Comparative genomics

Michelle M. Leger

IBE Course on Evolutionary Biology 2022

INSTITUT de BIOLOGIA EVOLUTIVA
CSIC
upf.

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

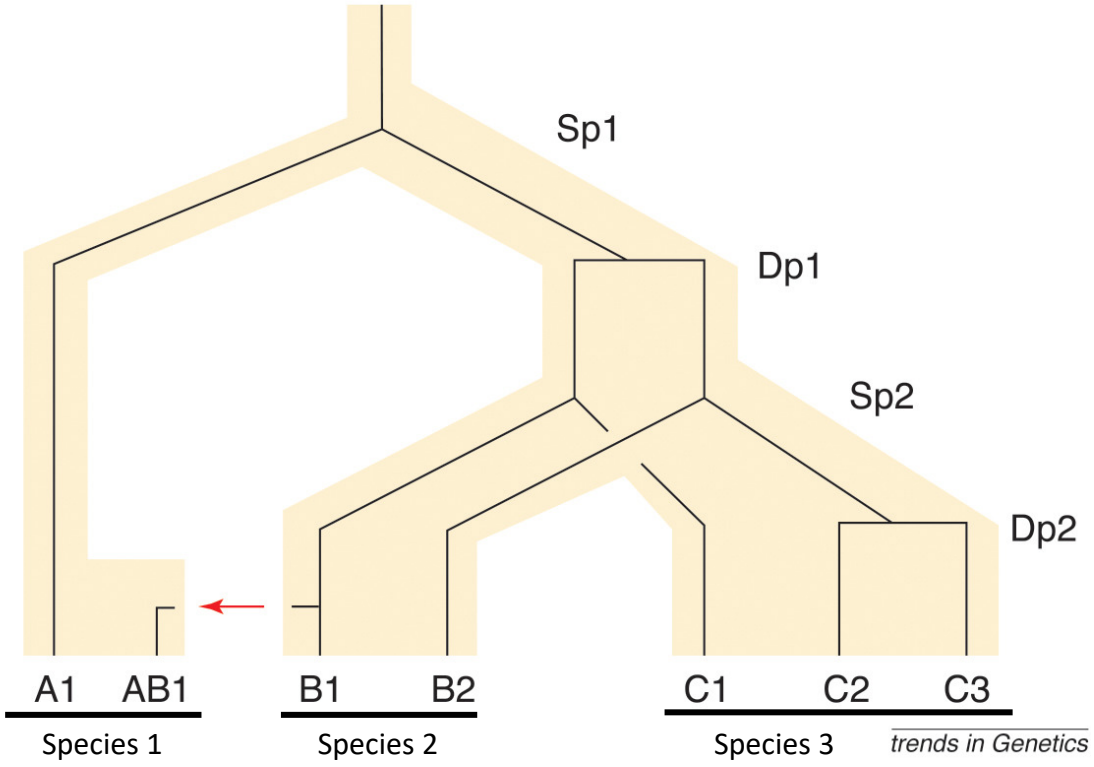
How did a sequence arise?

