

*Comparative Genome Browsing
with GBrowse_syn*

**Sheldon McKay,
Cold Spring Harbor Laboratory**

- A brief survey of synteny browsers
- A few challenges of rendering comparative data
- Comparative genome browsing with GBrowse_syn



What is Gbrowse?

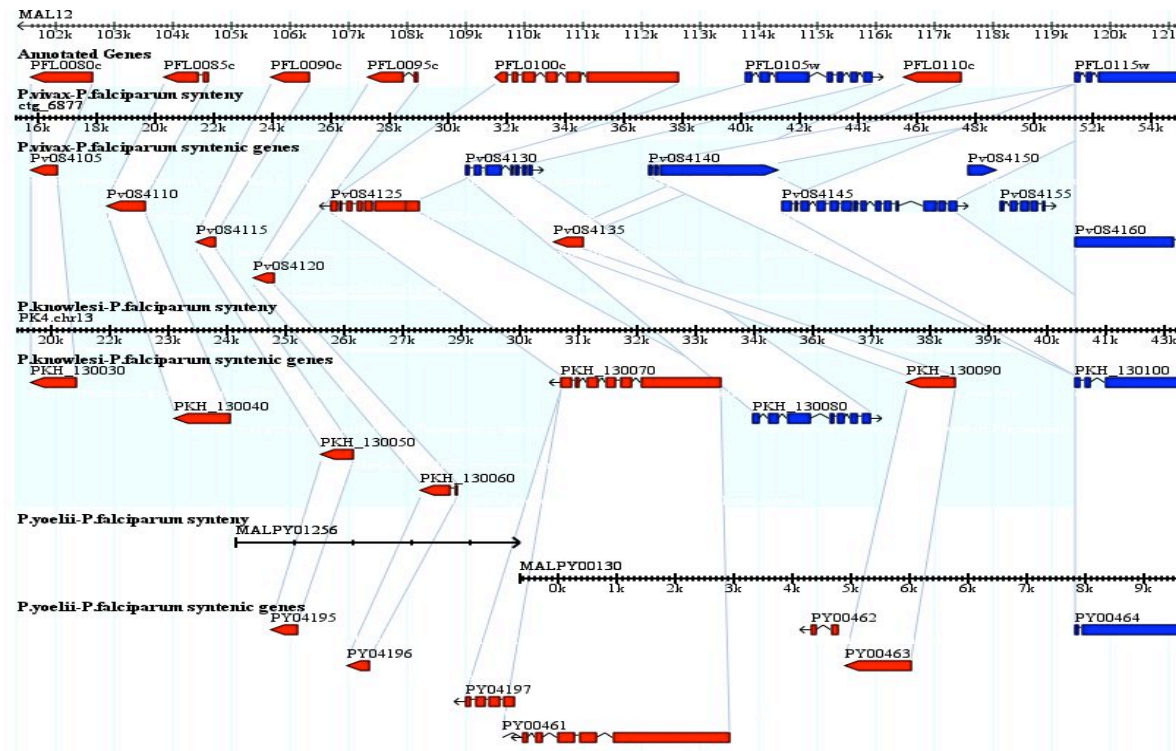
- Most popular software package from the Generic Model Organism Database (GMOD) project.
- Built on BioPerl and Bio::Graphics
- Relatively lightweight; designed for model organism databases (MODS) and groups with limited bioinformatics/IT support to display their own genomes

What is a Synteny Browser?

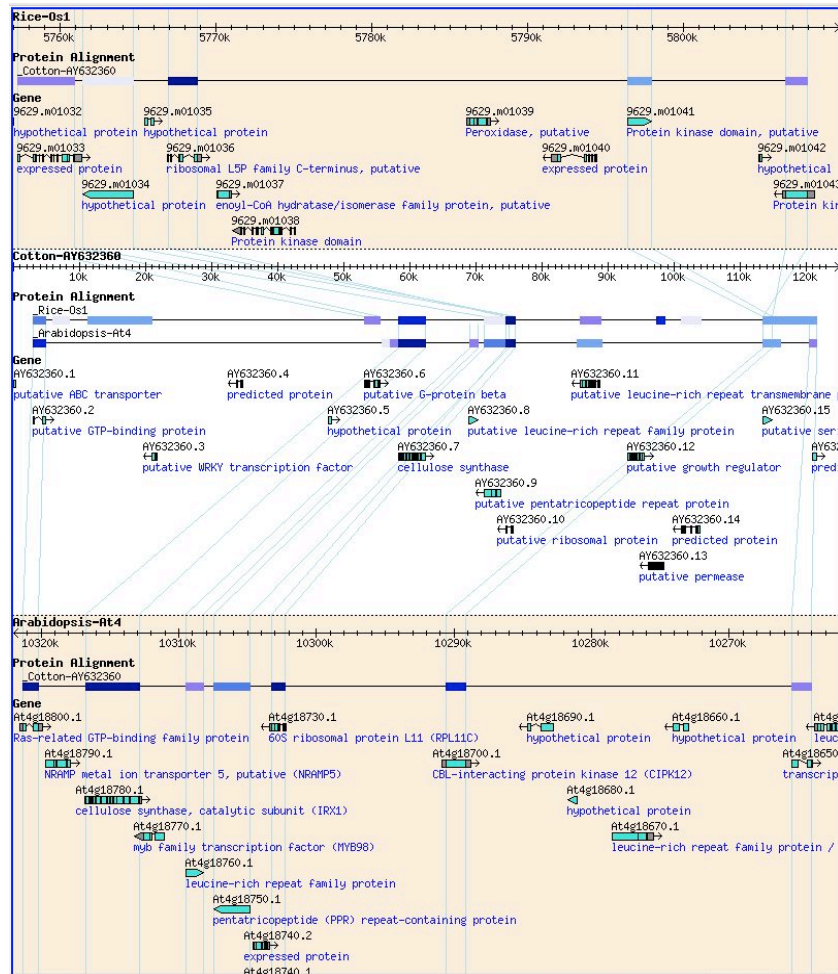
- Has display elements in common with genome browsers
- Uses sequence alignment or orthology to highlight co-linear segments of different genomes, strains, etc.
- Usually displays co-linearity relative to a reference genome.

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

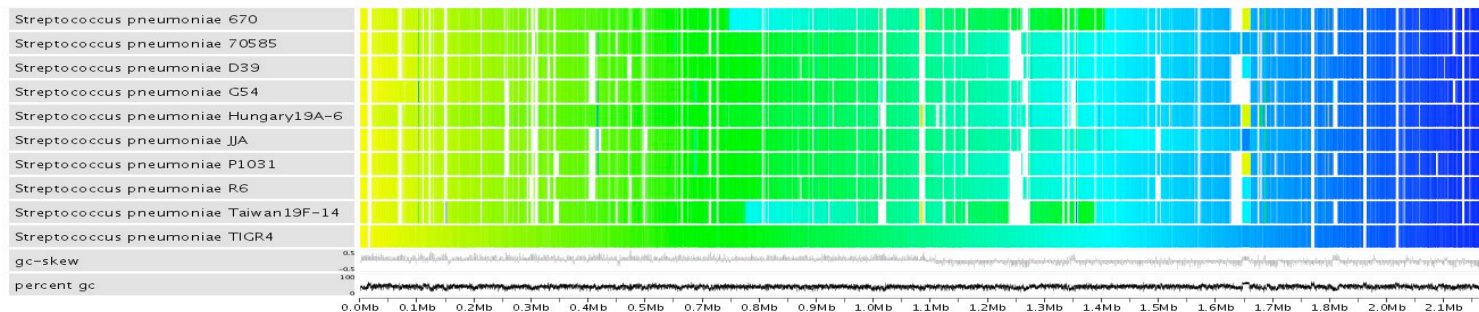
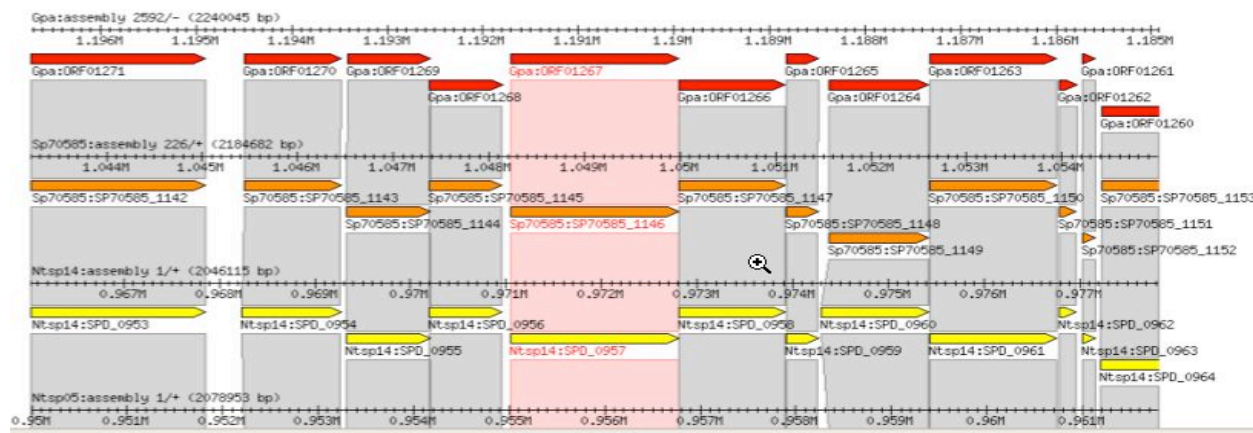


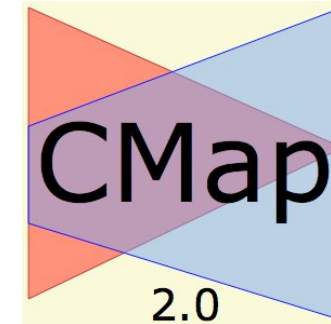
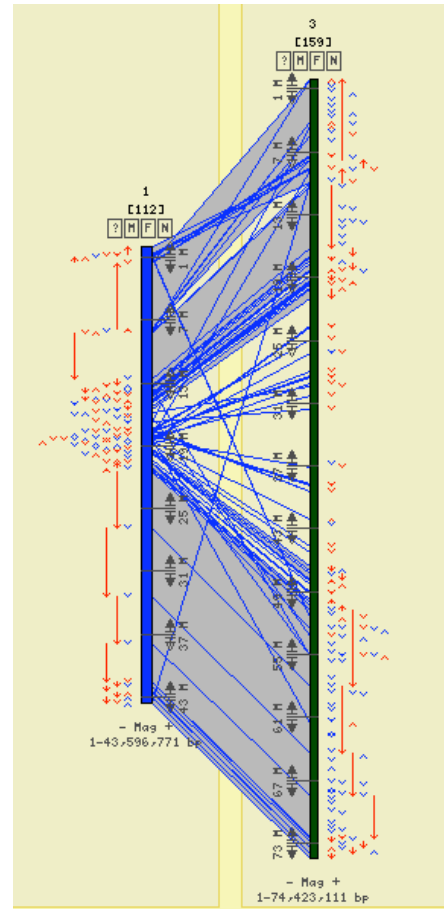
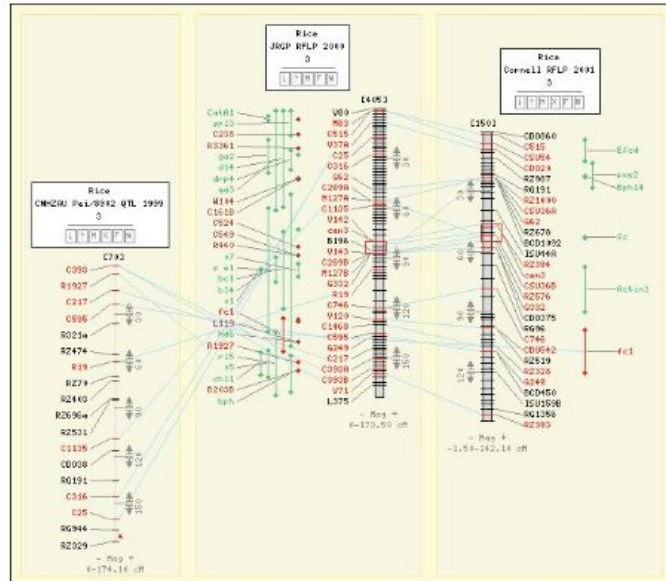
SynBrowse

