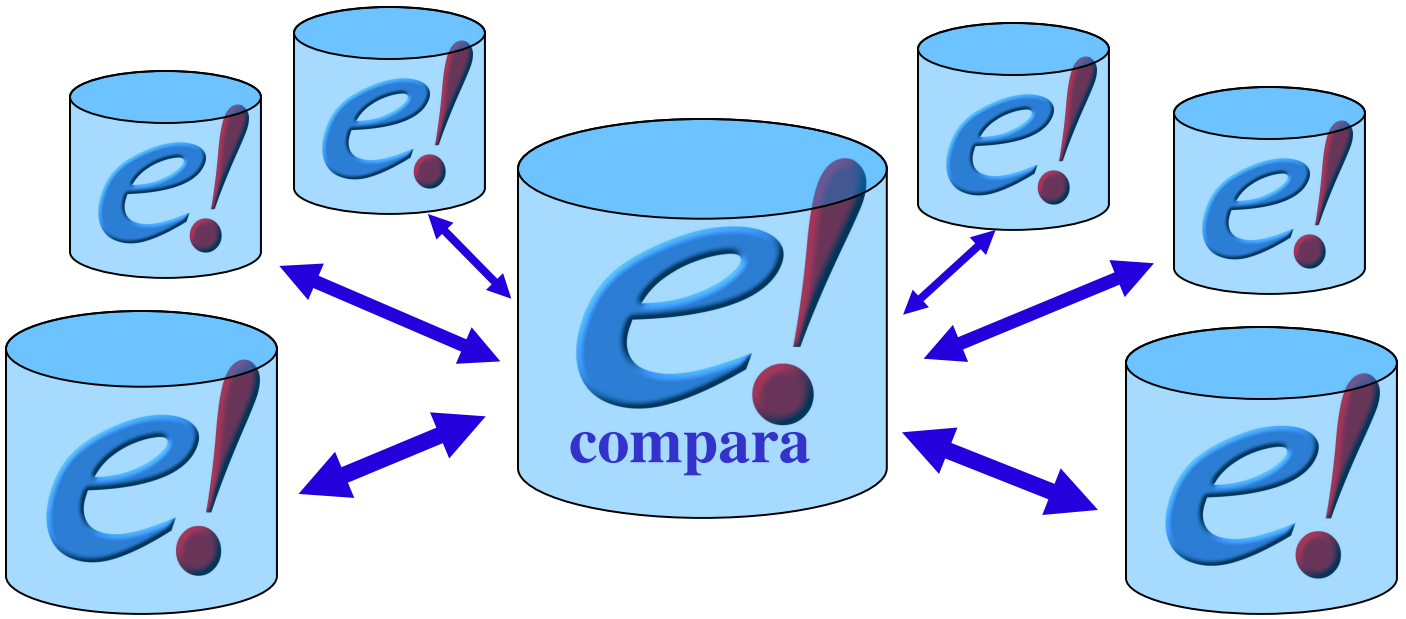
Ensembl Compara

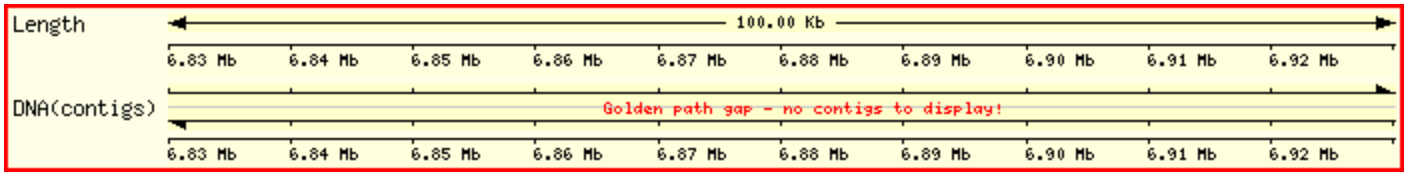


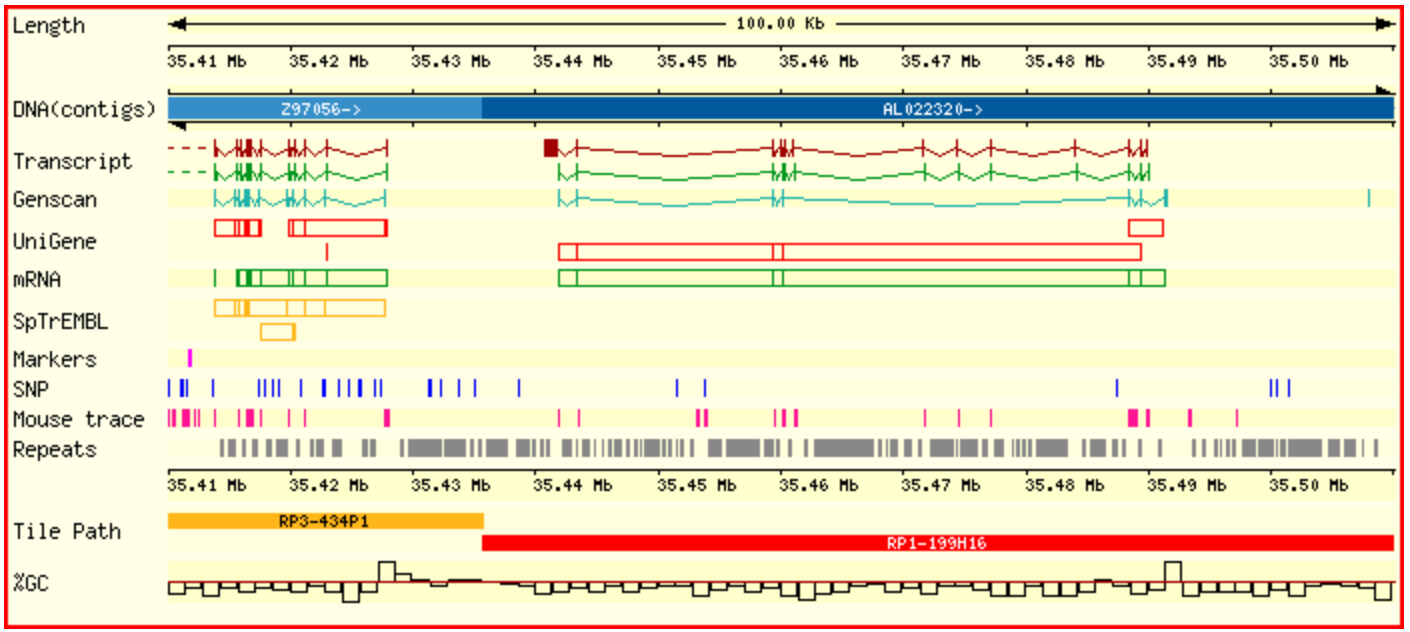
Javier Herrero

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

EBI - Wellcome Trust Genome Campus, UK

Ensembl goal





Ensembl Concept

- Collaborative project of the European Bioinformatics Institute and the Wellcome Trust Sanger Institute
- Provides annotation and analysis of chordate genomes
- Open by design
  - Code is BSD, not GNU
  - All data is freely available
- Continuously developed and comprehensively updated every two months
- Diverse skills across the project
- Technology adopted and used by many other projects

Timeline 1: Human Genome

2/98 Bermuda agreement for Human Genome

5/98 Celera announced

7/98 Tim Hubbard's annotation 'pipeline' of confirmed genscan predictions for all draft+finished sequence

2/99 G5 strategy of 'working draft' in 12 months'

5/99 Name 'EnsEMBL' dreamed up over afternoon email discussion, in order to have title for CSH talk

5/99 First code checked into CVS by Ewan/Michele for perl object relational database system

10/99 First public Ensembl datasets

1/00 Milestone 1 release (clone based web browser)

7/00 WT approves 1st Ensembl grant (£8m, 5yrs)

10/00 Release with Golden Path (gp) based web browser

11/00 (July gp); 1/01 (September gp); 2/01 (October gp); 4/01 (December gp) Ensembl 1.0; 7/01 (April gp)

- T. Hubbard

Timeline 2 : Genomes

9/01 Release with Mouse genome (gp)
2/03 Ensembl 10 (Fugu, worm, fly, Anopheles, briggsae; 1st major schema/API rewrite)
11/03 First Ensembl SAB
1/04 Ensembl 20 (2nd major schema/API rewrite)
12/04 Ensembl 27 (Variation schema/API rewrite)
2/05 Ensembl 28 (archive sites, bioMart replaces Mart)
7/05 Ensembl 32 (website redesign)
10/05 Ensembl 34 (multiple genome alignment data/AlignSlice API)
10/05 WT Funding approved (WTSI + Ensembl) until 2011
12/05 Switch from monthly to bi-monthly cycle
2/06 Ensembl 37 (mouse SNPs, transcriptSNPview)
4/06 Ensembl 38 (Combined Havana-Ensembl Human geneset)
12/06 Ensembl 42 (Platypus!)
4/07 Ensembl 44 (37 species, more than 500 Gb behind the website)

