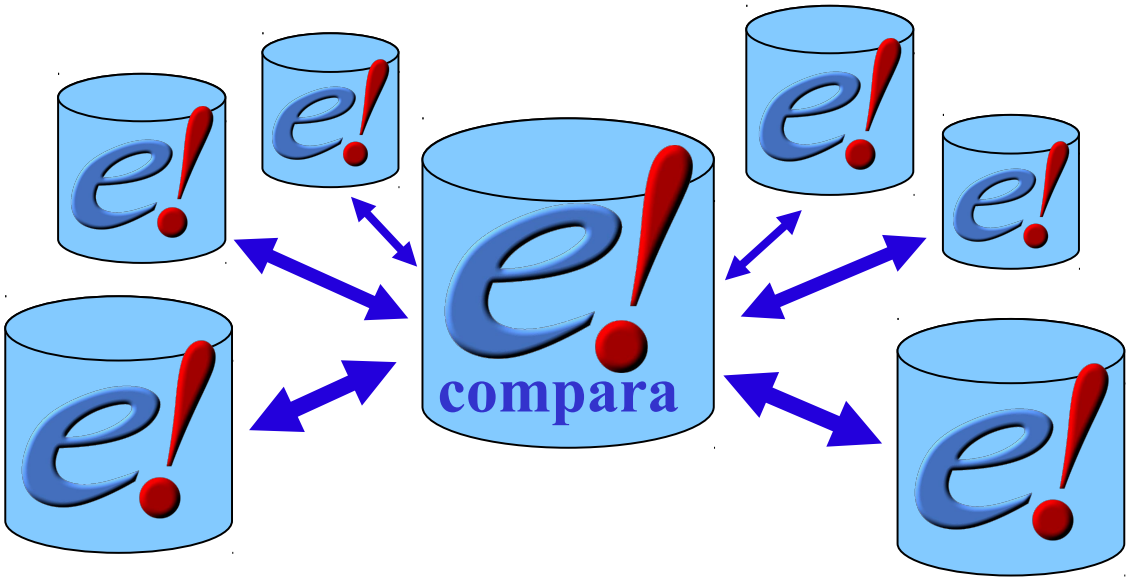
Ensembl Compara Perl API



Stephen Fitzgerald and Matthieu Muffato

EBI - Wellcome Trust Genome Campus, UK

Dec 2011

EMBL-EBI

wellcome trust sanger institute

What is Ensembl Compara?

