

# Automatic retrieval of orthologs and paralogs in databases of gene families

Laurent Duret, Simon Penel, Jean-François Dufayard,  
Julien Grassot, Guy Perrière and Manolo Gouy

Pôle BioInformatique Lyonnais  
CNRS - Université Lyon 1  
INRIA Groupe Helix

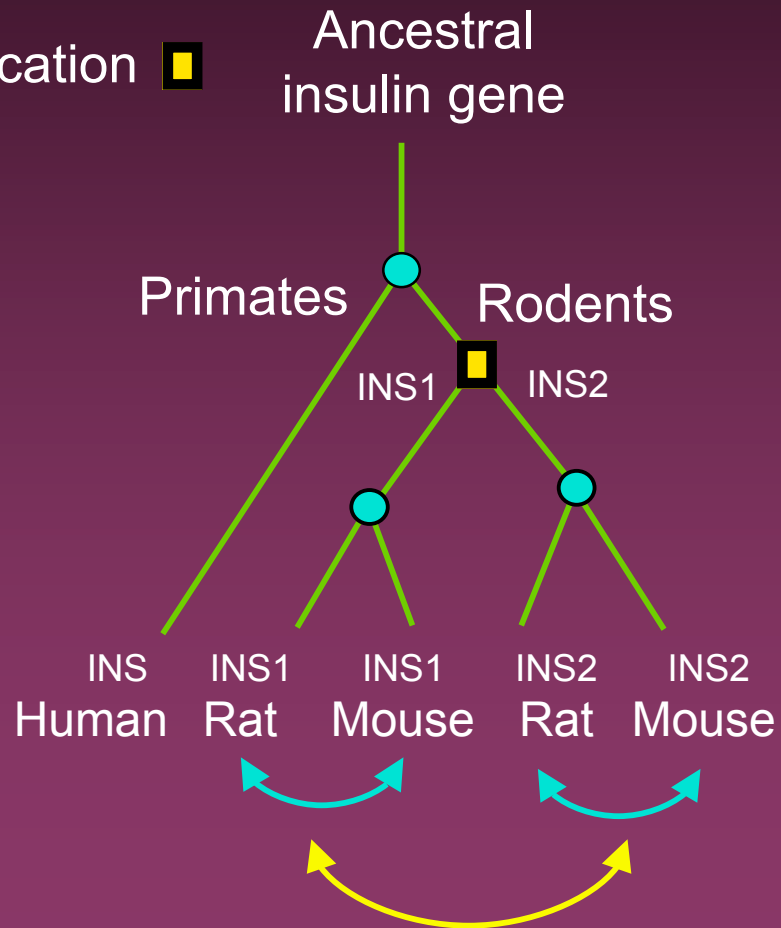
# Comparative genomics

- ✓ Functional genomics:
  - λ Prediction of gene function, protein structure
  - λ Identification of functional constraints
  - λ Identification of regulatory elements
- ✓ Molecular evolution studies:
  - λ Search for horizontal transfers
  - λ Species-specific metabolic pathways
  - λ Ancestral genome content
  - λ Chromosomal rearrangements
  - λ Gene, genome duplication and acquisition of novel functions
  - λ ...

# Orthology/Paralogy

Speciation ●

Duplication ■



Homology: two genes are homologous if they share a common ancestor

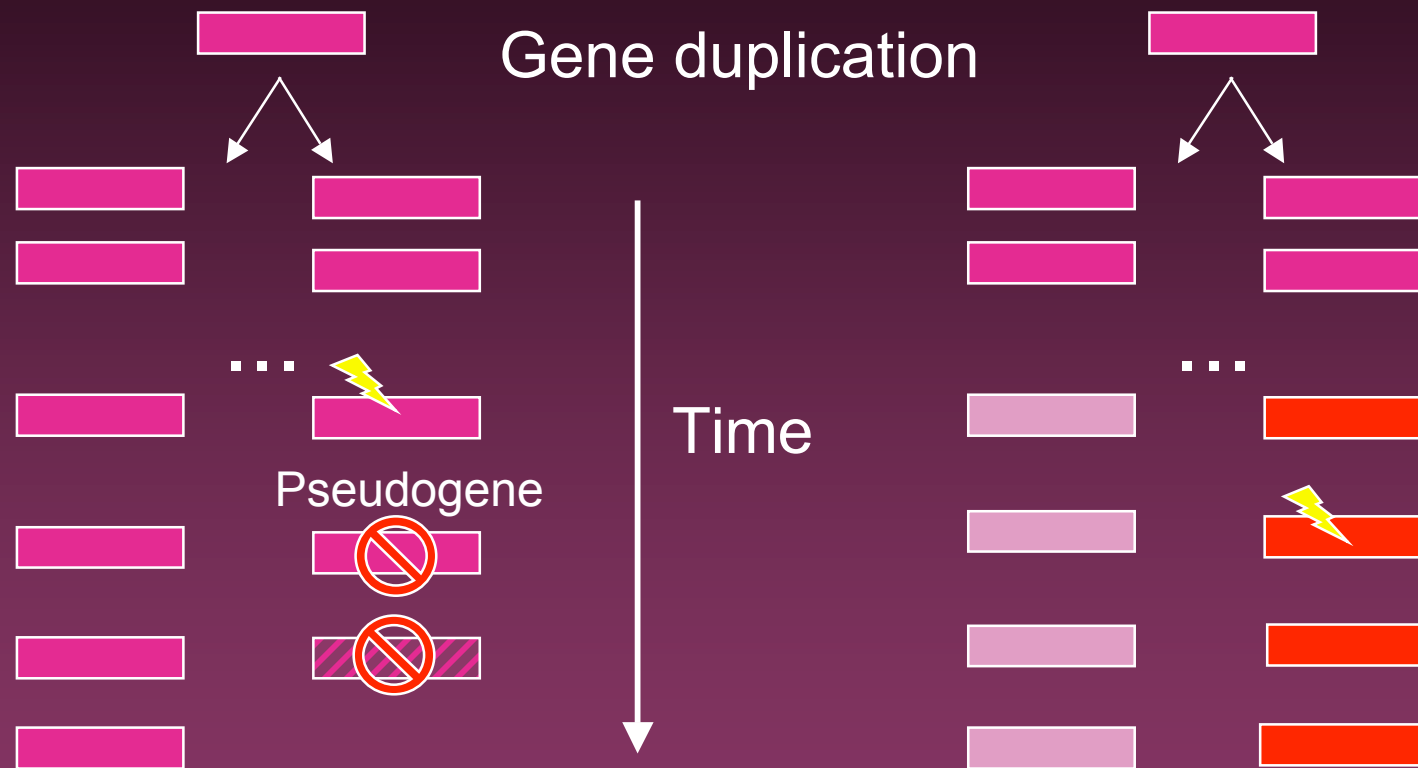
↔ Orthologs: homologs that have diverged after a speciation

↔ Paralogs: homologs that have diverged after a duplication

# Why is it important ?

- ✓ Distinguishing orthologs and paralogs is essential for:
  - λ Phylogeny: inference of the species tree from the gene tree
  - λ Comparative mapping, inference of genome rearrangements
  - λ Prediction of function by homology:
    - duplications promote the evolution of gene function
      - neo-functionalization
      - sub-functionalization

# Gene duplication and evolution of function



Ancient paralogs → Specific function

*e.g.* expression pattern, subcellular localisation, biochemical activity, ...

# Phylogenomic approach for function prediction

1) Identify homologs

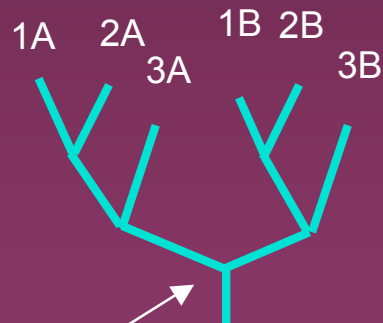
2A 2B  
1A 1B  
3B 3A

Species: 1, 2, 3

2) Align sequences

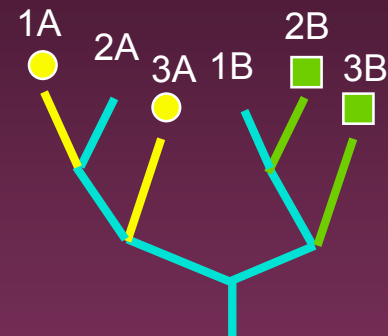


3) Compute phylogenetic tree

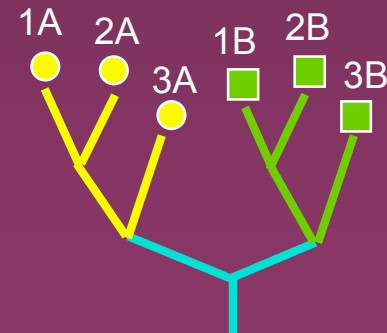


gene duplication

4) Place known functions in the tree



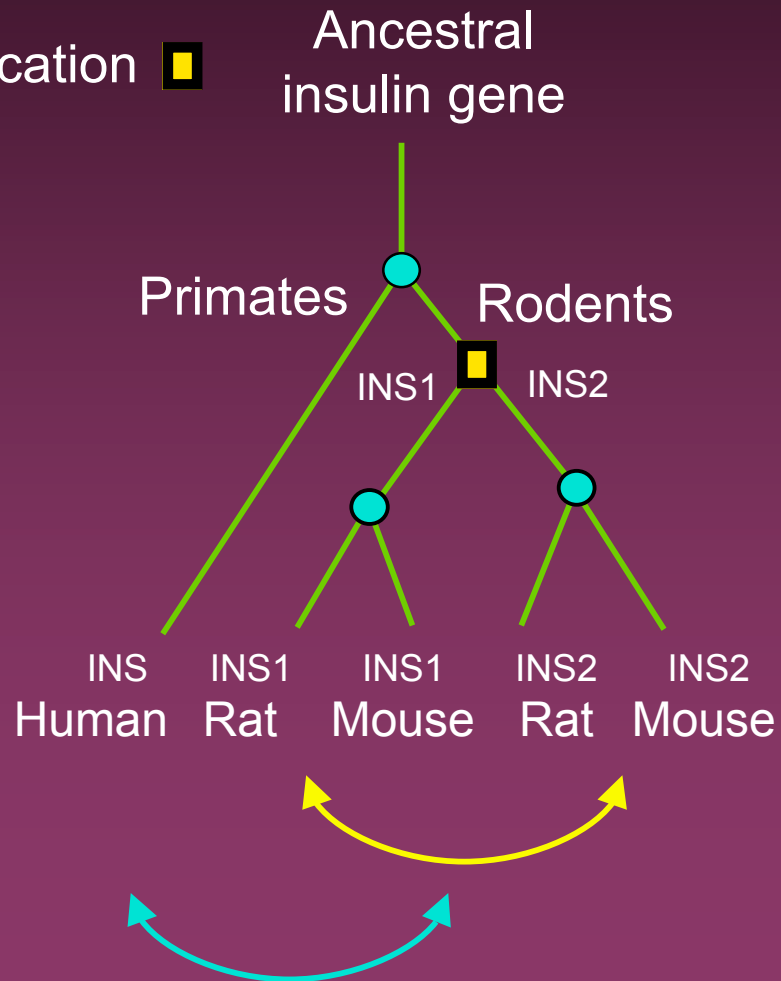
5) Infer the likely function of other genes



# Ortholog $\neq$ Functional equivalent !!

Speciation ●

Duplication ■



Orthology: not necessarily one-to-one relationship (one-to-many or many-to-many)

*e.g.:* the human *INS* gene has two orthologs in rodents (*Ins1* and *Ins2*)

The rodent *Ins1* gene is more closely related to its paralog *Ins2* than to its human ortholog *INS*.

