Comparative Genomics with GBrowse_syn

Sheldon McKay,
Cold Spring Harbor Laboratory

iPlant Collaborative™

Outline

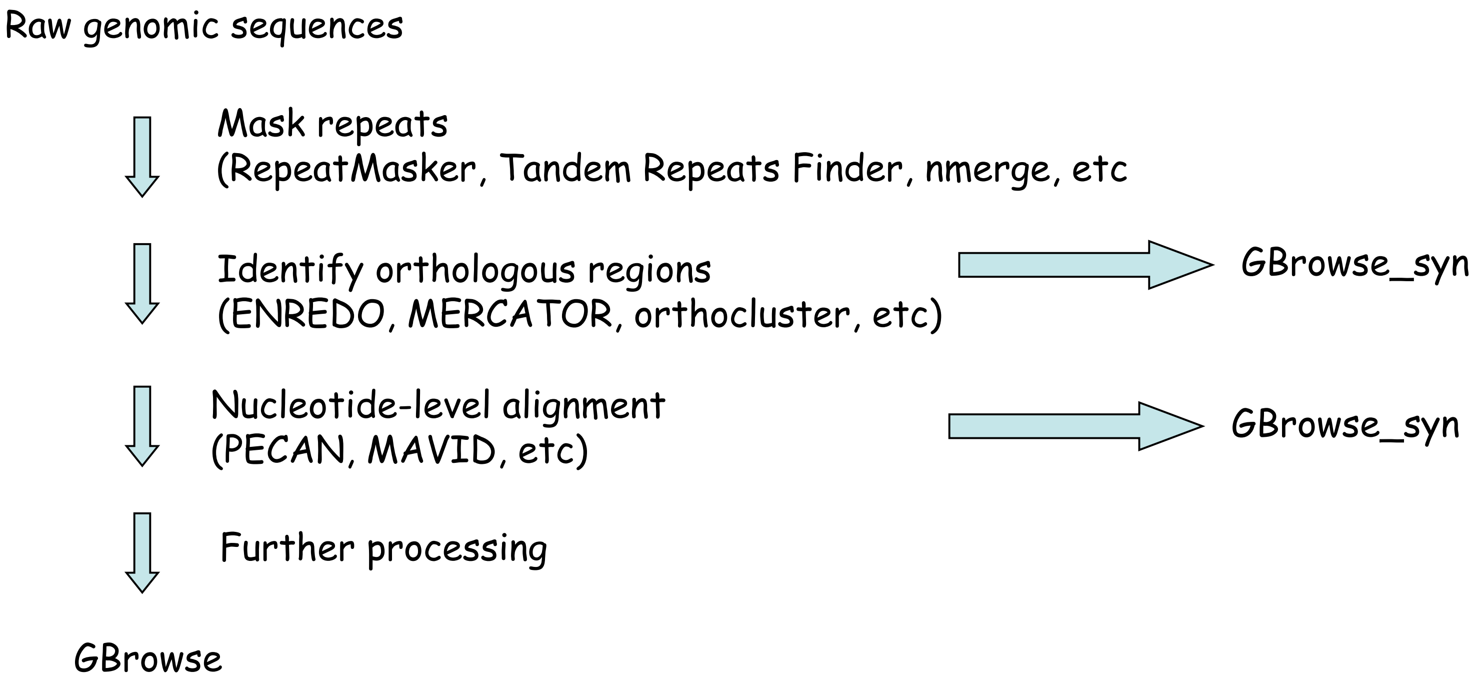
A few words on data

A brief survey of synteny browsers

A few challenges of rendering comparative data

Comparative genome browsing with GBrowse_syn

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

A Few Use Cases

- Multiple sequence alignment data from whole genome
- Synteny or Co-linearity data without alignments
- Gene orthology assignments based on proteins
- Self vs. Self comparison of duplications, homeologous regions, etc
- Others

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

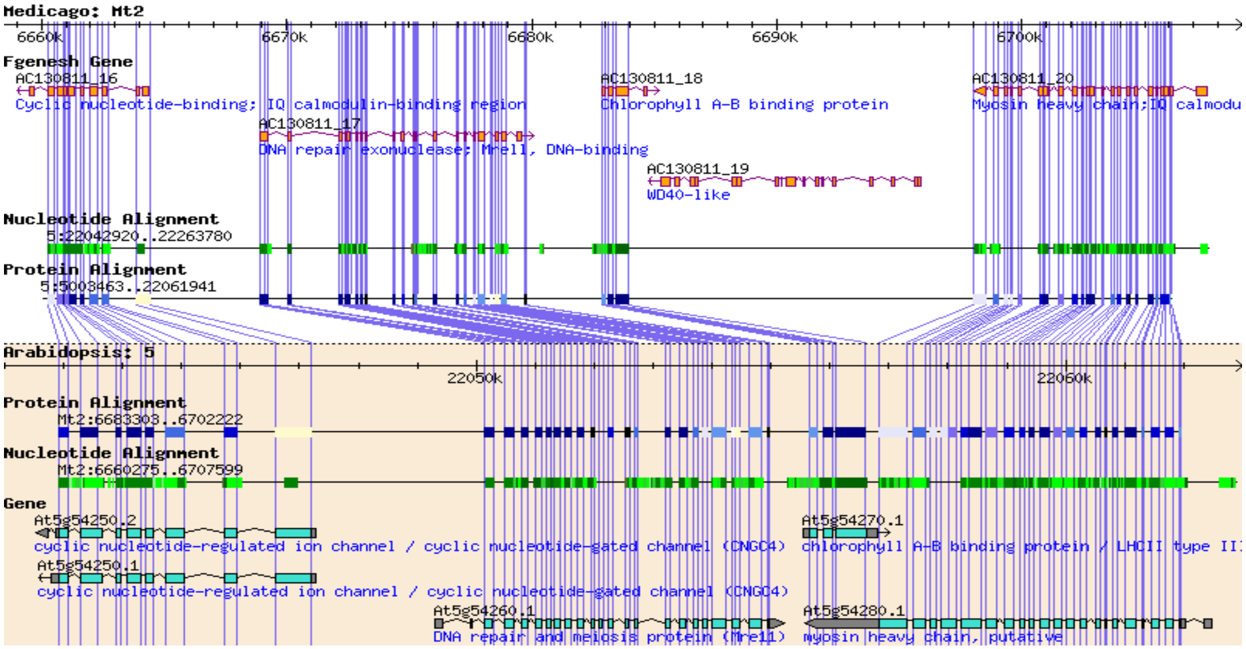
PopGenIE

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

iPlant Collaborative™

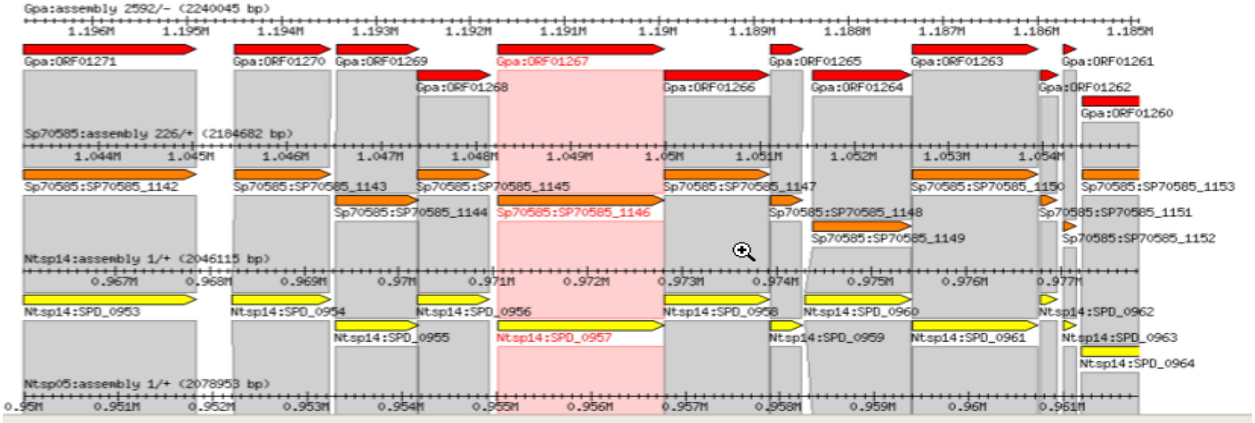
SynBrowse

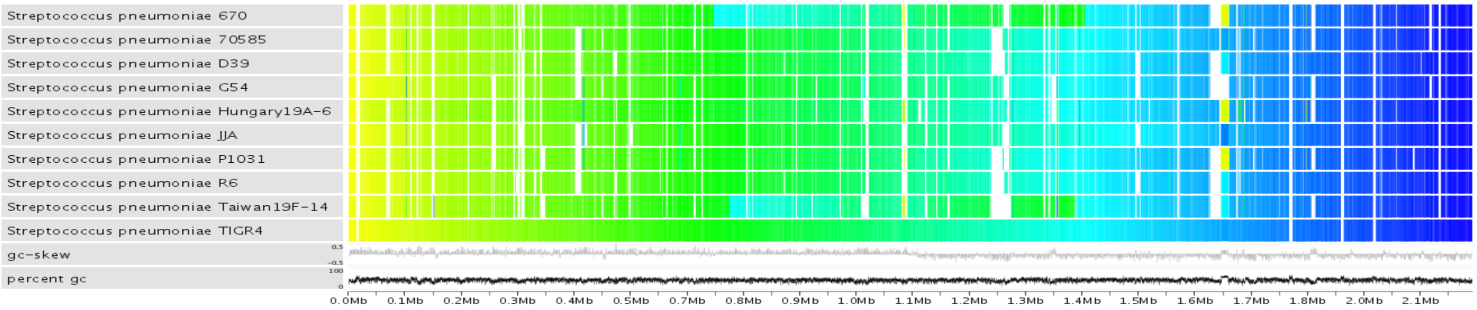
...A Synteny Browser for Comparative Sequence Analysis


Medicago: Mt2
6660k
6670k
6680k
6690k
6700k
Fgenesh Gene
AC130811_16
Cyclic nucleotide-binding; IQ calmodulin-binding region
AC130811_17
DNA repair exonuclease; Mre11, DNA-binding
AC130811_18
Chlorophyll A-B binding protein
AC130811_19
WD40-like
AC130811_20
Myosin heavy chain;IQ calmodu
Nucleotide Alignment
5:22042920..22263780
Protein Alignment
5:5003463..22061941
Arabidopsis: 5
22050k
22060k
Protein Alignment
Mt2:6683303..6702222
Nucleotide Alignment
Mt2:6660275..6707599
Gene
At5g54250.2
cyclic nucleotide-regulated ion channel / cyclic nucleotide-gated channel (CNGC4)
At5g54250.1
cyclic nucleotide-regulated ion channel / cyclic nucleotide-gated channel (CNGC4)
At5g54260.1
DNA repair and meiosis protein (Mre11)
At5g54270.1
chlorophyll A-B binding protein / LHCII type II
At5g54280.1
myosin heavy chain, putative