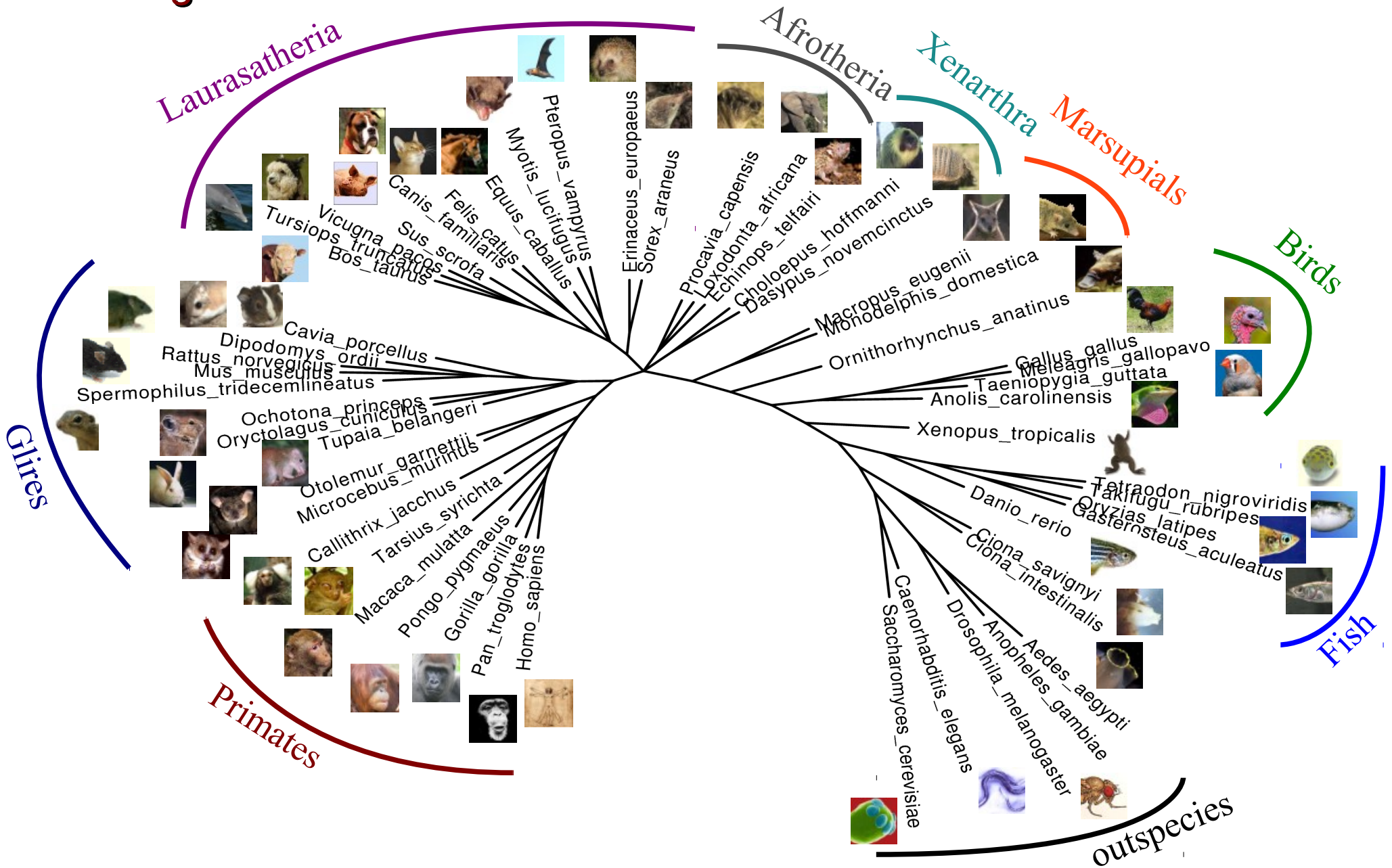
A single database which contains precalculated comparative genomics data and which is linked to all the Ensembl Species databases.

Access via perl API and mysql

A production system for generating that database

(not in this presentation)

The genomes in Ensembl



Help

- 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/main.html
- CVS
  - ensembl-compara/docs/ComparaTutorial.pdf
  - ensembl-compara/docs/protein_schema.png
  - ensembl-compara/docs/genomic_schema.png
- ensembl-dev mailing list:
  - ensembl-dev@ebi.ac.uk

Compara data

Genome level *(Stephen, this afternoon)*

Whole genome alignments (Pairwise and Multiple)

