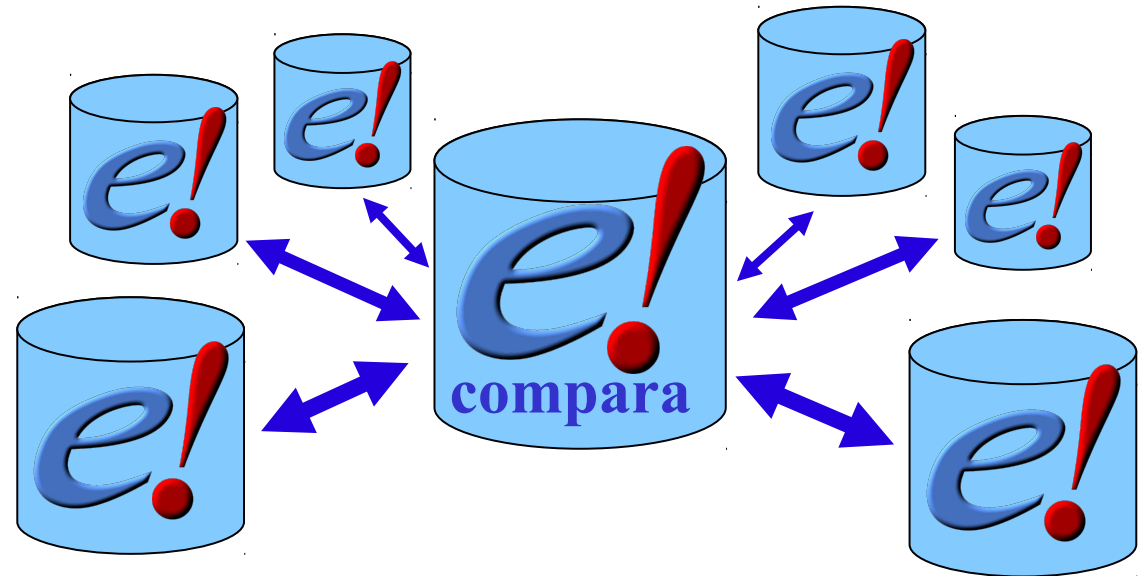


Ensembl Compara Perl API



Stephen Fitzgerald and Matthieu Muffato

EBI - Wellcome Trust Genome Campus, UK



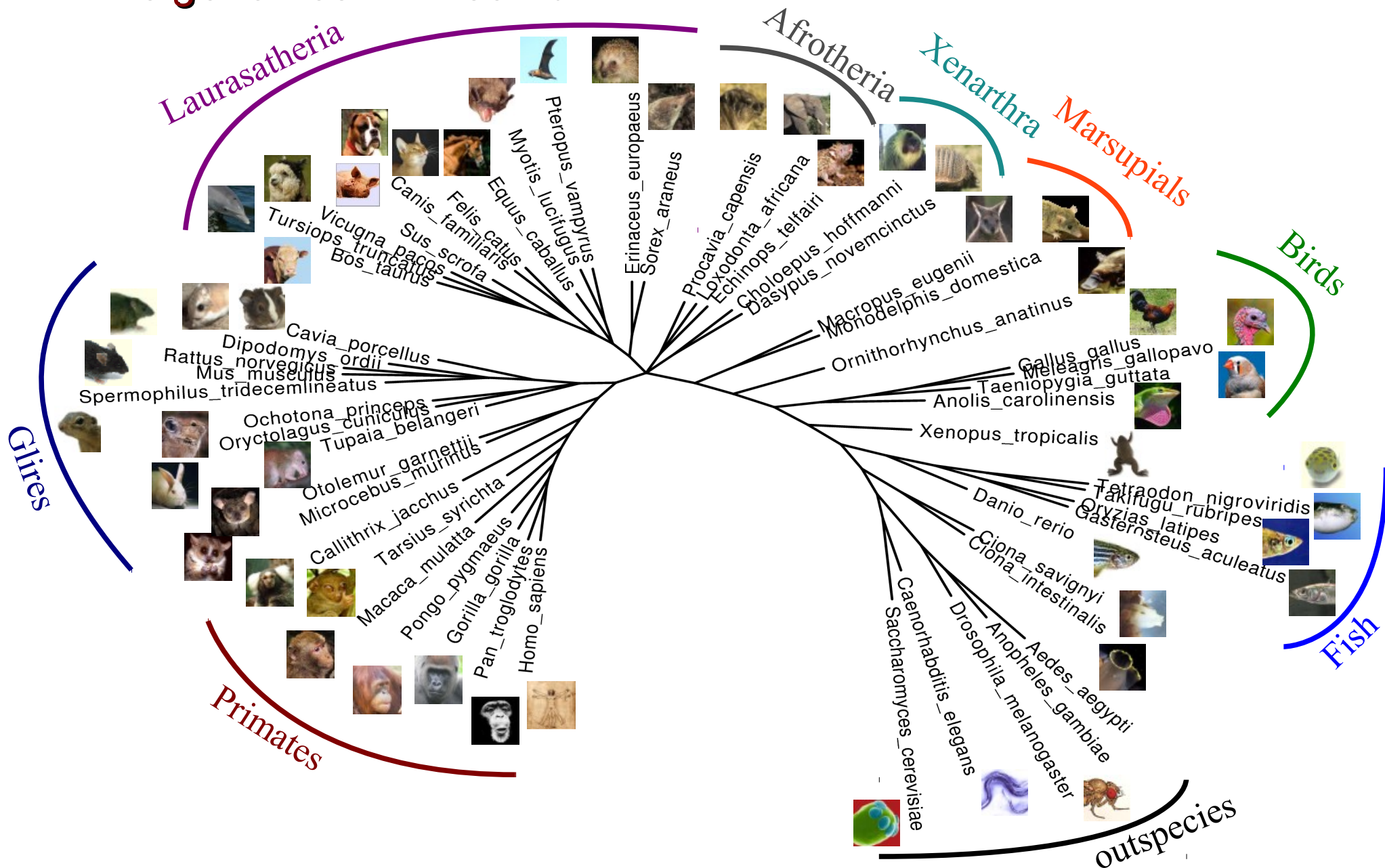
What is Ensembl Compara?

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



Compara data

Genome level *(Stephen, this afternoon)*

Whole genome alignments (Pairwise and Multiple)

Syntenic regions (based on pair-wise align.)

