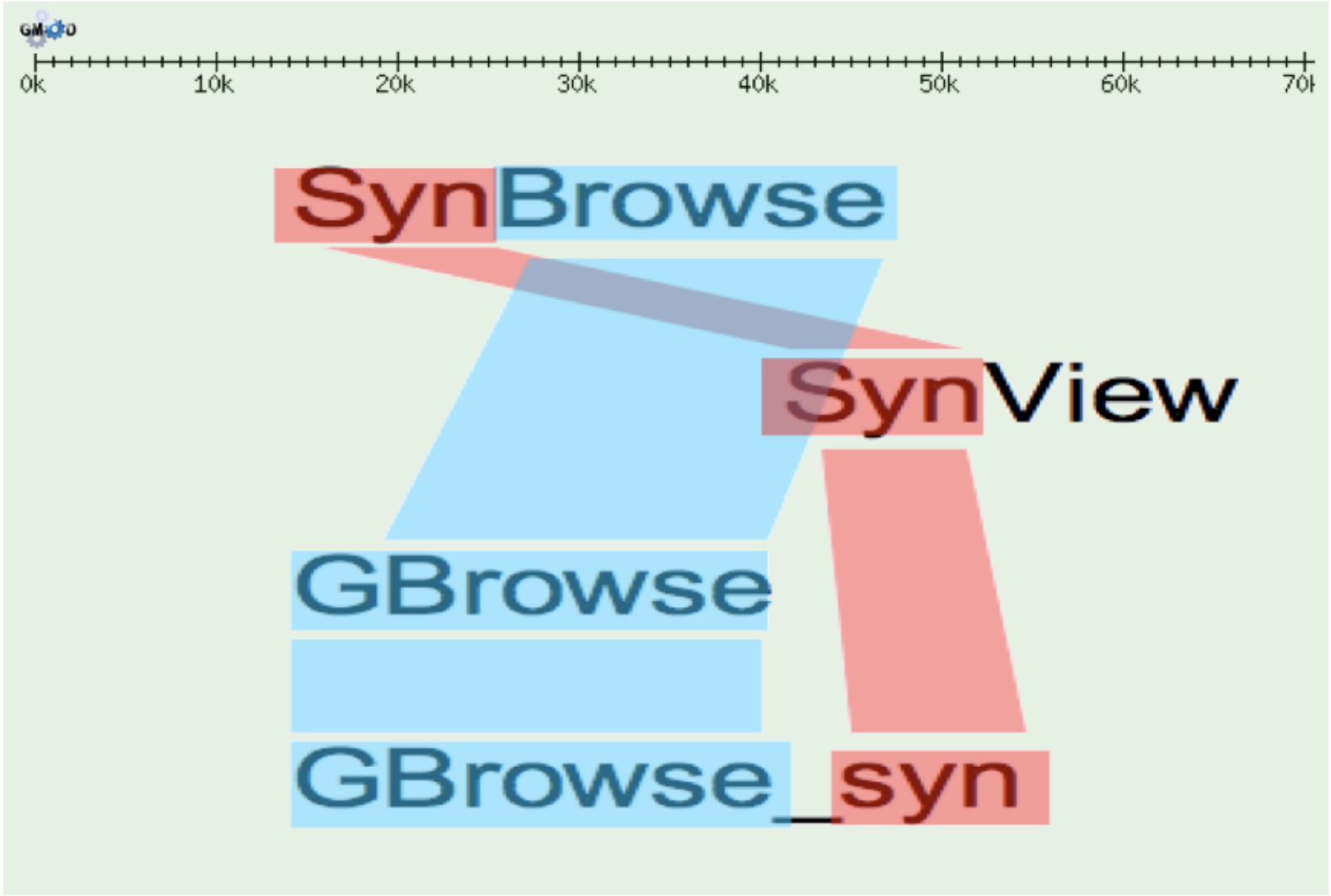


Desktop Synteny Viewers: Apollo and Artemis

Debating the relative merits of Apollo* and Artemis‡

*Apollo is better ‡Artemis is also better

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?

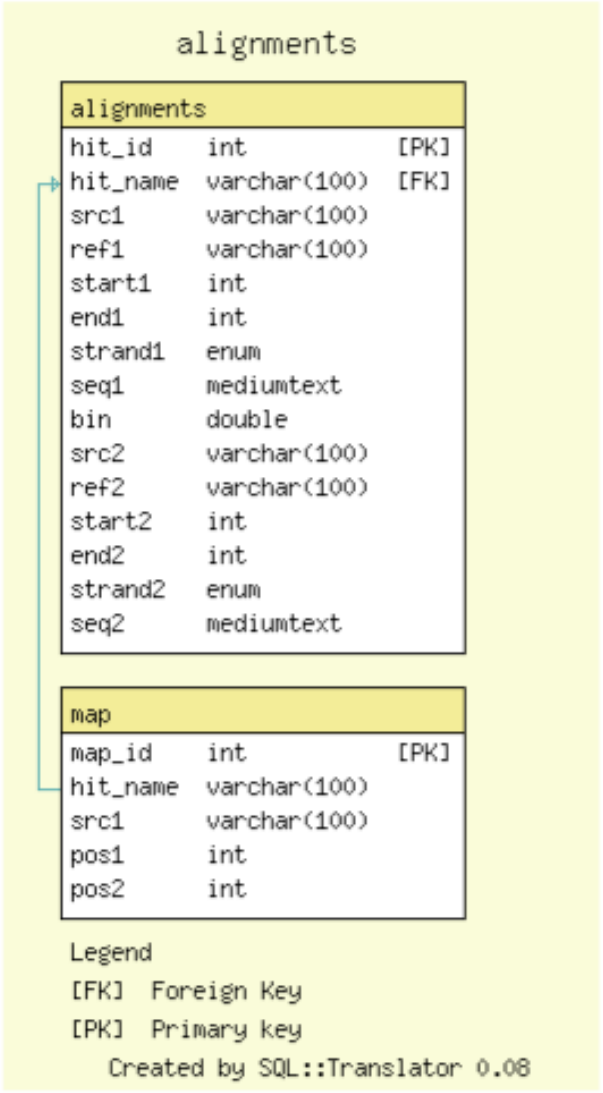
- Does not rely on perfect co-linearity across the entire displayed region (no orphan alignments)
- Offers on the fly alignment chaining
- No upward limit on the number of species
- Used grid lines to trace fine-scale sequence gain/loss
- Seamless integration with GBrowse data sources
- Ongoing support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

GBrowse_syn Architecture

[GBrowse]

[GBrowse]

Bio::DB::GFF
species1



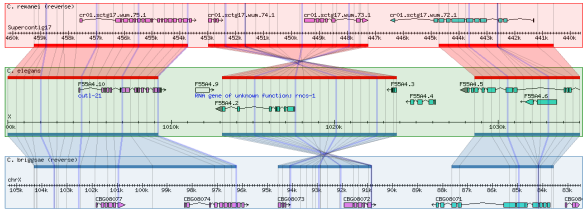
Bio::DB::GFF
species3

Bio::DB::GFF
species2

Bio::DB::GFF
species4

[GBrowse]

[GBrowse]



GBrowse_syn

GBrowse Databases*

GBrowse_syn alignment database

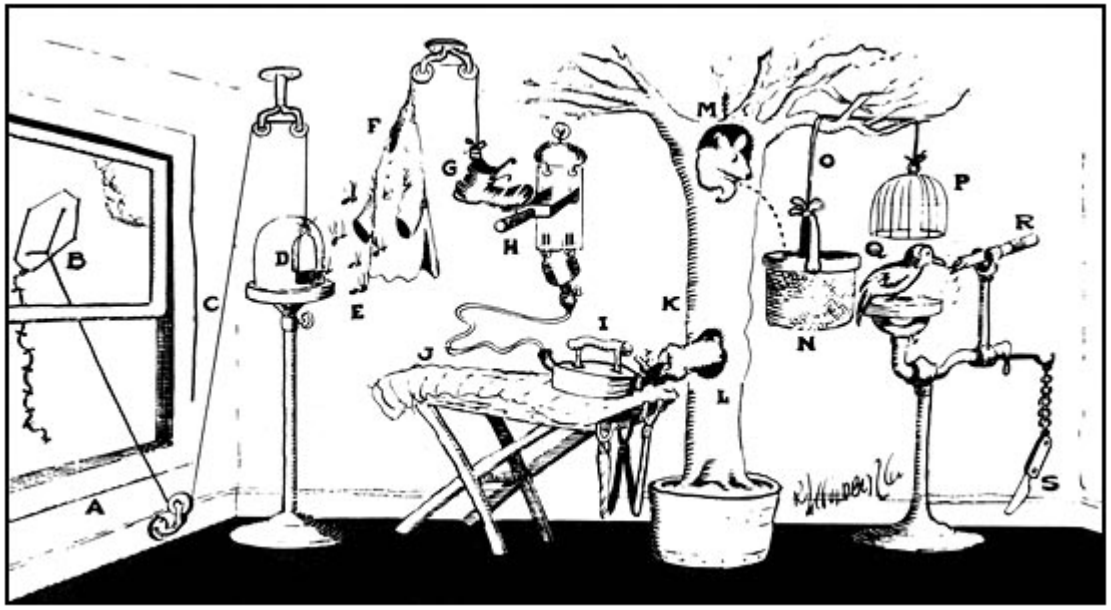
*.syn or *.conf

Species config.

