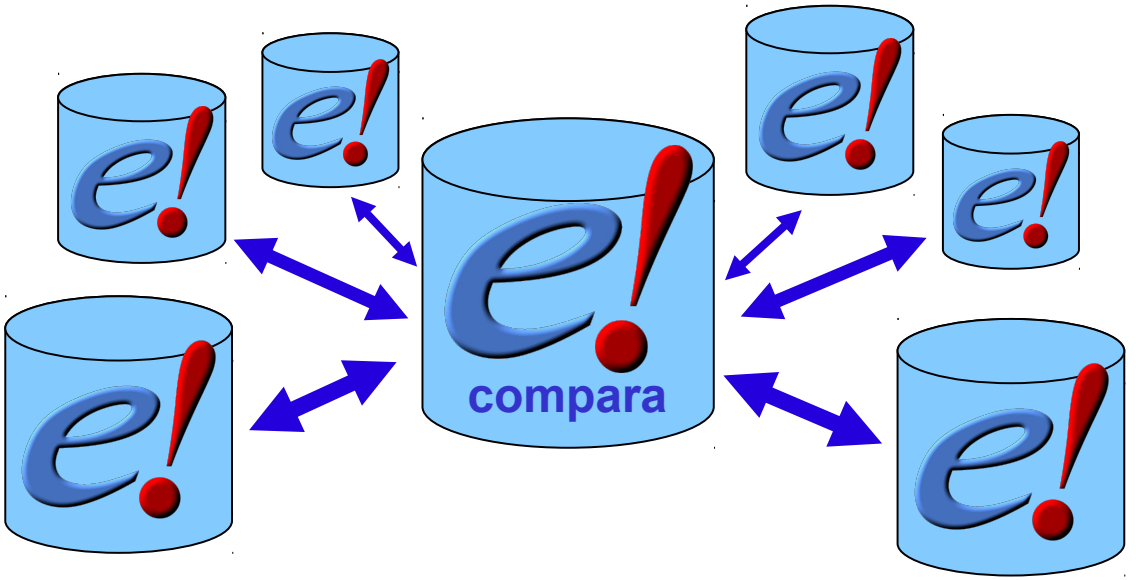
Ensembl Compara Perl API



Matthieu Muffato

API workshop - EBI

May 2013

EMBL-EBI

wellcome trust sanger institute

Outline of the course

- Introduction about Compara
  - Resources
  - API
- Base objects
  - Genes, peptides, RNAs
  - Multiple / pairwise alignments
- Data objects
  - Families
  - Gene trees
  - Homologies

Outline of the course

- Introduction about Compara
  - Resources
  - API
- Base objects
  - Genes, peptides, RNAs
  - Multiple / pairwise alignments
- Data objects
  - Families
  - Gene trees
  - Homologies

What is Ensembl Compara?

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

Access via perl API and mysql

A production system for generating that database

(not in this presentation)

The genomes in Ensembl

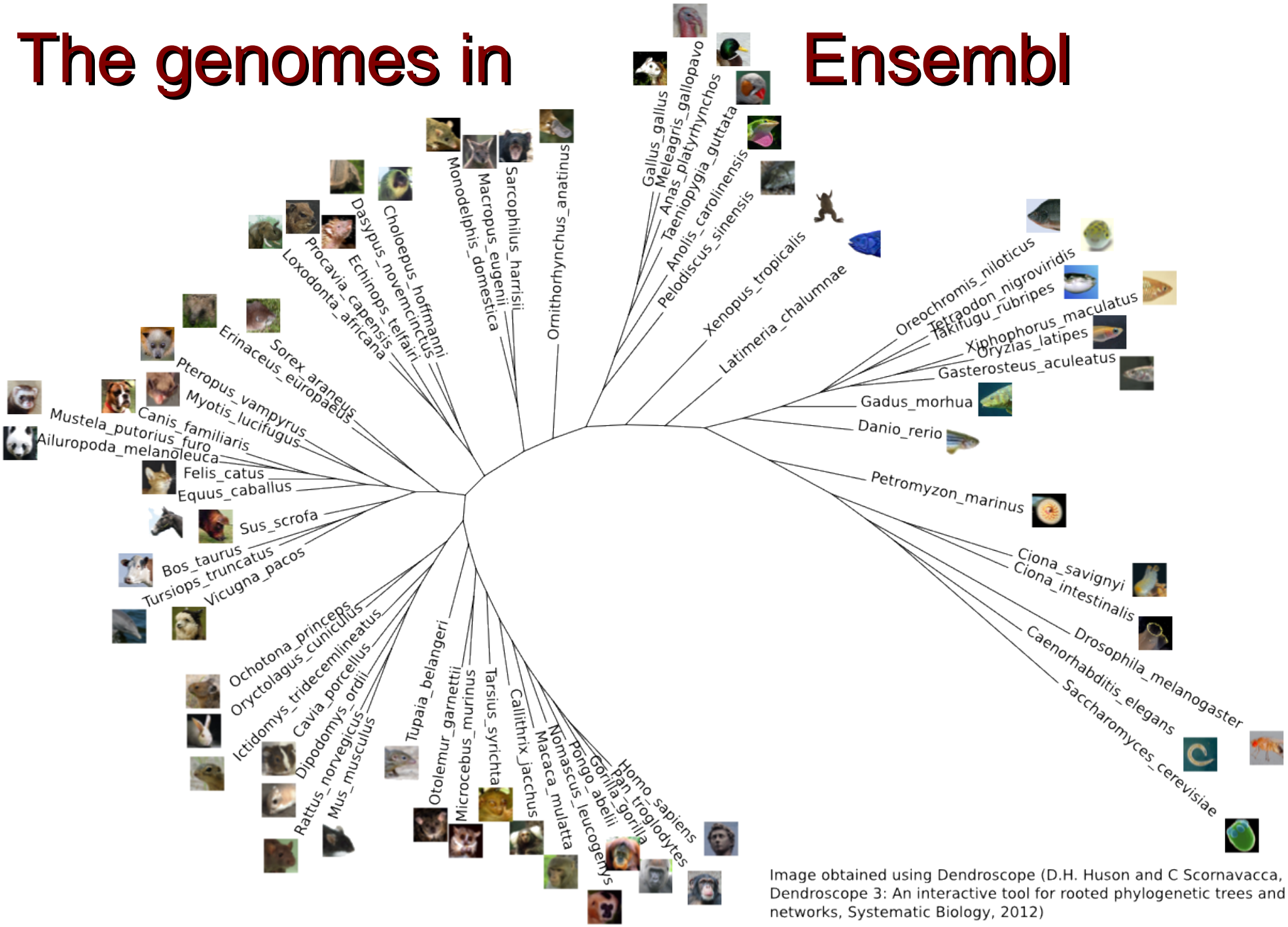


Image obtained using Dendroscope (D.H. Huson and C Scornavacca, Dendroscope 3: An interactive tool for rooted phylogenetic trees and networks, Systematic Biology, 2012)

Help & Useful documentation

- perldoc – Viewer for inline API documentation
  - `shell> perldoc Bio::EnsEMBL::Compara::GenomeDB`
  - `shell> perldoc Bio::EnsEMBL::Compara::DBSQL::MemberAdaptor`

- Online documents (website)
  - http://www.ensembl.org/info/docs/Doxygen/compara-api/index.html
  - http://www.ensembl.org/info/docs/api/compara/index.html

- CVS
  - ensembl-compara/docs/protein_schema.png
  - ensembl-compara/docs/genomic_schema.png

- ensembl-dev mailing list:
  - dev@ensembl.org
  - helpdesk@ensembl.org

Compara data

Genome level *(this afternoon)*

Whole genome alignments (pairwise and multiple)

Syntenic regions (based on pair-wise align.)

