Comparative genomics in e!Ensembl

EMBL-EBI

Training materials

- Ensembl training materials are protected by a CC BY license:

  creativecommons.org/licenses/by/4.0/

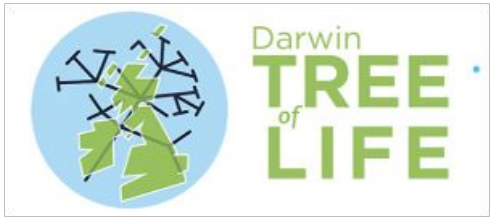
- If you wish to re-use these materials, please credit Ensembl for their creation
- If you use Ensembl for your work, please cite our papers:

  ensembl.org/info/about/publications.html

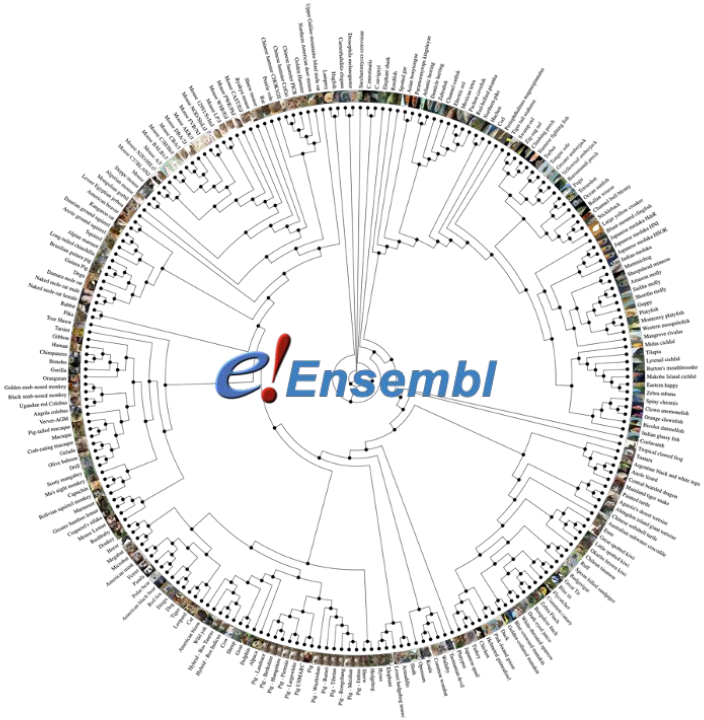
Applications

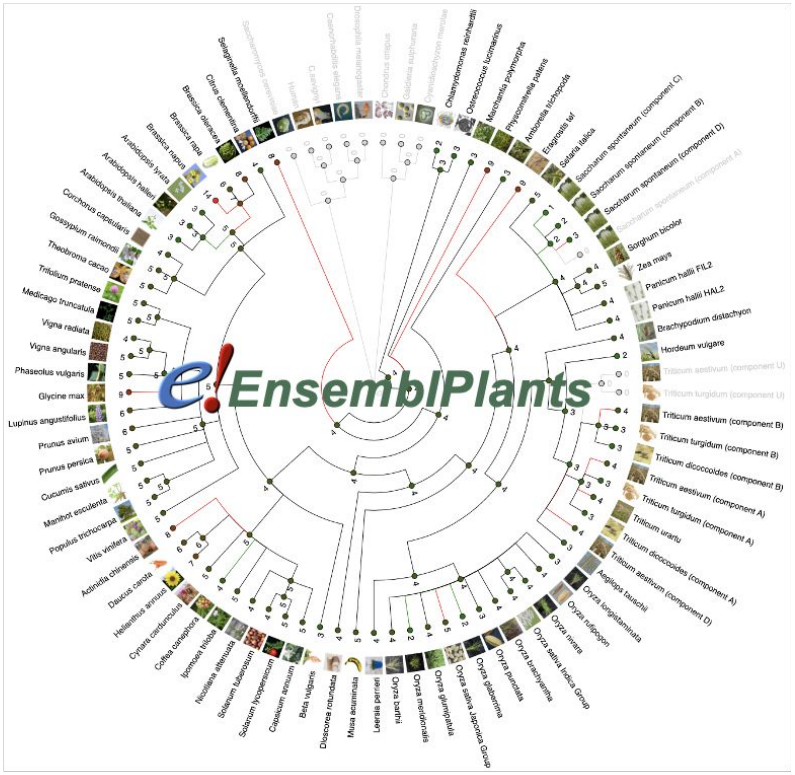
Comparative genomics allows us to:

- Examine evolutionary relationship
- Find differences between genomes
- Derive gene function based on homology
- Identify highly conserved regions