Gene level *(now !)*

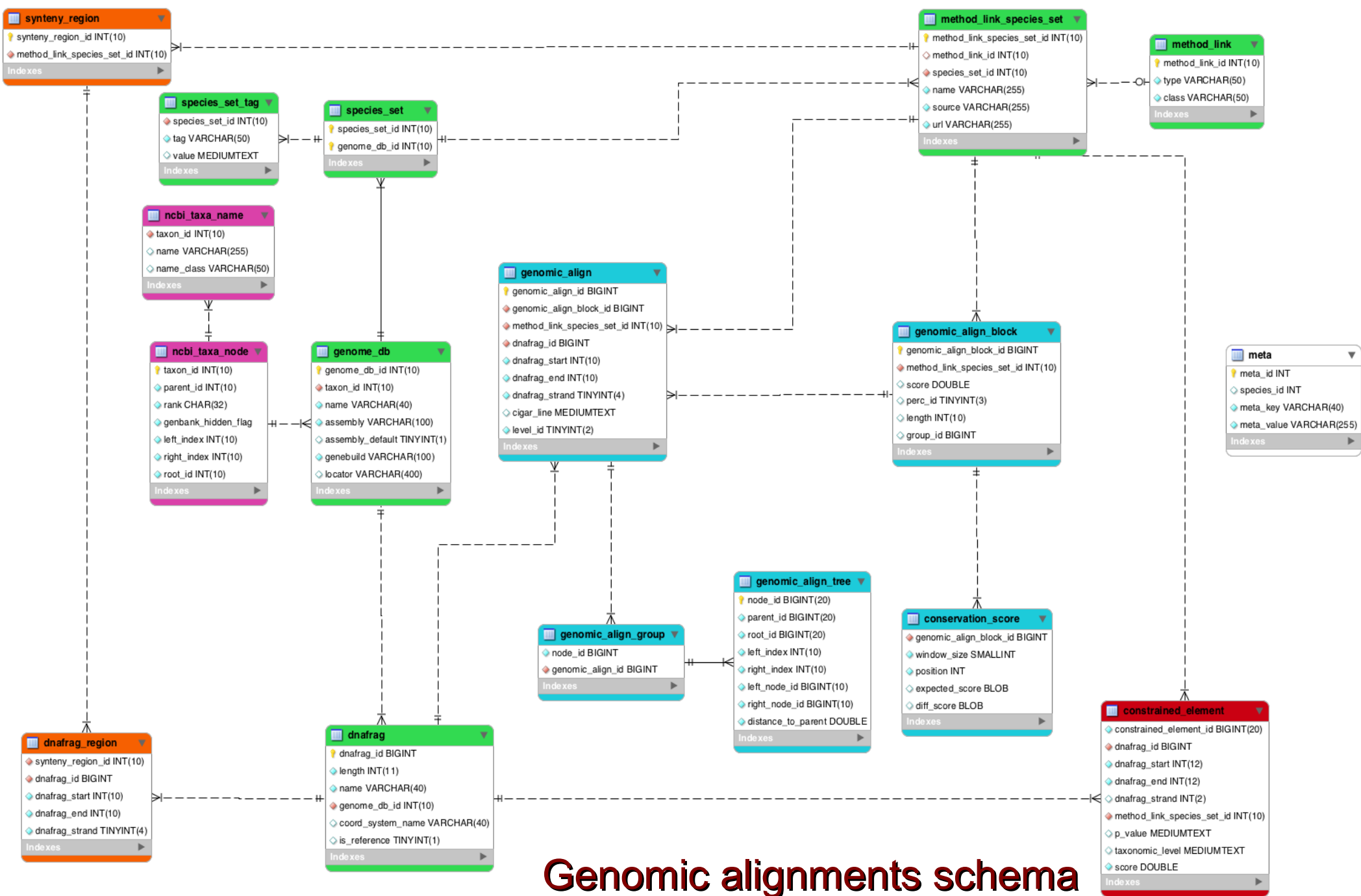
Raw protein alignments
Families (Clusters of proteins)