...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

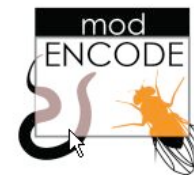
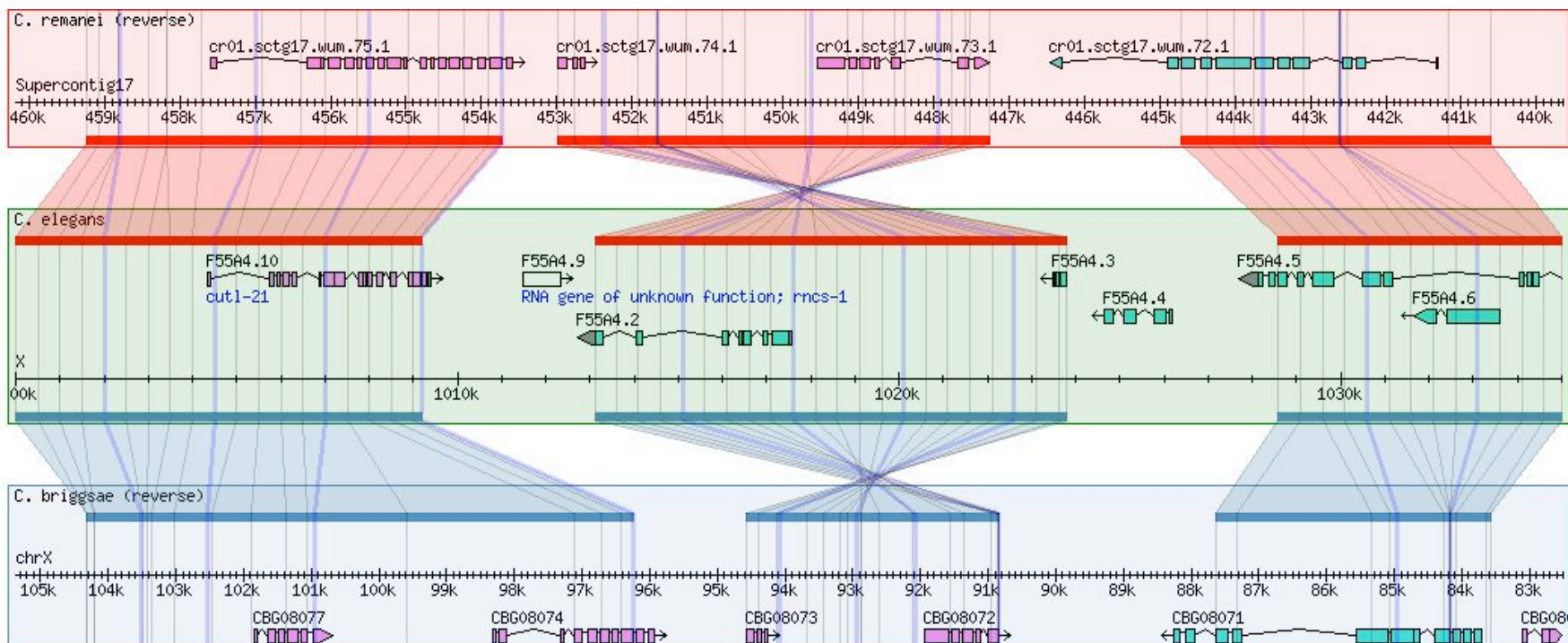
Sybil: Web-based software for comparative genomics





+ others...

GBrowse_syn





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 and overloaded 'source' tag
- 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



What is GBrowse_syn?

- Part of the **Generic Genome Browser Package (GBrowse)**
- A graphical multiple sequence alignment viewer
- Superimposes sequence alignment data on genes and other sequence features
- Compares two or more species to a central reference species

Hierarchical Genome Alignment Strategy

Raw genomic sequences



Mask repeats
(RepeatMasker, Tandem Repeats Finder, nmerge, etc)



Identify orthologous regions
(ENREDO, MERCATOR, orthocluster, etc)



GBrowse_syn



Nucleotide-level alignment
(PECAN, MAVID, etc)



GBrowse_syn



Further processing

GBrowse

```

Ca-CHROMOSOME_1(+)5195-16595 TTCTCTCAGATATTTTATAGAAATTACTGACTTTTCAGAAATAGATGAGCAATTTTG
Cb-chr1(-)4891925-4897143 -----
Cf-Contig8(+)571998-577344 ATGGTTTTGGTTTTGGAGCTGATTTTCGGGGTTTTTAAACGGGAAACAGAAATGTTT

Ca-CHROMOSOME_1(+)5195-16595 TTGTTTTAAAAATGAAATCTGAAATTTCCAAACAAAAACATGTCACACCAAGT
Cb-chr1(-)4891925-4897143 -----
Cf-Contig8(+)571998-577344 TGGTTTTCTGACTTCTATATCTGAAATTAAGCACCAGGACATTTGAAACTCGACAT

Ca-CHROMOSOME_1(+)5195-16595 TGGCAAAAATTTTGATTTCGGTTTTTCCTTTTCCTGGGAAAGTCAATTTTCGTAA
Cb-chr1(-)4891925-4897143 -----
Cf-Contig8(+)571998-577344 TTCGAAAC

Ca-CHROMOSOME_1(+)5195-16595 TTGGGCCATTTTCGAAATTTGAGCCAGCATAAAAAGTTTGACCAATTTTGGACAGTA
Cb-chr1(-)4891925-4897143 -----
Cf-Contig8(+)571998-577344 -----CAGAGAAACGAAACAATTTTA
* * * * *

Ca-CHROMOSOME_1(+)5195-16595 TTATTACGACATTCGTTTATTGAGCACAATTTGGCCCTATCTTCAAAATCGGGGTTT
Cb-chr1(-)4891925-4897143 -----TTCATGTCAA-----TCAT
Cf-Contig8(+)571998-577344 -----TTCTGAAACAGGTAGTATTATGTTCCGAGGGGTGTAGGGTTCCGAAACCGCCCTAG
* * *

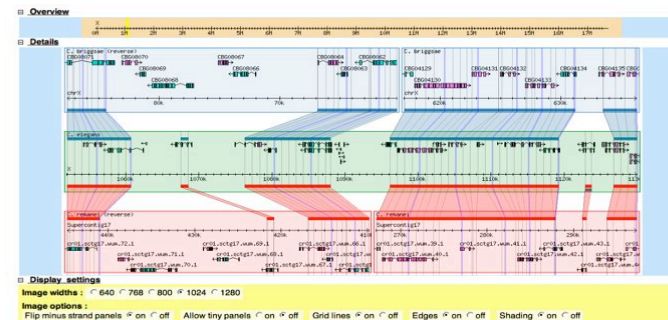
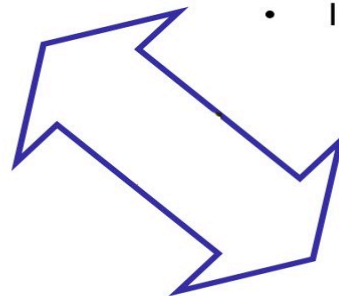
Ca-CHROMOSOME_1(+)5195-16595 GAAACCCCTATATGTTGAGCCGAAATGTTAACTCATAAAAAATTTGATGAAATAAAAT
Cb-chr1(-)4891925-4897143 -----CTAGCTCCCAATGATCACTCATAT-----ATT
Cf-Contig8(+)571998-577344 -----CGAACTTTTTGCATCTACTCTGGG-----TTT
* * * * *

Ca-CHROMOSOME_1(+)5195-16595 TTETACGGTCAATAAGATATAGCCCGGTCAGTCTCAAAATTTATAGATAGACACTTT
Cb-chr1(-)4891925-4897143 -----TCATCAACT-----
Cf-Contig8(+)571998-577344 TGGTACAACTCAACAGCCAGGGTTCGATCCCCACTGGTGGCCAACTCTTTTTATTTT
* * *

```

Goals

- More than two species
- Nucleotide-level resolution (gapped alignments)
- High-level resolution (synteny)
- Intuitive graphical rendering



GBrowse-like interface

PECAN alignments for *Caenorhabditis* (WS197)

Instructions

Select a Region to Browse and a Reference species:

Examples: [c_elegans X:1050001..1150000](#), [c_briggsae chrX:620000..670000](#), [c_elegans R193.2](#).

Search

Landmark:

X:1050001..1150000

Reference Species:

Aligned Species:

C. briggsae *C. remanei* *C. brenneri* *C. japonica*

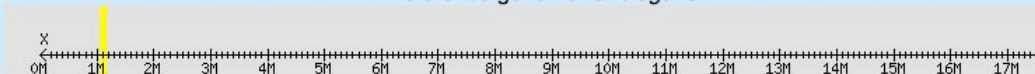
Data Source :

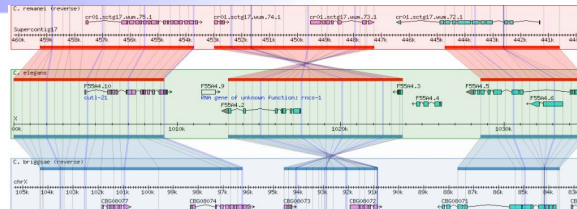
Display Mode :

Three species/panel [Click to show all species in one panel](#)

Overview

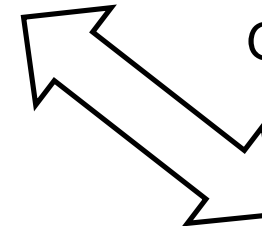
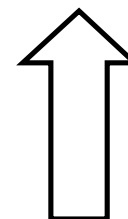
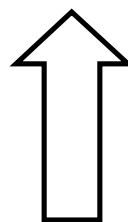
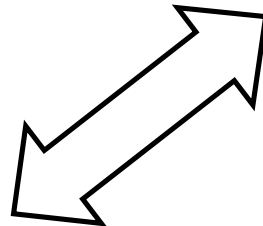
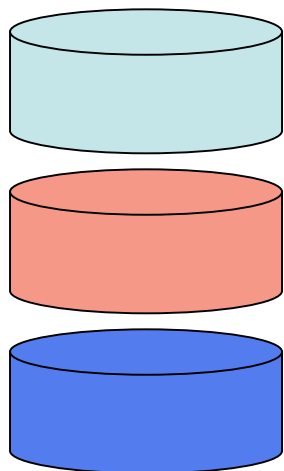
Reference genome: *C. elegans*



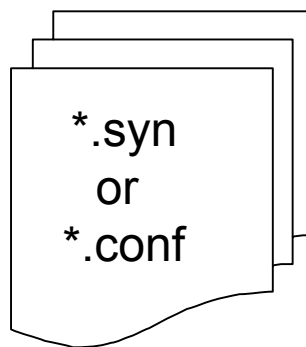


GBrowse_syn

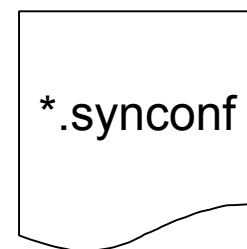
GBrowse
Databases*



GBrowse_syn
alignment
database



Species config.



Master config.

GBrowse_syn Architecture

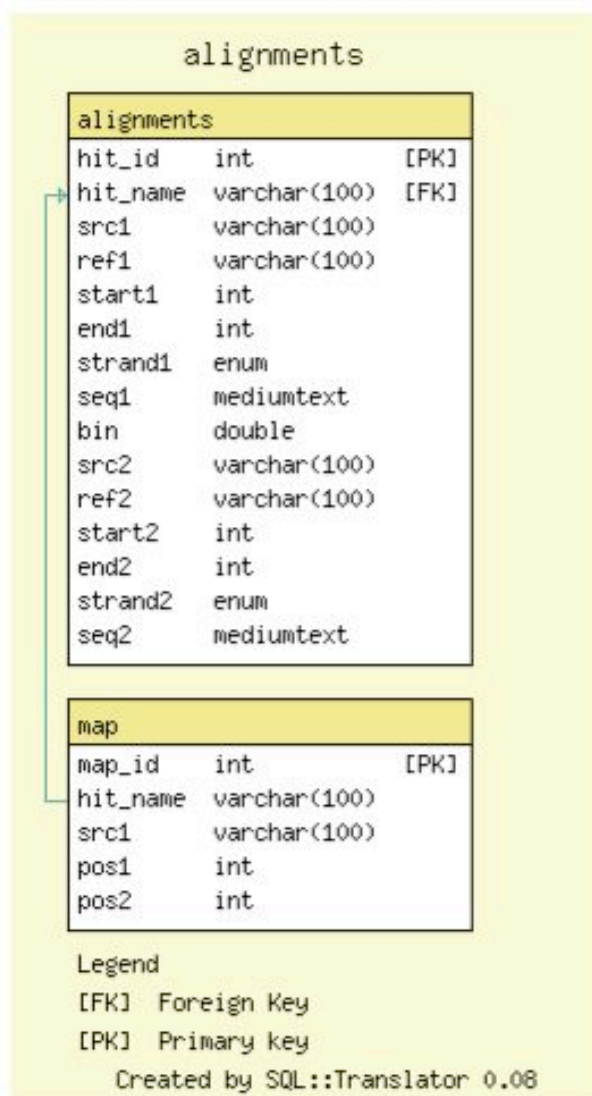
[GBrowse]

[GBrowse]

Bio::DB::GFF
species1



Bio::DB::GFF
species3



Bio::DB::GFF
species2

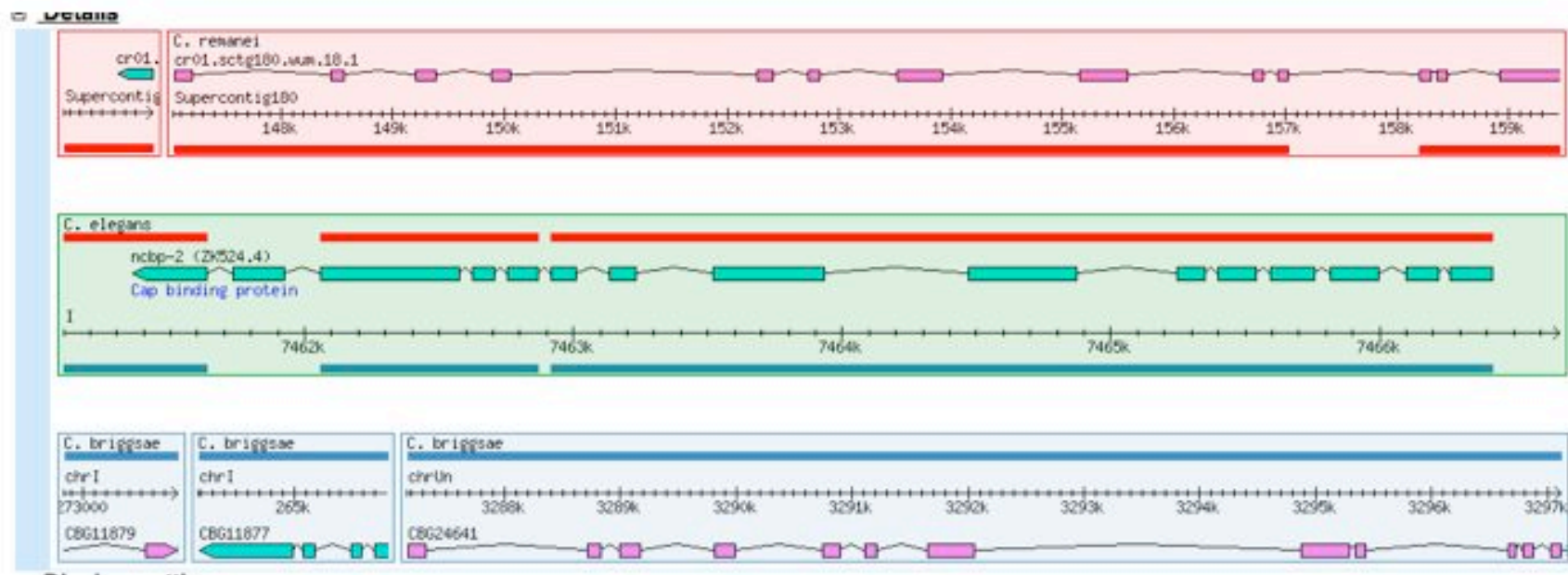


Bio::DB::GFF
species4

[GBrowse]

[GBrowse]

How to get the most information about the alignments?