Comparative analysis by taxa

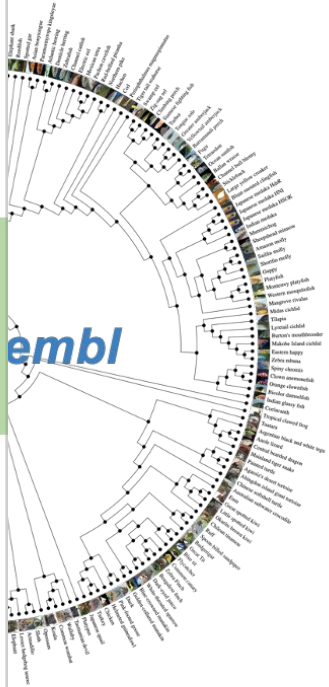


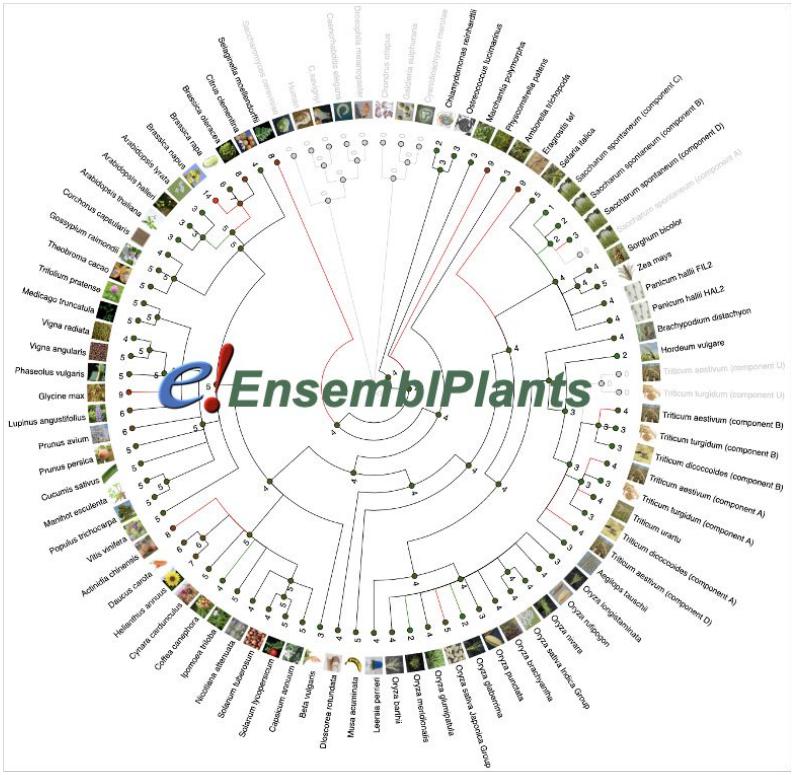


Comparative analysis by taxa

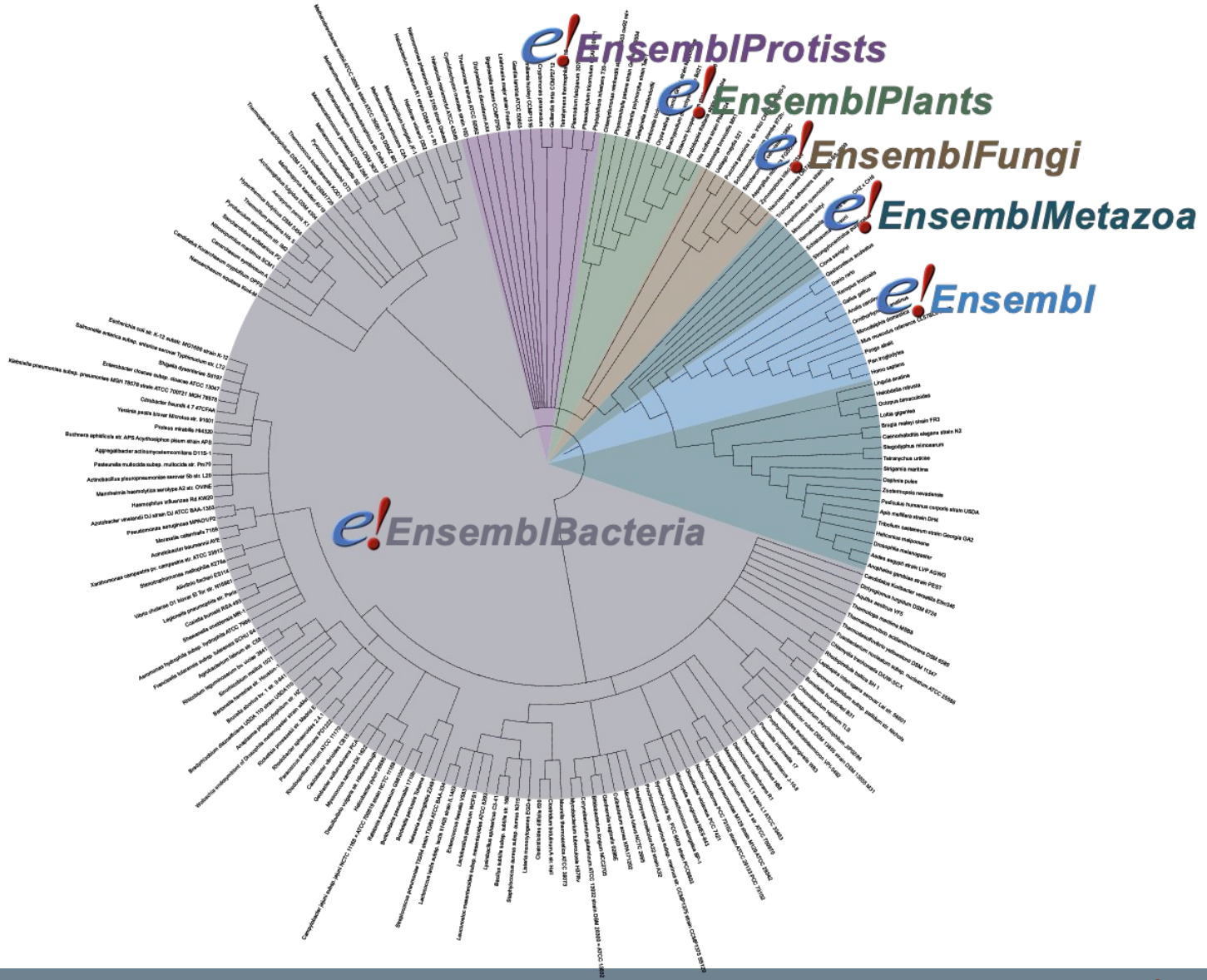
Gene-based displays

- **Summary**
  - Splice variants
  - Transcript comparison
  - Gene alleles
- **Sequence**
  - Secondary Structure
- Gene families
- Literature
- Plant Compara
  - Genomic alignments
  - Gene tree
  - Gene gain/loss tree
  - Orthologues
  - Paralogues
- Pan-taxonomic Compara
  - Gene Tree
  - Orthologues
- Ontologies
  - GO: Biological process
  - GO: Cellular component
  - GO: Molecular function
  - PO: Plant anatomical entity
  - PO: Plant structure development
- Phenotypes
- Genetic Variation
  - Variant table
  - Variant image
  - Structural variants
- Gene expression
- Pathway
- Regulation
- External references
- Supporting evidence
- ID History
  - Gene history