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

Gene orthology / paralogy predictions

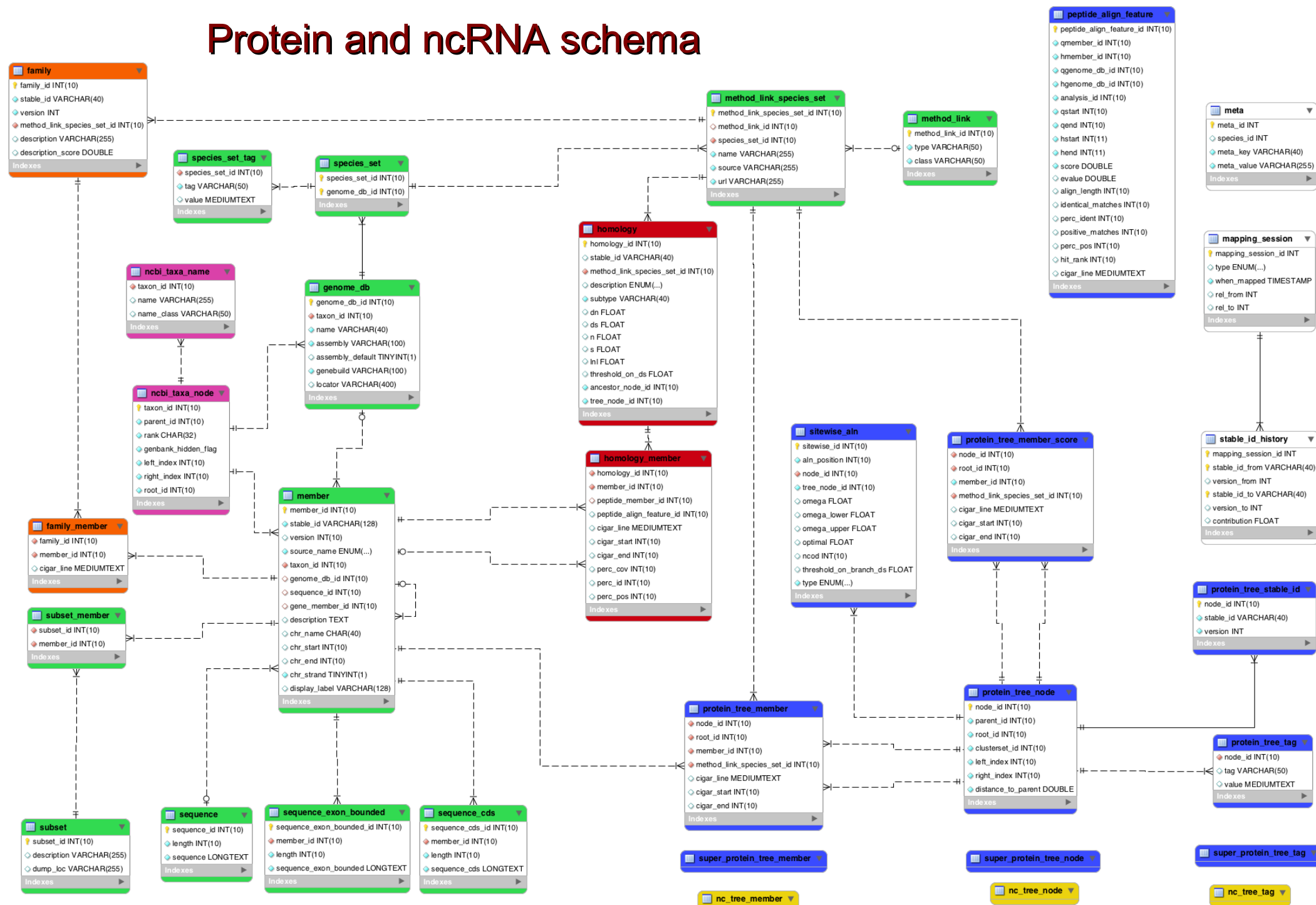
Help

- 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/Pdoc/ensembl-compara/>
 - <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



