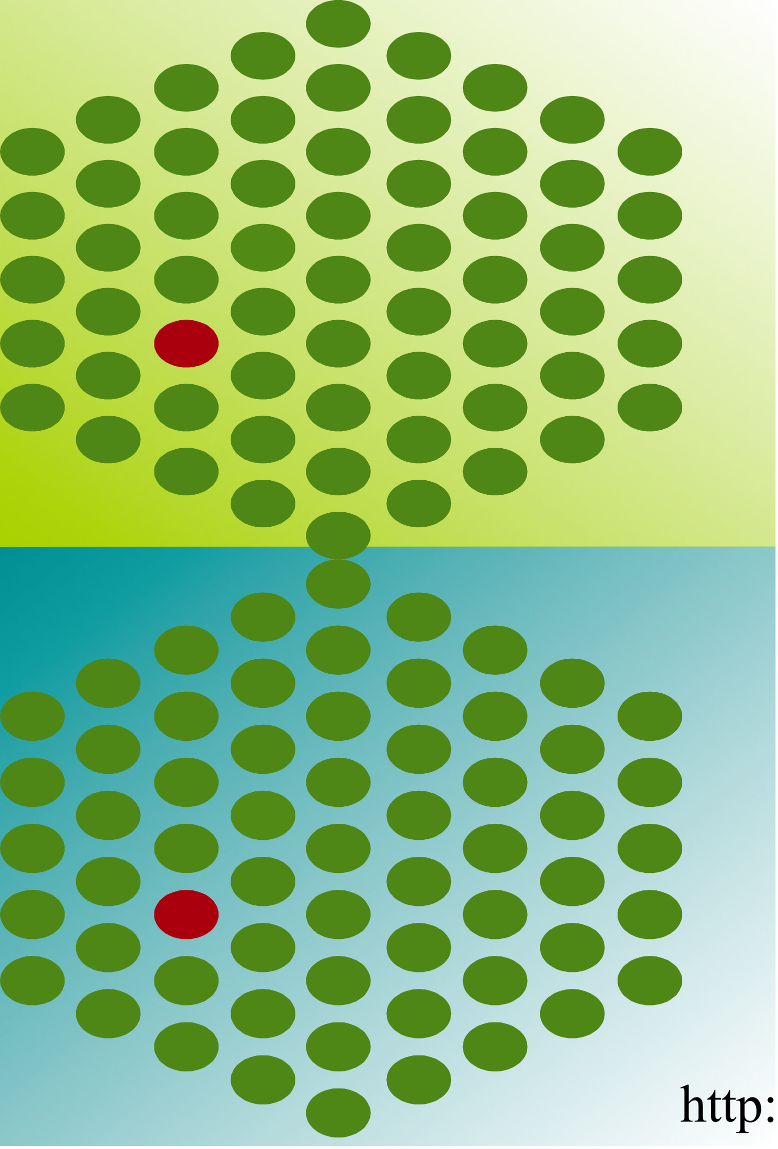


Ensembl Compara Perl API

Stephen Fitzgerald

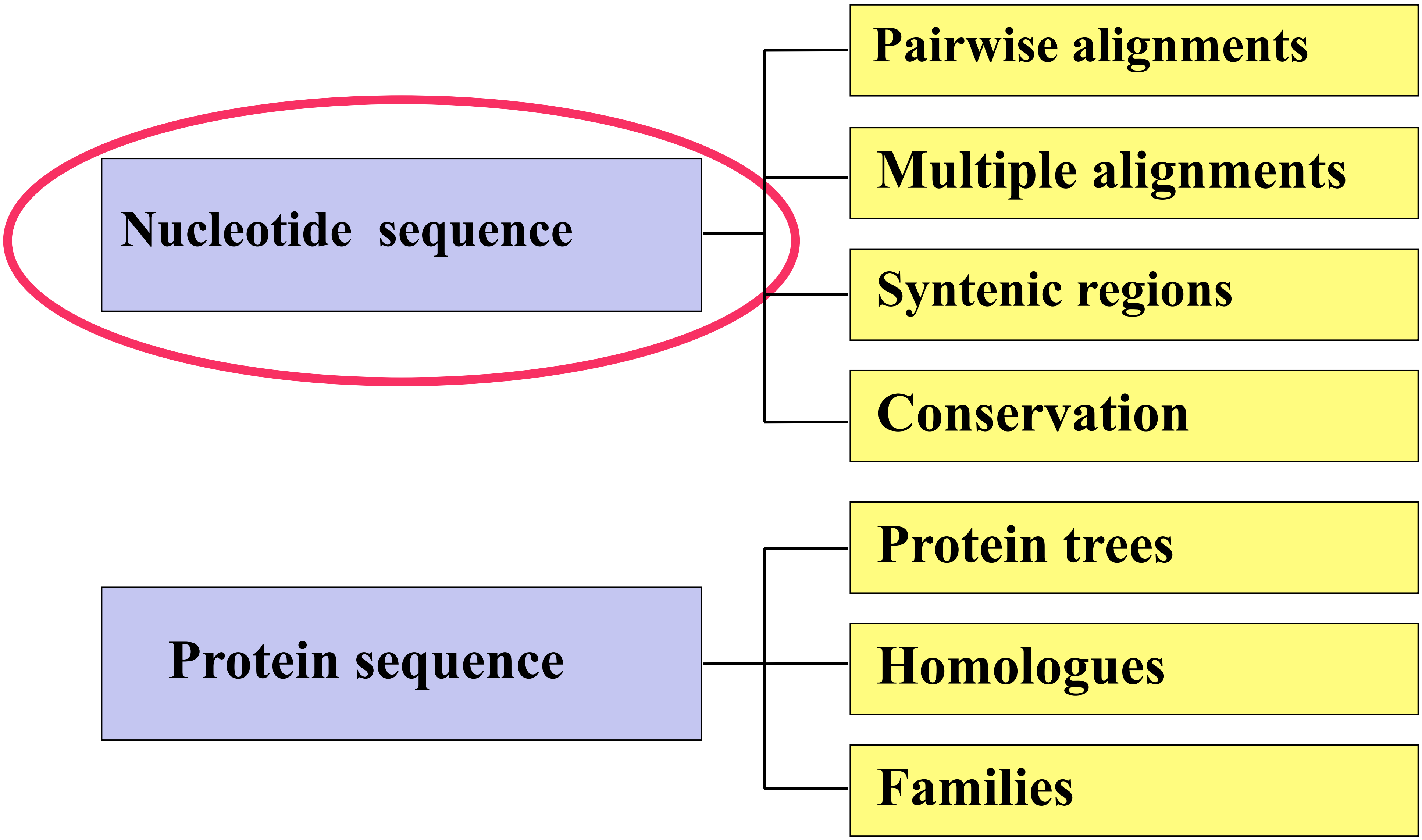
EBI - Wellcome Trust Genome Campus, UK

http://www.ebi.ac.uk/~stephenf/Workshops/EBI_may_2013/

EMBL-EBI