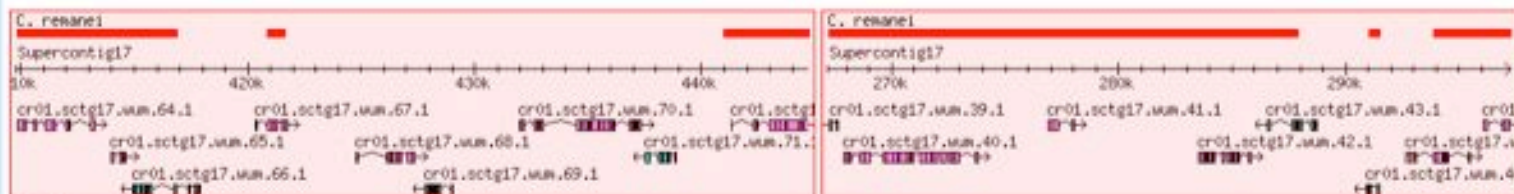
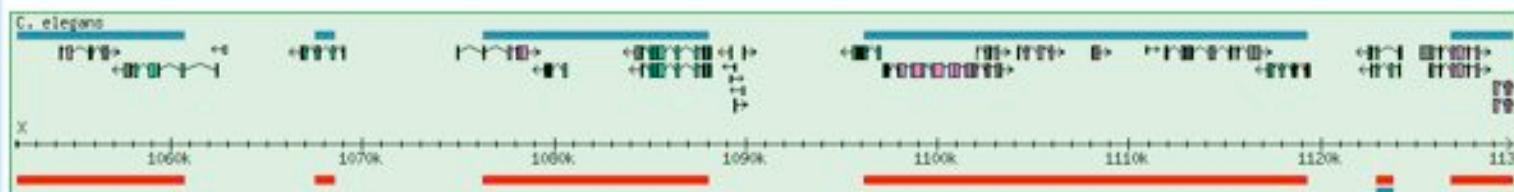


Gbrowse_syn: quick tour

Overview



Details



Display settings


Image widths : 640 768 800 1024 1280

Image options :

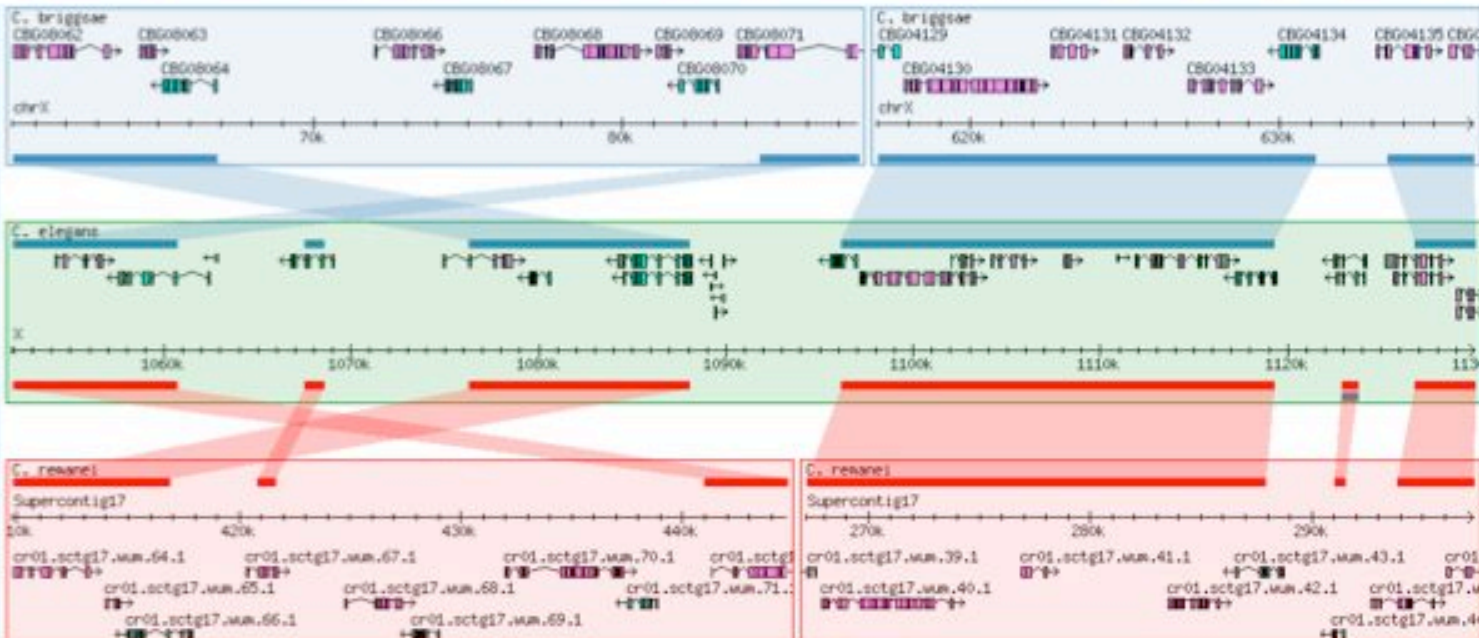
Flip minus strand panels on off Allow tiny panels on off Grid lines on off Edges on off Shading on off

Gbrowse_syn: quick tour (shaded alignments)

Overview



Details



Display settings

Image widths : 640 768 800 1024 1280

Image options :

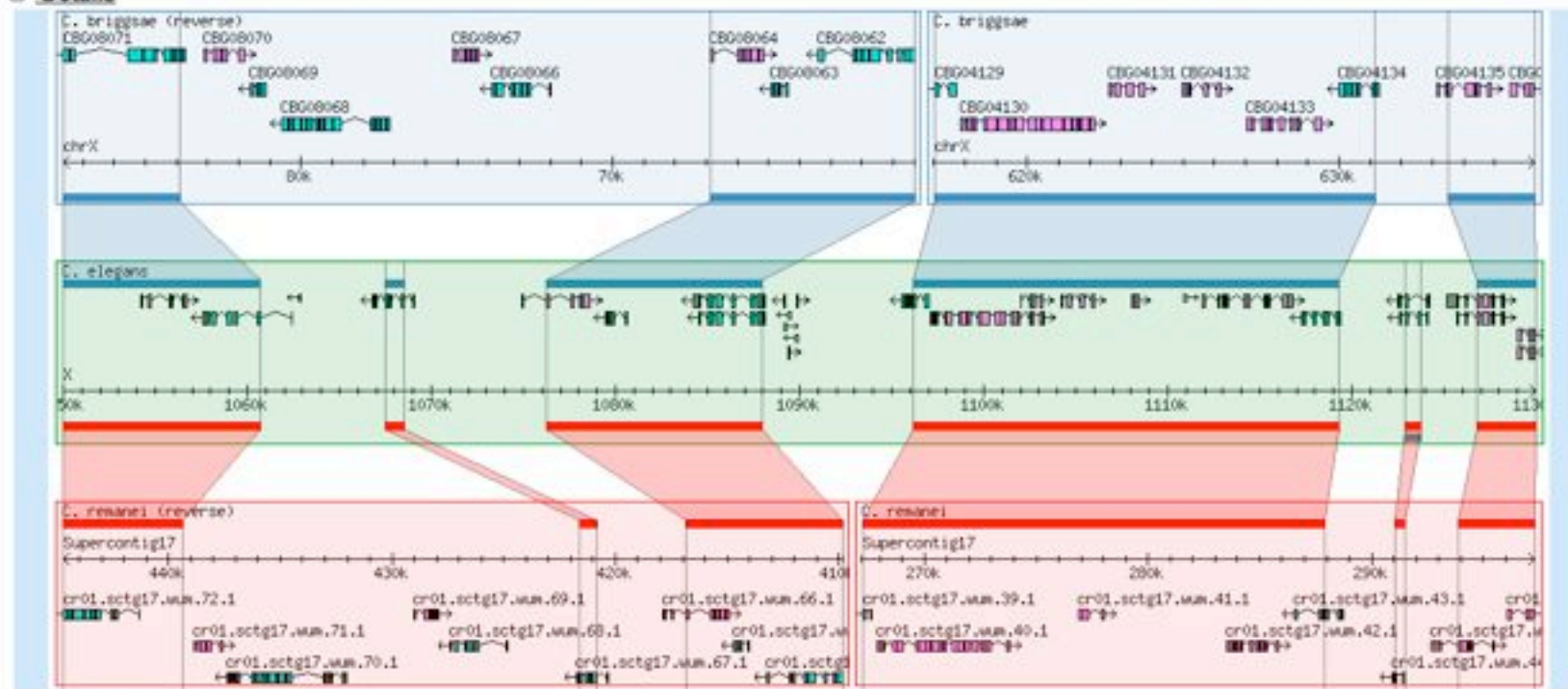
Flip minus strand panels on off Allow tiny panels on off Grid lines on off Edges on off Shading on off

Gbrowse_syn: quick tour (strand correction)