Genomic alignments schema

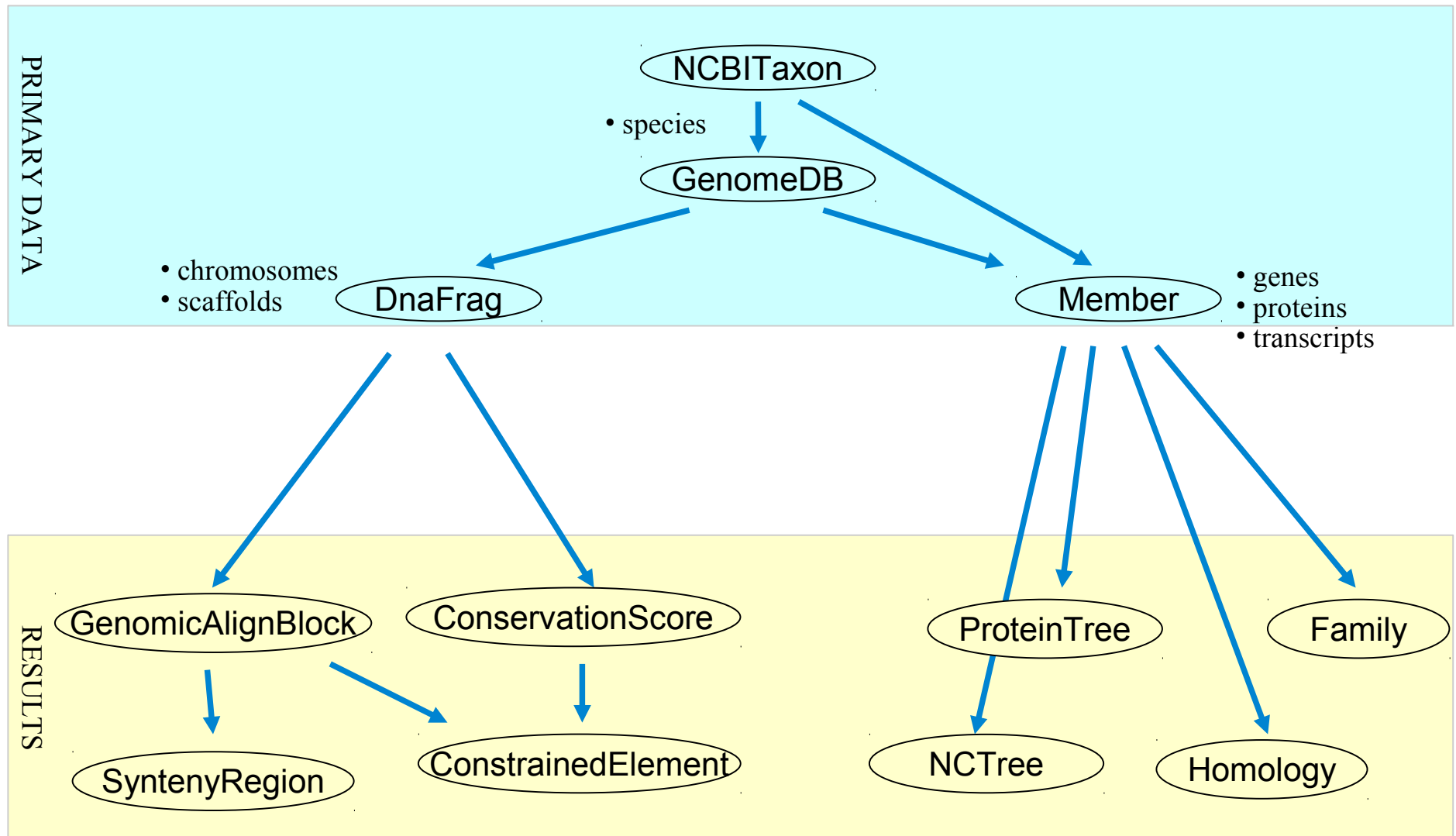
Protein and ncRNA schema



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



GenomeDB and NCBITaxon objects

- GenomeDB are used for Ensembl species
NCBITaxon can represent any species of the NCBI database

- Retrieved with GenomeDBAdaptor

```
$genomedb_adaptor->fetch_by_Slice(...)  
$genomedb_adaptor->fetch_all()
```

and NCBITaxonAdaptor

```
$ncbitaxon_adaptor->fetch_node_by_genome_db_id(...)  
$ncbitaxon_adaptor->fetch_node_by_name(...)  
$ncbitaxon_adaptor->fetch_node_by_taxon_id(...)
```

Attributes	Methods
Species name	<i>\$genomedb->name()</i> <i>\$taxon->genus()</i>
Ensembl properties	<i>\$genomedb->assembly()</i> <i>\$genomedb->genebuild()</i>
Taxon object	<i>\$genomedb->taxon()</i>

Example Compara script (with GenomeDB)

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

$reg->load_registry_from_db(
    -host => "ensembl.org",
    -user => "anonymous"
);

my $genomedb_adaptor = $reg->
    get_adaptor("Multi", "compara", "GenomeDB");

print "All Ensembl species:\n";
my $all_genomedb = $genomedb_adaptor->fetch_all();

foreach my $this_genomedb (@$all_genomedb) {
    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