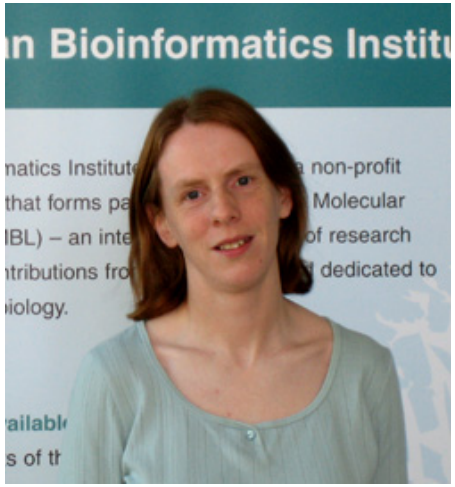
Gene level *(now !)*

Families (clusters of proteins + multiple align.)

Gene trees (proteins, non-coding RNAs)

Gene orthology / paralogy predictions

Who is in Ensembl Compara?



Kathryn Beal

Javier Hererro

Stephen Fitzgerald

Leo Gordon

+ me ! :)

Miguel Pignatelli

Outline of the course

- Introduction about Compara
  - Resources
  - API

- Base objects
  - Genes, peptides, RNAs
  - Multiple / pairwise alignments
- Data objects
  - Families
  - Gene trees
  - Homologies

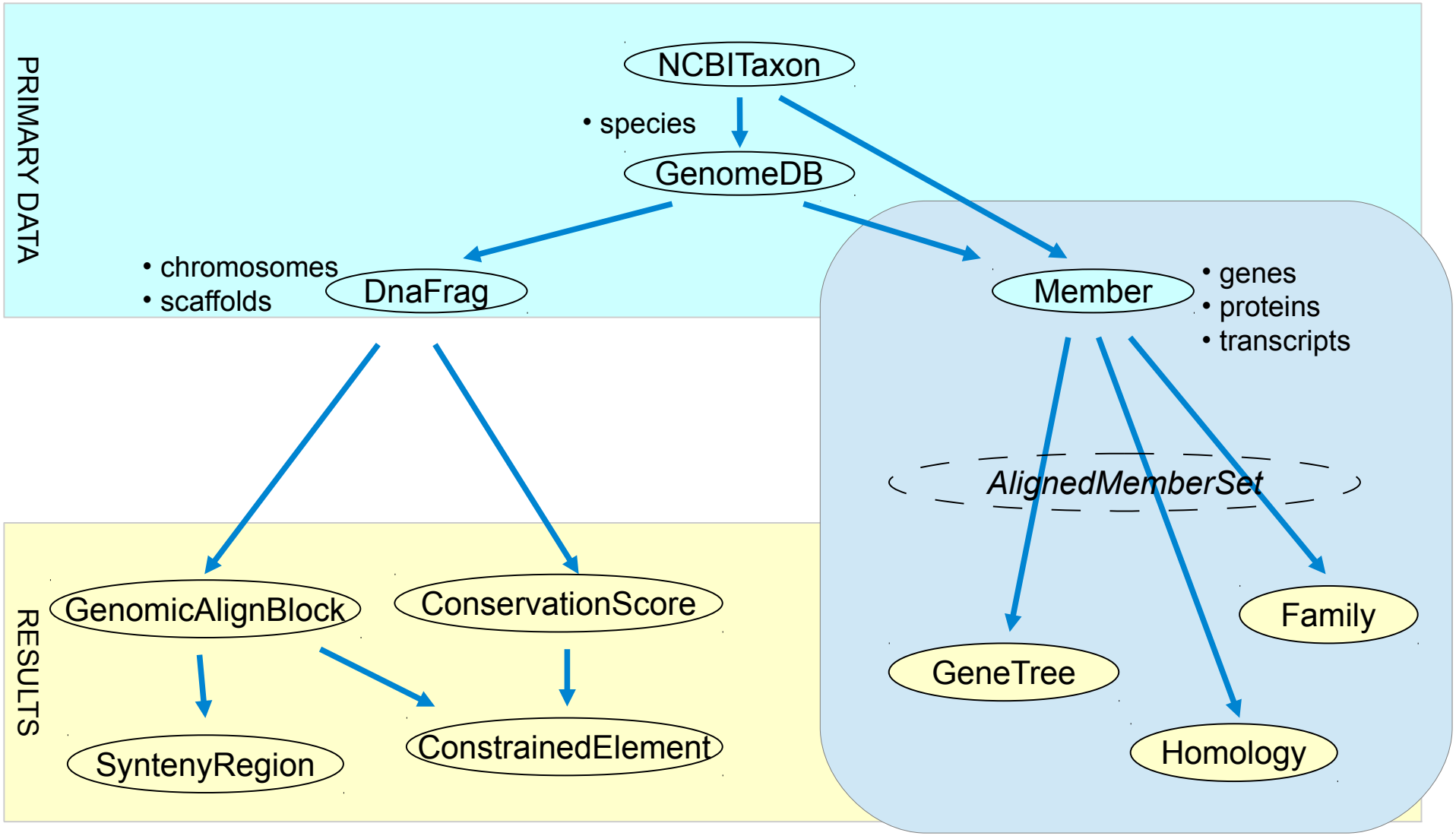
The Compara Perl API

- Written in Object-Oriented Perl
- Used to retrieve data from and store data into the Ensembl Compara database
- Links species together for Ensembl website
- Generalized to extend to non-Ensembl genomic data (Uniprot)
- Follows same ‘Data Object’ & ‘Object Adaptor’ DBAdaptor design as the other Ensembl APIs

Compara template script

```
use strict;
use Bio::EnsEMBL::Registry;
my $reg = "Bio::EnsEMBL::Registry";

# Auto-configure the registry
$reg->load_registry_from_db(
    -host => "ensembldb.ensembl.org",
    -user => "anonymous"
);


# Get the adaptor for the data type XX
# e.g. GeneTree, Family
my $xx_adaptor =  $reg->get_adaptor("Multi", "compara", "XX");


# Fetch the data objects using the adaptor
# e.g. get all the families that contain a given gene
my $all_interesting_xx = $xx_adaptor->fetch_all_by_YY();


print "All XX objects from E!Compara :\n";
foreach my $this_xx (@$all_interesting_xx) {
  # Do some stuff with the data object
  print "\t", $this_xx->stable_id, "\n";
}
```