- 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

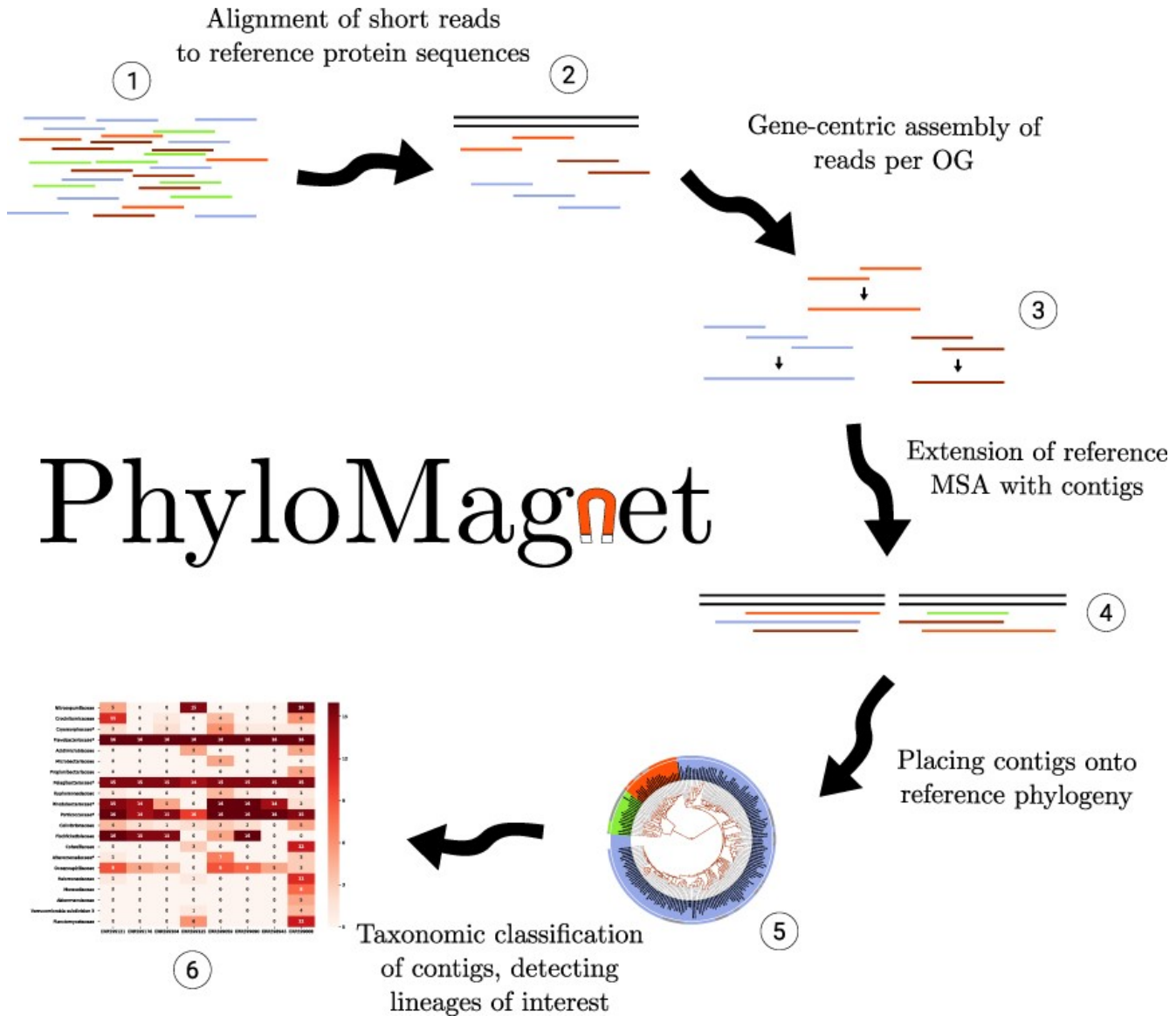
Sequence similarity searches – Basic Local Alignment Search Tool