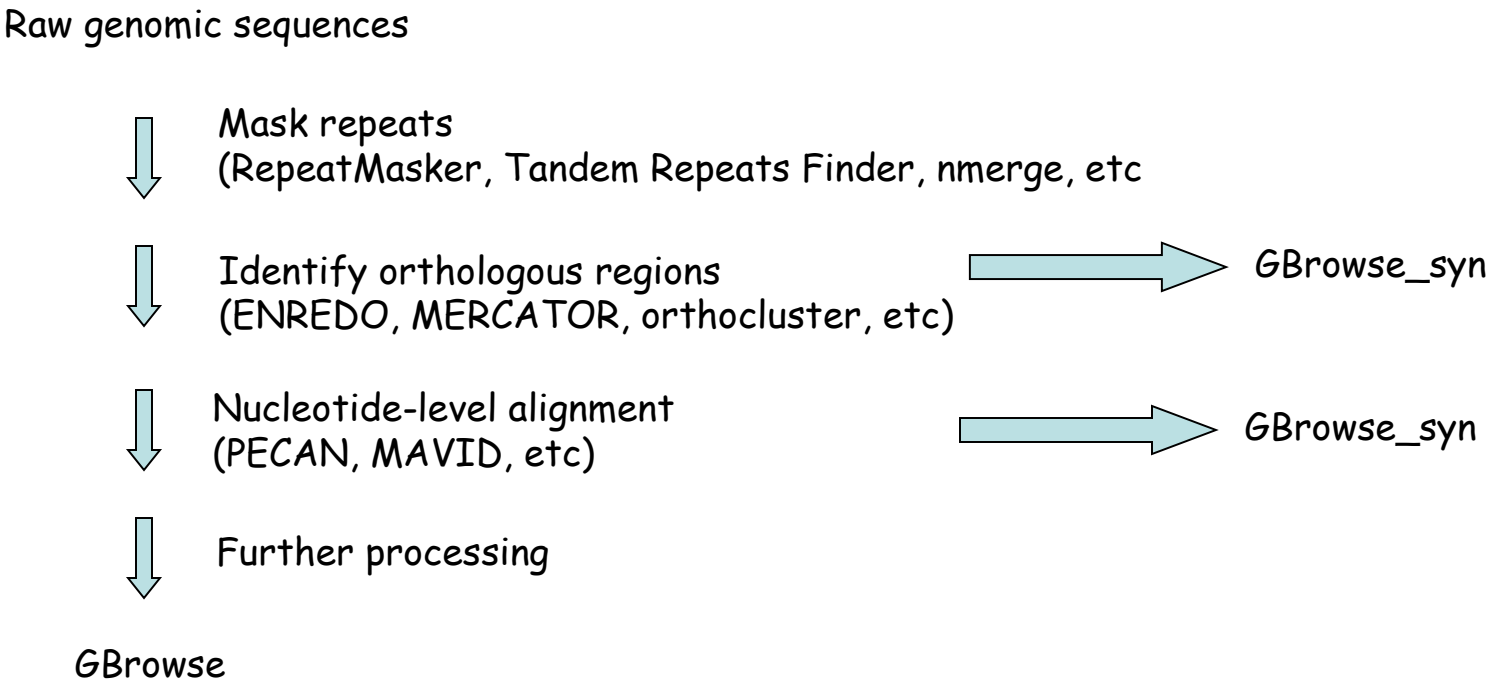
*.synconf

Master config.

Where do I get the data for GBrowse_syn?

