Comparative genomics

Michelle M. Leger

Learning outcomes

- Terminology and basic concepts revision
- Familiarisation with constructing a phylogeny for gene family evolution analysis
- Brief familiarisation with Orthofinder output
- Familiarisation with profile Hidden Markov Model searches
- **Comparative genomics approaches should be tailored to the question at hand**
- **Software output should be treated critically rather than accepted at face value ("garbage in, garbage out")**

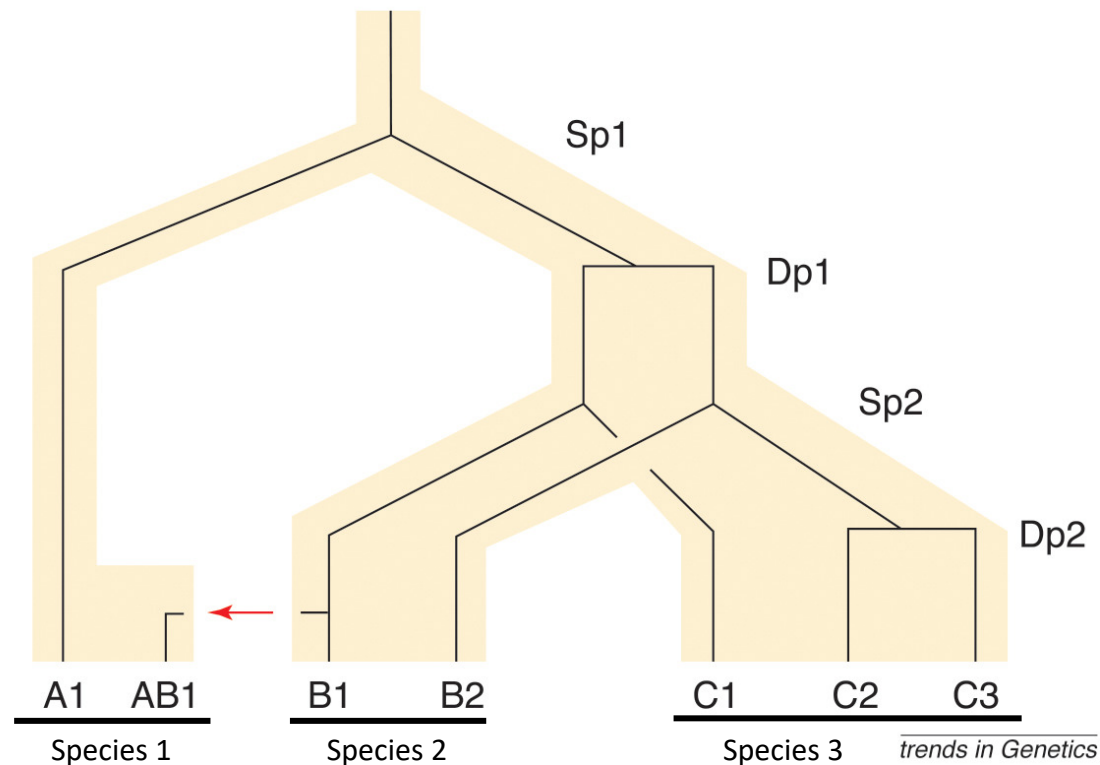# Infering evolutionary history and function

How did a sequence arise?

- Gene family evolution (gene duplication, loss, horizontal gene transfer)
- Synteny

What role is the protein playing in a given organism?

- Type of selection
- Domain architecture
- Functionally important residues
- Structure
- Regulatory patterns

Orthology

Fitch, 2000
Sp: speciation; Dp: duplication

- Homology: descent from a common ancestor
- Orthology: sequence divergence follows speciation
- Paralogy: sequence divergence follows duplication
- Xenology: sequence divergence follows lateral transfer
- Orthogroup (or hierarchical ortholog group): all genes descended from a single gene in the last common ancestor
- Strict ortholog group: set of genes which are all orthologs of each other
- The relationship between B1 and C1 is orthologous
- The relationship between A1 and any one of B1, B2, C1, C2 and C3 is orthologous
- The relationship between B1 and B2 is paralogous
- The relationship between C1, C2 and C3 is paralogous
- AB1 is a xenolog of all six other genes
- All of the genes shown here form an orthogroup

