

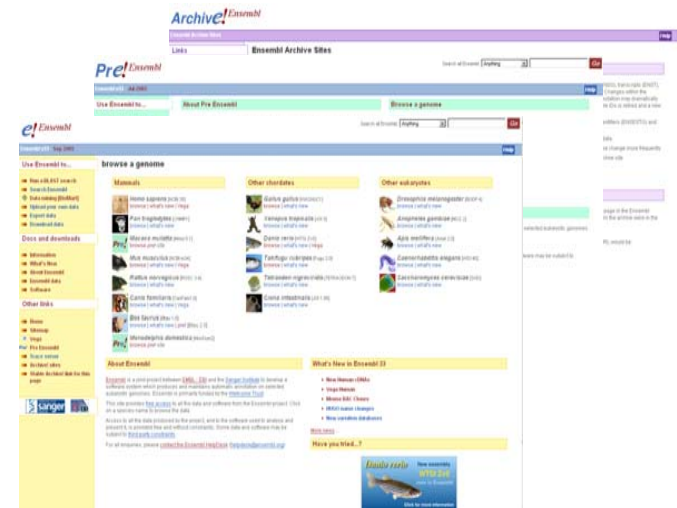
e!

Ensembl

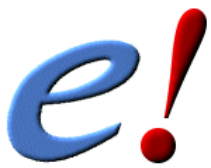
# Comparative proteomics in Ensembl

# Overview

- Rationale
- Comparative proteomics
  - Orthologues prediction
  - Protein clustering into families
- Outlook



The screenshot displays the Ensembl genome browser interface. At the top, there are navigation tabs for 'Archive!', 'Pre!', and 'Ensembl!'. Below the tabs, there is a search bar and a 'Browse a genome' button. The main content area is divided into several sections: 'Manuals', 'Other shortcuts', and 'Other ex-sites'. The 'Manuals' section includes links to 'How to use Ensembl', 'Genes', 'Variants', and 'Proteins'. The 'Other shortcuts' section includes links to 'Genes', 'Variants', and 'Proteins'. The 'Other ex-sites' section includes links to 'UniProt', 'NCBI', and 'PDB'. The page also features a 'What's New in Ensembl!' section and a 'How to use Ensembl!' section.