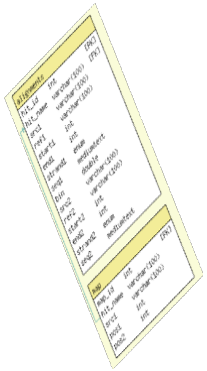


Hierarchical Genome Alignment Strategy

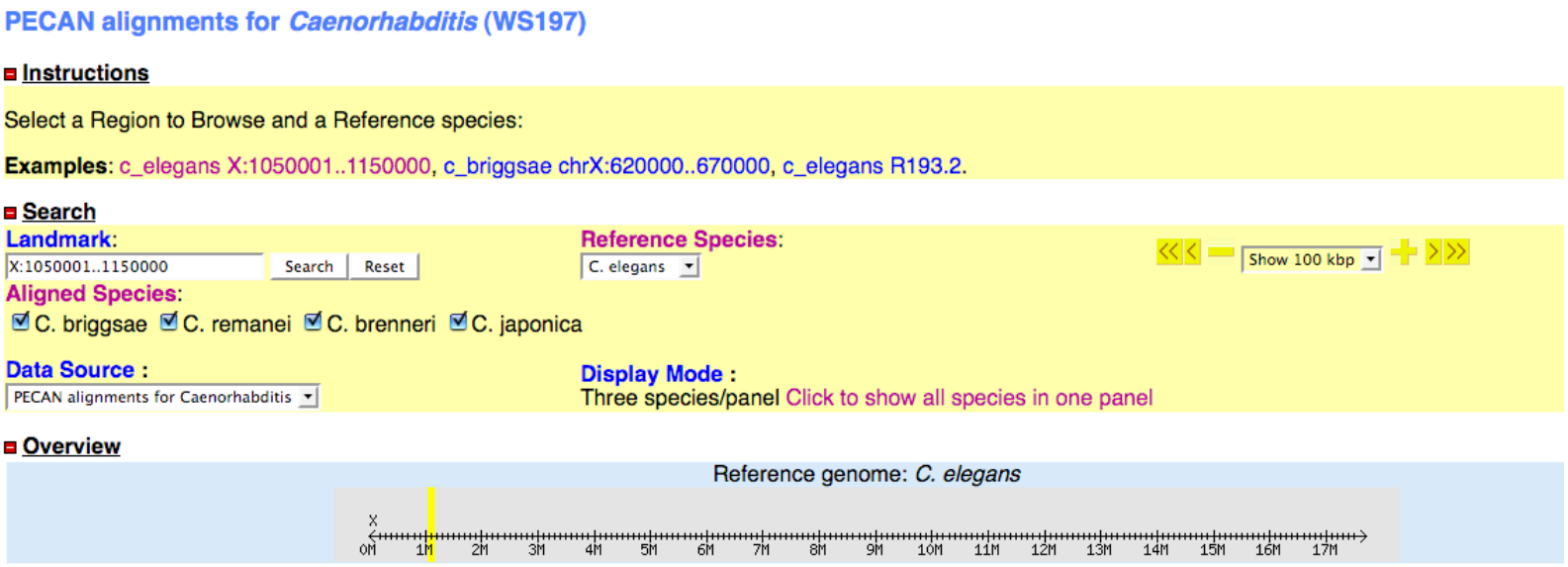


Getting Data into GBrowse_syn

CLUSTALW, FASTA, PECAN, MSF
SELEX, STOCKHOLM, GFF3,
TAB-DELIMITED,
etc…



GBrowse_syn interface



Gbrowse_syn: quick tour

Overview

Details

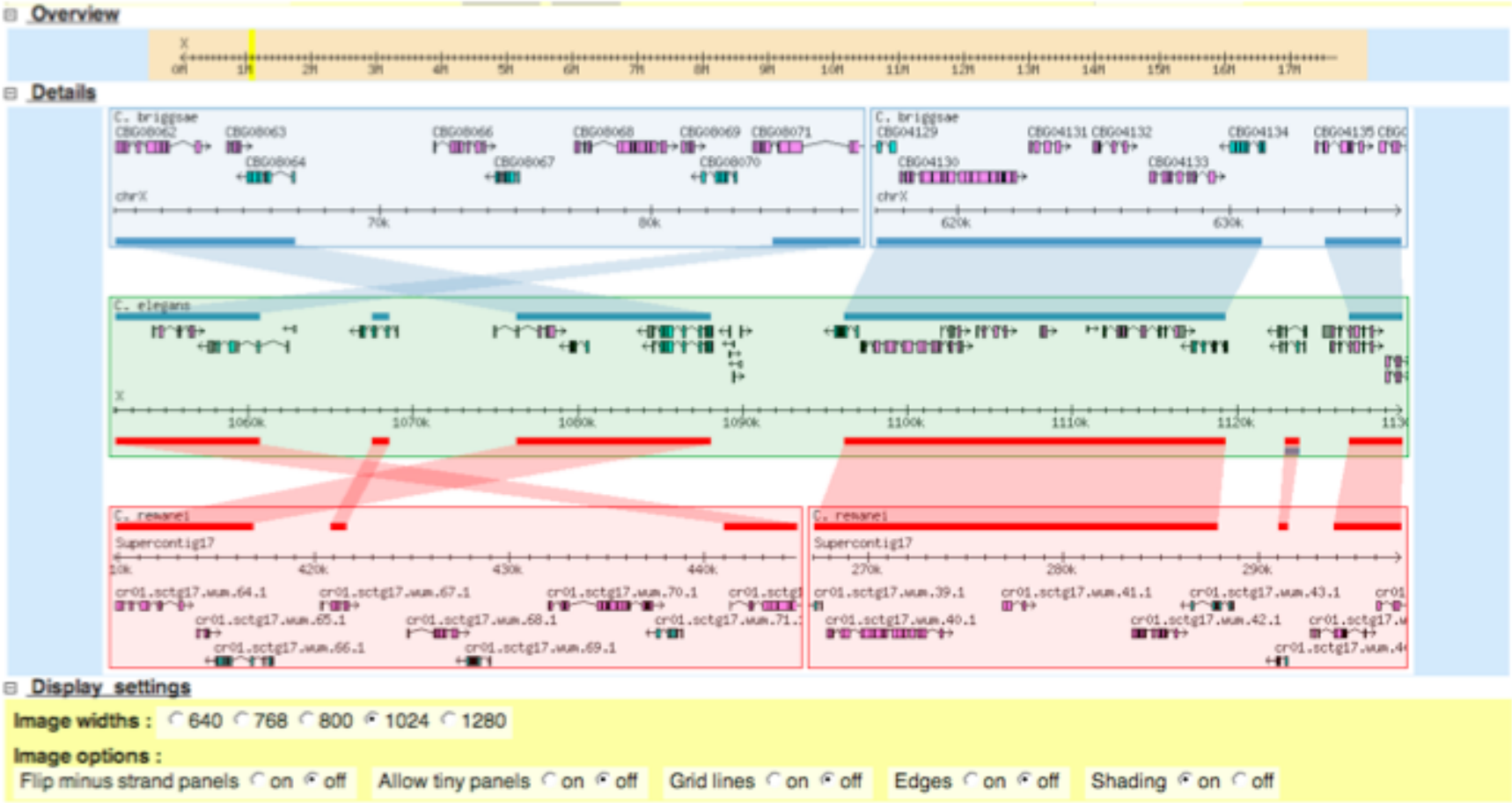
Display settings

Image widths : ○ 640 ○ 768 ○ 800 ◉ 1024 ○ 1280

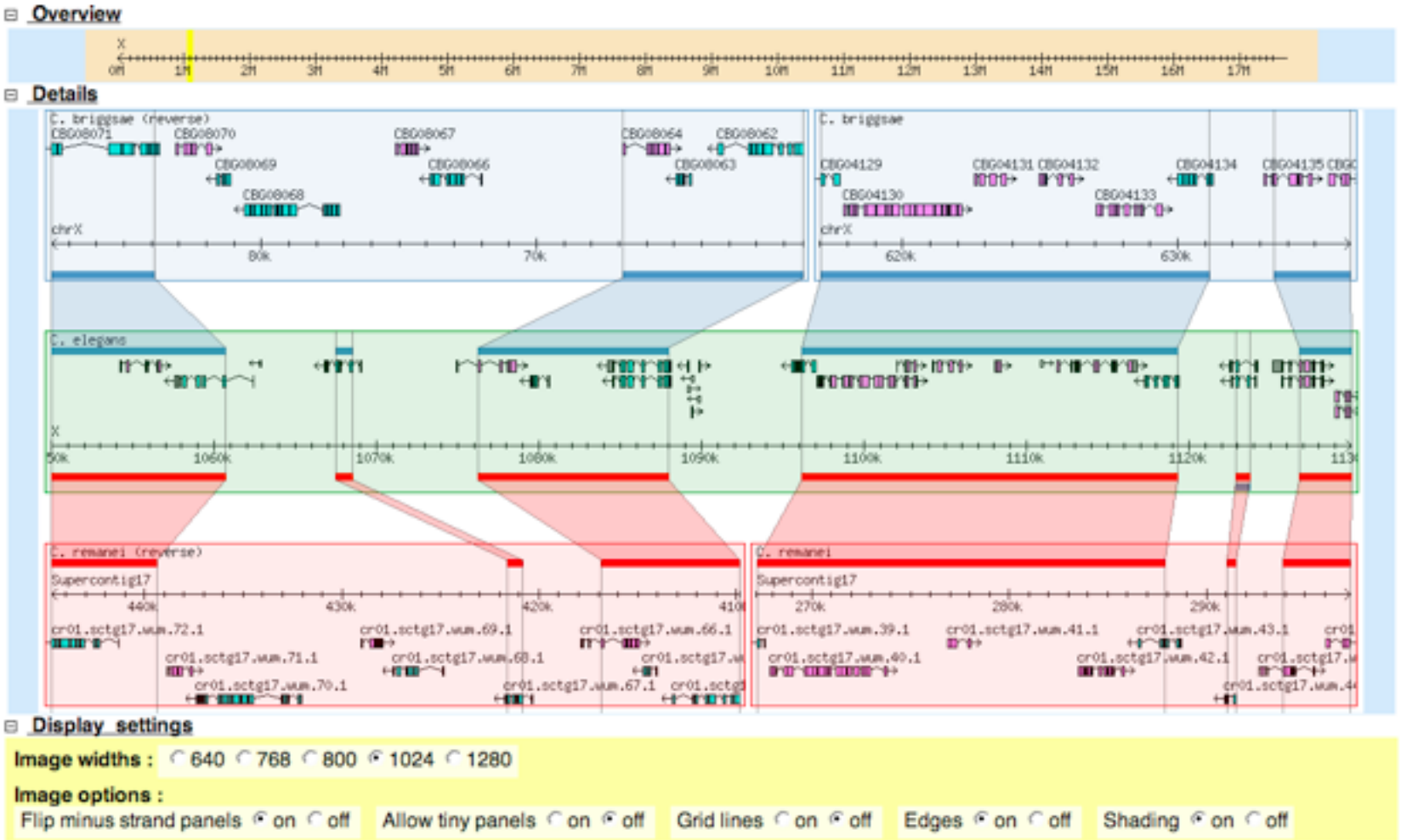
Image options :

Flip minus strand panels ○ on ◉ off | Allow tiny panels ○ on ◉ off | Grid lines ○ on ◉ off | Edges ○ on ◉ off | Shading ○ on ◉ off