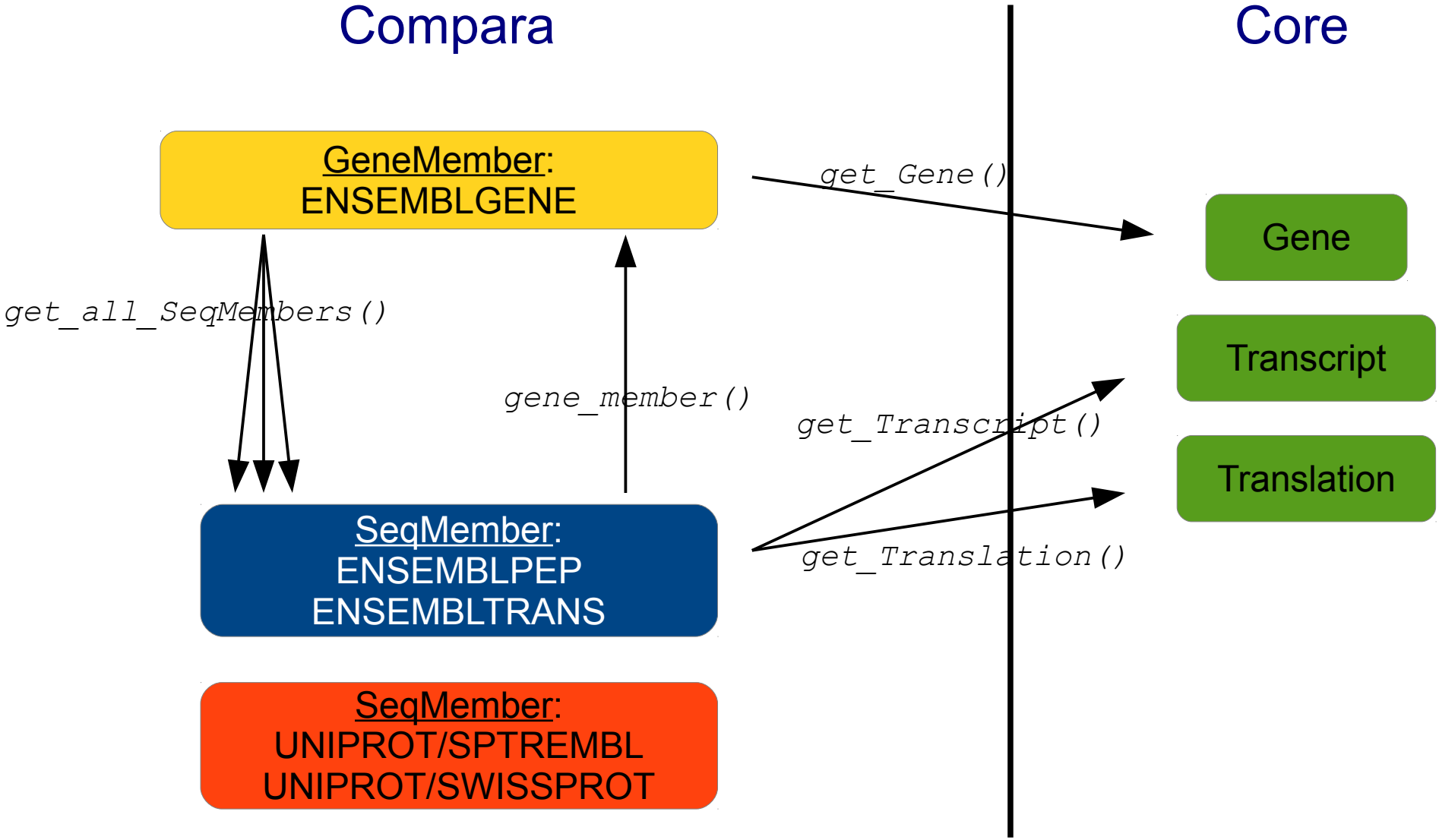
Compara object model overview



Outline of the course

- Introduction about Compara
  - Resources
  - API
- Base objects
  - Genes, peptides, RNAs
  - Multiple / pairwise alignments
- Data objects
  - Families
  - Gene trees
  - Homologies

Overview of *Member* objects



(*)Member object / (*)MemberAdaptor

- GeneMember / GeneMemberAdaptor for genes

```
ENSEMBLGENE
```

- SeqMember / SeqMemberAdaptor for RNAs and proteins

```
ENSEMBLPEP, ENSEMBLTRANS, Uniprot/SPTREMBL, Uniprot/SWISSPROT
```

- Objects share most their properties:

```
$seq_member_adaptor->fetch_by_source_stable_id(...)
$gene_member_adaptor->fetch_all_by_source_taxon(...)
```

Attributes	Methods
Stable ID	*$member->stable_id()*
Coordinates	*$member->chr_name()* *$member->chr_start()* …
Sequence (SeqMember only)	*$member->sequence()*
Function	*$member->description()*

HOWTO: get a GeneMember from a gene symbol

- Compara only references genes by their Ensembl stable ID
- From a gene symbol, you first have to use the core API to get the stable id(s)
- Gene symbols may not be unique (for instance: U6)

```
# Get the Human gene adaptor
my $hg_adaptor = $reg->get_adaptor("human", "core", "Gene");

# Get all the genes
my $all_genes = $hg_adaptor->fetch_all_by_external_name(XX);

# For each gene
foreach my $gene (@{$all_genes}) {
    do some stuff with $gene->stable_id();
}
```

Exercises - *Member*

- Print the sequence of the Member corresponding to SwissProt protein O93279

- Find the Member(s) for the human ncRNA gene(s) FAM41C

- Find and print the sequence of all the peptide Members corresponding to the human protein-coding gene(s) FRAS1

Outline of the course

- Introduction about Compara
  - Resources
  - API
- Base objects
  - Genes, peptides, RNAs
  - Multiple / pairwise alignments 
- Data objects
  - Families
  - Gene trees
  - Homologies

AlignedMemberSet object

- Base object that represents a set of members aligned together, e.g. a multiple alignment of peptides / ncRNAs
- “Applied” in gene trees, families, and homologies
- No specific adaptor

Attributes	Methods
List of members	`$aln->get_all_Members()`
Alignment (BioPerl object)	`$aln->get_SimpleAlign()`
Description (if available)	`$aln->description()`
Stable ID (if available)	`$aln->stable_id()`

HOWTO: print a BioPerl alignment

- Compara objects return alignments as BioPerl instances

```
$aln->get_SimpleAlign()
```

- BioPerl provides an AlignIO object to format the actual output in various formats (fasta, clustalw, phylip …)

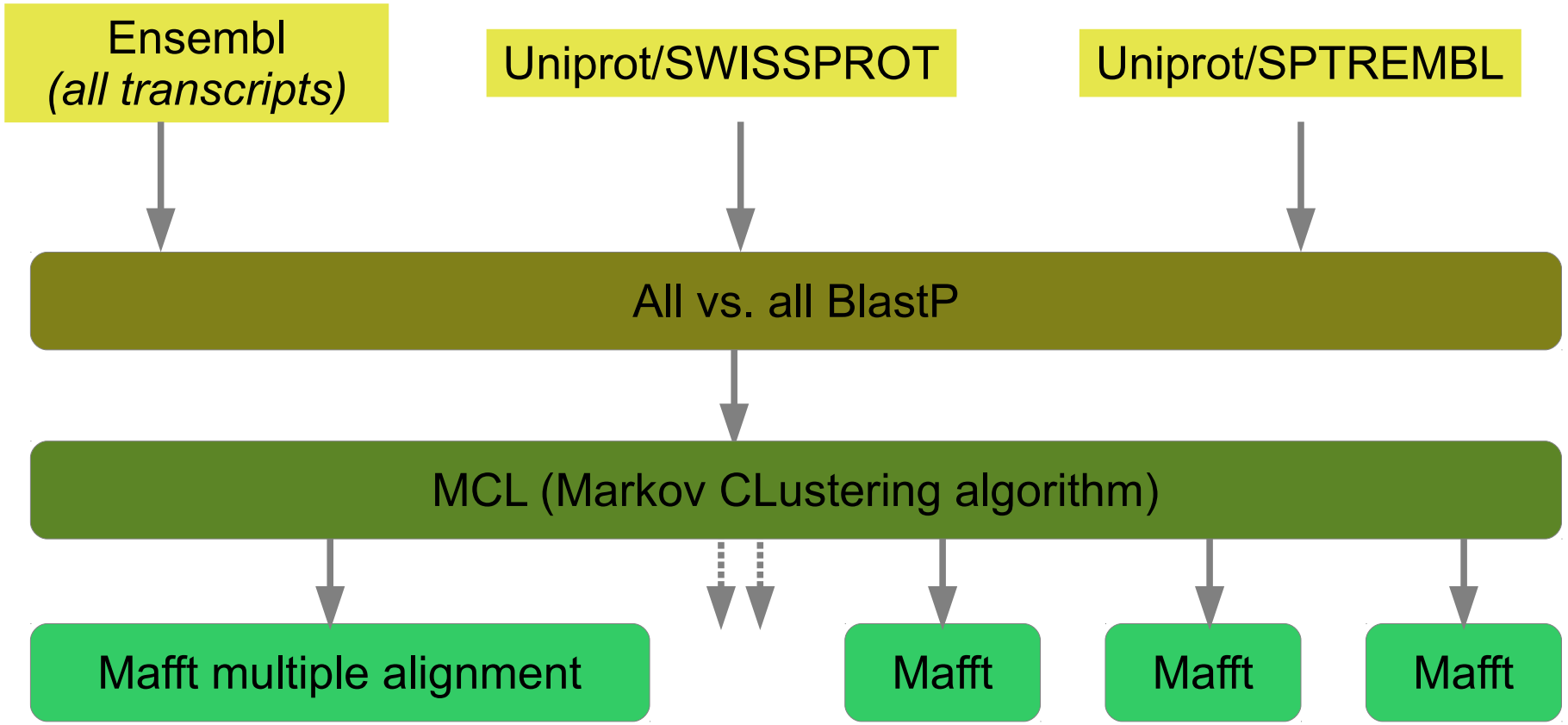
```
use Bio::AlignIO;

# Get the alignIO object from BioPerl
my $alignIO = Bio::AlignIO->newFh(-format => "fasta");

# Print the alignment
print $alignIO $aln;
```