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

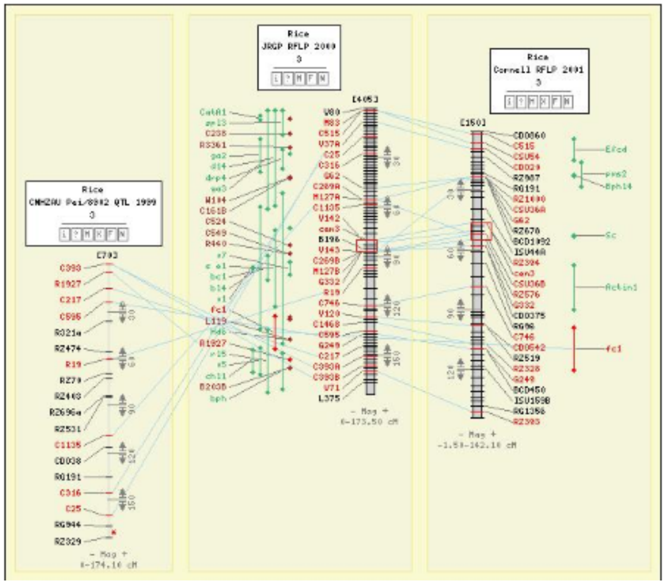


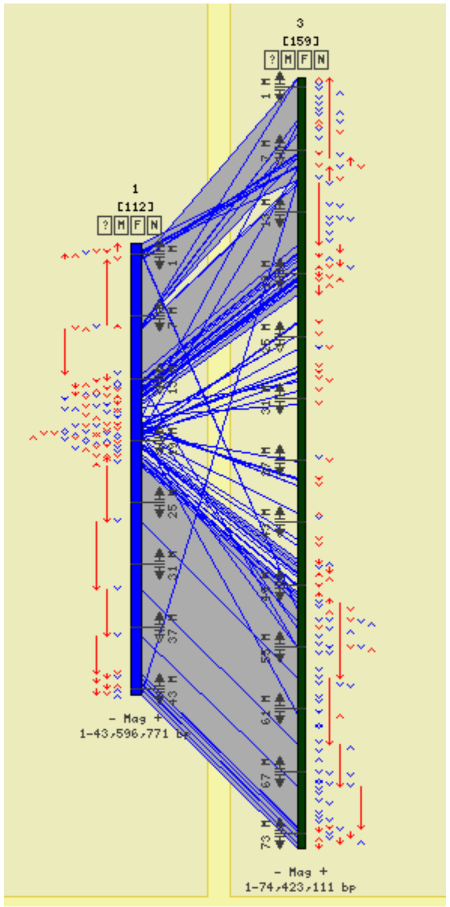

Streptococcus pneumoniae 670
Streptococcus pneumoniae 70585
Streptococcus pneumoniae D39
Streptococcus pneumoniae G54
Streptococcus pneumoniae Hungary19A-6
Streptococcus pneumoniae JJA
Streptococcus pneumoniae P1031
Streptococcus pneumoniae R6
Streptococcus pneumoniae Taiwan19F-14
Streptococcus pneumoniae TIGR4
gc-skew
percent gc
0.5
-0.5
100
0
0.0Mb
0.1Mb
0.2Mb
0.3Mb
0.4Mb
0.5Mb
0.6Mb
0.7Mb
0.8Mb
0.9Mb
1.0Mb
1.1Mb
1.2Mb
1.3Mb
1.4Mb
1.5Mb
1.6Mb
1.7Mb
1.8Mb
2.0Mb
2.1Mb

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.

iPlant Collaborative™





CMap

2.0

GRAMENE

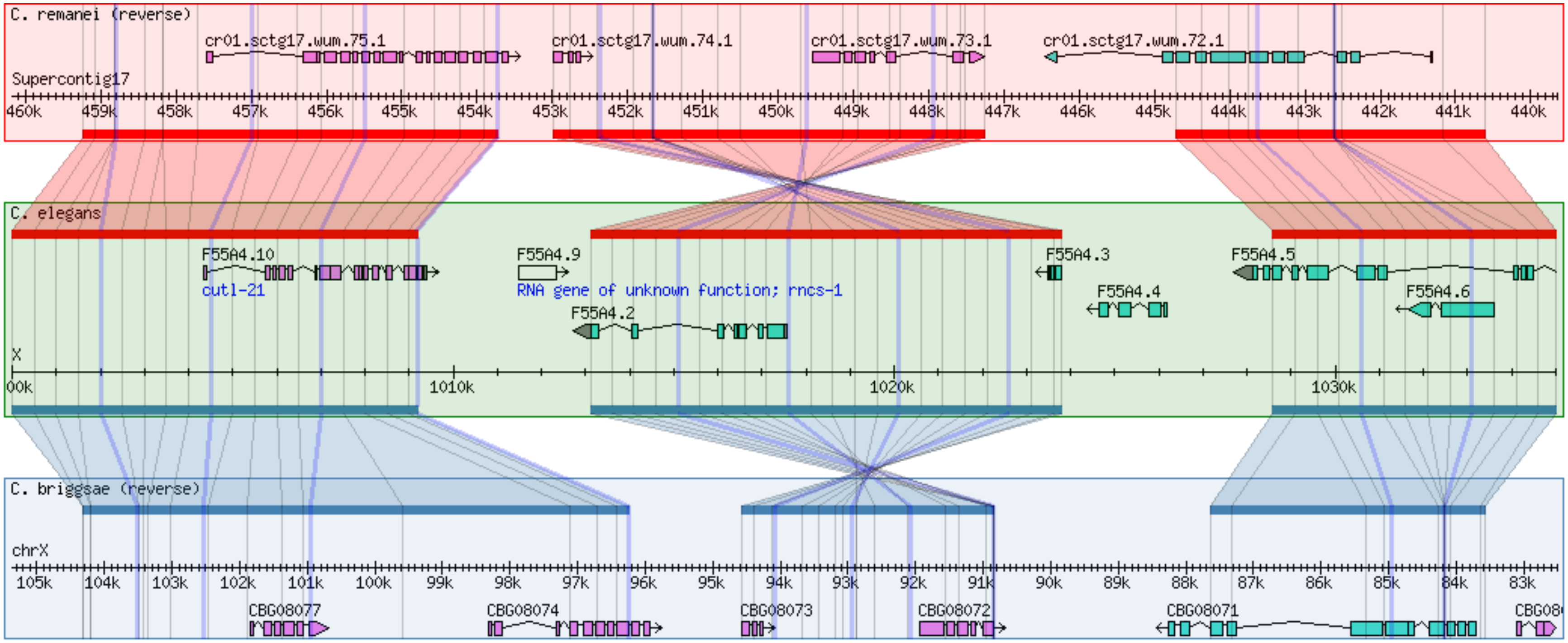
The Legume Information System

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...

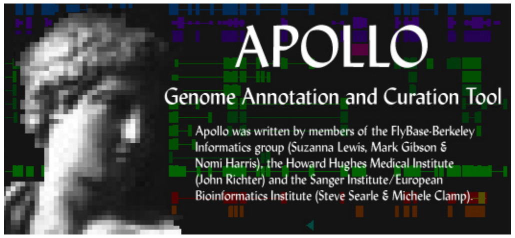
Pseudomonas Genome Database v2

iPlant Collaborative

NSF

Branding ideas…

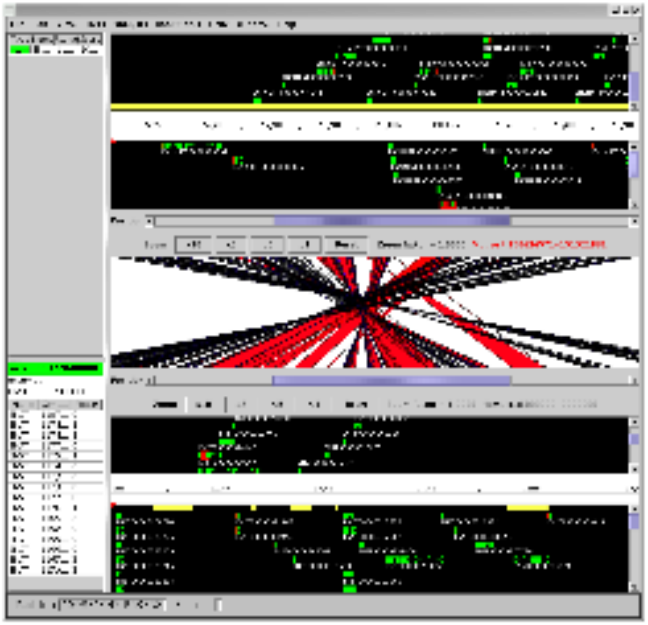
iPlant Collaborative™

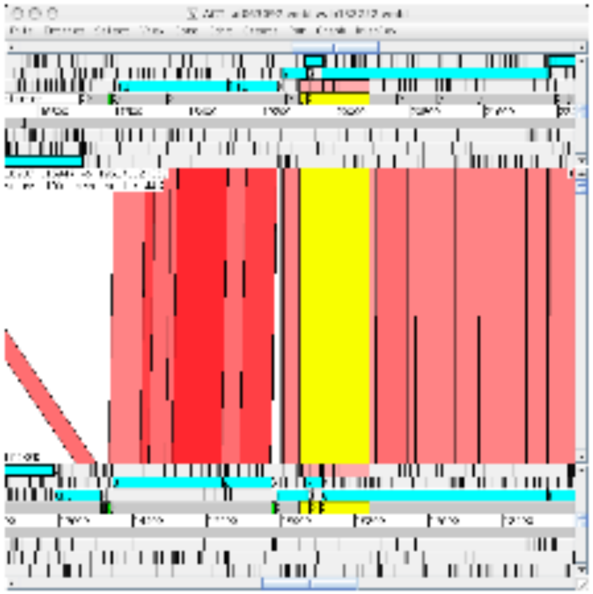


APOLLO

Genome Annotation and Curation Tool

Apollo was written by members of the FlyBase-Berkeley Informatics group (Suzanna Lewis, Mark Gibson & Nomi Harris), the Howard Hughes Medical Institute (John Richter) and the Sanger Institute/European Bioinformatics Institute (Steve Searle & Michele Clamp).





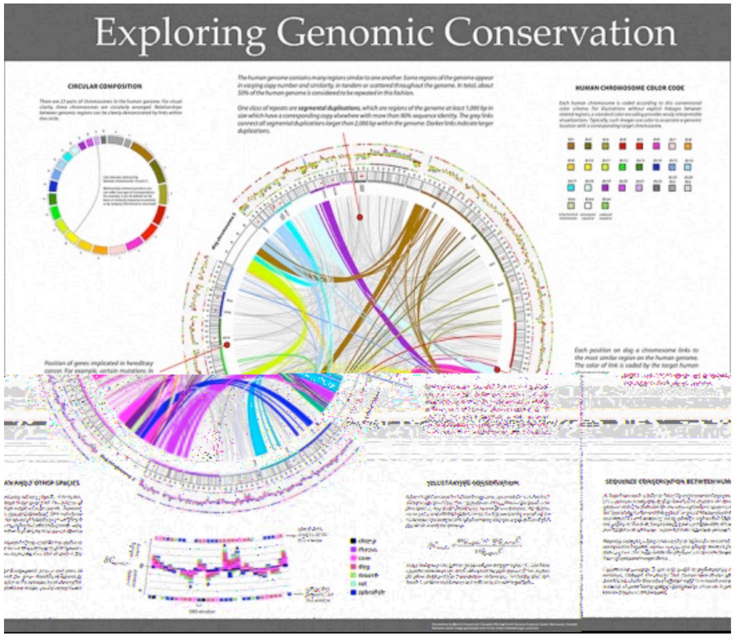
Desktop Synteny Viewers: Apollo and Artemis