Sequence similarity searches – Basic Local Alignment Search Tool

- Identify short sequences (called words) identical between query and subject

- expand alignments from the regions that match these words

- generate a score for the resulting alignments, using a matrix of expected amino acid substitution rates

- Many faster or more sensitive variations
    - BLAT
    - PLAST
    - DIAMOND
    - DELTA-BLAST

Gene-centric assembly

- PhyloMagnet
- MEGAN
- GRASP2

...
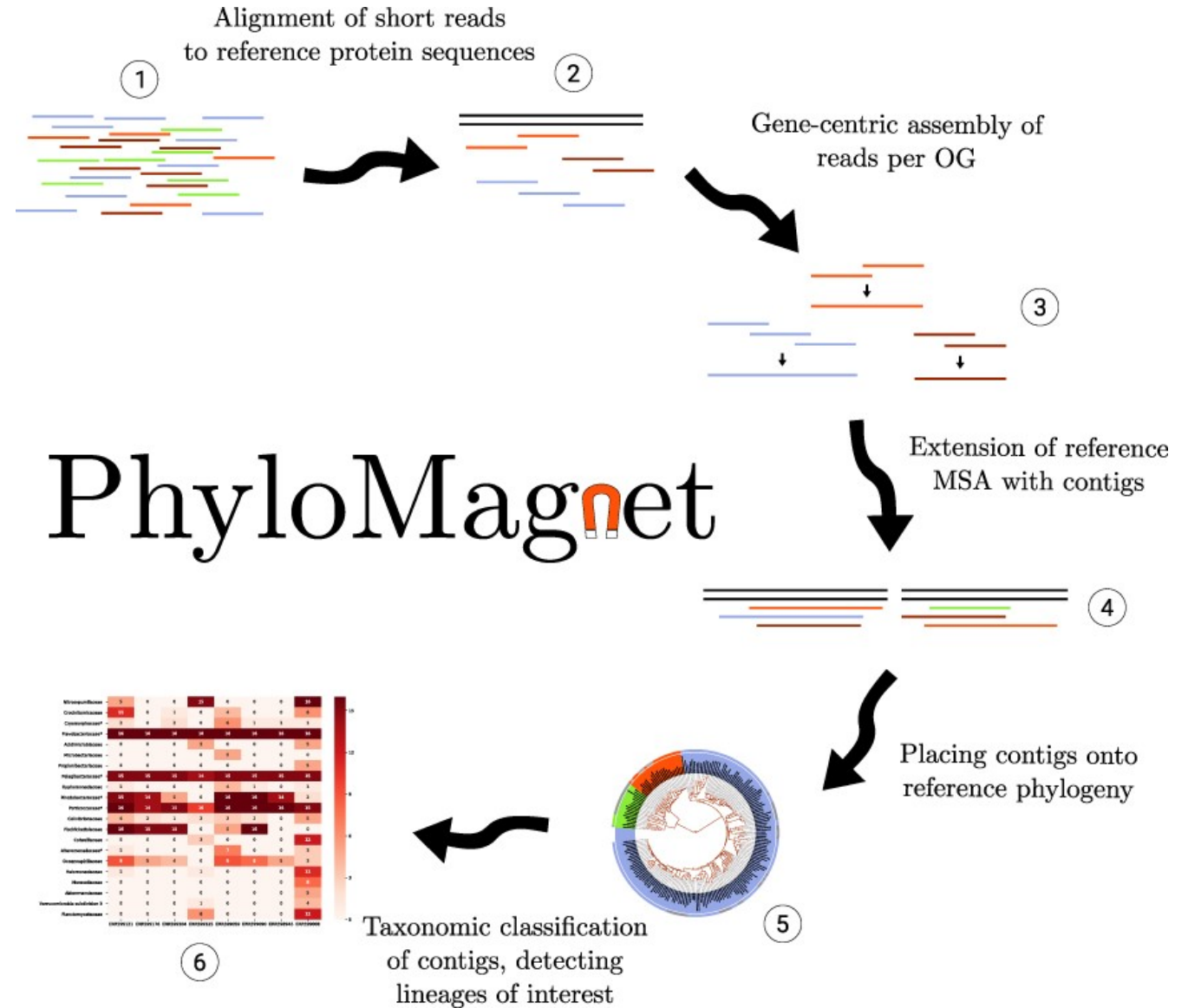
Profile Hidden Markov Model searches