Outline of the course

- Introduction about Compara
  - Resources
  - API
- Base objects
  - Genes, peptides, RNAs
  - Multiple / pairwise alignments
- Data objects
  - Families
  - Gene trees
  - Homologies

Families

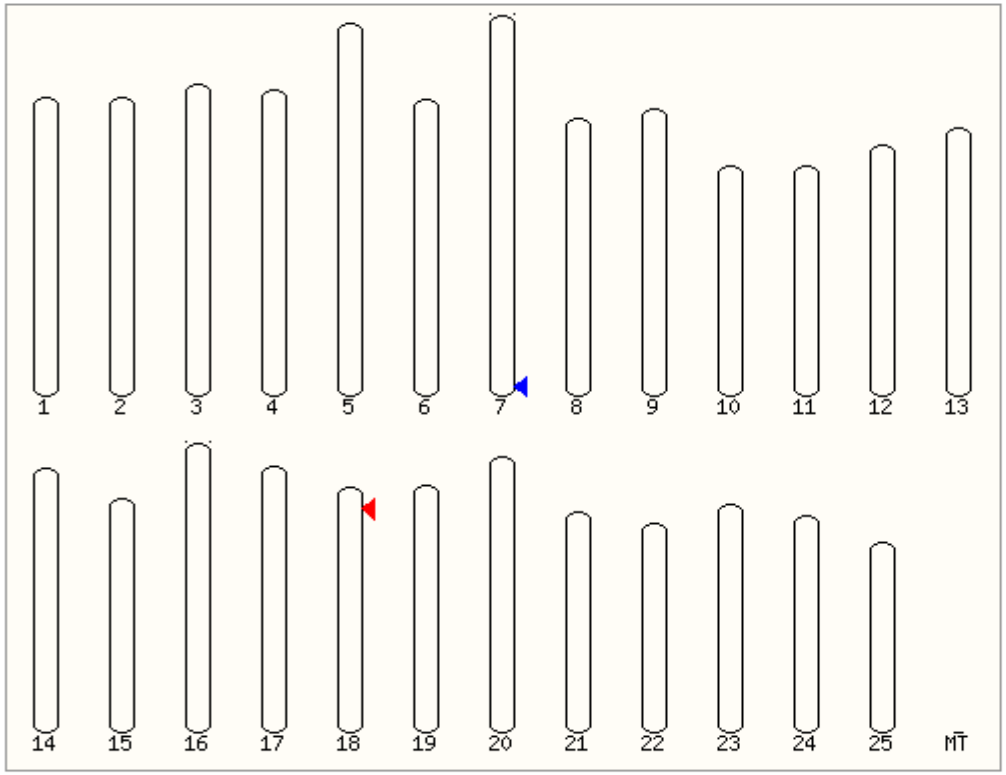
Families are clusters of similar peptides



Example on the web: ENSFM00500000271501 in Zebrafish

ZEBRAFISH genes in this family

Ensembl genes containing proteins in family ENSFM00500000271501



Gene ID and Location	Gene Name	Description(if known)
ENSDARG00000041086 Chromosome 7: 75.50m	CABZ01071177.1	Uncharacterized protein [Source:UniProtKB/TrEMBL;Acc:F1QUF1]
ENSDARG00000024771 Chromosome 18: 4.64m	slc24a5	solute carrier family 24, member 5 [Source:ZFIN;Acc:ZDB-GENE-031210-1]

Family object / *FamilyAdaptor*

- Represents a group of similar peptide members

```
$family_adaptor->fetch_all_by_Member(...)
$family_adaptor->fetch_by_stable_id(...)
```

- Alternative transcripts can belong to different families !
  Families contain both SeqMembers and GeneMembers

- (almost) the same methods as in *AlignedMemberSet*

Attributes	Methods
Alignment	*$family->get_SimpleAlign()*
Biological function	*$family->description()* *$family->description_score()*
Gene content	*$family->get_all_Members()*

Exercises - Families

- Get the multiple alignment corresponding to the family with the stable id ENSFM00250000006121

- Get the families predicted for the human gene ENSG00000139618. What do you notice ?

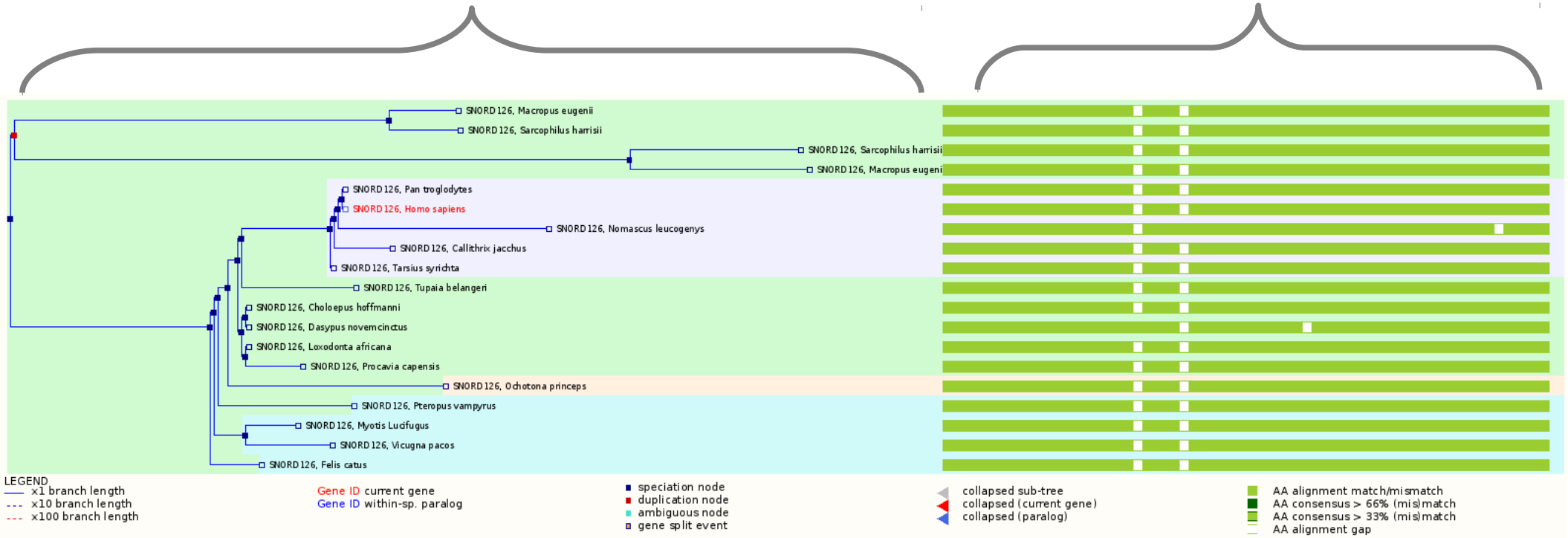
Outline of the course

- Introduction about Compara
  - Resources
  - API
- Base objects
  - Genes, peptides, RNAs
  - Multiple / pairwise alignments
- Data objects
  - Families
  - Gene trees
  - Homologies