- T. Hubbard

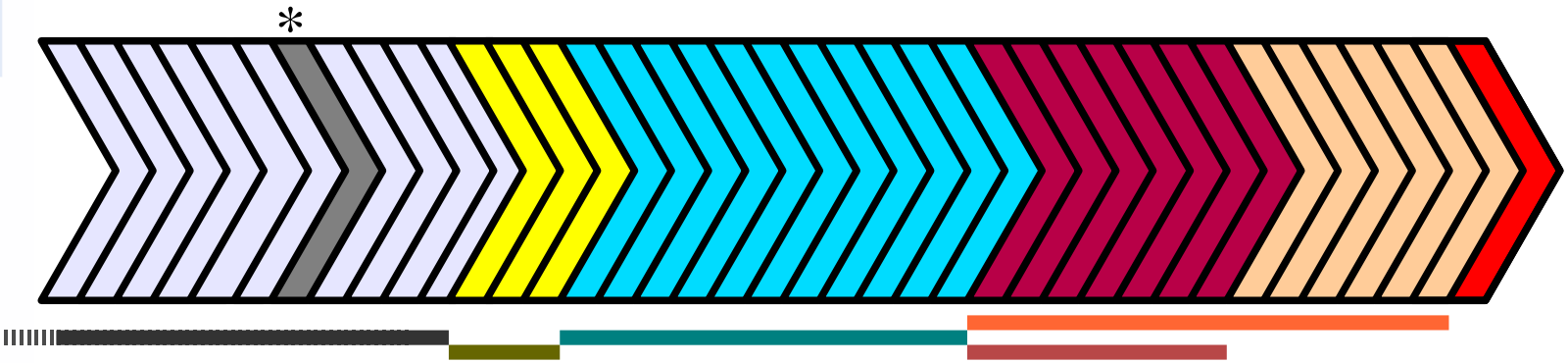
Ensembl: What do you get?

- **Genome Annotation**
  - Protein coding gene structure
    - Consistent with genome, predicted across all vertebrates
    - Manual annotations (human, mouse, zebrafish, MHC)
  - RNA genes (including miRNA)
    - Consistent with genome, predicted in across mammals
  - Additional identifiers per genes
    - Affymetrix, EntrezGene, Uniprot…
- **Comparative & Functional Genomics**
  - Genome alignments
    - Blastz, Blat, coordinated with UCSC
  - Homologues between genomes
  - Protein trees
  - Variants (SNPs), strains, genotypes
  - Tiling array data
- **Infrastructure**
  - Website, Data mining tool, database and data dump
  - Portable, extendable, open source system with database, API, website, pipeline

Ensembl groups

- **GeneBuilders**: sequence masking, gene building
- **Core**: database schema, stable id mapping
- **Compara**: protein homology, genomic sequence alignments
- **Functional Genomics**: SNPs, probe mapping, functional data
- **Mart**: martification of the previous data for data mining
- **Web**: web site, new views for new data
- **User support**: help, workshops, tutorials
- More people: Research, DAS, VectorBase, Zebrafish, Systems...

Ensembl release cycle



genebuilder (10+3+4)
~ 3 months

core (4)

compara (6)

functional genomics (5)

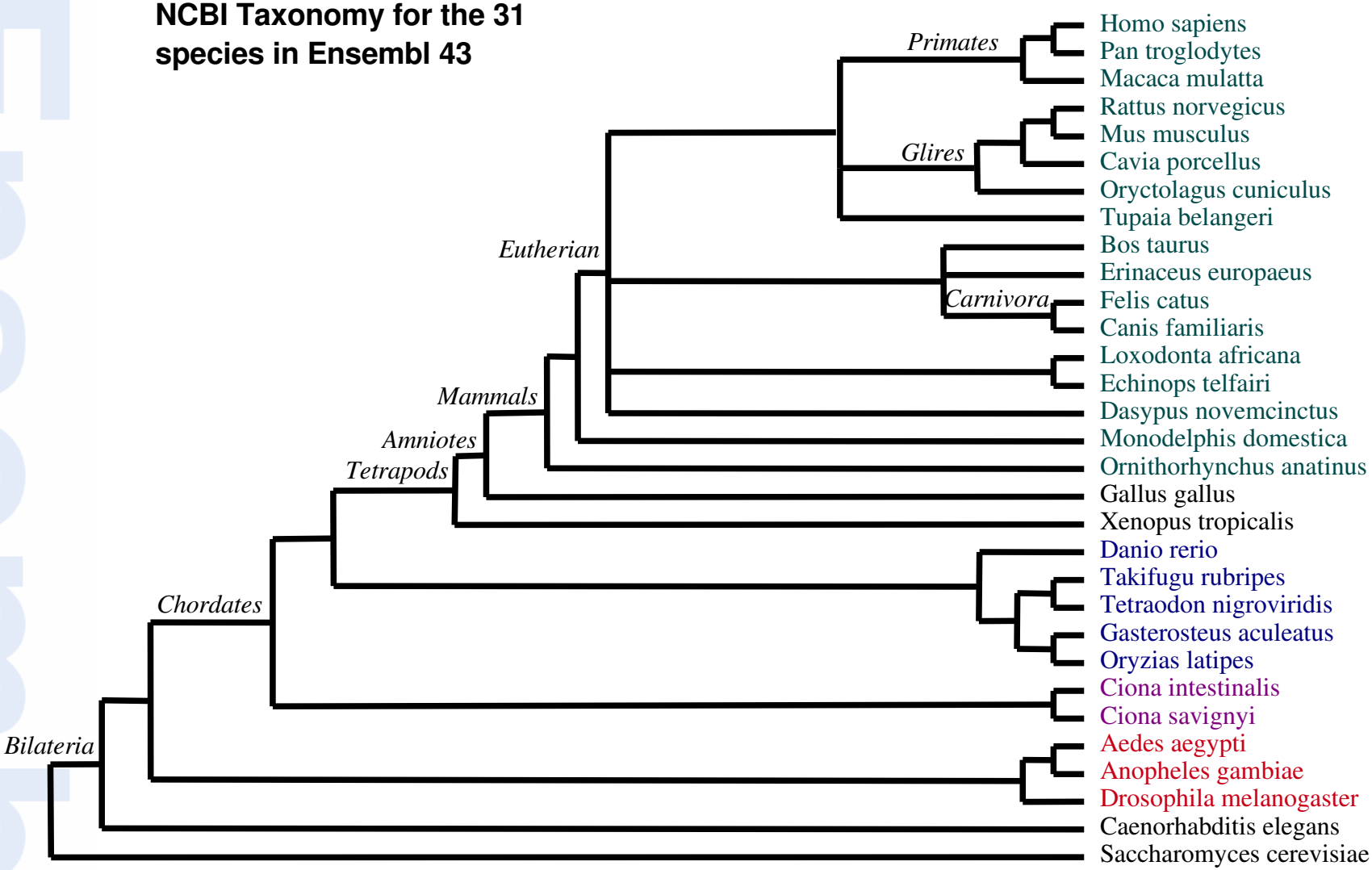
mart (4)

web (7)

release!!

1 release coordinator
+ 1 assistant

NCBI Taxonomy for the 31 species in Ensembl 43



Genomes: the future

- Another 28 mammals being sequenced
  - Main driver is evolution, but perhaps opening up biology
- New technology sequencing
  - 454 - currently about 5 fold cheaper
  - Solexa - currently about 100 fold cheaper

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

Compara data

Raw genomic sequence

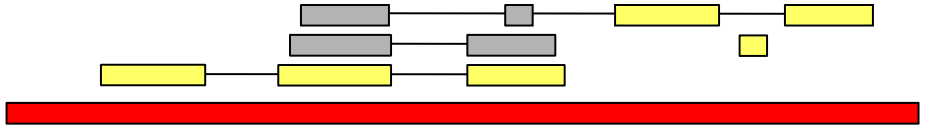
- Whole genome alignments
  (tBLAT, BlastZ-net, PECAN)
- Syntenic regions (based on BlastZ-net)

Protein sequence

- Raw protein alignments (wublastp)
- Protein Family clusters
- Protein trees (since Jun 2006 - v39)
- Gene orthology / paralogy predictions