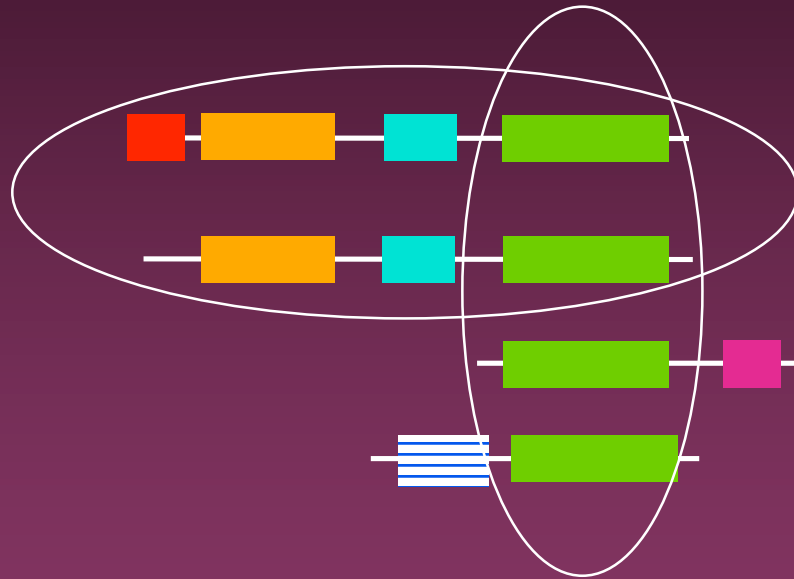
# Databases of homologous genes

- v Goal :
  - λ Provide a simple access to data required for phylogenomics (alignments, trees, taxonomic data, sequence annotations)
  
- v Databases of homologous genes at PBIL:
  - λ HOVERGEN (1994): vertebrates
  - λ HOBACGEN (2000): prokaryotes =>
  - λ HOGENOM: complete genomes (EBI complete proteomes)
  - λ HOMOLENS: Ensembl complete genomes (animals)



# Domain vs. gene families

## v Modular evolution of protein genes



Gene families homologous protein domains:

- Evolution by domain shuffling (by insertion, loss, or gene duplication)
- Sequences are homologous over their entire length (or almost)

# Databases for comparative genomics

- Databases of homologous protein domains
  - λ PROSITE
  - λ PFAM
  - λ PRODOM
  - λ ...
  - λ InterPro
- √ Databases of gene families
  - λ COG
  - λ HOVERGEN, HOGENOM
  - λ ...

# Different databases for different purposes

- √ Databases of protein domains (InterPro, etc.)
  - λ Prediction of the biochemical activity of proteins:
    - λ Does this protein have a kinase catalytic site ?
    - λ Does it contain a DNA binding domain ?
    - λ ...
  - λ Prediction of protein structures
    - λ Does this protein contain a domain homologous to an already known 3D structure ?
    - λ ...

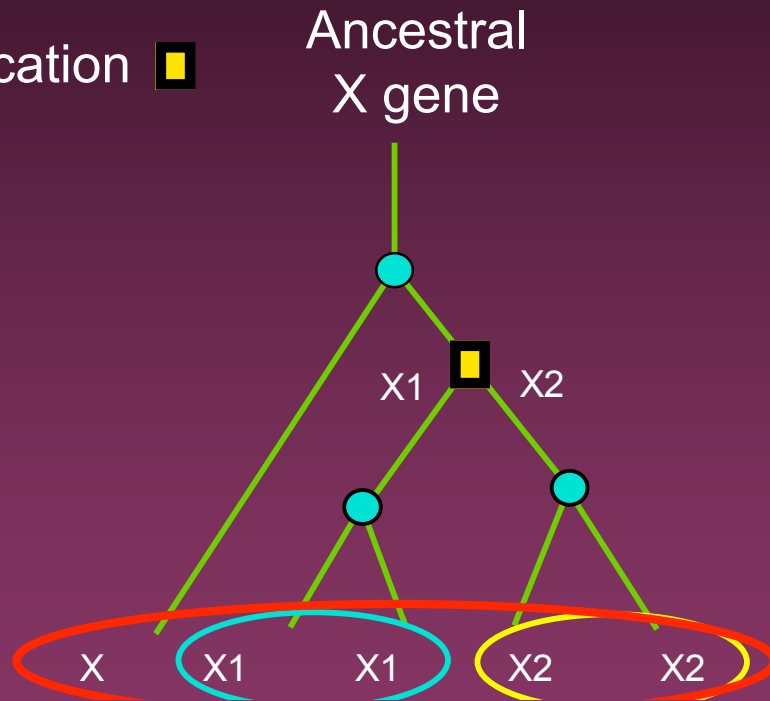
# Different databases for different purposes

- ✓ Databases of gene families (HOGENOM, etc.): identify **orthologues** or **paralogues** within a given set of taxa
  - λ Identify all orthologues between human, mouse and zebrafish
    - Prediction of gene function
    - Phylogenetics
    - Comparative mapping
  - λ Identify all paralogous genes originating from a duplication in the last common ancestor of vertebrates
    - Evolution of the function of duplicated genes
    - Analysis of genome duplications
  - λ Identify all the genes that are specific to a pathogenic strain of *E. coli*
  - λ ...

# Why not a database of *ORTHOLOGOUS* genes ?

Speciation ●

Duplication ■



Retrieve all orthologs  
between rat and mouse

Retrieve all orthologs  
between primates and  
rodents

- The clustering of homologs into groups of orthologs depends on the taxa being considered.
- No "universal" clustering of orthologous genes !

# Databases of homologous genes

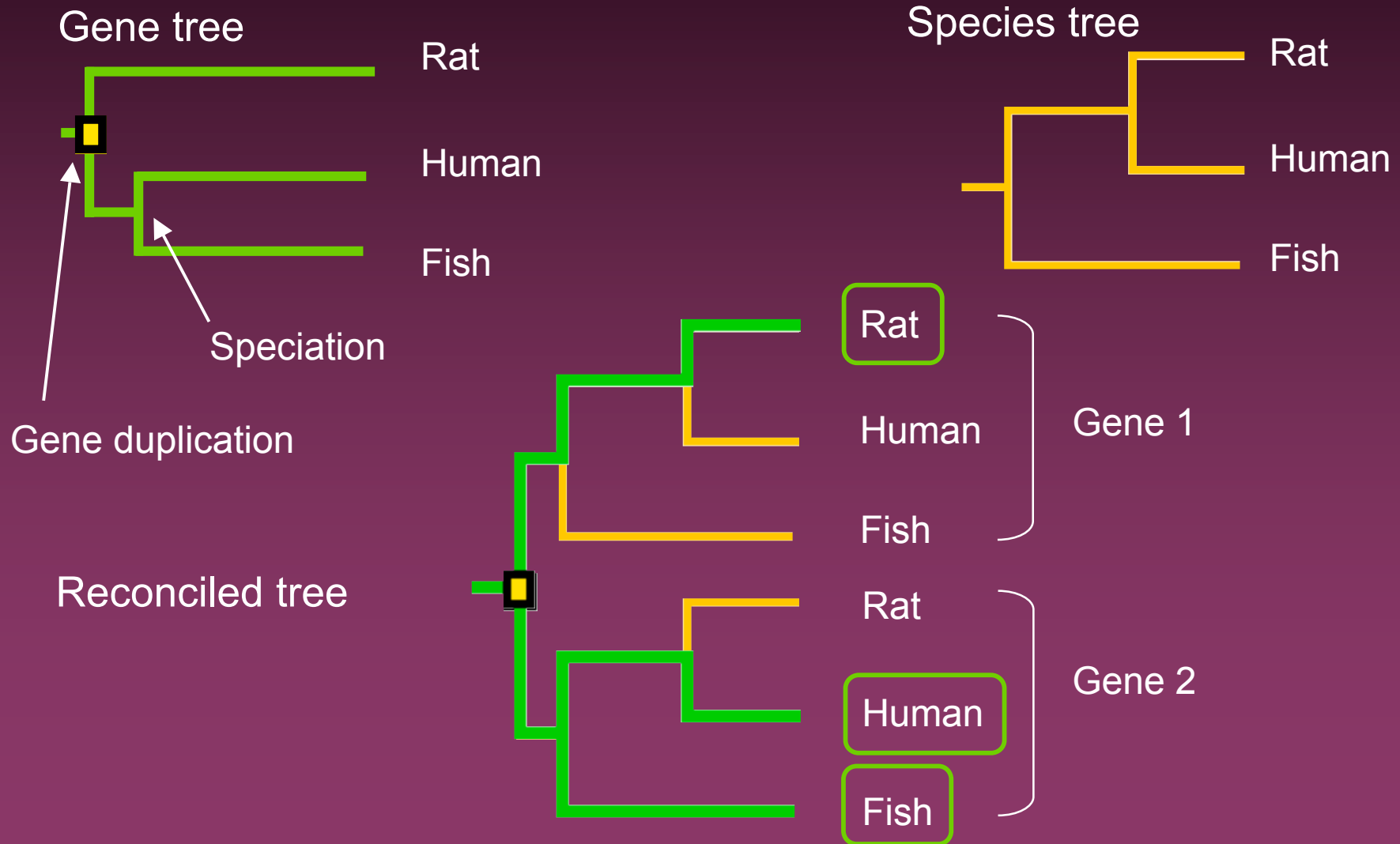
- √ Gene families include both orthologs and paralogs
- √ Development of tools for automatic retrieval of orthologs for a given set of taxa