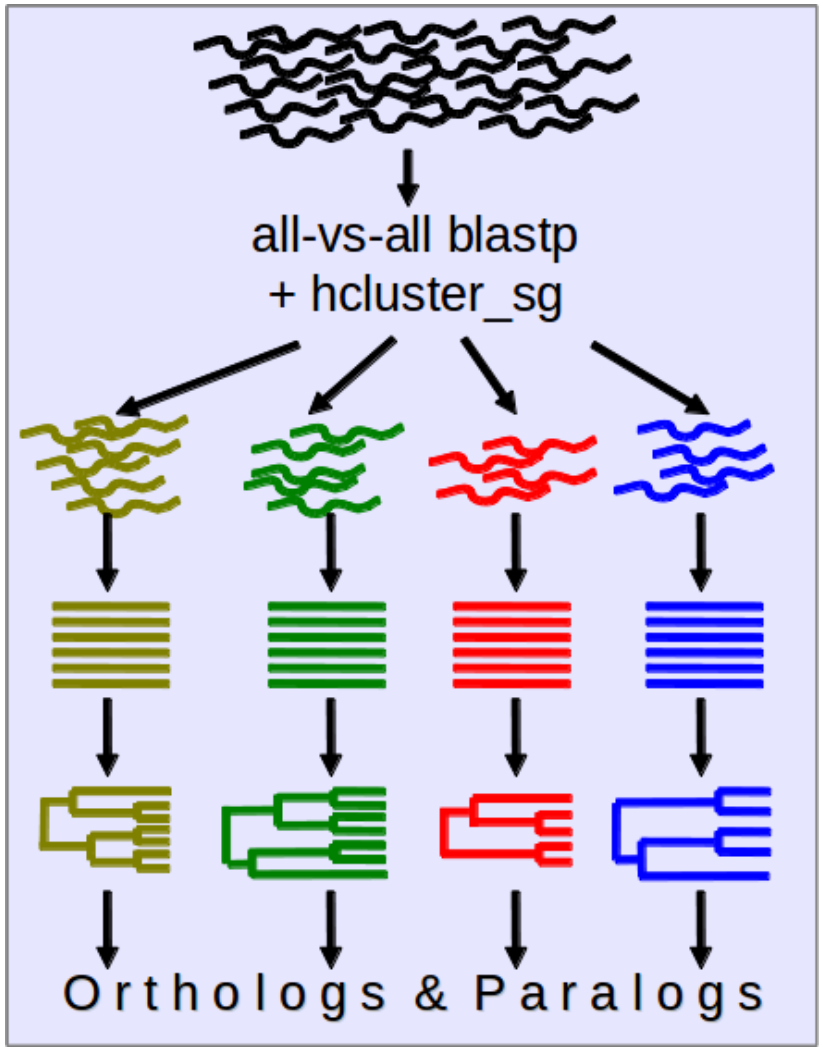
GeneTree example on the website

Tree

Multiple alignment



Protein-Tree pipeline overview



All *e!* genes – canonical prot.

BLAST

hcluster_sg

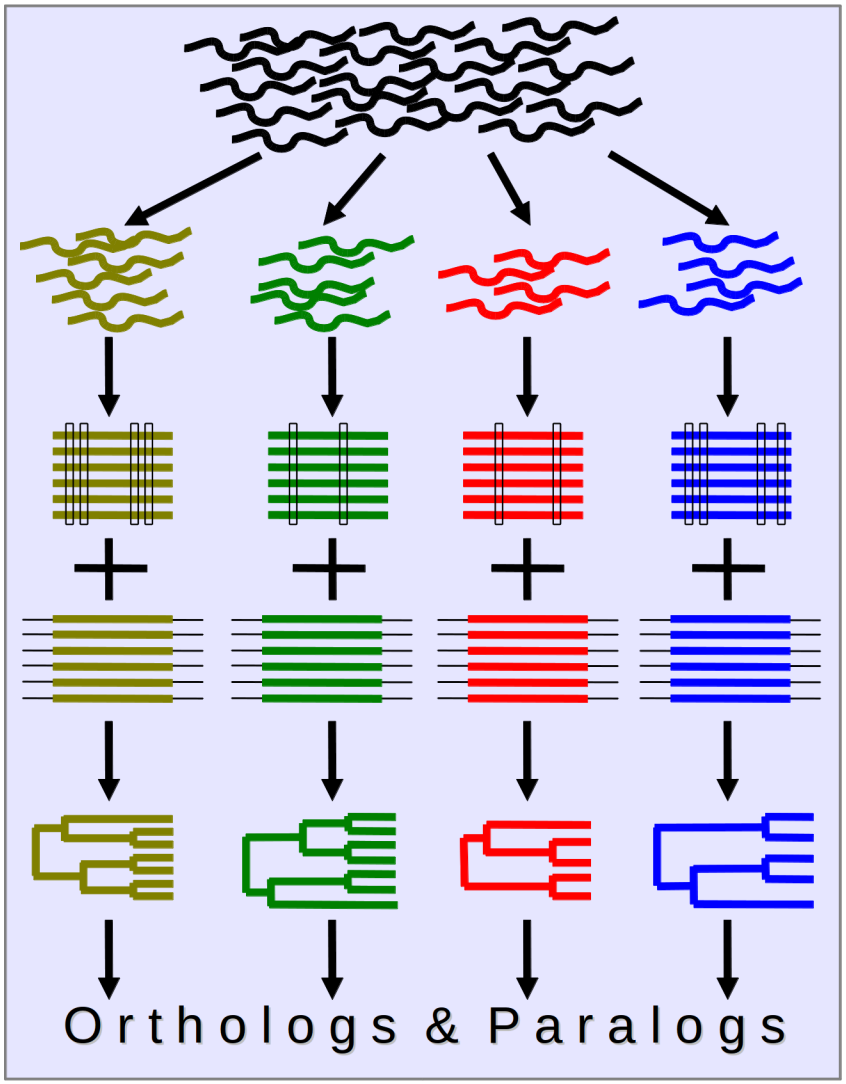
MCoffee: MSA

TreeBeST: (+ reconciliation)

Ortholog/Paralog inference

Vilella et al., Genome Res. 2009

ncRNA-Tree pipeline overview



All *e!* ncRNA genes

Grouped in Family Models - RFAM

Infernal alignment + RaxML trees

PRANK alignment + NJ/ML trees

TreeBeST (tree reconciliation)

Ortholog/Paralog inference

Pignatelli et al., in preparation

GeneTree object / *GeneTreeAdaptor*

- Represents a set of members, in a phylogenetic tree

```
$genetree_adaptor->fetch_by_stable_id(...)
$genetree_adaptor->fetch_default_for_Member(...)
```

- *fetch_all** methods require some more arguments:

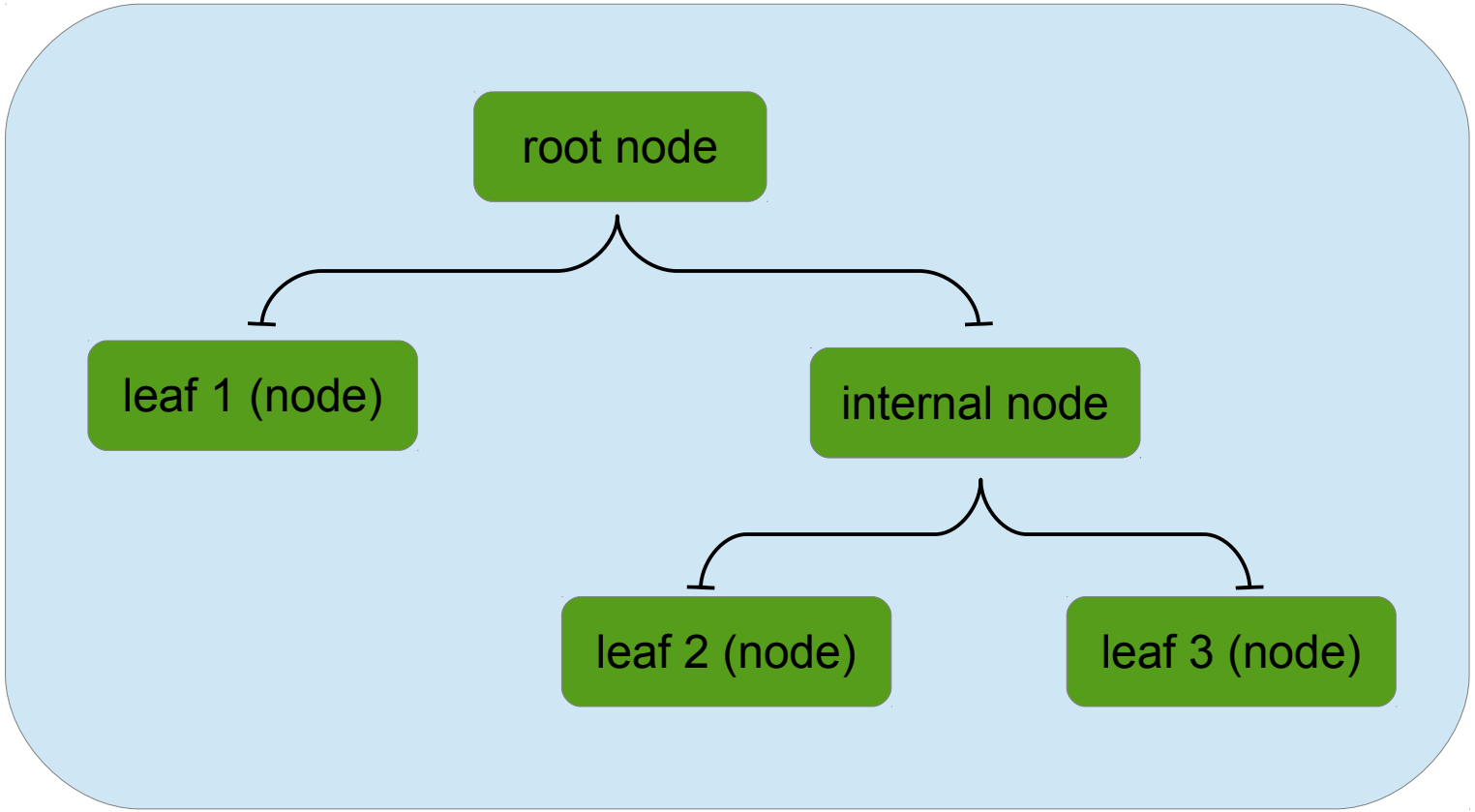
```
-clusterset_id => 'default'
-tree_type => 'tree'
-member_type => 'protein' or 'ncrna'
```

Attributes	Methods
Alignment	*$tree->get_SimpleAlign()*
Stable ID	*$tree->stable_id()*
Tree export	*$tree->newick_format('simple')* *$tree->nhx_format('full')* *$tree->print_tree()*

GeneTreeNode object

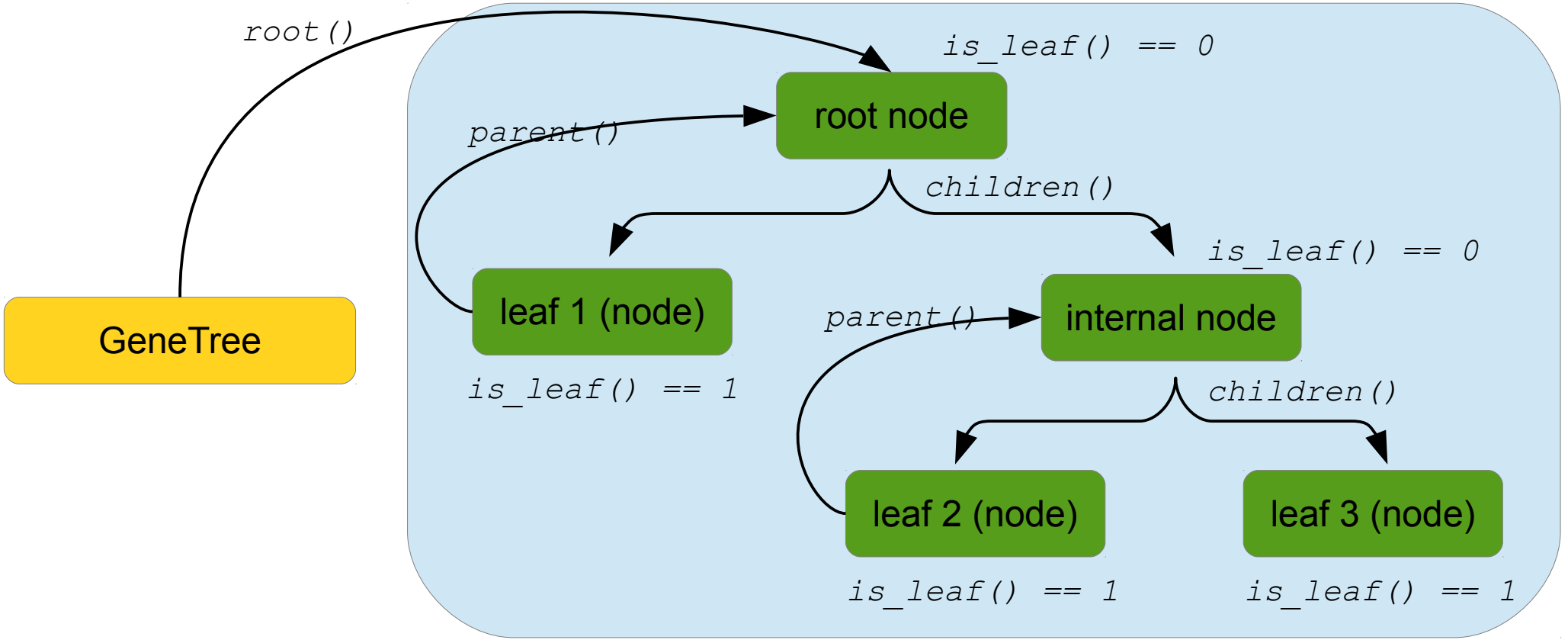
The actual tree structure is a hierarchy of *GeneTreeNode* objects

GeneTree



GeneTreeNode object

The actual tree structure is a hierarchy of *GeneTreeNode* objects



Additional information is stored with “tags”

```
$node->get_all_tags()
$node->get_tagvalue('node_type') or 'taxon_name', 'bootstrap'
```