Debating the relative merits of Apollo* and Artemis‡

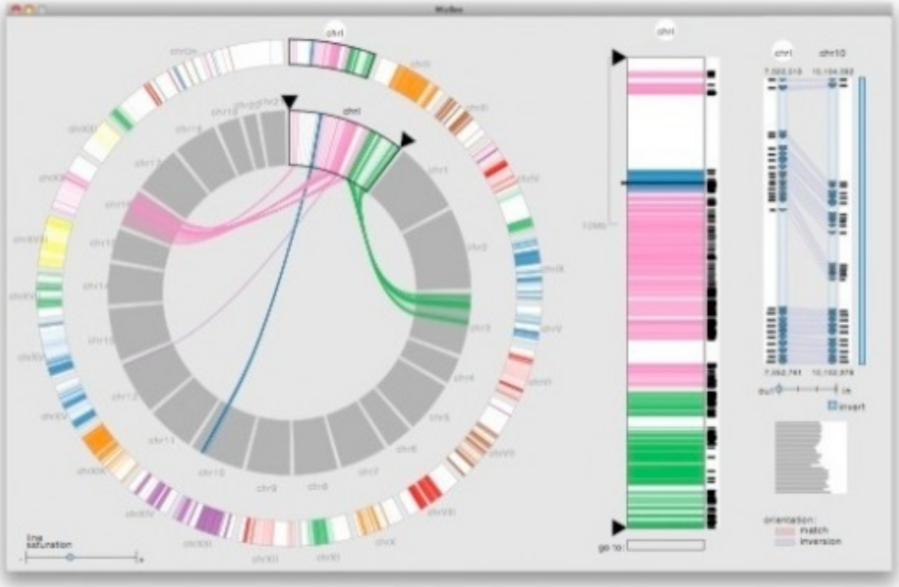
*Apollo is better

‡Artemis is also better

Other non-GMOD Browsers

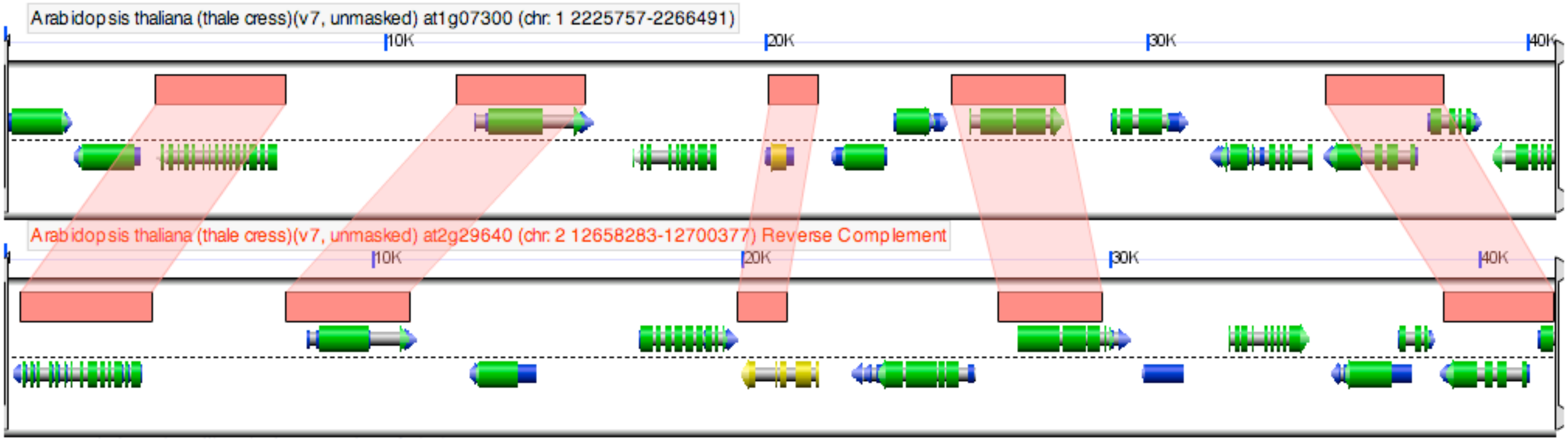

Exploring Genomic Conservation
CIRCULAR COMPOSITION
HUMAN CHROMOSOME COLOR CODE
SEQUENCE CONSERVATION BETWEEN HUM

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



http://www.mizbee.org

Other non-GMOD Browsers



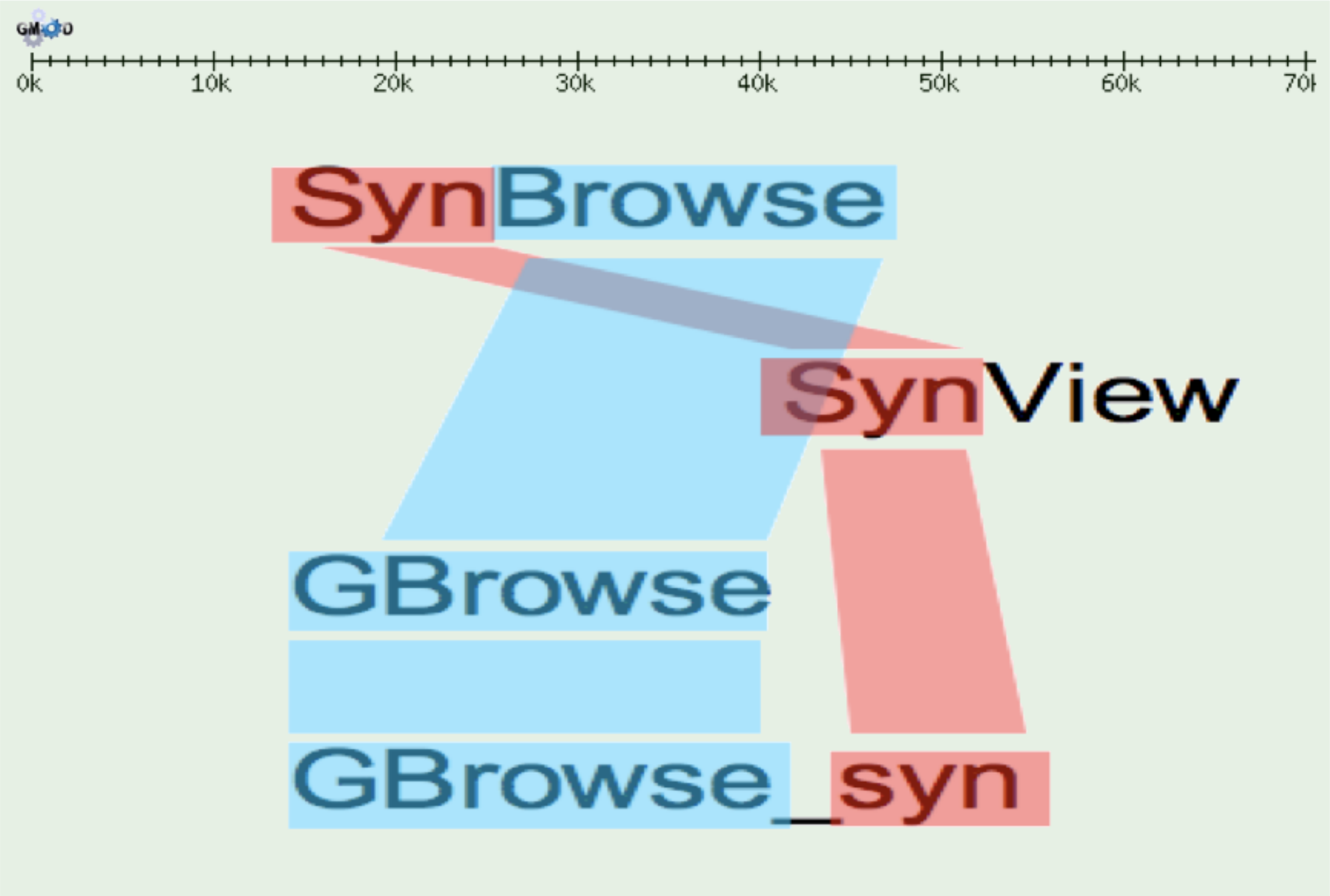
Arabidopsis thaliana (thale cress)(v7, unmasked) at1g07300 (chr: 1 2225757-2266491)

Arabidopsis thaliana (thale cress)(v7, unmasked) at2g29640 (chr: 2 12658283-12700377) Reverse Complement

CoGe *A new kind of Comparative Genomics*

http://synteny.cnr.berkeley.edu/CoGe/

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

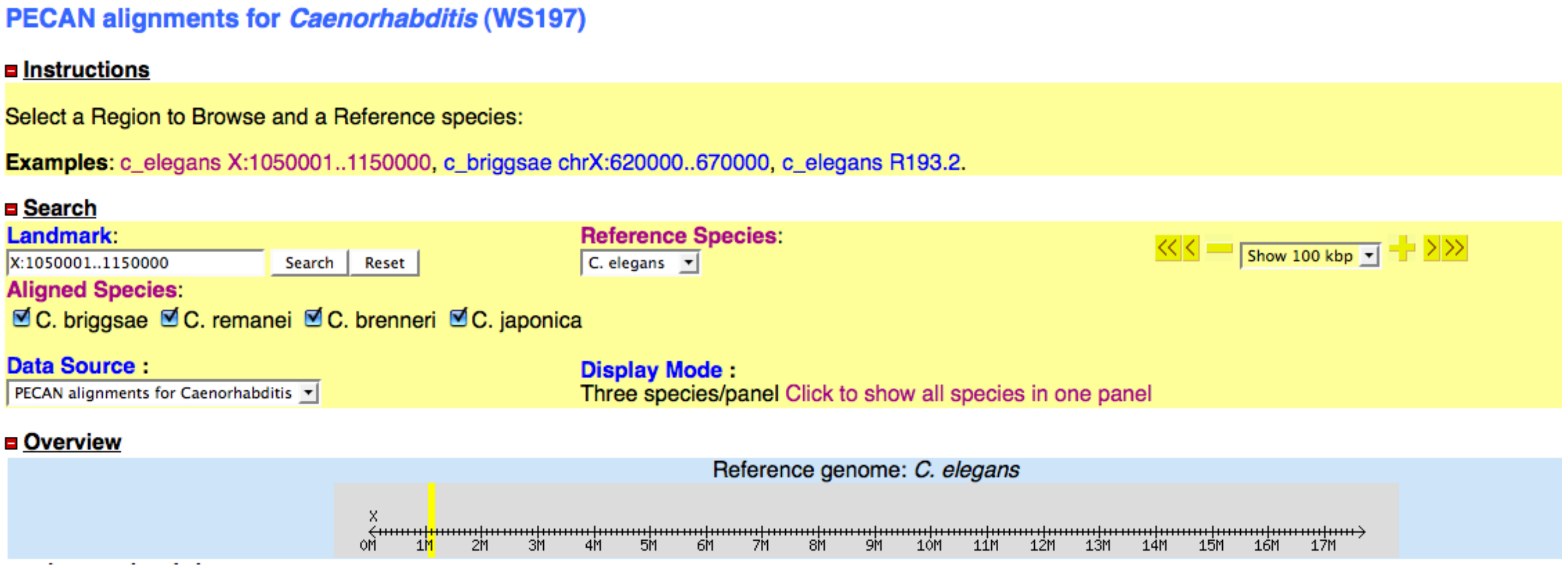
GBrowse_syn:

- 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-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: C. elegans

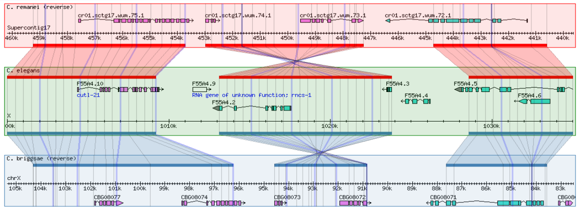
Aligned Species: ☑ C. briggsae ☑ C. remanei ☑ C. brenneri ☑ C. japonica

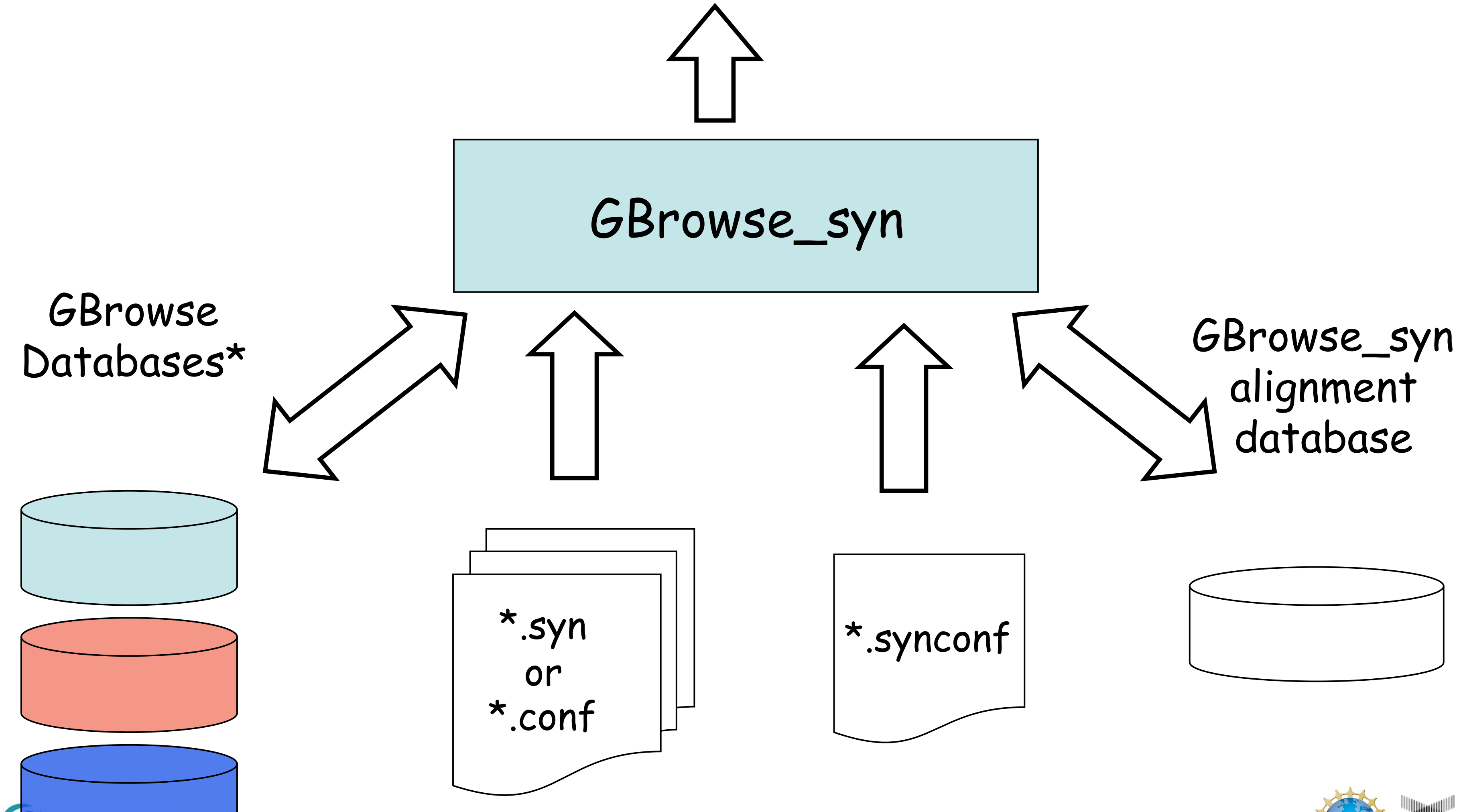
Data Source : PECAN alignments for Caenorhabditis

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

*.syn or *.conf

*.synconf

Species config.

