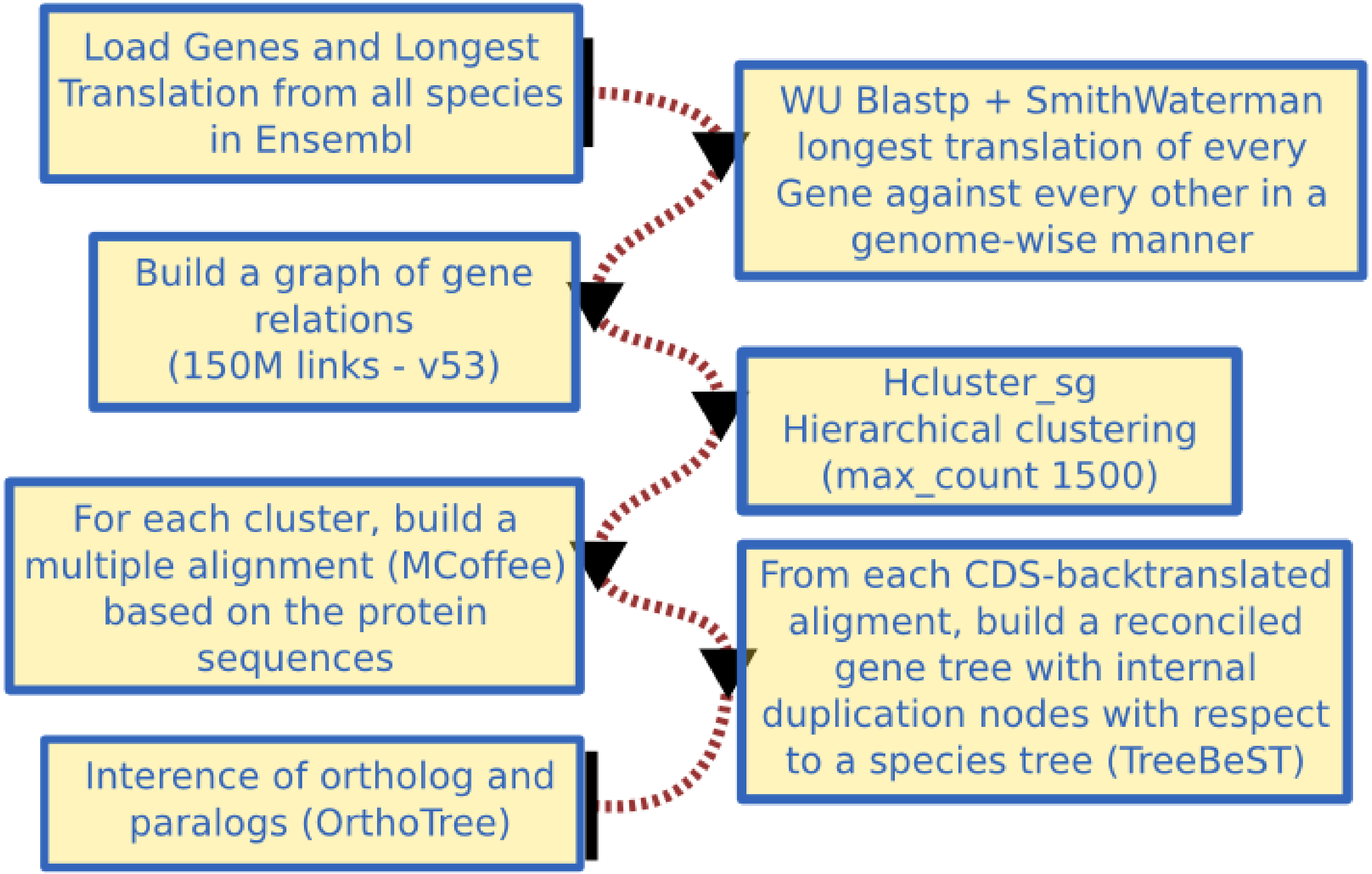
Ensembl Gene Trees & Annotation propagation

Javier Herrero & Glenn Proctor
Vertebrate Genomics Team
EMBL-EBI
Wellcome Trust Genome Campus
Hinxton CB10 1SD, UK

GeneTree pipeline overview



Load Genes and Longest Translation from all species in Ensembl

WU Blastp + SmithWaterman longest translation of every Gene against every other in a genome-wise manner

Build a graph of gene relations (150M links - v53)

Hcluster_sg Hierarchical clustering (max_count 1500)

For each cluster, build a multiple alignment (MCoffee) based on the protein sequences

From each CDS-backtranslated aligment, build a reconciled gene tree with internal duplication nodes with respect to a species tree (TreeBeST)