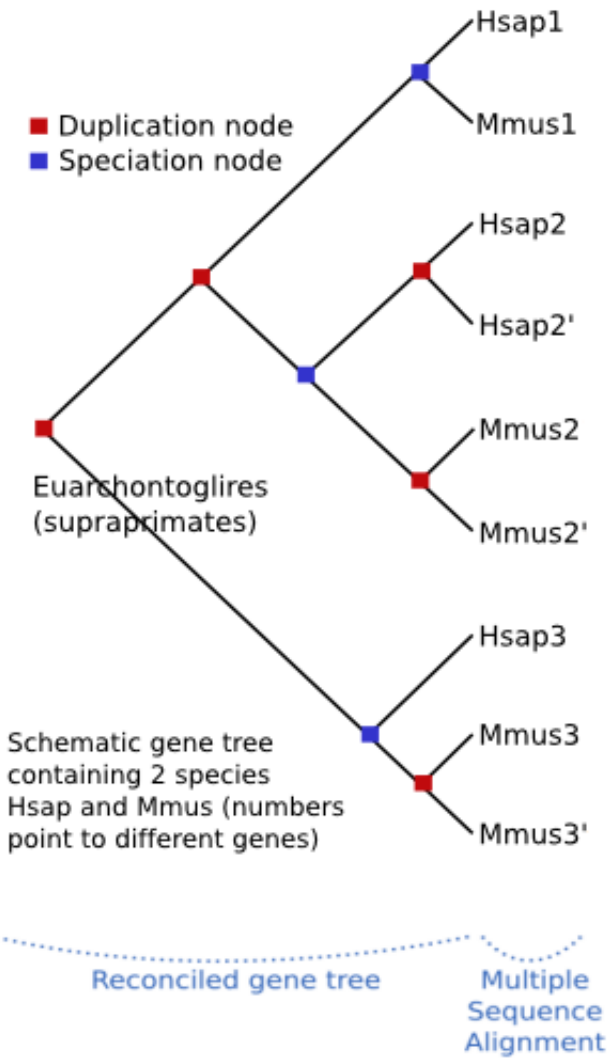
Exercises – Protein and ncRNA trees

- Print the protein tree with the stable id ENSGT00390000003602
- Print all the members of the tree containing the human ncRNA gene ENSG00000238344
- Count the number of duplication events in the tree of the zebrafish protein-coding gene ENSDARG00000003399

Outline of the course

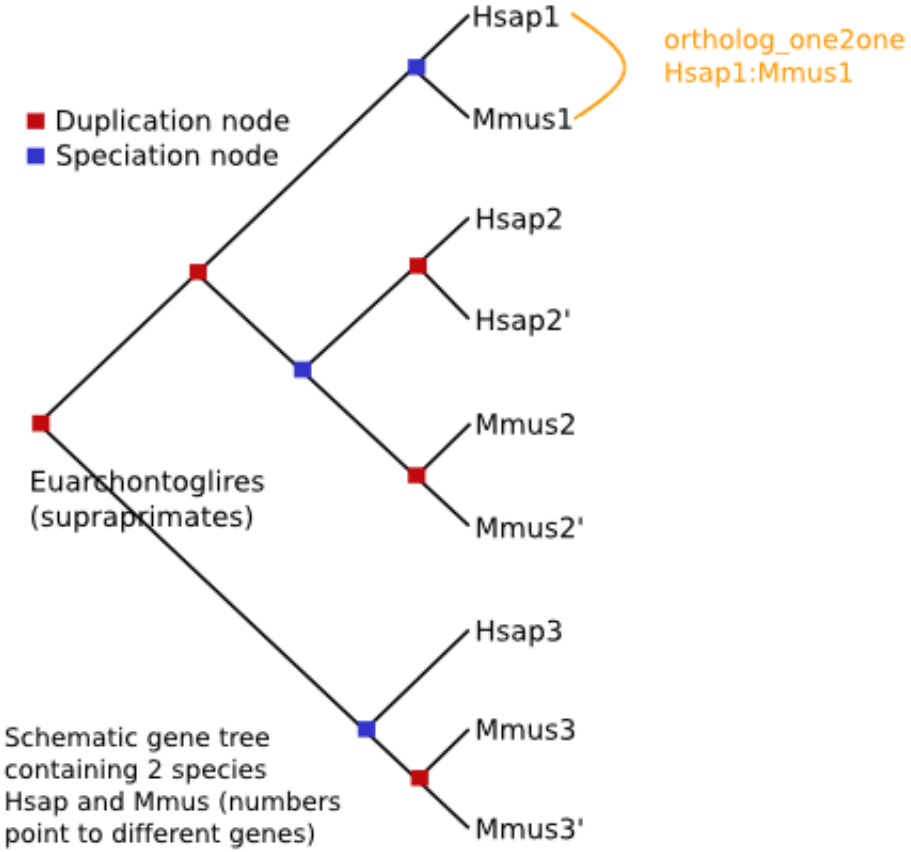
- Introduction about Compara
  - Resources
  - API
- Base objects
  - Genes, peptides, RNAs
  - Multiple / pairwise alignments
- Data objects
  - Families
  - Gene trees
  - Homologies

Homology inference

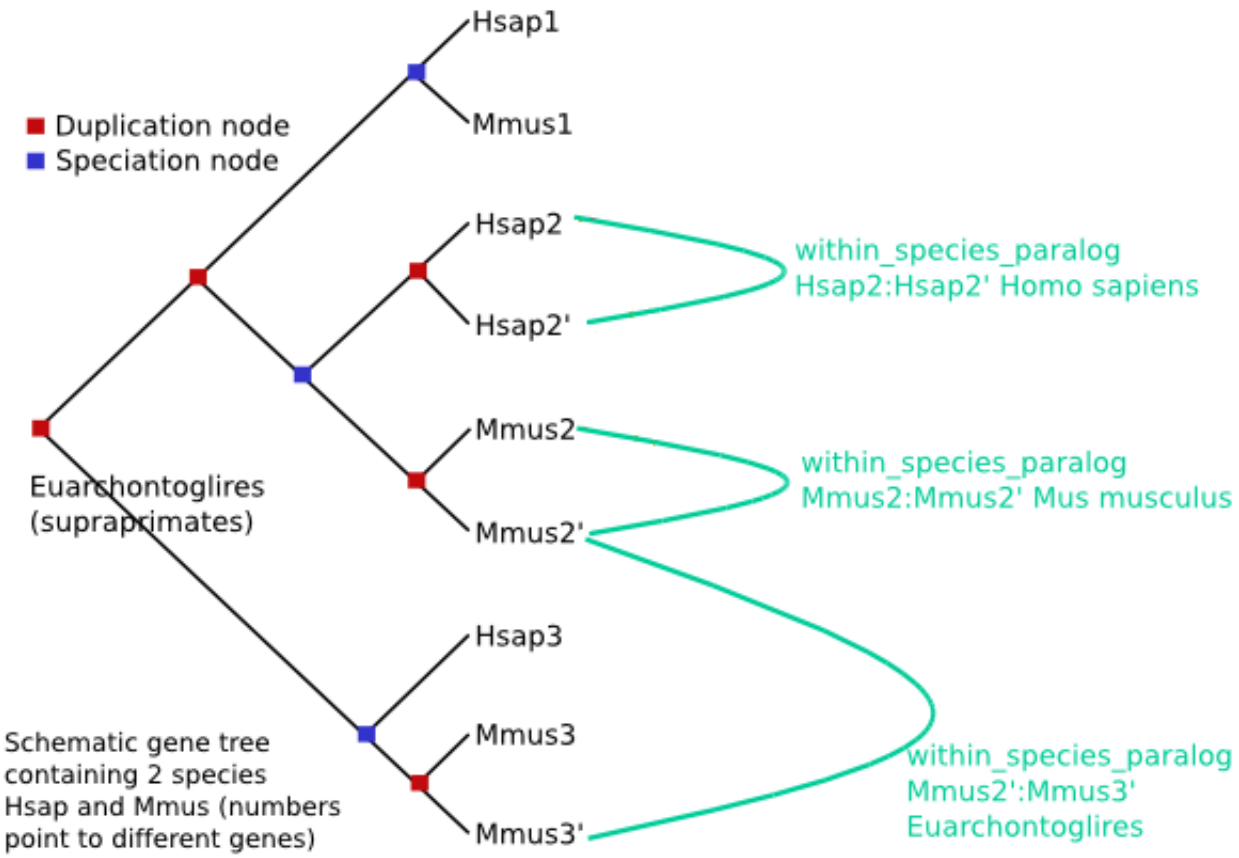


Consists in tagging the pairs of genes of all the trees with a relation type, depending on the tree topology.