# Orthology / Paralogy: automatic detection

---

- √ Comparison of gene and species trees: tree reconciliation
- √ Identification of speciation and duplication nodes

# Tree reconciliation





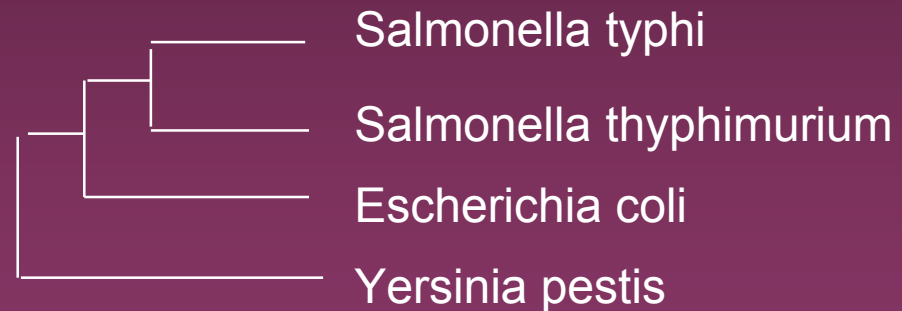
# Orthology / Paralogy: automatic detection

- ✓ Improvement of existing algorithms (Dufayard et al. *Bioinformatics* 2005):
  - λ Unresolved nodes in species or gene trees
  - λ Branch lengths
  - λ Tree rooting
- ✓ HOVERGEN, HOGENOM, HomolEns: systematic reconciliation of all phylogenetic trees
- ✓ Limitations (bacteria !): species tree, horizontal transfer

# Generalization: searching patterns in trees

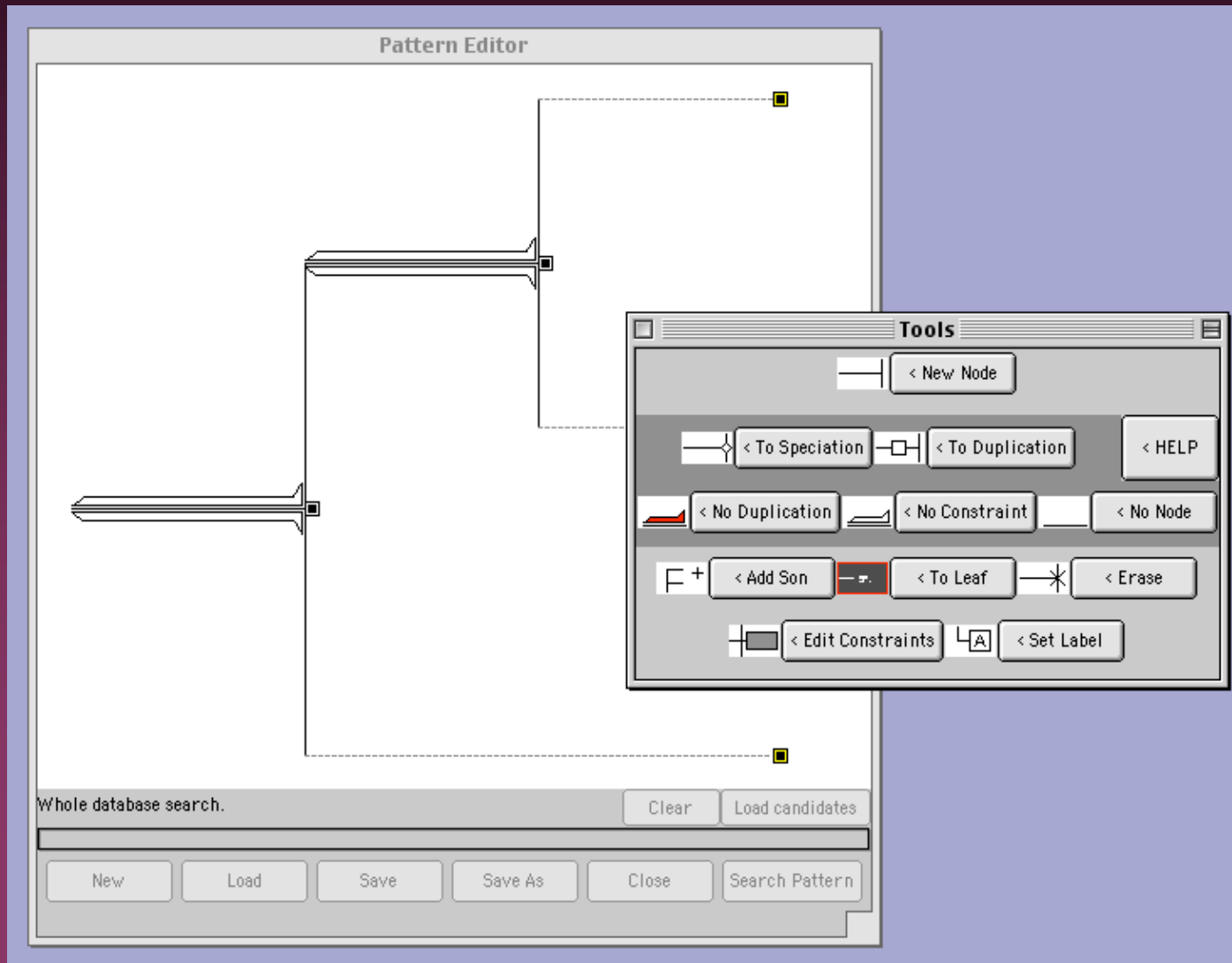
- ✓ Automatic detection of any pattern in gene trees
- ✓ Example: Search for orthologs present in completely sequenced in enterobacteriaceae

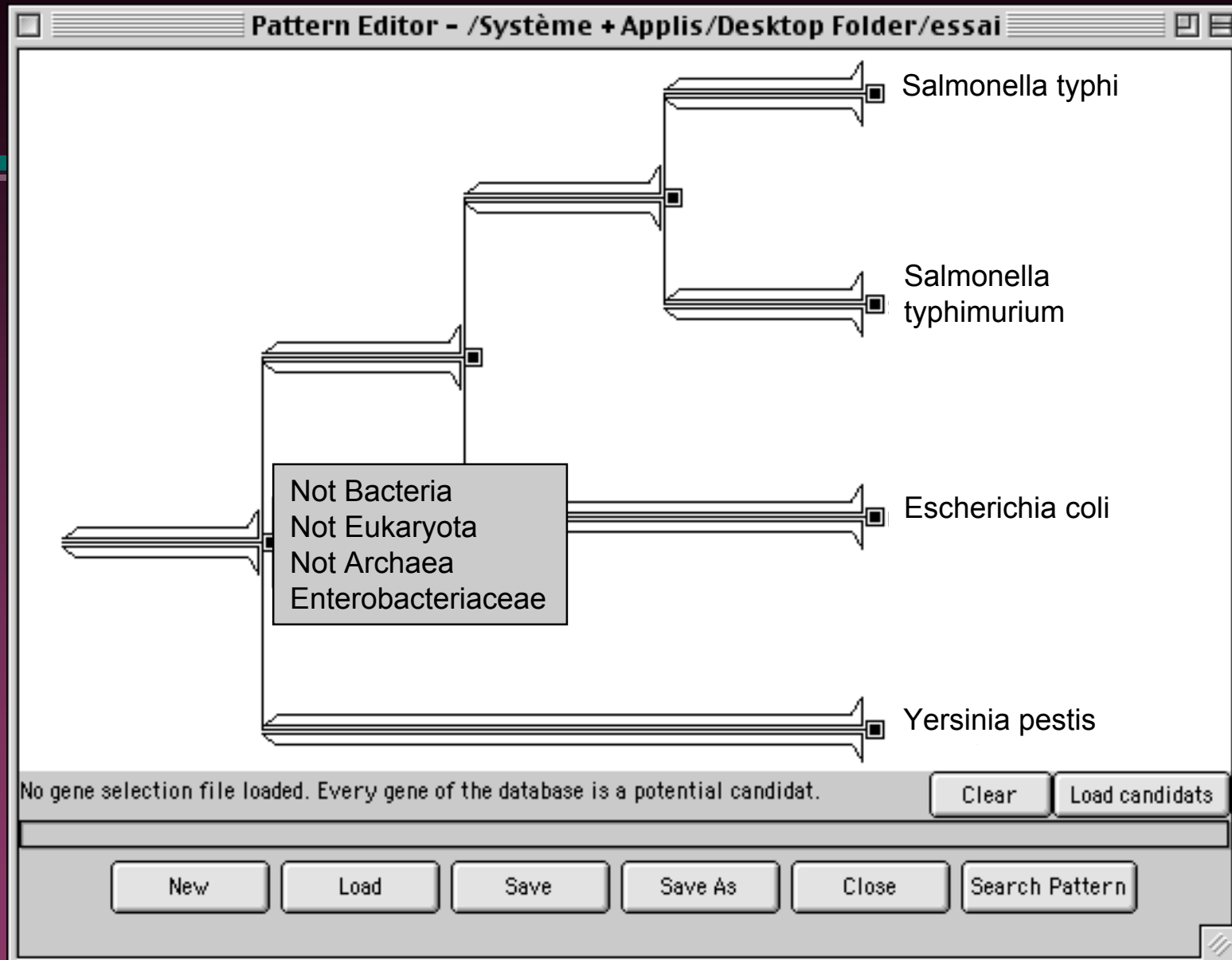
λ Known species phylogeny:



- λ Select all families containing this tree pattern
- λ Only enterobacteriaceae species in this subtree