- 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

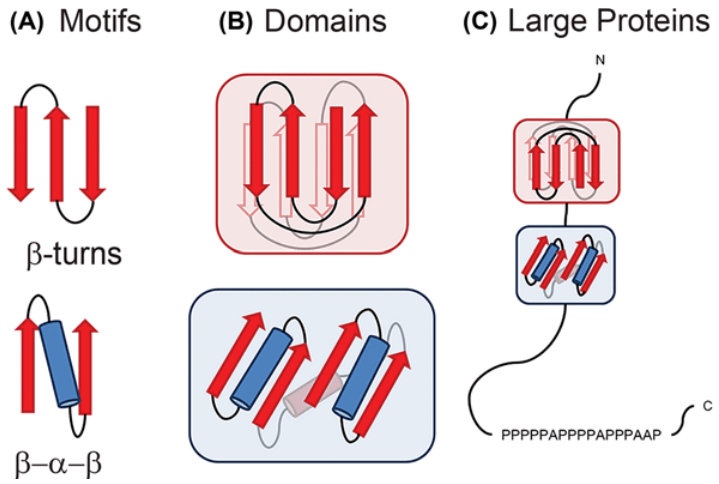
- PhyloMagnet
- MEGAN
- GRASP2

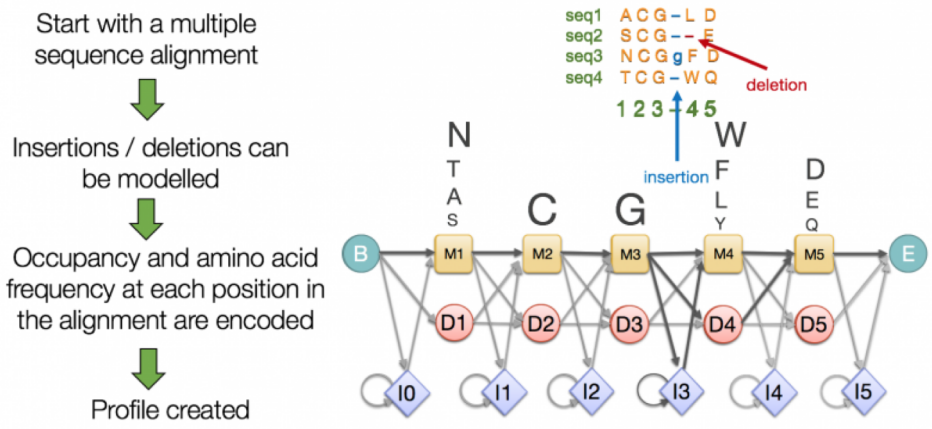
…



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…

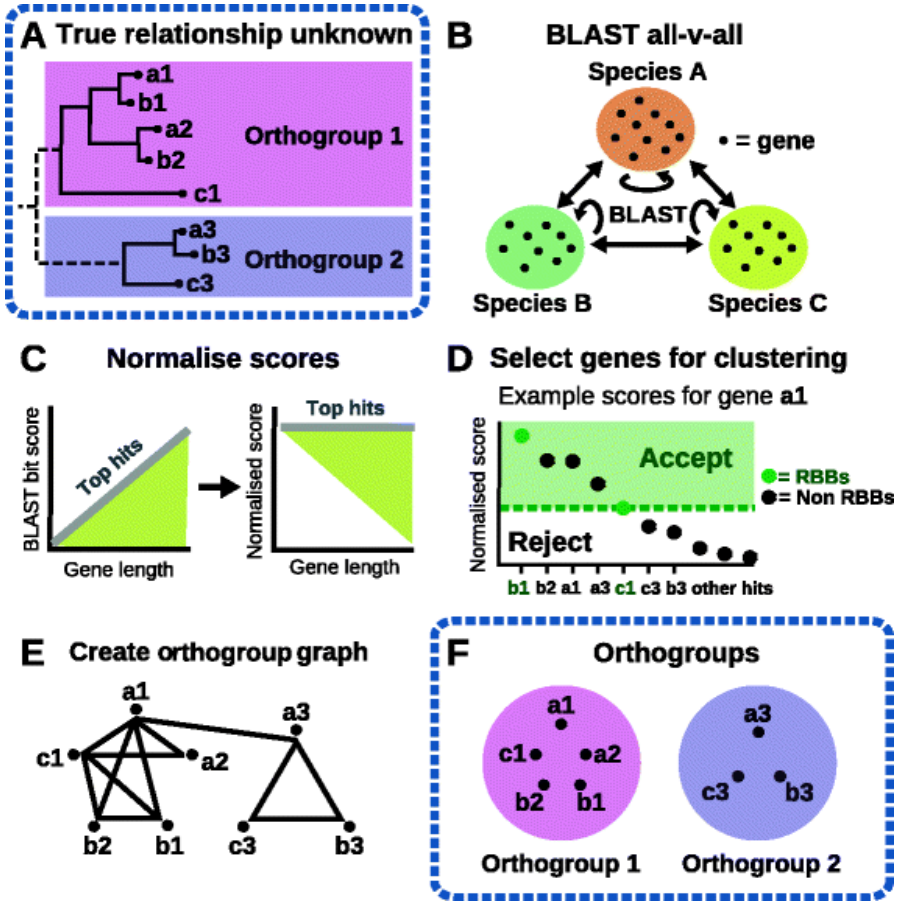




Tree-based orthogroup inference

- Automated sequence similarity search, alignment, phylogeny reconstruction
- PhylomeDB
- GIGA
- Implemented as a step in some clustering-based software