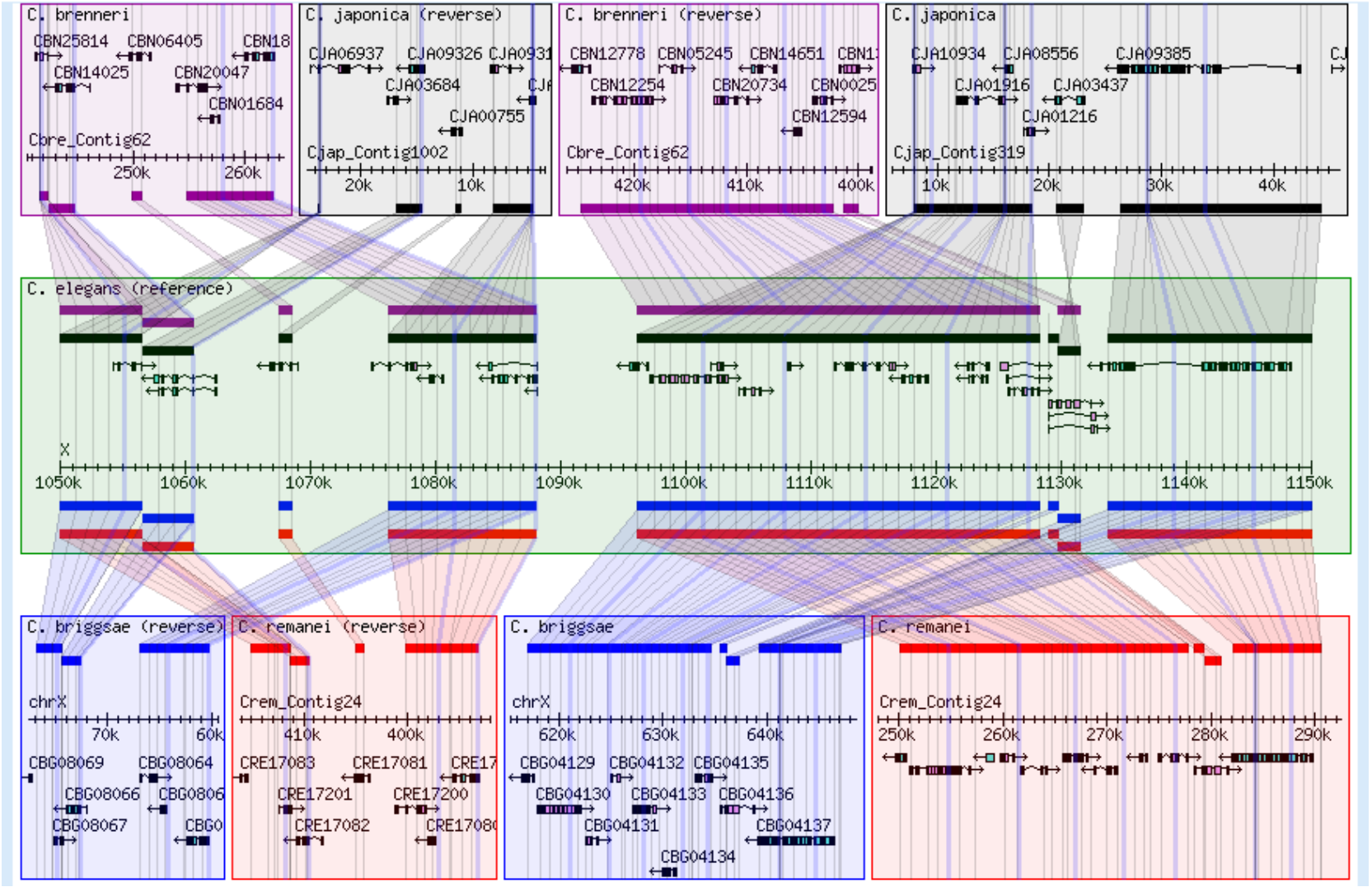
Gbrowse_syn: quick tour (shaded alignments)



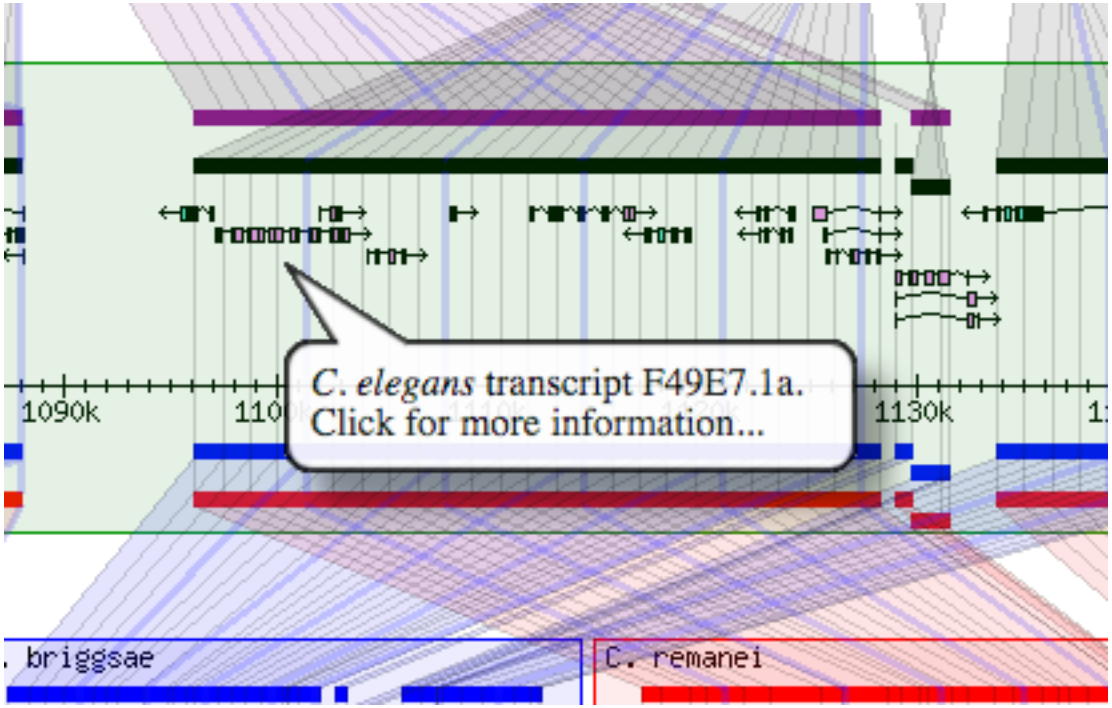
Gbrowse_syn: quick tour (strand correction)



Optional "All in one" view



Adding markup to the annotations



How to use Insertions/Deletion data

A

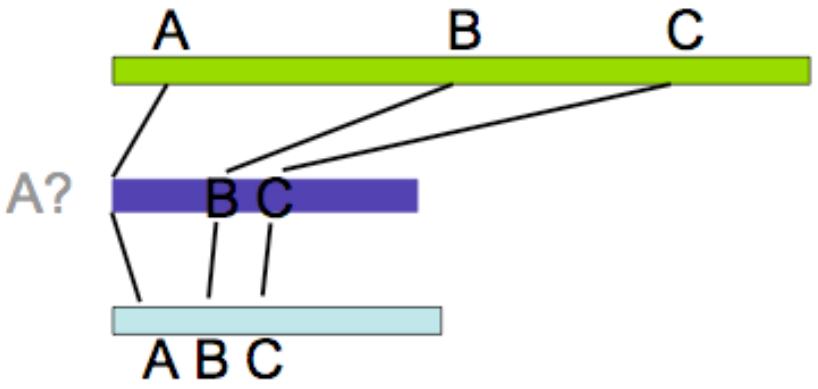
```
Ce-CHROMOSOME_I(+)/5195-16585 TGGCAAAAATATTTTGCATTTGCCGTTTTTCCCGTTTGCCGAAAAGTCTAATTTCGGTAA
Cb-chrI(-)/4091935-4097143    ------------------------------------------------------------
Cr-Contig8(+)/571990-577344   TTCGAAAC--------------------------------------------------
```