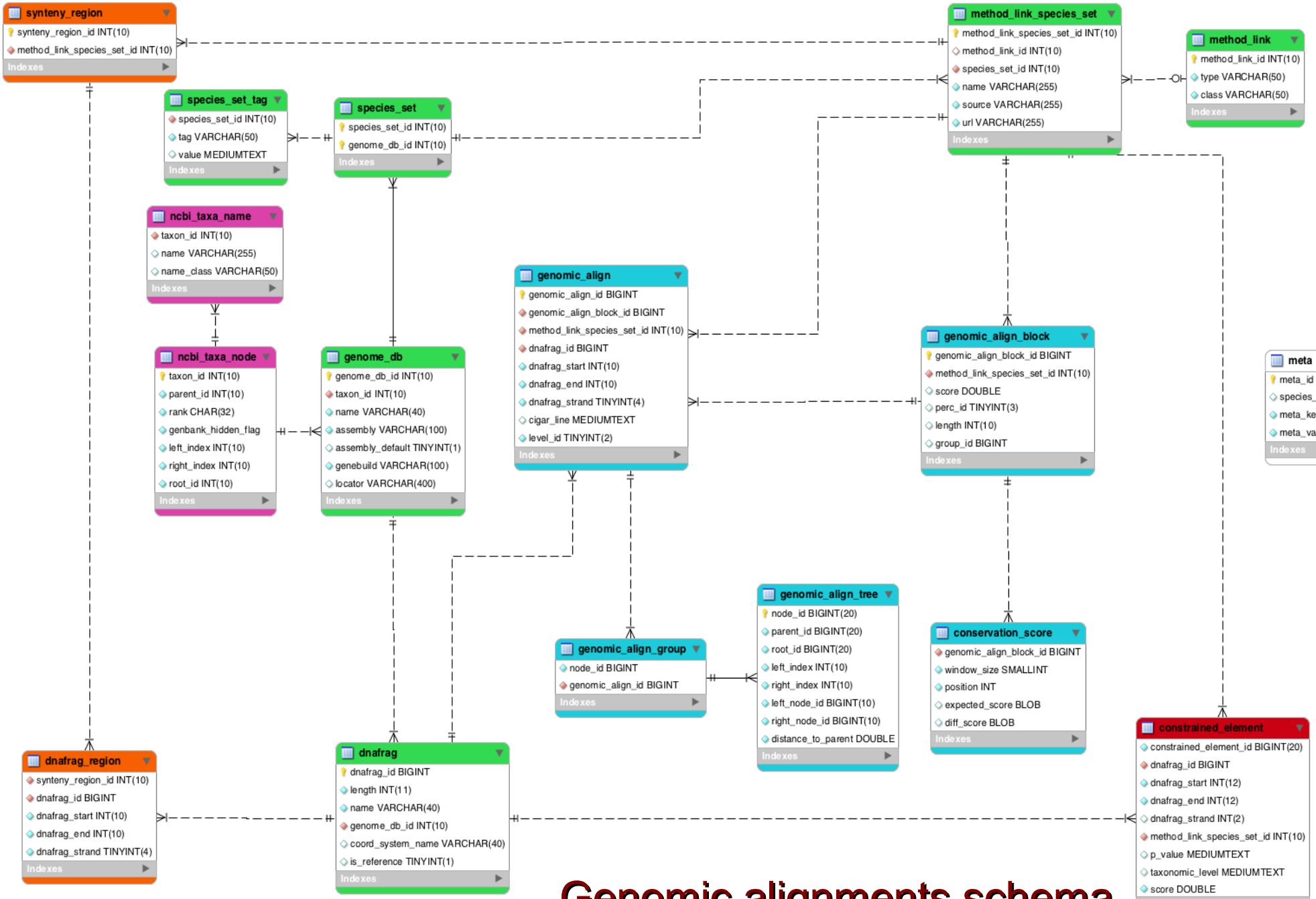
Syntenic regions (based on pair-wise align.)

Gene level *(now !)*

Raw protein alignments
