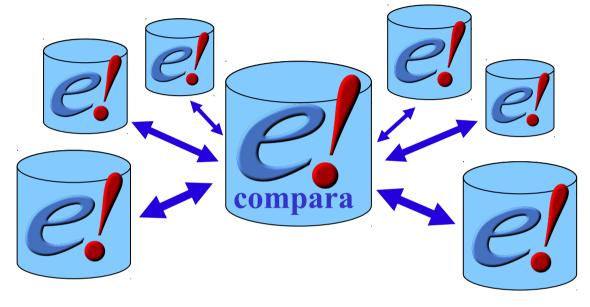


Ensembl Compara Perl API



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

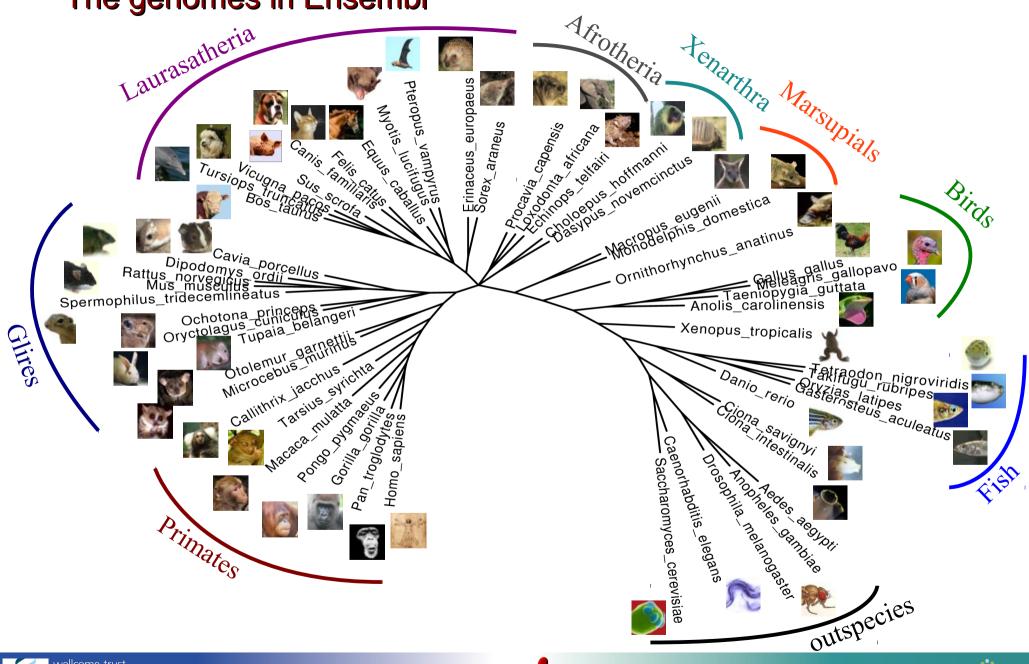
(not in this presentation)