Master config.

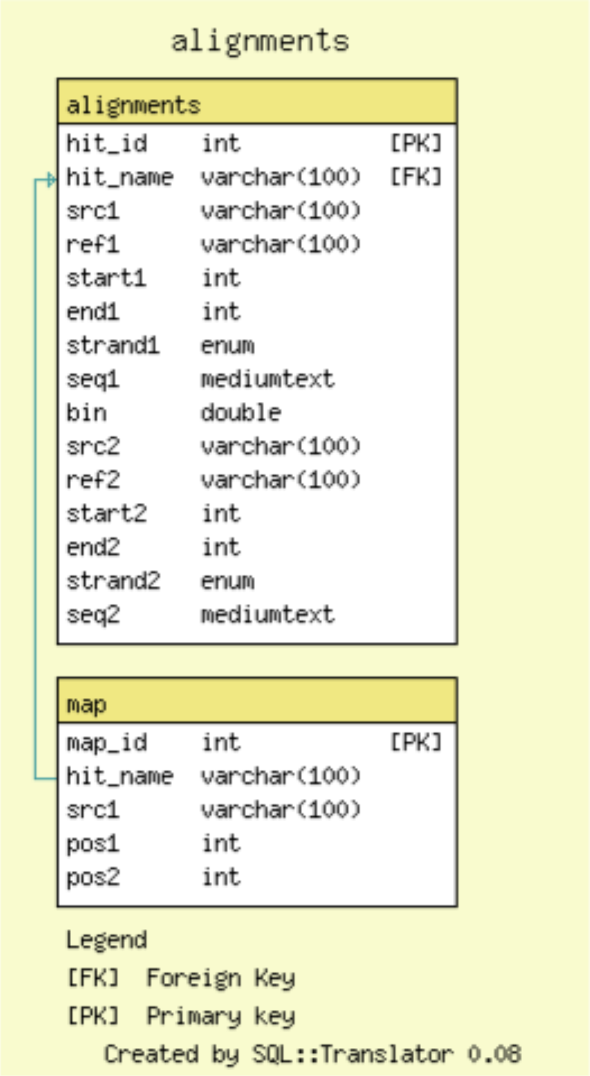
GBrowse_syn Architecture

[GBrowse]

[GBrowse]

Bio::DB::GFF
species1

Bio::DB::GFF
species3



alignments

alignments		
hit_id	int	[PK]
hit_name	varchar(100)	[FK]
src1	varchar(100)	
ref1	varchar(100)	
start1	int	
end1	int	
strand1	enum	
seq1	mediumtext	
bin	double	
src2	varchar(100)	
ref2	varchar(100)	
start2	int	
end2	int	
strand2	enum	
seq2	mediumtext	

map		
map_id	int	[PK]
hit_name	varchar(100)	
src1	varchar(100)	
pos1	int	
pos2	int	

Legend
[FK] Foreign Key
[PK] Primary key

Created by SQL::Translator 0.08

Bio::DB::GFF
species2

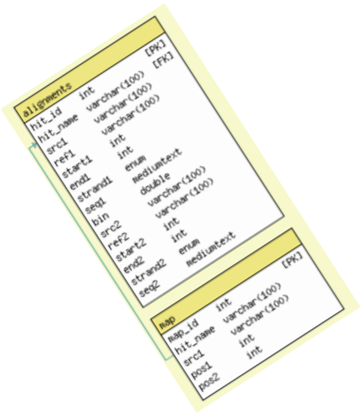
Bio::DB::GFF
species4

[GBrowse]

[GBrowse]

Getting Data into GBrowse_syn

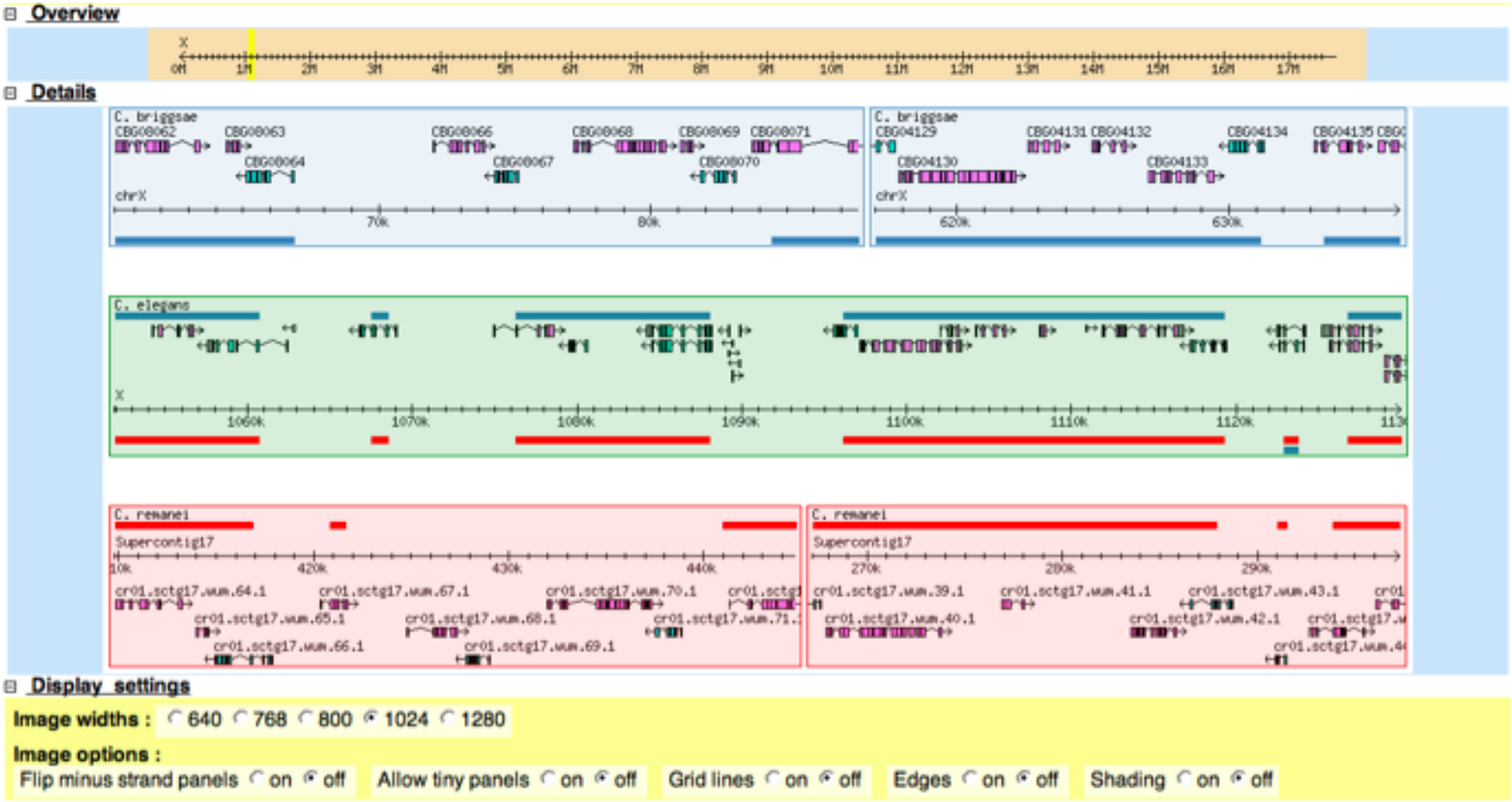
CLUSTALW PECAN
MSF *ad hoc* tab-delimited
FASTA STOCKHOLM
GFF3 etc…



alignments		
hit_id	int	[PK]
hit_name	varchar(100)	[FK]
src1	varchar(100)	
ref1	varchar(100)	
start1	int	
end1	int	
strand1	enum	
seq1	mediumtext	
bin	double	
src2	varchar(100)	
ref2	varchar(100)	
start2	int	
end2	int	
strand2	enum	
seq2	mediumtext	

map		
map_id	int	[PK]
hit_name	varchar(100)	
src1	varchar(100)	
pos1	int	
pos2	int	

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)

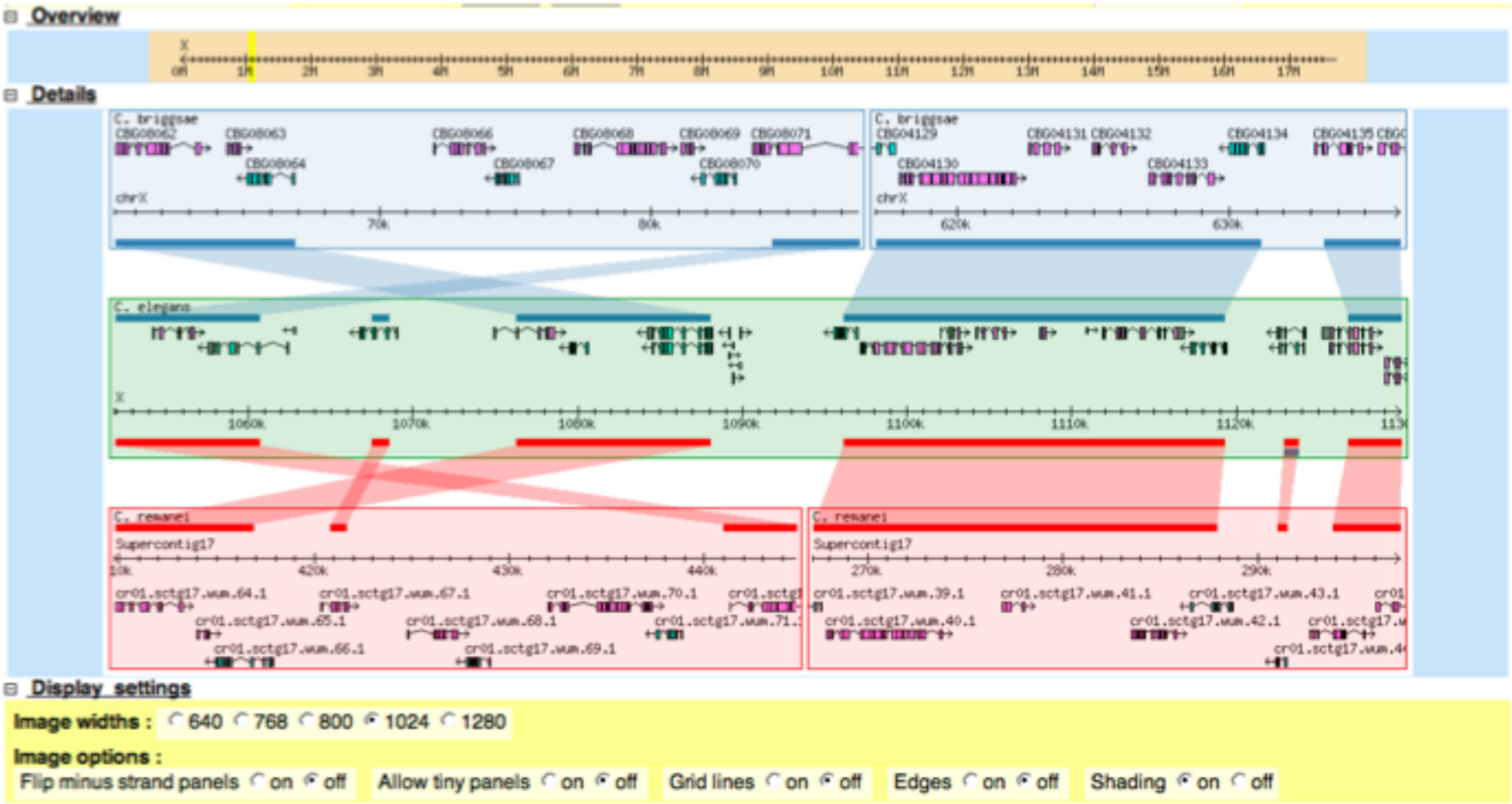


Image widths : ○ 640 ○ 768 ○ 800 ● 1024 ○ 1280

Image options :