Families (Clusters of proteins)

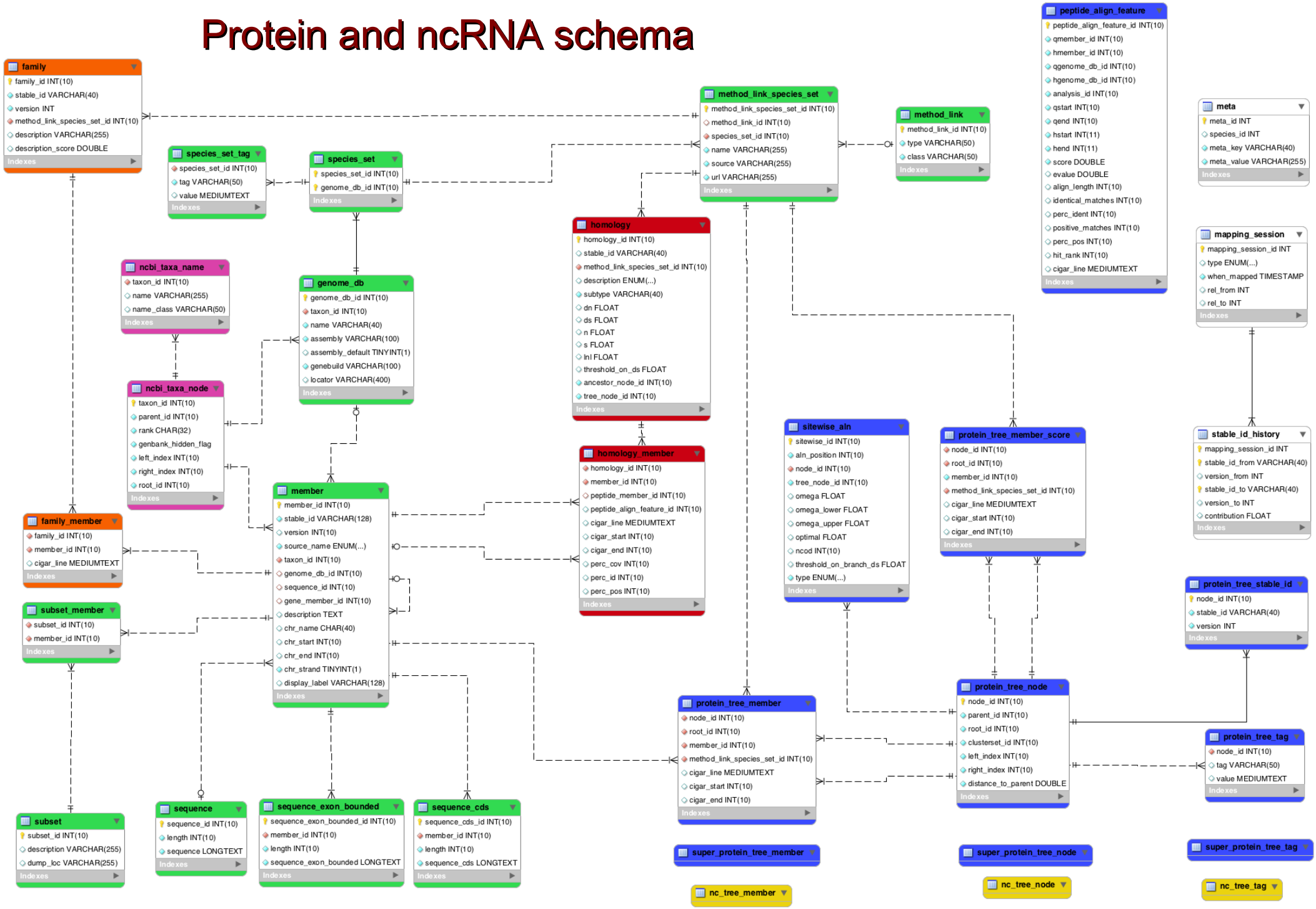
Protein trees (since June 2006, rel. 39)
Non-coding RNA trees (since May 2010, rel. 58)