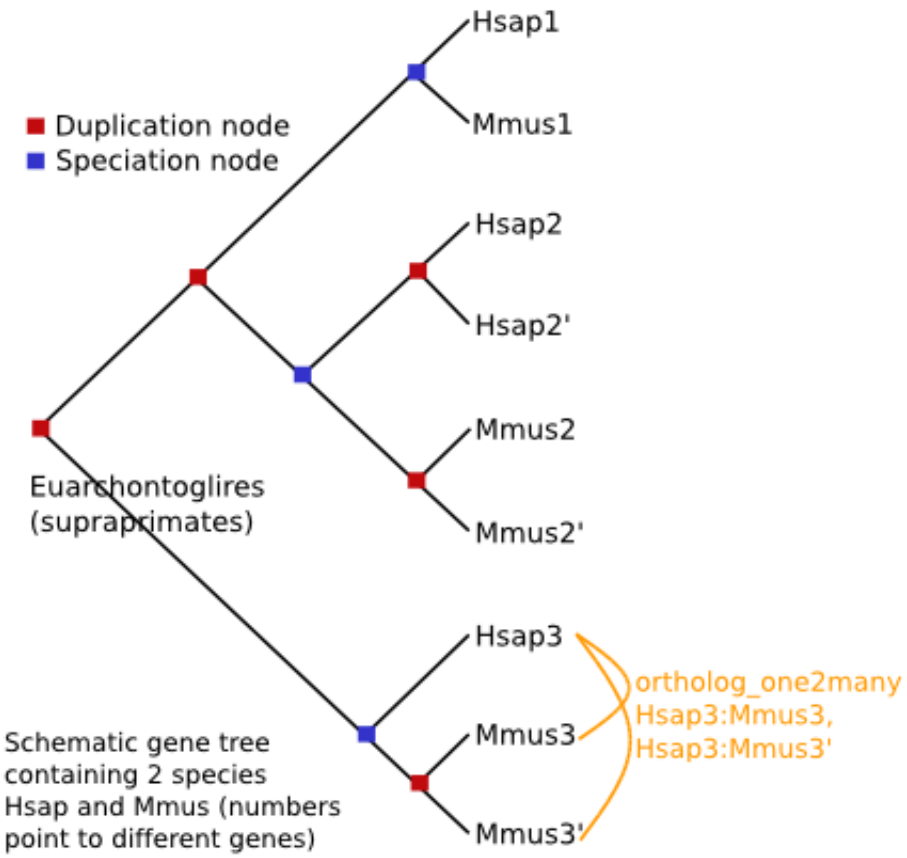
Homology inference



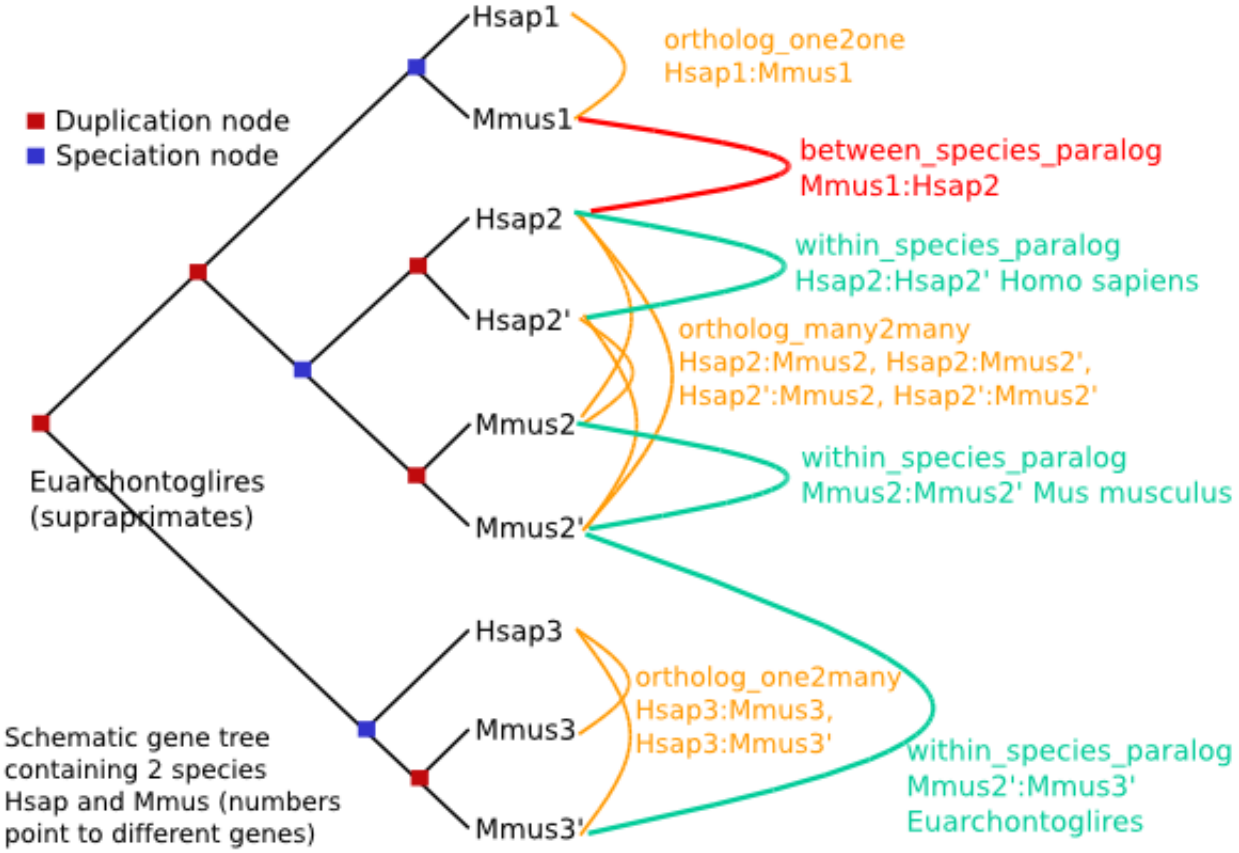
Homology inference



Homology inference



Homology inference



Homology object / *HomologyAdaptor*

- Represents a relationship between two members

```
$homology_adaptor->fetch_all_by_Member(...)
$homology_adaptor->fetch_all_by_MethodLinkSpeciesSet(..)
$homology_adaptor->fetch_all_by_Member_paired_species(..)
```

- One-to-many relationships are split: 
  - "H ortholog to M1" and "H ortholog to M2" are different objects

Attributes	Methods
Alignment	*$homology->get_SimpleAlign()*
Natural selection	*$homology->dn() / $homology->ds()*
Gene content	*$homology->get_all_GeneMembers()*
Homology characteristics	*$homology->description()* *$homology->taxonomy_level()*
Node in the gene tree	*$homology->node_id()*

Exercises - Homologies

- Get all the homologues for the human gene ENSG00000229314
- Count the number of “one2one” homologues between human and mouse
- Find the human orthologues of ENSMUSG00000004843 and ENSMUSG00000025746. For each homology, display the alignment and the dn value. Comment on the divergence