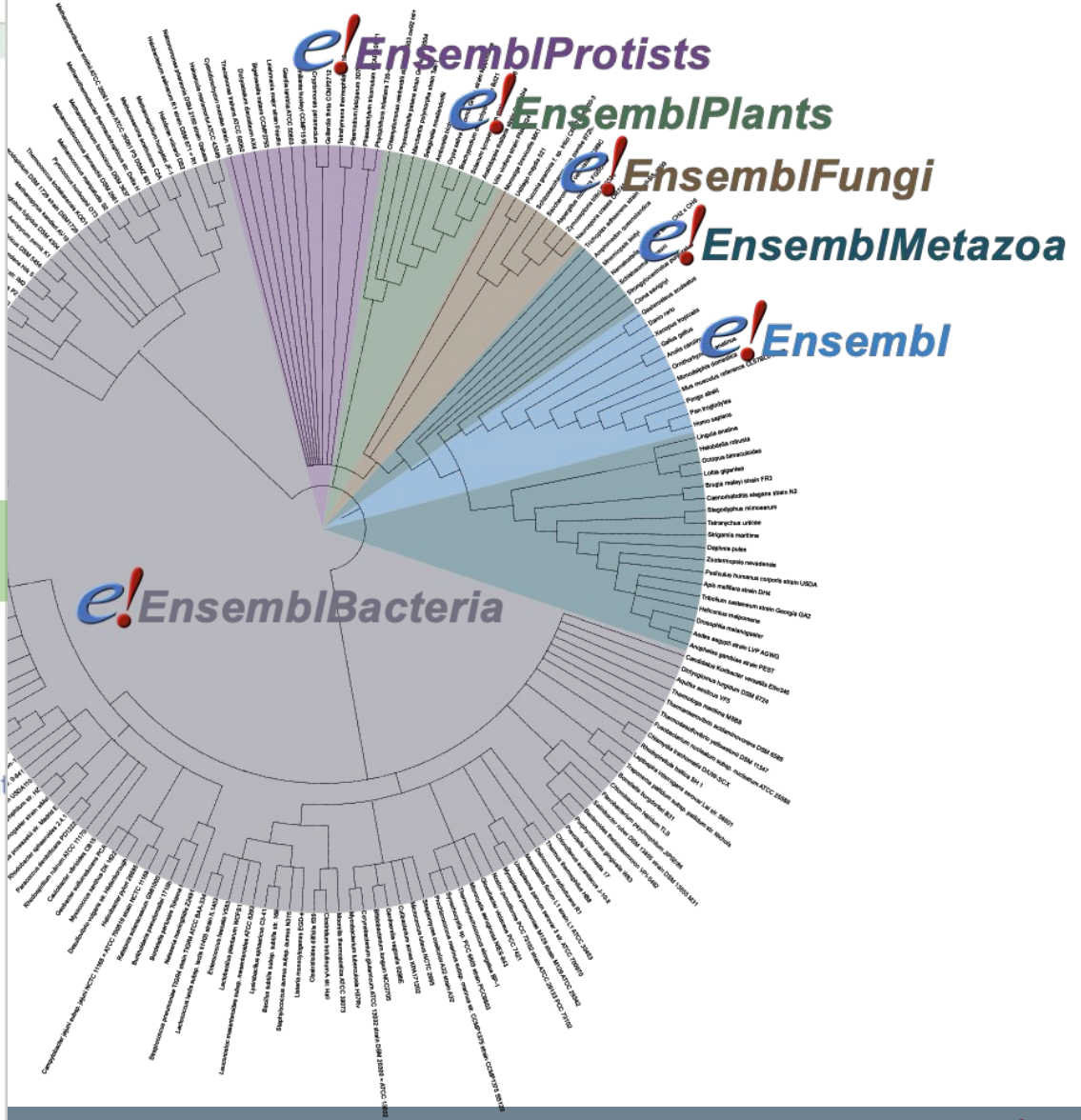
Pan-taxonomic compara



Pan-taxonomic compara

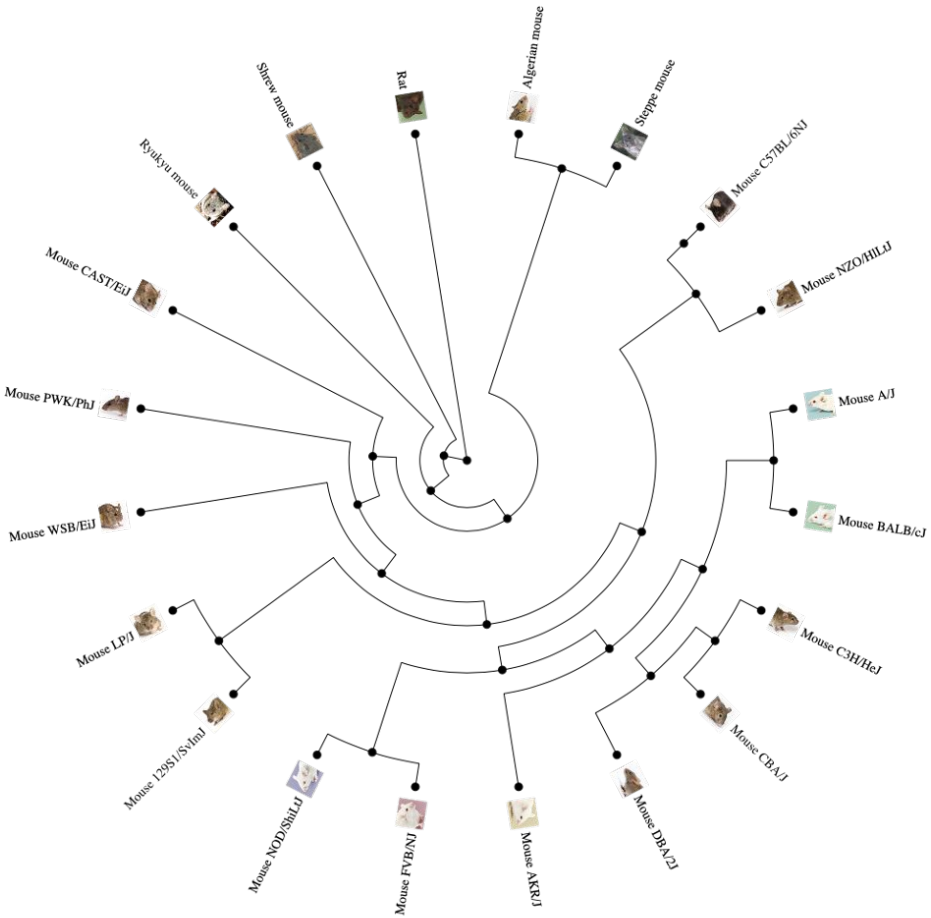
Gene-based displays

- **Summary**
  - Splice variants
  - Transcript comparison
  - Gene alleles
- Sequence
  - Secondary Structure
- Gene families
- Literature
- Plant Compara
  - Genomic alignments
  - Gene tree
  - Gene gain/loss tree
  - Orthologues
  - Paralogues
- Pan-taxonomic Compara
  - Gene Tree
  - Orthologues
- Ontologies
  - GO: Biological process
  - GO: Cellular component
  - GO: Molecular function
  - PO: Plant anatomical entity
  - PO: Plant structure development
- Phenotypes