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







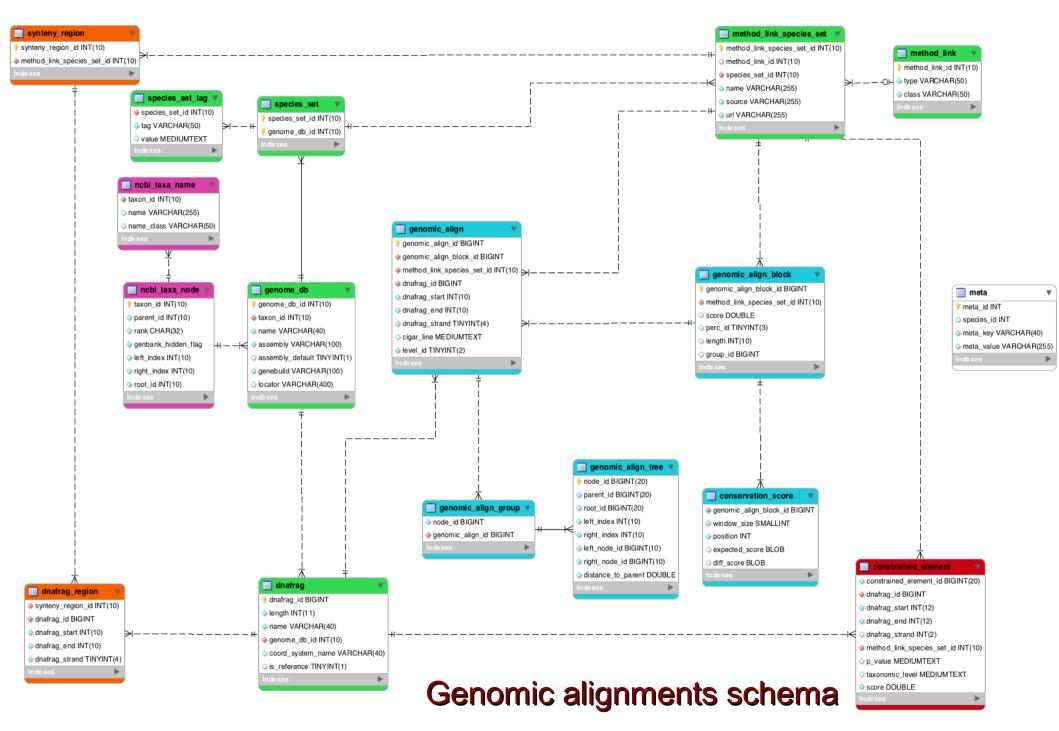
Help

- 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











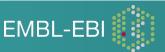




Protein and ncRNA schema peptide_align_feature_id INT(10) amember_id INT(10) hmember_id INT(10) family ggenome db id INT(10) hgenome_db_id INT(10) family_id INT(10) method_link_species_set v analysis_id INT(10) stable_id VARCHAR(40) meta astart INT(10) version INT method link species set id INT(10) method link method_link_species_set_id INT(10) qend INT(10) meta id INT method_link_id INT(10) method link id INT(10) description VARCHAR(255) hstart INT(11) species_id INT species set id INT(10) type VARCHAR(50) description_score DOUBLE hend INT(11) meta_key VARCHAR(40) name VARCHAR(255) species_set_tag w class VARCHAR(50) species_set A SOURCE VARCHAR(255) score DOUBLE meta_value VARCHAR(255) species_set_id INT(10) species_set_id INT(10) evalue DOUBLE url VARCHAR(255) ag VARCHAR(50) genome_db_id INT(10) align_length INT(10) value MEDIUMTEXT identical_matches INT(10) perc_ident INT(10) positive_matches INT(10) mapping session homology_id INT(10) perc pos INT(10) mapping_session_id INT stable_id VARCHAR(40) hit_rank INT(10) type ENUM(...) ncbi taxa name method_link_species_set_id INT(10) cigar_line MEDIUMTEXT when_mapped TIMESTAMP taxon_id INT(10) odescription ENUM(...) genome_db rel from INT name VARCHAR(255) subtype VARCHAR(40) genome db id INT(10) rel_to INT on FLOAT name_class VARCHAR(50) taxon id INT(10) ds FLOAT name VARCHAR(40) n FLOAT assembly VARCHAR(100) s FLOAT assembly_default TINYINT(1) genebuild VARCHAR(100) threshold_on_ds FLOAT locator VARCHAR(400) ancestor node id INT(10) ncbi_taxa_node tree_node_id INT(10) taxon_id INT(10) parent_id INT(10) sitewise_aln rank CHAR(32) stable_id_history protein_tree_member_score sitewise_id INT(10) ogenbank hidden flag node_id INT(10) mapping_session_id INT aln_position INT(10) left_index INT(10) root_id INT(10) rstable_id_from VARCHAR(40) homology_id INT(10) node_id INT(10) right_index INT(10) version_from INT member_id INT(10) member_id INT(10) tree_node_id INT(10) oroot_id INT(10) method_link_species_set_id INT(10) * stable_id_to VARCHAR(40) peptide_member_id INT(10) omega FLOAT cigar_line MEDIUMTEXT version to INT member id INT(10) peptide_align_feature_id INT(10) omega_lower FLOAT contribution FLOAT stable_id VARCHAR(128) cigar_start INT(10) maily_member cigar line MEDIUMTEXT omega_upper FLOAT cigar_end INT(10) version INT(10) family_id INT(10) cigar start INT(10) optimal FLOAT source_name ENUM(...) member id INT(10) cigar_end INT(10) ncod INT(10) taxon id INT(10) cigar line MEDIUMTEXT perc_cov INT(10) threshold_on_branch_ds FLOAT genome_db_id INT(10) perc_id INT(10) type ENUM(...) protein_tree_stable_id > sequence_id INT(10) perc_pos INT(10) node_id INT(10) gene_member_id INT(10) subset_member v stable_id VARCHAR(40) description TEXT subset_id INT(10) version INT chr_name CHAR(40) member_id INT(10) chr_start INT(10) chr end INT(10) chr strand TINYINT(1) display_label VARCHAR(128) protein_tree_node node id INT(10) protein tree member narent id INT(10) node id INT(10) oroot_id INT(10) oroot_id INT(10) protein tree tag member_id INT(10) clusterset_id INT(10) node_id INT(10) left_index INT(10) method_link_species_set_id INT(10) ag VARCHAR(50) right_index INT(10) cigar_line MEDIUMTEXT value MEDIUMTEXT distance to parent DOUBLE cigar_start INT(10) cigar_end INT(10) sequence_exon_bounded sequence_cds sequence sequence_exon_bounded_id INT(10) subset sequence_cds_id INT(10) sequence_id INT(10) subset_id INT(10) lenath INT(10) member_id INT(10) member_id INT(10) super_protein_tree_tag description VARCHAR(255) length INT(10) length INT(10) sequence LONGTEXT super_protein_tree_member super_protein_tree_node sequence_exon_bounded LONGTEXT dump_loc VARCHAR(255) sequence_cds LONGTEXT nc_tree_node i nc_tree_tag ▼ nc_tree_member v