B

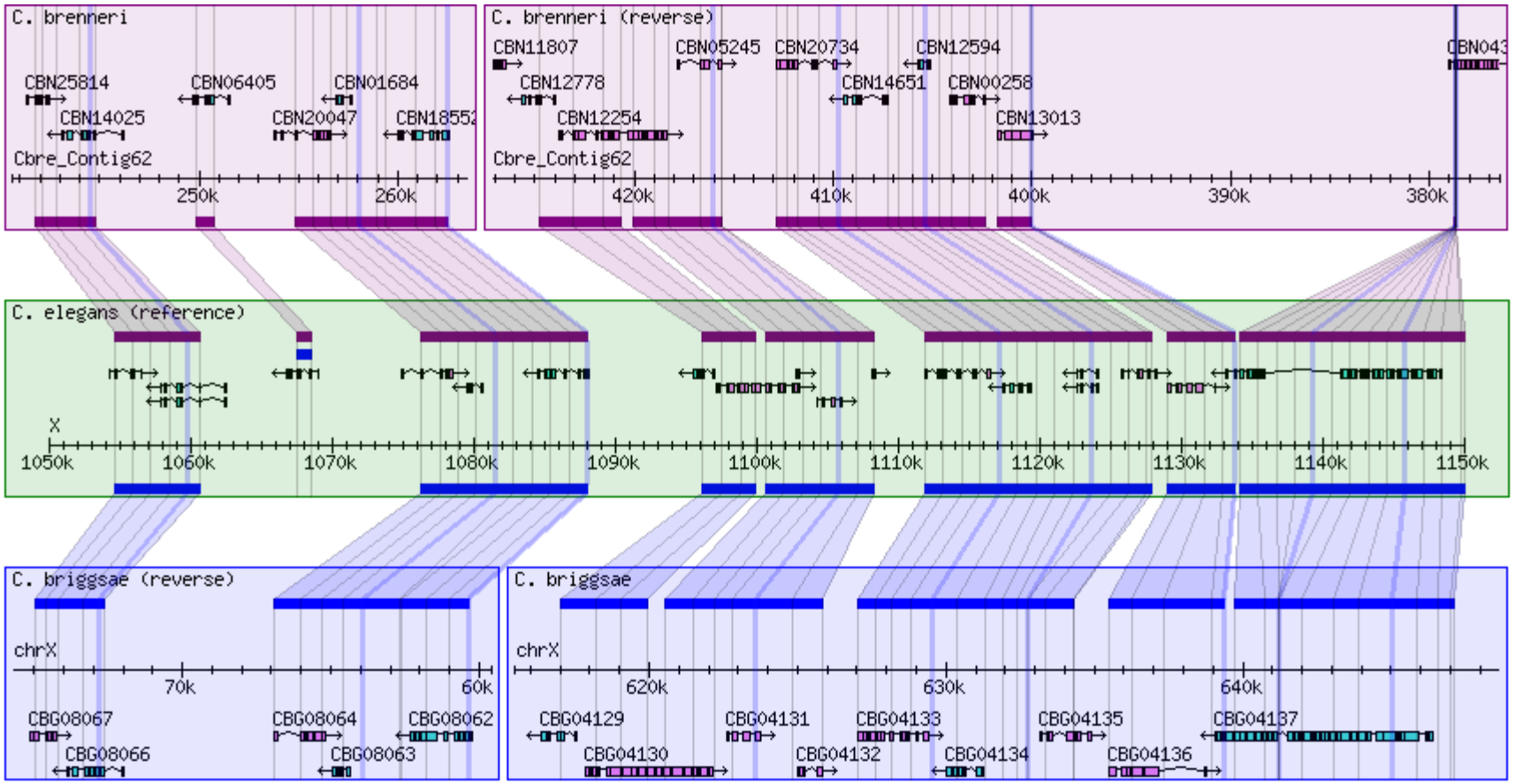
```
Ce-CHROMOSOME_I(+)/5195-16585 TTGGGCCATTTTTCGAAATTTTGAGCCACATAAAAAACTTTGAACCATTTTTGAGAAGTA
Cb-chrI(-)/4091935-4097143    --------------------AGAAGAATGTGAAGATCTTCA-----------------
Cr-Contig8(+)/571990-577344   --------------------CAGAGAAACAGAAACAATTTTA-----------------
                          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

C

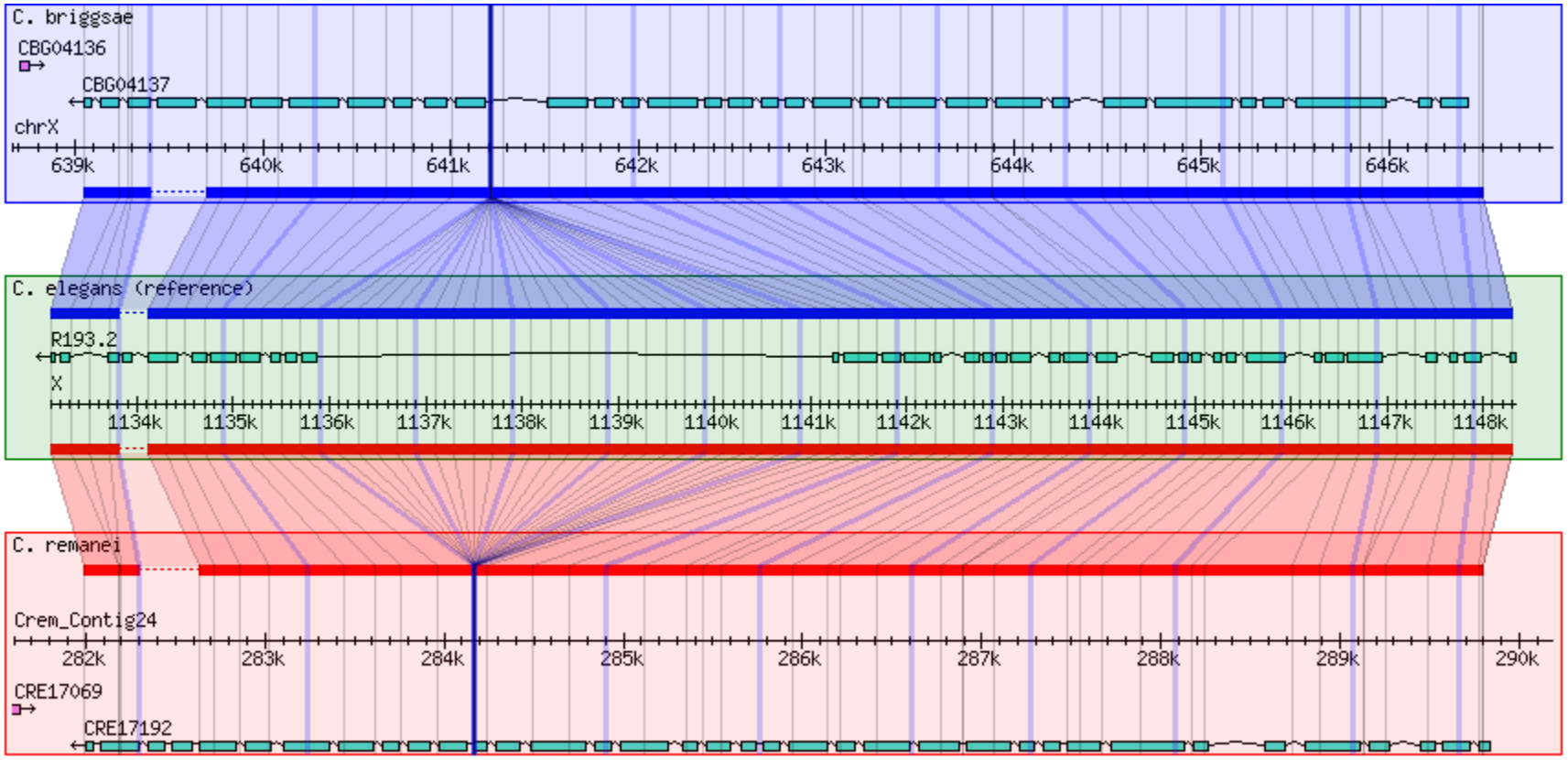
```
Ce-CHROMOSOME_I(+)/5195-16585 TTATTACGACATTCGTTTATTTGAGCACAATTTGGGCCTATACTTTCAAAATCGGGGTTT
Cb-chrI(-)/4091935-4097143    --TTCATGTCAA-------------------------------------------TCAT
Cr-Contig8(+)/571990-577344   --TTTCTGAAAACAGGTAGTATTATGGTTCCGAGGGTGTAGGGTTCGAAACCGGGCCTAG
                        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