- Distinct conserved units of tertiary structure that fold independently, generally with distinct functions

- Motifs: shorter supersecondary structural units

- Predicted using profile Hidden Markov Models: models of amino acids at different positions, based on alignments of known examples

- Search tools: PSI-BLAST, PHI-BLAST, DELTA-BLAST, HMMer, HH-suite, MMseqs2, MasterBlaster…
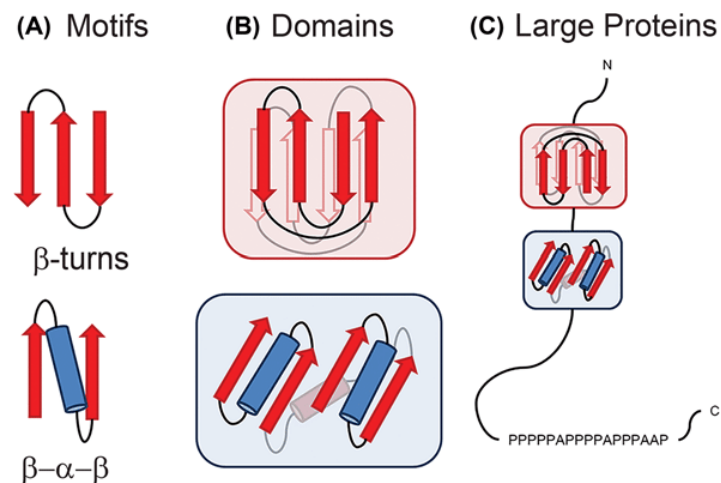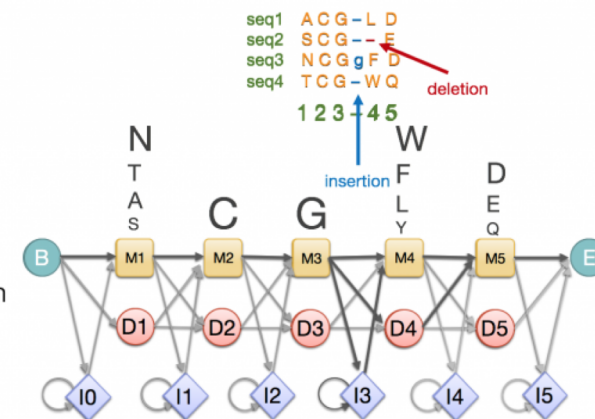
# Tree-based orthogroup inference