Flip minus strand panels ○ on ● off

Allow tiny panels ○ on ● off

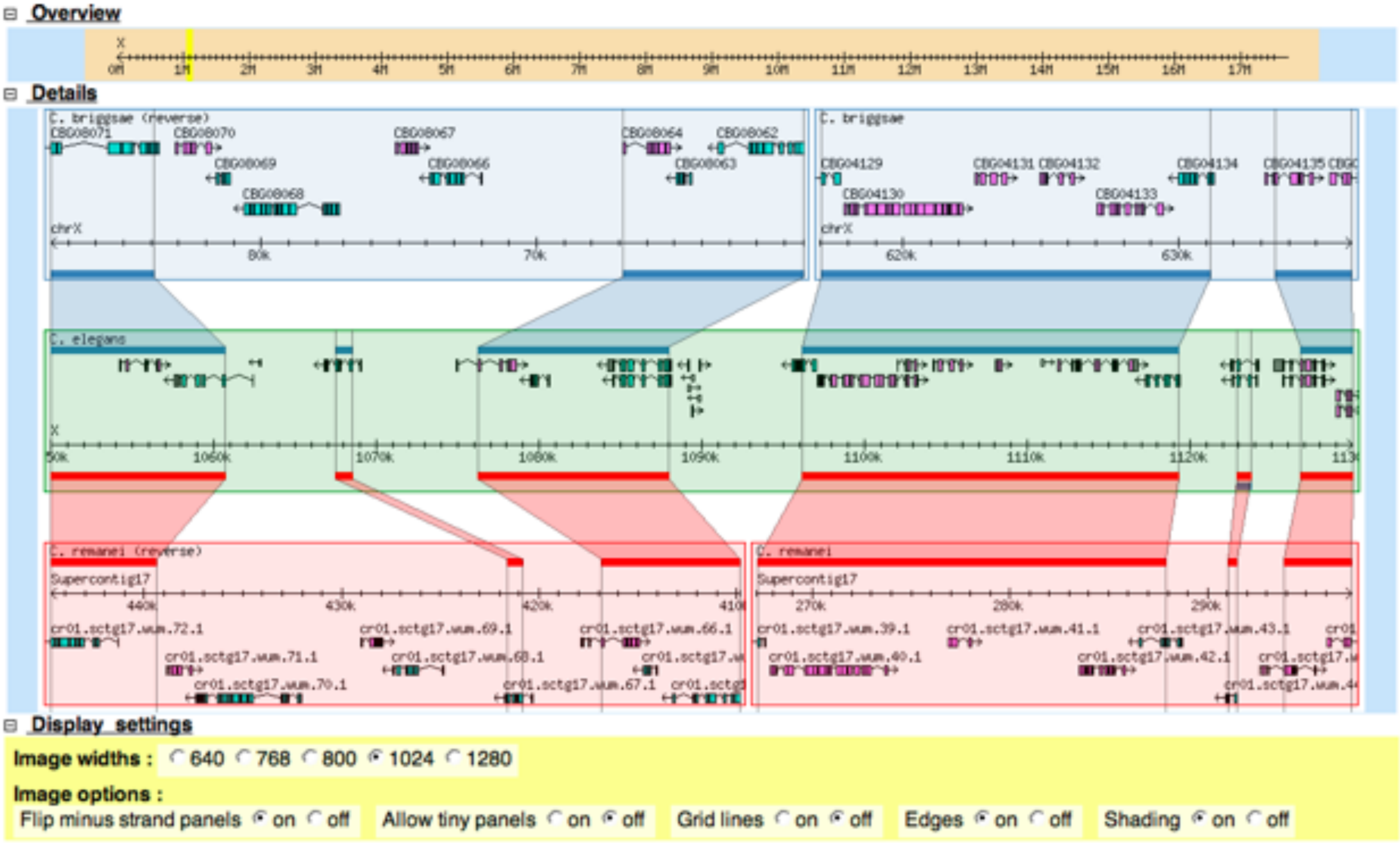
Grid lines ○ on ● off

Edges ○ on ● off

Shading ● on ○ off

iPlant Collaborative™

Gbrowse_syn: quick tour (strand correction)



⊟ Overview

⊟ Details

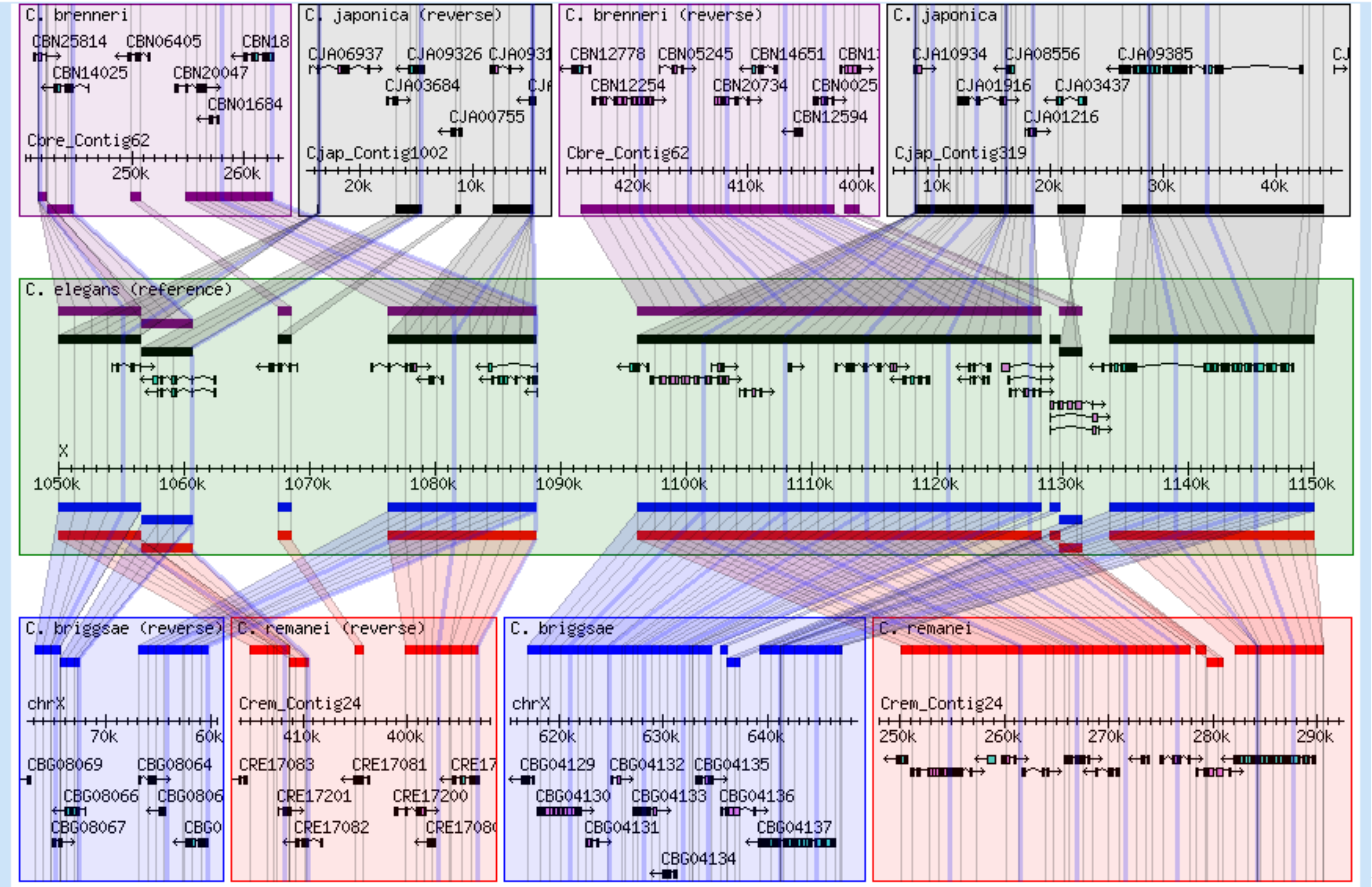
⊟ Display settings

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

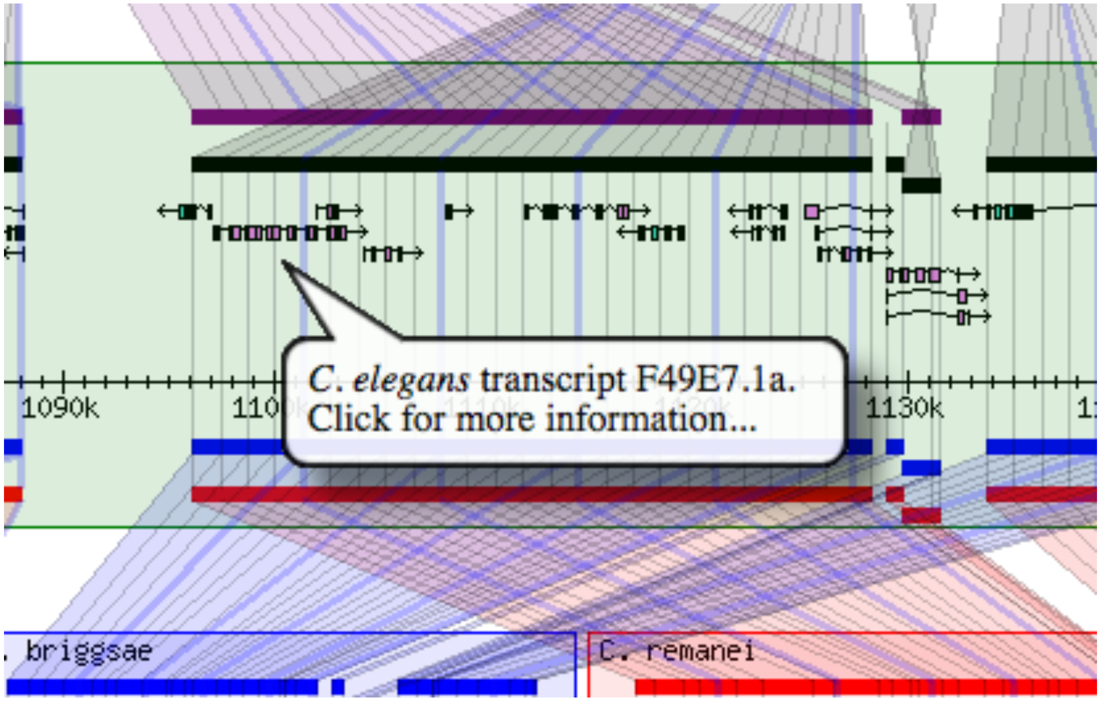
Image options :