Overview



Details



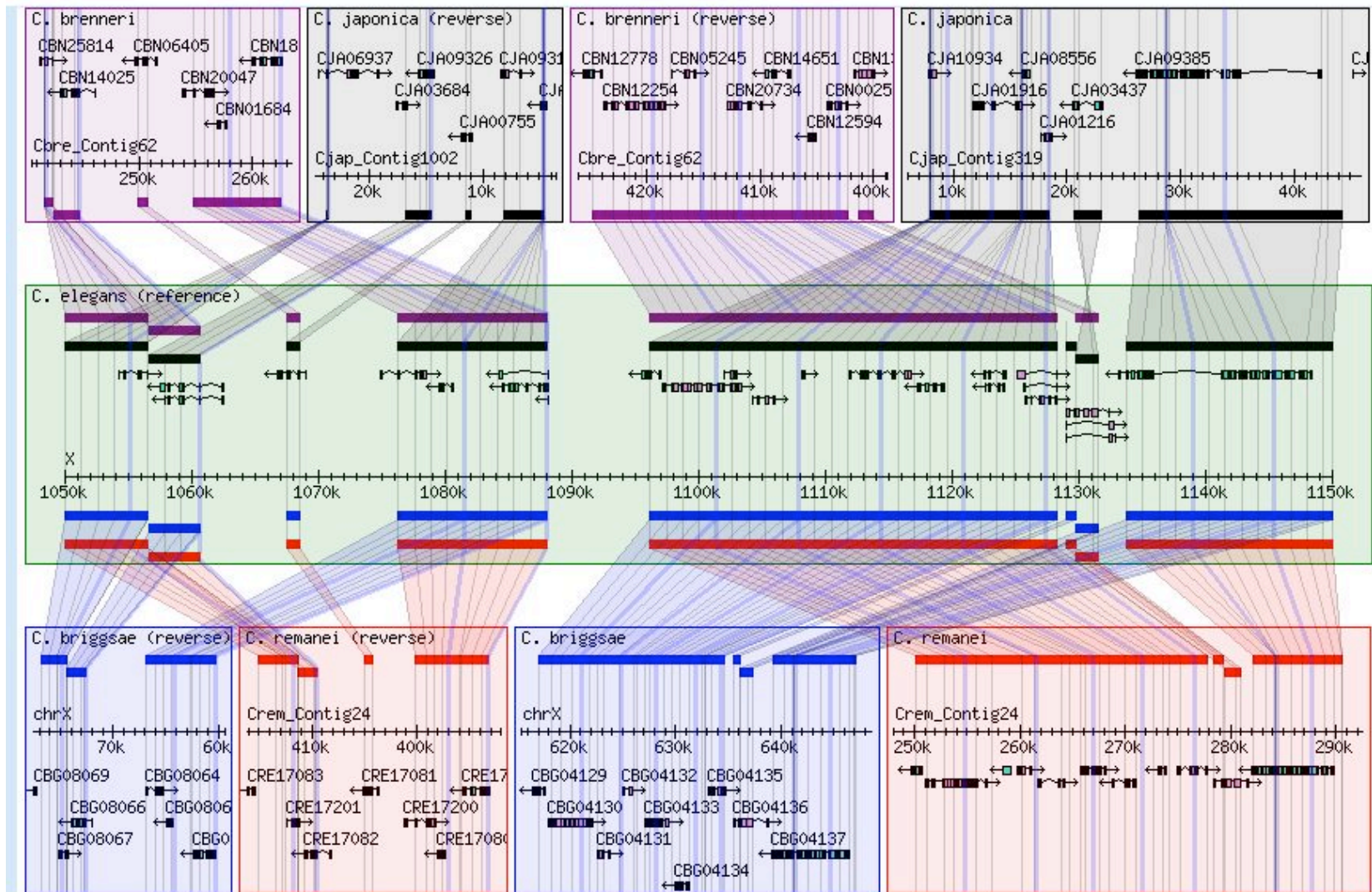
Display settings

Image widths : 640 768 800 1024 1280

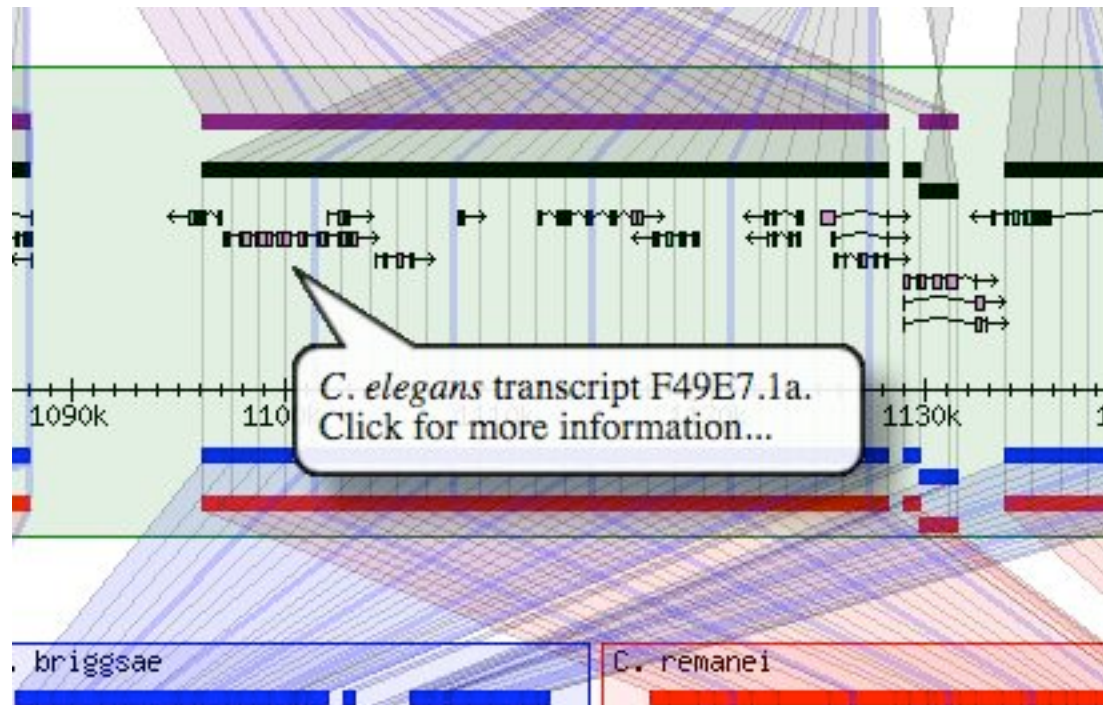
Image options :

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

All in one view



Adding markup to the annotations



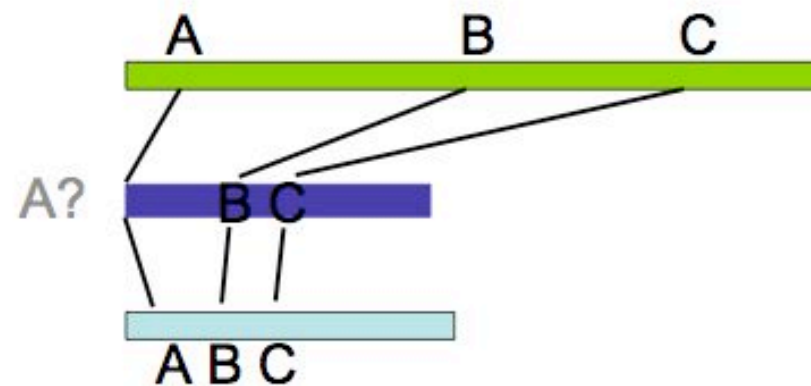
Problem : How to use Insertions/Deletion data

```

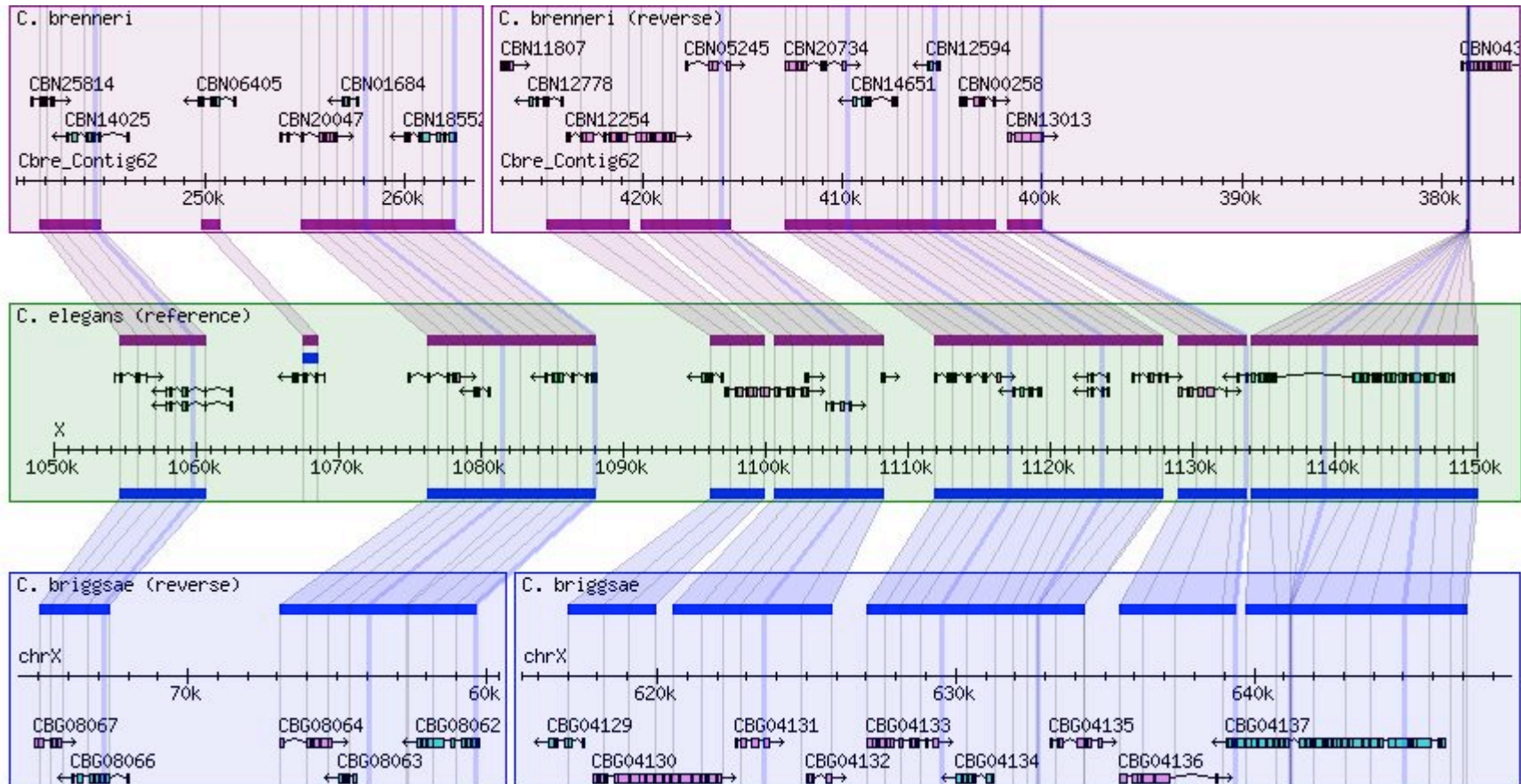
A
Ce-CHROMOSOME_I(+)/5195-16585 TGGCAAAAATATTTTGCATTTGCCGTTTTTCCCGTTTGCCGAAAAGTCTAATTTGCGTAA
Cb-chrI(-)/4091935-4097143 -----
Cr-Contig8(+)/571990-577344 TTCGAAAC-----

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

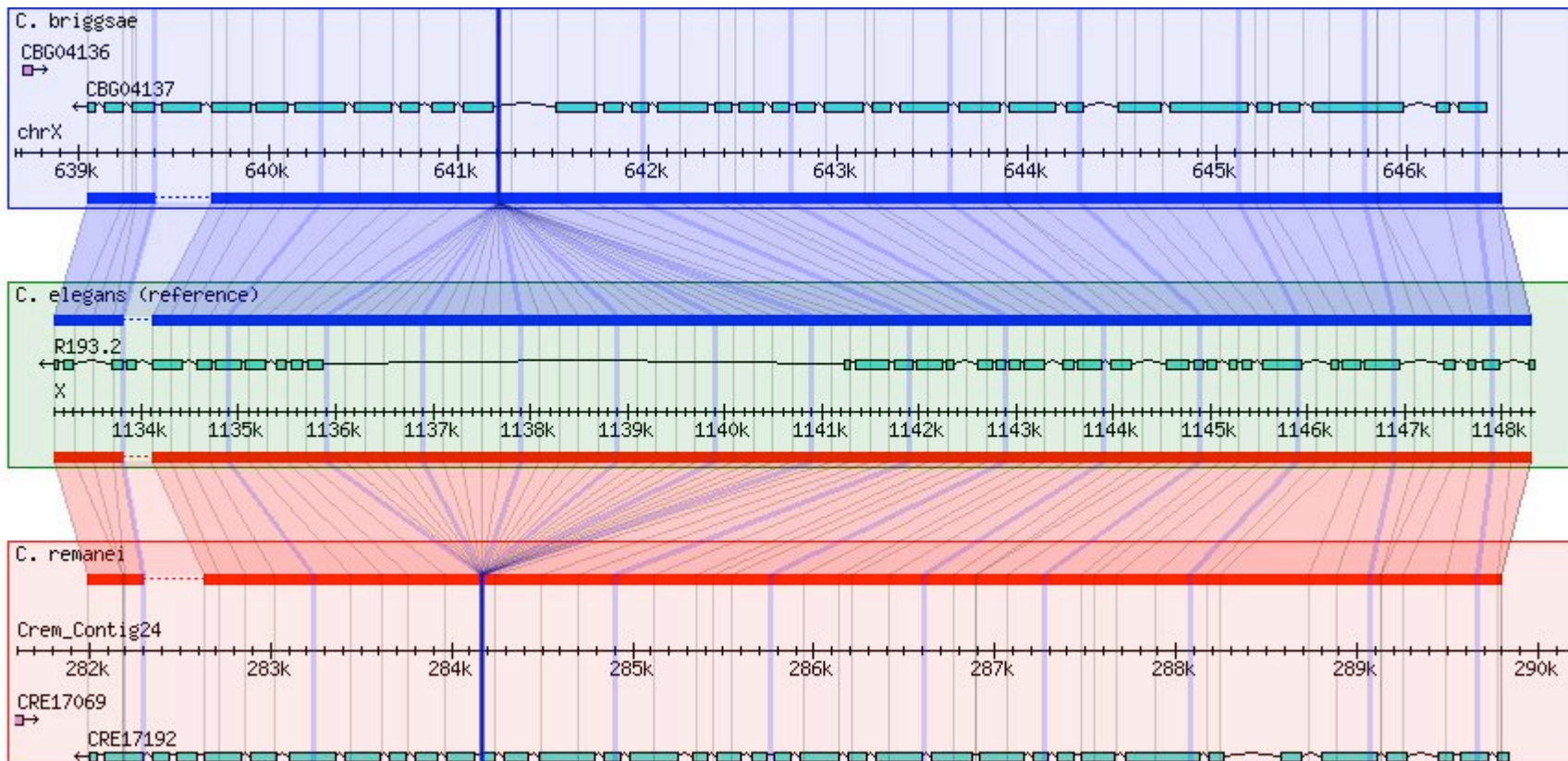
C
Ce-CHROMOSOME_I(+)/5195-16585 TTATTACGACATTCGTTTTATTTGAGCACAATTTGGGCCTATACTTTCAAAATCGGGGTTT
Cb-chrI(-)/4091935-4097143 --TTCATGTCAA-----TCAT
Cr-Contig8(+)/571990-577344 --TTTCTGAAAACAGGTAGTATTATGGTTCCGAGGGTGTAGGGTTTCGAAACCGGCCTAG
                                * * *
  
