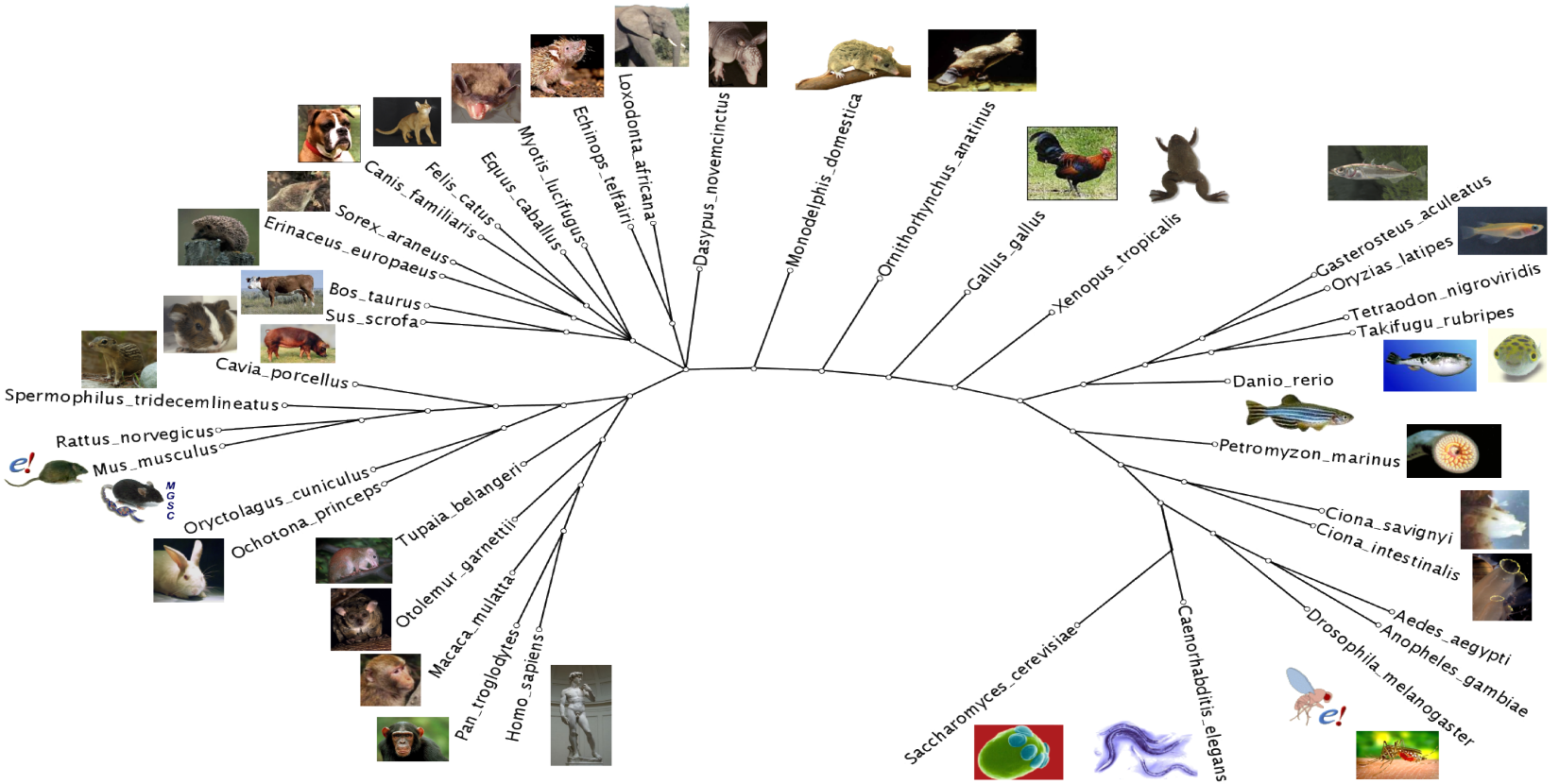
Comparative Genomics in Ensembl



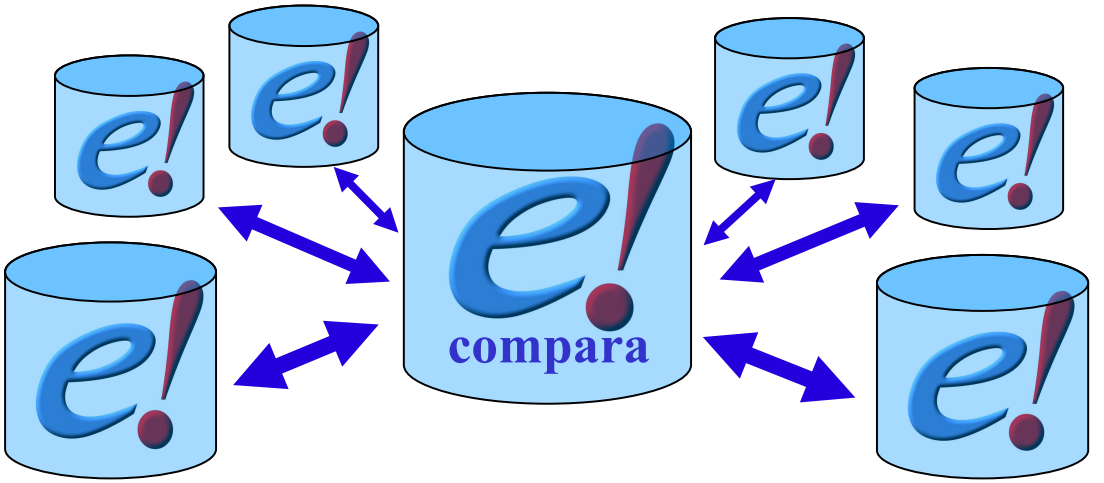
Javier Herrero

http://www.ebi.ac.uk/~jherrero/

EBI - Wellcome Trust Genome Campus, UK

Sep 2008

Ensembl Compara



A single database which contains precalculated comparative genomics data and which is linked to all the Ensembl Species databases.

Access via web interface, perl API and mysql

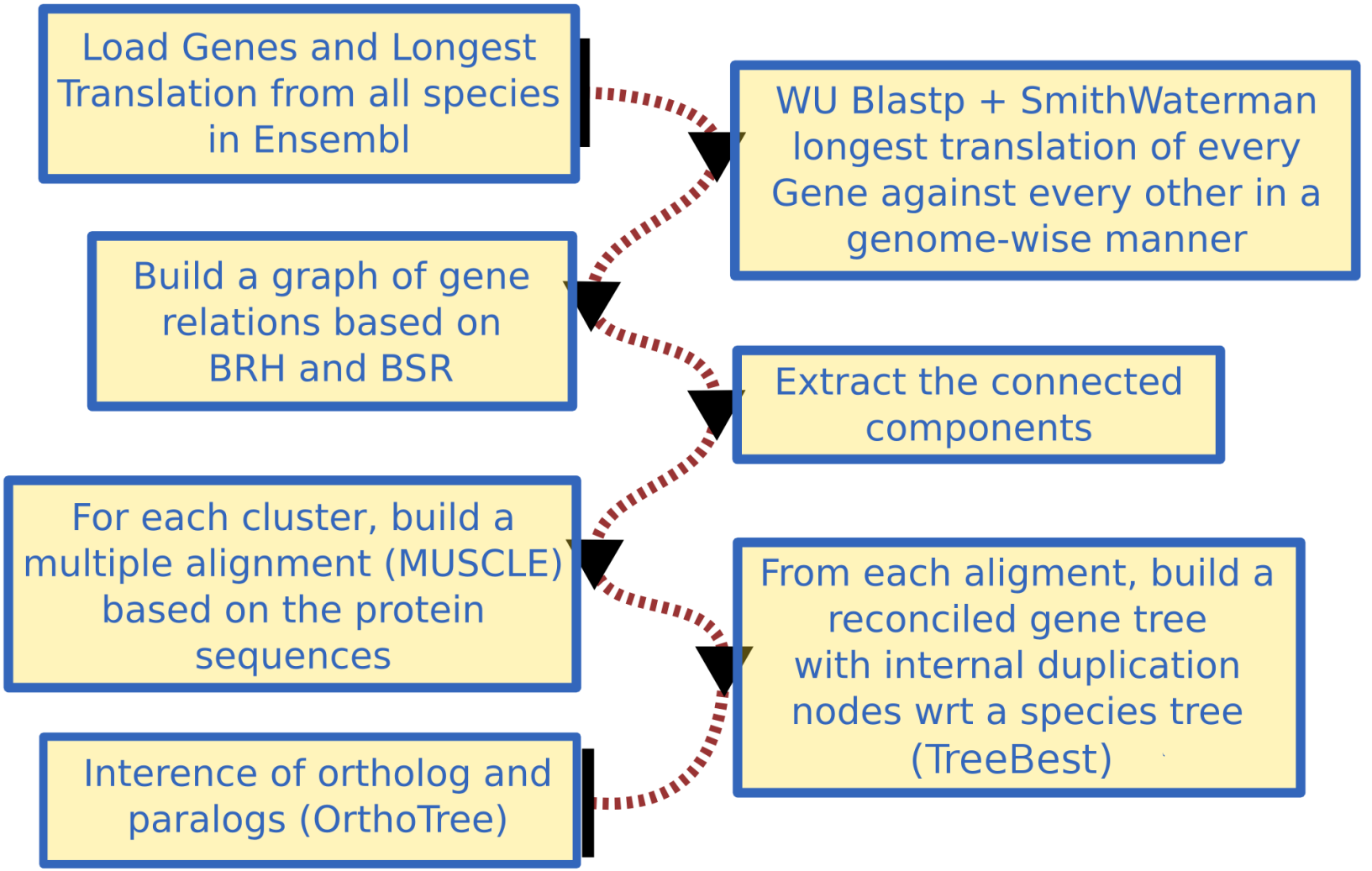
A production system for generating that database

Studying the evolution

- Comparing extant species
- Protein level
  - Multiple alignments
  - Gene Trees (protein trees)
- Genomic level
  - Pairwise alignments
  - Multiple alignments
  - Syntenies

- Conserved regions
- Non-conserved-regions
- Lineage-specific changes

Protein homology

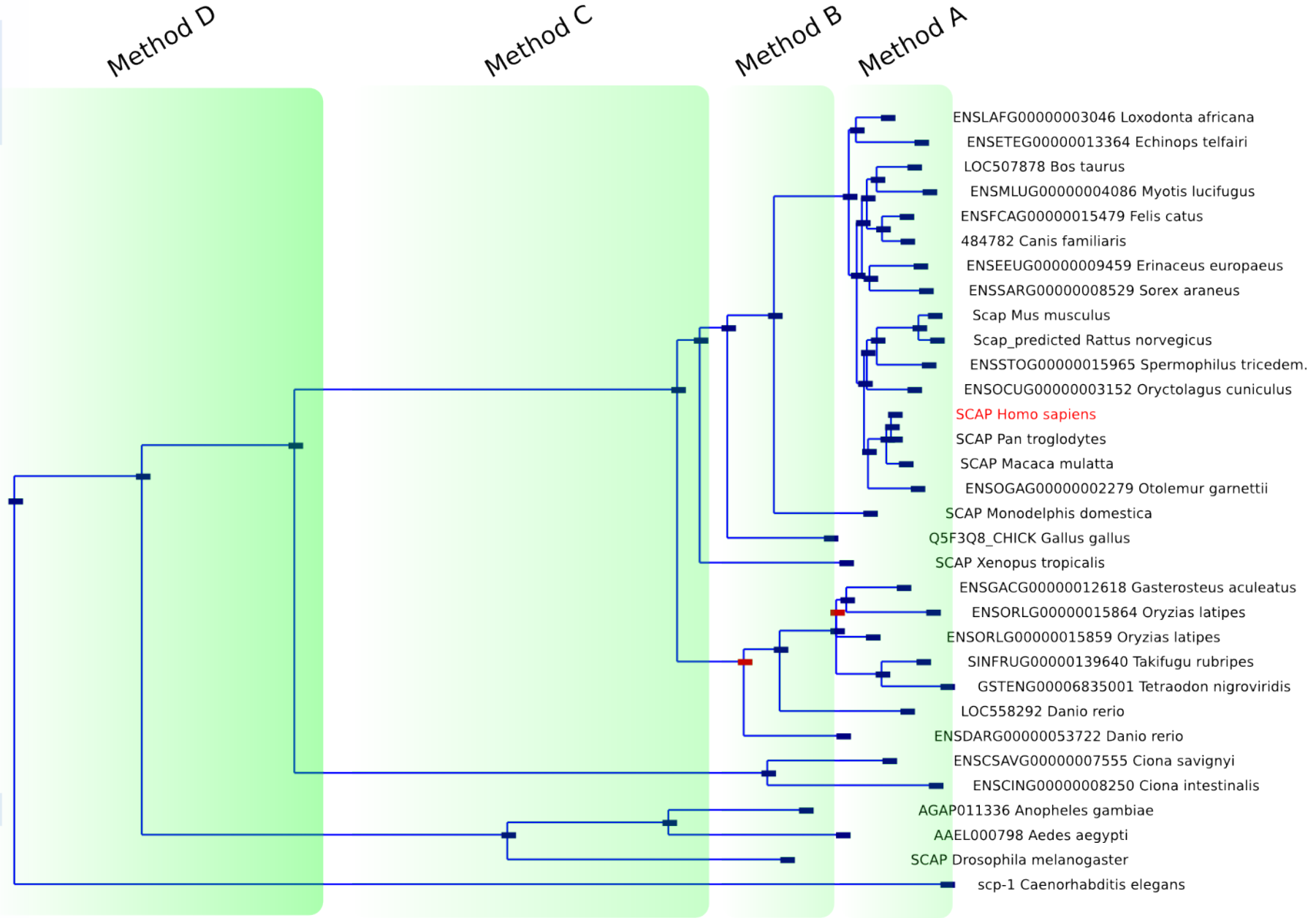


BSR: Blast Score Ratio. When 2 proteins P1 and P2 are compared, BSR=scoreP1P2/max(self-scoreP1 or self-scoreP2). The default threshold used in the initial clustering step is 0.33.