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

Optional “All in one” view

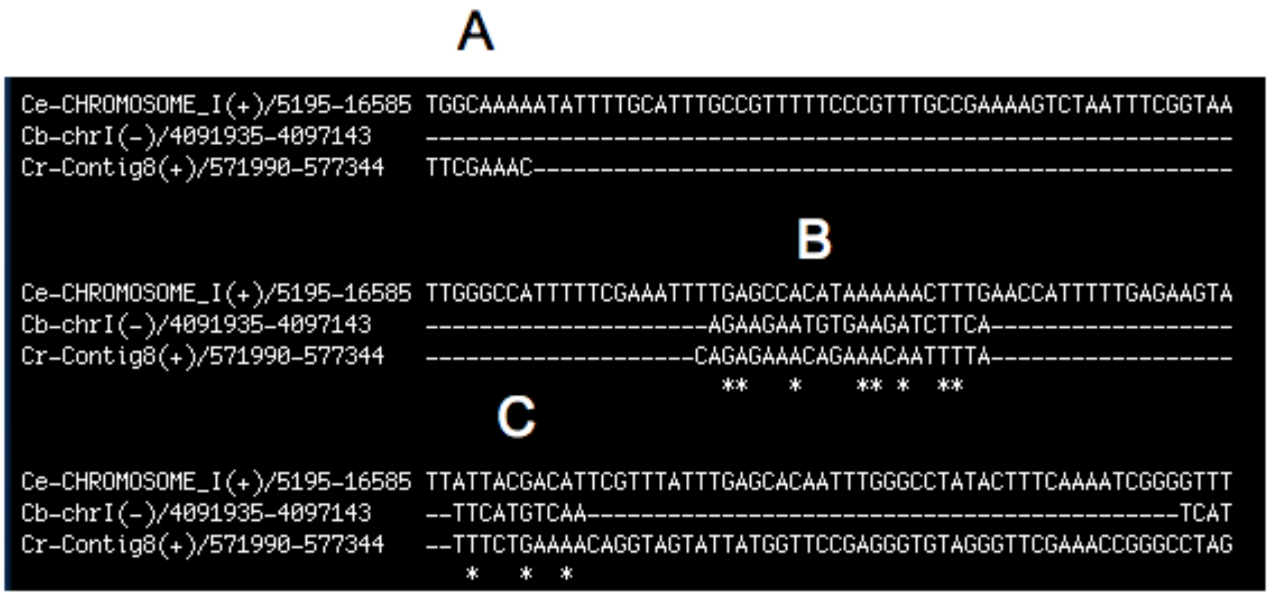


Adding markup to the annotations



C. elegans transcript F49E7.1a.
Click for more information...

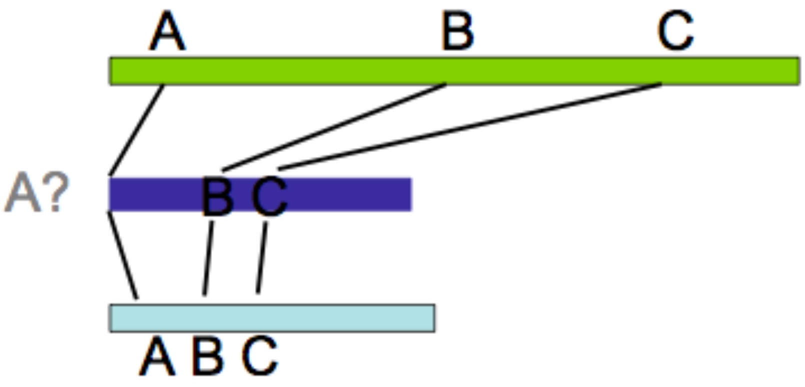
Problem : How to use Insertions/Deletion data



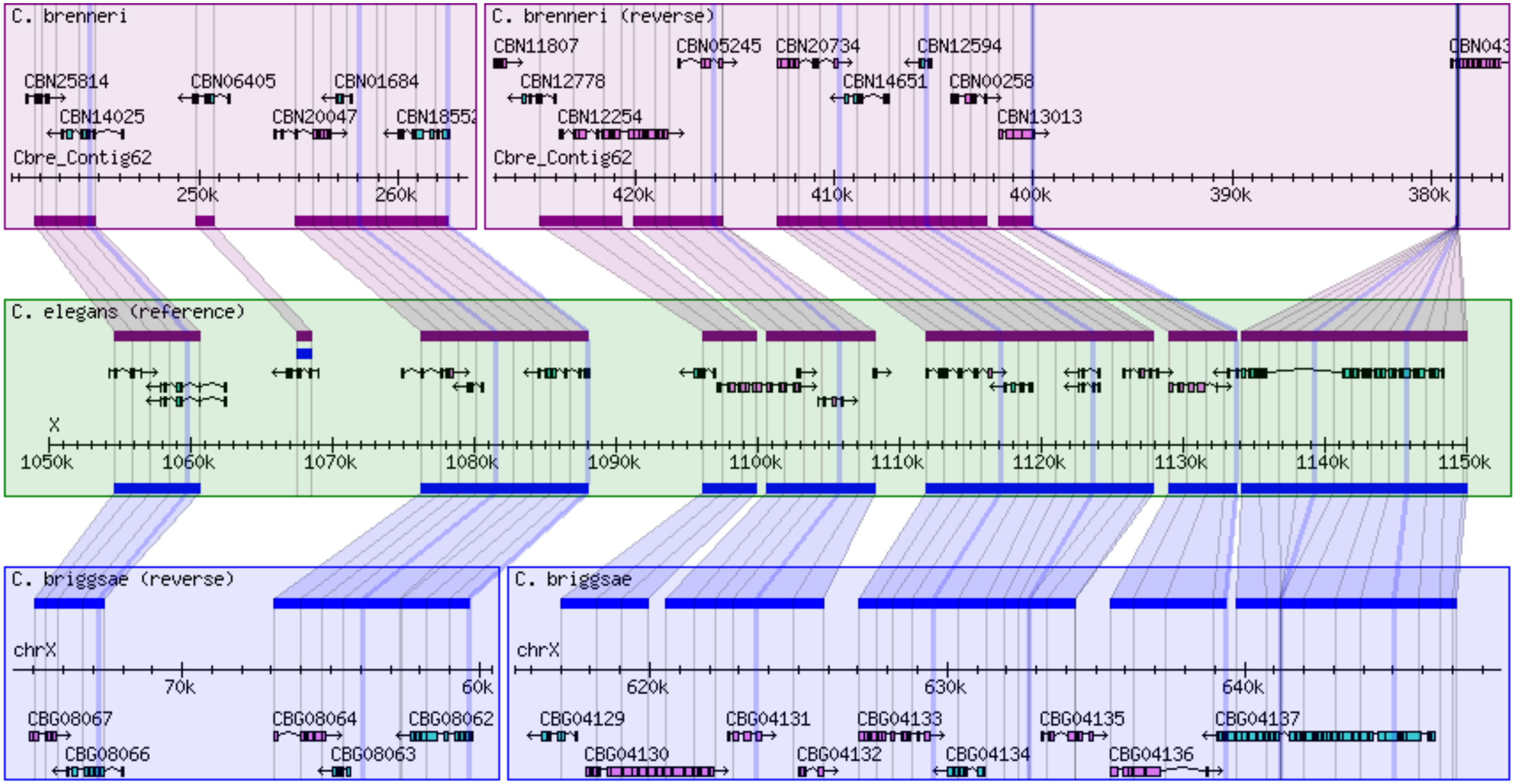
```
Ce-CHROMOSOME_I(+)/5195-16585 TGGCAAAAATATTTTGCATTTGCCGTTTTTCCCGTTTGCCGAAAAGTCTAATTTCGGTAA
Cb-chrI(-)/4091935-4097143    ------------------------------------------------------------
Cr-Contig8(+)/571990-577344   TTCGAAAC----------------------------------------------------
```