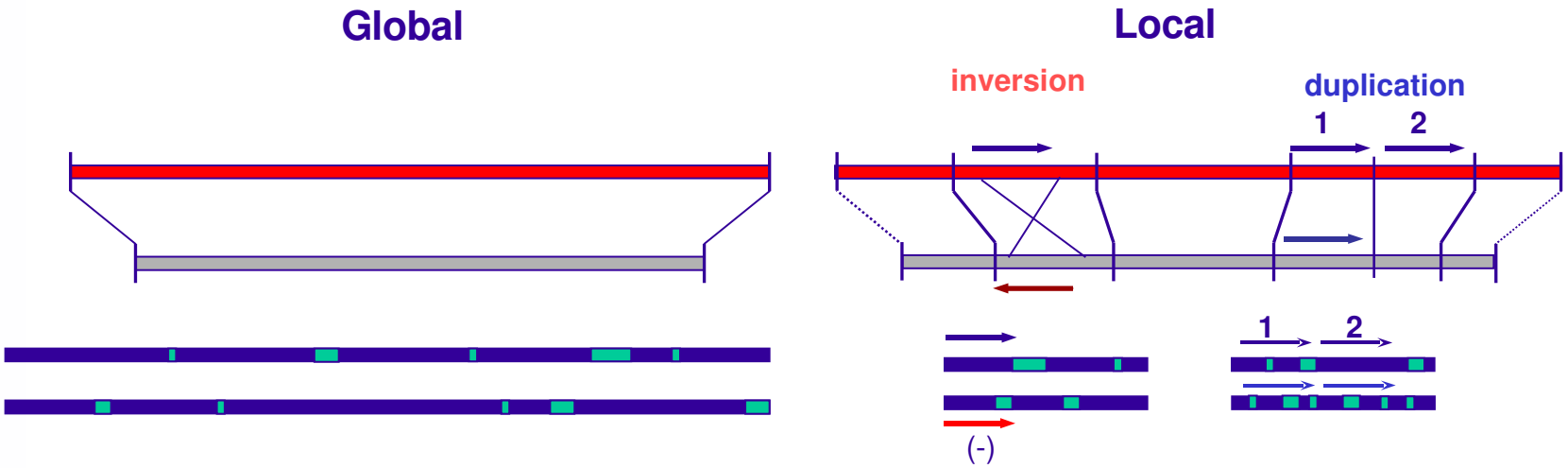
Genomic Alignments

- 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
- Pecan
  - multiple global alignments
  - all vs all coding exons wublastp -> Mercator -> Pecan on each syntenic block

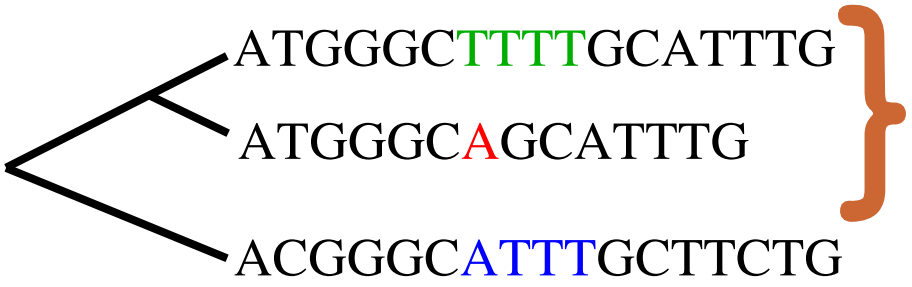
Global vs. Local Alignments



	Advantages	Disadvantages
Local	For large genomic regions Can identify inversions	Fails to identify insertions or deletions
Global	Can detect insertions or deletions	Fails to detect inversions

Pecan

a consistency based multiple-alignment program




```
ATGGGCTTTTGCATTTG
ATGGGC---AGCATTTG
```

vs

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

```
ATGGGCTTTTGCATTTG
ACGGGCATTTGCTTCTG
```

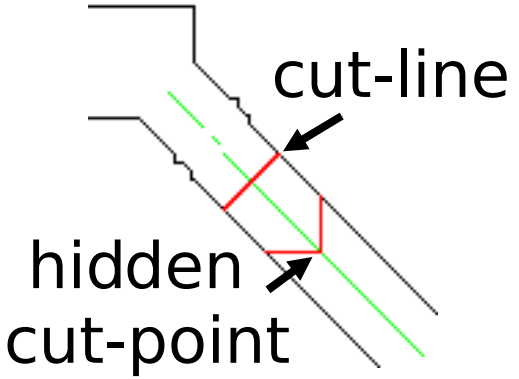
```
ATGGGCA---GCATTTG
ACGGGCATTTGCTTCTG
```

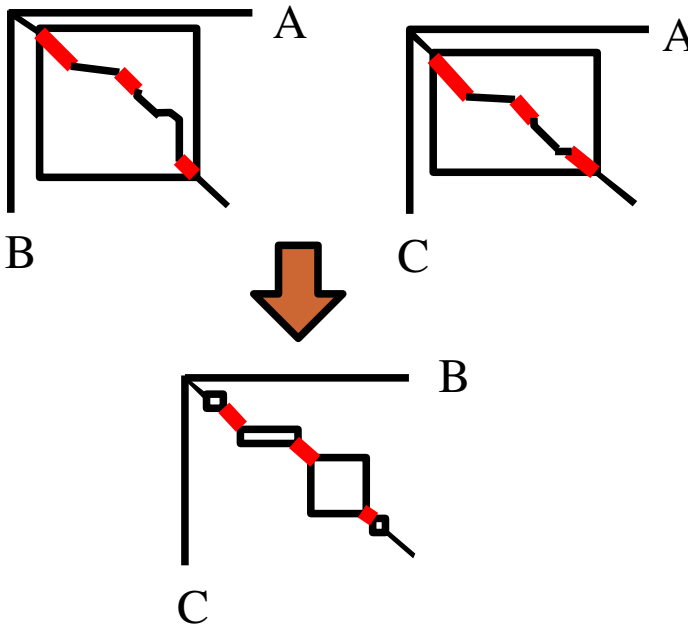
```
ATGGGCTTTTGCATTTG
ATGGGCA---GCATTTG
ACGGGCATTTGCTTCTG
```

Takes into account all pairwise alignments, across the entire tree

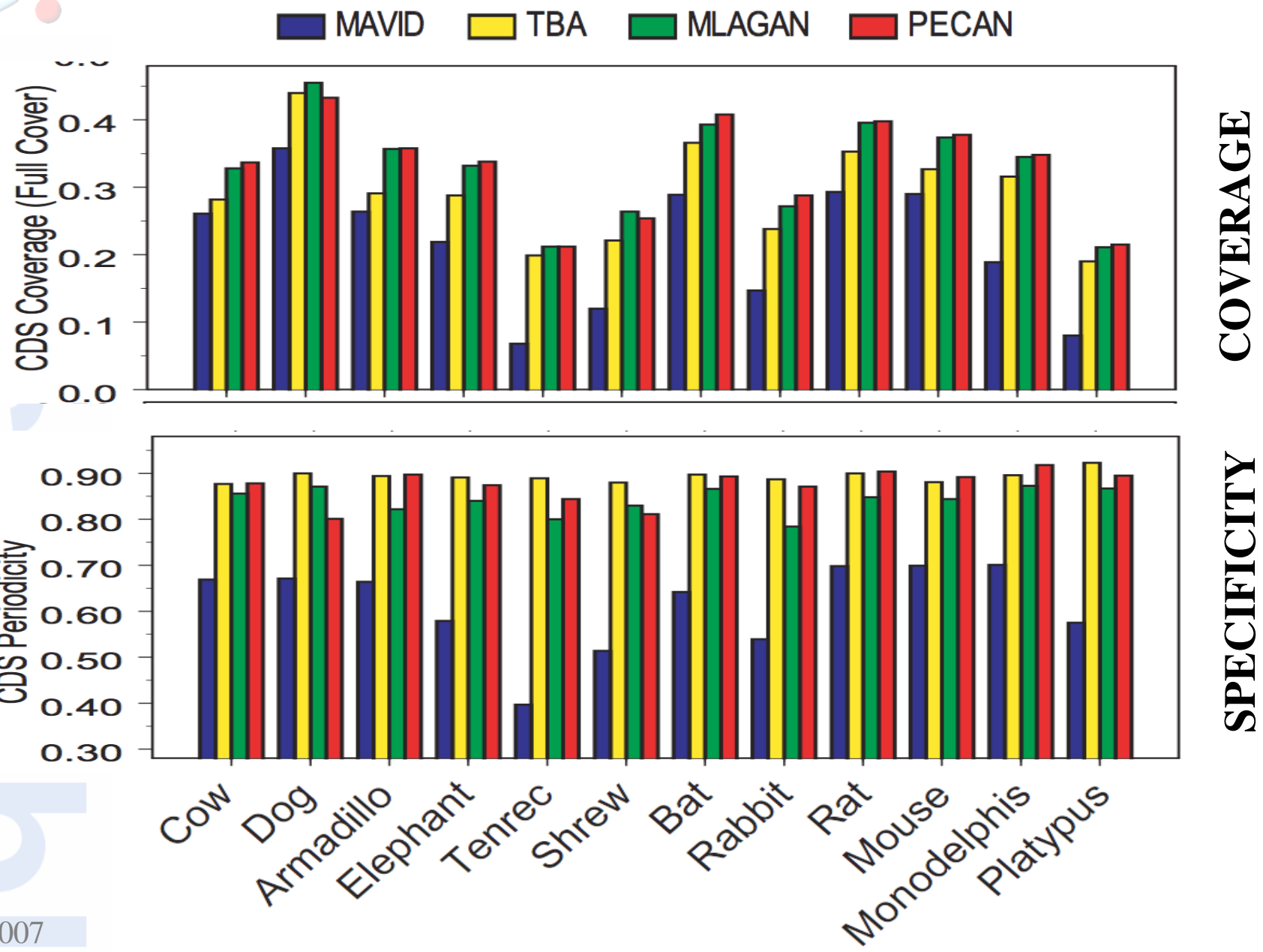
Pecan optimizations

- Look for anchors (exonerate)
- perform a banded alignment
- Break-up alignments into fragments
- Much redundancy between pairwise alignments: use transitive anchors