Gene orthology / paralogy predictions



Genomic alignments schema

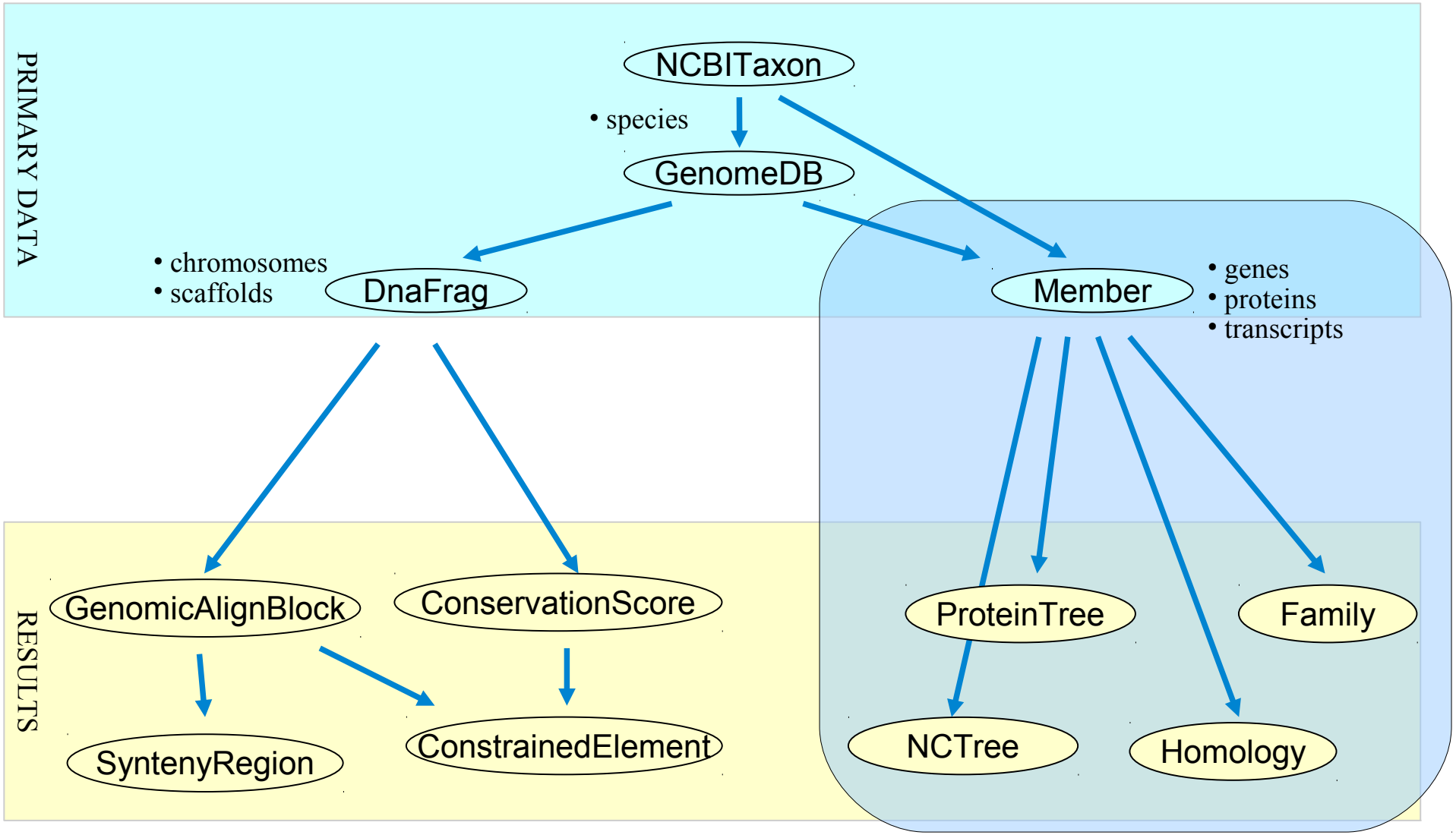
Protein and ncRNA schema



The Compara Perl API

- Written in Object-Oriented Perl
- Used to retrieve data from and store data into ensembl-compara database (only the production pipeline generates the alignments)
- 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 object model overview



Compara template script (with GenomeDB)

```
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 object for the data type you want
# Corresponds to a table in the compara database
my $genomedb_adaptor =
   $reg->get_adaptor("Multi", "compara", "GenomeDB");

print "All Ensembl species:\n";
# Fetch the data objects using the adaptor
# Corresponds to a row entry (tuple) in the table
my $all_genomedb = $genomedb_adaptor->fetch_all();

foreach my $this_genomedb (@$all_genomedb) {
  # Do some stuff with the data object
  # Corresponds to fields (attributes) in the row
  print "full name: ",
    $this_genomedb->taxon ? $this_genomedb->taxon->binomial : "?";
  print ", short name: ", $this_genomedb->short_name;
  print ", assembly: ", $this_genomedb->assembly, "\n";
}
```

Member object

- Represents a gene, a transcript, or a protein
- Retrieved with MemberAdaptor

```
$member_adaptor->fetch_by_source_stable_id(...)
$member_adaptor->fetch_all_by_source_taxon(...)
```

- Possible sources are ENSEMBLGENE, ENSEMBLPEP, ENSEMBLTRANS, Uniprot/SPTREMBL, Uniprot/SWISSPROT

Attributes	Methods

Gene properties	`$member->chr_start()`
	`$member->stable_id()`
	`$member->sequence()`

Species	`$member->taxon()`
Linked members	`$member->get_Translation()` `$member->get_all_peptide_Members()`