# Tree pattern editor





Search for orthologs in enterobacteriaceae

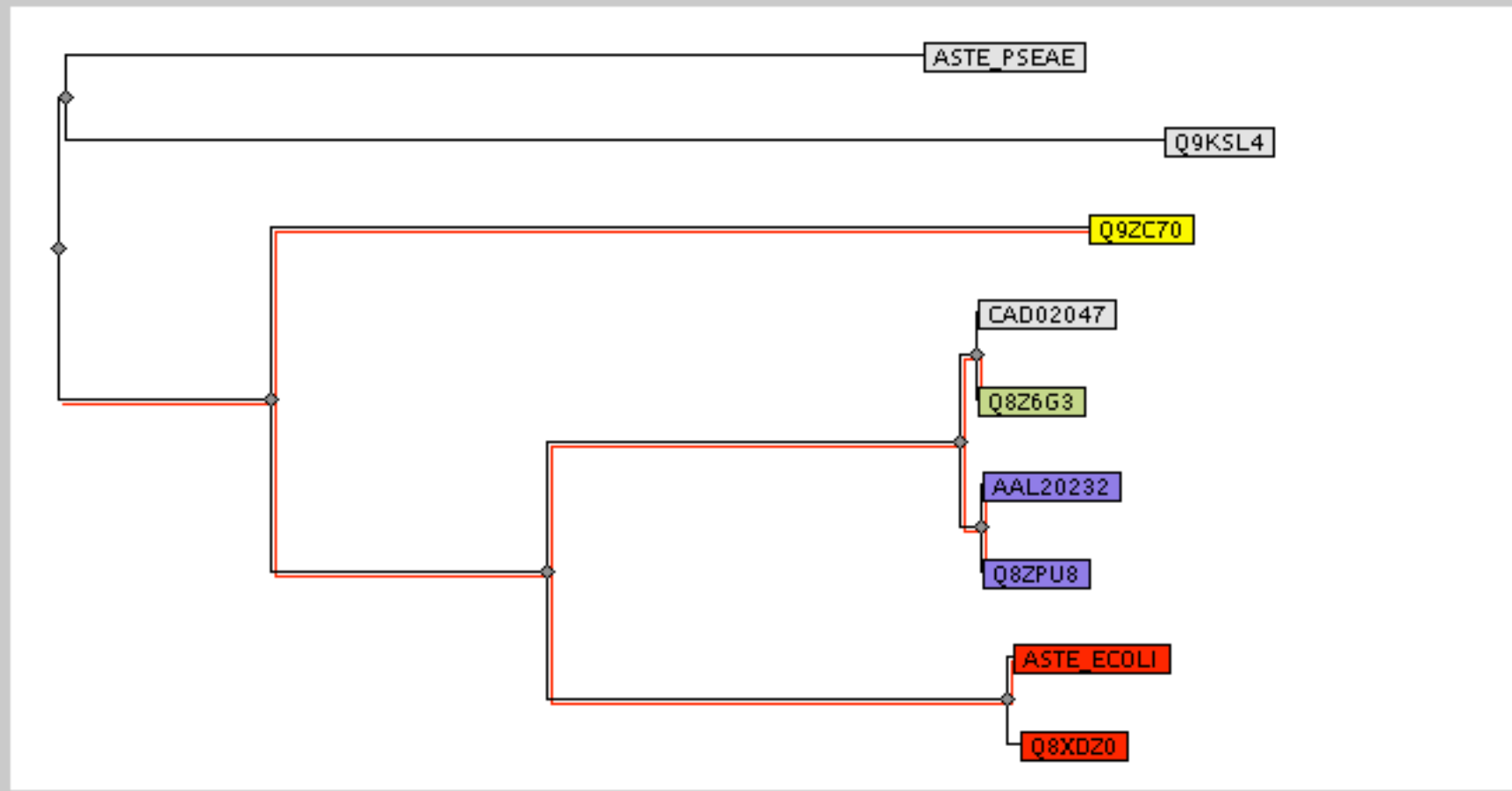
## Families

Number of selected families in HoGenom: 1116

|           |     |    |   |
|-----------|-----|----|---|
| HBG000005 | 50  | 42 | 6-PHOSPHOGLUCONATE DEHYDROGENASE FAMILY                     |
| HBG000011 | 19  | 19 | ACNB; ACONITATE HYDRASE B; ACONITATE HYDRATASE 2 TRANSMEMB  |
| HBG000012 | 49  | 42 | ACONITASE/IPM ISOMERASE FAMILY                              |
| HBG000030 | 29  | 20 | N-ACETYLMURAMIDYL-L-ALANINE AMIDASE AMIB                    |
| HBG000037 | 24  | 20 | 2-DEHYDROPANTOATE 2-REDUCTASE; KETOPANTOATE REDUCTASE PANE  |
| HBG000042 | 53  | 49 | ADENINE PHOSPHORIBOSYLTRANSFERASE 1; ADENINE PHOSPHORIBOSY  |
| HBG000043 | 40  | 39 | NAGSA DEHYDROGENASE FAMILY                                  |
| HBG000046 | 57  | 54 | 3-PHOSPHOSHIKIMATE 1-CARBOXYVINYLTRANSFERASE, CHLOROPLAST   |
| HBG000048 | 61  | 60 | AGR_C_1368P; AROF PROTEIN; CHORISMATE SYNTHASE; CHLOROPLAS  |
| HBG000052 | 82  | 55 | FUMARATE HYDRATASE C 1; FUMARATE HYDRATASE, MITOCHONDRIAL   |
| HBG000054 | 10  | 9  | ARUB PROTEIN; ORF; HYPOTHETICAL PROTEIN; SUCCINYLGARGININE  |
| HBG000056 | 9   | 9  | ARUE PROTEIN; ORF; HYPOTHETICAL PROTEIN; SUCCINYLGUTAMATE   |
| HBG000059 | 28  | 29 | KDPA FAMILY   |
| HBG000060 | 28  | 29 | CATION TRANSPORT ATPASES FAMILY (E1-E2 ATPASES). SUBFAMILY  |
| HBG000061 | 29  | 29 | KDPC FAMILY   |
| HBG000069 | 55  | 53 | ATPASE GAMMA CHAIN FAMILY                                   |
| HBG000089 | 140 | 35 | 6-PHOSPHO-BETA-GLUCOSIDASE ASCB; 6-PHOSPHO-BETA-GLUCOSIDAS  |
| HBG000091 | 40  | 29 | DETHIOBIOTIN SYNTHETASE 1; DETHIOBIOTIN SYNTHETASE 2        |
| HBG000092 | 41  | 40 | BIOTIN AND LIPOIC ACID SYNTHETASES FAMILY                   |
| HBG000093 | 11  | 11 | METHYLTRANSFERASE SUPERFAMILY                               |
| HBG000106 | 31  | 15 | CYTOCHROME B561 HOMOLOG 1; CYTOCHROME B561 HOMOLOG 2        |
| HBG000110 | 27  | 22 | PEPCASE FAMILY; PHOSPHOENOLPYRUVATE CARBOXYLASE PROTEIN; P  |
| HBG000111 | 56  | 54 | CARBAMIDYL-PHOSPHATE SYNTHASE, PYRIMIDINE-SPECIFIC, SMALL C |
| HBG000114 | 17  | 14 | TRANSCRIPTIONAL REGULATOR CBL                               |
| HBG000130 | 70  | 48 | 2-METHYLCITRATE SYNTHASE; CITRATE SYNTHASE 1                |

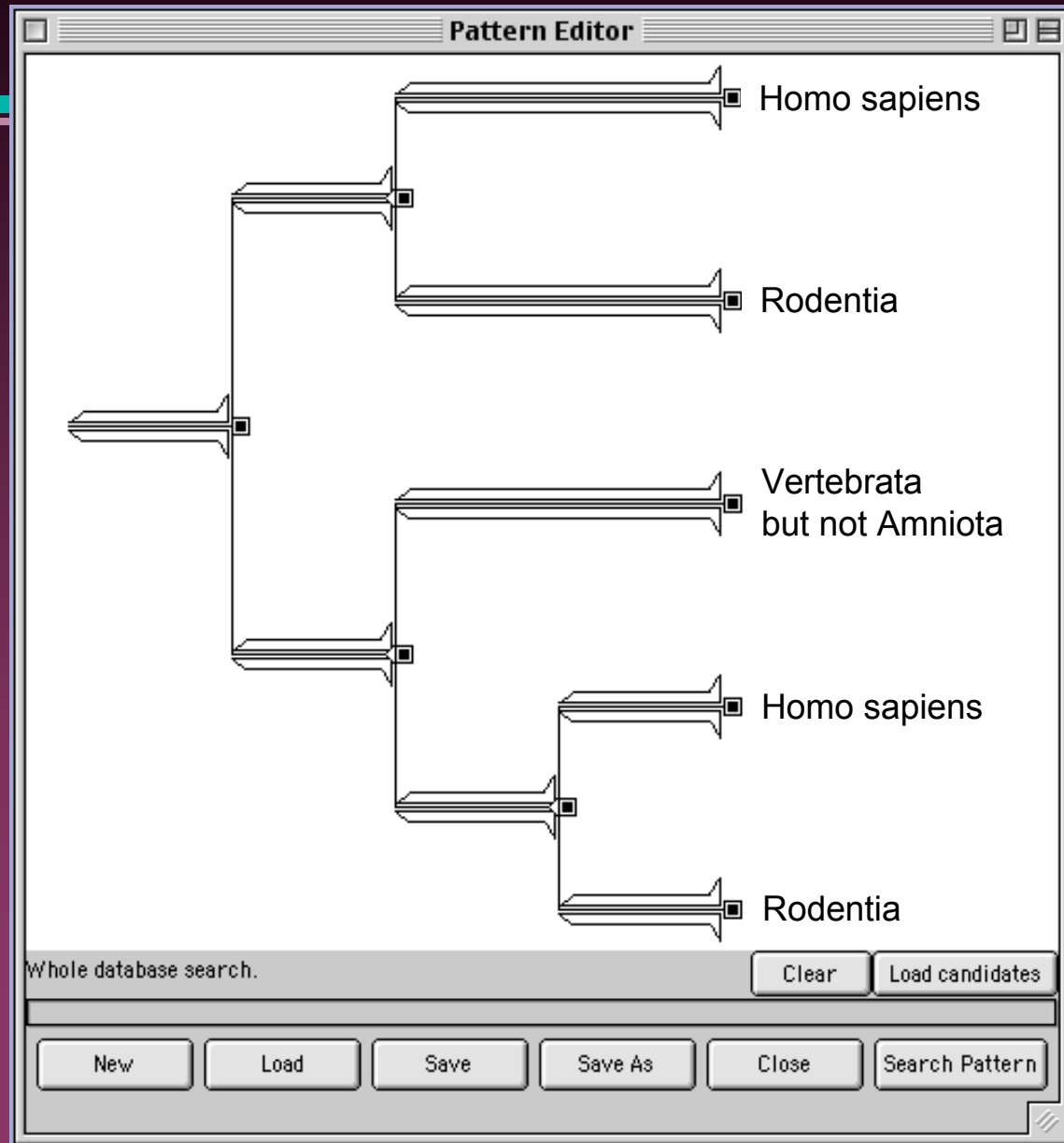
# Tree

Family: HBG000056



Select  Subtree  Outgroup  Swap nodes  Partial  Length   Medium

# Search for ancient gene duplications in vertebrates



Number of selected families in Hovergen: 886

|           |     |    |              |
|-----------|-----|----|--------------|
| HBG000002 | 58  | 10 | 14-3-3 FAMIL |
| HBG000009 | 30  | 6  | PHOSPHATASE  |
| HBG000011 | 18  | 6  | PHOSPHATASE  |
| HBG000012 | 31  | 8  | PHOSPHATASE  |
| HBG000014 | 43  | 17 | 3BETA-HSD FA |
| HBG000020 | 83  | 16 | ALDO/KETO RE |
| HBG000034 | 16  | 4  | Clathrin coa |
| HBG000036 | 13  | 6  | ADAPTOR COMP |
| HBG000038 | 8   | 4  | ADAPTOR COMP |
| HBG000040 | 33  | 6  | ADAPTOR COMP |
| HBG000041 | 32  | 7  | ADAPTOR COMP |
| HBG000051 | 52  | 17 | APP FAMILY   |
| HBG000054 | 100 | 25 | TROPOMYOSIN  |
| HBG000055 | 132 | 40 | TRANSFERRIN  |
| HBG000057 | 52  | 15 | Ubiquitin cr |
| HBG000060 | 19  | 6  | Ectonucleoti |
| HBG000062 | 49  | 30 | CU-ZN SUPERO |
| HBG000063 | 24  | 16 | OPIOIDS NEUR |
| HBG000064 | 85  | 17 | LIGAND-GATED |
| HBG000067 | 62  | 31 | Alpha enolas |
| HBG000069 | 10  | 5  | Alpha-1 cate |
| HBG000071 | 38  | 14 | Eosinophil p |