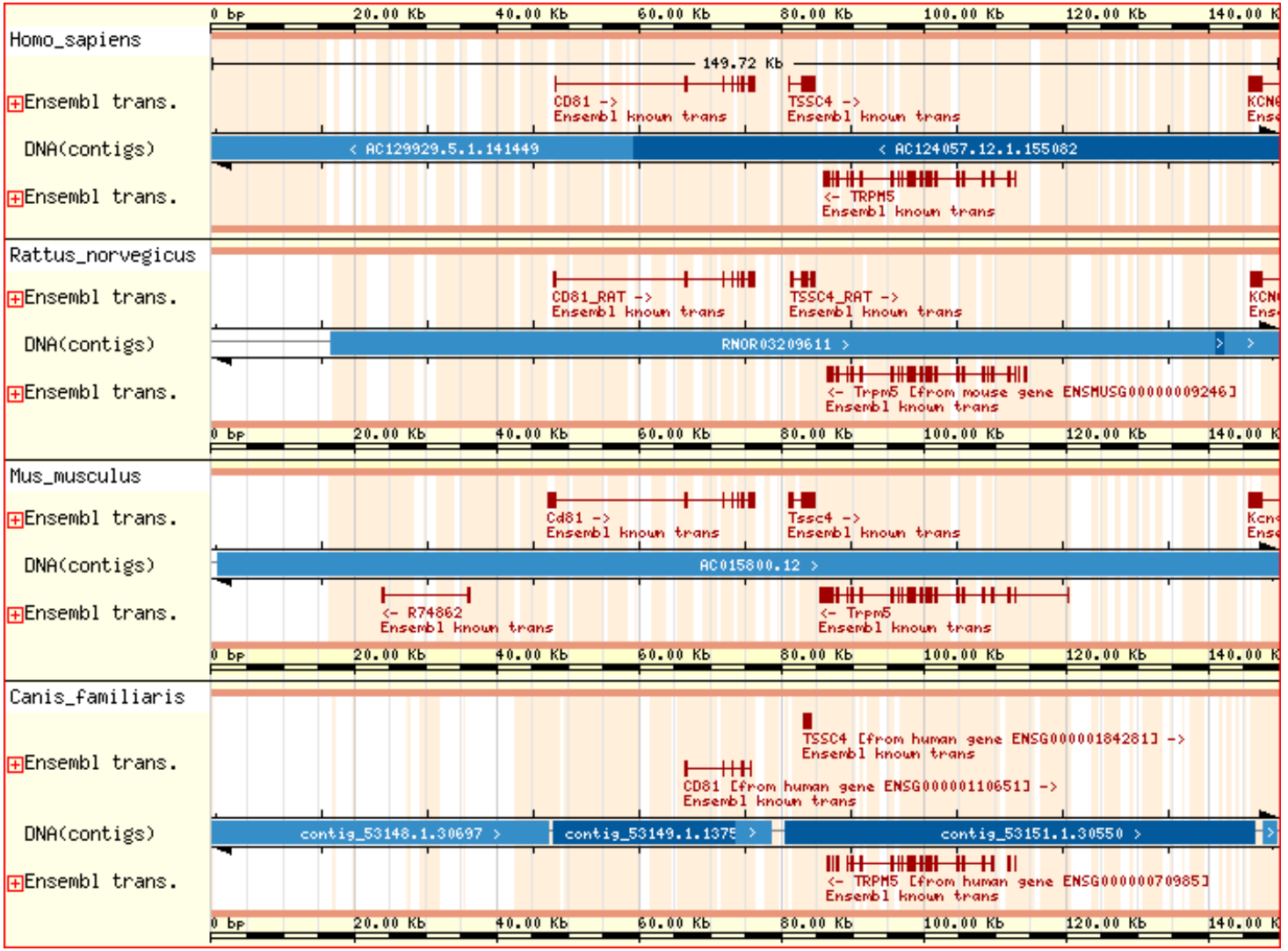




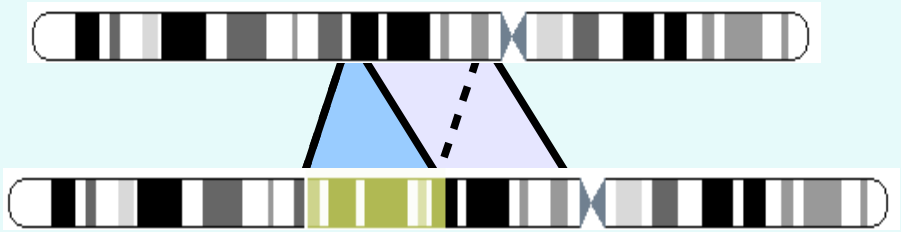
Encode Comparison



Example of AlignSliceView between Human/Mouse/Rat/Dog with PECAN

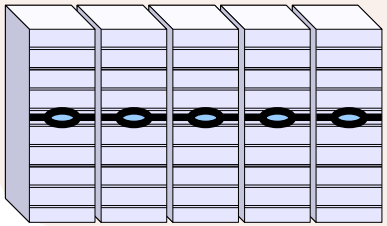


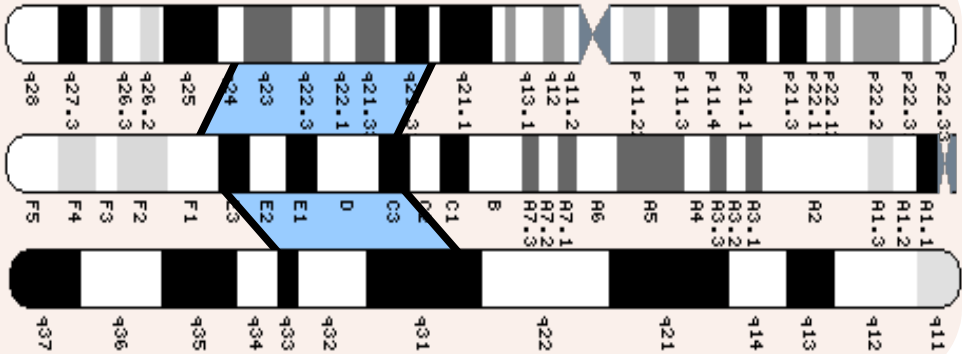
New challenges

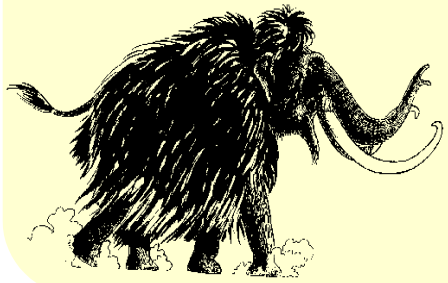


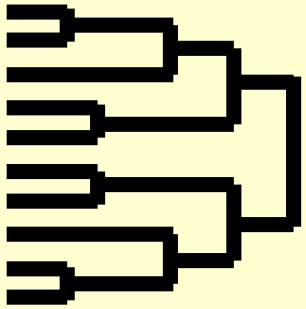
How to deal with duplications?

How to align such large segments?

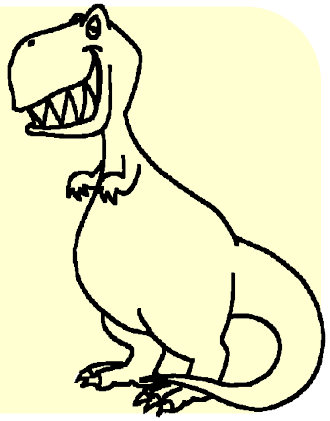






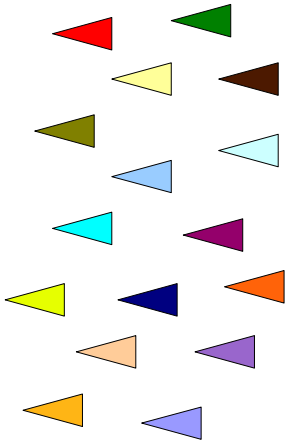


How to predict ancestral genomes?