wellcome trust sanger institute

Sequence types and outputs



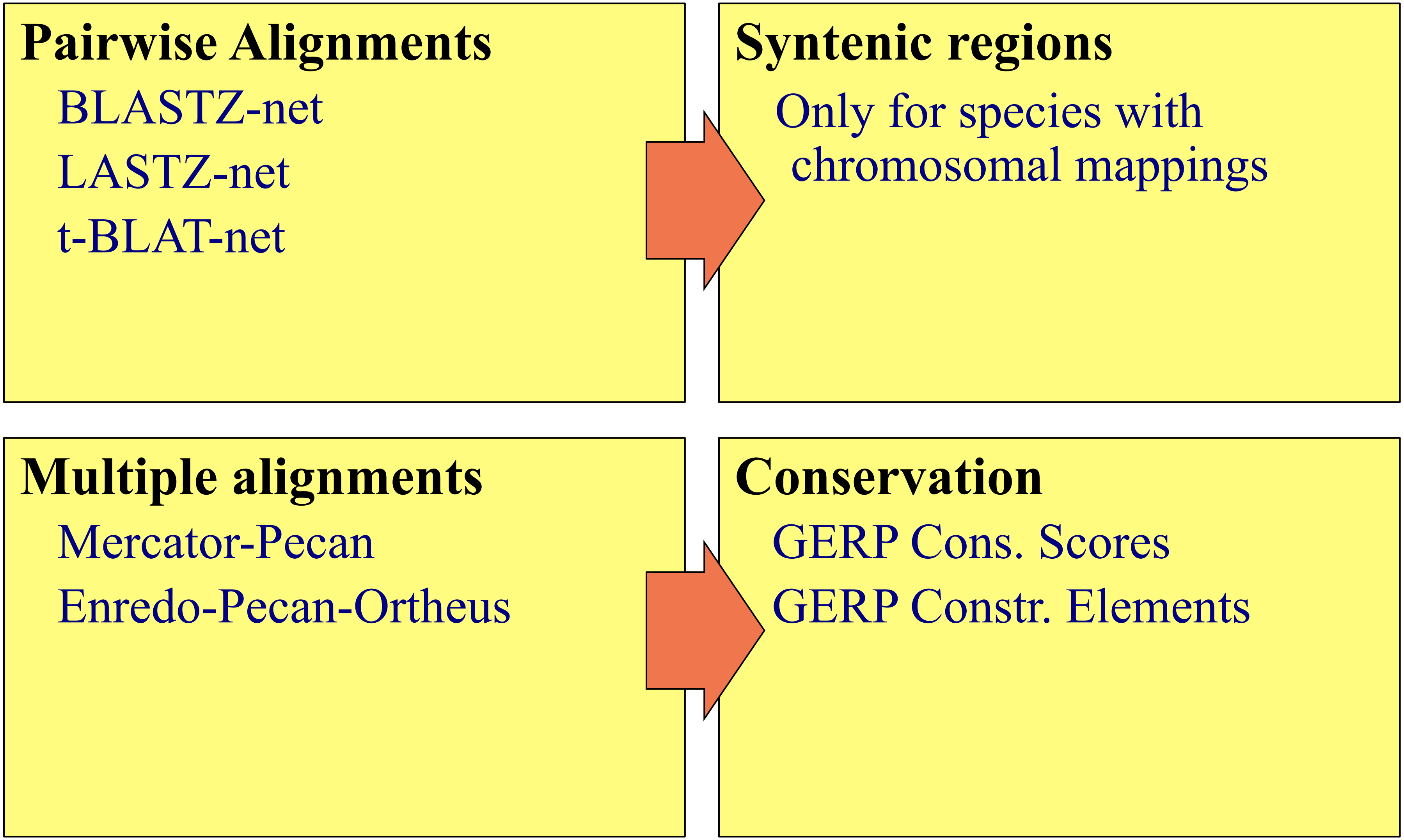
Nucleotide sequence

- Pairwise alignments
- Multiple alignments
- Syntenic regions
- Conservation

Protein sequence