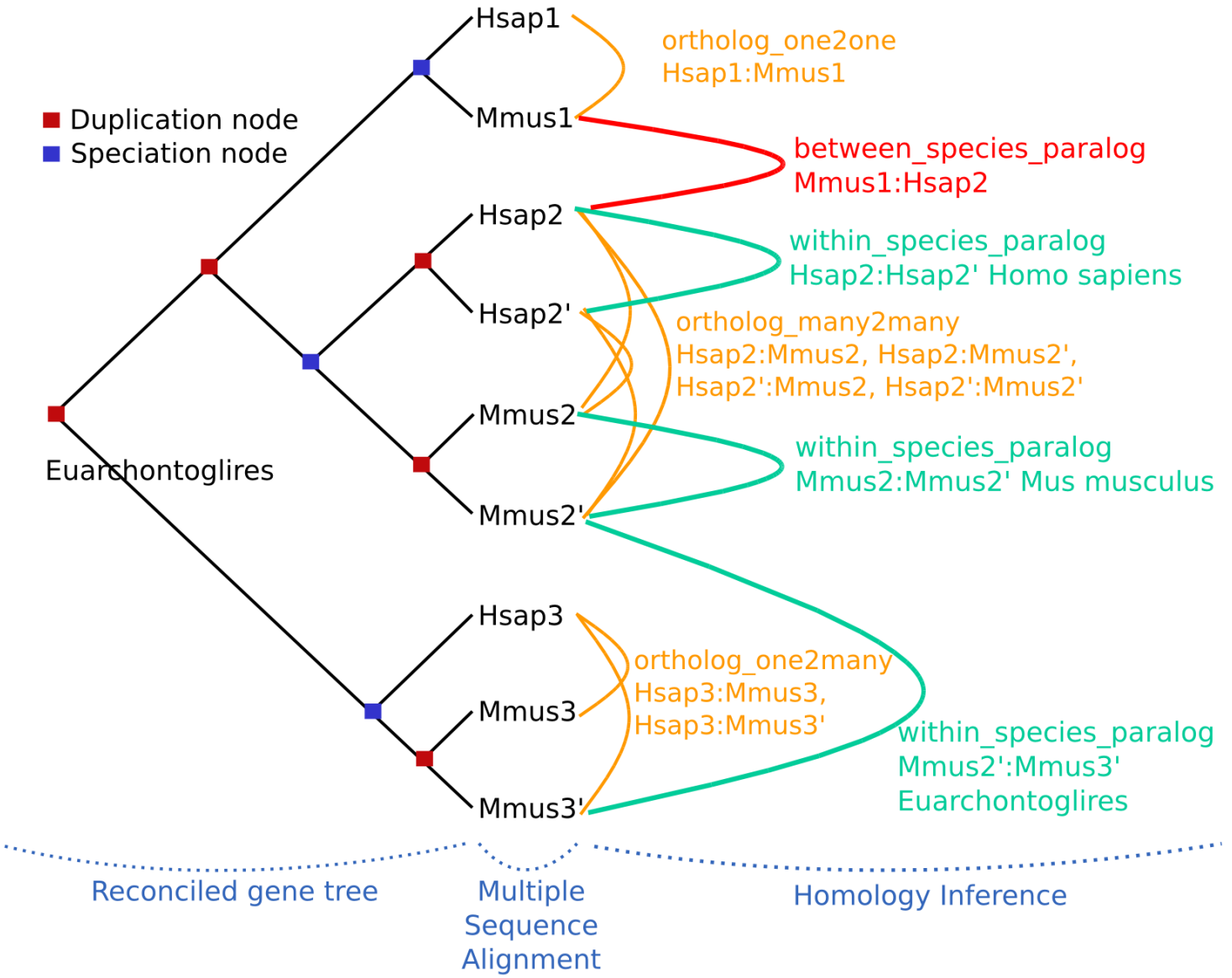
TreeBeST – treemerge algorithm

- ML-AA-WAG4 – WAG matrix aminoacidic model – Maximum Likelihood (PHYML)
- ML-NT-HKY85 – Hasegawa-Kishino-Yano nucleotidic model – Maximum Likelihood (PHYML)
- NJ-NT-p-distance – any substitutions – neighbor-joining with bootstrap
- NJ-NT-dN – non-syn substitutions – neighbor-joining with bootstrap
- NJ-NT-dS – synonymous substitutions – neighbor-joining with bootstrap
- Curated tree topology (if provided)

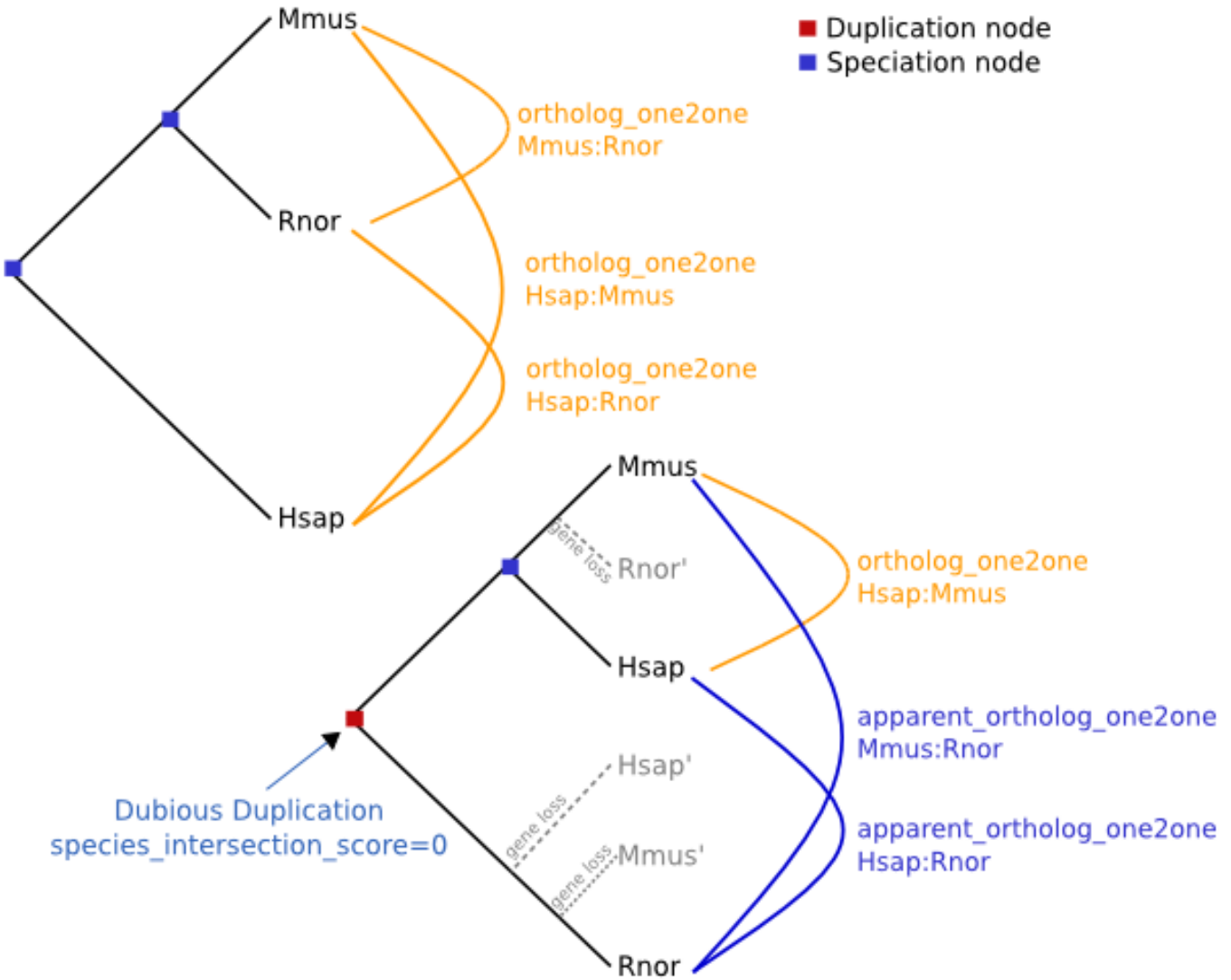
TreeBeST approach



Homology inference

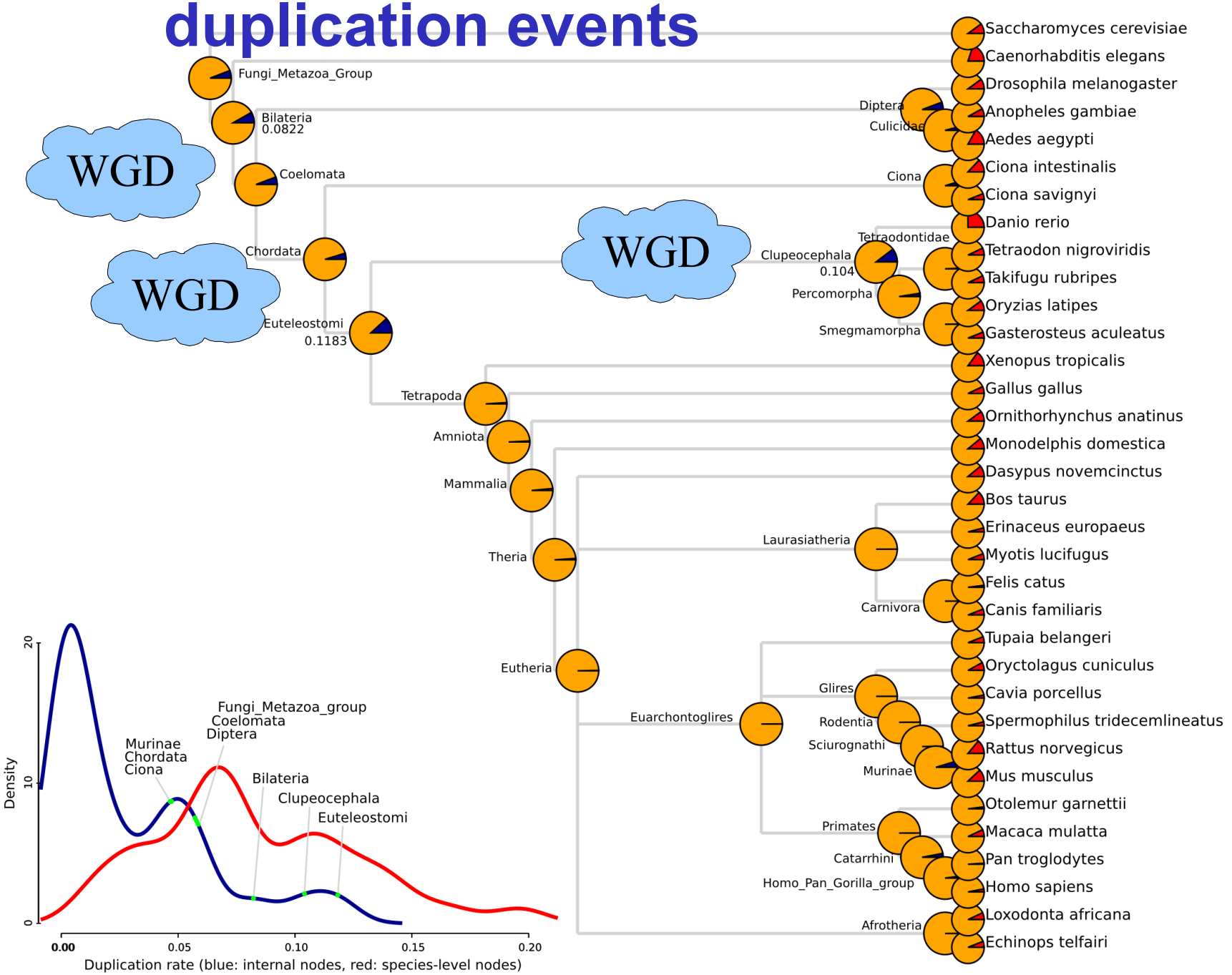


Dubious duplications



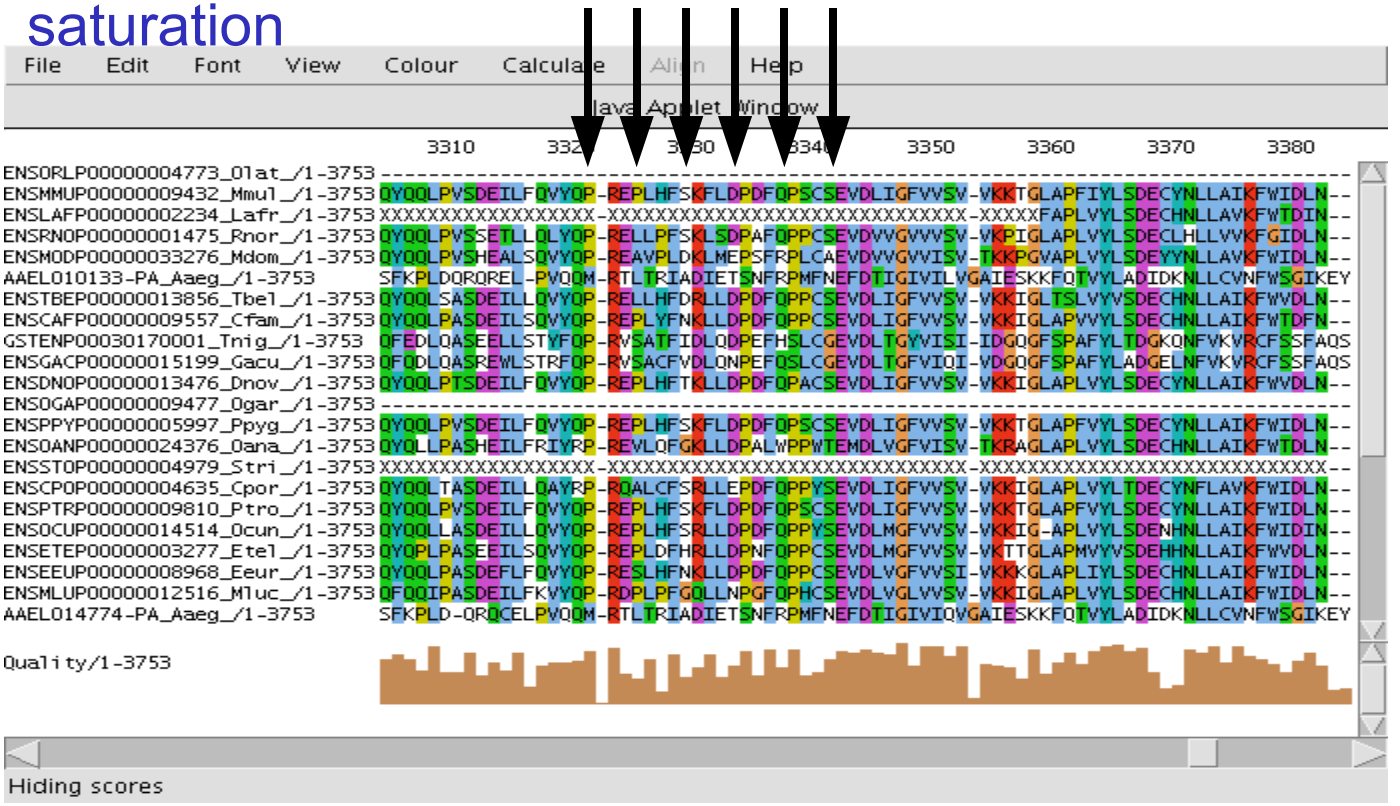
Orthologues : any gene pairwise relation where the ancestor node is a SPECIATION event.
Paralogues : any gene pairwise relation where the ancestor node is a DUPLICATION event.