- Genetic Variation
  - Variant table
  - Variant image
  - Structural variants
- Gene expression
- Pathway
- Regulation
- External references
- Supporting evidence
- ID History
  - Gene history



EMBL-EBI

Mouse compara



Types of data

Gene-based resources (found in the gene tab)

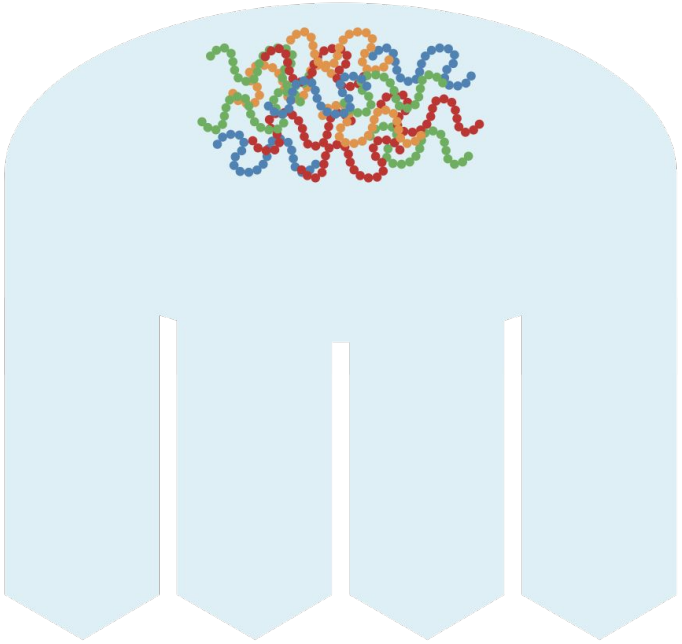
- Phylogenetic trees and tree-inferred homology
- Protein trees
- ncRNA trees
- Stable ID mapping
- Protein families

Sequence-based resources (found in the location tab)

- Whole genome alignments
- Ancestral sequences
- Age of base
- Conservation scores and constrained elements
- Syntenies