- Protein trees
- Homologues
- Families

Nucleotide sequence analyses



Pairwise Alignments

BLASTZ-net

LASTZ-net

t-BLAT-net

Syntenic regions

Only for species with chromosomal mappings

Multiple alignments

Mercator-Pecan

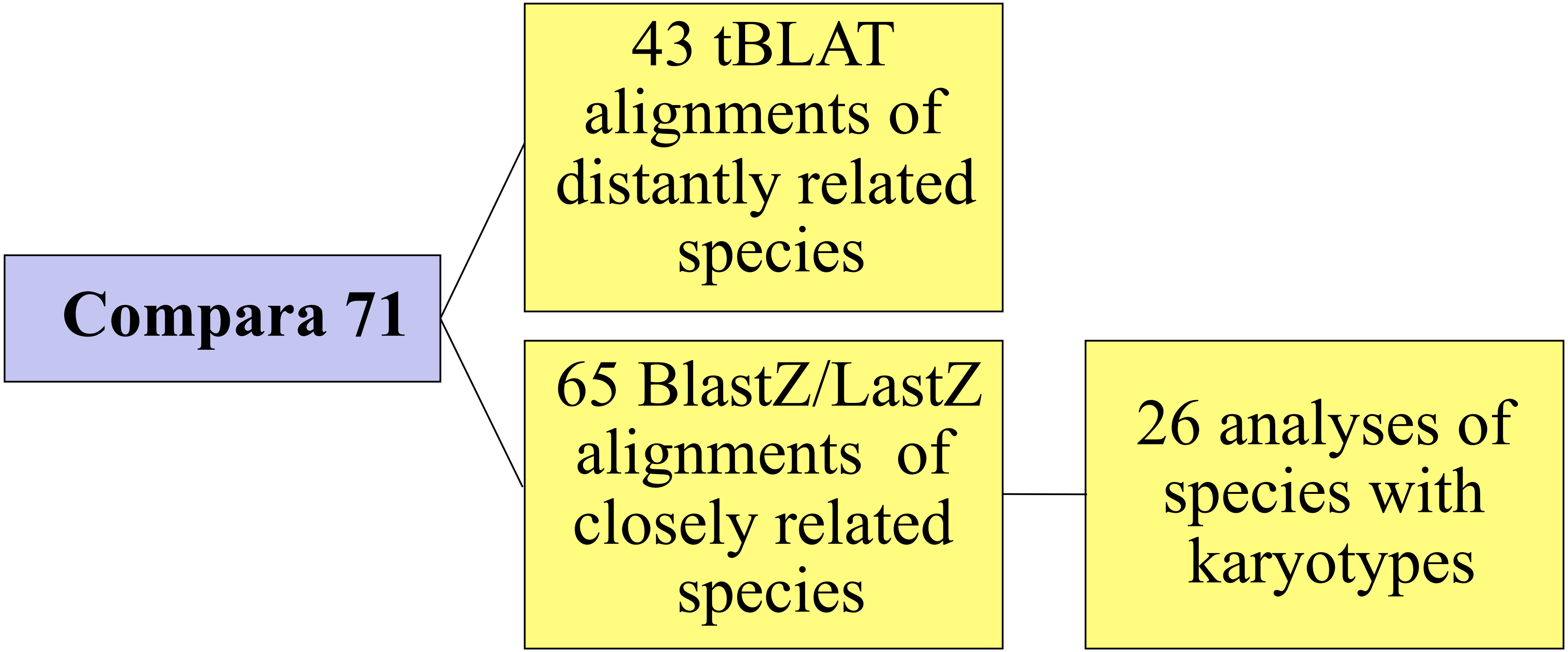
Enredo-Pecan-Ortheus

Conservation

GERP Cons. Scores

GERP Constr. Elements

Pairwise Alignments in Compara 71



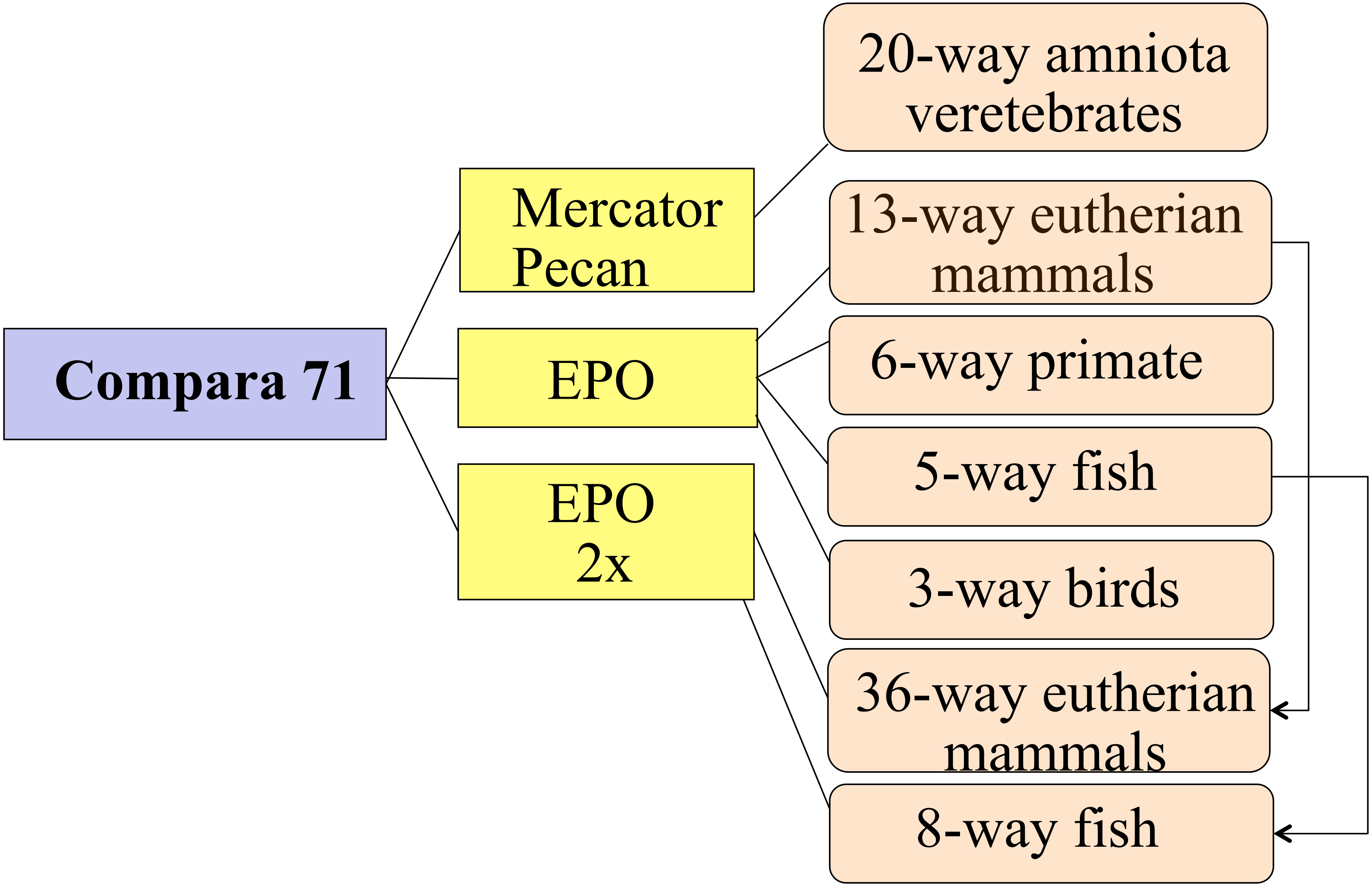
Compara 71

43 tBLAT alignments of distantly related species