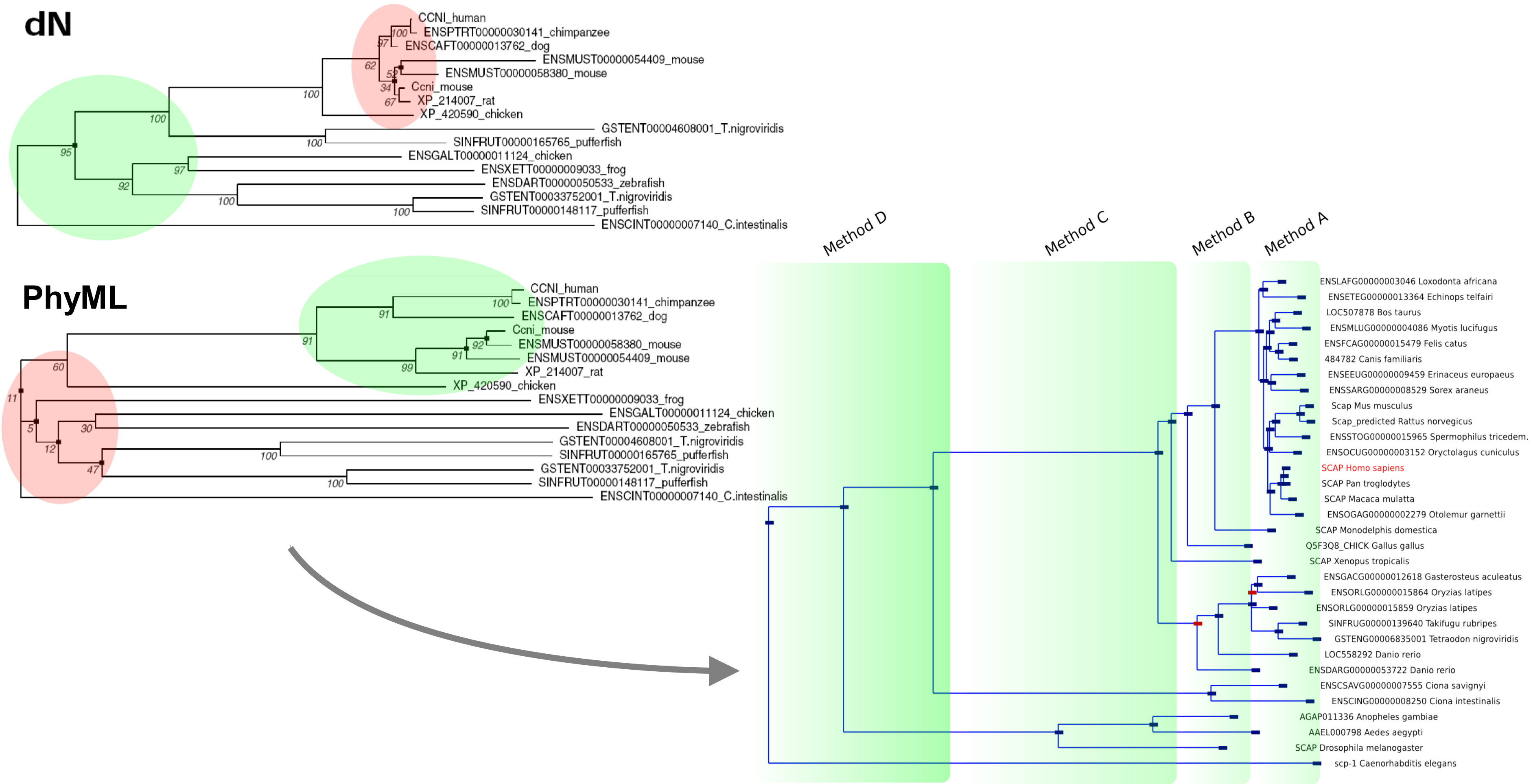
Interence of ortholog and paralogs (OrthoTree)