Topological timing of duplication events

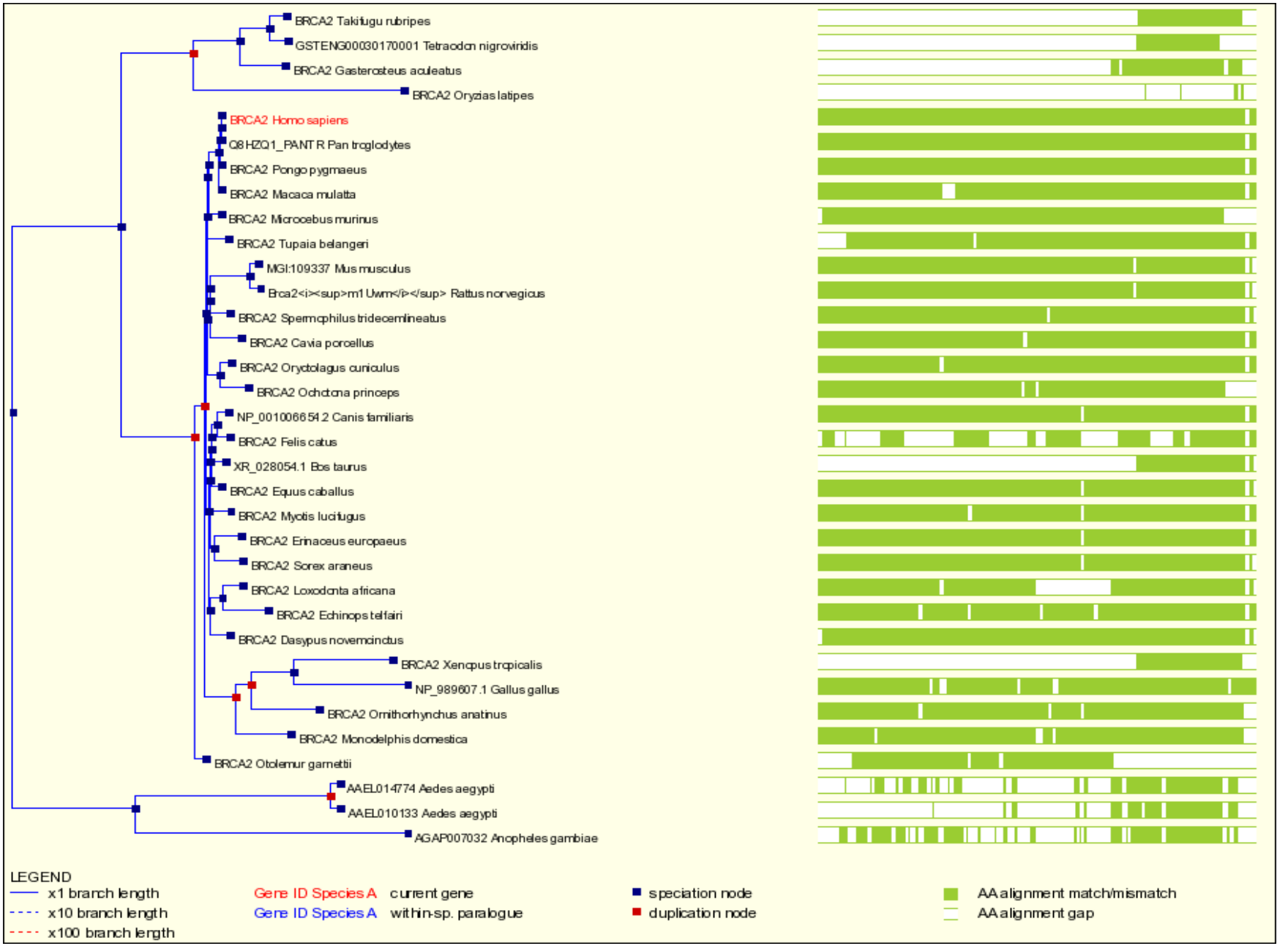


Sitewise dN/dS

- Looking for **nonneutral** evolution at **specific codons** in the alignments
- SLR by Massingham and Goldman (EBI)
  - Doable in 24hr x 400CPUs
  - SLREnsembl -- Choosing subtrees based on dS saturation

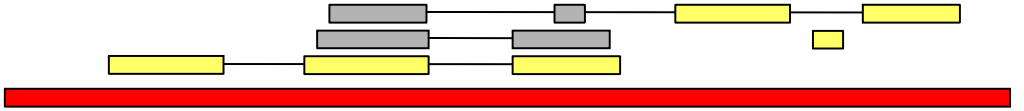


Gene Tree in Ensembl



Genomic Alignments

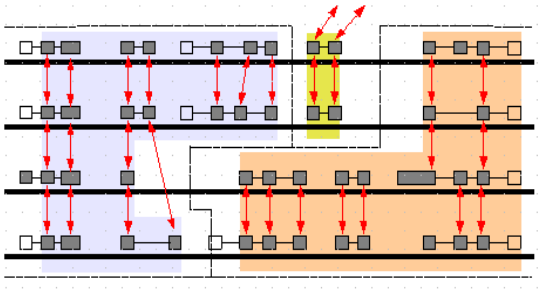
- BlastZ-Net
  - used to compare closely related pair of species
  - BlastZ-raw → BlastZ-chain → BlastZ-net



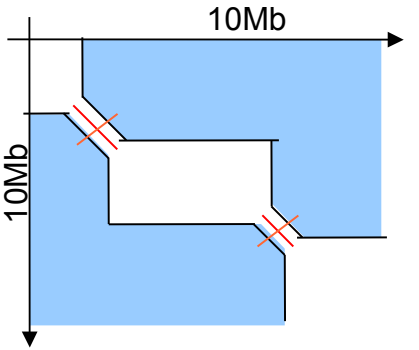
- Translated BLAT
  - used to compare more distant pair of species
  - we use the same approach (chain & net) starting from 50!
- Pecan (Mercator-Pecan)
  - multiple global alignments
  - all vs all coding exons wublastp → Mercator → Pecan on each syntenic block
- EPO (Enredo-Pecan-Ortheus)
  - Segmental duplications + multiple alignments + ancestral sequences inference
  - Anchors → Enredo → Pecan → Ortheus
- GERP (G. Cooper *et al.*, Stanford)
  - Scores the conservation of each col. in the alignment
  - Define constrained elements as stretches of high scores

Mercator-Pecan Pipeline Overview

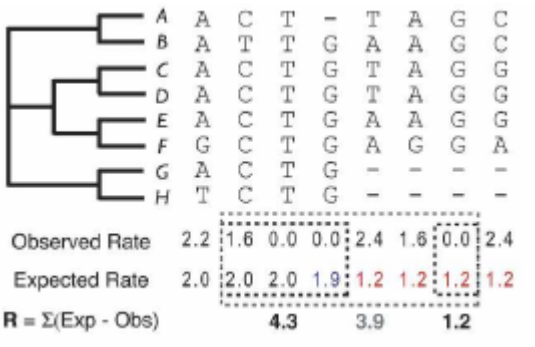
- **Mercator**
  - Defines blocks of orthologous sequences based on coding exon similarities



- **Pecan**
  - Consistency based multiple aligner
  - Optimized to cope with long genomic sequences



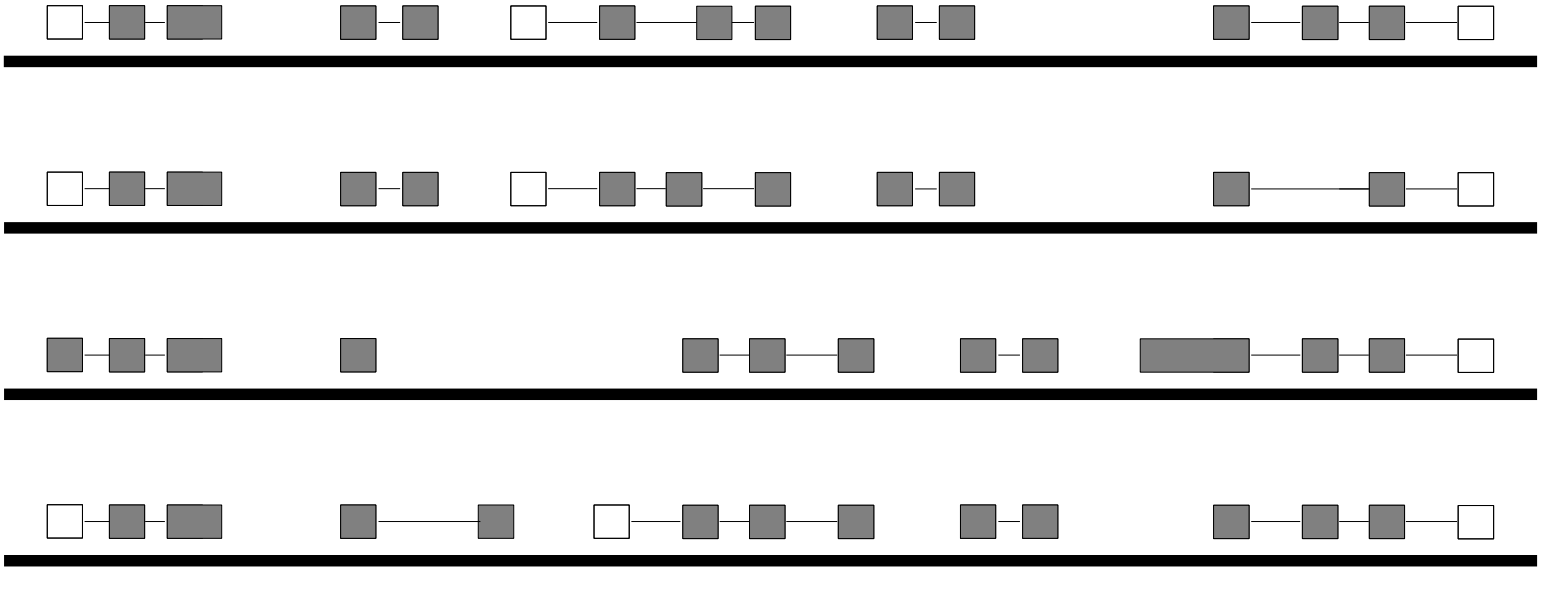
- **GERP**
  - Estimates the conservation of each position in the alignment by looking at the expected and observed number of mutations



Strategy

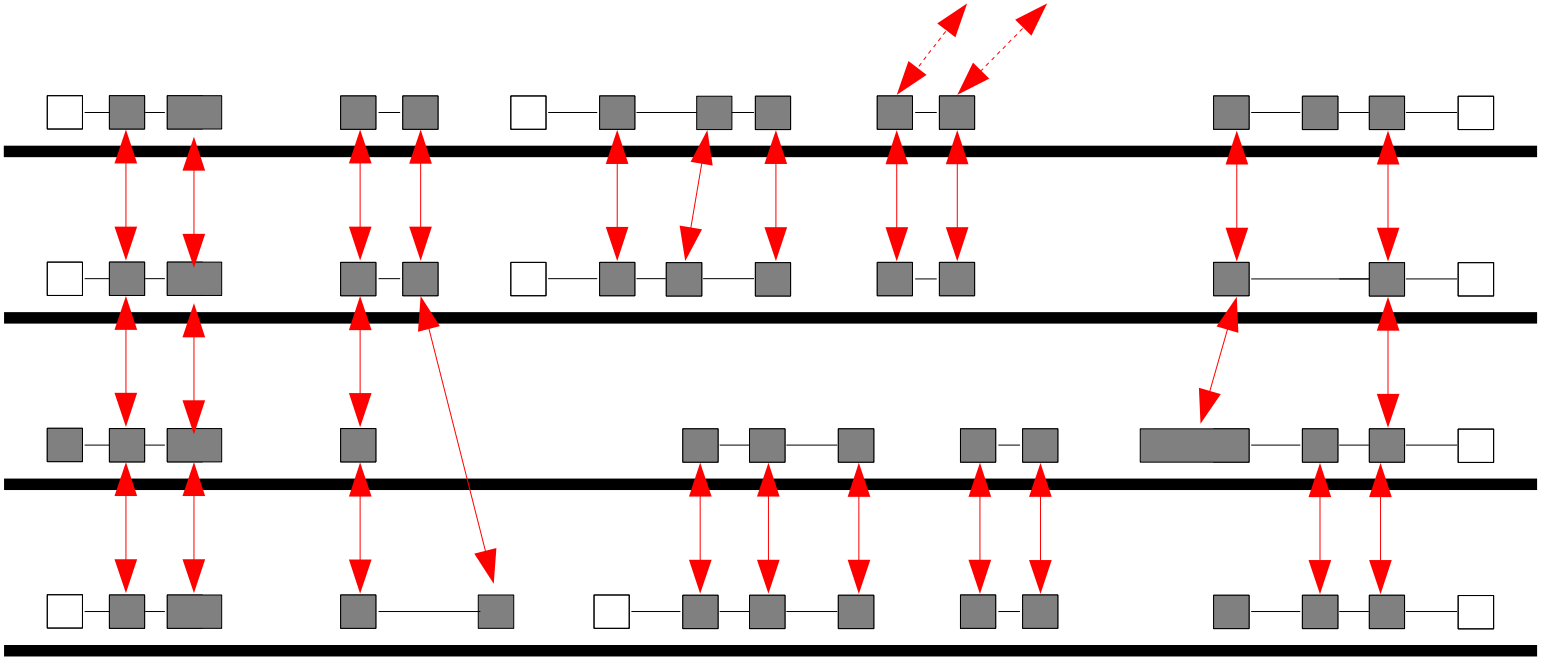
- Global aligner needs orthology maps
- Mercator-Pecan pipeline:
  - 1. Get all coding exons
  - 2. all-vs-all blastp
  - 3. Mercator => strict maps
  - 4. Pecan => multiple alignments