65 BlastZ/LastZ alignments of closely related species

26 analyses of species with karyotypes

Multiple Alignments in Compara 71



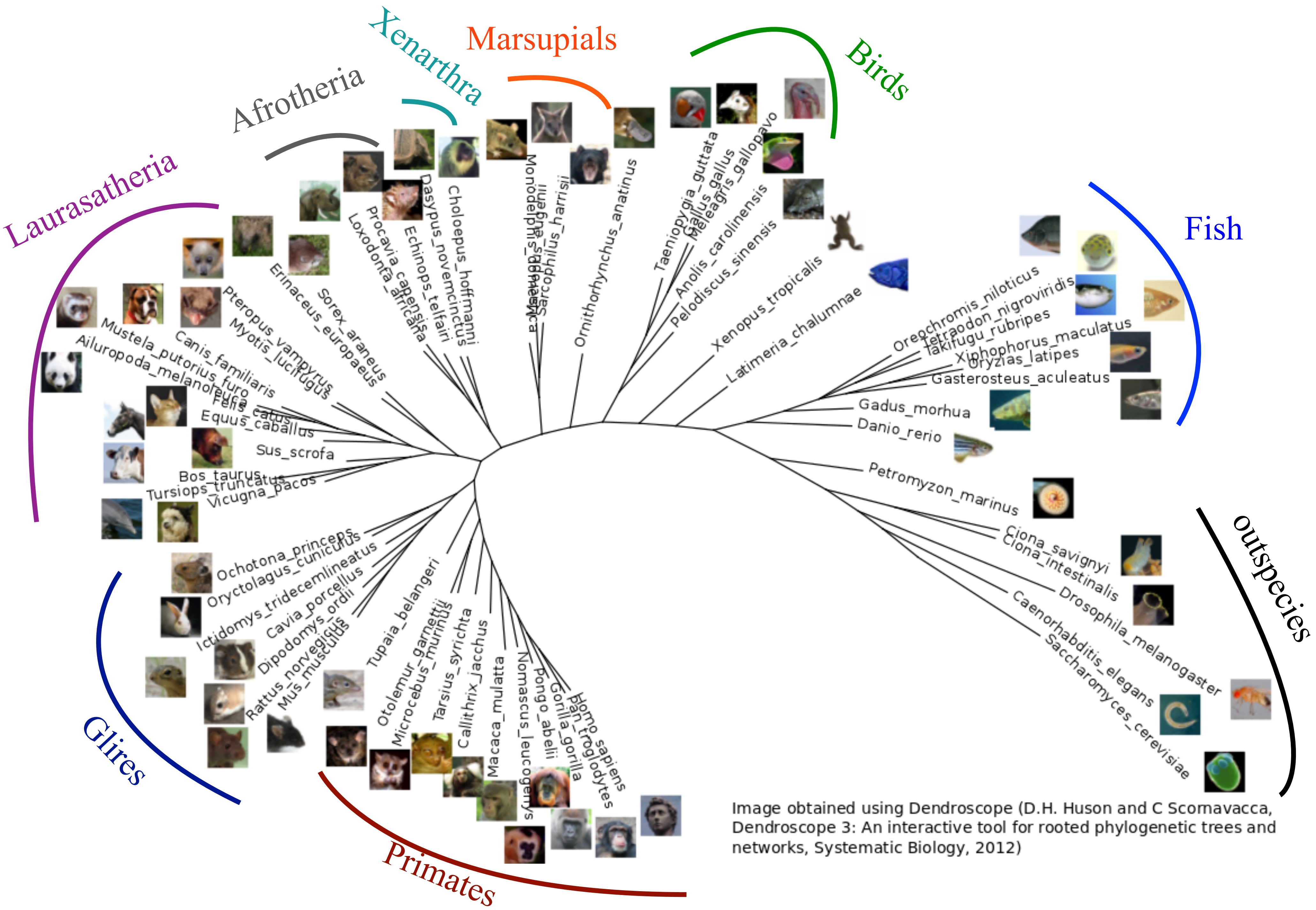
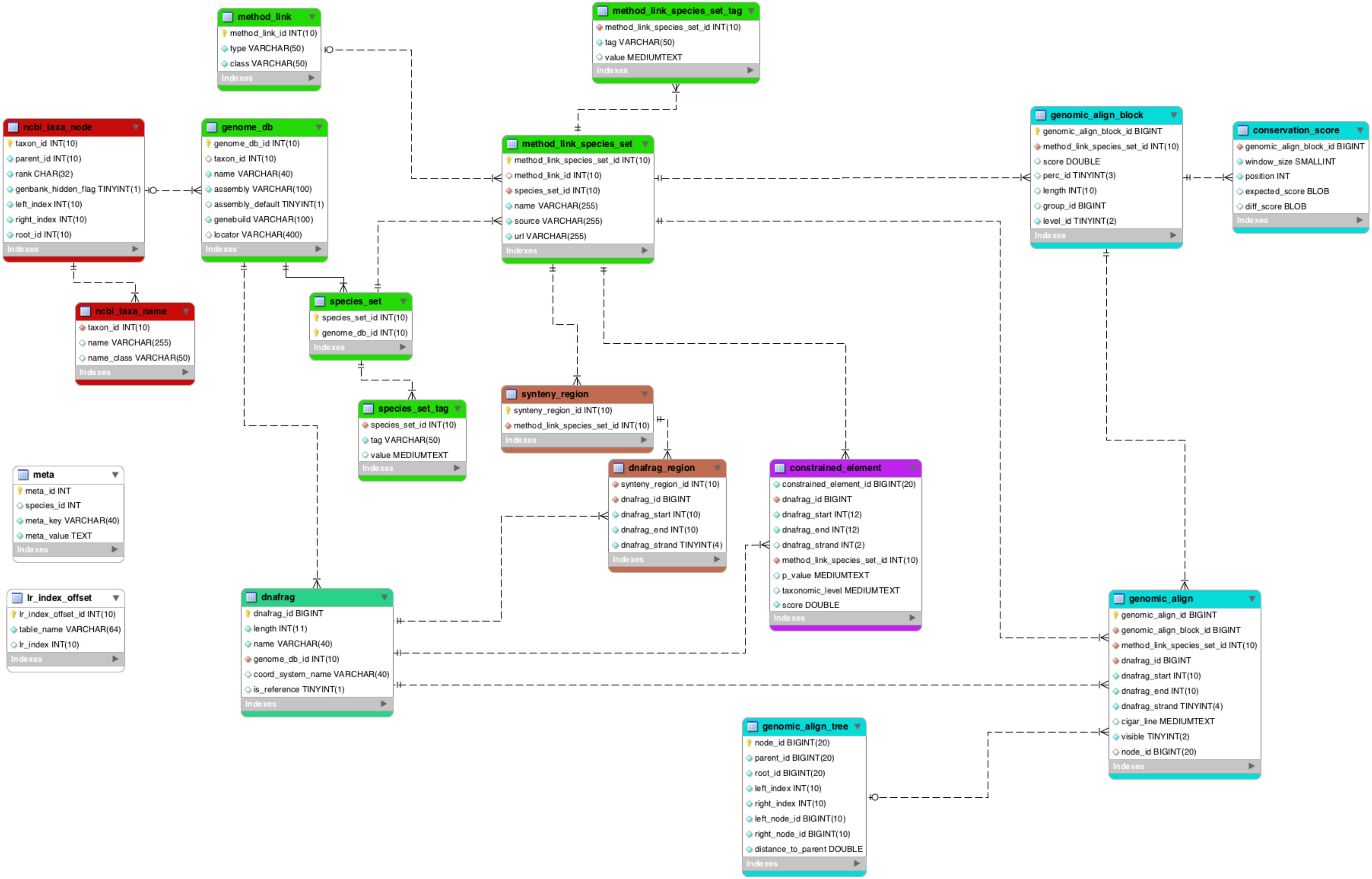