Exercises - Members

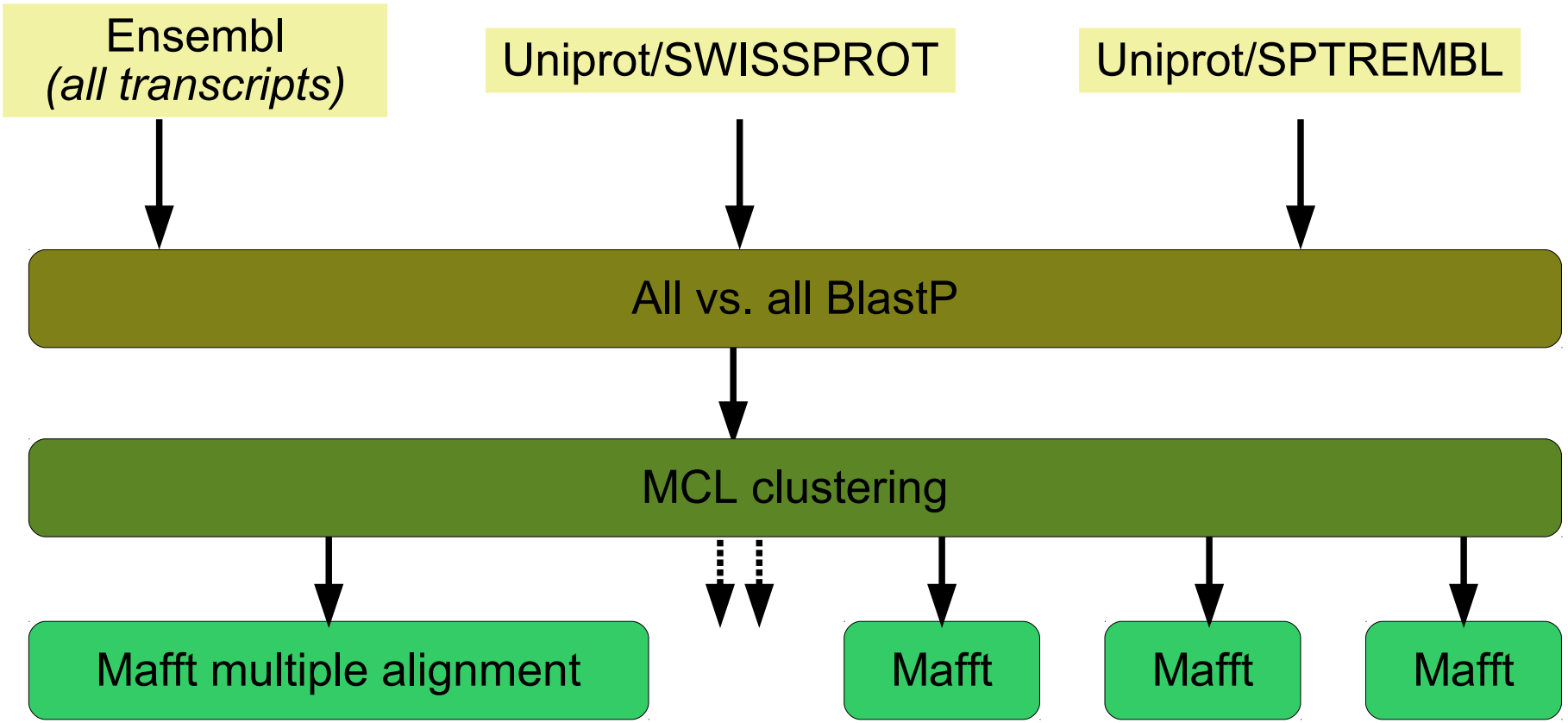
- Print the sequence of the Member corresponding to SwissProt protein O93279

- Find the Member(s) for the human gene(s) BRCA2

- Find and print the sequence of all the peptide Members corresponding to the human gene(s) CTDP1

Families

Gene family clustering predictions



Family object

- Represents a group of members
- Retrieved with FamilyAdaptor

  `$family_adaptor->fetch_all_by_Member(...)`
-  Transcripts from a given gene can belong to different families !