Vilella et al., Genome Res. 2009

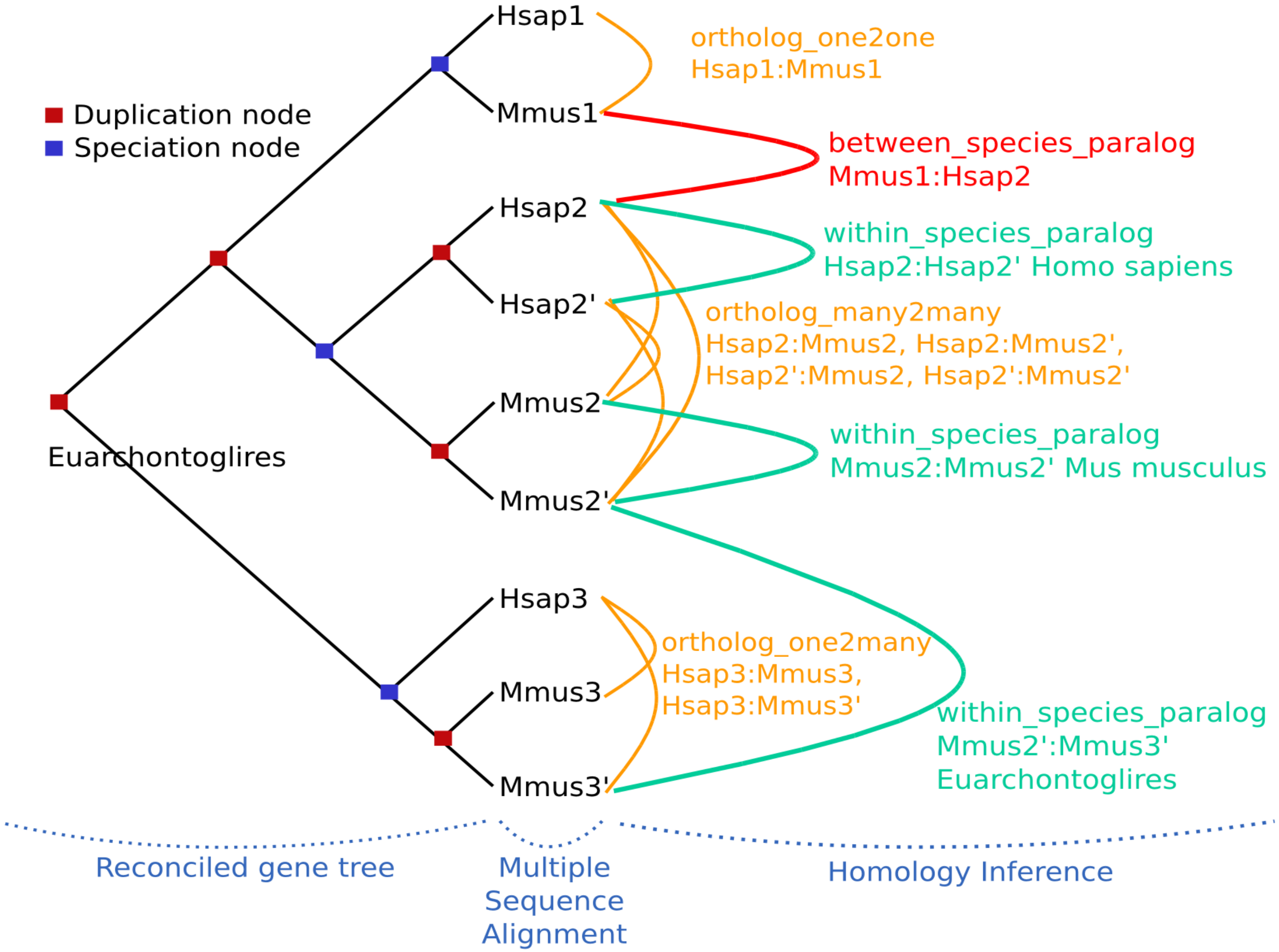
TreeBeST - treemerge

- 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)