- Represent 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

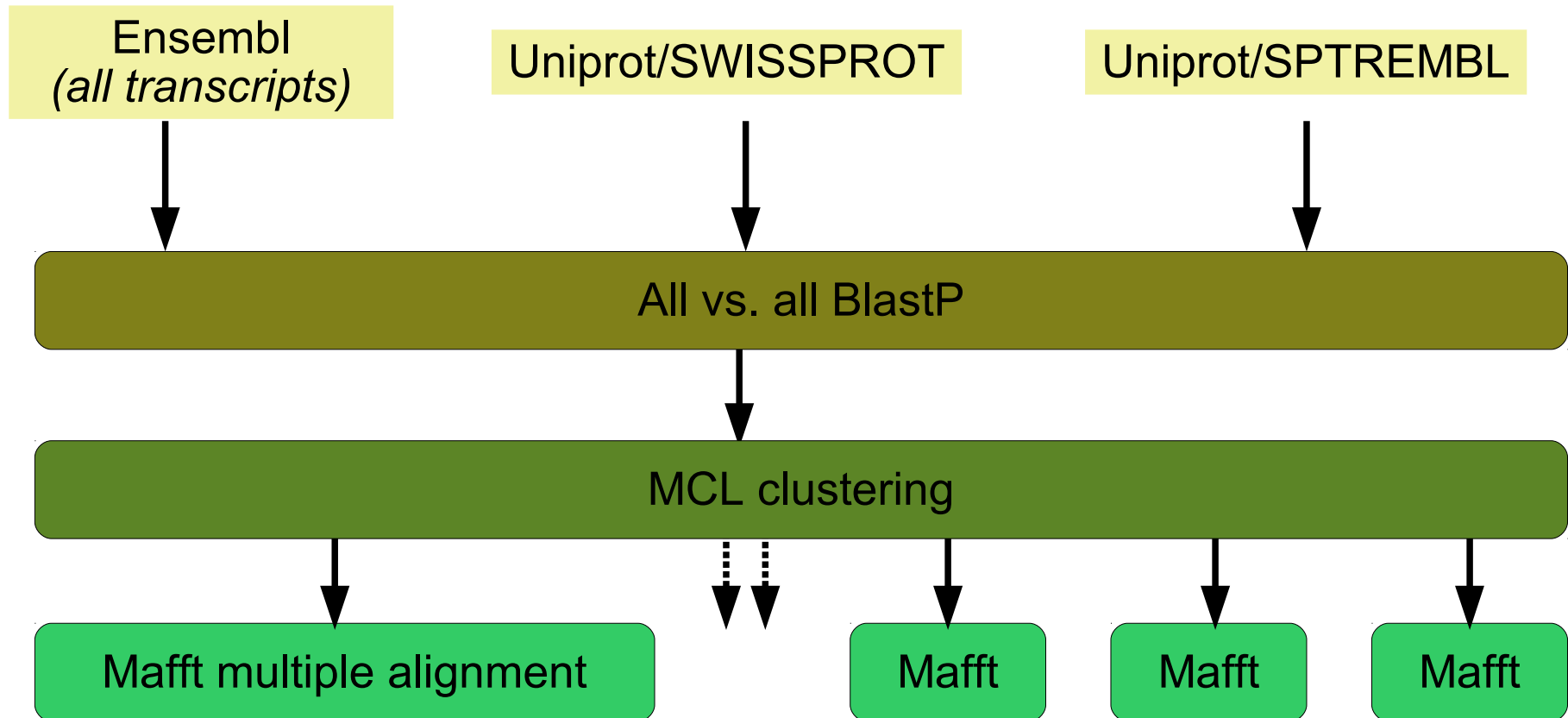
Attributes	Methods
Gene properties	<i>\$member->chr_start()</i> <i>\$member->stable_id()</i> <i>\$member->sequence()</i>
Species	<i>\$member->taxon()</i>
Linked members	<i>\$member->get_Translation()</i> <i>\$member->get_all_peptide_Members()</i>