Image obtained using Dendroscope (D.H. Huson and C Scornavacca, Dendroscope 3: An interactive tool for rooted phylogenetic trees and networks, Systematic Biology, 2012)



method_link
method_link_id INT(10)
type VARCHAR(50)
class VARCHAR(50)
Indexes

method_link_species_set_tag
method_link_species_set_id INT(10)
tag VARCHAR(50)
value MEDIUMTEXT
Indexes

ncbi_taxa_node
taxon_id INT(10)
parent_id INT(10)
rank CHAR(32)
genbank_hidden_flag TINYINT(1)
left_index INT(10)
right_index INT(10)
root_id INT(10)
Indexes

genome_db
genome_db_id INT(10)
taxon_id INT(10)
name VARCHAR(40)
assembly VARCHAR(100)
assembly_default TINYINT(1)
genebuild VARCHAR(100)
locator VARCHAR(400)
Indexes

method_link_species_set
method_link_species_set_id INT(10)
method_link_id INT(10)
species_set_id INT(10)
name VARCHAR(255)
source VARCHAR(255)
url VARCHAR(255)
Indexes

genomic_align_block
genomic_align_block_id BIGINT
method_link_species_set_id INT(10)
score DOUBLE
perc_id TINYINT(3)
length INT(10)
group_id BIGINT
level_id TINYINT(2)
Indexes

conservation_score
genomic_align_block_id BIGINT
window_size SMALLINT
position INT
expected_score BLOB
diff_score BLOB
Indexes

ncbi_taxa_name
taxon_id INT(10)
name VARCHAR(255)
name_class VARCHAR(50)
Indexes

species_set
species_set_id INT(10)
genome_db_id INT(10)
Indexes

species_set_tag
species_set_id INT(10)
tag VARCHAR(50)
value MEDIUMTEXT
Indexes

synteny_region
synteny_region_id INT(10)
method_link_species_set_id INT(10)
Indexes

dnafrag_region
synteny_region_id INT(10)
dnafrag_id BIGINT
dnafrag_start INT(10)
dnafrag_end INT(10)
dnafrag_strand TINYINT(4)
Indexes

constrained_element
constrained_element_id BIGINT(20)
dnafrag_id BIGINT
dnafrag_start INT(12)
dnafrag_end INT(12)
dnafrag_strand INT(2)
method_link_species_set_id INT(10)
p_value MEDIUMTEXT
taxonomic_level MEDIUMTEXT
score DOUBLE
Indexes

meta
meta_id INT
species_id INT
meta_key VARCHAR(40)
meta_value TEXT
Indexes

lr_index_offset
lr_index_offset_id INT(10)
table_name VARCHAR(64)
lr_index INT(10)
Indexes

dnafrag
dnafrag_id BIGINT
length INT(11)
name VARCHAR(40)
genome_db_id INT(10)
coord_system_name VARCHAR(40)
is_reference TINYINT(1)
Indexes

genomic_align
genomic_align_id BIGINT
genomic_align_block_id BIGINT
method_link_species_set_id INT(10)
dnafrag_id BIGINT
dnafrag_start INT(10)
dnafrag_end INT(10)
dnafrag_strand TINYINT(4)
cigar_line MEDIUMTEXT
visible TINYINT(2)
node_id BIGINT(20)
Indexes

genomic_align_tree
node_id BIGINT(20)
parent_id BIGINT(20)
root_id BIGINT(20)
left_index INT(10)
right_index INT(10)
left_node_id BIGINT(10)
right_node_id BIGINT(10)
distance_to_parent DOUBLE
Indexes

Genomic alignments schema

Exercises – GenomeDB and DnaFrag

- A GenomeDB is used to link the Compara database to each of the Core species databases.

➢ Print the name, assembly version and genebuild version for all the GenomeDBs in the compara db

- A DnaFrag represents a top-level SeqRegion in the Compara database.

➢ Print all the DnaFrags for chimp

Exercises – MethodLinkSpeciesSet

- The MethodLinkSpeciesSet is a central component in the Compara database, it stores information connecting the various analyses (method_link_type) with a set of species (species_set).

➢ Print the total number of MethodLinkSpeciesSet entries stored in the database.

➢ Print a unique list of method_link_types and a count of their number in the database.

➢ Print a list of the species and their internal ids (dbIDs) for the 12 eutherian mammal EPO alignments

Compara database is coupled to Ensembl core databases

Compara stores relationships between the genomes by loading references or ‘handles’ to external data.