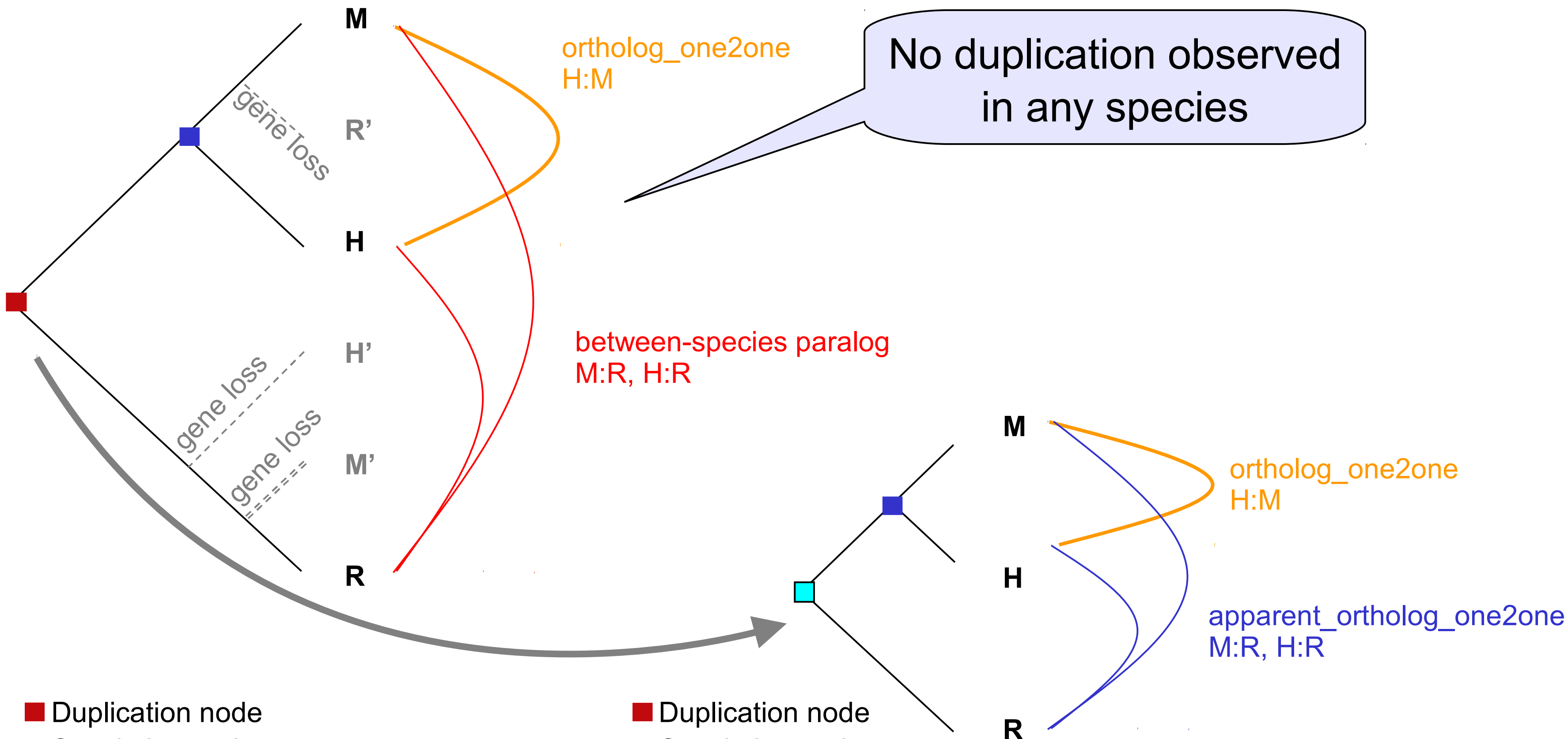
Each method performs best at a given level


dN
CCNI_human
100
ENSPTRT00000030141_chimpanzee
97
ENSCAFT00000013762_dog
62
ENSMUST00000054409_mouse
52
ENSMUST00000058380_mouse
34
Ccni_mouse
67
XP_214007_rat
100
XP_420590_chicken
100
GSTENT00004608001_T.nigroviridis
100
SINFRUT00000165765_pufferfish
95
ENSGALT00000011124_chicken
97
ENSXETT00000009033_frog
92
ENSDART00000050533_zebrafish
100
GSTENT00033752001_T.nigroviridis
100
SINFRUT00000148117_pufferfish
ENSCINT00000007140_C.intestinalis
PhyML
CCNI_human
100
ENSPTRT00000030141_chimpanzee
91
ENSCAFT00000013762_dog
91
Ccni_mouse
92
ENSMUST00000058380_mouse
91
ENSMUST00000054409_mouse
99
XP_214007_rat
60
XP_420590_chicken
11
ENSXETT00000009033_frog
ENSGALT00000011124_chicken
5
30
ENSDART00000050533_zebrafish
12
GSTENT00004608001_T.nigroviridis
100
SINFRUT00000165765_pufferfish
47
GSTENT00033752001_T.nigroviridis
100
SINFRUT00000148117_pufferfish
ENSCINT00000007140_C.intestinalis
Method D
Method C
Method B
Method A
ENSLAFG00000003046 Loxodonta africana
ENSETEG00000013364 Echinops telfairi
LOC507878 Bos taurus
ENSMLUG00000004086 Myotis lucifugus
ENSFCAG00000015479 Felis catus
484782 Canis familiaris
ENSEEUG00000009459 Erinaceus europaeus
ENSSARG00000008529 Sorex araneus
Scap Mus musculus
Scap_predicted Rattus norvegicus
ENSSTOG00000015965 Spermophilus tricedem.
ENSOCUG00000003152 Oryctolagus cuniculus
SCAP Homo sapiens
SCAP Pan troglodytes
SCAP Macaca mulatta
ENSOGAG00000002279 Otolemur garnettii
SCAP Monodelphis domestica
Q5F3Q8_CHICK Gallus gallus
SCAP Xenopus tropicalis
ENSGACG00000012618 Gasterosteus aculeatus
ENSORLG00000015864 Oryzias latipes
ENSORLG00000015859 Oryzias latipes
SINFRUG00000139640 Takifugu rubripes
GSTENG00006835001 Tetraodon nigroviridis
LOC558292 Danio rerio
ENSDARG00000053722 Danio rerio
ENSCSAVG00000007555 Ciona savignyi
ENSCING00000008250 Ciona intestinalis
AGAP011336 Anopheles gambiae
AAEL000798 Aedes aegypti
SCAP Drosophila melanogaster
scp-1 Caenorhabditis elegans

Homology inference



A special case of homology



M

R’

H

H’