```



Tracking Indels with grid lines

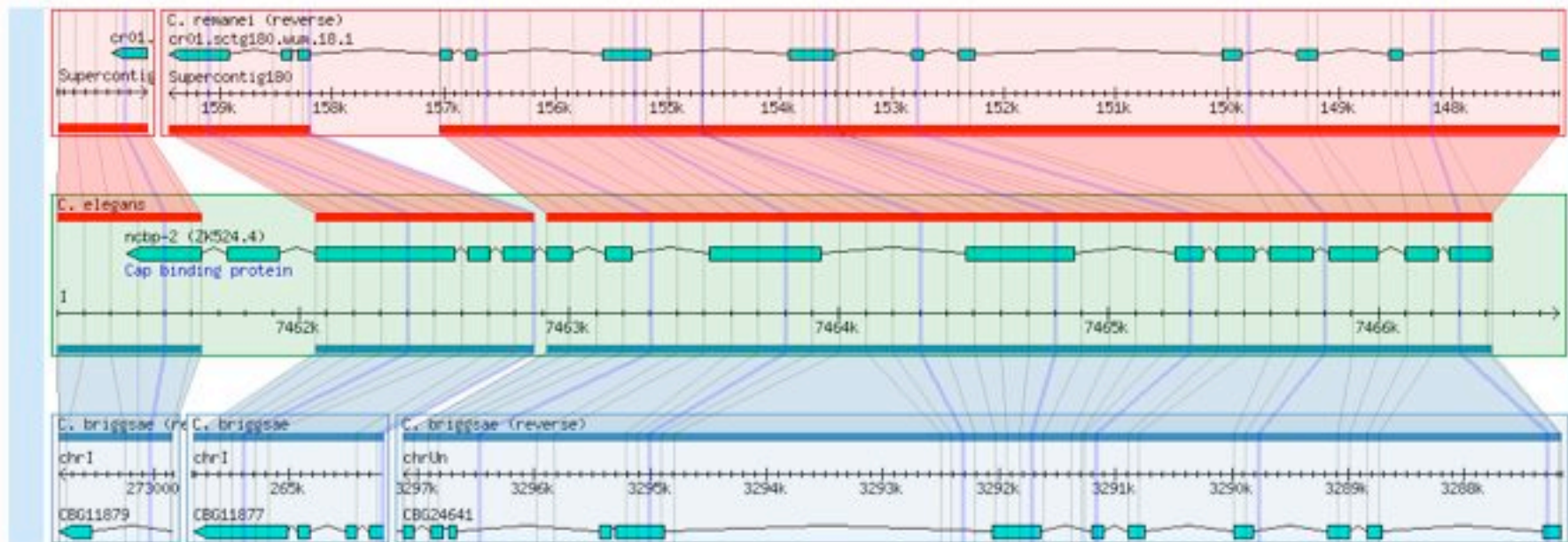


Evolution of Gene Structure

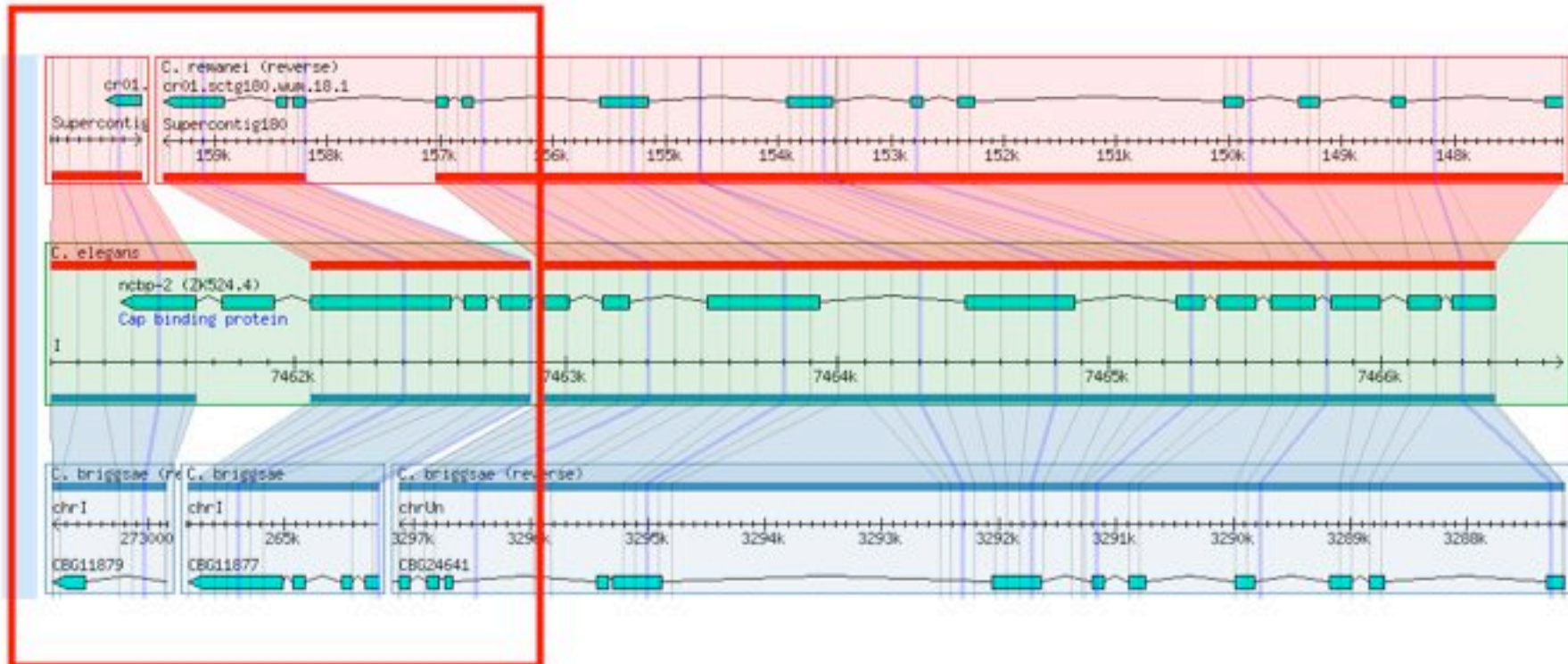


Comparing gene annotations:

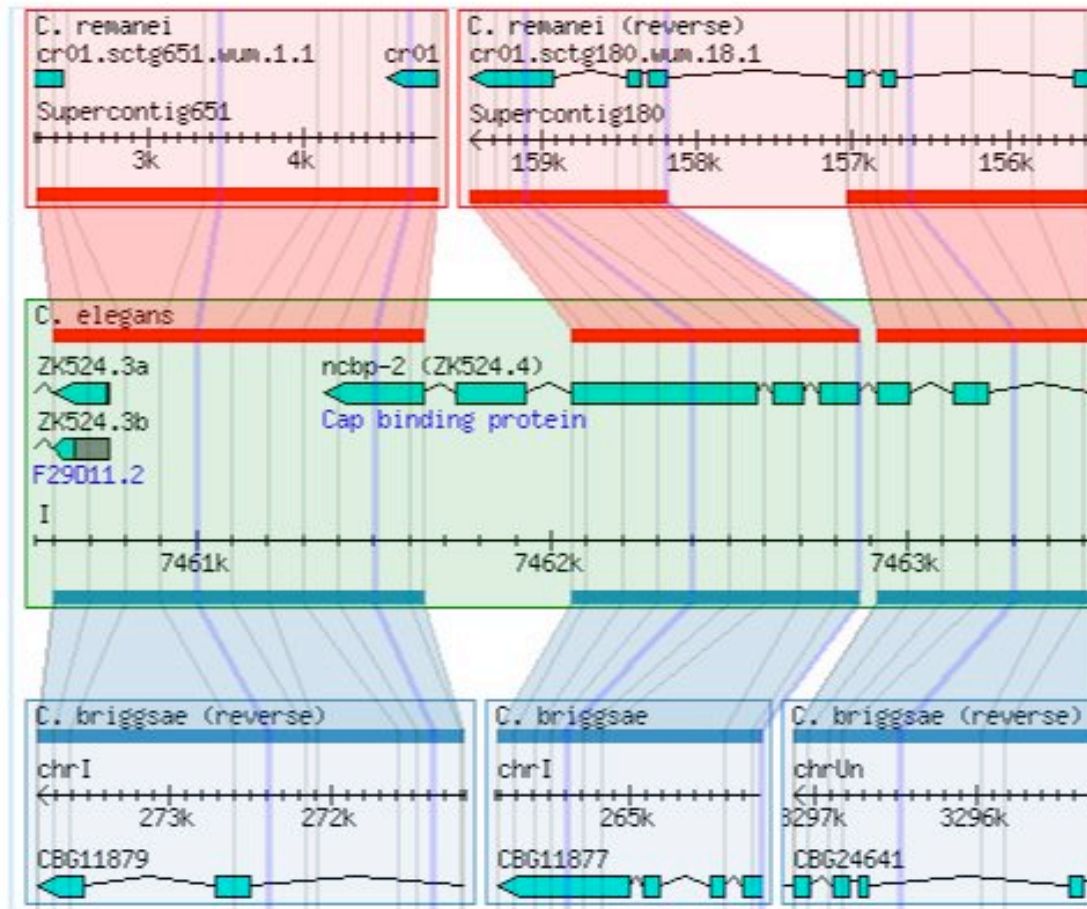
C. elegans ZK524.4 gene



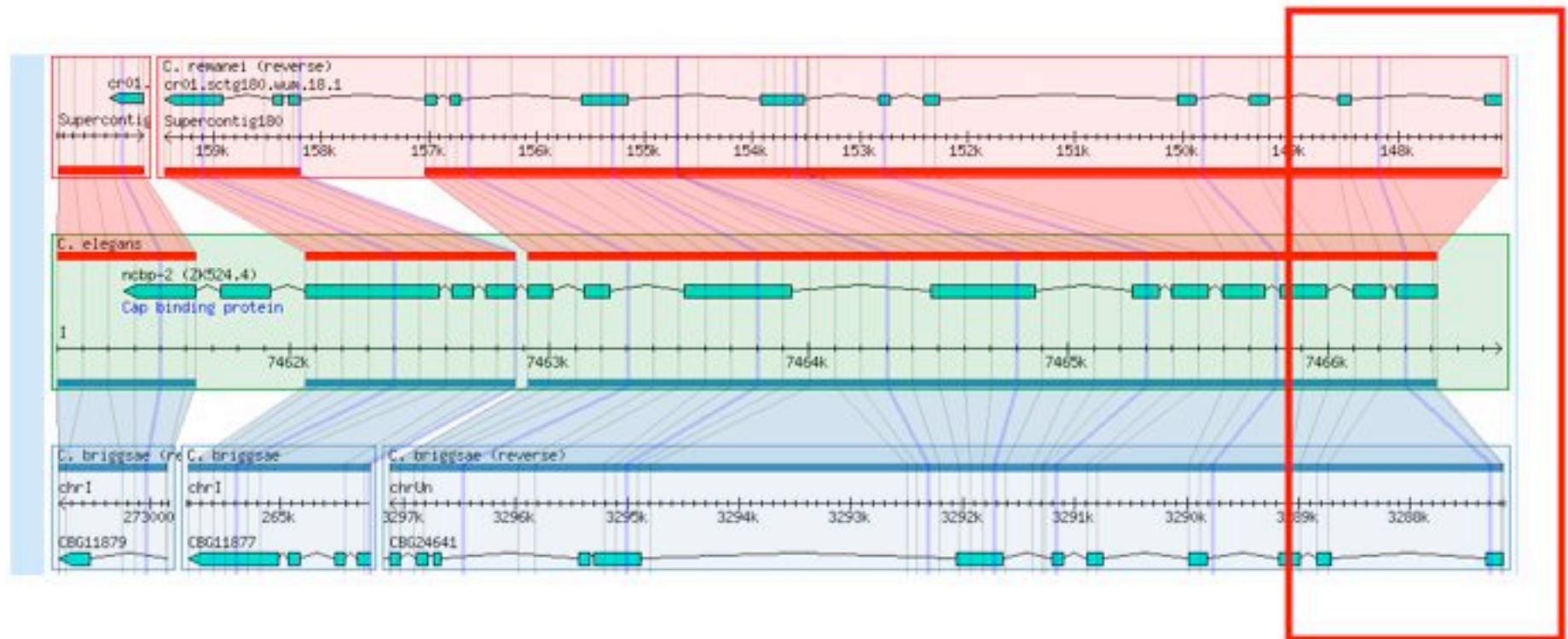
C. elegans ZK524.4 gene



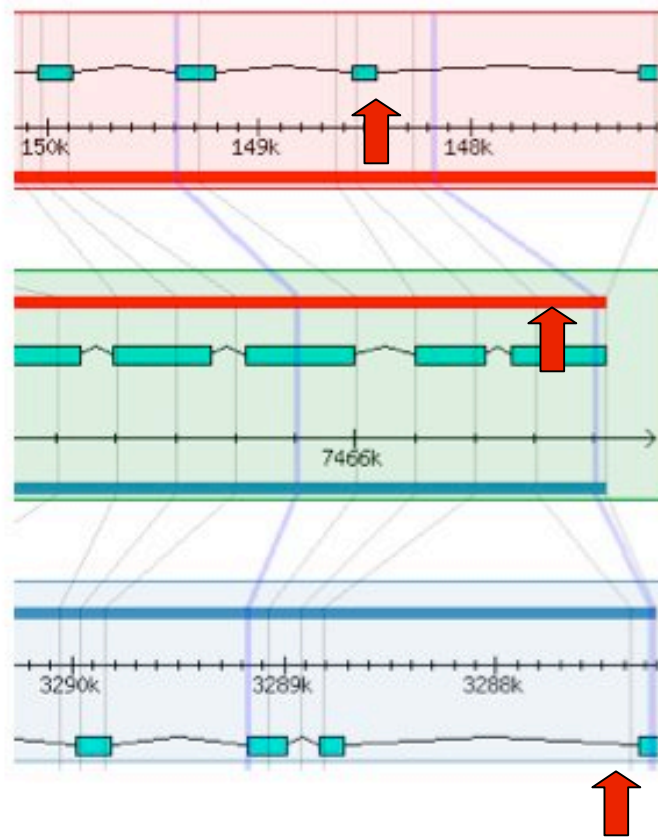
3' end



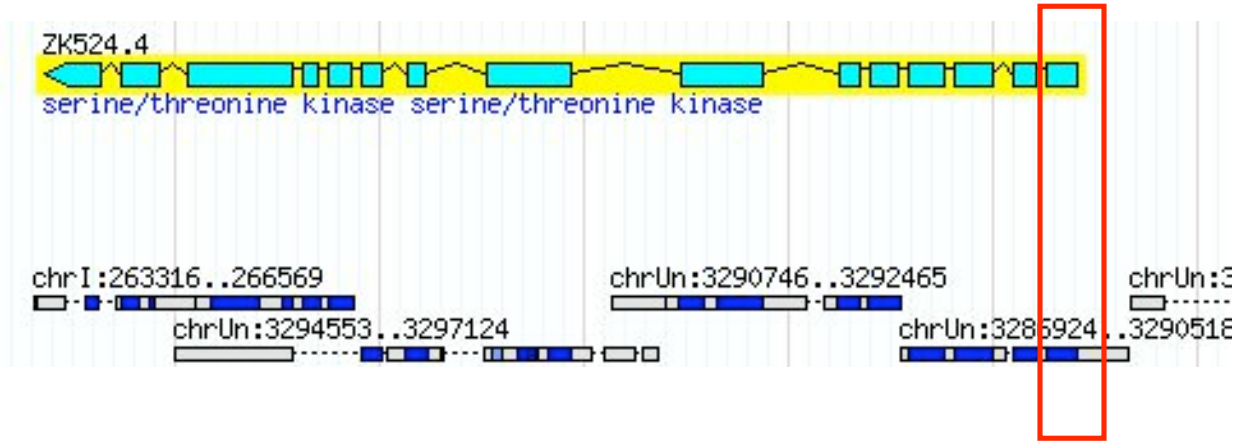
C. elegans ZK524.4 gene



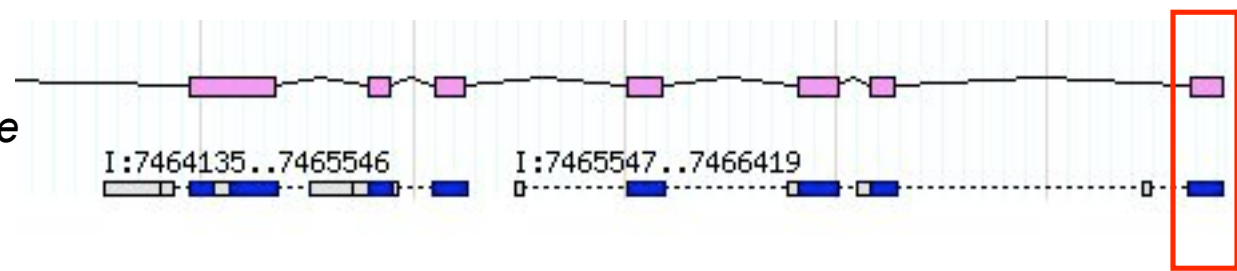
5' end



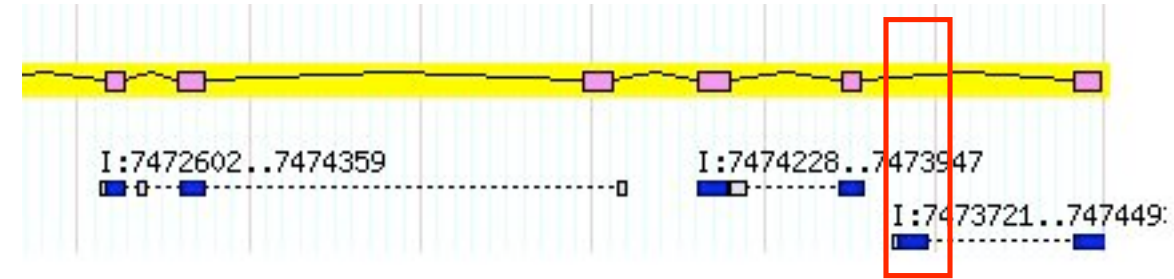
C. elegans



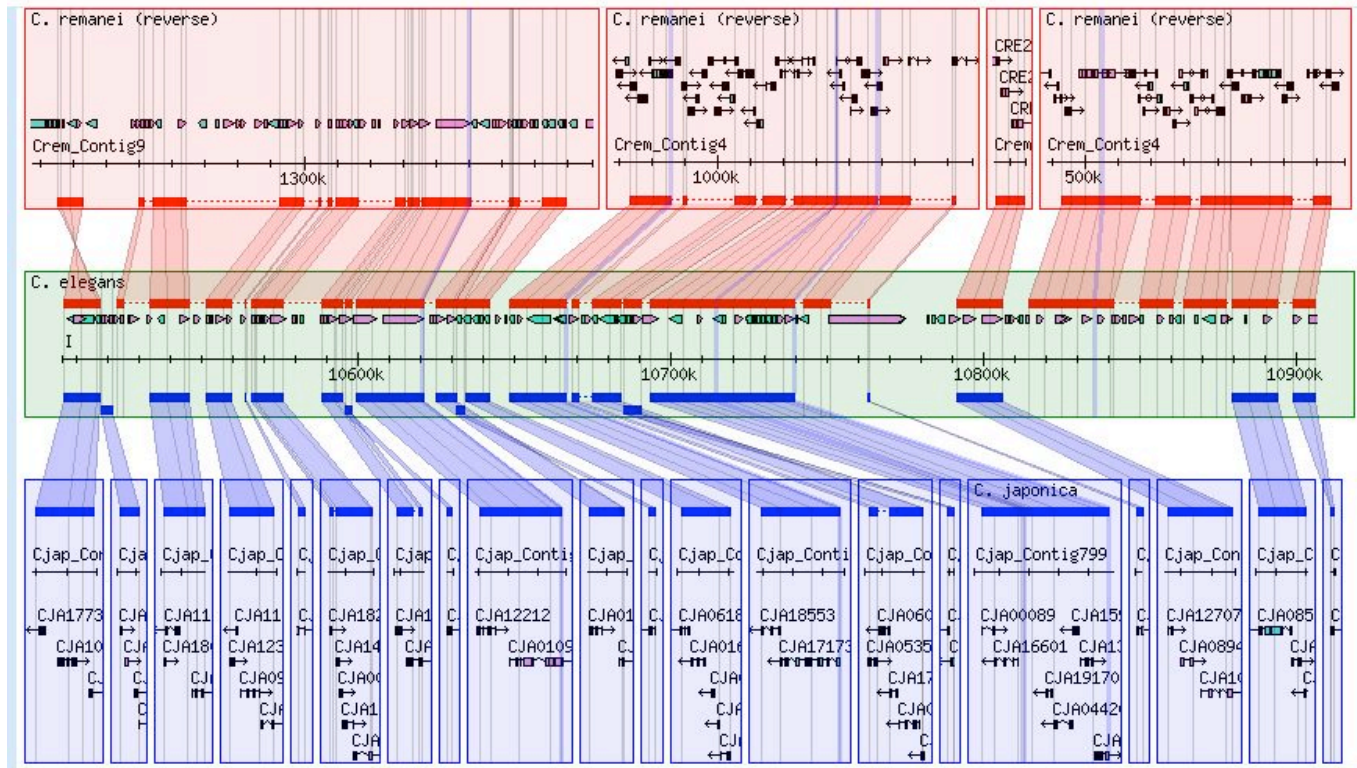
C. briggsae



C. remanei



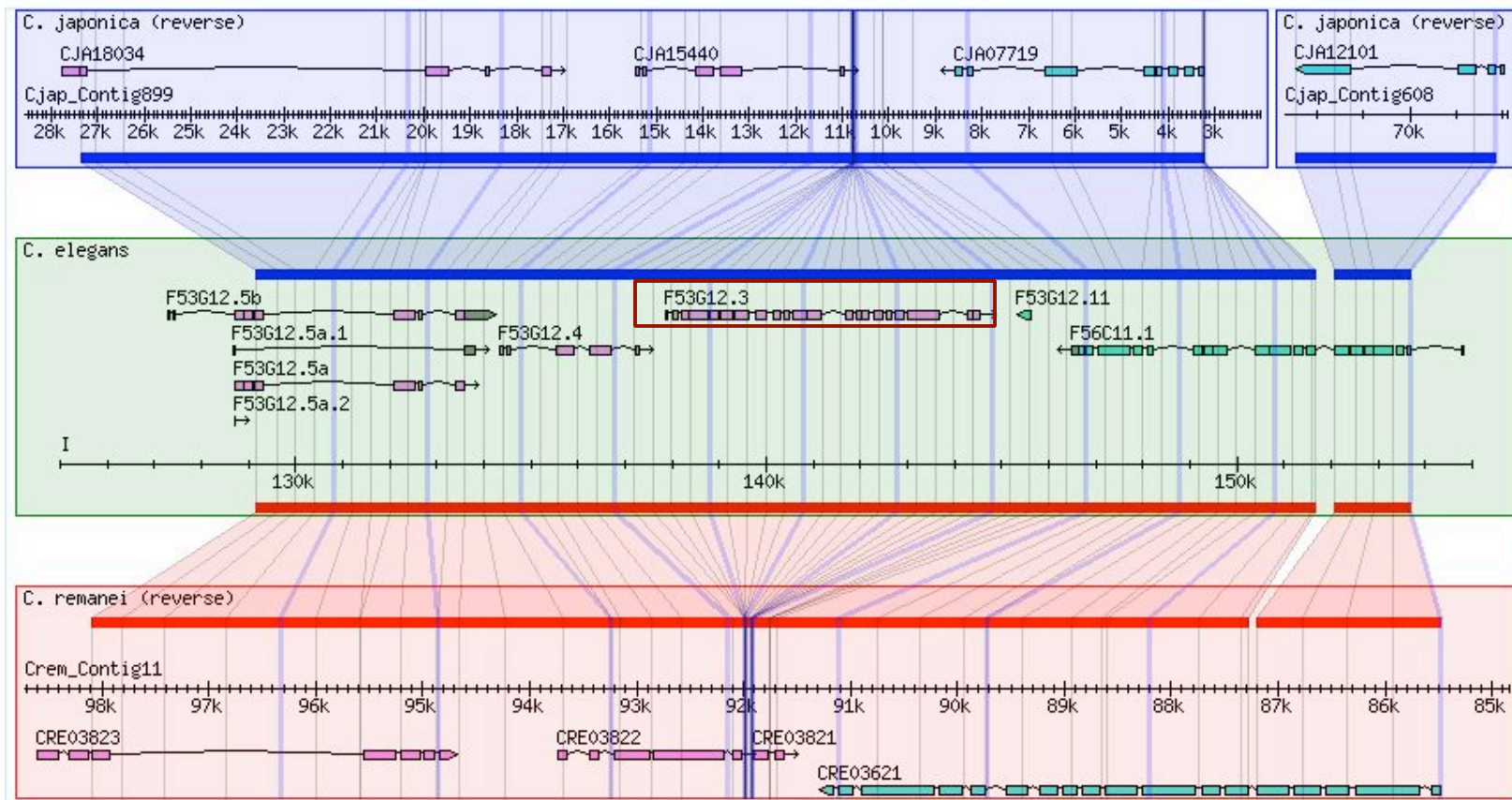
Comparing assemblies



Not bad

Needs work

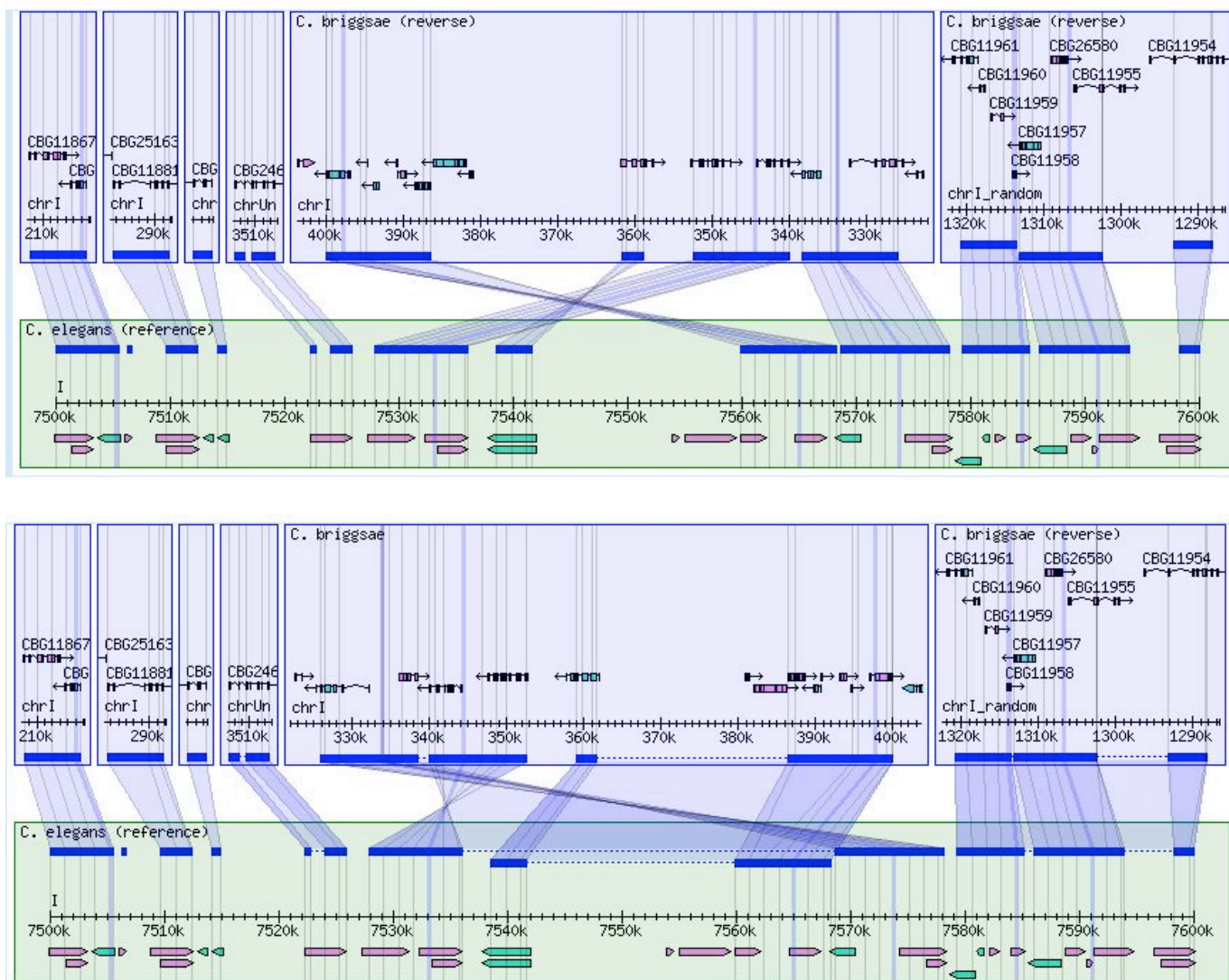
Putative gene or loss



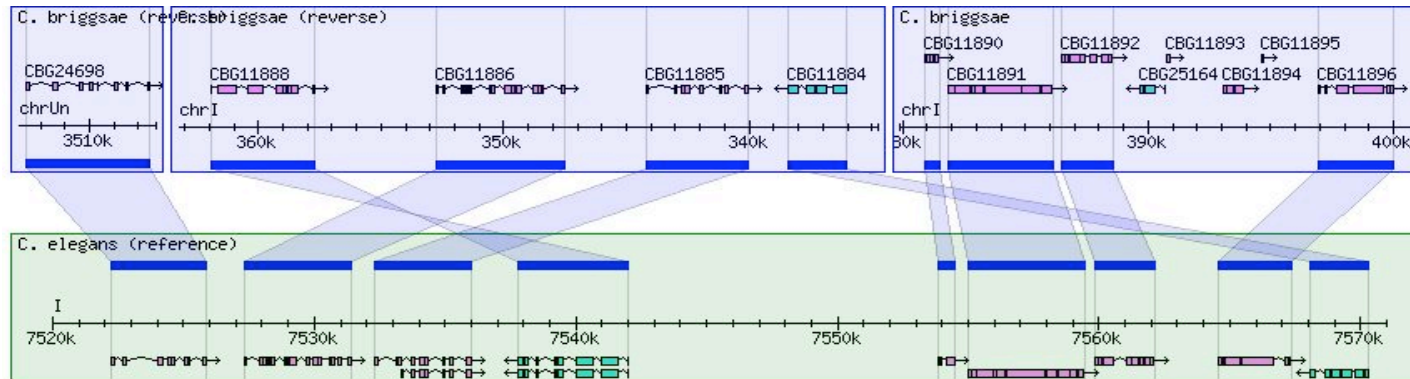


Problem: Getting the most out of small aligned regions