Strategy



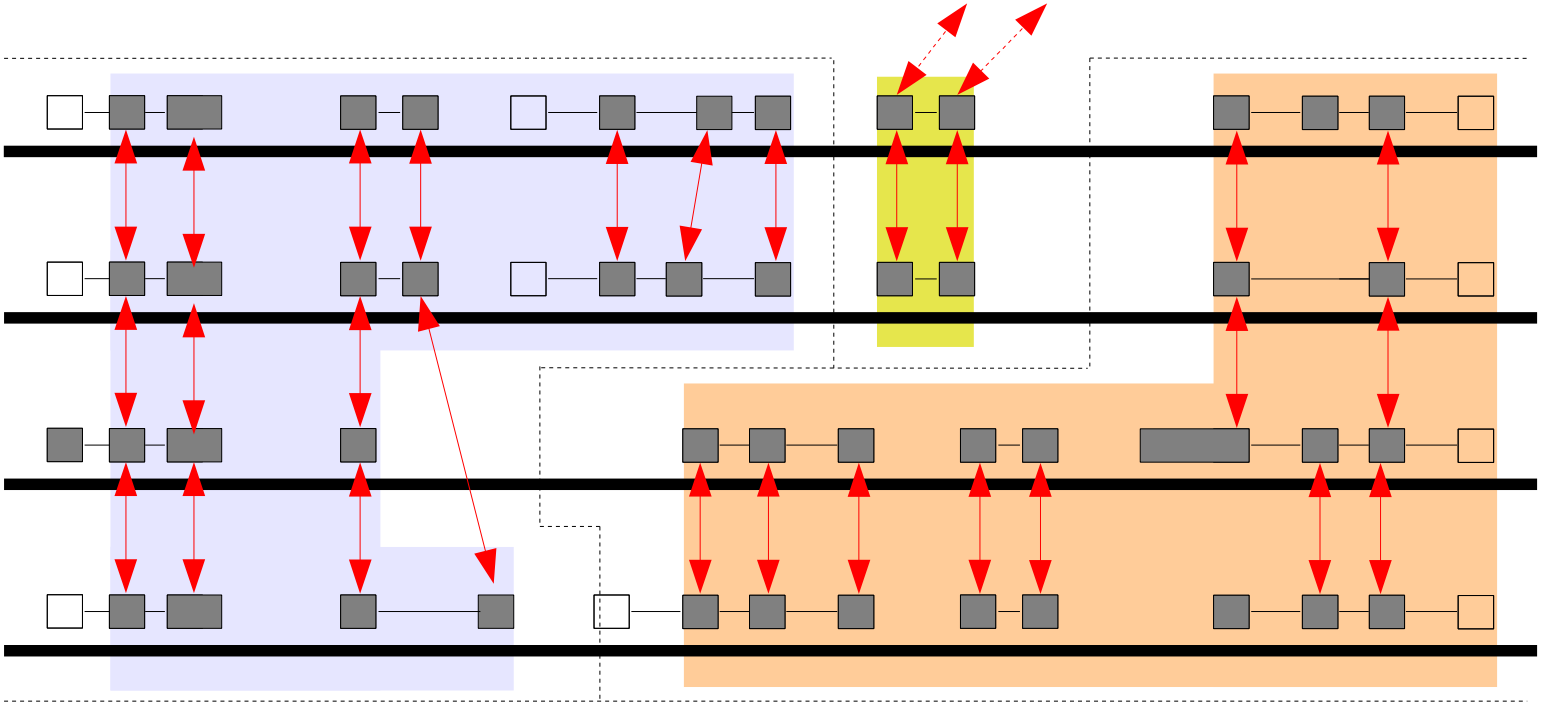
- Use all coding exons

Strategy



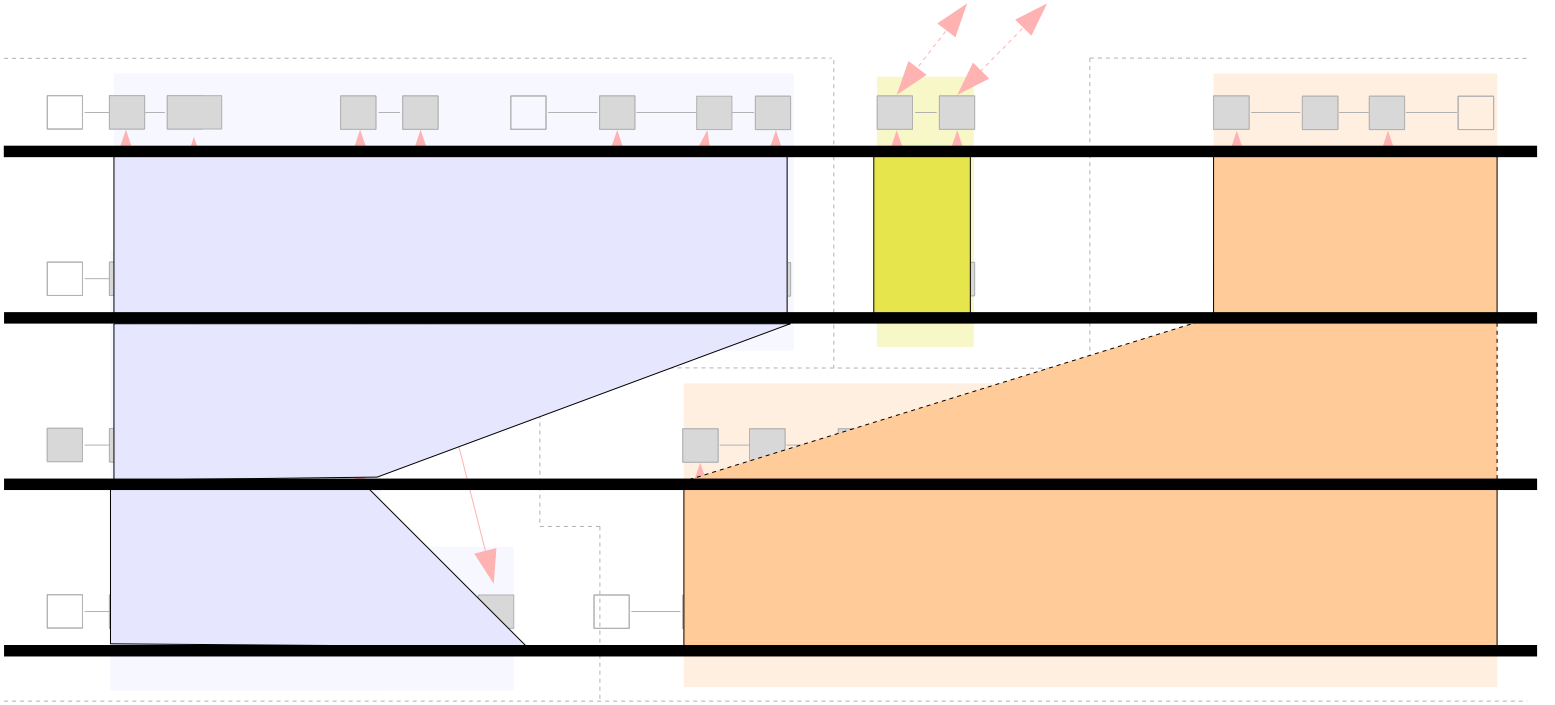
- Use all coding exons
- Get sets of best reciprocal hits

Strategy



- Use all coding exons
- Get sets of best reciprocal hits
- Create orthology maps

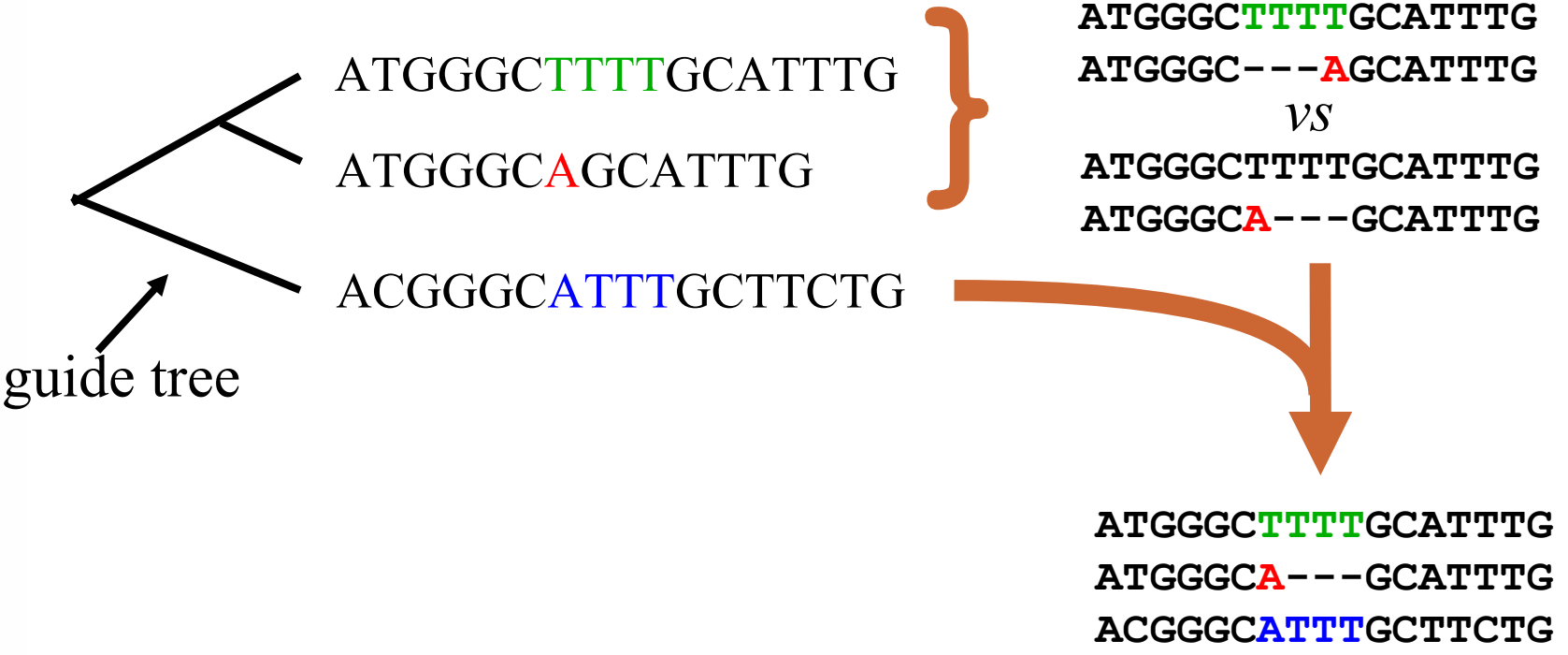
Strategy



- Use all coding exons
- Get sets of best reciprocal hits
- Create orthology maps
- Build multiple global alignments

Pecan

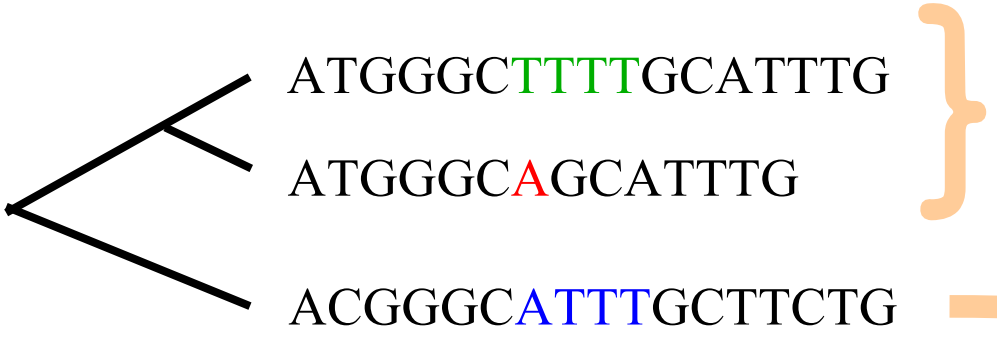
a consistency based multiple-alignment program



Progressive aligner

Pecan

a consistency based multiple-alignment program




```
ATGGGCTTTTGCATTTG
ATGGGC---AGCATTTG
   vs
ATGGGCTTTTGCATTTG
ATGGGCA---GCATTTG
```

```
ATGGGCTTTTGCATTTG
ACGGGCATTTGCTTCTG
```

```
ATGGGCA---GCATTTG
ACGGGCATTTGCTTCTG
```

```
ATGGGCTTTTGCATTTG
ATGGGCA---GCATTTG
```

Takes into account all pairwise alignments, across the entire tree

```
ATGGGCTTTTGCATTTG
ATGGGCA---GCATTTG
ACGGGCATTTGCTTCTG
```