M’

R

gene loss

gene loss

gene loss

ortholog_one2one
H:M

No duplication observed
in any species

between-species paralog
M:R, H:R

M

H

R

ortholog_one2one
H:M

apparent_ortholog_one2one
M:R, H:R

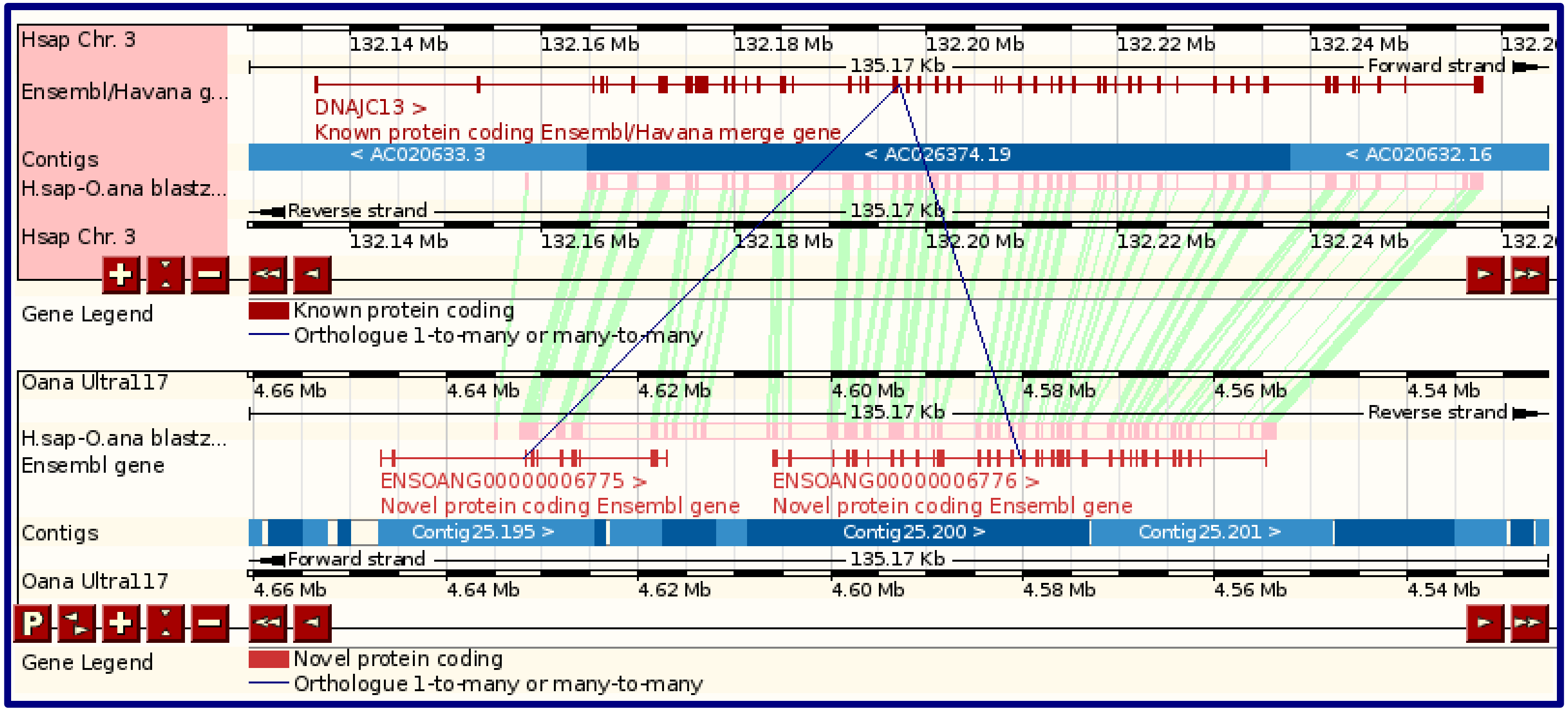
🟥 Duplication node
🟦 Speciation node

🟥 Duplication node
🟦 Speciation node
□ Ambiguous node

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.