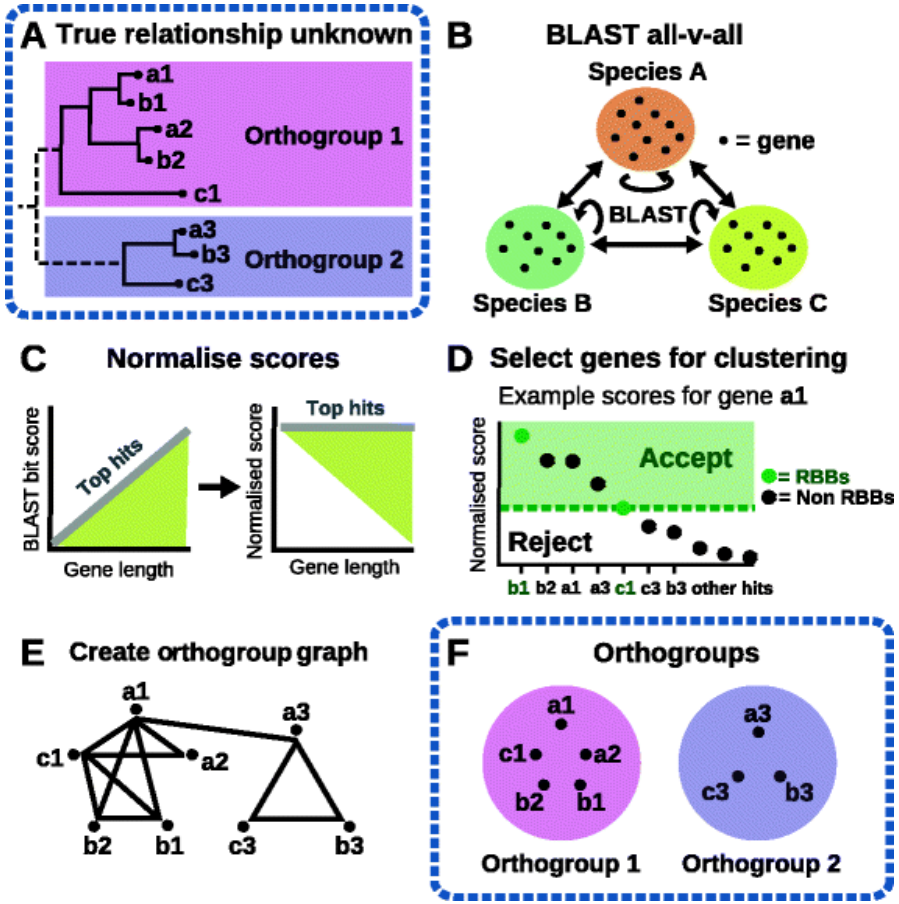
Clustering/graph-based orthogroup inference