peptide_align_feature

The Compara Perl API

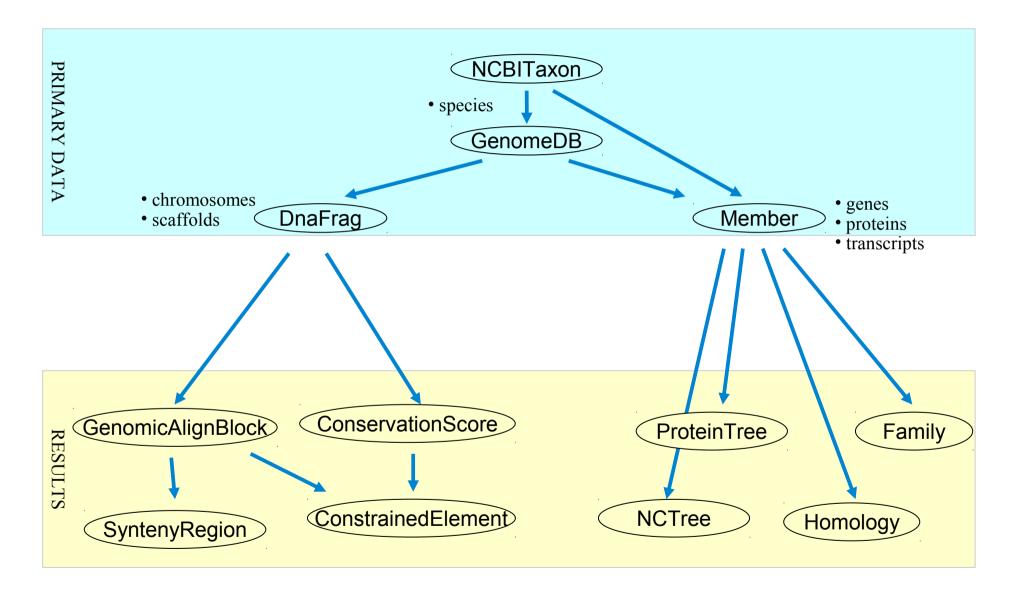
- Written in Object-Oriented Perl
- Used to retrieve data from and store data into ensemblcompara database (only the production pipeline generates the alignments)
- Links species together for Ensembl website
- Generalized to extend to non-Ensembl genomic data (Uniprot)
- Follows same 'Data Object' & '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	\$genomedb->name() \$taxon->genus()
Ensembl properties	<pre>\$genomedb->assembly() \$genomedb->genebuild()</pre>
Taxon object	\$genomedb->taxon()