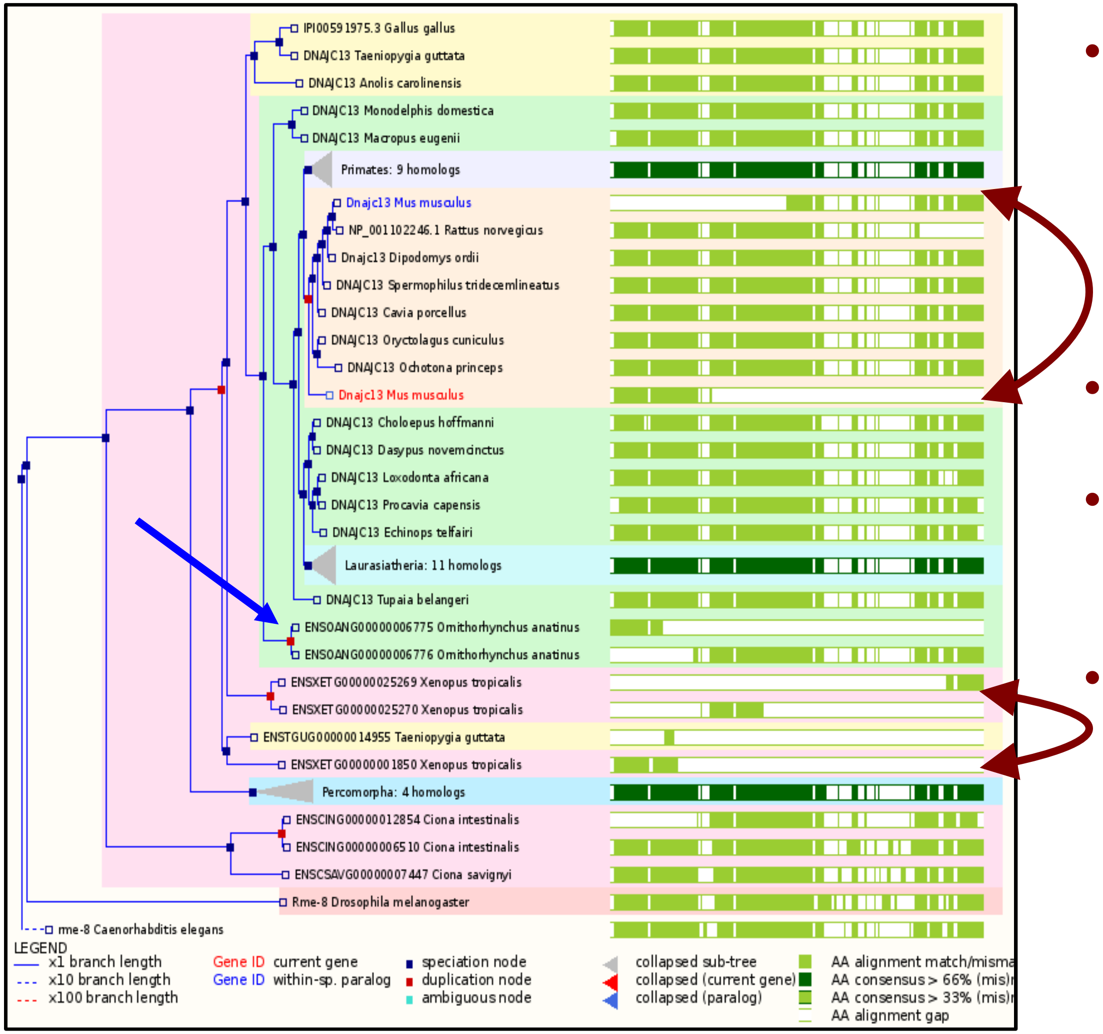
Gene splits

- DNAJC13 is split in platypus:



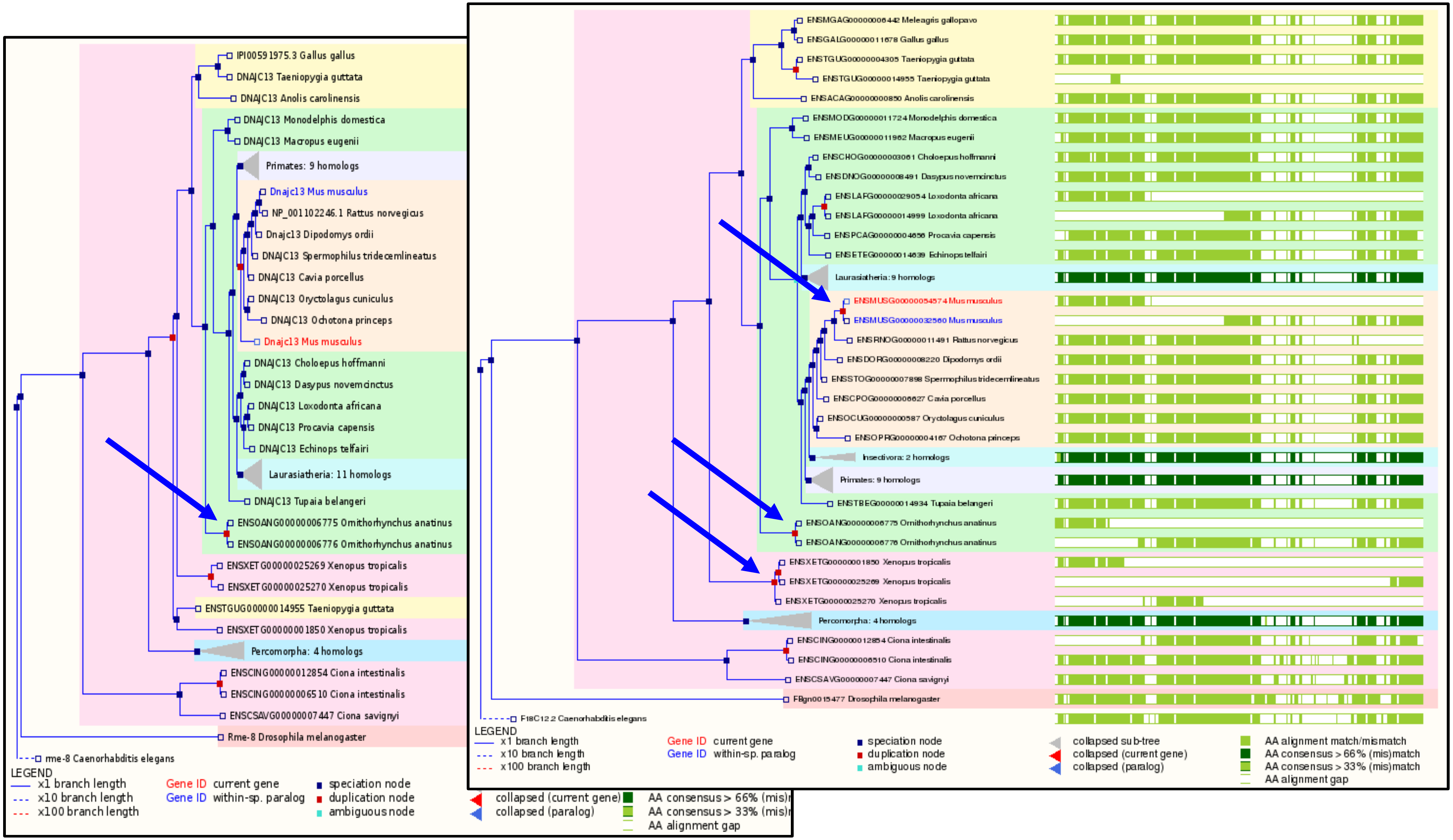
- We annotate them as gene split instead of within-species paralogues

Gene splits


IPI00591975.3 Gallus gallus
DNAJC13 Taeniopygia guttata
DNAJC13 Anolis carolinensis
DNAJC13 Monodelphis domestica
DNAJC13 Macropus eugenii
Primates: 9 homologs
Dnajc13 Mus musculus
NP_001102246.1 Rattus norvegicus
Dnajc13 Dipodomys ordii
DNAJC13 Spermophilus tridecemlineatus
DNAJC13 Cavia porcellus
DNAJC13 Oryctolagus cuniculus
DNAJC13 Ochotona princeps
Dnajc13 Mus musculus
DNAJC13 Choloepus hoffmanni
DNAJC13 Dasypus novemcinctus
DNAJC13 Loxodonta africana
DNAJC13 Procavia capensis
DNAJC13 Echinops telfairi
Laurasiatheria: 11 homologs
DNAJC13 Tupaia belangeri
ENSOANG00000006775 Ornithorhynchus anatinus
ENSOANG00000006776 Ornithorhynchus anatinus
ENSXETG00000025269 Xenopus tropicalis
ENSXETG00000025270 Xenopus tropicalis
ENSTGUG00000014955 Taeniopygia guttata
ENSXETG00000001850 Xenopus tropicalis
Percomorpha: 4 homologs
ENSCING00000012854 Ciona intestinalis
ENSCING00000006510 Ciona intestinalis
ENSCSAVG00000007447 Ciona savignyi
Rme-8 Drosophila melanogaster
rme-8 Caenorhabditis elegans
LEGEND
x1 branch length
x10 branch length
x100 branch length
Gene ID current gene
Gene ID within-sp. paralog
speciation node
duplication node
ambiguous node
collapsed sub-tree
collapsed (current gene)
collapsed (paralog)
AA alignment match/misma
AA consensus > 66% (mis)
AA consensus > 33% (mis)
AA alignment gap

- Most of the splits are in:
  - anole lizard, platypus, zebrafish, macaque, frog, dog, zebra finch
- Feedback to genebuilders
- Ideally, we want to see them together in the tree
- Copy & paste the missing sequence to the other part of the gene