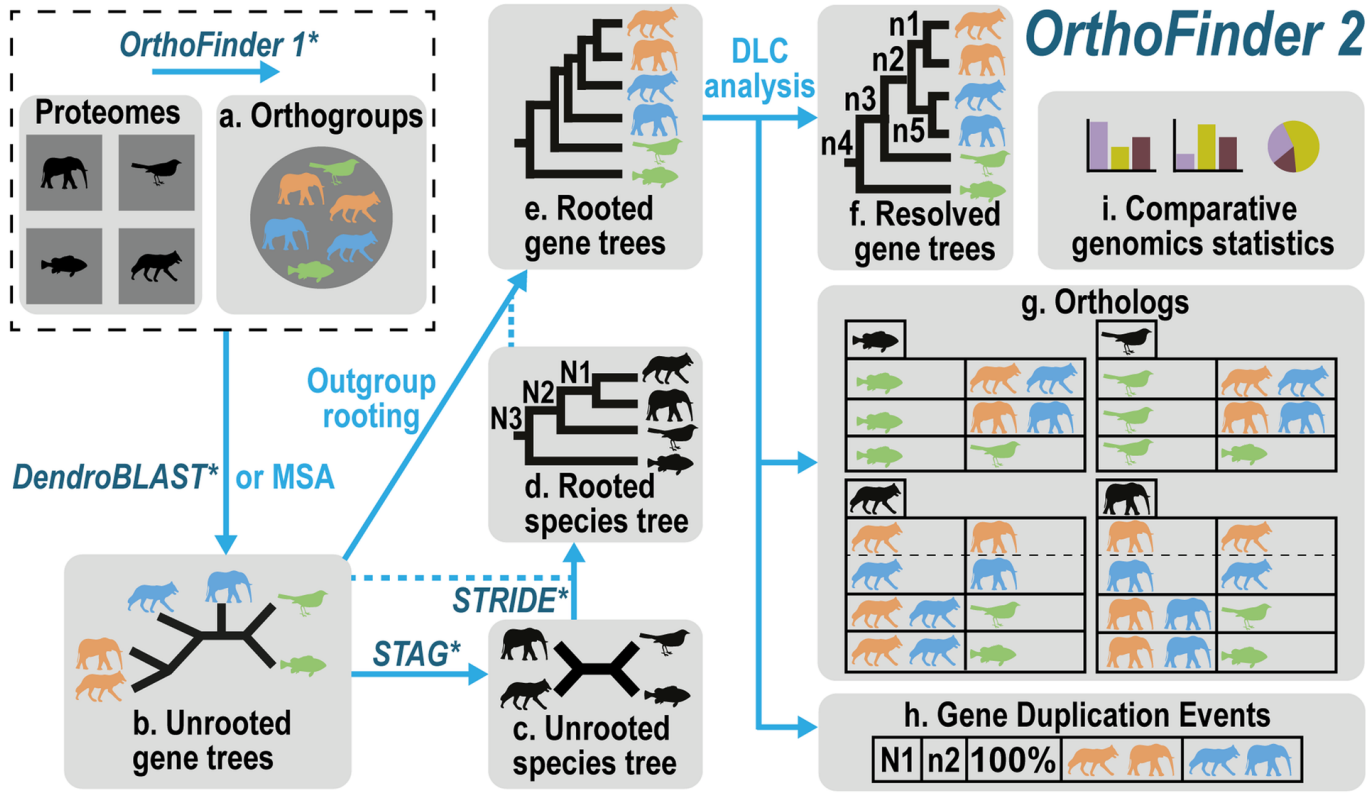
OrthoFinder 1 overview (Emms & Kelly 2015)