Colors

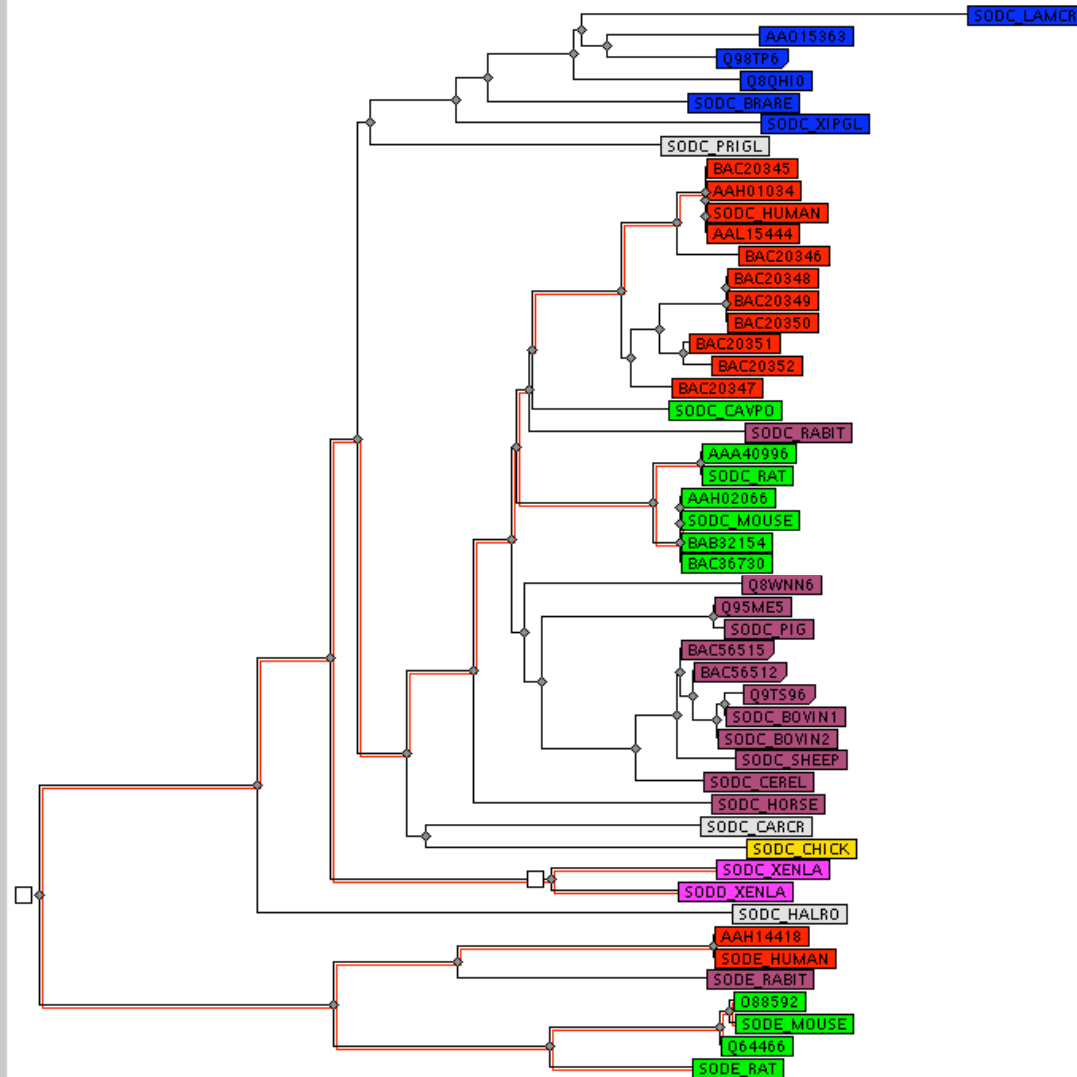
- Primates
- Rodentia
- Mammalia
- Aves
- Amphibia
- Actinopterygii

OK

Close

Tree

Family: HBG000062



Select  Subtree  Outgroup  Swap nodes  Partial  Length   Medium

Use leaf

Up

Colors

Reset

Valid

Close

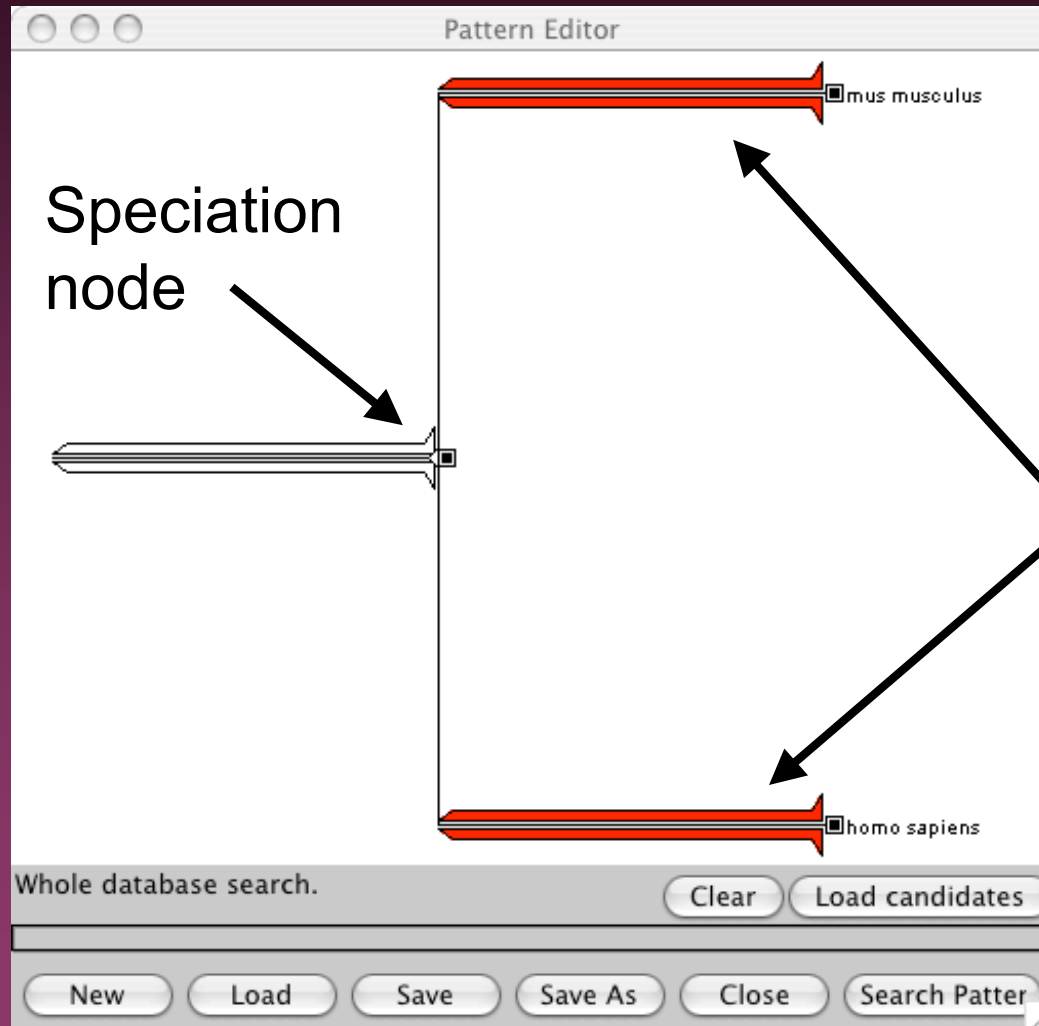
Help



# Tree Pattern vs. Reciprocal Best Hits

- √ Search all 1:1 orthologs between human and mouse
- √ Gene set: Ensembl Rel. 24
  - λ 22,077 human protein genes
  - λ 24,132 mouse protein genes
  - λ longest CDS of each gene
- √ Reciprocal Best Hits:
  - λ NCBI BLASTP, default filtering parameters,  $E < e^{-04}$