or orthology-only data

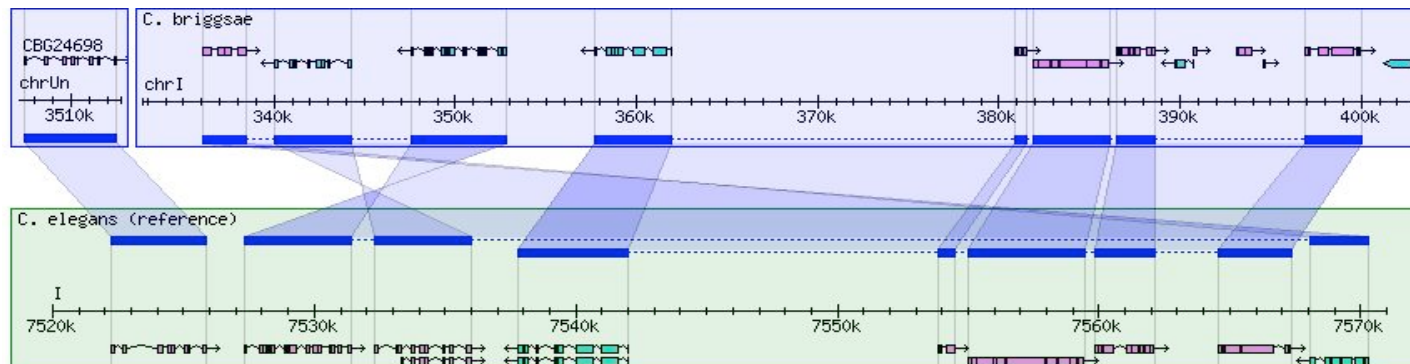
Chaining Alignments

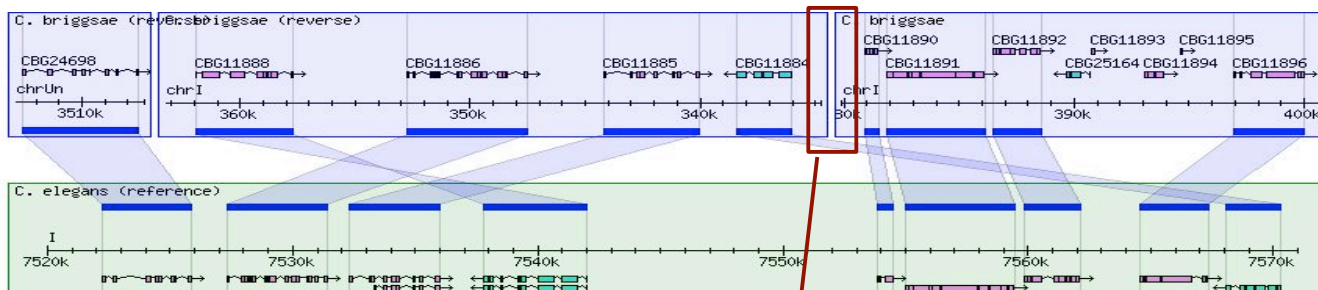


Gene Orthology



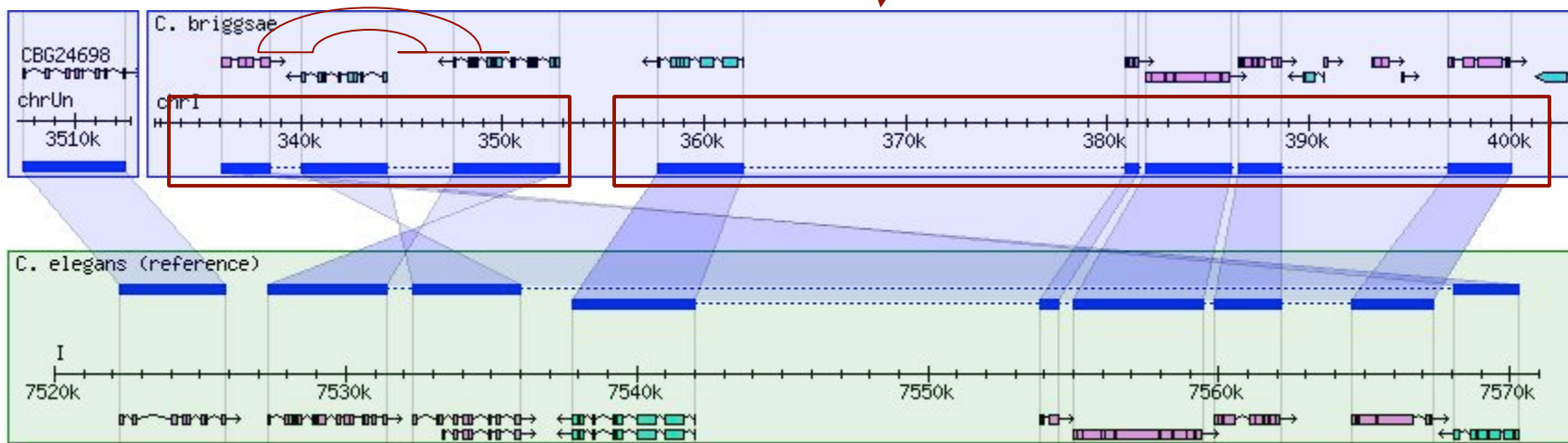
Chained Orthologs





Inversion + translocation?

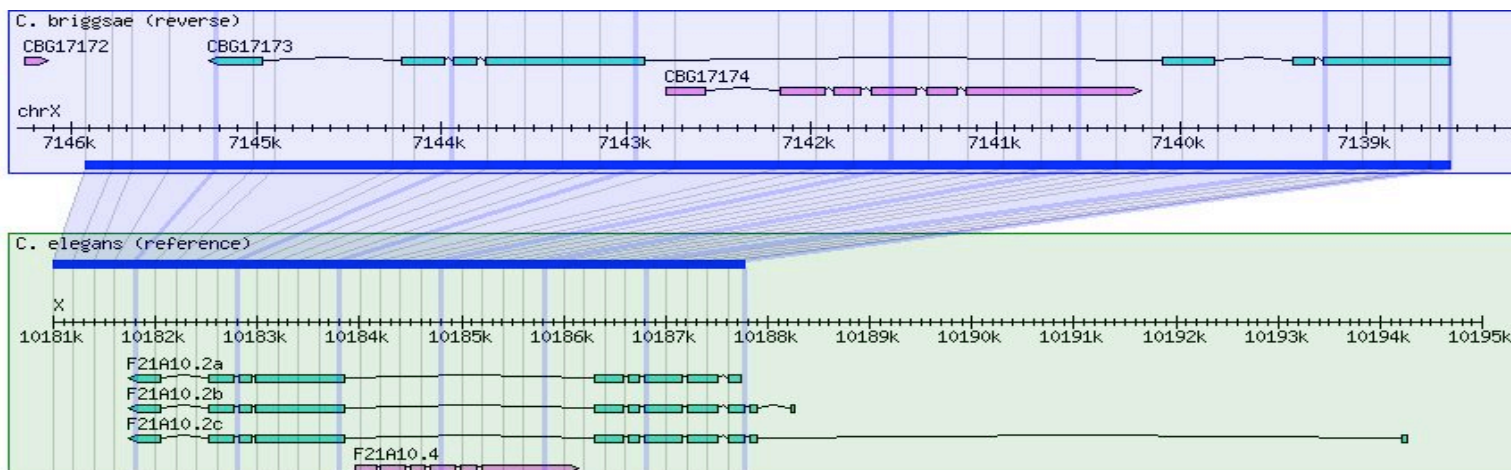
2 panels merged



Gene Orthology



PECAN Alignments



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

- Synteny blocks may only have two anchor points
- Synteny blocks may not have a 1:1 length ratio



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.

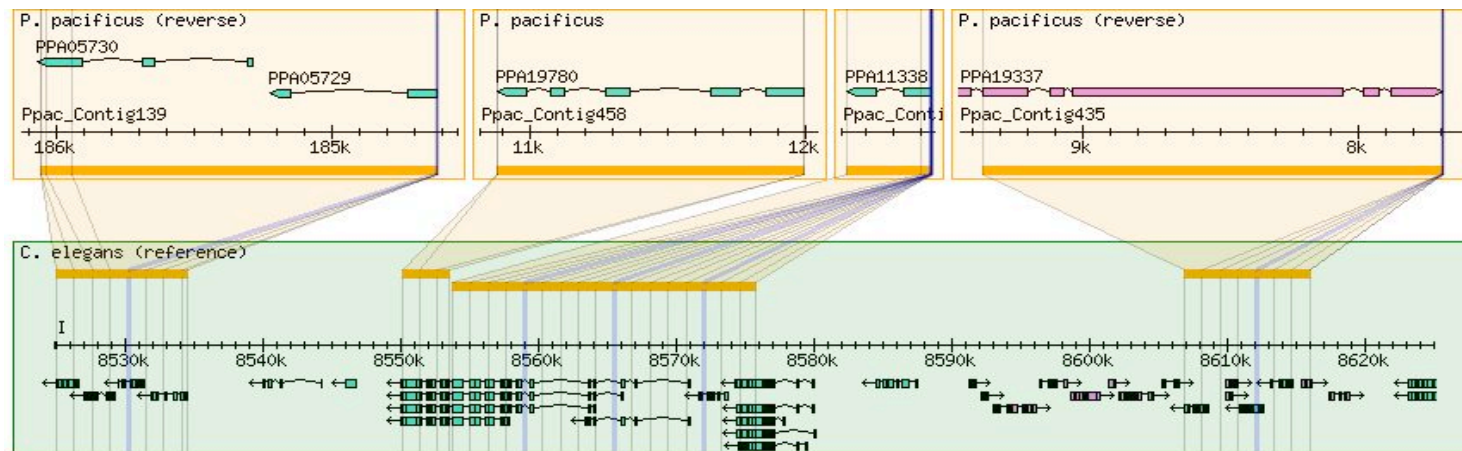
Solution 3: If all else fails, guess the end of the target block based on the overall length ratio.