- Automated sequence similarity search, alignment, phylogeny reconstruction

- PhylomeDB

- GIGA

- Implemented as a step in some clustering-based software

Clustering/graph-based orthogroup inference

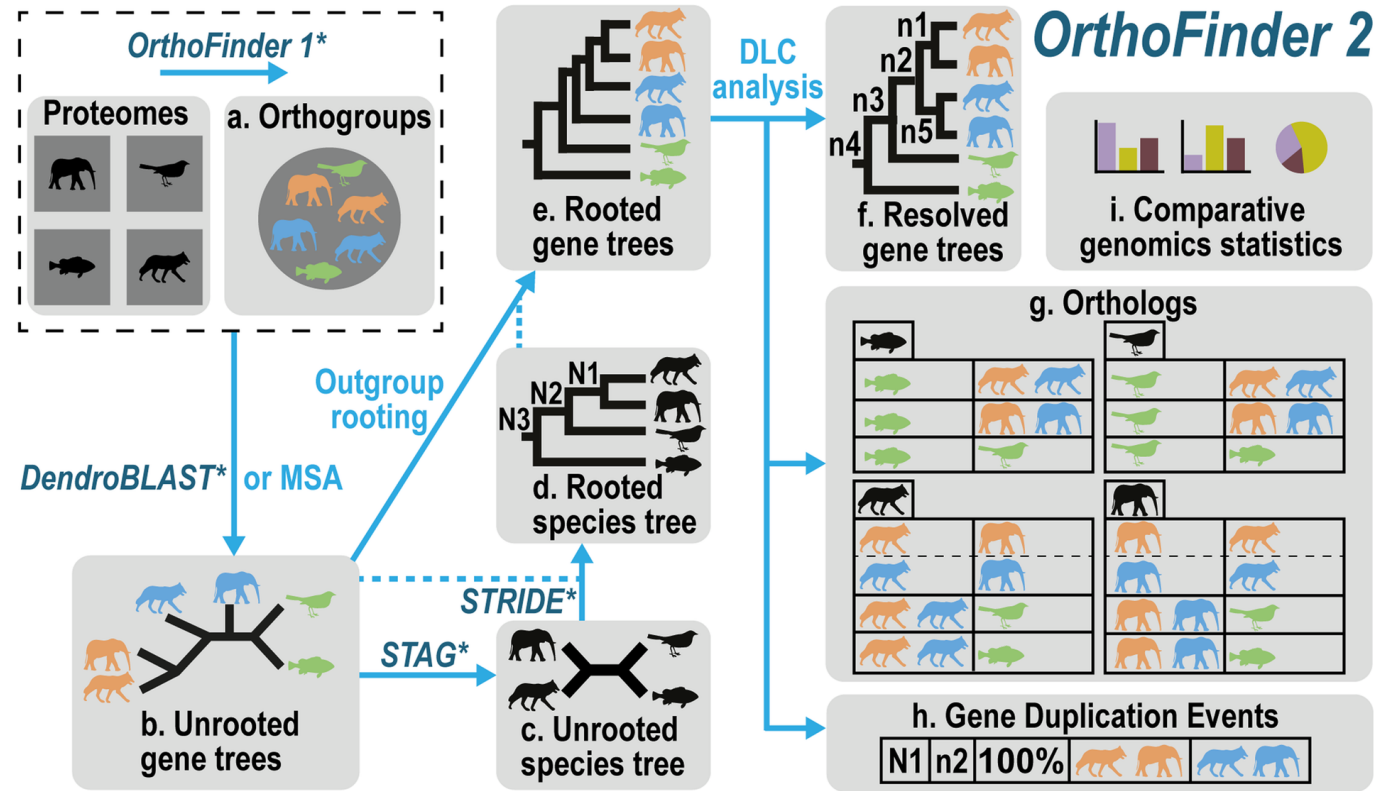OrthoFinder 1 overview (Emms & Kelly 2015)