Since there is minimal primary data inside Compara, to gain full access to the data these external links must be re-established

Example: compara_70 must be linked with the Ensembl core_70 databases

Proper REGISTRY configuration is needed

Alignments are stored in the genomic_align and genomic_align_block tables

A small example :

```
gorilla_gorilla/MT/935-953   gacat-ttaactaaaac-ccc
macaca_mulatta/MT/1469-1488  aacatcttaactaaacg-ccc
pan_troglodytes/MT/934-953   gatac-ttaacttaaaccccc
pongo_pygmaeus/MT/940-958    actac-ctaactaaaac-ccc
homo_sapiens/MT/1516-1534    gacat-ttaactaaaac-ccc
                                *   ***** **   ***
```

```
GACATTTAACTAAAACCCC     5MD11MD3M
AACATCTTAACTAAACGCCC    17MD3M
GATACTTAACTTAAACCCCC    5MD15M
ACTACCTAACTAAAACCCC     5MD11MD3M
GACATTTAACTAAAACCCC     5MD11MD3M
```

5 genomic_align entries
1 genomic_align_block

Sequences from core

Exercises - GenomicAlignBlock

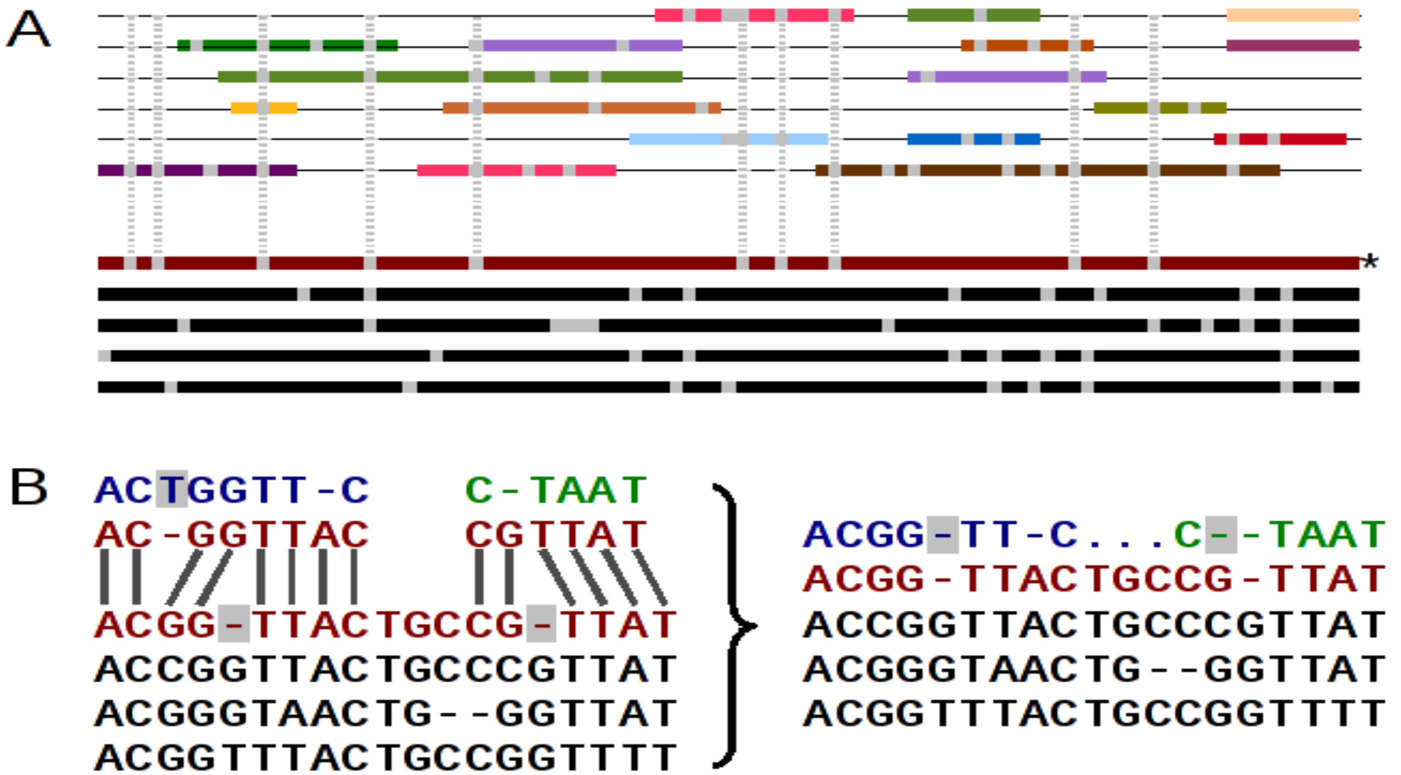
- A GenomicAlignBlock represents an alignment between two or more regions of genomic DNA. Within these blocks every region of genomic DNA is represented by a GenomicAlign object.

➢ Print the LASTZ-NET alignments for pig chromosome 15 with cow (using pig coordinates 105734307 and 105739335).

➢ Change the above example so that it prints the 13-way eutherian mammal (EPO) multiple alignments.

Adding low-coverage (2X) genomes

- Low coverage genomes cannot be fully assembled
- Resulting assembly is too scattered to be used with Enredo
- Run EPO on high-coverage genomes only
- Map 2X genomes using pairwise alignments



Exercises – GenomicAlignBlock (Constrained elements)

- A Constrained Elements represent regions in the multiple alignment which appear to be under functional constraint.