$\text{length displayed target} = (\text{length target} / \text{length reference}) * \text{length displayed reference}$

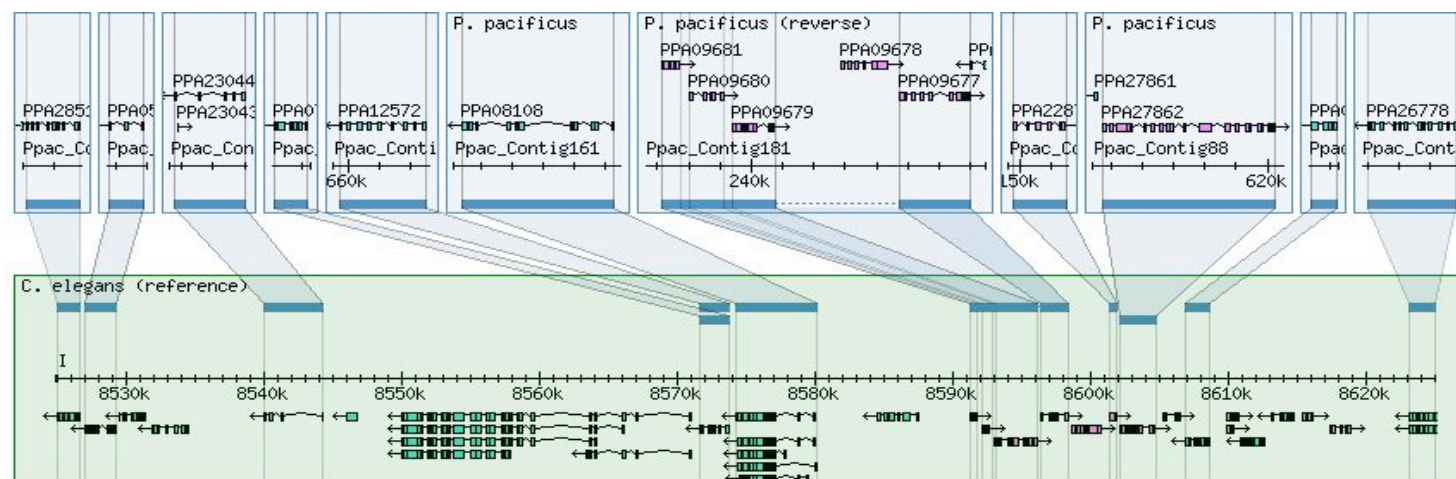


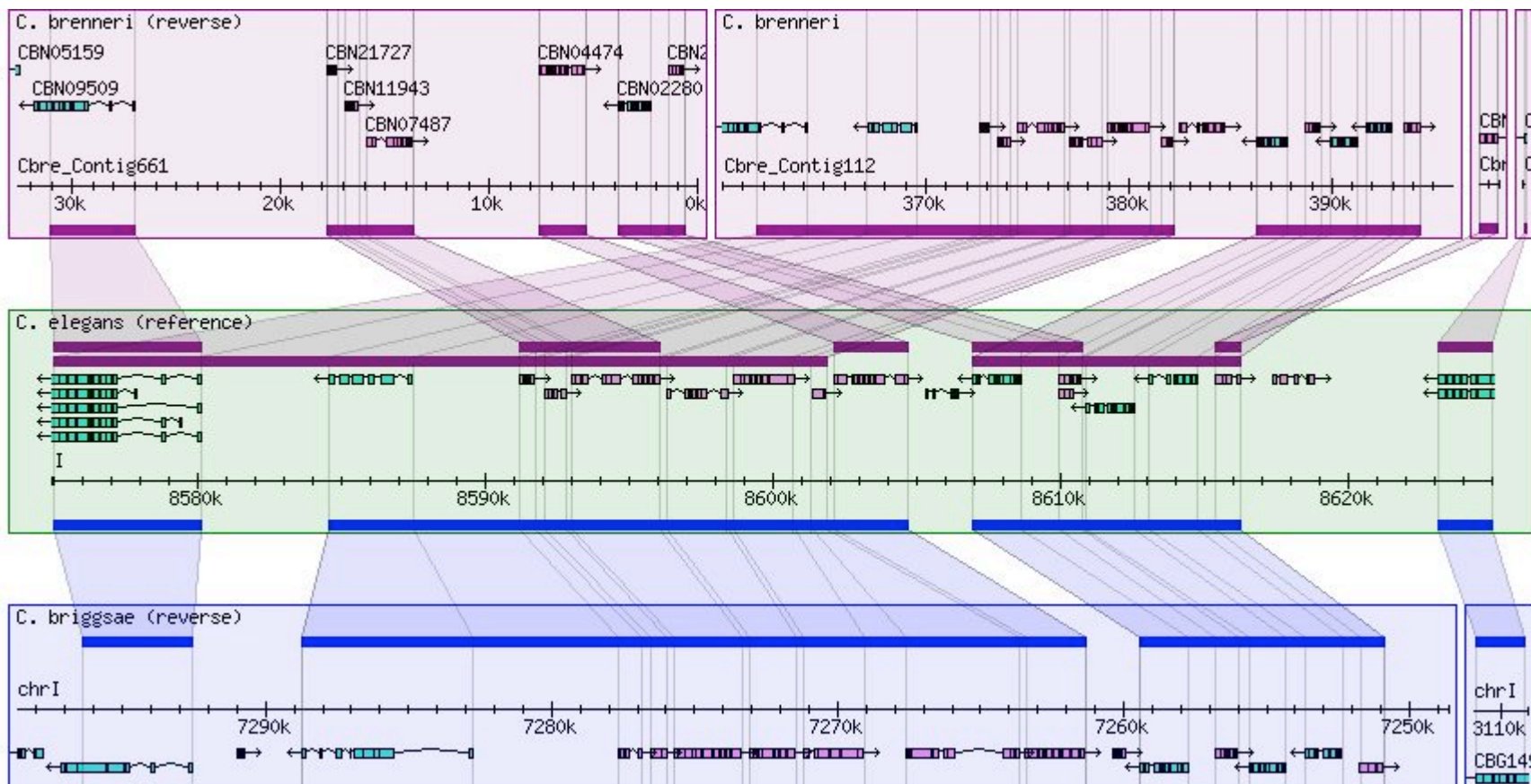
Problem: What if the aligned DNA sequences are too distant?

Pecan alignments

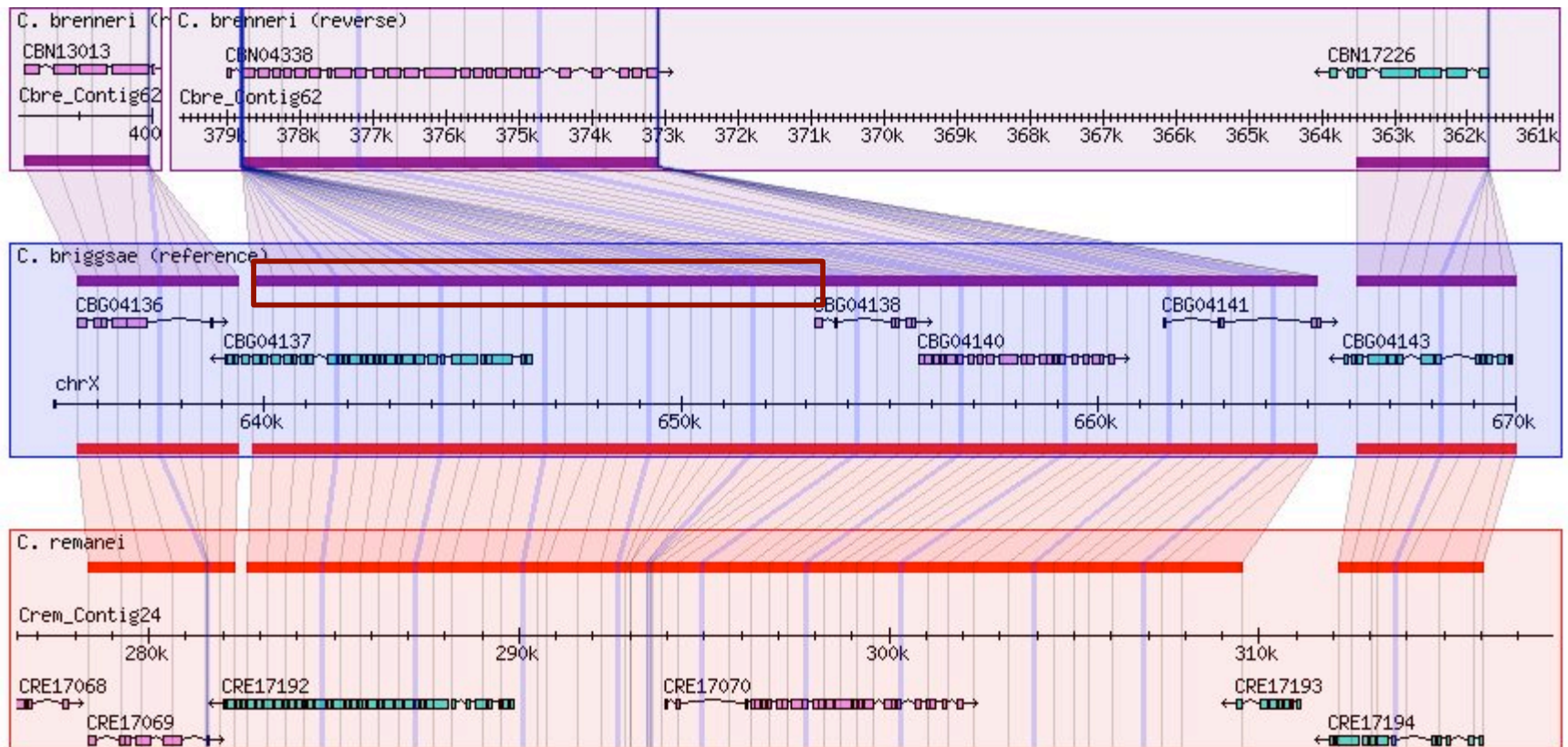


Orthocluster Synteny blocks



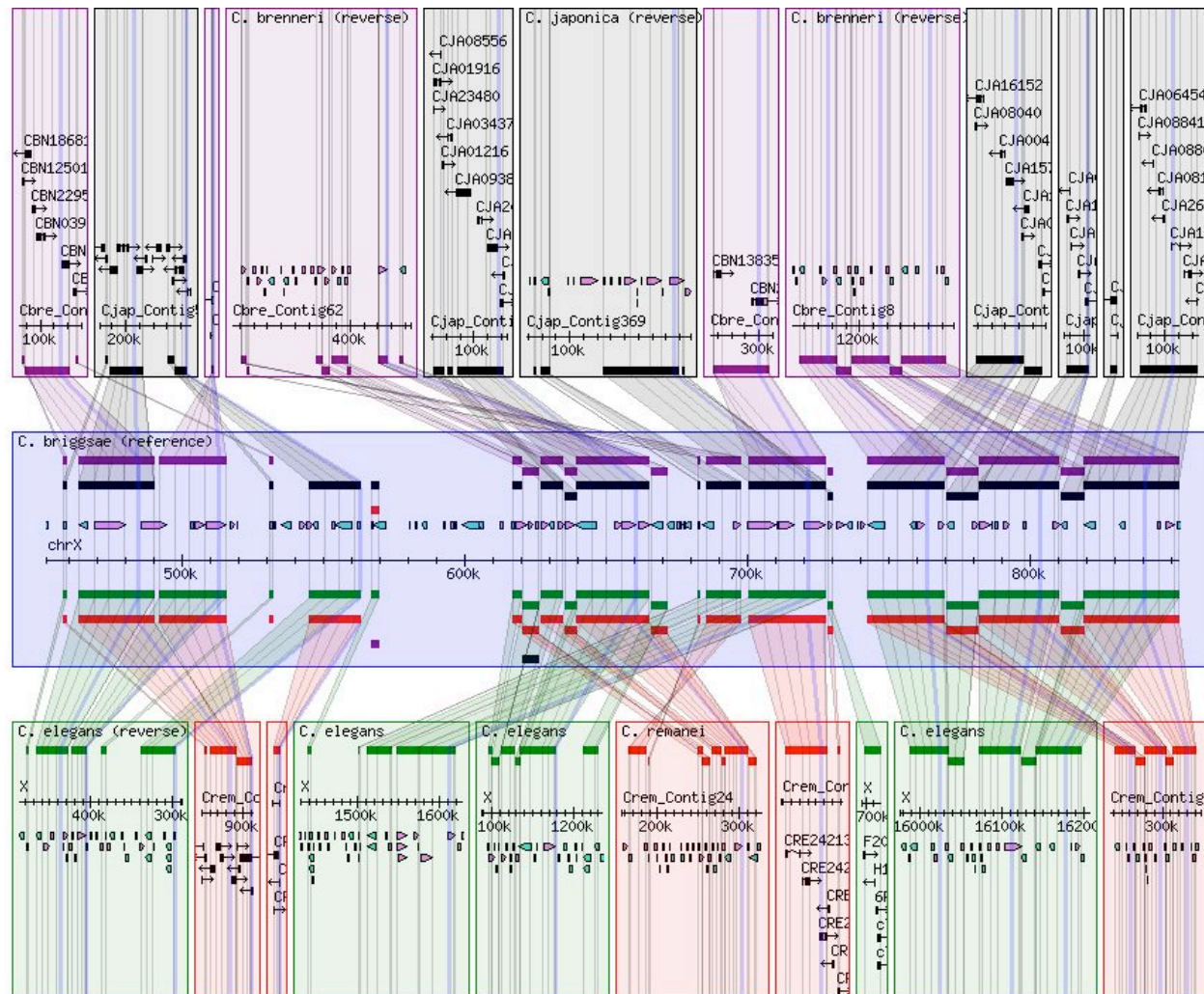


Problem: terminal insertion/deletions in multiple sequence alignments



Solution: use pair-wise alignments?

Problem: Space!





Future Improvements for GBrowse_syn

- “On the fly” sequence alignment view (definitely)
- AJAX-based image configuration (probably)
- 3D image rendering? (maybe)
- Suggestions?



Acknowledgments

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