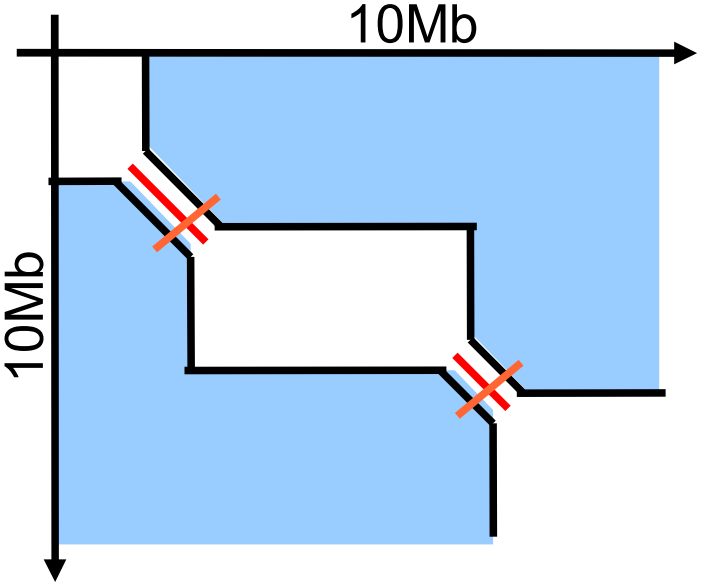
Consistency based aligner

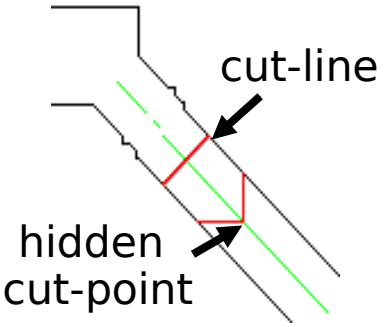
```
ATGGGCTTTTGCATTTG
ATGGGCA---GCATTTG
ACGGGCATTTGCTTCTG
```

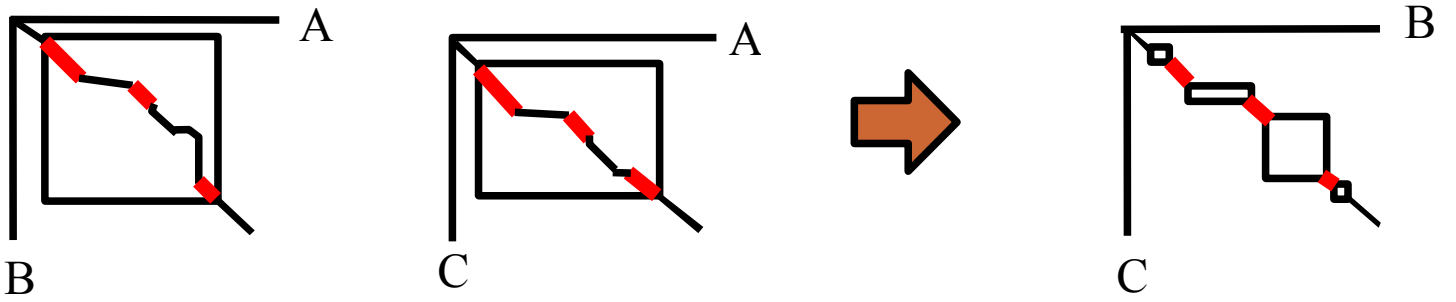
Progressive aligner

Pecan optimizations

- Look for anchors (regions of high similarity)
- perform a banded posterior alignment
- Use cut lines and points to generate effective sub problems for each pairwise alignment simultaneously
- Much redundancy between pairwise alignments: use transitive anchors

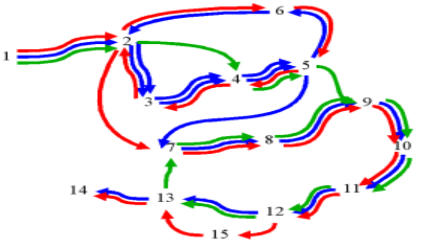


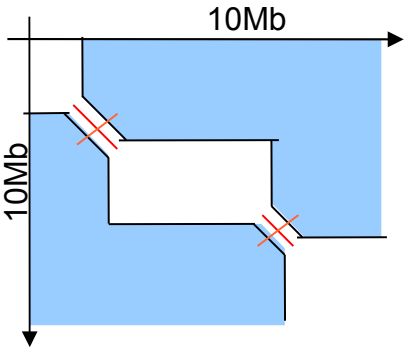


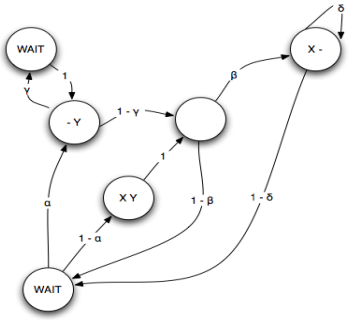


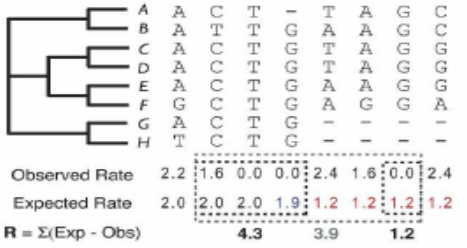
EPO Pipeline Overview

- **Enredo**
  - Defines blocks of collinear sequences
  - Supports segmental duplications
- **Pecan**
  - Consistency based multiple aligner
  - Optimized to cope with long genomic sequences
- **Ortheus**
  - Ancestral sequences reconstructor (Tree Aligner)
  - Infers the history of insertion and deletions
- **GERP**
  - Estimates the conservation of each position in the alignment







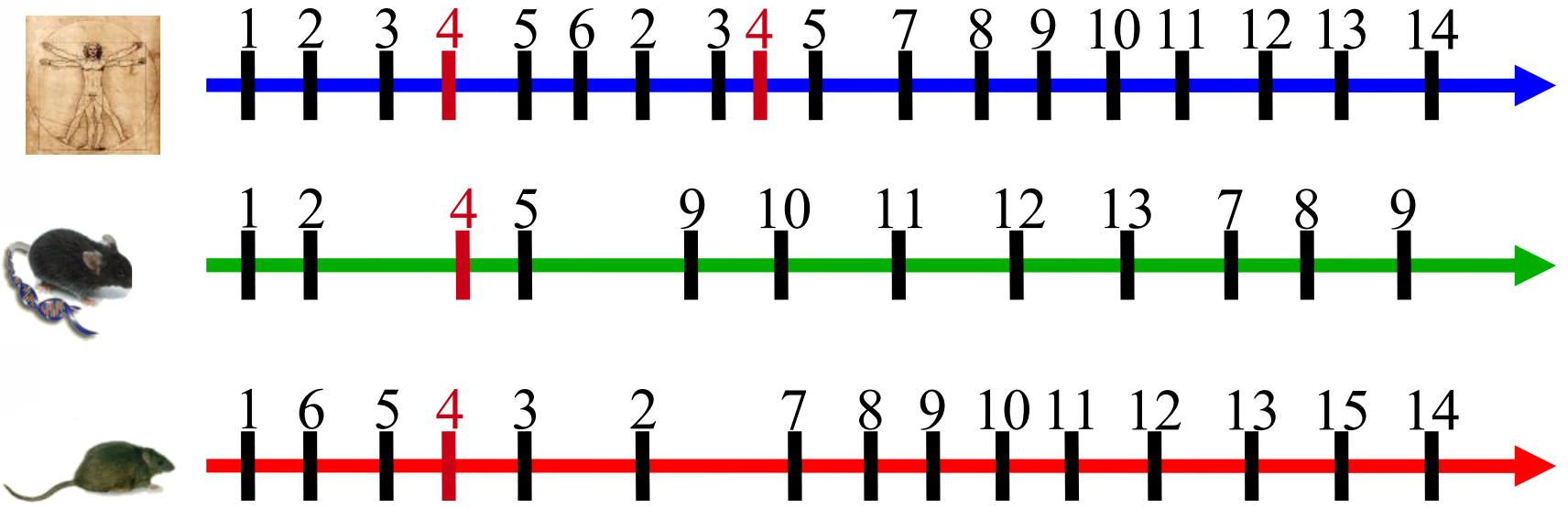


ENREDO graph

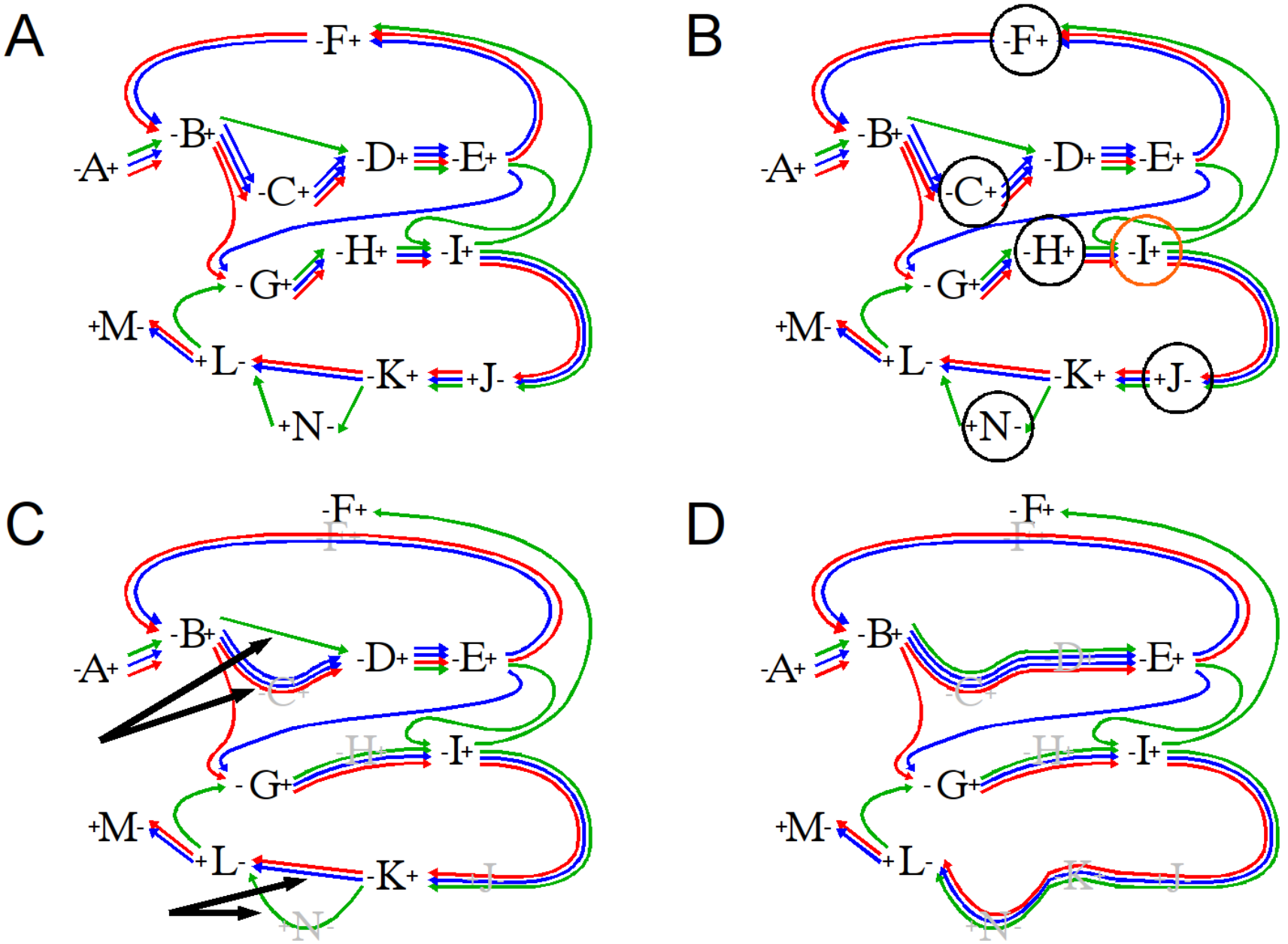
- Similar in spirit to a De Bruijn graph of sequences used for assembly
  - homologous regions between genomes will be represented as one edge
- Formed by creating a set of non-redundant anchors (short regions) which are present 0, 1 or multiple times in each extant genome
- Anchors could be all coding exons, made non-redundant to handle duplications
- In our case, a series of pairwise alignments defines short regions of high homology between genomes

ENREDO: Mapping the anchors

- Mapping the anchors
- Cleaning up the anchor set
  - Removal of overlapping anchors
  - Removal of anchors mapping too many times