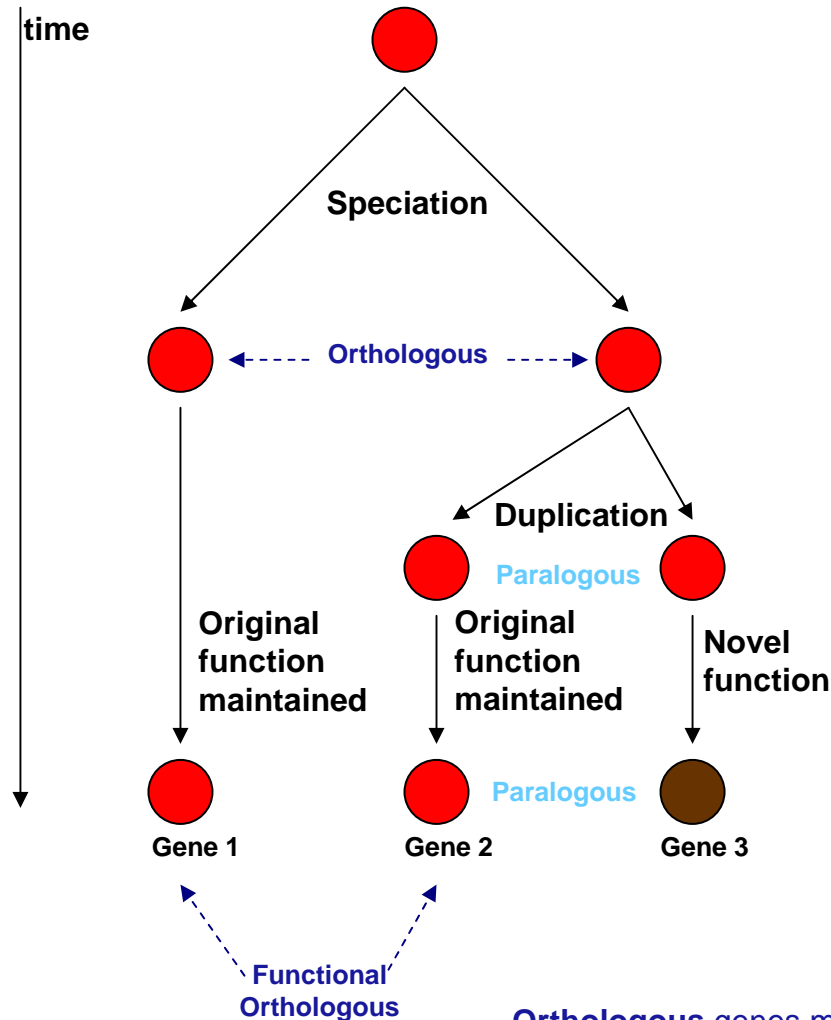
# Comparing different species

- From the Ensembl perspective joins species through
  - orthologous/paralogous genes links
  - protein family links
- From a broader perspective
  - How many genes are conserved?
  - Where are orthologous/paralogous genes?
  - Is gene order conserved?
  - Where are potential regulatory regions?
  - What is missing in one species, present only in another?

Ensembl

# Identifying orthologous genes

Ensembl



**Orthologous** genes most often have equivalent functions.  
**Paralogous** are genes related via duplication.

# Gene Evolution

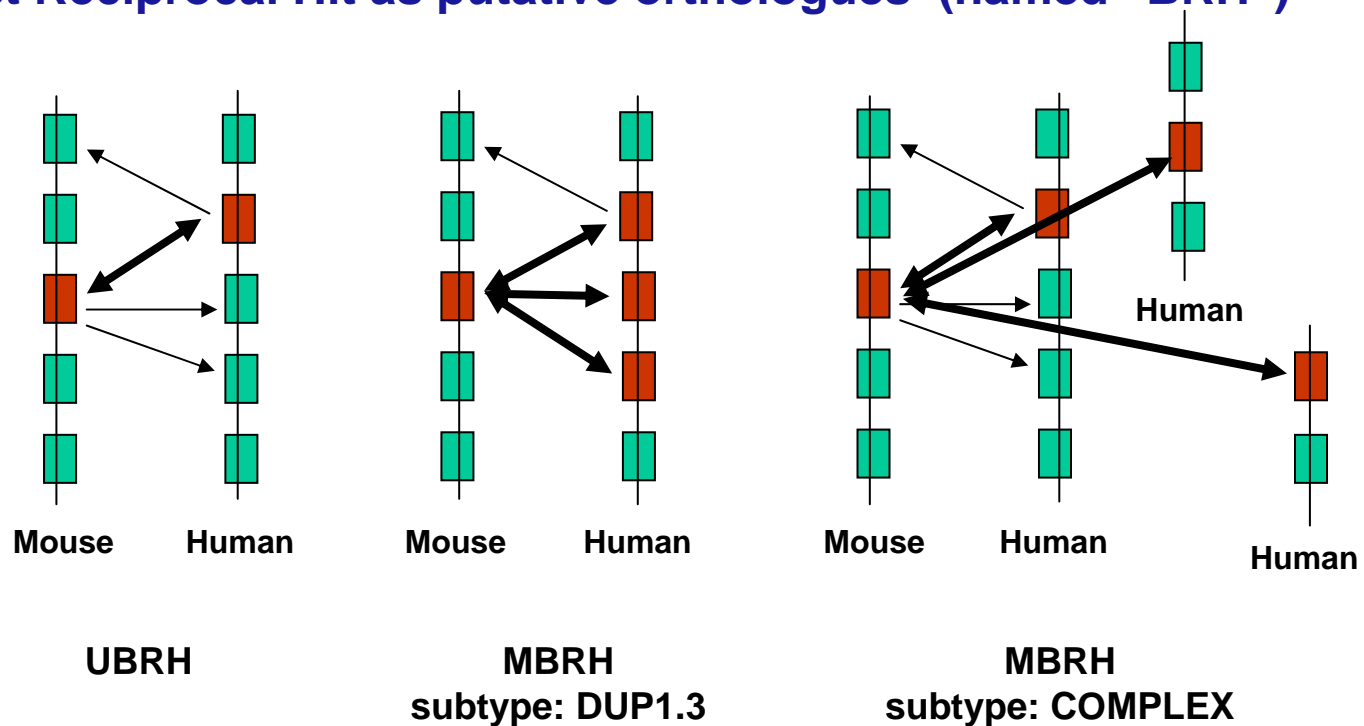
## *Orthologues and Paralogues*

Reconstruct the Molecular Evolutionary history from the evidence visible within the known extant genes