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

Orthofinder overview

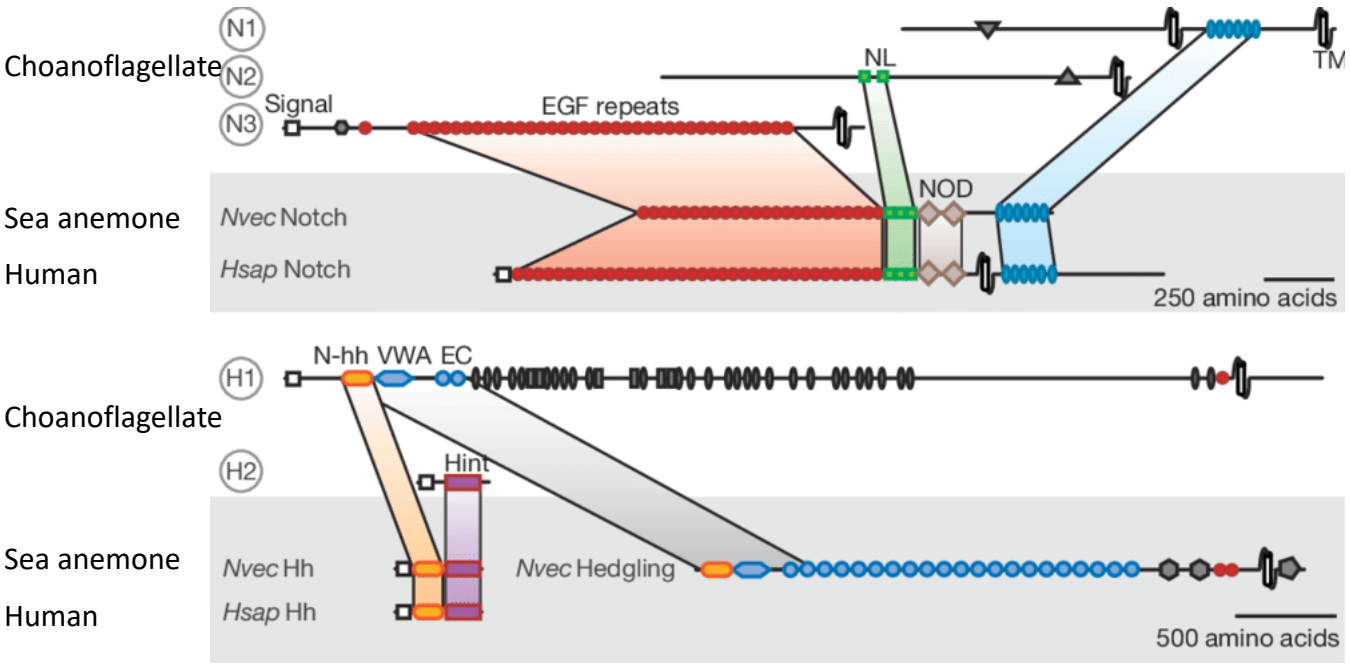


OrthoFinder 1 overview (Emms & Kelly 2015)

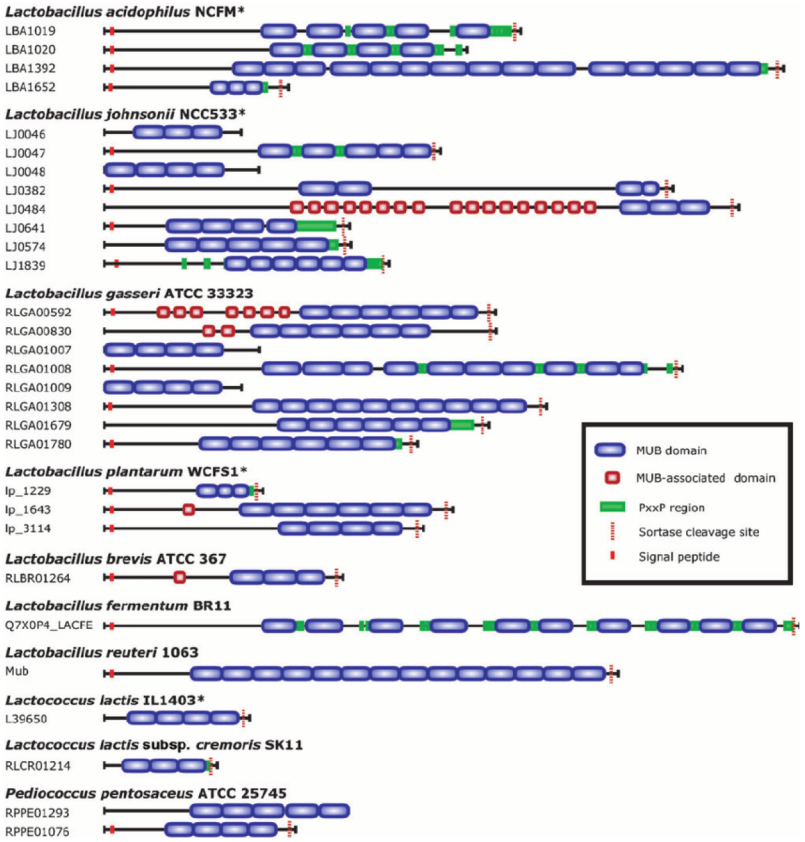


OrthoFinder 2 overview (Emms & Kelly 2019)