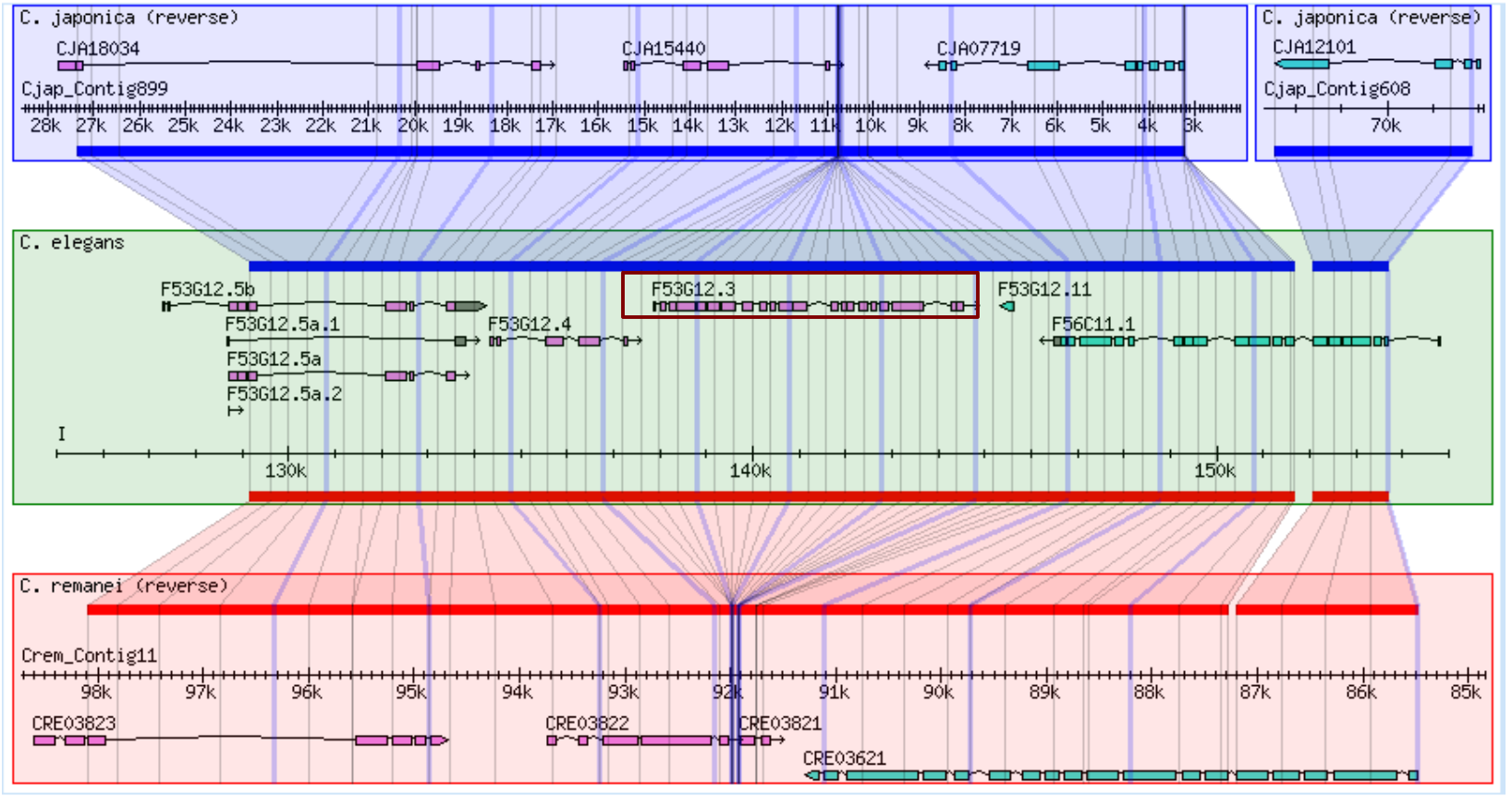
Tracking Indels with grid lines



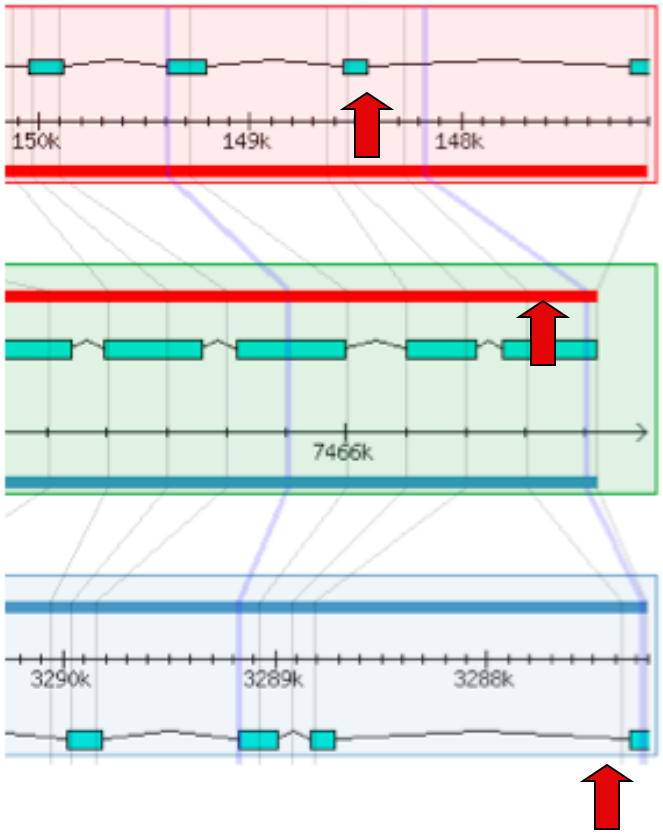
Evolution of Gene Structure



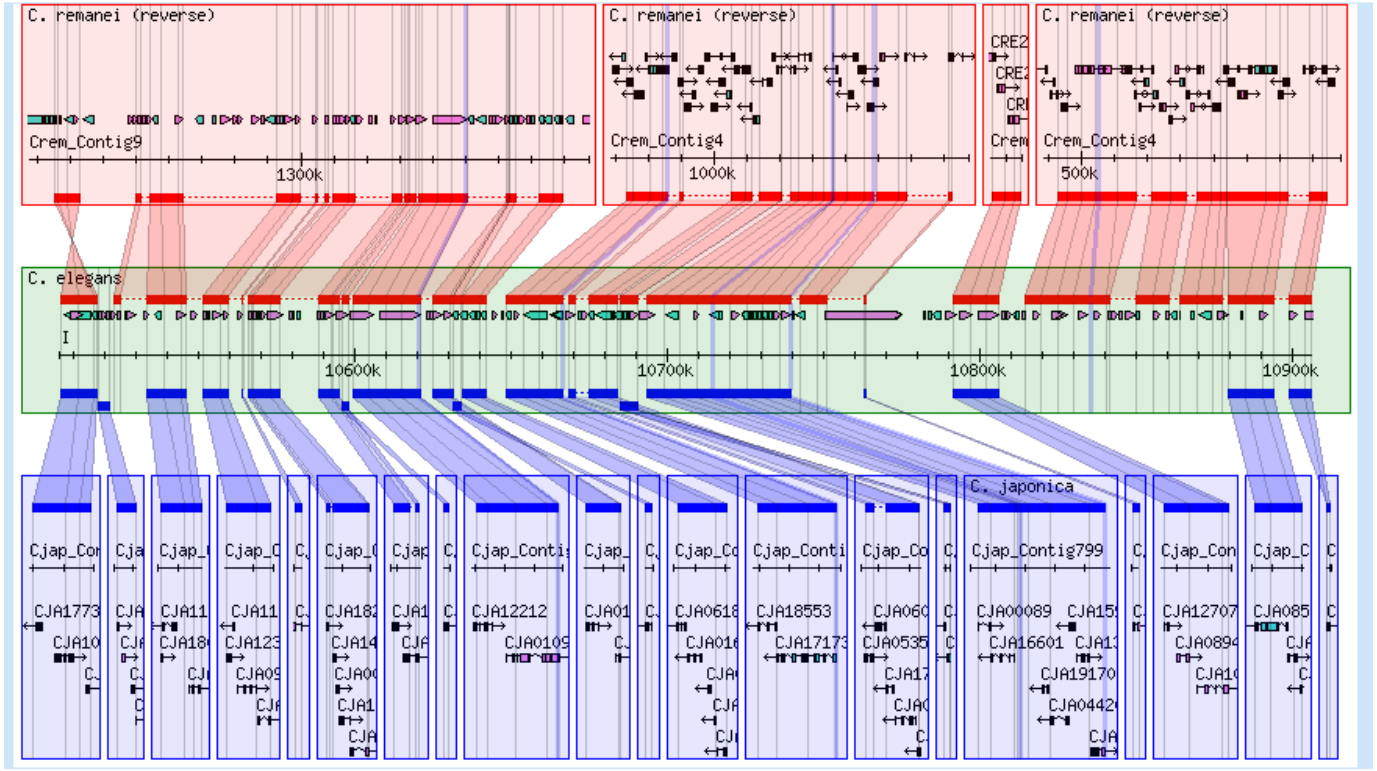
Putative gene or loss



Comparing gene models



Comparing assemblies

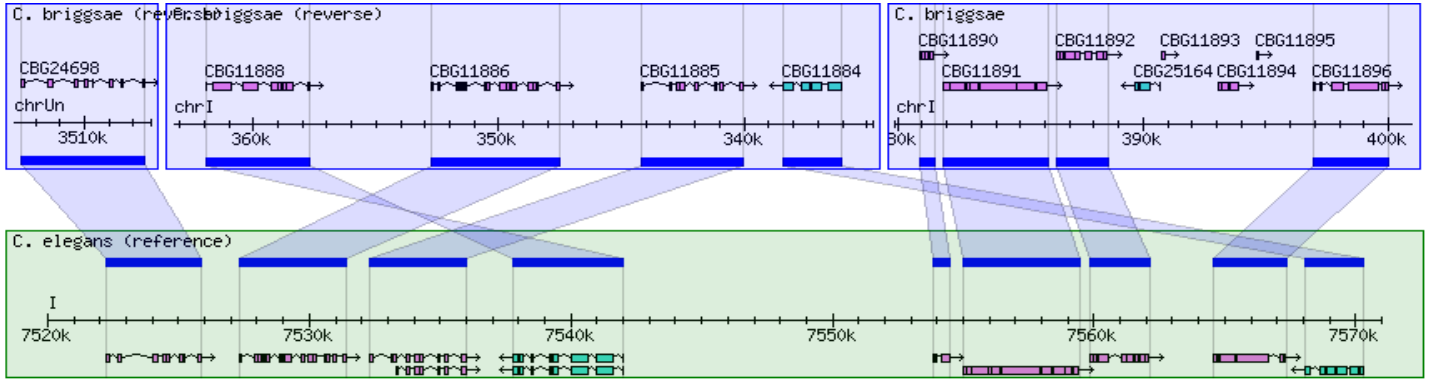


Not bad

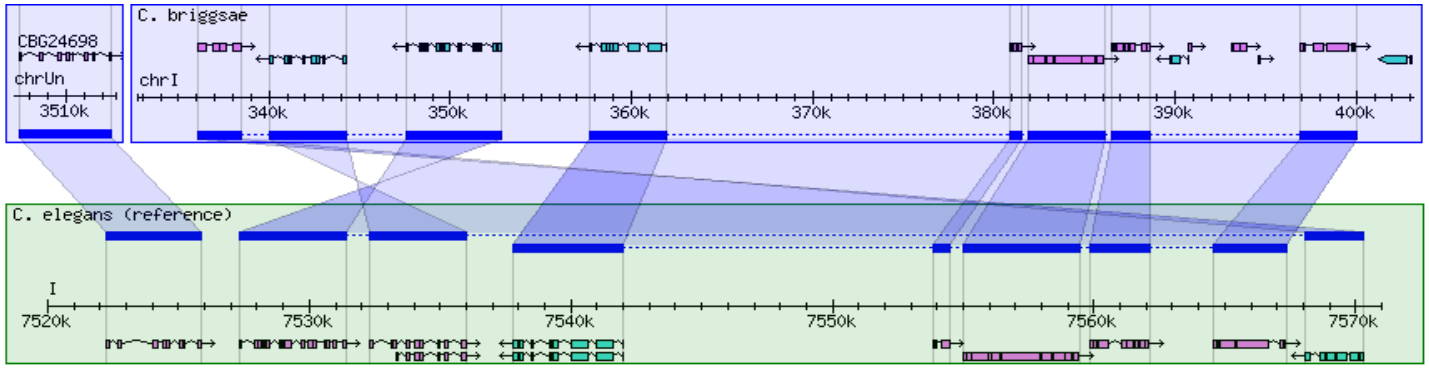
Needs work

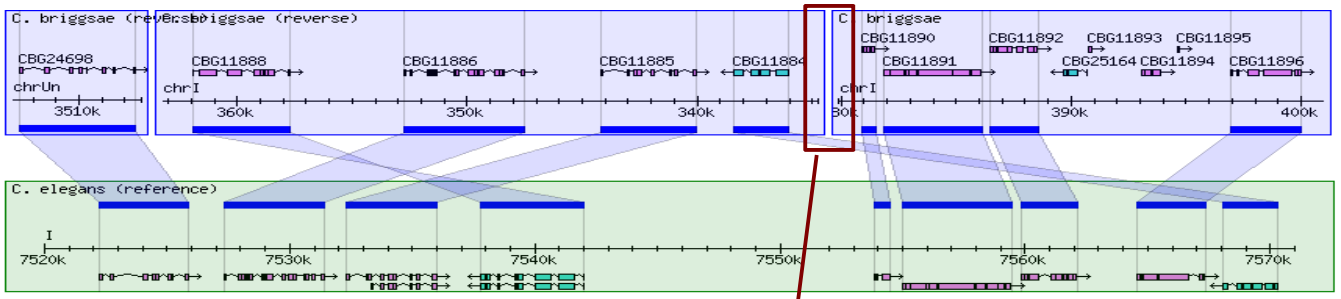
Getting the most out of small aligned regions or orthology-only data

Gene Orthology



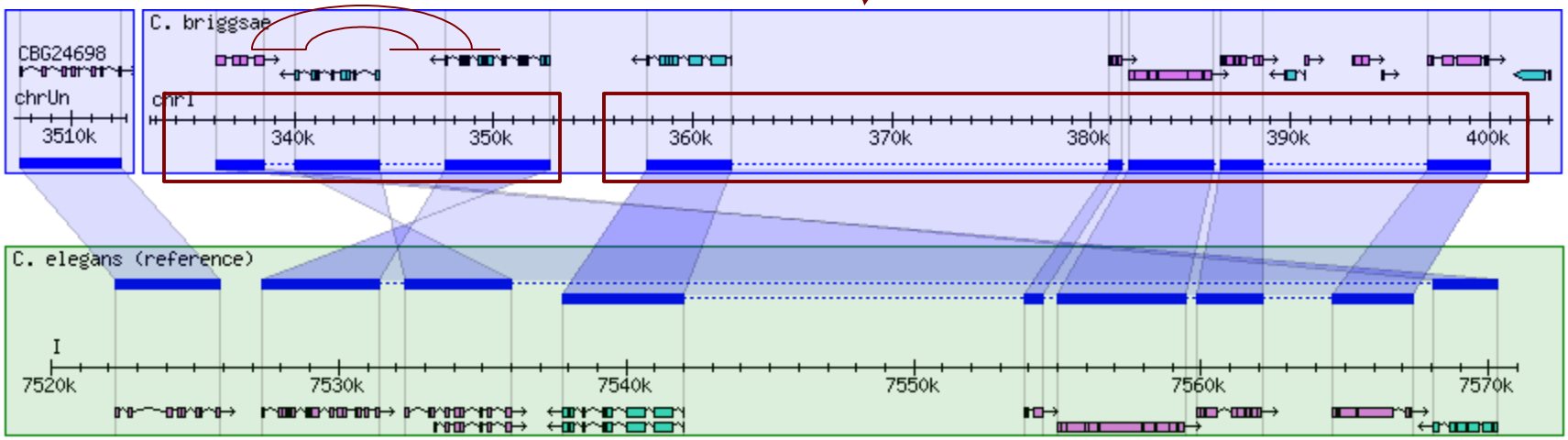
Chained Orthologs





2 panels merged

Inversion + translocation?



What about synteny blocks that fall off the ends of the displayed reference sequence?

Solution 1 : With multiple sequence alignment data, calculate many anchor points (done anyway for grid lines)

Solution 2 : For orthology-based synteny blocks, use individual start and end coordinates of orthologs as anchor points.

Solution 3: If all else fails, guess the end of the target block based on the overall length ratio.

length displayed target = (length target/length reference)* length displayed reference

What if the aligned DNA sequences are too distant?