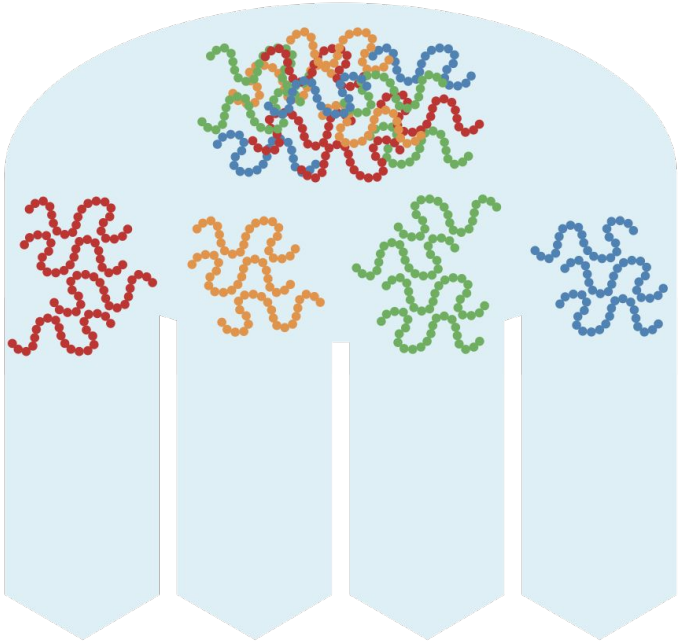
Gene/protein trees

1. Representative translation of each gene from all species



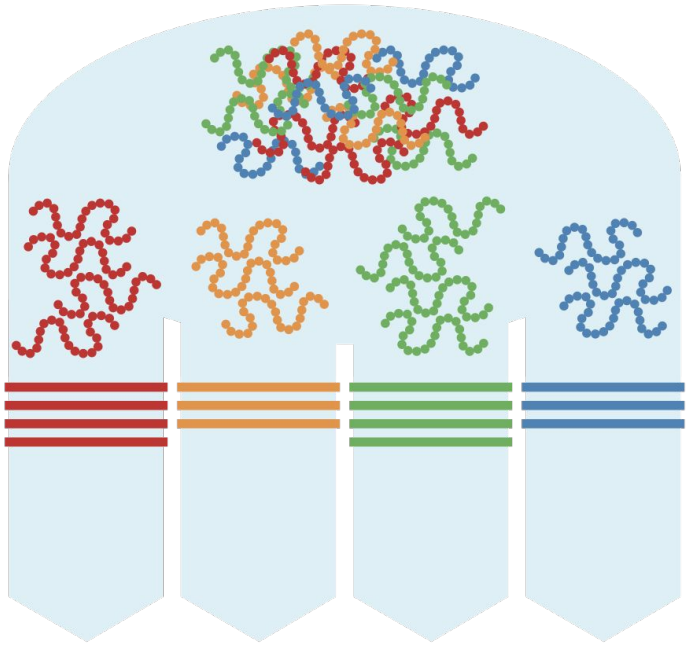
Gene/protein trees

1. Representative translation of each gene from all species
2. All-vs-all HMM search to classify into families or clustering



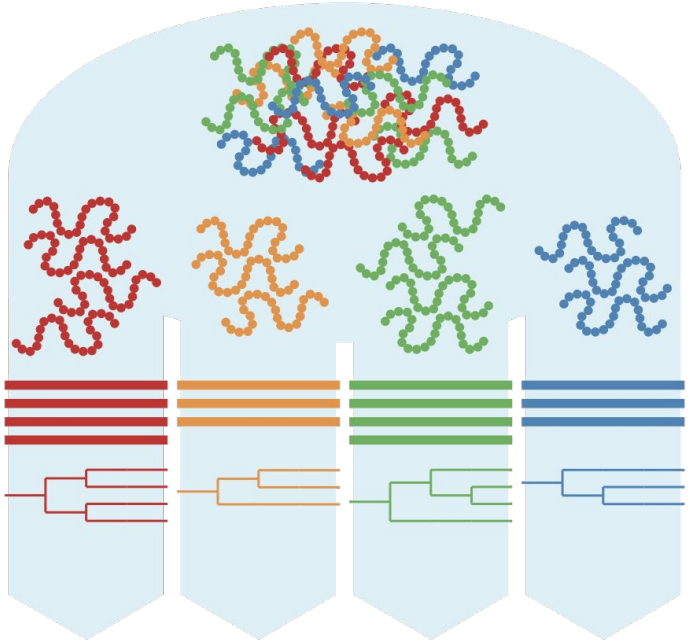
Gene/protein trees

1. Representative translation of each gene from all species
2. All-vs-all HMM search to classify into families or clustering
3. Multiple protein alignment



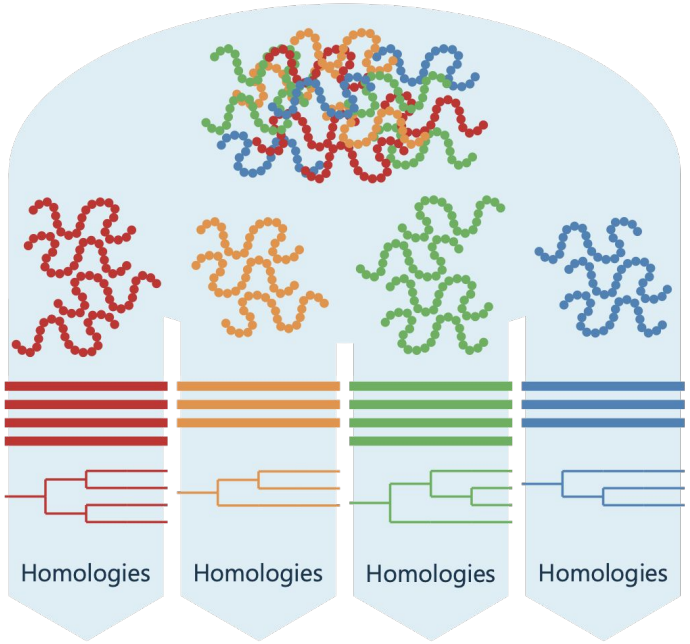
Gene/protein trees

1. Representative translation of each gene from all species
2. All-vs-all HMM search to classify into families or clustering
3. Multiple protein alignment
4. Phylogenetic tree for each aligned cluster and reconciliation against NCBI taxonomy