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

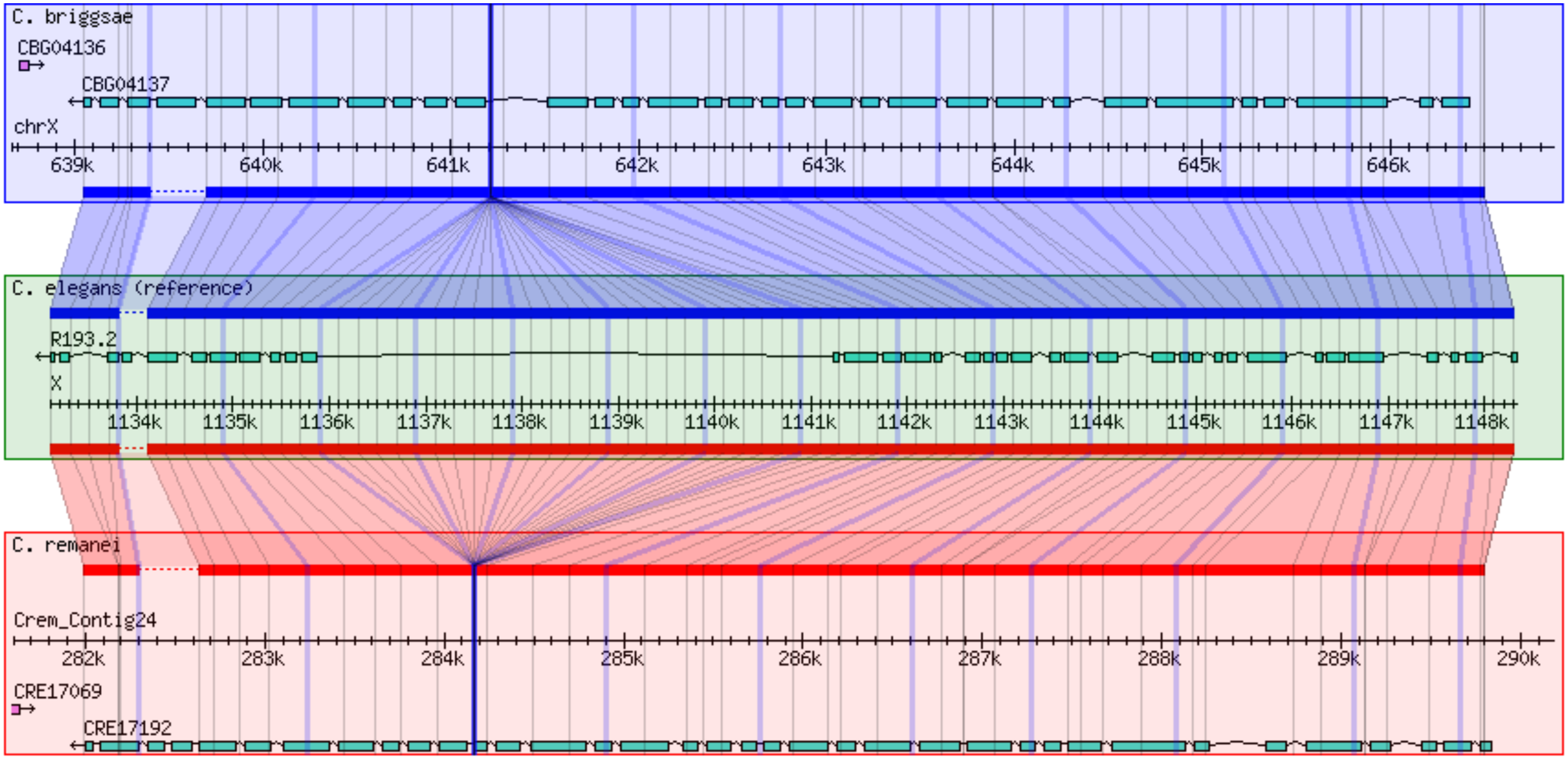
```
Ce-CHROMOSOME_I(+)/5195-16585 TTATTACGACATTCGTTTATTTGAGCACAATTTGGGCCTATACTTTCAAAATCGGGGTTT
Cb-chrI(-)/4091935-4097143    --TTCATGTCAA--------------------------------------------TCAT
Cr-Contig8(+)/571990-577344   --TTTCTGAAAACAGGTAGTATTATGGTTCCGAGGGTGTAGGGTTCGAAACCGGGCCTAG
                       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



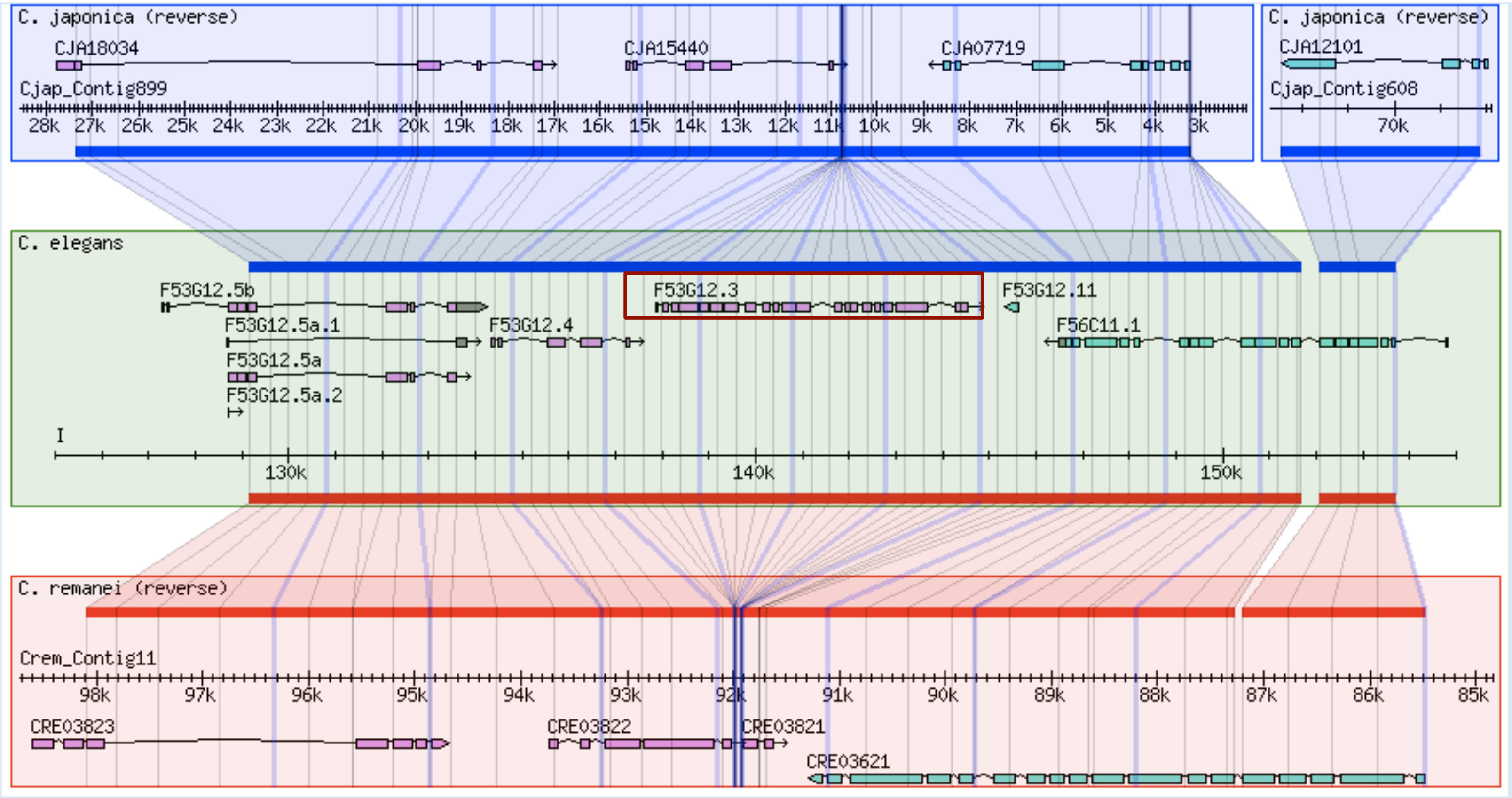
Tracking Indels with grid lines



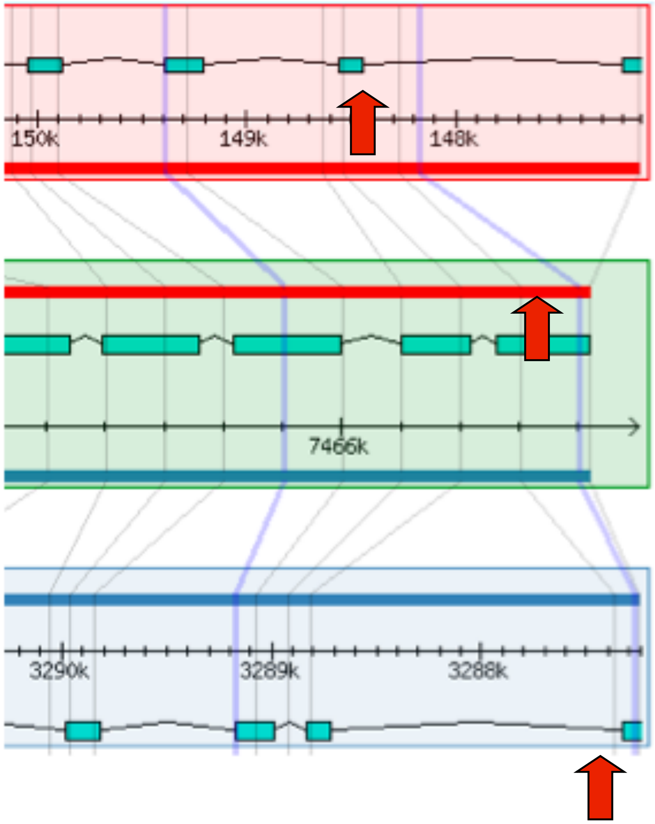
Evolution of Gene Structure



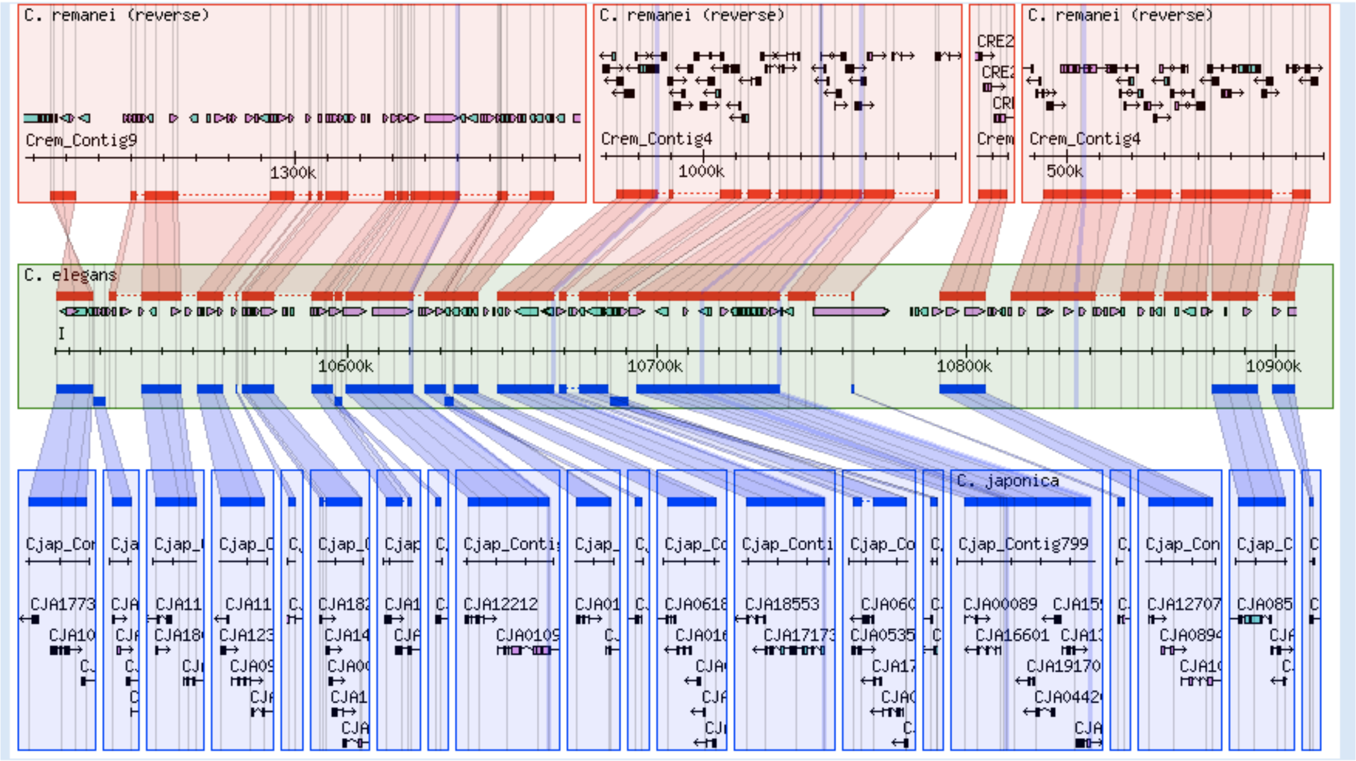
Putative gene or loss



Comparing gene models



Comparing assemblies

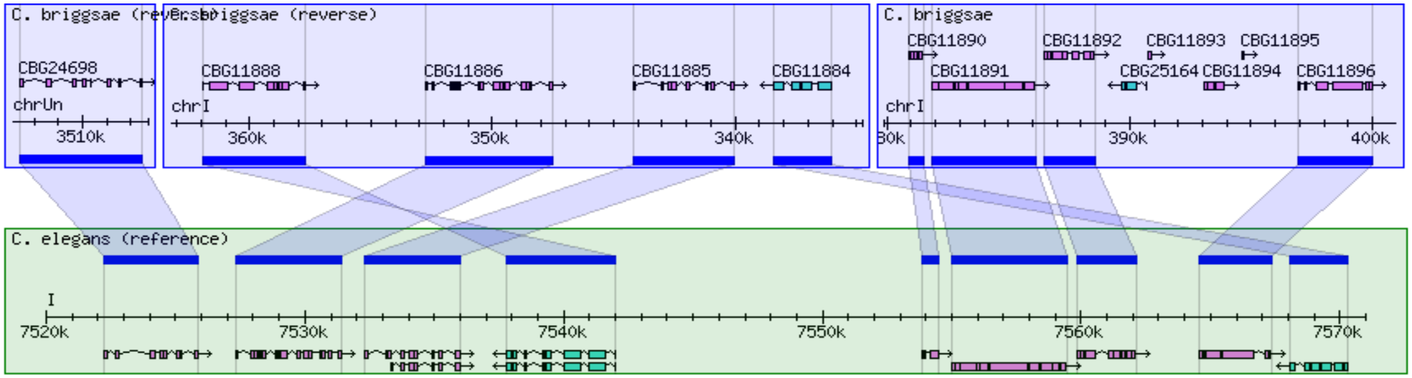


Not bad

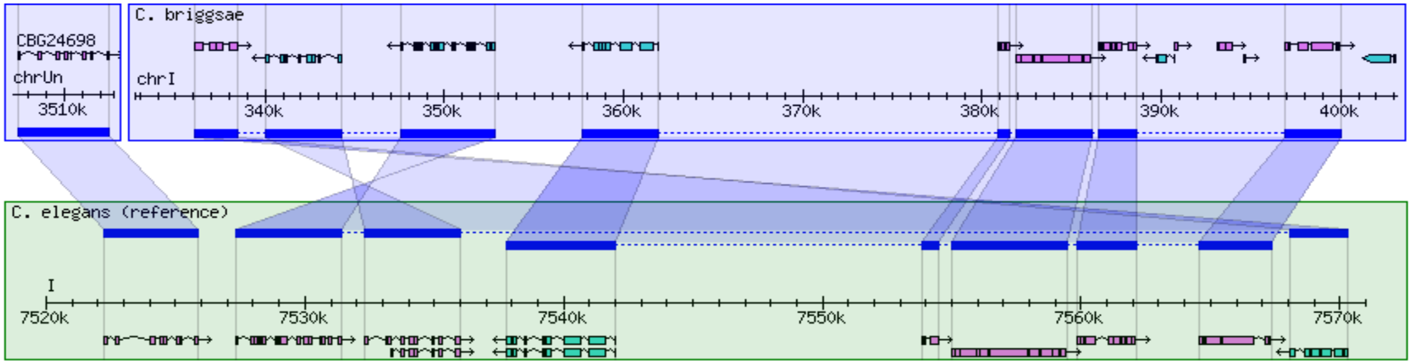
Needs work

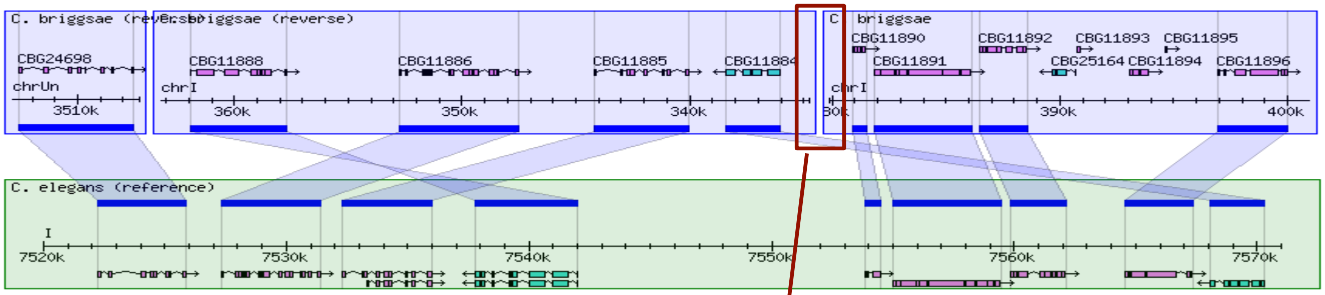
Getting the most out of small aligned regions or orthology-only data

Gene Orthology



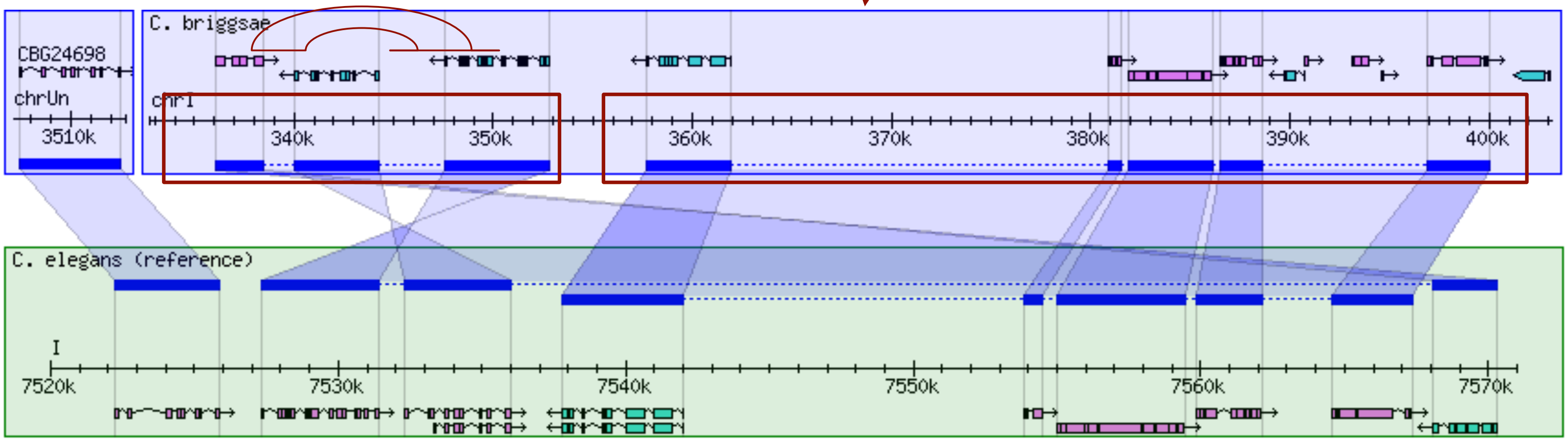
Chained Orthologs