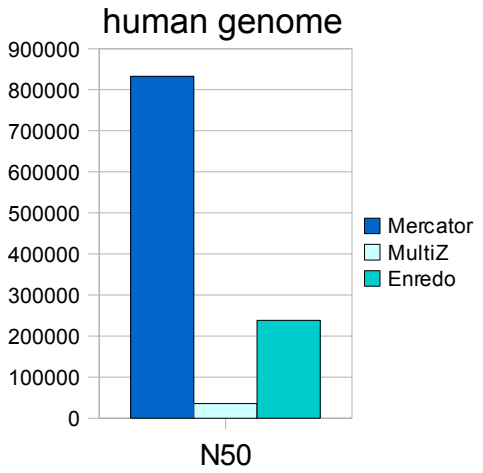


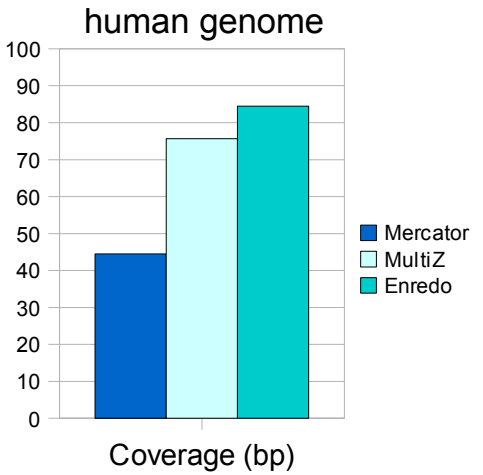
ENREDO Graph

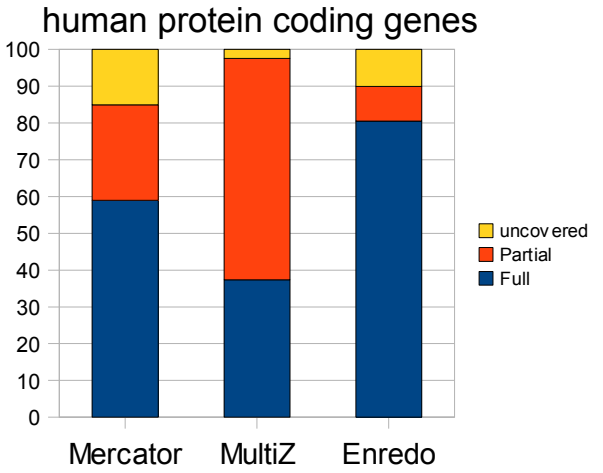


Enredo assessment

- Human, Mouse, Rat, Dog and Cow
- Mercator, MultiZ and Enredo coverage





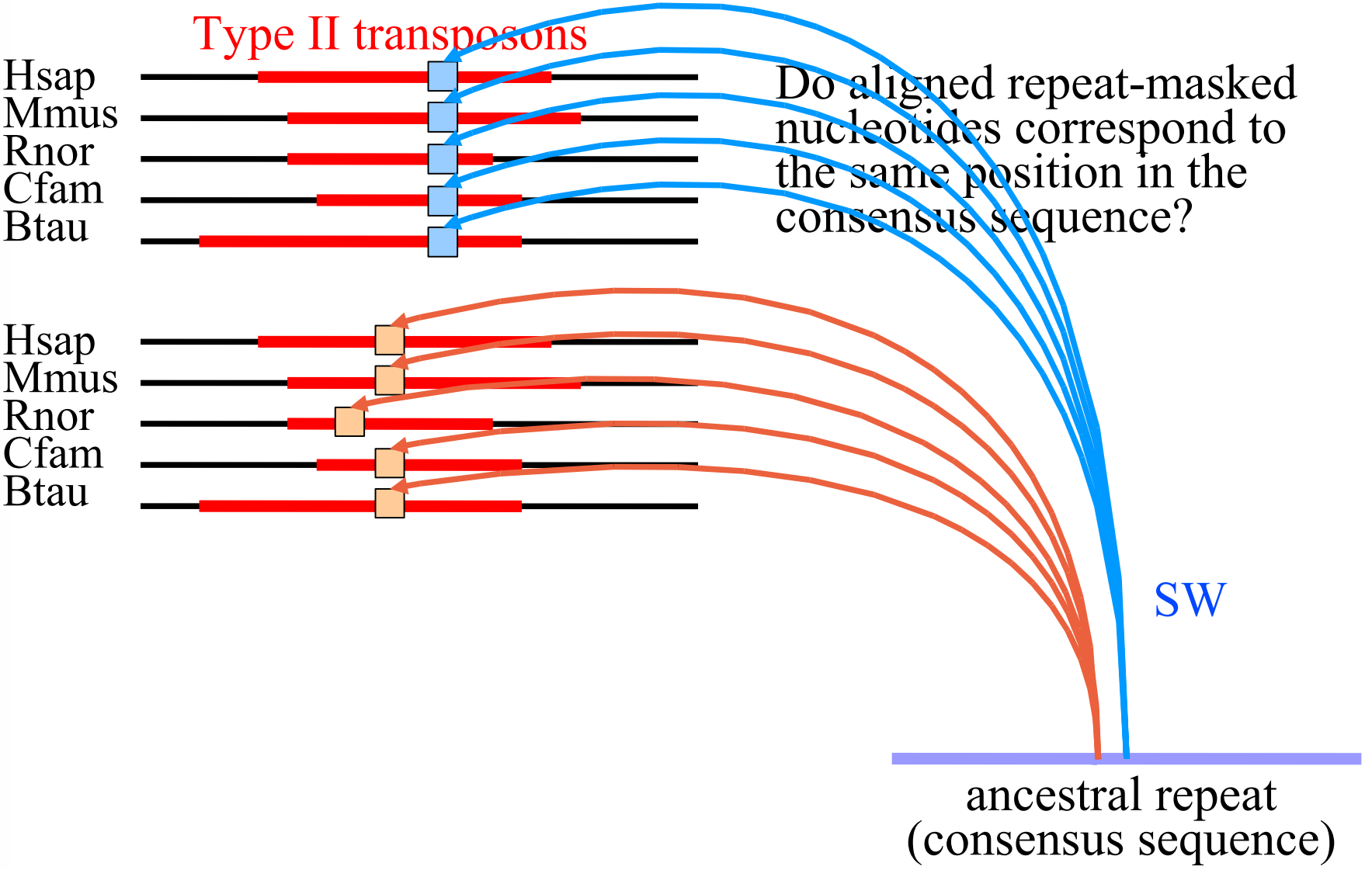


- Putative rearrangements between human chromosome X and any autosome in another species

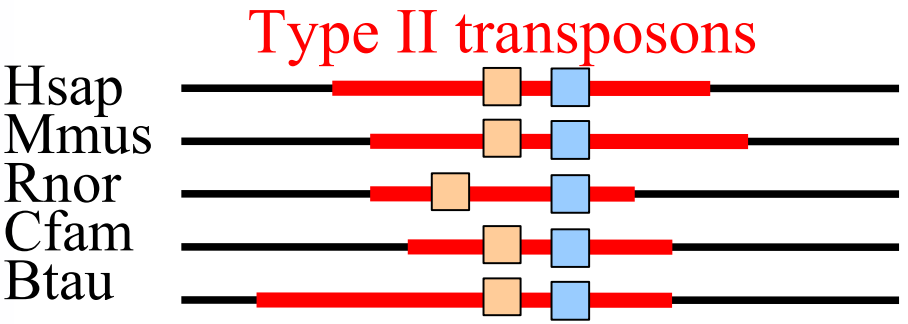
Method	blocks		length	
Mercator	15	6.7%	2750241	4.0%
MultiZ*	211117	28.0%	25785059	19.0%
Enredo	19	1.3%	1168017	1.0%

* from UCSC 17 way MultiZ

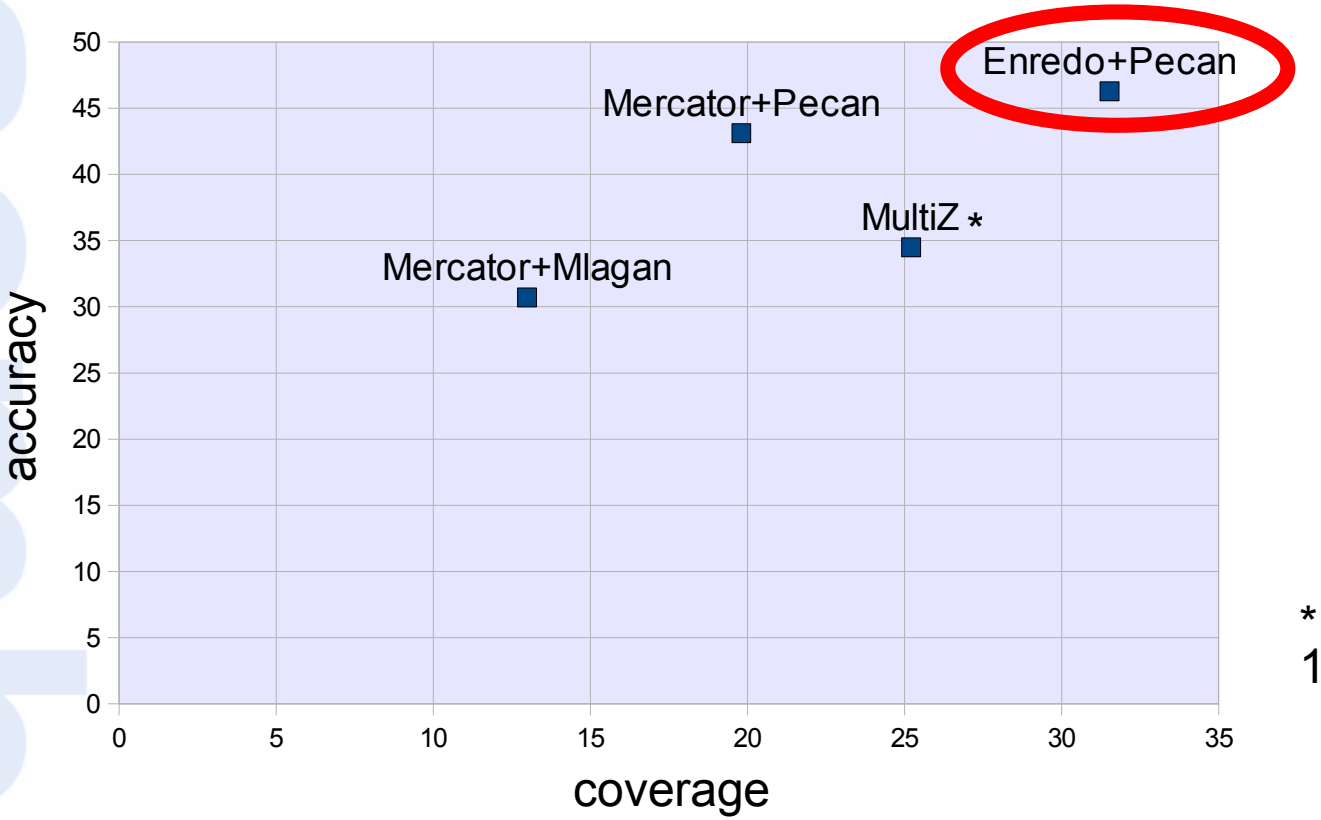
Multiple aligner assessment: ancestral repeats



Multiple aligner assessment: ancestral repeats



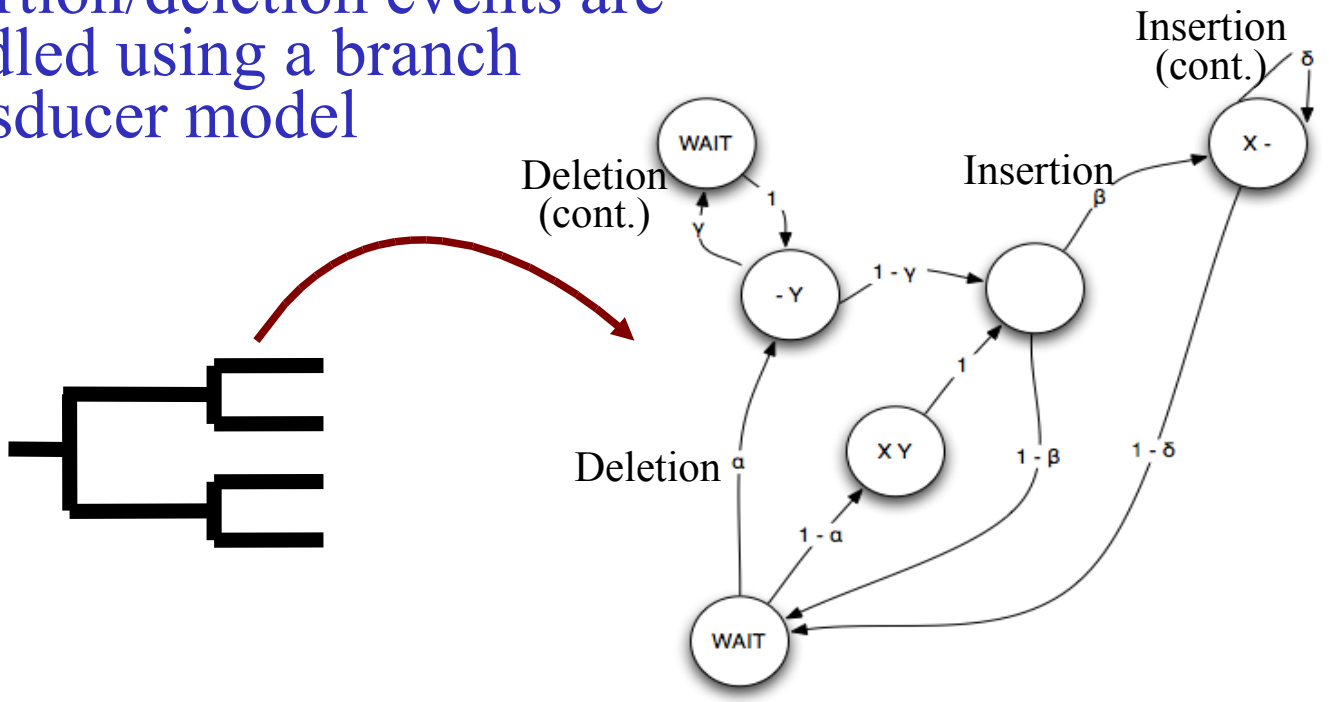
Do aligned repeat-masked nucleotides correspond to the same position in the consensus sequence?