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

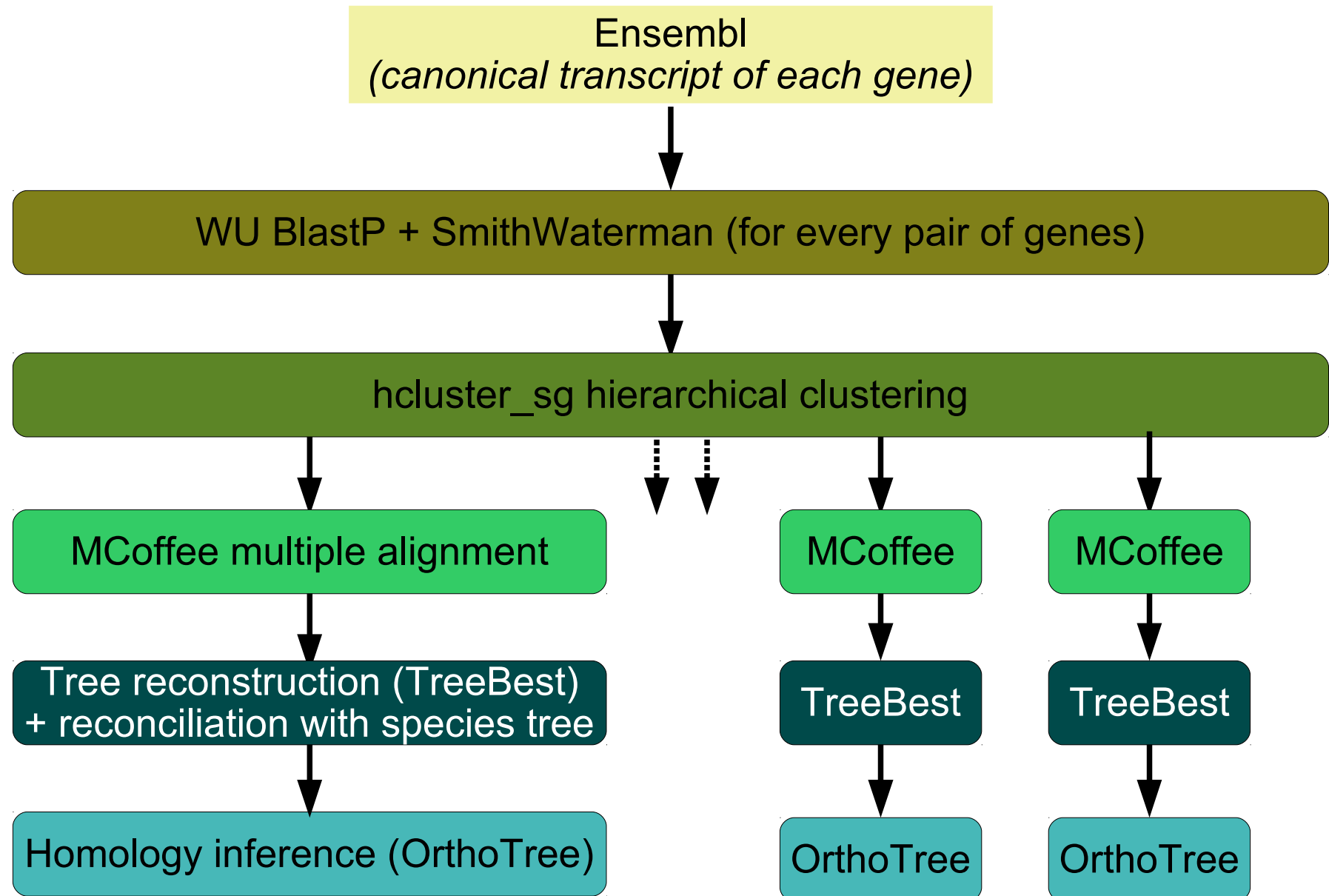
- Represent a group of members
- Retrieved with FamilyAdaptor
\$family_adaptor->fetch_all_by_Member(...)
-  Transcripts from a given gene can belong to different families !

Attributes	Methods
Alignment	<i>\$family->get_SimpleAlign()</i>
Biological function	<i>\$family->description()</i>
Gene content	<i>\$family->get_all_Members()</i> <i>\$family->get_all_Member_Attribute()</i>

Exercises – Families

- Get the family predicted for the human gene(s) BRCA2
- Get the alignments corresponding to the family of the human gene(s) HBEGF