- **Divergence**
  - Speciation / Duplication
- **Change within allelic population**
  - Point Mutations / Selection / Drift
  - Exon/domain shuffling
  - Transposition / Translocation
  - Retroposition (reverse transcription)
- **Horizontal gene transfer?**

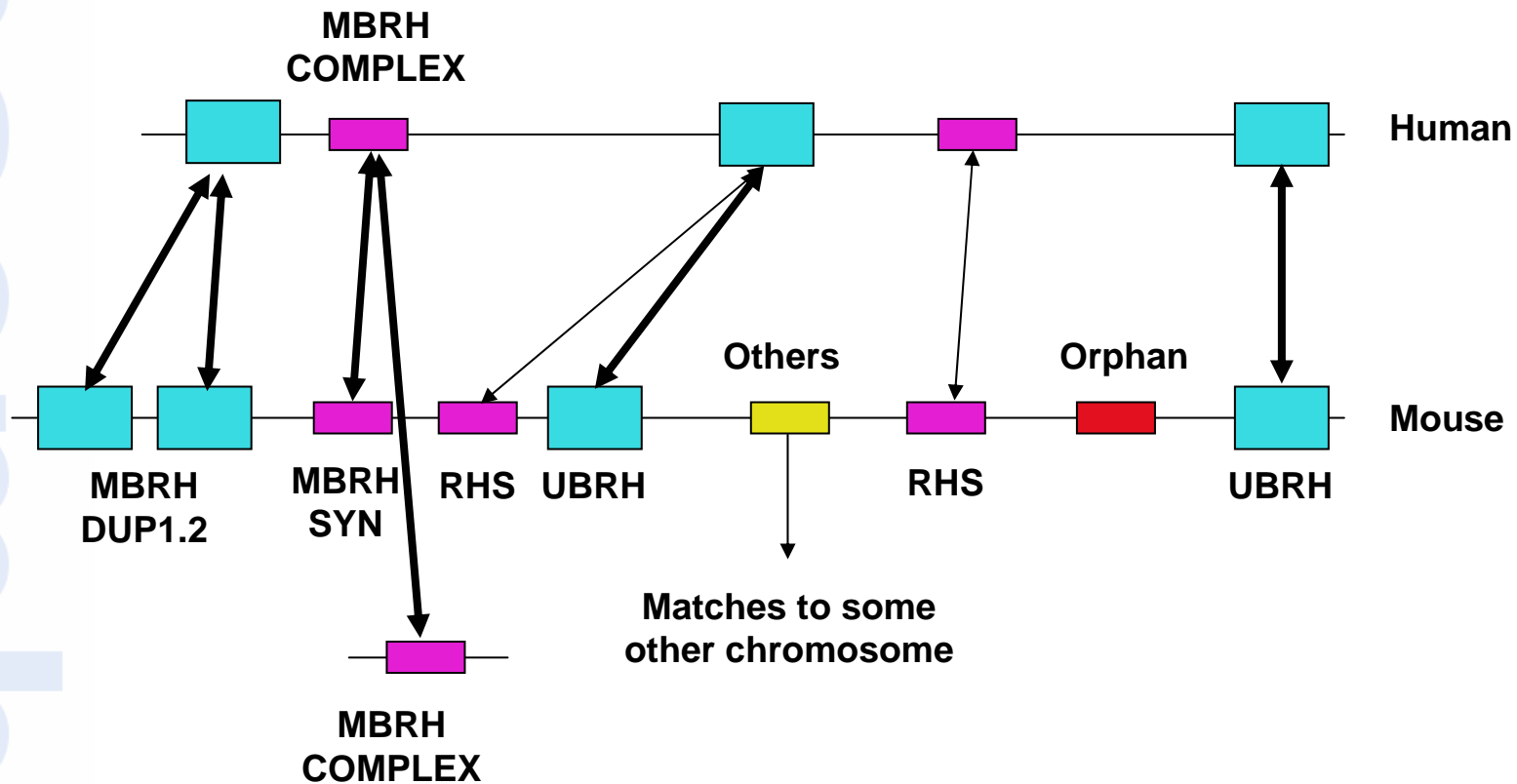
# Orthologues prediction

- Find orthologous genes by comparing the protein sets of two species (only the longest peptide considered).
- `blastp+sw` all *versus* all (on a paired species basis)
- Best Reciprocal Hit as putative orthologues (named “BRH”)