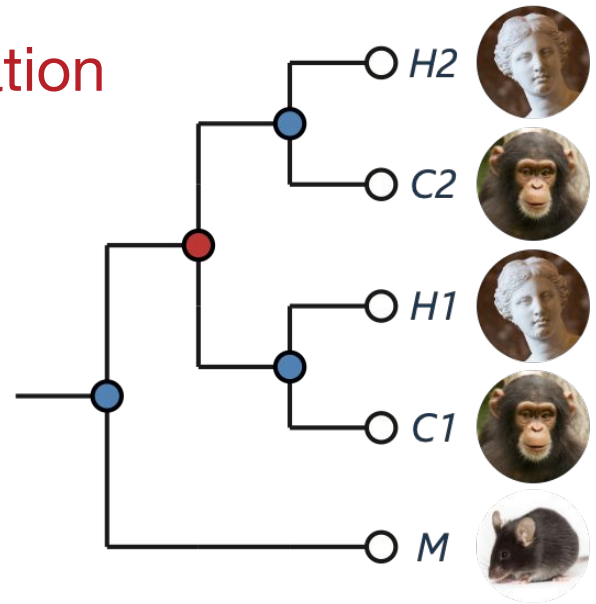
Gene/protein trees

1. Representative translation of each gene from all species
2. All-vs-all HMM search to classify into families or clustering
3. Multiple protein alignment
4. Phylogenetic tree for each aligned cluster and reconciliation against NCBI taxonomy
5. Ortho-/paralogue inference