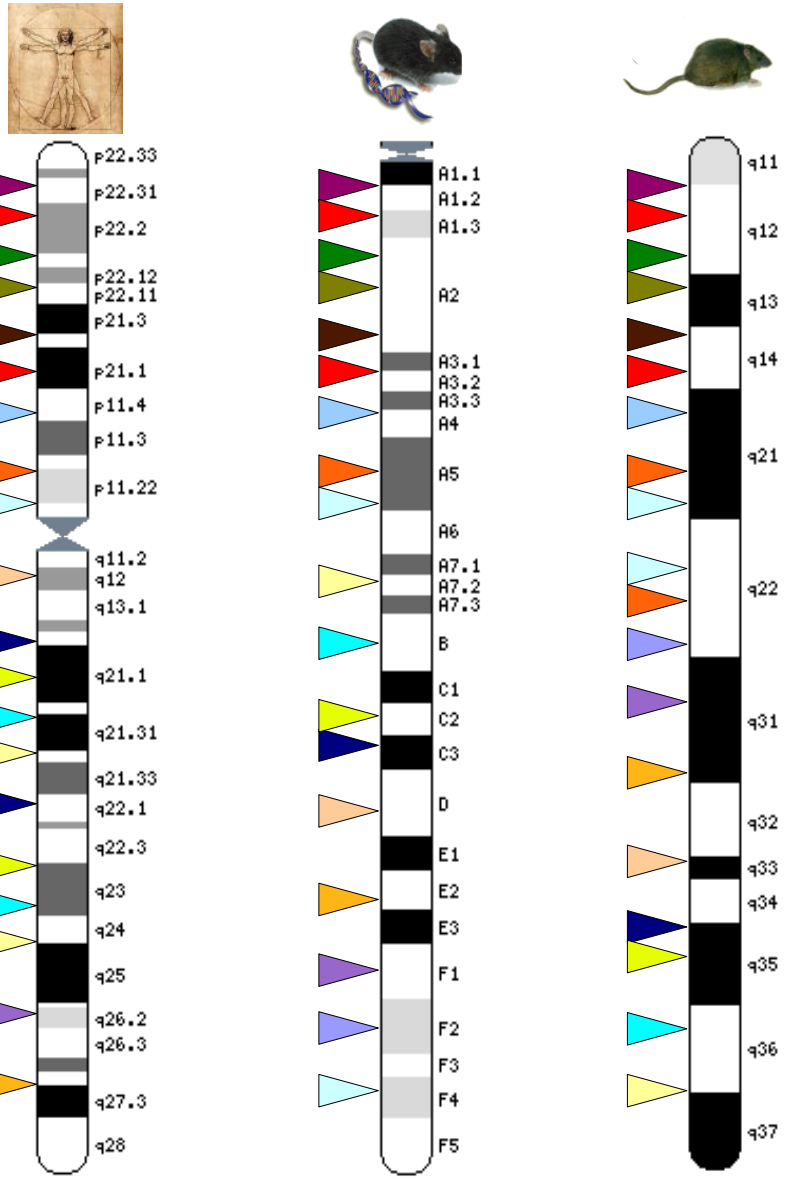


Enredo

Anchors

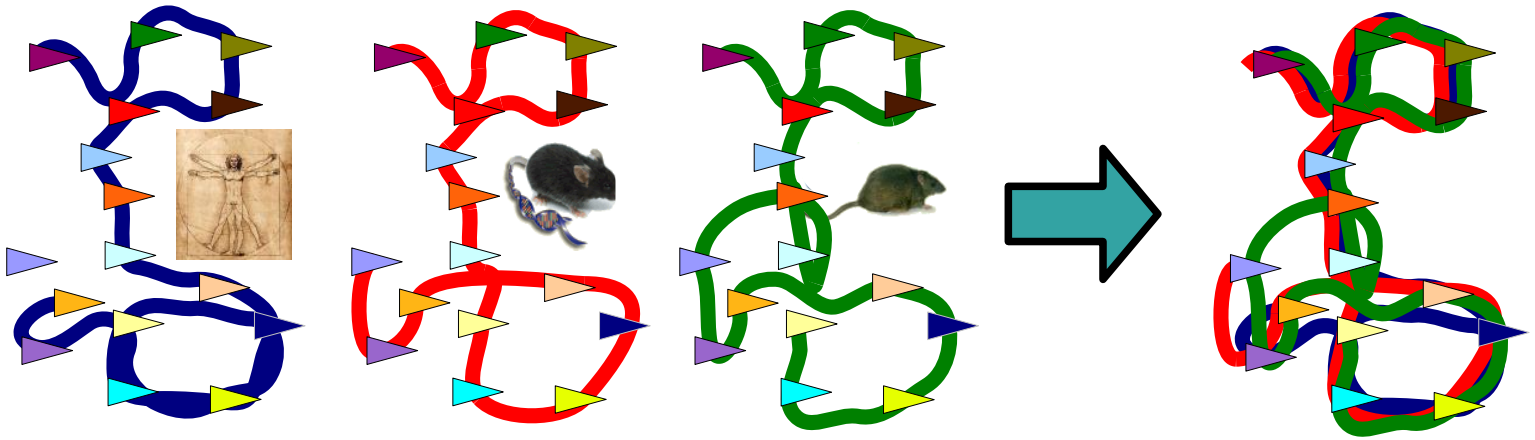


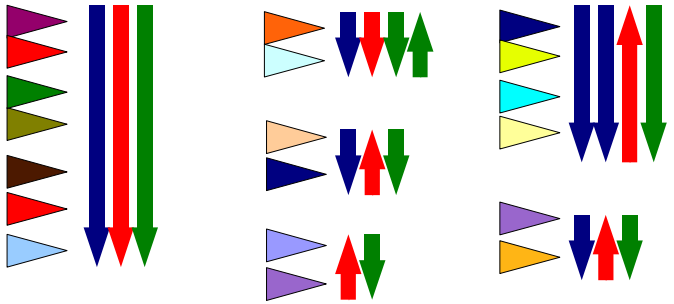
500.000 anchors
for mammals
---
more than 1 anchor
per 10Kb



Enredo

Solving the graph





Common paths define co-linear regions

This process allows for mismatches in the paths

ORTHEUS

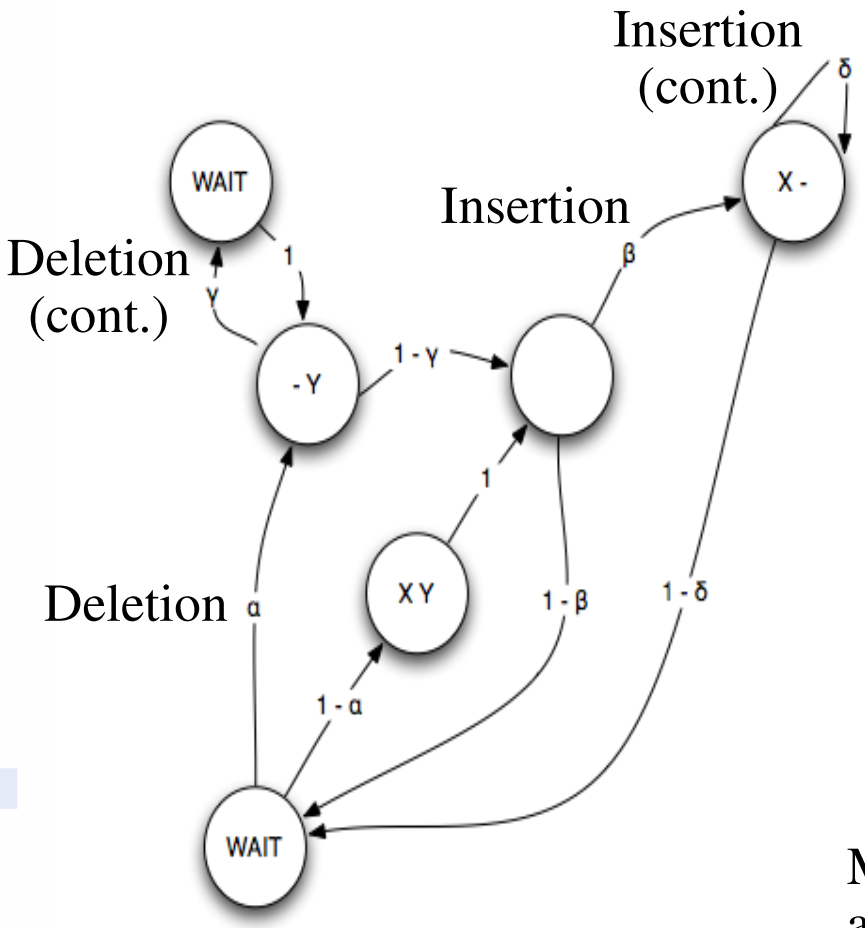
an ancestral sequence inference program

- Addresses the inference of insertion-deletion histories and substitution events
- Starts from a (multiple) alignment and assumes a fixed tree
- Reconstructs the ancestral sequences in the tree and refines the input alignment
- Insertion/deletion events are handled using a branch transducer model
- Substitution are handled using Tamura-Nei nucleotide substitution model
- Ancestral sequence are inferred using weighted sequence graph