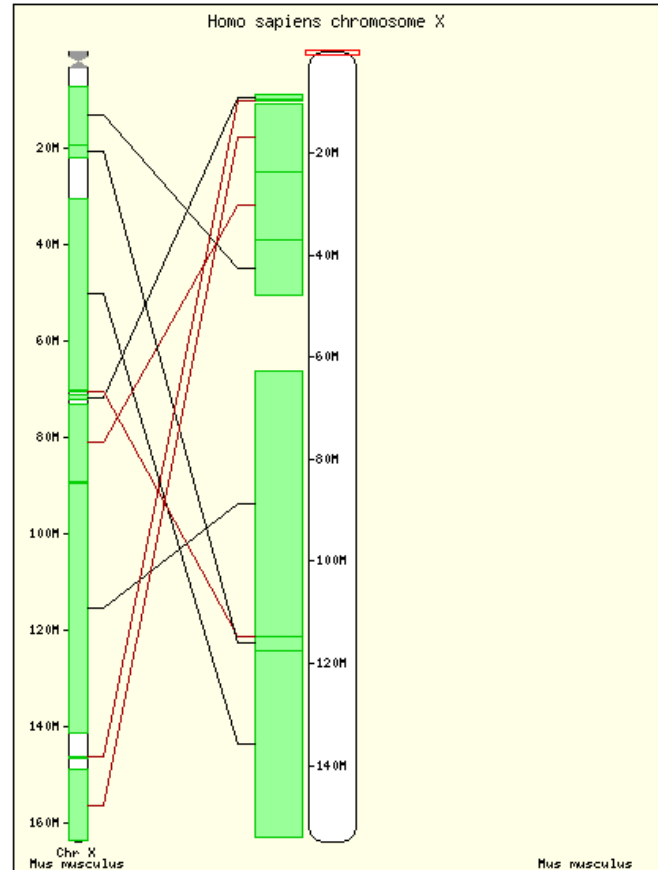
# RHS, Orphans and Others

Based on UBRH and MBRH-DUPs genomic coordinates in both species compared and gene order conservation, we identify additional orthologues or RHS for Reciprocal Hit supported by Synteny.



**Chromosome X**  
138,229,875 - 138,686,223

- View of Chromosome X
- Graphical view
- Graphical overview
- Export information about region
- Export sequence as FASTA
- Export EMBL file
- Export Gene info in region
- Export SNP info in region
- Export Vega info in region
- View alongside ...
- View syntonic regions
- View ... with *Canis familiaris*
- View ... with *Gallus gallus*
- View ... with *Mus musculus*
- View ... with *Pan troglodytes*
- View ... with *Rattus norvegicus*



**Homology Matches**

<i>Homo sapiens</i> Genes		<i>Mus musculus</i> Homologues
<a href="#">MCF2</a> (138.39 Mb) [ContigView]	->	<a href="#">Mcf2</a> [ContigView] [MultiContigView]
<a href="#">ATP11C</a> (138.53 Mb) [ContigView]	->	<a href="#">Atp11c</a> [ContigView] [MultiContigView]
<a href="#">XP_372255.1</a> (138.90 Mb) [ContigView]	->	<a href="#">XP_205232.2</a> [ContigView] [MultiContigView]
<a href="#">SOX3</a> (139.31 Mb) [ContigView]	->	<a href="#">Sox3</a> [ContigView] [MultiContigView]
<a href="#">LDOC1</a> (140.00 Mb) [ContigView]	->	<a href="#">Ldoc1</a> [ContigView] [MultiContigView]

**Navigate Homology**

[Upstream](#) (<138.39 Mb)   [Downstream](#) (>140.51 Mb)

**Change Chromosome**

Chromosome

Fields marked with \* are required

# For each orthologous gene pair

- **We store**
  - %identity, %positivity, %coverage, cigar lines, description (UBRH, MRHS), subtype (DUP1.2, SYN, COMPLEX), dN, dS
  - All the blastp+sw results are provided
- **Using the compara perl API**
  - Protein or cDNA protein-based alignment
  - 4D, 2D sites can also be easily retrieved
- **Future developments**
  - UBRH:SYN or UBRH:NON-SYN
  - Consider all isoforms for each gene
  - Build clusters of orthologues
  - Phylogenetic trees

# Protein clustering into families

- **Cluster proteins from different organisms that may share the same function**
- **Obtain some kind of description for 'novel' genes/proteins**
- **Locate family members over the whole genome**
- **Identify possible orthologues and paralogues in other species**