- OrthoMCL
- OrthoFinder
- OrthoDB
- INPARANOID (within species), Hieranoid (multiple species)
- HaMSTR
- eggNOG
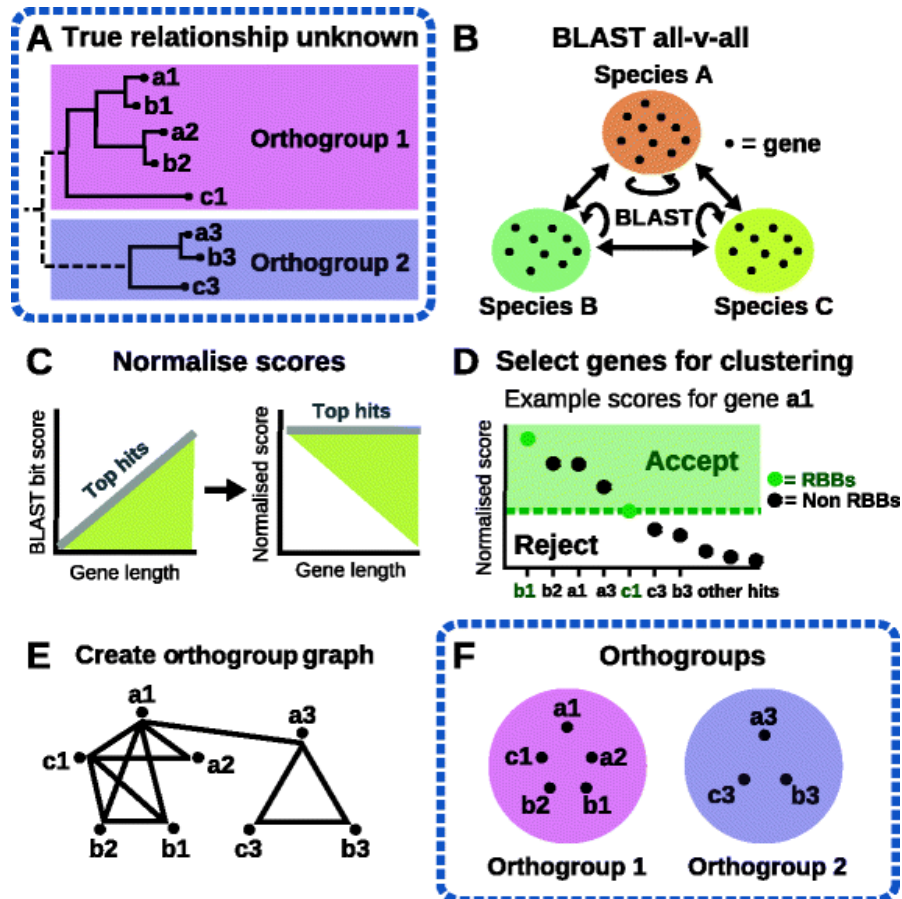
Orthofinder overview
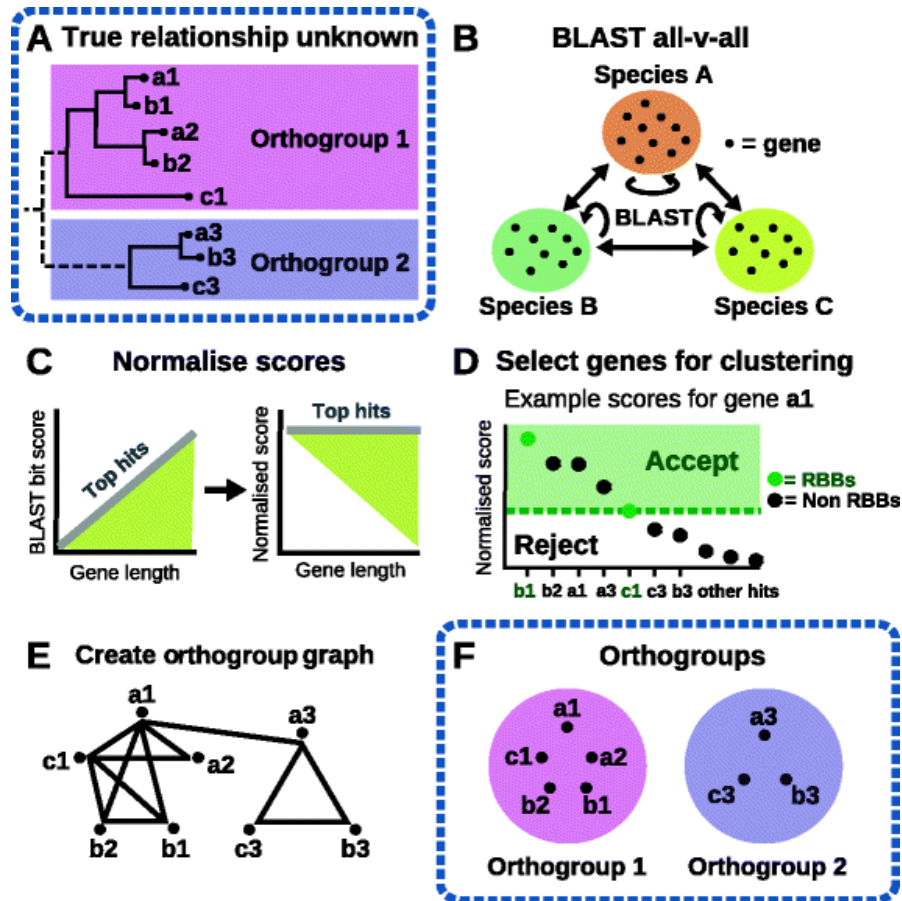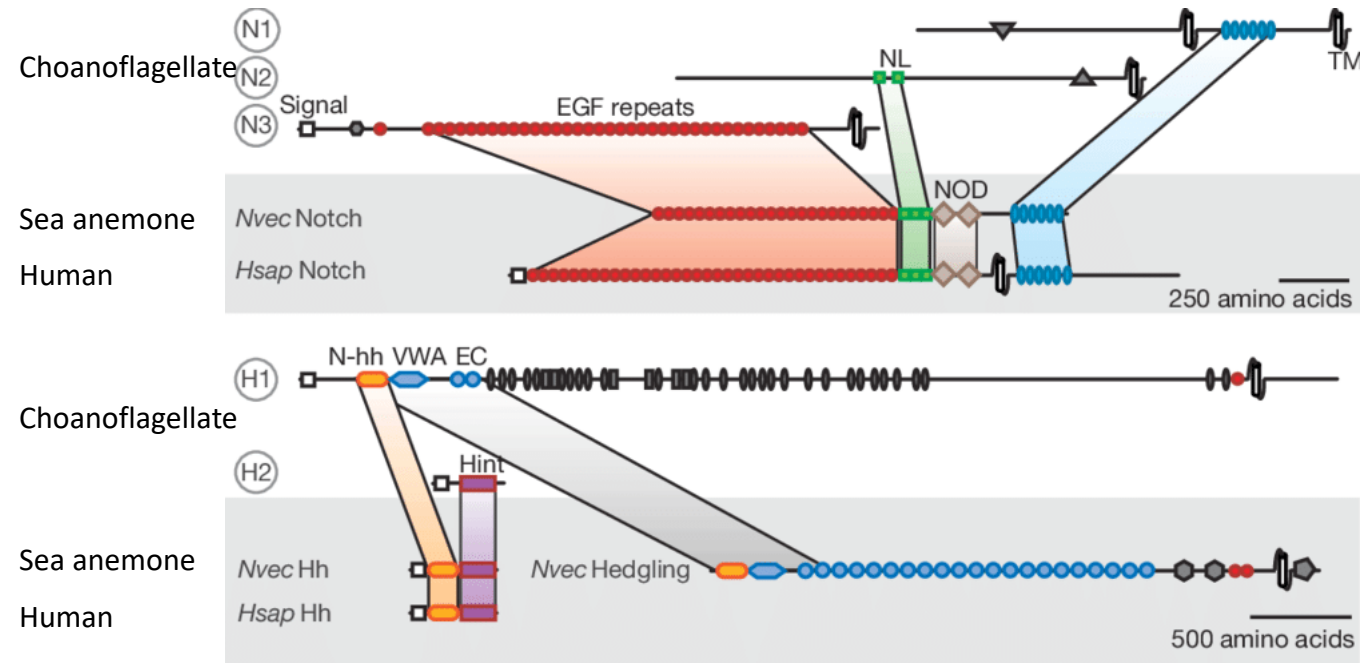