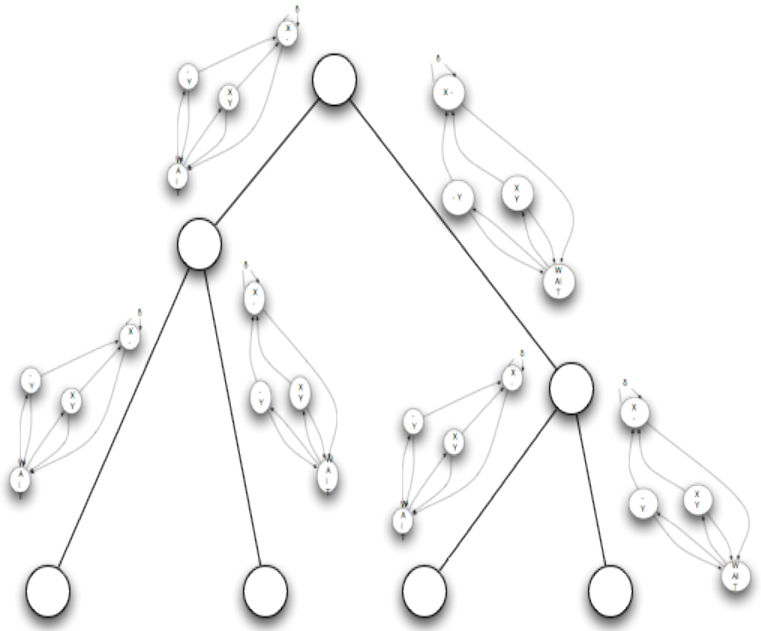
ORTHEUS

Branch Transducer Model

Four transition parameters model:



Applied to a tree:



Matches and insertion/deletion events are propagated down the tree

Protein Homology

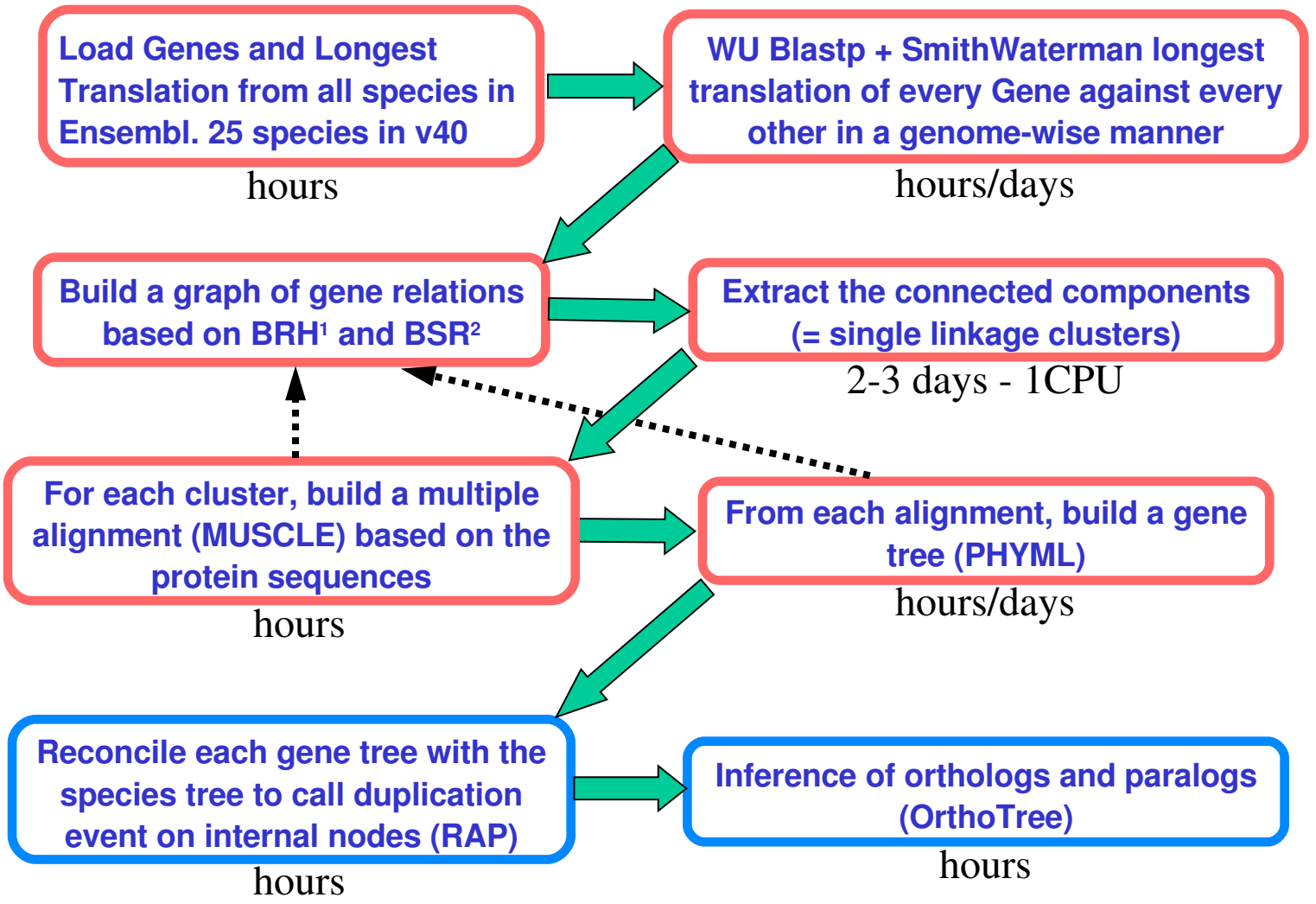
- (e! 38):
  - Orthologue predictions based on ‘best reciprocal blast hits’ and ‘synteny extensions’
  - Young paralogues for a selected set of species

- *e! 39+:*
  - *orthologues and paralogues are inferred from protein trees*

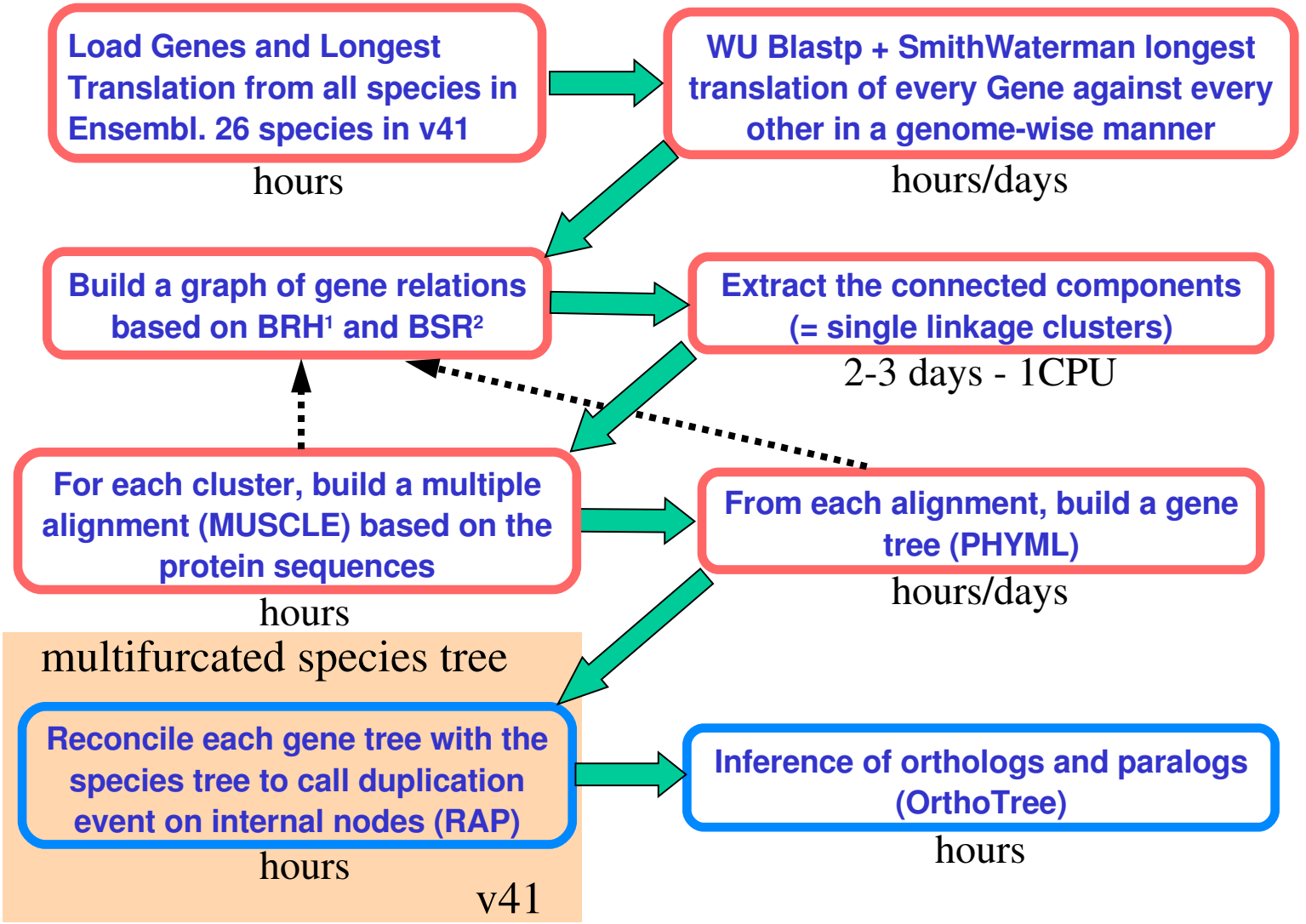
Homology

- BRH/RHS -- pairs of species (n*(n-1)/2)
- Cases of fast evolving genes
- Global view of the evolution history of the gene considered