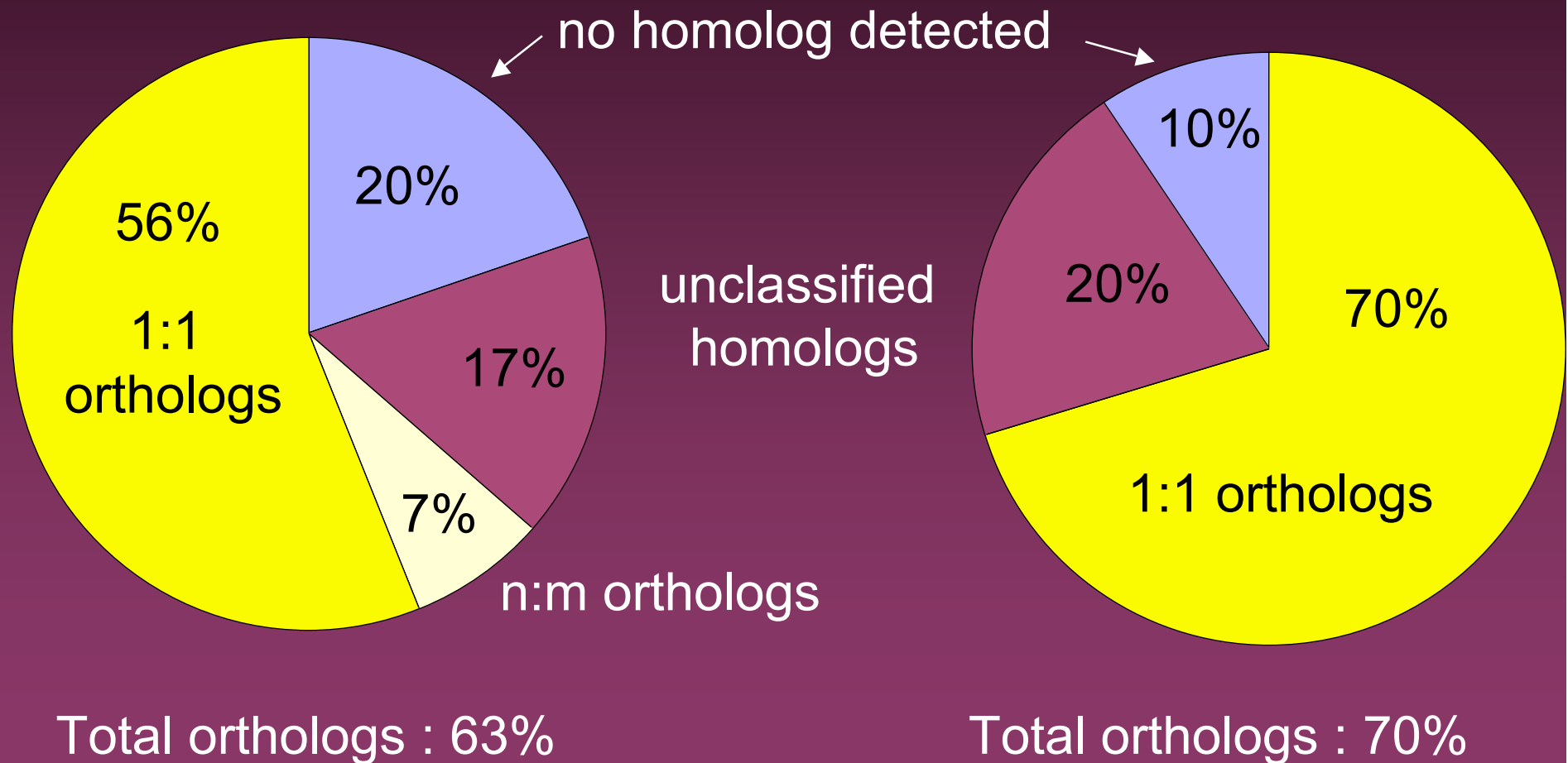
# Tree Pattern



No duplication

# Tree Pattern

# BRH

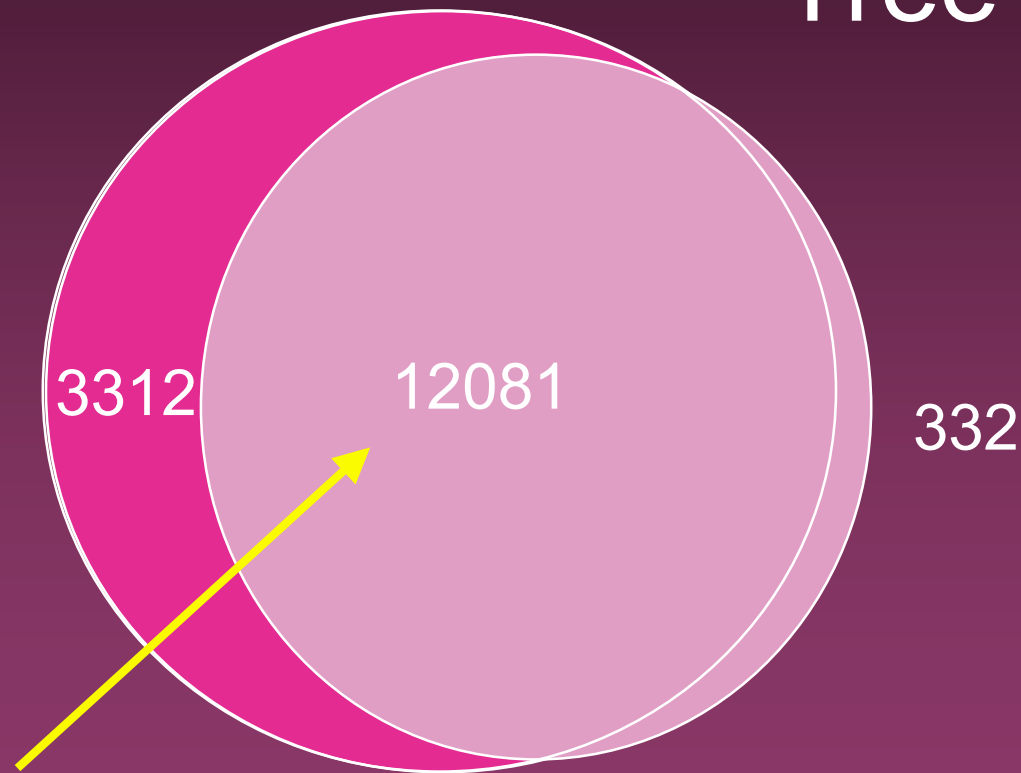


# Tree Pattern vs. Reciprocal Best Hits

BRH

1:1 orthologs

Tree Pattern

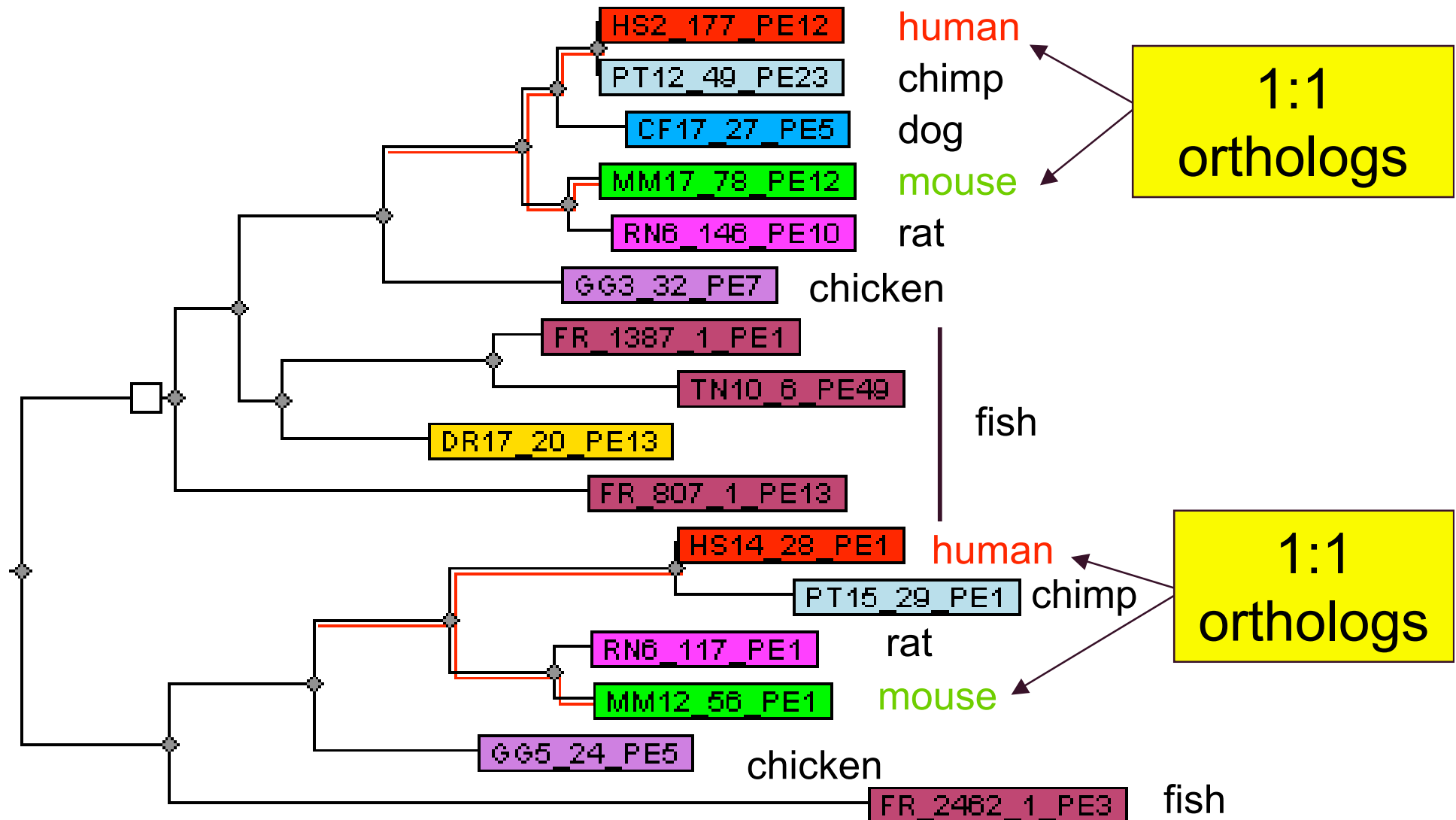


Predicted by both methods

# Predicted by Tree Pattern + BRH

- √ N=12081 predicted 1:1 orthologs
- √ Manual expertise of 50 genes (randomly sampled):
  - λ 98% true positive
  - λ 0% false positive
  - λ 2% unsure (need more expertise)
- √ True positive :
  - λ Gene tree is consistent with the species tree
  - λ Orthologs are also found in non-mammalian vertebrates