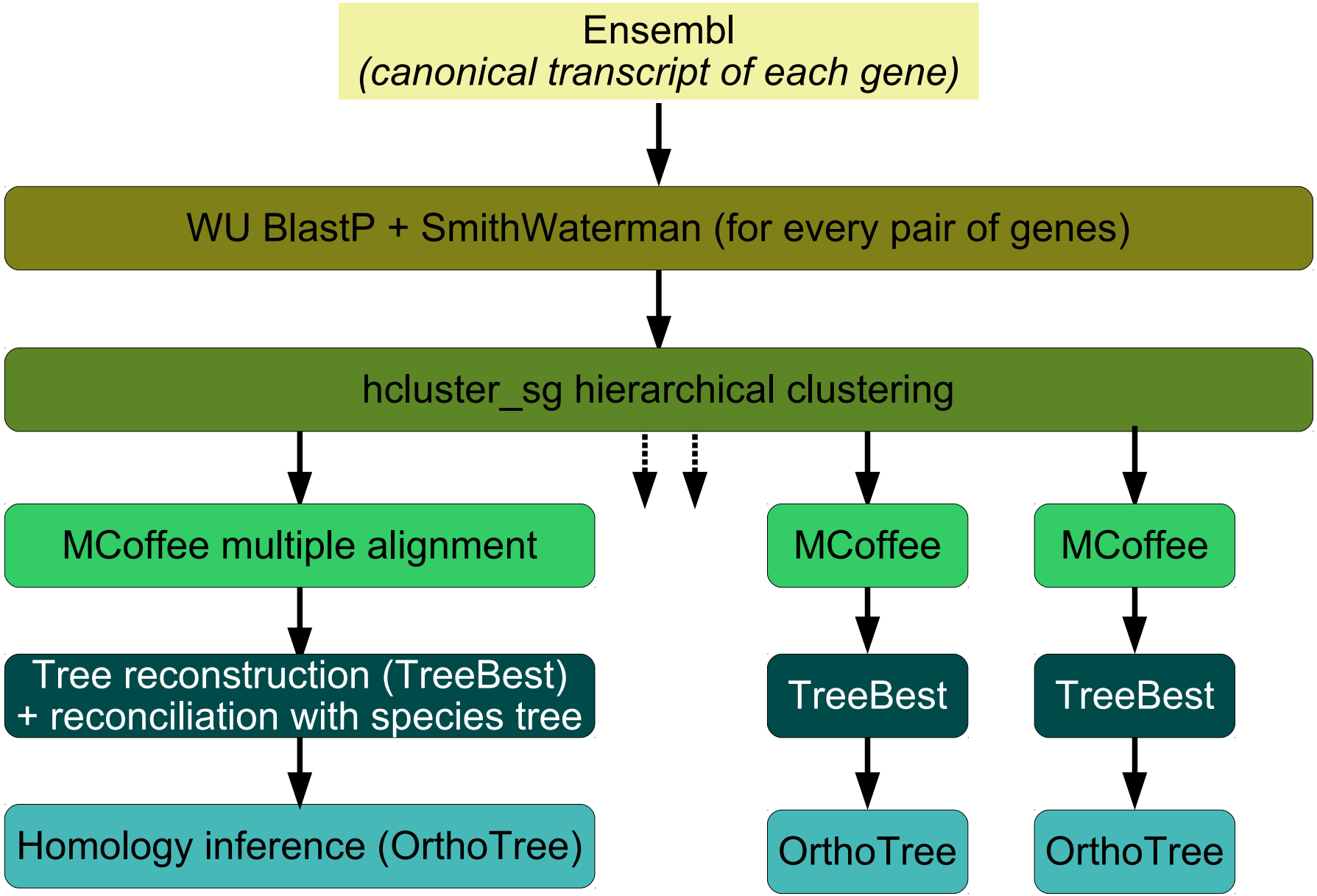
Attributes	Methods
Alignment	`$family->get_SimpleAlign()`
Biological function	`$family->description()`
Gene content	`$family->get_all_Members()` `$family->get_all_Member_Attribute()`

Exercises – Families

- Get the family predicted for the human gene ENSG00000139618

- Get the alignments corresponding to the family of the human gene ENSG00000113070

Gene trees


Ensembl
(canonical transcript of each gene)
WU BlastP + SmithWaterman (for every pair of genes)
hcluster_sg hierarchical clustering
MCoffee multiple alignment
MCoffee
MCoffee
Tree reconstruction (TreeBest)
+ reconciliation with species tree
TreeBest
TreeBest
Homology inference (OrthoTree)
OrthoTree
OrthoTree

ProteinTree & NCTree objects

- Represents a group of members, organised in a phylogenetic tree (one set of trees for protein coding genes, and one for non-coding RNAs)
- Retrieved with ProteinTreeAdaptor and NCTreeAdaptor

```
$proteintree_adaptor->fetch_by_Member_root_id(...)
$proteintree_adaptor->fetch_all()
```

- All trees are linked to a huge tree (root_id = 1)

Attributes	Methods
The transcript used for the trees	`$member->get_canonical_peptide_Member()`
Alignment	`$proteintree->get_SimpleAlign()`
Children (in the tree structure)	`$proteintree->children()` `$proteintree->get_all_leaves()`
Tree export	`$proteintree->newick_simple_format()` `$proteintree->print_tree()`