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.

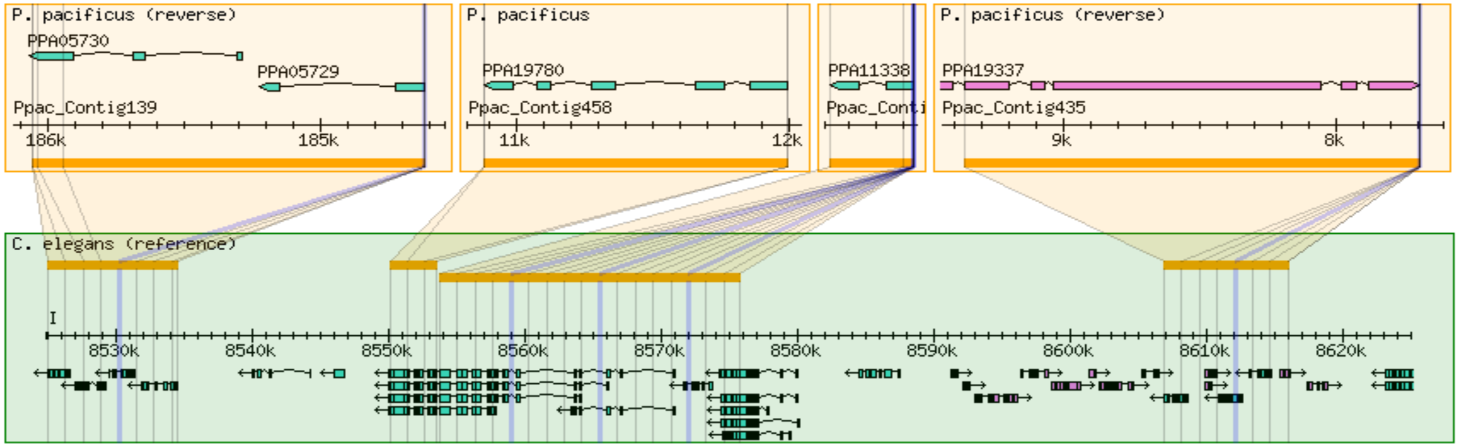
length displayed target = (length target/length reference)* length displayed reference

What if the aligned DNA sequences are too distant?

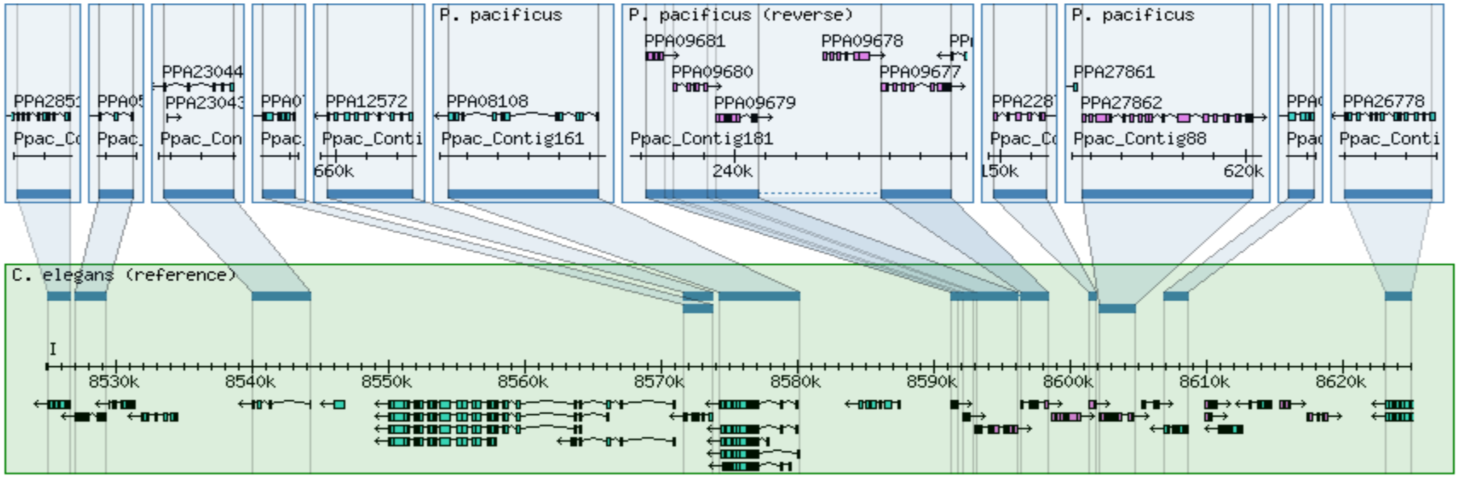
!=

*

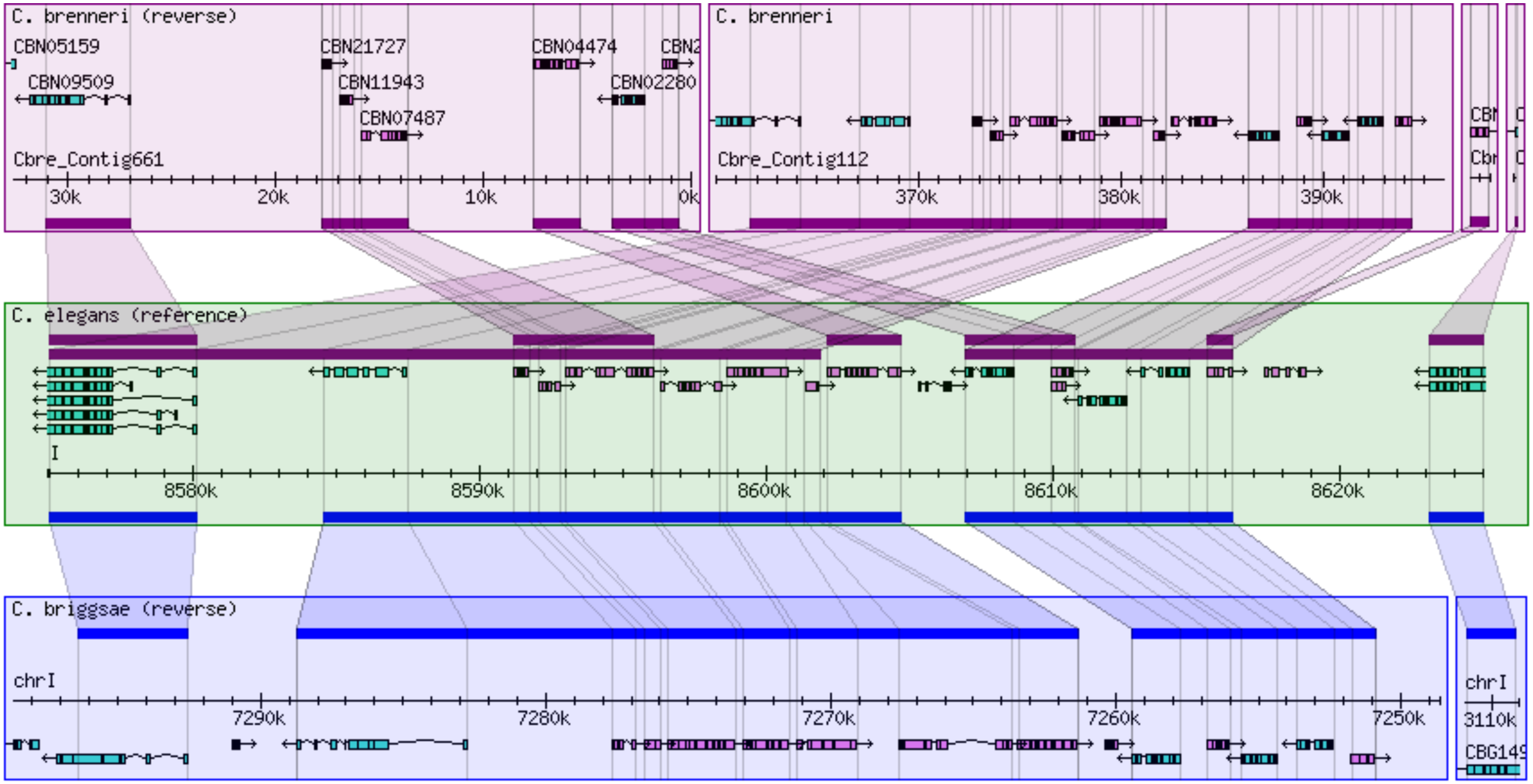
Pecan alignments



Protein orthology based Synteny blocks



What about segmental duplications?


C. brenneri (reverse)
CBN05159
CBN09509
CBN21727
CBN11943
CBN07487
CBN04474
CBN2
CBN02280
Cbre_Contig661
30k
20k
10k
0k
C. brenneri
Cbre_Contig112
370k
380k
390k
CBN
Cbr
C
C
C. elegans (reference)
I
8580k
8590k
8600k
8610k
8620k
C. briggsae (reverse)
chrI
7290k
7280k
7270k
7260k
7250k
chrI
3110k
CBG149

The Future of GBrowse_syn*

- Integration with GBrowse 2.0
- "On the fly" sequence alignment view
- AJAX-based user interface and navigation (Jbrowse_syn)
- 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

Projects

iPlant Collaborative™ modENCODE WormBase GMOD

Funding

NSF