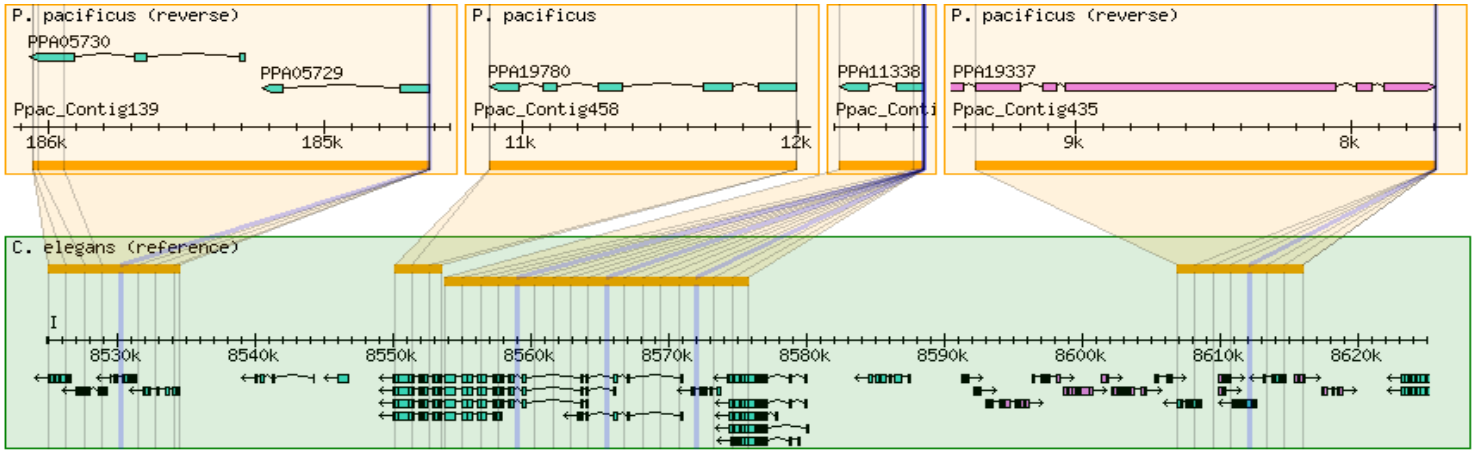
!=

*

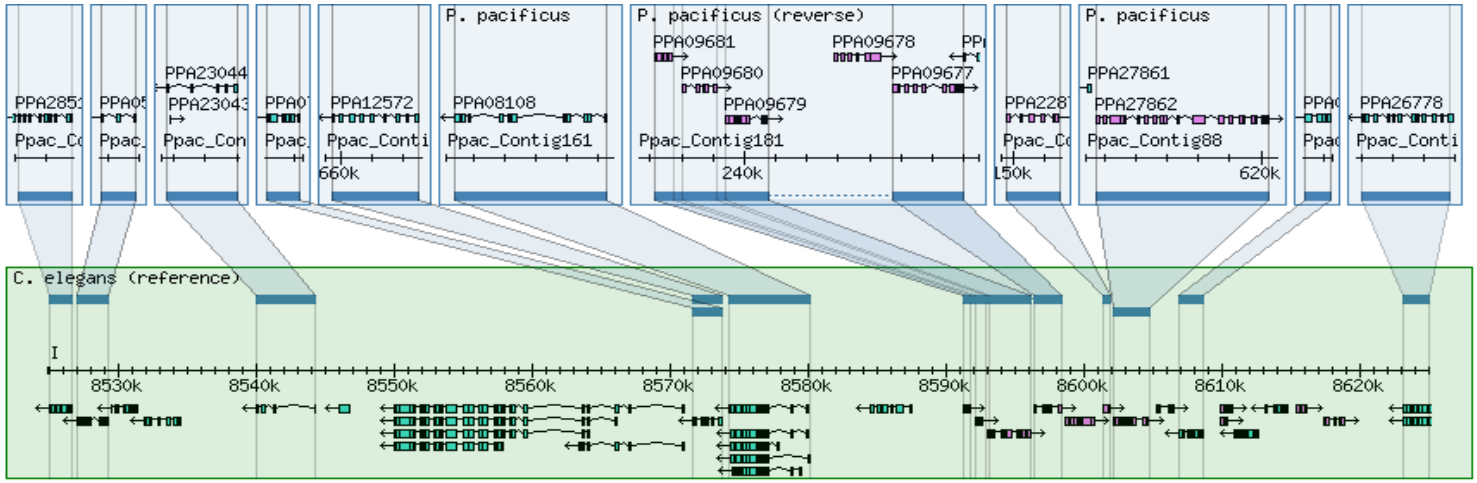
* For our continental European friends:

!=

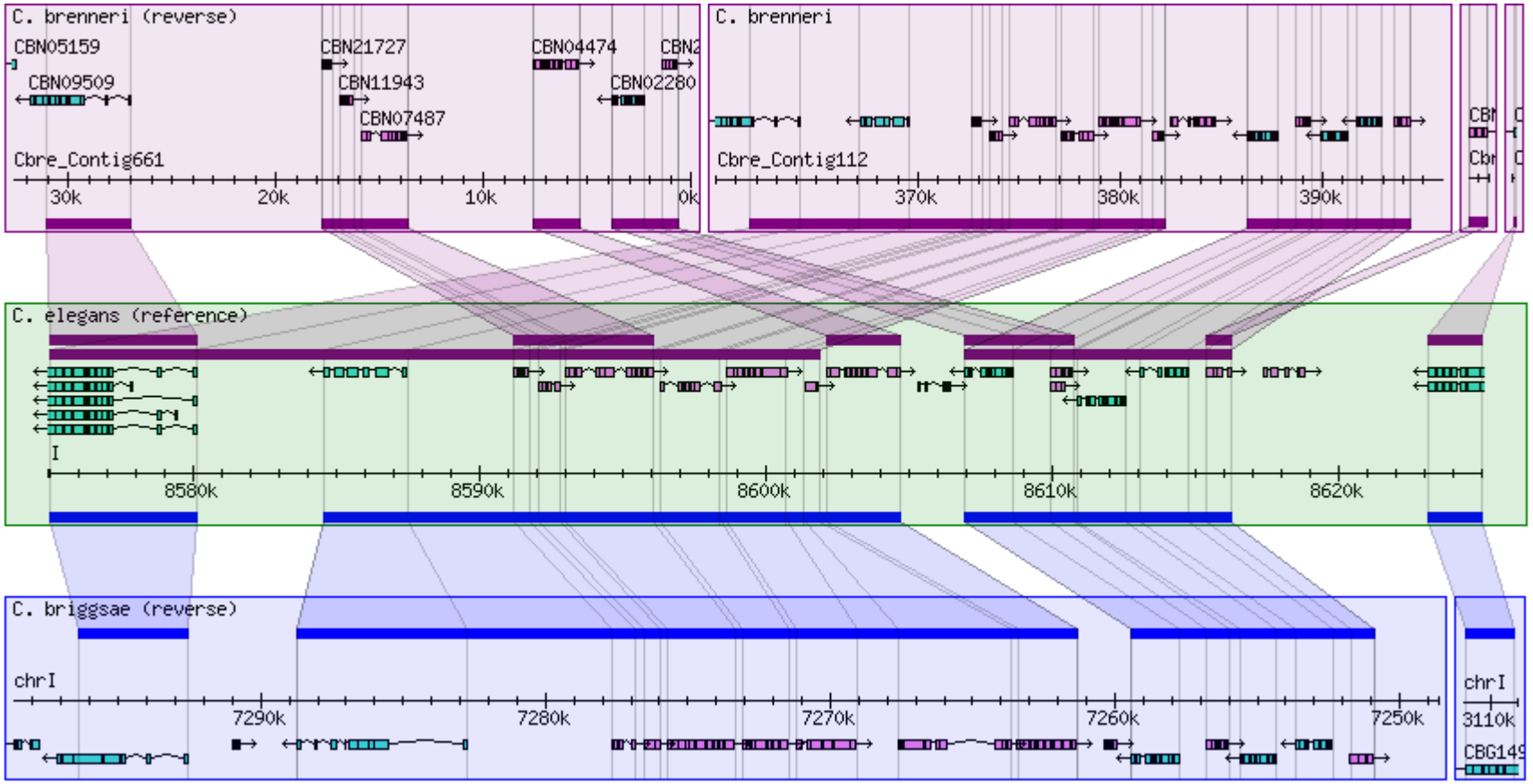
Pecan alignments



Protein orthology based Synteny blocks



What about segmental duplications?


C. brenneri (reverse)
CBN05159
CBN09509
CBN21727
CBN11943
CBN07487
CBN04474
CBN02280
Cbre_Contig661
30k
20k
10k
0k
C. brenneri
Cbre_Contig112
370k
380k
390k
C. elegans (reference)
I
8580k
8590k
8600k
8610k
8620k
C. briggsae (reverse)
chrI
7290k
7280k
7270k
7260k
7250k
chrI
3110k

The Future of GBrowse_syn*

- Integration with GBrowse 2.0
- "On the fly" sequence alignment view
- AJAX-based user interface and navigation
- High-level graphical overviews
- Suggestions?

*GBrowse-syn has a future

Acknowledgments

Lincoln Stein
Dave Clements
Scott Cain
Jason Stajich
Bonnie Hurwitz
Eva Huala
Cynthia Lee
Jack Chen
Ismael Verga
Michael Han
WormBase Curators