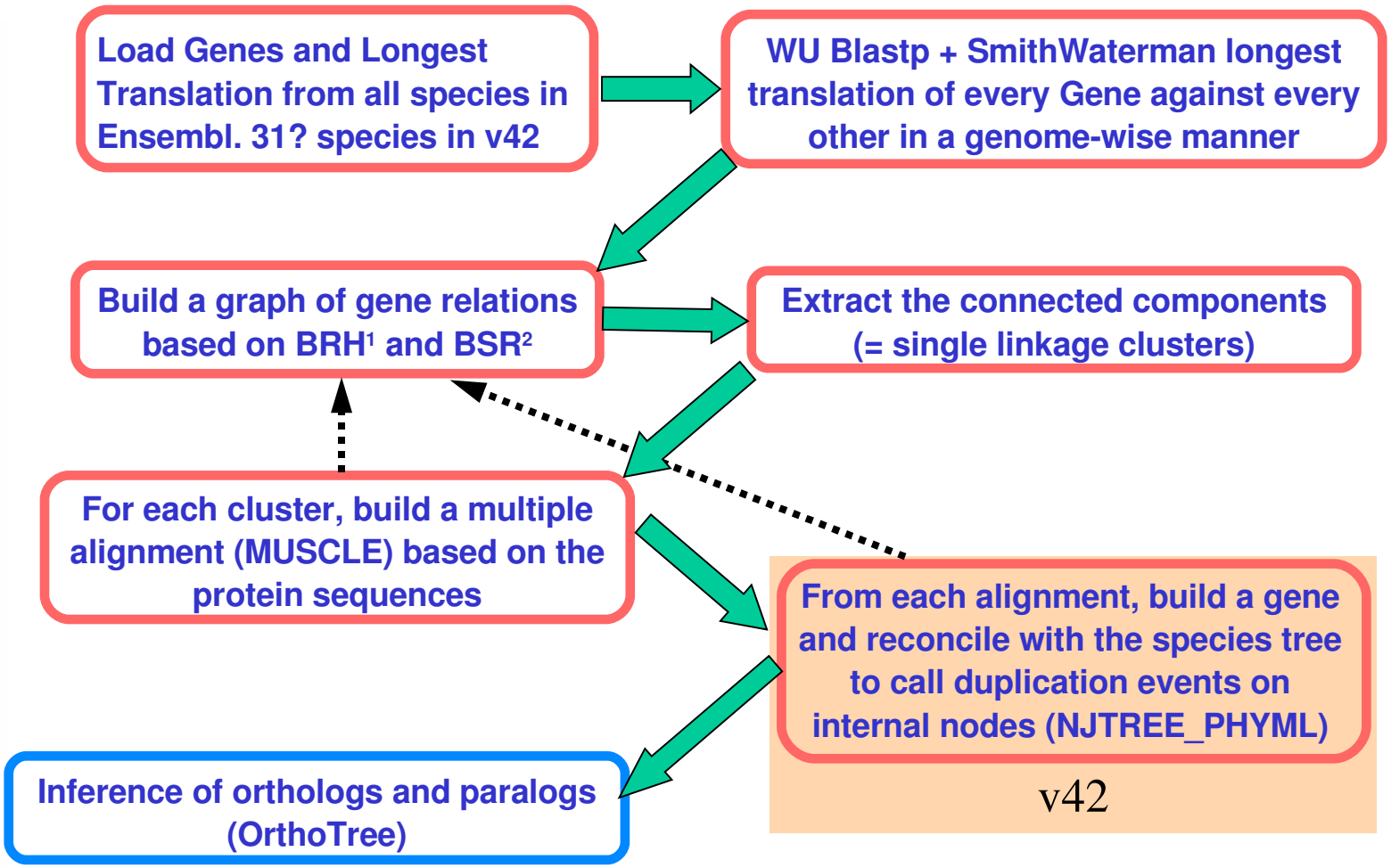
- Phylogeny: Orthology/Paralogy in one go



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.

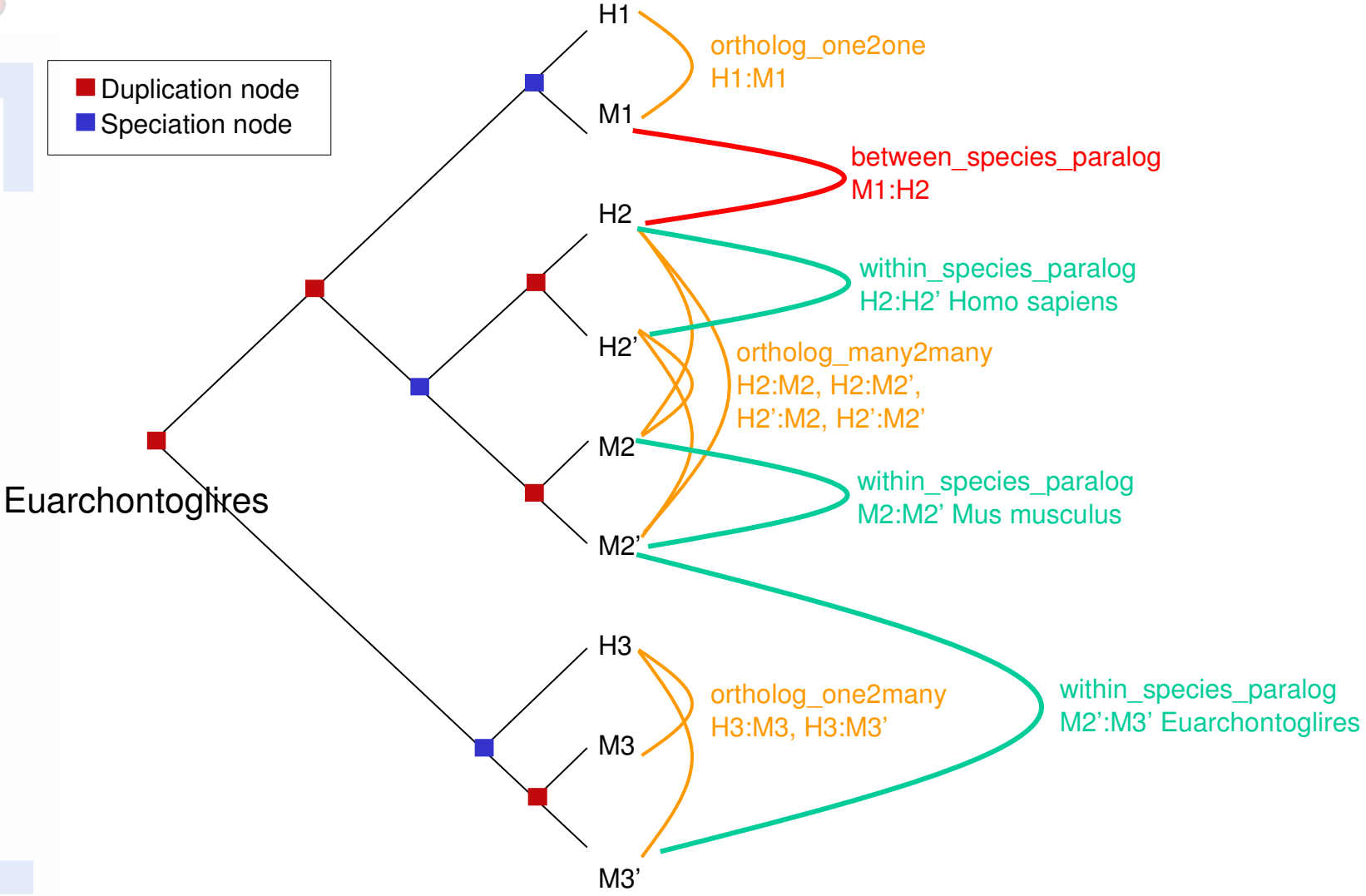


BSR: Blast Score Ratio. When 2 proteins P1 and P2 are compared, BSR=scoreP1P2/max(self-scoreP1 or self-scoreP2). The default threshold used in the initial clustering step is 0.33.