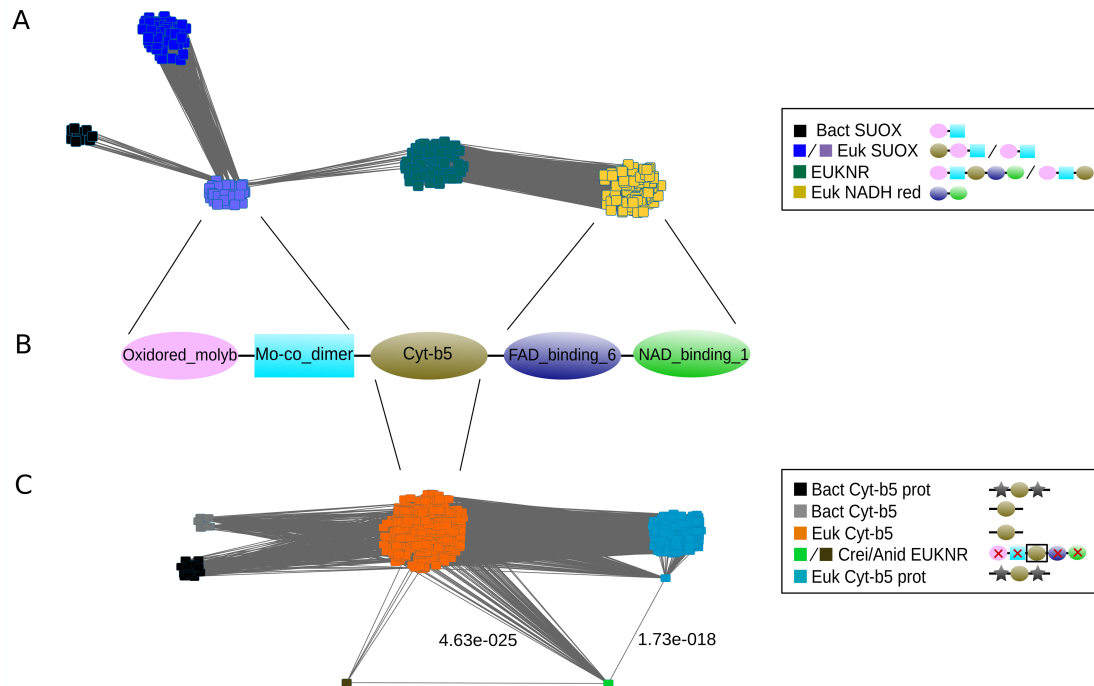OrthoFinder 1 overview (Emms & Kelly 2015)