Gene splits


IPI00591975.3 Gallus gallus
DNAJC13 Taeniopygia guttata
DNAJC13 Anolis carolinensis
DNAJC13 Monodelphis domestica
DNAJC13 Macropus eugenii
Primates: 9 homologs
Dnajc13 Mus musculus
NP_001102246.1 Rattus norvegicus
Dnajc13 Dipodomys ordii
DNAJC13 Spermophilus tridecemlineatus
DNAJC13 Cavia porcellus
DNAJC13 Oryctolagus cuniculus
DNAJC13 Ochotona princeps
Dnajc13 Mus musculus
DNAJC13 Choloepus hoffmanni
DNAJC13 Dasypus novemcinctus
DNAJC13 Loxodonta africana
DNAJC13 Procavia capensis
DNAJC13 Echinops telfairi
Laurasiatheria: 11 homologs
DNAJC13 Tupaia belangeri
ENSOANG00000006775 Ornithorhynchus anatinus
ENSOANG00000006776 Ornithorhynchus anatinus
ENSXETG00000025269 Xenopus tropicalis
ENSXETG00000025270 Xenopus tropicalis
ENSTGUG00000014955 Taeniopygia guttata
ENSXETG00000001850 Xenopus tropicalis
Percomorpha: 4 homologs
ENSCING00000012854 Ciona intestinalis
ENSCING00000006510 Ciona intestinalis
ENSCSAVG00000007447 Ciona savignyi
Rme-8 Drosophila melanogaster
rme-8 Caenorhabditis elegans
LEGEND
x1 branch length
x10 branch length
x100 branch length
Gene ID current gene
Gene ID within-sp. paralog
speciation node
duplication node
ambiguous node
collapsed (current gene)
collapsed (paralog)
AA consensus > 66% (mis)
AA consensus > 33% (mis)
AA alignment gap
ENSMGAG00000006442 Meleagris gallopavo
ENSGALG00000011678 Gallus gallus
ENSTGUG00000004305 Taeniopygia guttata
ENSTGUG00000014955 Taeniopygia guttata
ENSACAG00000000850 Anolis carolinensis
ENSMODG00000011724 Monodelphis domestica
ENSMEUG00000011962 Macropus eugenii
ENSCHOG00000003061 Choloepus hoffmanni
ENSDNOG00000008491 Dasypus novemcinctus
ENSLAFG00000029054 Loxodonta africana
ENSLAFG00000014999 Loxodonta africana
ENSPCAG00000004656 Procavia capensis
ENSETEG00000014639 Echinops telfairi
Laurasiatheria: 9 homologs
ENSMUSG00000054574 Mus musculus
ENSMUSG00000032560 Mus musculus
ENSRNOG00000011491 Rattus norvegicus
ENSDORG00000008220 Dipodomys ordii
ENSSTOG00000007898 Spermophilus tridecemlineatus
ENSCPOG00000006627 Cavia porcellus
ENSOCUG00000000587 Oryctolagus cuniculus
ENSOPRG00000004167 Ochotona princeps
Insectivora: 2 homologs
Primates: 9 homologs
ENSTBEG00000014934 Tupaia belangeri
ENSOANG00000006775 Ornithorhynchus anatinus
ENSOANG00000006776 Ornithorhynchus anatinus
ENSXETG00000001850 Xenopus tropicalis
ENSXETG00000025269 Xenopus tropicalis
ENSXETG00000025270 Xenopus tropicalis
Percomorpha: 4 homologs
ENSCING00000012854 Ciona intestinalis
ENSCING00000006510 Ciona intestinalis
ENSCSAVG00000007447 Ciona savignyi
FBgn0015477 Drosophila melanogaster
F18C12.2 Caenorhabditis elegans
LEGEND
x1 branch length
x10 branch length
x100 branch length
Gene ID current gene
Gene ID within-sp. paralog
speciation node
duplication node
ambiguous node
collapsed sub-tree
collapsed (current gene)
collapsed (paralog)
AA alignment match/mismatch
AA consensus > 66% (mis)match
AA consensus > 33% (mis)match
AA alignment gap

Orthologues in forthcoming e! 58

- Human-dog

Orthology type	Pairwise events
1:1 orthologues	14932
Apparent 1:1 orthologues	445
1:many orthologues	2193 (1520 human genes)
many:many orthologues	2095 (358 human genes)

- Human-zebrafish

Orthology type	Pairwise events
1:1 orthologues	8779
Apparent 1:1 orthologues	187
1:many orthologues	8633 (4582 human genes)
many:many orthologues	5034 (933 human genes)

NB: These values include ncRNA genes as well

Gene name projections

- From human and mouse to all vertebrates
  - Use the 1:1 orthologs
- From human to fish
  - Use the 1:many orthologs
  - Names become NAME (1 of 3), NAME (2 of 3), etc.
- Rules:
  - if source gene has an HGNC name, and target gene has no name or only a RefSeq predicted name, add name to target gene.
  - Change status to “KNOWN_BY_PROJECTION”

GO term projections

- Use the 1:1 orthologs only
- From human and mouse to all vertebrates
- From rat to human and mouse
- From zebrafish to other fish
- From human to zebrafish
- Rules:
  - add GO terms from source gene to target gene, avoiding duplicates. Only project source GO terms with evidence codes IDA, IEP, IGI, IMP, IPI.
  - Projected GO terms on target gene are given evidence code IEA.