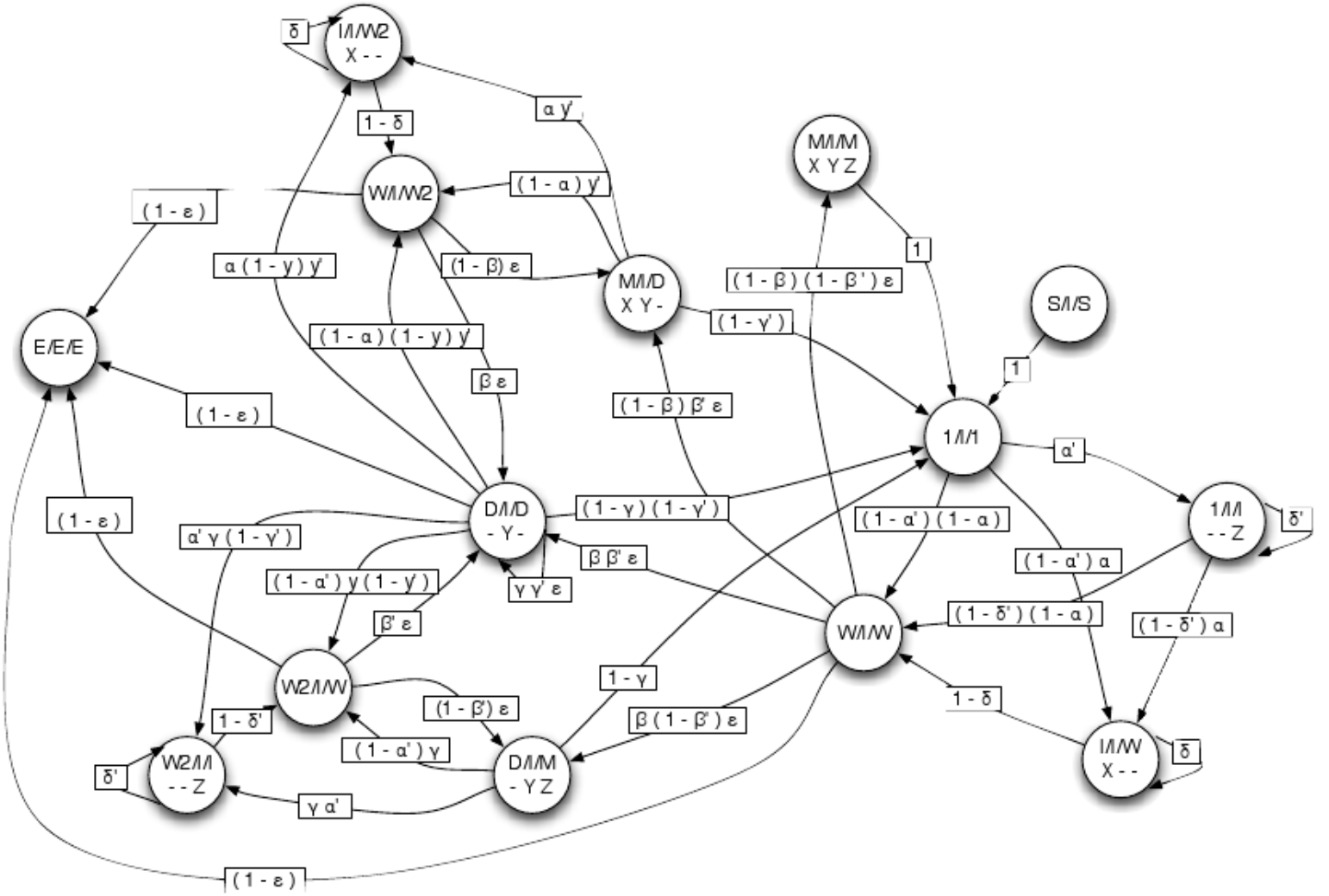
* from UCSC 17way multiZ

Ortheus

- Addresses the inference of insertion-deletion histories and substitution events
- Uses a multiple alignment as guiding input
- Reconstructs the ancestral sequences in the tree and refines the input alignment
- Insertion/deletion events are handled using a branch transducer model

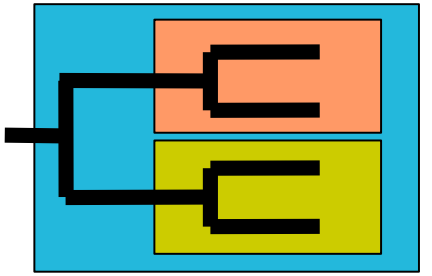


Ortheus transducer model for 2 descendants and 1 ancestor

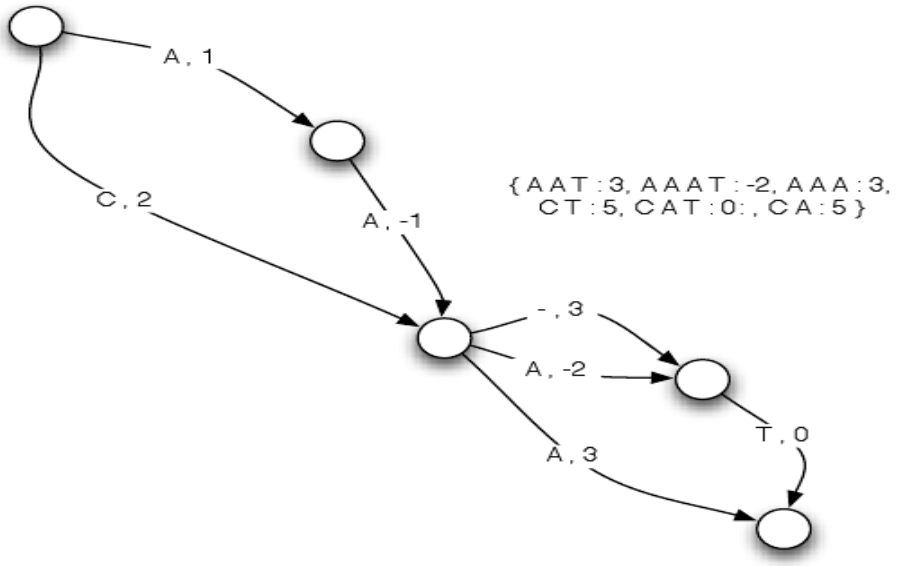


Ortheus: inference of the ancestral sequence

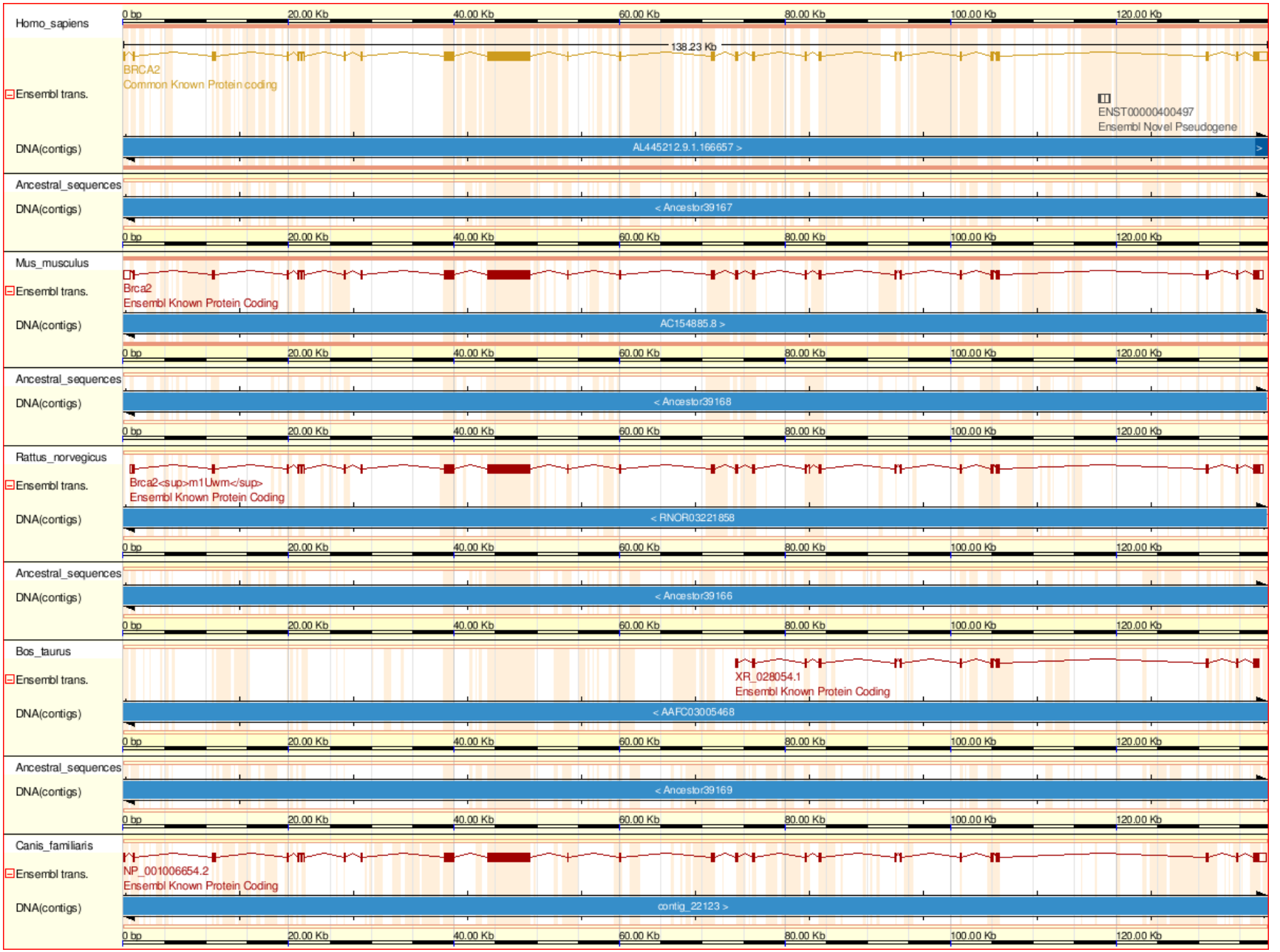
- Substitution are handled using Tamura-Nei nucleotide substitution model.
- Works in a progressive manner:



- Ancestral sequences are represented using weighted sequence graphs



Display on AlignSliceView

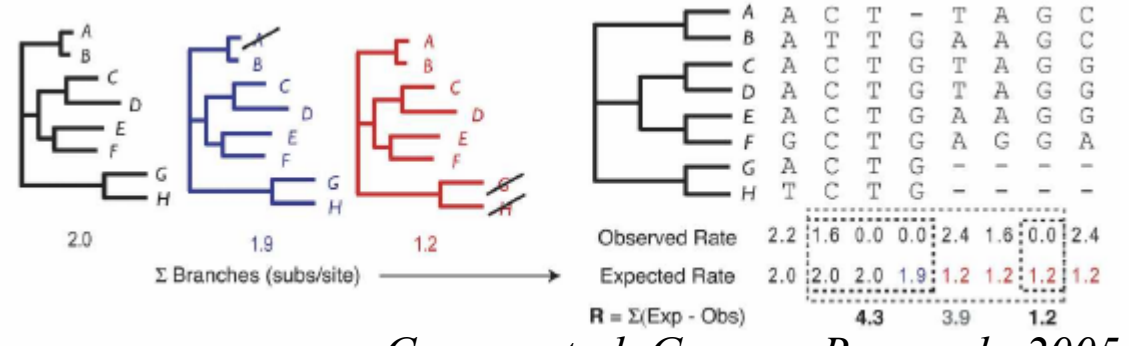


Current sets of alignments

- Primates:
  - 4-way EPO alignments (high-coverage genomes only)
- Mammals
  - 9-way EPO alignments (high-coverage genomes only)
  - 23-way EPO alignments (including 2X genomes)
- Amniota
  - 12-way Mercator-Pecan alignments (high-cov. Only)
- *Fish*
  - *Planning a 5-way EPO alignments set for 2009*

Gerp Constrained Elements

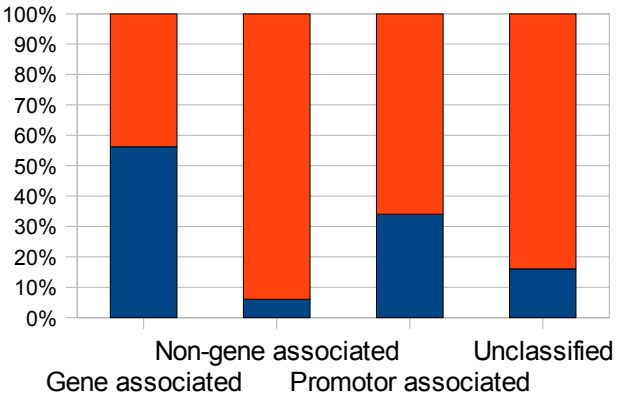
- Stretches of the alignment with a high conservation



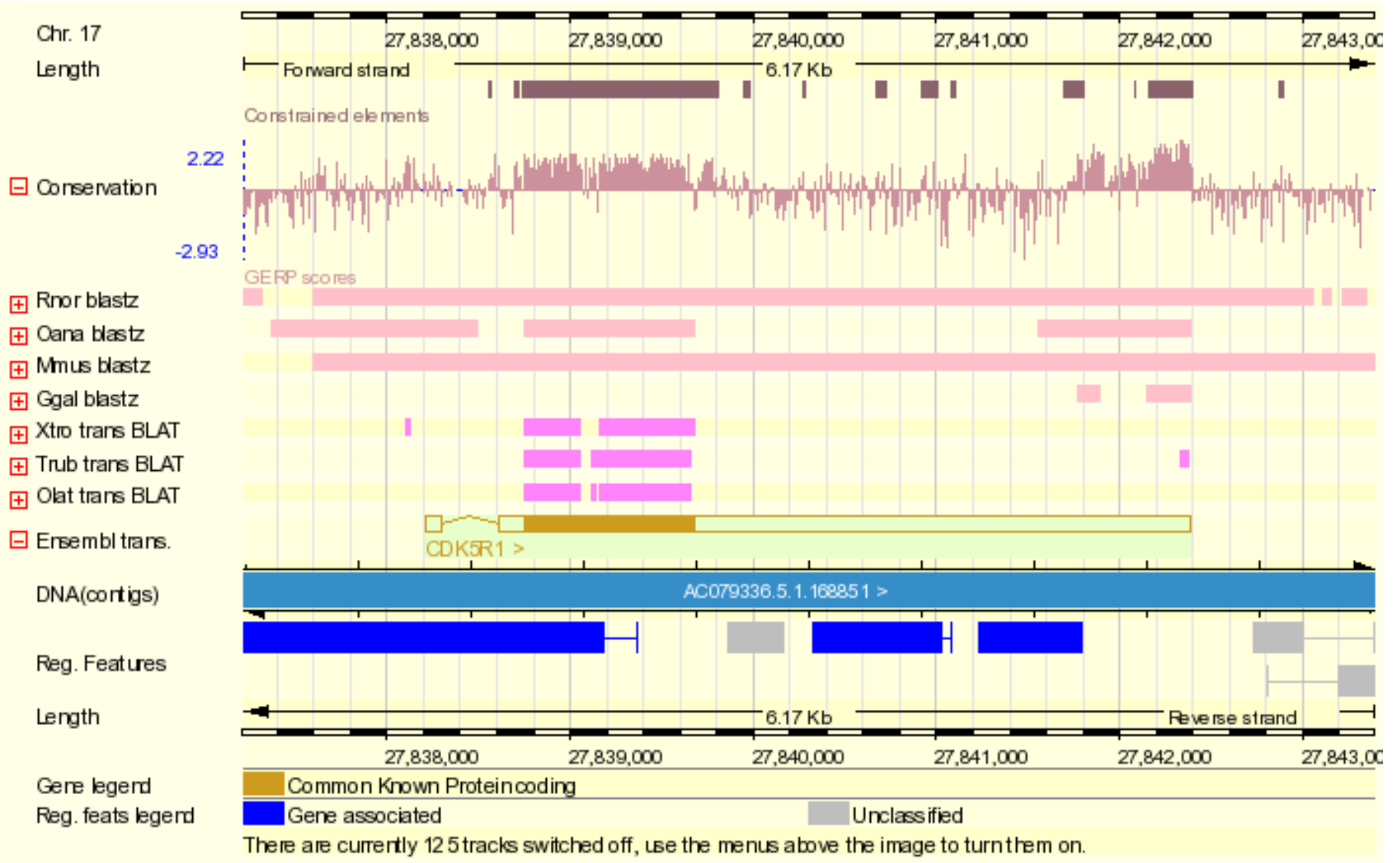
Cooper et al. Genome Research, 2005

- Constrained elements and coding exons
  - 74% of coding exons are associated with constr. elem.
  - 22% of constr. elem. are associated with coding exons