# Predicted by Tree Pattern + BRH

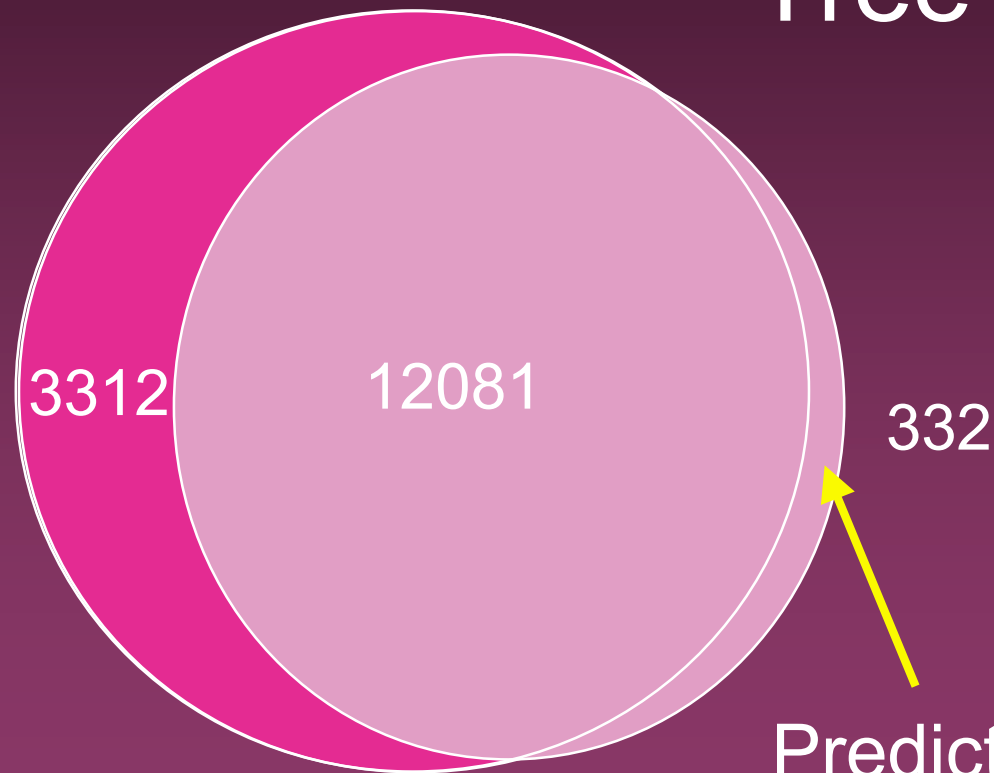


# Tree Pattern vs. Reciprocal Best Hits

BRH

1:1 orthologs

Tree Pattern



Predicted by Tree Pattern only

# Predicted by Tree Pattern but not BRH

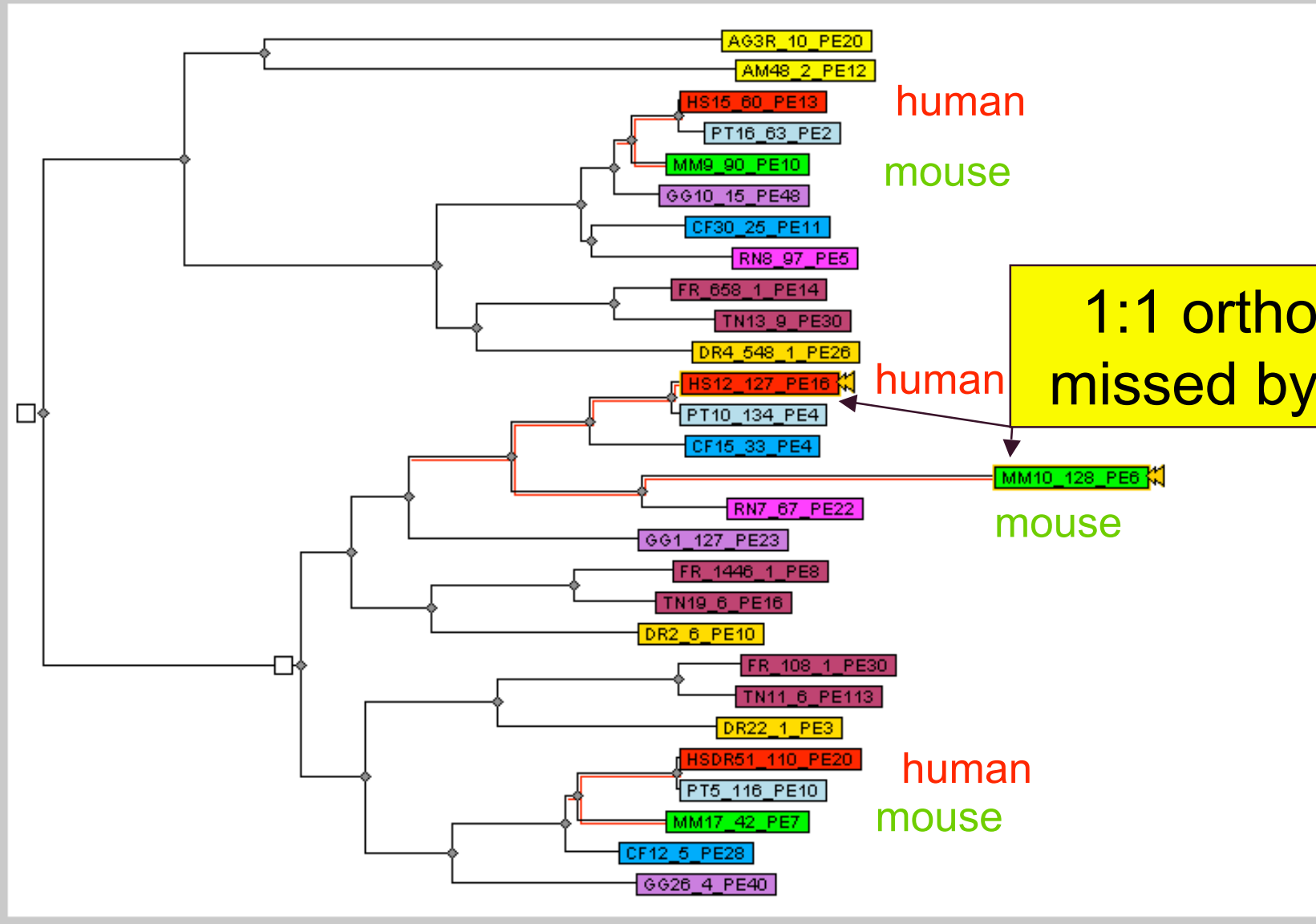
- √ N=332 predicted 1:1 orthologs
- √ Manual expertise of 47 genes:
  - λ 64% true positive
  - λ 23% false positive
  - λ 13% unsure (need more expertise)
- √ True positive = orthologs missed by BRH:
  - λ fast evolving gene in one lineage





Tree

Family: HBG060621



1:1 orthologs missed by BRH

human

mouse

human

mouse

human

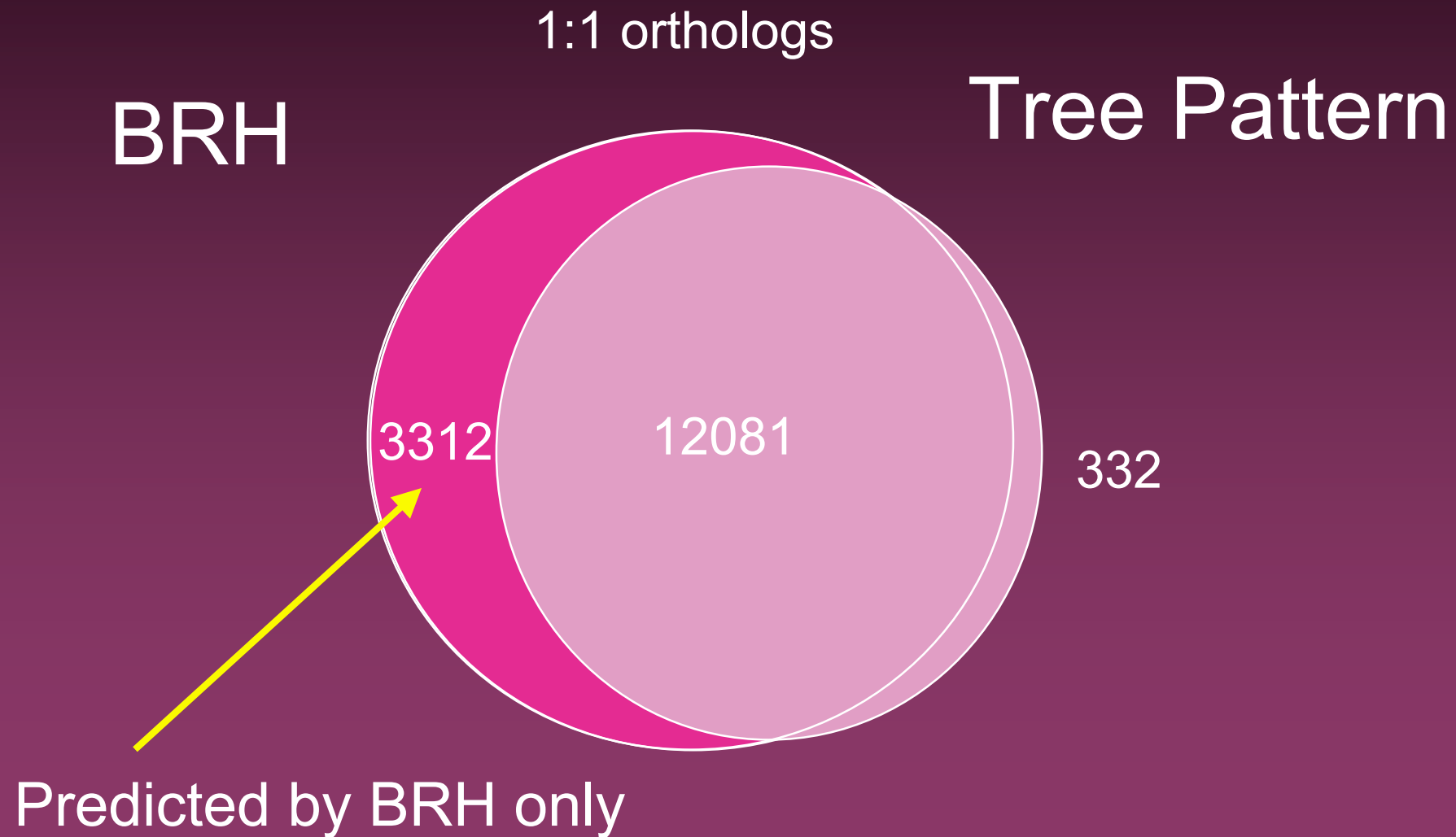
mouse

Select  Subtree  Outgroup  Swap nodes  Partial  Length   Medium

# Predicted by Tree Pattern but not BRH

- √ N=332 predicted 1:1 orthologs
- √ Manual expertise of 47 genes:
  - λ 64% true positive
  - λ 23% false positive
  - λ 13% unsure (need more expertise)
- √ True positive = orthologs missed by BRH:
  - λ fast evolving gene in one lineage
  - λ incomplete or incorrect BLAST alignment => wrong evolutionary distance
- √ False positive :
  - λ n:m orthologs