# Dataset used and comparisons

- Nearly a million proteins clustered:
  - All Ensembl proteins from all species in Ensembl
    - 513,860 predicted proteins
  - All metazoan (animal) proteins in UniProt
    - 53,224 UniProt/Swiss-Prot
    - 435,226 UniProt/TrEMBL
- Blastp all *versus* all, then clustering with MCL

### Ensembl Family ENSF00000000497

Family ID	ENSF00000000497	
Consensus annotation	AMBIGUOUS	
Prediction method	Protein families were generated using the MCL (Markov Clustering) package available at <a href="http://micans.org/mcl/">http://micans.org/mcl/</a> . The application of MCL to biological graphs was initially proposed by Enright A.J., Van Dongen S. and Ouzounis C.A. (2002) "An efficient algorithm for large-scale detection of protein families." Nucl. Acids. Res. 30, 1575-1584.	
Multiple alignments	Click to view multiple alignments of the 203 Ensembl members of this family.	<a href="#">JaView</a>
	Click to view multiple alignments of the 249 members of this family.	<a href="#">JaView</a>
Ensembl genes containing peptides in family ENSF00000000497		

**JaView alignment editor**

File Edit Font View Colour Calculate Align Help

10 20 30 40 50 60 70 80

```

C07E3_9/1-457      -----MMFILLVFLAALSITCVCLNF
CG14507-PE/1-457   MEDLHMVQVVIARUNDFFQYSMEKGSFTNDSREKELLPE-----GD
ENSANGP00000018874/1-457  ~RDARHEDVFIARANNDFGHSTKRWSETEQEDFLKQEVQSSRDG
ENSAPHP00000017092/1-457  -----MMFILLILNIIQEIICSEK
ENSAPHP00000020248/1-457  -----QEEQNDERMVEQSLTHK
ENSAPHP00000024732/1-457  -----FSUMFTYMERNN
ENSAPHP00000031716/1-457  -----
ENSETAPO00000028684/1-457  -----MAPLELCRQGLLLLELLGSEPL
ENSETAPO00000037286/1-457  -----KUGLLLLLELLGSEPL
ENSETAPO00000037960/1-457  -----IFEL--MLCPFAVGAGQAGLNS
ENSCAFP00000015064/1-457  -----MKFLV--LAALLTVAARAEGLSP
ENSCIMP00000004699/1-457  -----MKISTTYNSINIIISLIASNTTQAFSSRL
ENSCIMP00000013364/1-457  -----MSFLTAPFLCLUCSIVLTSAEERK
ENSCIMP00000013369/1-457  -----MERVLEFAALLILIVGLSQAMANTVFSIPEDRHVSRVKEVMSTDKFYRNTSUSFMRITVLR
ENSDARP00000016680/1-457  ----------TULATL
ENSGALP00000011693/1-457  -----MKTLG--QLFLLSVVAIAAISL
ENSGALP00000011694/1-457  ----------SP
ENSGALP00000011695/1-457  -----MKTLG--QLFLLSVVAIAAISL
ENSMODP00000003899/1-457  -----MMLIL--LAALLAVCTVSDARP
ENSMUSP00000031495/1-457  -----MKLLL--LAALLTVAARAEGLSP
ENSP000000312286/1-457   -----MKLLV--LAALLTVAARAEGLSP
  