Example Compara script (with GenomeDB)

```
use strict;
use Bio::EnsEMBL::Registry;
my $reg = "Bio::EnsEMBL::Registry";
$req->load registry from db(
    -host => "ensembldb.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: ", $thīs genomedb->short name;
  print ", assembly: ", \$this \overline{genomedb}->assembl\overline{y}, "\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	<pre>\$member->chr_start() \$member->stable_id() \$member->sequence()</pre>
Species	<pre>\$member->taxon()</pre>
Linked members	<pre>\$member->get_Translation() \$member->get_all_peptide_Members()</pre>







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

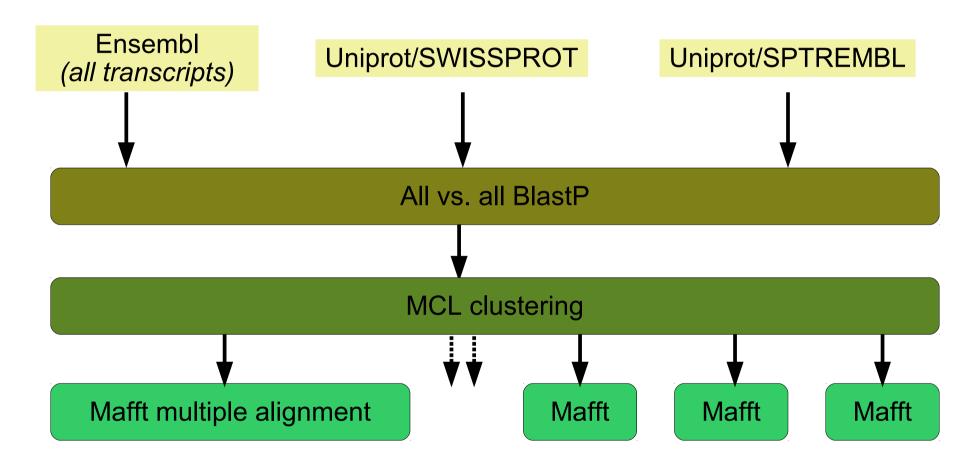






Families

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	<pre>\$family->get_SimpleAlign()</pre>
Biological function	\$family->description()
Gene content	<pre>\$family->get_all_Members() \$family->get_all_Member_Attribute()</pre>







Exercises – Families

Get the family predicted for the human gene(s) BRCA2

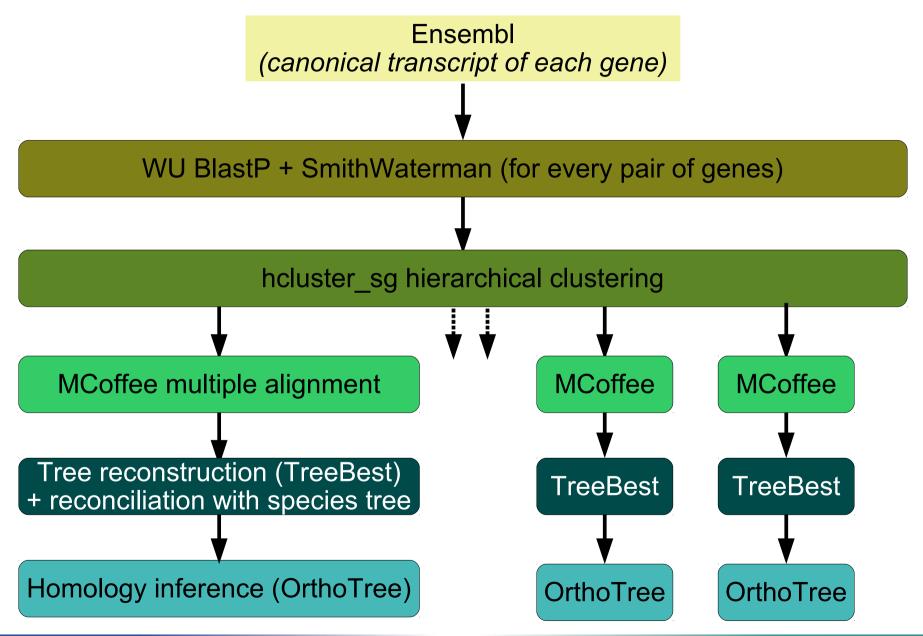
 Get the alignments corresponding to the family of the human gene(s) HBEGF







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	<pre>\$member->get_canonical_peptide_Member()</pre>
Alignment	<pre>\$proteintree->get_SimpleAlign()</pre>
Children (in the tree structure)	<pre>\$proteintree->children() \$proteintree->get_all_leaves()</pre>
Tree export	<pre>\$proteintree->newick_simple_format() \$proteintree->print_tree()</pre>







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)

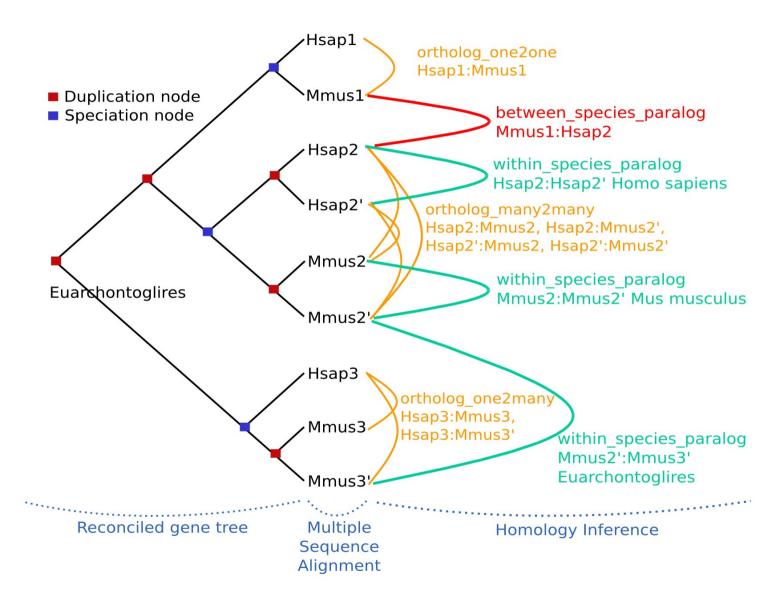
Print all the members of the tree containing the human gene
 CTDP1 (by recursively going into the tree structure)







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







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