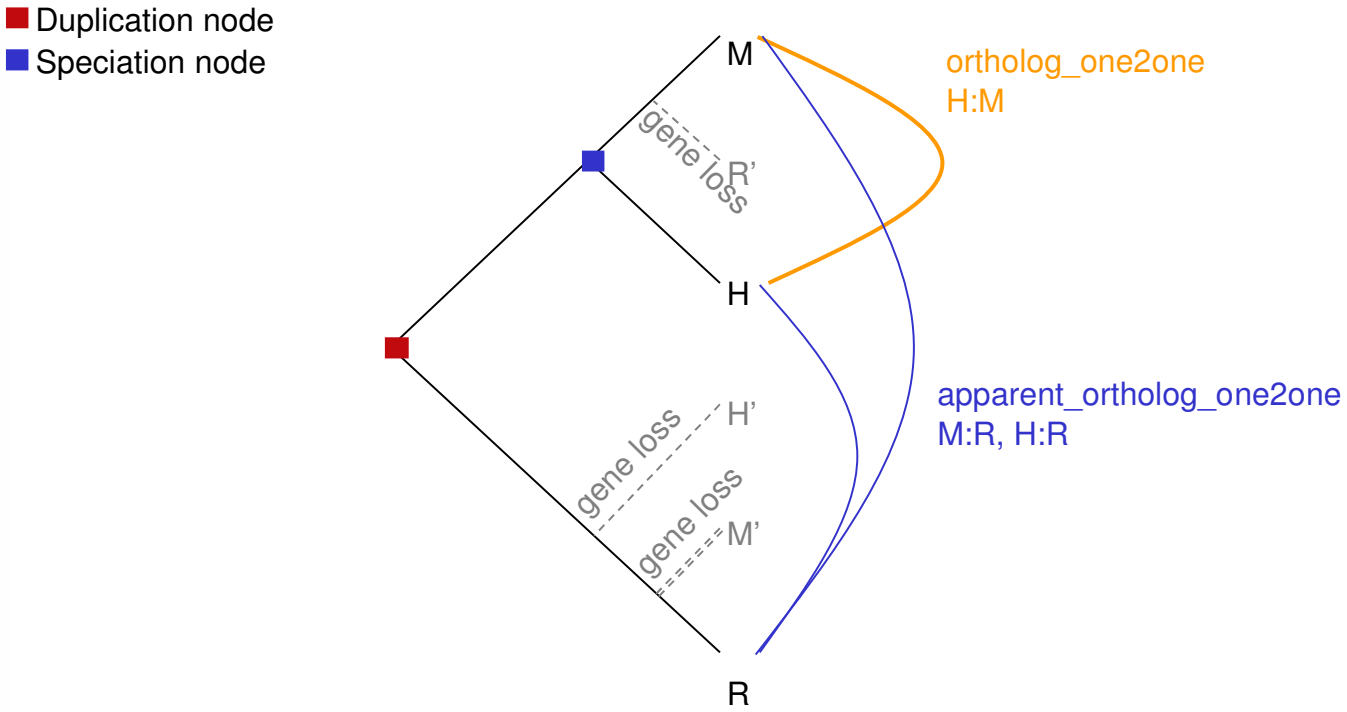
BSR: Blast Score Ratio. When 2 proteins P1 and P2 are compared, BSR=scoreP1P2/max(self-scoreP1 or self-scoreP2). The default threshold used in the initial clustering step is 0.33.

Protein trees and homology



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.

A special case of homology



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.

Gene tree : 1st data assessment

Good concordance with the classical BRH/RHS paired species approach (RHS are based on gene order conservation)
Find more complex one-to-many and many-to-many relations

Human/Mouse

	RHS	BRH	NEW
many2many	177	113	1,439
one2many	725	1,309	2,815
one2one	205	10,736	109
apparent one2one	78	1,571	104
lost	2,027	2,060	

19,381

Human/Drosophila

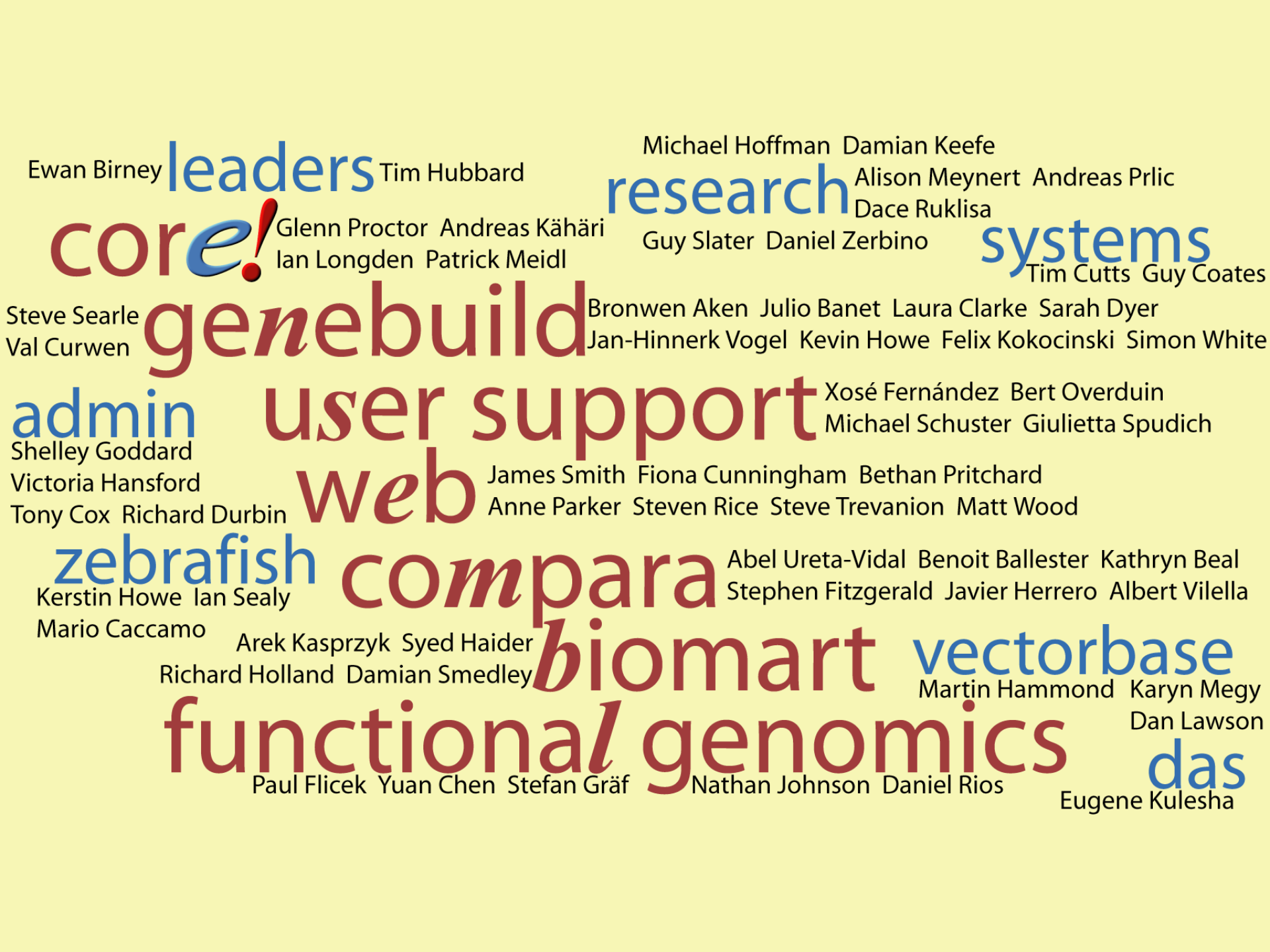
	BRH	NEW
many2many	170	1,599
one2many	1,870	4,563
one2one	880	80
apparent one2one	2,040	241
lost	620	

11,443

Future plans: convergence with TreeFam

Family

- Gene family clustering predictions
- Runs on **all Ensembl transcripts** plus all Uniprot/SWISSPROT and Uniprot/SPTREMBL metazoan proteins
- Algorithm is based on all vs all blastp, MCL clustering, Muscle multiple aligner


Ewan Birney
leaders
Tim Hubbard
Michael Hoffman
Damian Keefe
research
Alison Meynert
Andreas Prlic
Dace Ruklisa
core!
Glenn Proctor
Andreas Kähäri
Ian Longden
Patrick Meidl
Guy Slater
Daniel Zerbino
systems
Tim Cutts
Guy Coates
Steve Searle
Val Curwen
genebuild
Bronwen Aken
Julio Banet
Laura Clarke
Sarah Dyer
Jan-Hinnerk Vogel
Kevin Howe
Felix Kokocinski
Simon White
admin
user support
Xosé Fernández
Bert Overduin
Michael Schuster
Giulietta Spudich
Shelley Goddard
Victoria Hansford
Tony Cox
Richard Durbin
web
James Smith
Fiona Cunningham
Bethan Pritchard
Anne Parker
Steven Rice
Steve Trevanion
Matt Wood
zebrafish
compara
Abel Ureta-Vidal
Benoit Ballester
Kathryn Beal
Stephen Fitzgerald
Javier Herrero
Albert Vilella
Kerstin Howe
Ian Sealy
Mario Caccamo
Arek Kasprzyk
Syed Haider
Richard Holland
Damian Smedley
biomart
vectorbase
Martin Hammond
Karyn Megy
Dan Lawson
functional genomics
das
Paul Flicek
Yuan Chen
Stefan Gräf
Nathan Johnson
Daniel Rios
Eugene Kulesha