# Tree Pattern vs. Reciprocal Best Hits

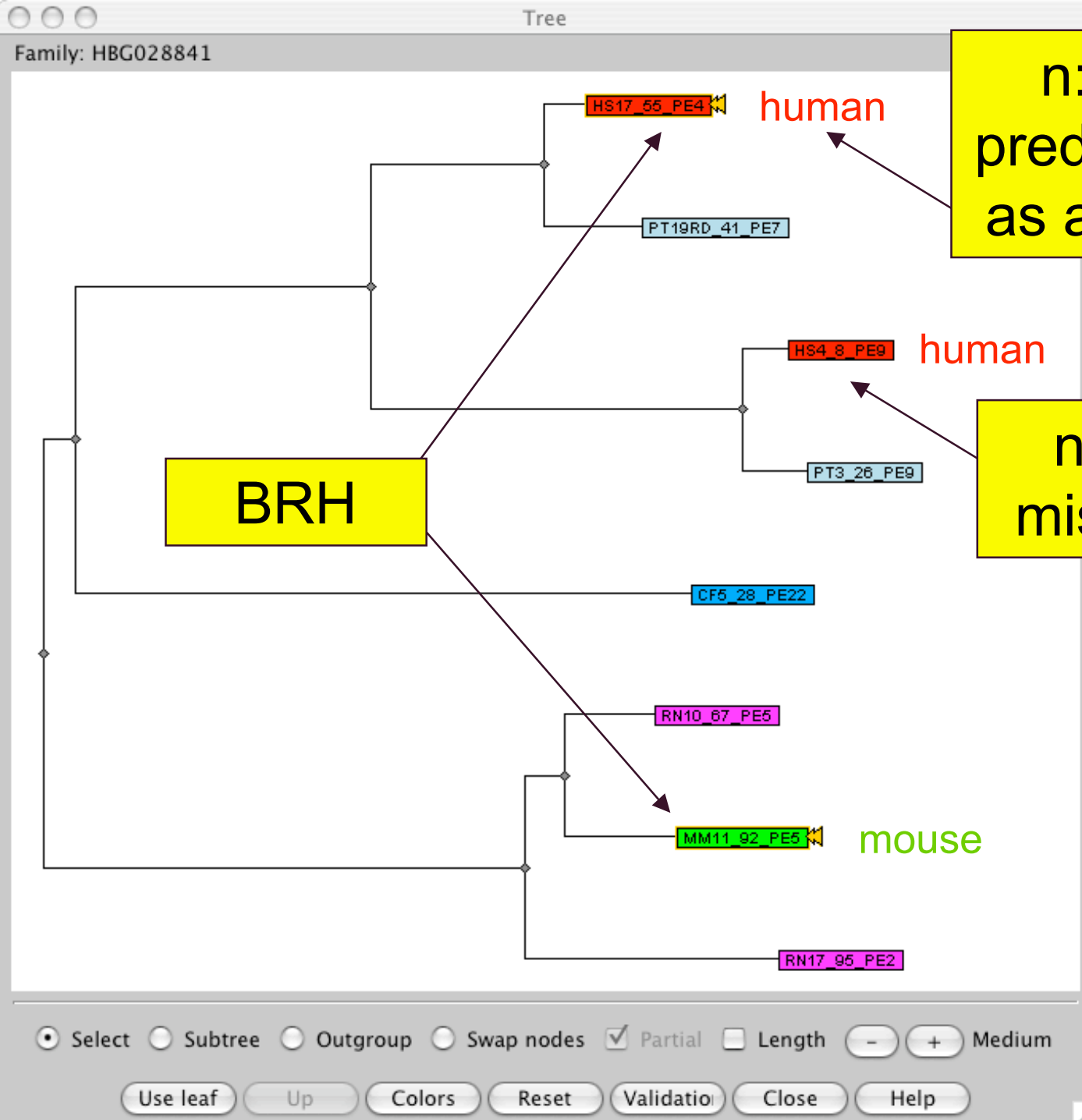


# Predicted by BRH but not by Tree Pattern

- √ N=3,312 predicted 1:1 orthologs
- √ 1,099 (33%) do not fit the clustering criteria (alignment cover 80% of protein length)
  - λ real changes in gene structure ?
  - λ incorrect/incomplete gene annotation ?
- √ 1,059 (32%) are in families with > 500 genes
  - λ no tree computed in HOGENOM => missed by Tree Pattern

# Predicted by BRH but not by Tree Pattern

- ✓ 1,154 (35%)
  - λ alignment  $\geq$  80% protein length
  - λ families with less than 500 genes  $\Rightarrow$  tree
- ✓ Manual expertise of 25 gene families:
  - λ 6/25 (24%) True positive
  - λ 19/25 (76%) False positive = n:m orthologs



n:m ortholog  
predicted by BRH  
as a 1:1 ortholog

n:m ortholog  
missed by BRH

# Tree pattern advantages: good specificity

- ✓ Explicitly based on phylogenetic trees
  - λ better estimates of evolutionary relationships than simple BLAST scores
- ✓ Does not require an exhaustive gene set
  - λ incomplete genome sequences, missing annotations, gene losses
  - λ NB: human and mouse gene sets are the most complete and accurate ones
- ✓ Distinguish 1:1 and n:m orthologs
  - λ essential for the comparison of duplicated genomes (e.g. fish vs. tetrapodes, vertebrates vs. invertebrates)
- ✓ Possibility to search for complex tree patterns
  - λ more than two species
  - λ search for gene duplications, gene losses, horizontal transfers, ...

# Tree pattern searches: Limitations

- √ Large gene families: presently, phylogenetic trees are not computed for families with  $> 500$  genes
  - λ possible improvements
- √ Classification criteria (alignment  $\geq 80\%$  length)
  - λ highly divergent orthologs, or orthologs with important differences in gene length (possibly annotation problems) are missed
- √ Quality of phylogenetic trees
  - λ quality of multiple alignments
  - λ possible improvements: GBlocks, PHYML
- √  $\Rightarrow$  problems of sensitivity for distantly related species

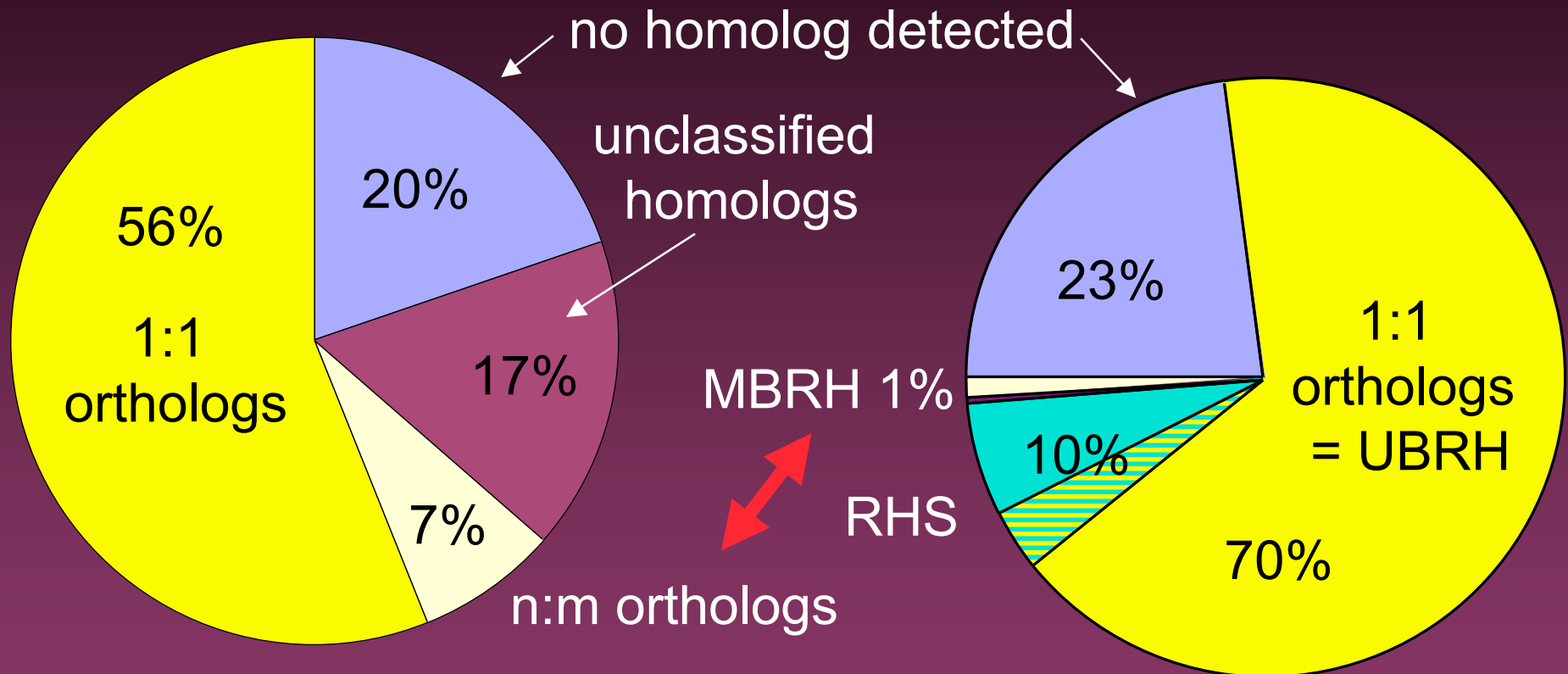


# Reciprocal Best Hits

- ✓ Easy to implement
- ✓ Good sensitivity for 1:1 orthologs (but systematically miss n:m orthologs)
- ✓ Miss some 1:1 orthologs (not many)
  - λ differences in evolutionary rates
  - λ incorrect alignment (possible improvement)
- ✓ Requires a complete gene set
  - λ problems with incomplete or not fully annotated genomes
- ✓ Problems of specificity
  - λ many n:m orthologs predicted as 1:1
- ✓ Difficult to extend to more than 2 species

# Tree Pattern

# Ensembl orthologs



Total orthologs : 63%

Total orthologs : 77%

UBRH: unique reciprocal best hit

MBRH: multiple BRH

RHS: reciprocal hit based on synteny

# Perspectives

- √ Tree Pattern : possible improvements
  - λ multiple alignment (muscle, Gblocks)
  - λ phylogenetic tree (phym1)
  - λ tree reconciliation: combine tree reconstruction and tree reconciliation
- √ Combine Tree Pattern with informations based on synteny

# People

- √ PBIL (Lyon)

- λ S. Penel

- λ G. Perrière

- λ M. Gouy

- λ J. Grassot

- λ L. Duret

- √ INRIA (Grenoble)

- λ J.F. Dufayard

- √ IN2P3 (Lyon)

- λ P. Calvat