- Co-occurrence of features
  - Annotation of constr. elements
    - genes, TSS, Reg. features...
  - Annotation of SNPs
    - in constrained elements or not



ContigView: p23



GeneSeqAlingView: p23

THIS STYLE: Location of conserved regions (where >50% of bases in alignments match)

THIS STYLE: Location of START/STOP codons

THIS STYLE: Location of selected exons

THIS STYLE: Location of SNPs

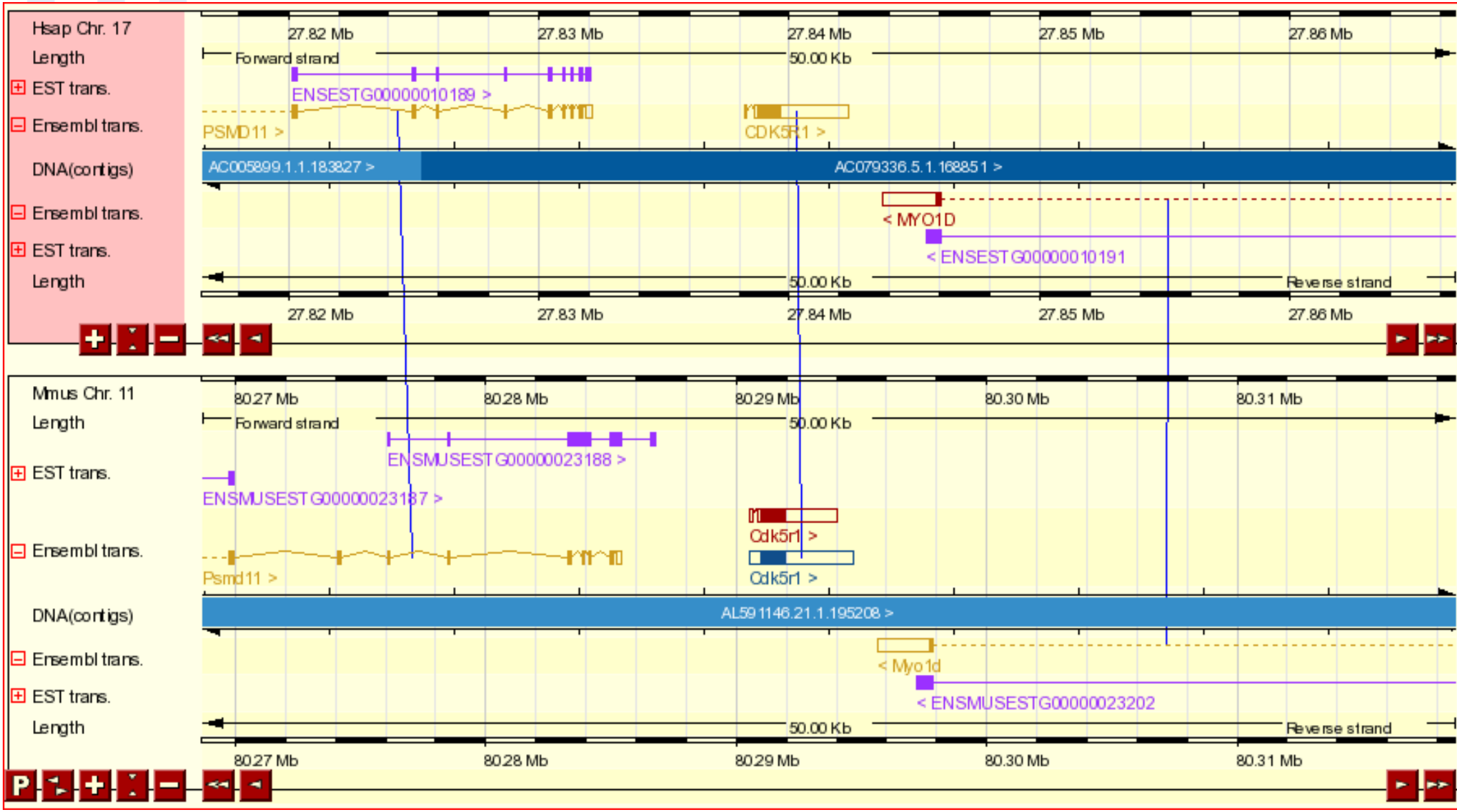
THIS STYLE: Location of deletions

Homo_sapiens > chromosome:NCBI36:17:27838014:27842583:1

Macaca_mulatta > chromosome:MMUL_1:16:27819375:27823565:1

```
Homo_sapiens        481 AGGGCCCGGGACTGGGGCGGCGGGGTGCGCCGAGGCGCGGGGCGGAGGGGCGCAGGGGCG 540
Macaca_mulatta      481 AGGGCCCGGGACTGGGGCCGCGGGGTCCGCGGAGGCGCGGGGCGGAGGGGCGC------- 540

Homo_sapiens        541 CAGGGGCGCGGCGCGGAGCCCAGCCTGGCGCTAAGAACCATCTTGTTTTCCAGGCAGATC 600
Macaca_mulatta      541 ---------GGCGCGGAGCCCAGCTTGGCGCTAAGAACCATCTTGTTTTCCAGGCAGATC 600

Homo_sapiens        601 CAAGGGGGCAGCACGCTTCCCGGGAGCGCCCCCGCCTCCTCTCCGGGGCCGCCGCAGGCT 660
Macaca_mulatta      601 CAAGGGGGCAGCACGCTTCCCGGGAGCGCCCCCGCCTCCTCCCCGGGGCCACCGCAGGCT 660

Homo_sapiens        661 CGGTGAGCGGTTTTATCCYTCCGGCCGGCAGGCTGGGCGCGCAGGGGCGCGAGCCCCCGC 720
Macaca_mulatta      661 CCGTGAGTGGTTTTATCCCTCCGGCCGGCAGGCTGGGCGCGCAGGGGCGCGAGCCCCCGC 720

Homo_sapiens        721 CCGGCGCGCAGCAGCACCATGGGCACGGTGCTGTCCCTGTCTCCCAGCTACCGGAAGGCC 780
Macaca_mulatta      721 CCGGCGCGCAGCGGCACCATGGGCACGGTGCTGTCCCTGTCCCCCAGCTACCGGAAGGCC 780

Homo_sapiens        781 ACGCTGTTTGAGGATGGCGCGGCCACCGTGGGCCACTATACGGCCGTACAGAACAGCAAG 840
Macaca_mulatta      781 ACGCTGTTTGAGGATGGCGCGGCCACCGTGGGCCACTATACGGCCGTACAGAACAGCAAG 840

Homo_sapiens        841 AACGCCAAGGACAAGAACCTGAAGCGCCACTCCATCATCTCCGTGCTGCCTTGGAAGAGA 900
Macaca_mulatta      841 AACGCCAAGGACAAGAACCTGAAGCGCCACTCCATCATCTCCGTGCTGCCTTGGAAGAGA 900

Homo_sapiens        901 ATCGTGGCCGTGTCGGCCAAGAAGAAGAACTCMAAGAAGGTGCAGCCYAACAGCAGCTAC 960
Macaca_mulatta      901 ATCGTGGCCGTGTCGGCCAAGAAGAAGAACTCCAAGAAGGTGCAGCCCAACAGCAGCTAC 960

Homo_sapiens        961 CAGAACAACATCACGCACCTCAACAATGAGAACCTGAAGAAGTCGCTGTCRTGYGCCAAC 1020
Macaca_mulatta      961 CAGAACAACATCACGCACCTCAACAATGAGAACCTGAAGAAGTCGCTGTCGTGCGCCAAC 1020
```

MultiContigView

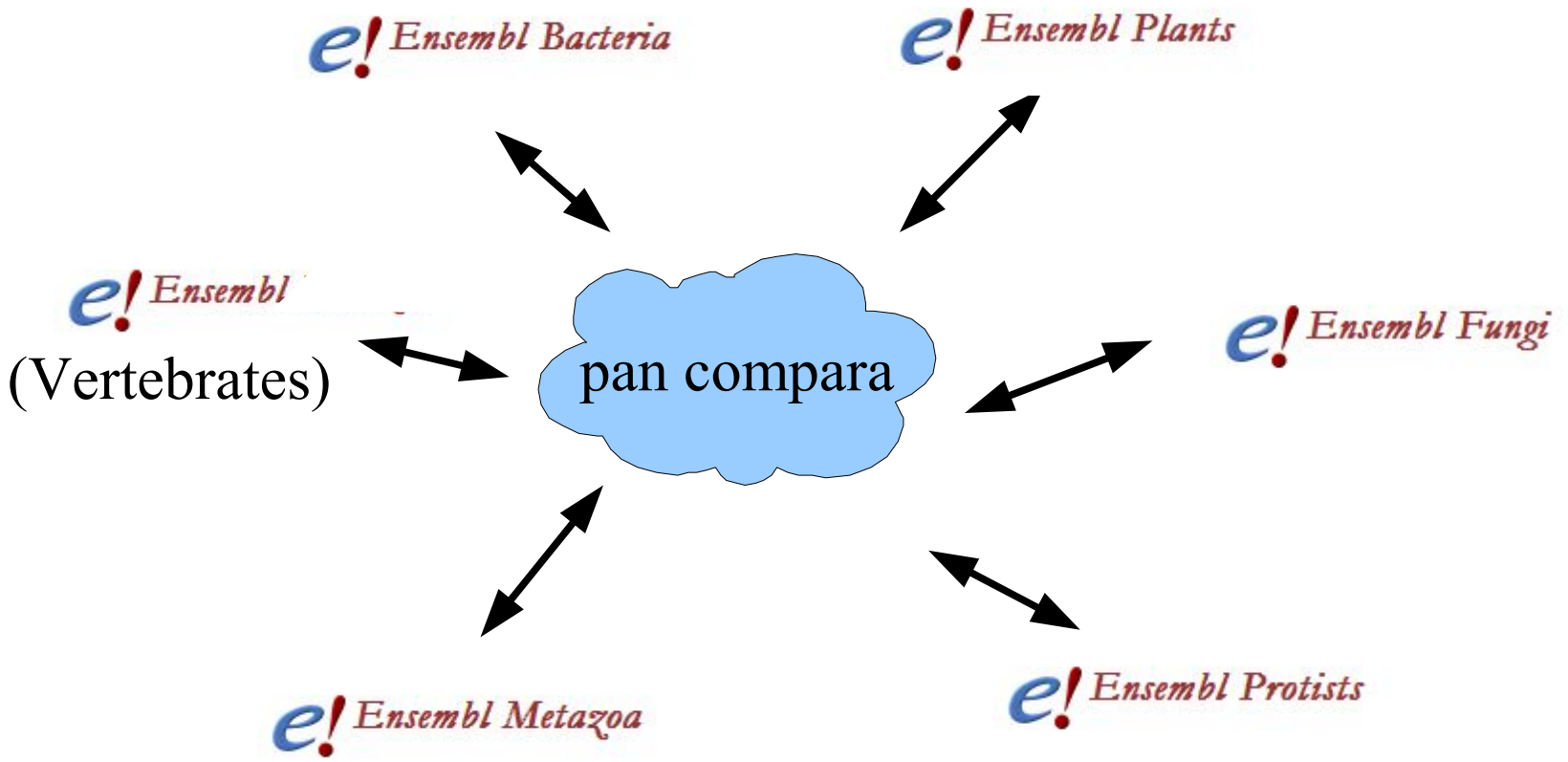


Summary

- Ensembl is a system created for the study and analysis of the genomes
- Comparative genomics
  - Protein tree and inference of homologues
  - Genomic alignments, conserved regions
- Many views to match different usages
  - ContigView: genomic region
  - MultiContigView: side-by-side comparison
  - AlignSliceView: alignment in genomic context
  - GeneSeqAlignView: alignment of genomic regions
  - GeneTreeView: protein trees, homologues
  - many other views...
- All data accessible through the web and the Perl API

Pan-Ensembl compara

- Take advantage of the whole new span of Ensembl Genomes
- Link the projects together
- Breakout session after the coffee/tea break!!



EMBL-EBI

wellcome trust sanger institute

Ensembl

Ensembl	Paul Flicek (*EBI*), Steve Searle (Sanger Institute)
Vertebrate Genomics	Mario Caccamo, Laura Clark, Jonathan Hinton, Zam Iqbal, Vasudev Kumanduri, Ilkka Lappalainen
Software	Glenn Proctor, Syed Haider, Andrew Jenkinson, Andreas Kähäri, Stephen Keenan, Rhoda Kinsella, Eugene Kulesha, Ian Longden, Daniel Rios
Comparative Genomics	Javier Herrero, Kathryn Beal, Benoît Ballester, Stephen Fitzgerald, Leo Gordon, Albert Vilella
Functional Genomics	Nathan Johnson, Stefan Gräf, Steven Wilder
Variation	Fiona Cunningham, Yuan Chen
Analysis and Annotation	Bronwen Aken, Julio Banet, Susan Fairley, Jan-Hinnerck Vogel, Simon White, Amonida Zadissa
Web Team	James Smith, Eugene Bragin, Anne Parker, Bethan Pritchard, Steve Trevanion (VEGA)
Zebrafish	Kerstin Howe, Britt Reimholz, James Torrance
VectorBase	Dan Lawson, Martin Hammond, Karyn Megy
Outreach	Xosé M Fernández, Bert Overduin, Michael Schuster (QC), Giulietta Spudich
Systems & Support	Guy Coates, Tim Cutts, Shelley Goddard
Research	Ian Dunham, Damian Keefe, Alison Meynert, Dace Ruklisa, Guy Slater, Daniel Zerbino
Ensembl Strategy	Ewan Birney, Richard Durbin, Tim Hubbard