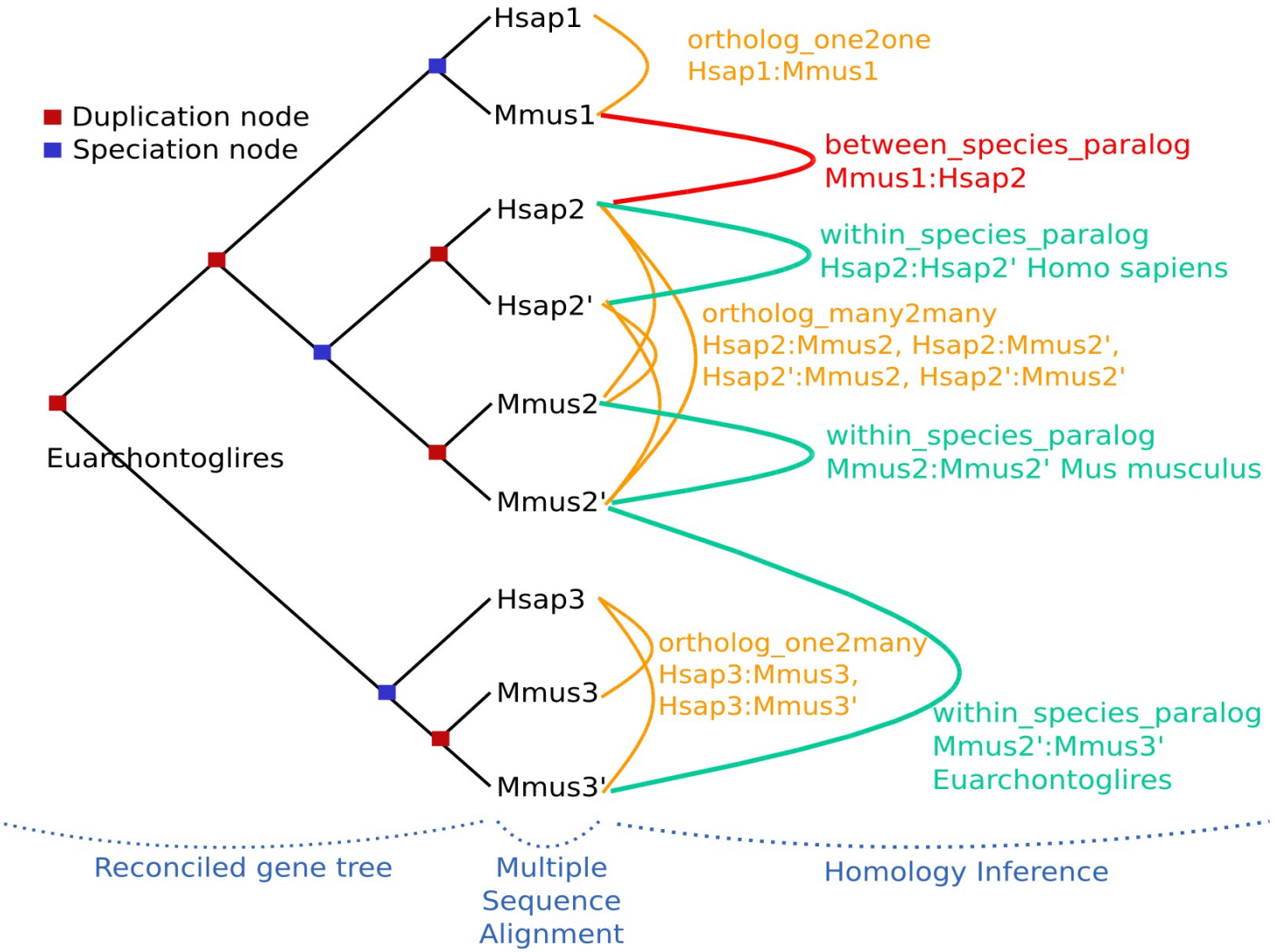
Exercises – ProteinTrees

- Print the protein tree for the human gene ENSG00000139618

- Print all the members of the tree containing the human gene ENSG00000060069 (with the direct API function)

- Print all the members of the tree containing the human gene ENSG00000060069 (by recursively going into the tree structure)

Homology inference



Homology object

- Represents a relationship between two members
- Retrieved with HomologyAdaptor

```
$homology_adaptor->fetch_all_by_Member(...)
$homology_adaptor->fetch_all_by_genome_pair(...)
```

-  One-to-many relationships (H ortholog to M1 and M2) appear are split (H ortholog to M1 and H ortholog to M2)

Attributes	Methods
Alignment	`$homology->get_SimpleAlign()`
Natural selection	`$homology->dn() / $homology->ds()`
Gene content	`$homology->gene_list()`
Homology characteristics	`$homology->description()` `$homology->taxonomy_level()`

Exercises - Homology

- Get all the predicted homologues for the human gene ENSG00000170037

- Get all the mouse homologues predicted for the human gene ENSG00000060069