Gene trees



ProteinTree & NCTree objects

- Represent 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

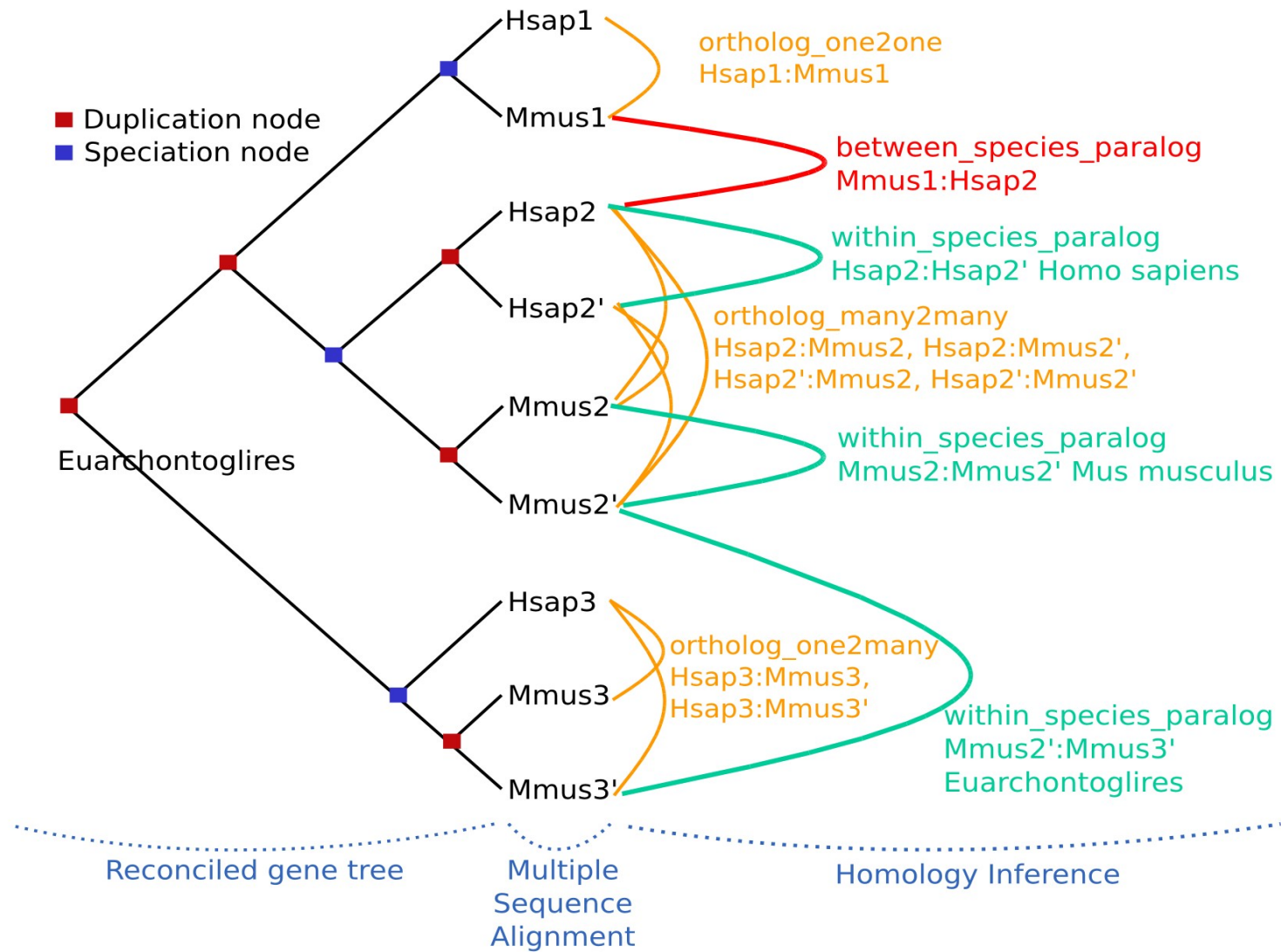
```
$proteintree_adaptor->fetch_by_Member_root_id(...)  
$proteintree_adaptor->fetch_all()
```
- All trees are linked to a huge tree (root_id = 1)

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

Exercises – ProteinTrees


- Print the protein tree for the human gene BRCA2
- Print all the members of the tree containing the human gene CTDP1 (with the direct API function)
- Print all the members of the tree containing the human gene CTDP1 (by recursively going into the tree structure)

Homology inference



Homology object

- Represent 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 as distinct (H ortholog to M1 and H ortholog to M2)

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

Exercises - Homology

- Get all the predicted homologues for the human gene(s) CNTROB
- Get all the mouse homologues predicted for the human gene(s) CTDP1