OrthoFinder 2 overview (Emms & Kelly 2019)

Domain shuffling

King et al. 2008 Nature

Domain shuffling

A

B

C

Ocaña-Pallarès et al. 2019 PLoS Biol

- Treat each domain separately for phylogenies

- Sequence similarity network

- More difficult if there is poor taxon sampling

- More difficult in the case of short domains

Profile HMM databases/prediction tools

- Pfam: https://pfam.xfam.org/
  Search tool PfamScan: http://ftp.ebi.ac.uk/pub/databases/Pfam/current_release/

- The Simple Modular Architecture Research Tool (SMART): http://smart.embl-heidelberg.de/

- ProSite: https://prosite.expasy.org/

- The NCBI Conserved Domain Database (CDD): https://www.ncbi.nlm.nih.gov/cdd/ ; search tool CD-Search can be used in batch format for up to 4000 sequences

- InterPro: https://www.ebi.ac.uk/interpro/
  Search tool InterProScan: https://www.ebi.ac.uk/interpro/download/
  InterProScan searches sets of sequences using not only InterPro profiles, but also those from several other databases, including Pfam, CDD, ProSite, and SMART. This effectively makes it a more sensitive search tool

Useful databases

- National Center for Biotechnology Information (NCBI) GenBank (https://www.ncbi.nlm.nih.gov/genbank/)
  - e. g. nt (nonredundant nucleotide), nr (nonredundant amino acid), RefSeq (nonredundant, well annotated set of reference sequences; taxonomically limited), Transcriptome Shotgun Archive (TSA), Whole Genome Shotgun archive, Short Reads Archive (SRA), Microbial Genomes…

- EukProt (https://figshare.com/articles/dataset/EukProt_a_database_of_genome-scale_predicted_proteins_across_the_diversity_of_eukaryotic_life/12417881/3 , evocellbio.com/eukprot/): curated database designed to represent every major eukaryotic group, without overrepresenting animals or plants

- UniProt (https://www.uniprot.org/)
  - UniProtKB: protein sequences, including predicted functional annotations based on (and therefore biased toward) model organisms such as yeast or mouse.
  - Proteomes from individual species

- Uniclust (https://uniclust.mmseqs.com): proteins from UniProt clustered at different levels of sequence identity (30%, 50%, 90%) and including functional annotations

- Ensembl Genomes: by group (e.g. Ensembl Protists, Ensembl Bacteria etc. – N.B. Ensembl Bacteria includes archaea). http://ensemblgenomes.org/ (Bonus question: do you feel that the taxonomic basis for Ensembl's grouping is appropriate?)

- GenomeArk (https://vgp.github.io/genomeark): repository for vertebrate genome sequence data

- PhylomeDB (http://www.phylomedb.org/): repository of single-gene phylogenies and orthology predictions

This is by no means exhaustive: if you are interested in a specific group, it's worth searching for specialist repositories for their genome and transcriptome data

Additional resources and further reading

- Fernández et al. 2020. Orthology: Definitions, Prediction, and Impact on Species Phylogeny Inference. https://hal.archives-ouvertes.fr/hal-02535414/

- Tekaia 2016. Inferring Orthologs: Open Questions and Perspectives https://journals.sagepub.com/doi/full/10.4137/GEI.S37925

- https://bioinformaticsworkbook.org/ : helpful introductions to basic concepts in bioinformatics, including comparative genomics, data acquisition best practices, and project management

- https://holtlab.net/tag/comparative-genomics/ : Bacterial comparative genomics resources

- https://davidemms.github.io/ : Orthofinder tutorials