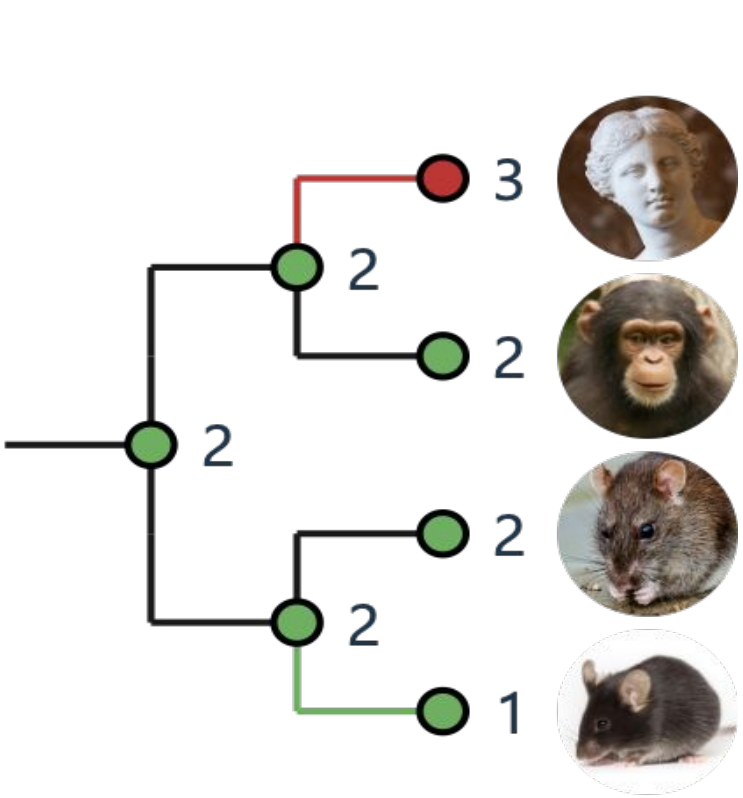


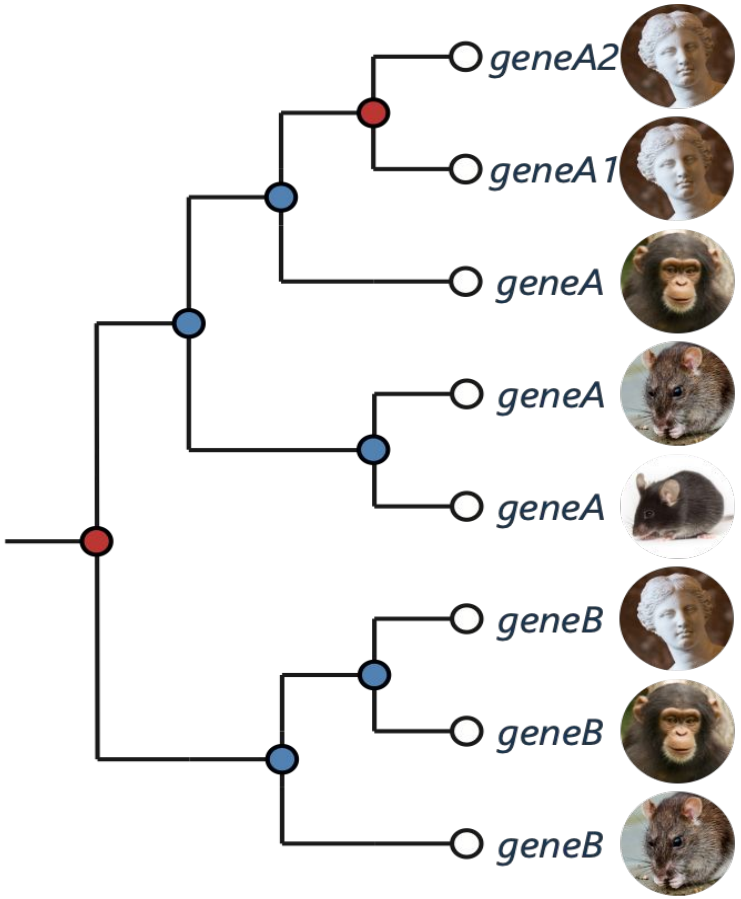
Homology relationships

- Orthologues
  - Genes emerged through a speciation event, e.g. *C1* and *H1*; *C2* and *M*; *H2* and *M*
  - 1-to-1: *C1* and *H1*
  - 1-to-many: *M* and *H1*, *H2*
- Paralogues
  - Genes emerged through a duplication event, e.g. *C1* and *C2*, *H1* and *H2*



Gene tree vs gene gain/loss tree





Whole genome alignments: pairwise vs multiple

- To identify highly conserved regions
  - Sequences that evolve slowly
  - Regions likely to be functional
  - Both coding and non-coding
- To support problematic gene predictions
- To define syntenic regions

Pairwise alignments

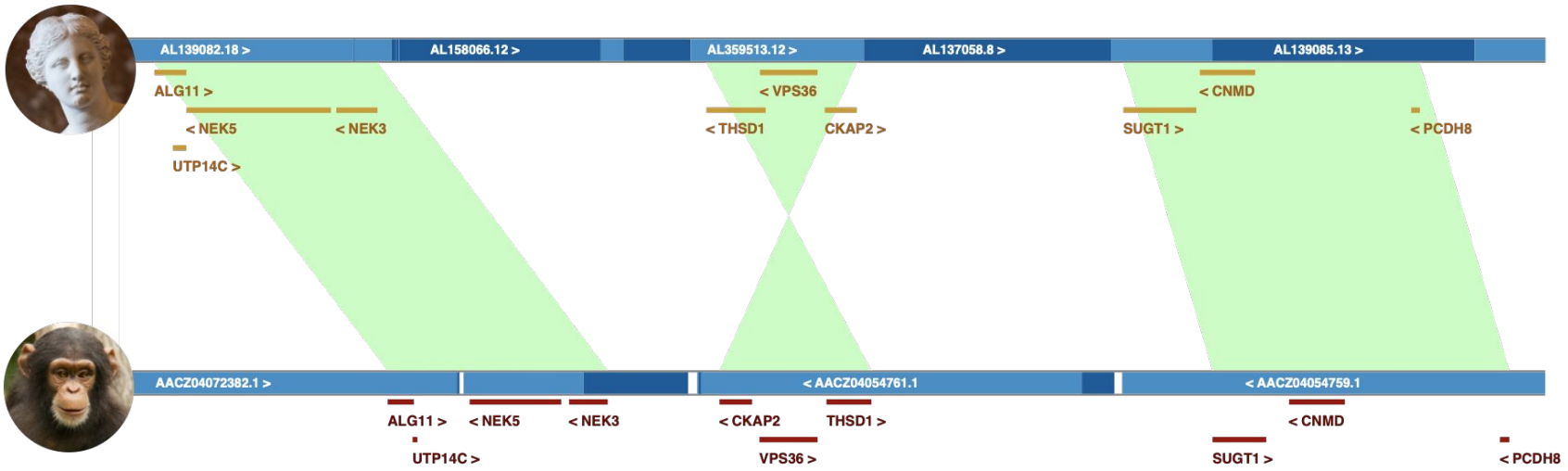
Pairwise alignments with BLASTZ (older) or LASTZ (newer):

- Human vs everything
- Model organisms vs related species
- Agricultural mammals vs each other




Shared synteny

Conserved order of aligned homologous genomic blocks between species (irrespective of orientation):



Multiple alignments

- EPO (Enredo-Pecan-Ortheus)

38 fish; 17 sauropsids; 46 eutherian mammals; 12 primates, 21 murinae

- EPO Extended (formerly “low-coverage”)

Allows fragmented assemblies

65 fish; 27 sauropsids; 99 eutherian mammals; 24 primates; 16 pig breeds and other agricultural mammals

- Mercator-Pecan

65 amniota vertebrates (mammals, birds, reptiles)

More information

Herrero et al.

Ensembl comparative genomics resources

Database: the Journal of Biological Databases and Curation (2016)

epmc.org/abstract/MED/26896847