```

Quality/1-457

done

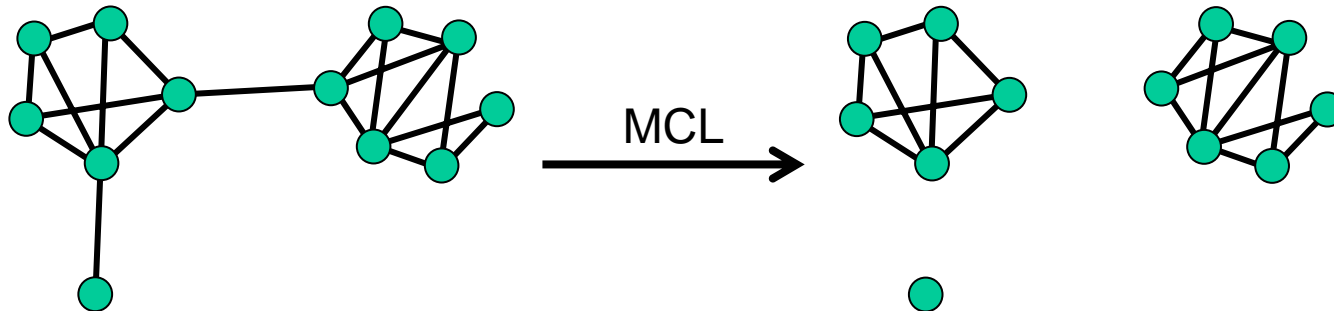
Java Applet Window

source:Uniprot/SPTREMBL;Acc:G96IX9

source:Uniprot/SWISSPROT;Acc:Q86VFX

# Clustering with MCL

- MCL for Markov CLustering algorithm, based on flow simulation in graphs (<http://micans.org/mcl/>)
- Keeps into the same graph/cluster only very well inter-connected nodes/protein



- Allows rapid and accurate detection of protein families on large-scale.
- Automatic description and clustalw multiple alignment applied on each cluster

## For each cluster

- **We store**
  - Description and score
  - Multiple alignment
- **Future extensions**
  - Improving descriptions
  - Multiple alignment assessment
  - Build phylogeny on each cluster
    - Using the multiple alignment
    - Using dS values (mainly inside mammals)
    - Extend paralogous prediction

# Para/Orthologue predictions

- Homology system (BRH/RHS)
  - **Pairwise** species analysis
  - Conservative / missed predictions
- Family system
  - **MCL Clustering** all species (+UniProt)
  - No ranking of relationships within cluster
  - Liberal / over predict
- △ **Tree** based analysis to get best of both worlds

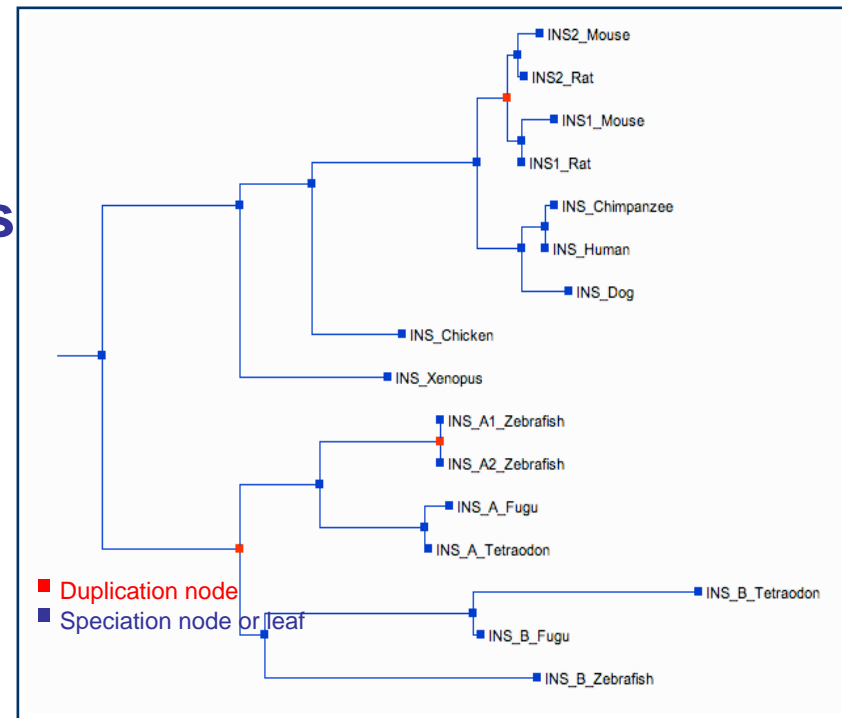
# Molecular Phylogenetics

- Protein sequences in different species, both:
  - Provide information about the history of evolution
  - Reconstruct evolution
- We are after an alignment that equally reflects all species:
  - Modeling the branching processes by comparing gene and species trees (*tree reconciliation*)

# Phylogenies

Revealing the evolutionary history that has led to the organisms at the current stage.

- Leaves are real genomes
- Internal nodes are ancestors

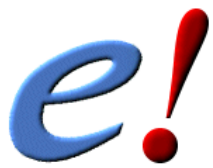


e!

Ensembl

# Outlook

- OrthoView
- Displaying alignments both from whole genome alignments and on orthologues
- Phylogentic trees



Ensembl

# Acknowledgements

- **Abel Ureta-Vidal**
- **Javier Herrero**
- **Kathryn Beal**
- **Albert Vilella**
  
- **Ensembl team**