➢ Print the constrained element alignments from the above pig locus (use the constrained elements generated from the EPO_LOW_COVERAGE mammals alignments)

Exercises - Synteny

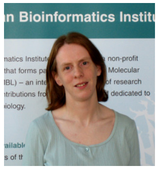
- Synteny blocks are derived from Lastz-net alignments
  - group syntenic alignments closer than 200Kb
  - link syntenic groups closer than 3Mb
  - minimum length of the syntenic block: 100 kb

- Print the pig-cow synteny map using pig chromosome 15 as a reference

Acknowledgements

Compara Team

Javier



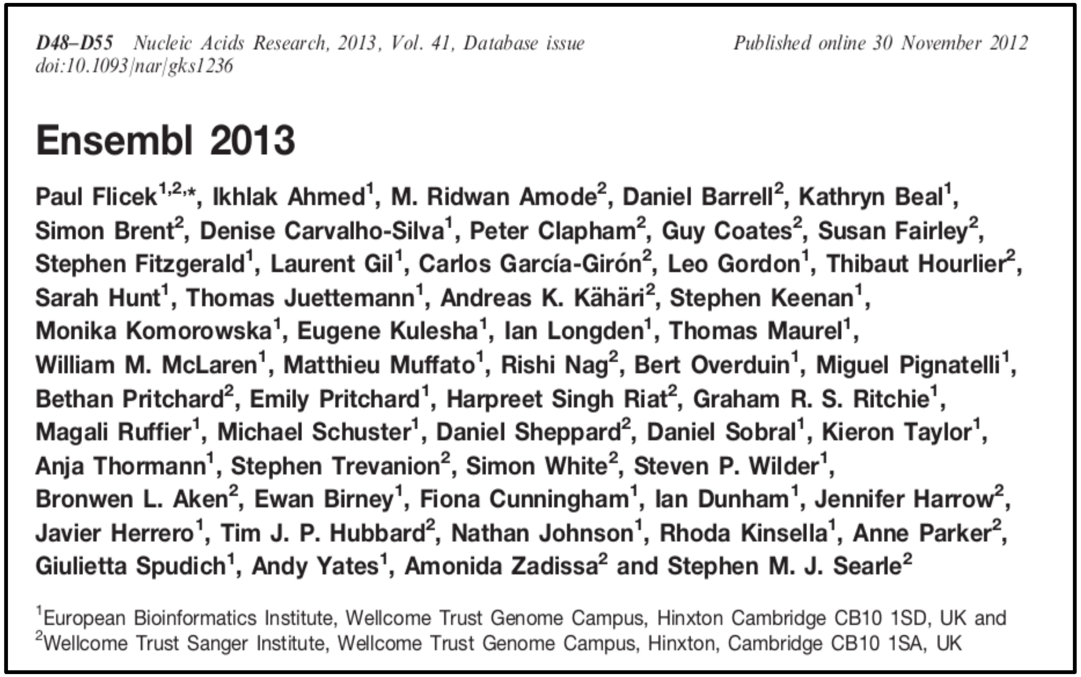
Kathryn

Stephen

Leo

Matthieu

Miguel



D48–D55 Nucleic Acids Research, 2013, Vol. 41, Database issue *Published online 30 November 2012*
doi:10.1093/nar/gks1236

Ensembl 2013

Paul Flicek[1,2,*], Ikhlak Ahmed[1], M. Ridwan Amode[2], Daniel Barrell[2], Kathryn Beal[1], Simon Brent[2], Denise Carvalho-Silva[1], Peter Clapham[2], Guy Coates[2], Susan Fairley[2], Stephen Fitzgerald[1], Laurent Gil[1], Carlos García-Girón[2], Leo Gordon[1], Thibaut Hourlier[2], Sarah Hunt[1], Thomas Juettemann[1], Andreas K. Kähäri[2], Stephen Keenan[1], Monika Komorowska[1], Eugene Kulesha[1], Ian Longden[1], Thomas Maurel[1], William M. McLaren[1], Matthieu Muffato[1], Rishi Nag[2], Bert Overduin[1], Miguel Pignatelli[1], Bethan Pritchard[2], Emily Pritchard[1], Harpreet Singh Riat[2], Graham R. S. Ritchie[1], Magali Ruffier[1], Michael Schuster[1], Daniel Sheppard[2], Daniel Sobral[1], Kieron Taylor[1], Anja Thormann[1], Stephen Trevanion[2], Simon White[2], Steven P. Wilder[1], Bronwen L. Aken[2], Ewan Birney[1], Fiona Cunningham[1], Ian Dunham[1], Jennifer Harrow[2], Javier Herrero[1], Tim J. P. Hubbard[2], Nathan Johnson[1], Rhoda Kinsella[1], Anne Parker[2], Giulietta Spudich[1], Andy Yates[1], Amonida Zadissa[2] and Stephen M. J. Searle[2]

[1]European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton Cambridge CB10 1SD, UK and [2]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK

wellcometrust

EMBL

National Human Genome Research Institute

BBSRC bioscience for the future

European Commission
Framework Programme 7

SEVENTH FRAMEWORK PROGRAMME

GEN2PHEN

BLUE PRINT epigenome

Quantomics
From Sequence to Consequence :
Tools for the Exploitation of Livestock Genomes