Domain shuffling

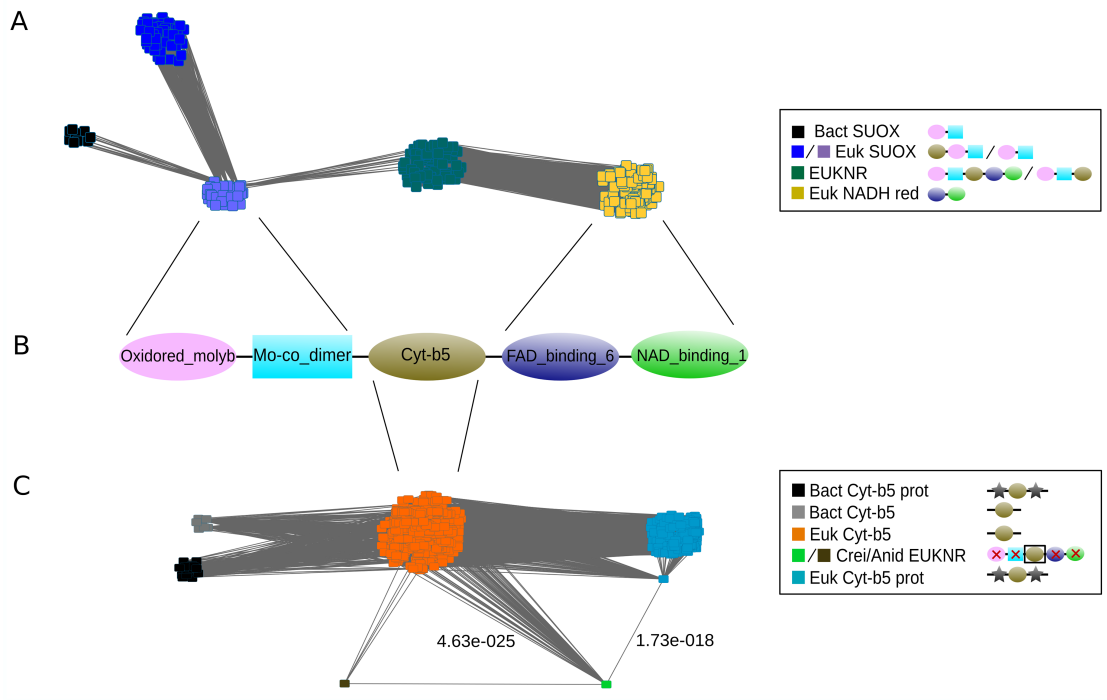


King et al. 2008 Nature



Boekhorst et al. 2006 Microbiology

Domain shuffling



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

Some good tools for phylogenetic reconstruction

- To make the multiple sequence alignment (play around with these – they're pretty comparable, but most people have a favourite our of habit or superstition)
  - MAFFT (in automatic mode; or modes adapted for multiple well-aligned regions alternating with poorly aligned regions (E-INS-I) or a single well-conserved área of alignment (L-INS-I); global alignment mode; option to add short sequences to an existing alignment….) (https://mafft.cbrc.jp/alignment/software/manual/manual.html)
  - MUSCLE (https://www.ebi.ac.uk/Tools/msa/muscle/)
  - T-COFFEE (https://tcoffee.crg.eu/)
  - Clustal series (ClustalX has a user interface, ClustalW runs through the terminal, ClustalOmega is adapted to large quantities of data) (http://www.clustal.org/)
  - PROBCONS (http://probcons.stanford.edu/)

  Most these tools are also implemented in platforms such as Geneious or MEGA, and used by programs such as Orthofinder

  Pre-2010 or so, it was normal to correct the alignment by eye. With the quantity of data now available, and the quality of aligners, that's rarely feasible or necessary.

- To trim the alignment (each of these has different possible levels of stringency)
  - trimAl (http://trimal.cgenomics.org/) (written by Toni Gabaldón while at the CRG, now at the Barcelona Supercomputing centre)
  - BMGE (https://bioweb.pasteur.fr/packages/pack@BMGE@1.12)
  - GBlocks (http://molevol.cmima.csic.es/castresana/Gblocks.html - currently not working)
  - For very well conserved alignments involving closely related sequences, may not be necessary at all

- To view or edit the alignments (trimmed or untrimmed)
  - So many! (https://en.wikipedia.org/wiki/List_of_alignment_visualization_software) personally I like Jalview

- To construct the phylogeny
  - Maximum Likelihood: IQ-TREE and RaxML are the most popular. Others: PhyML, Garli, PAUP* (LOL nobody actually uses PAUP* any more, because the latest version is 20 years in the making), FastTree (good only for very fast, inaccurate initial phylogenies, and kind of obsolete now that IQ-TREE is so fast)
  - PAY ATTENTION TO WHICH MODEL YOU'RE USING. Use a mixture model, which takes into account the fact that not all sites in a gene/protein evolve at the same rate.
  - Bayesian: MrBayes, PhyloBayes
  - An explanation of Maximum Likelihood vs. Bayesian methods: https://towardsdatascience.com/maximum-likelihood-vs-bayesian-estimation-dd2eb4dfda8a
  - "What about maximum parsimony?" – No.

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