PROVIDED STARTING DATABASE

258 TAXONOMICALLY BALANCED GENE DATA SETS FROM 277 TAXA COVERING KNOWN EUKARYOTIC DIVERSITY AND KNOWN PARALOGS
HMM PROFILES OF ALL 258 GENES

USERS PROVIDE PREDICTED PROTEOME THEY WISH TO ADD TO THE DB

[CD-HIT]

CLUSTERED PROTEINS FOR EACH ORGANISM (98% SIMILARITY, GLOBAL)

[HMMER3]

FOR GENE 1 ... 258 DO:

COLLECT HMM HITS FROM PREDICTED PROTEINS USING PROVIDED HMM PROFILES

PHYLOGENY-DIRECTED SETTING (OPTIONAL)

USE SPECIFIC QUERIES FROM RELATED ORGANISM(S) ALREADY IN STARTING DATABASE

IS THE PEPTIDE PRESENT IN THE RELATED ORGANISM?

NO

HOW MANY RELATED ORGANISMS WERE CHOSEN?

ONE

PROCESS GENE USING DEFAULT SETTING

MORE THAN ONE

CHECK NEXT ORGANISM

YES

[BLASTP]

ARE ANY OF THE FOLLOWING TRUE FOR FIRST 5 BLAST HITS?

NO SIGNIFICANT BLAST HIT (<1e-10)
BLAST HIT IS NOT IN HMM HITS
BEST BLAST HIT DOES BLAST TO A BACTERIAL SEQUENCE IN ORTHOMCL
BEST BLAST HIT DOES BLAST TO INCORRECT ORTHOGROUP

COLLECT UP TO 5 THAT PASS THOSE CRITERIA

BLAST HIT DOES BLAST TO THE CORRESPONDING GENE IN STARTING DATABASE

YES

SKIP AND CHECK THE OTHER HIT

NO

SKIP AND CHECK THE OTHER HIT

YES

PHYLOGENETIC ANALYSIS OF UP TO 5 BEST BLAST HITS, AND CORRESPONDING GENE SET

[PREQUAL][MAFFT][BMGE][FASTTREE]

REMOVE SEQS <30% OF THE TRIMMED ALIGNMENT LENGTH

DO ANY OF SURVIVING BLAST HITS BRANCH WITHIN THE EXPECTED TAXONOMIC GROUP?

YES

ADD POSITIVE HITS TO CORRESPONDING GENE SET

NO

ADD BEST BLAST HIT TO CORRESPONDING GENE SET

DEFAULT SETTING

USER DEFINES NUMBER OF HMM HITS TO BE CONSIDERED
DEFAULT =5

[BLASTP]

BLAST HIT DOES BLAST TO A BACTERIAL SEQUENCE IN ORTHOMCL

NO

ARE BOTH OF THE FOLLOWING TRUE?

HMM HIT DOES BLAST TO THE CORRESPONDING ORTHOGROUP IN ORTHOMCL
HMM HIT DOES BLAST TO THE CORRESPONDING GENE SET IN STARTING DATABASE

YES

ADD GIVEN PEPTIDE TO CORRESPONDING GENE SET AND PROCESS NEXT HMM HIT

NO

YES

HOW MANY HITS WERE TO BE CONSIDERED?

ONE

SKIP THIS GENE

MORE THAN ONE

PROCESS NEXT HMM HIT

ADDITION OF SELECTED PEPTIDES FROM GIVEN ORGANISM TO STARTING SINGLE GENE DATA SETS

SINGLE GENE DATA SETS INCLUDING SELECTED SEQUENCES FROM ALL NEWLY ADDED TAXA

[PREQUAL] **[MAFFT]**

SINGLE GENE ALIGNMENTS
ADD KNOWN PARALOGS FROM STARTING DATASET [ON/OFF]

Marked all len(protein) >= len(aln) BMGE every time
[BMGE / ?DIVVIER-TRIMAL? / NO TRIMMING]

REMOVAL OF SHORT SEQUENCES
<30% OF THE MEANINGFUL ALIGNMENT LENGTH ("-","X")

SINGLE GENE ALIGNMENTS OF SELECTED TAXA FOR TREE CONSTRUCTION
+
SUMMARY STATISTICS PER ORGANISM

[RAxML] or [IQ-TREE]

MANUAL STEP

ML SINGLE GENE TREES WITH BOOTSTRAPS

MANUAL STEP

VISUAL INSPECTION OF TREES COLOURED BY TAXONOMIC AFFILIATION AND WITH SUSPICIOUS CLADES HIGHLIGHTED (BOOT >70, MIXED TAXONOMY)

MANUAL STEP

MARKING CORRECT ORTHOLOGS, PARALOGS, AND SEQUENCES TO BE ELIMINATED
[MODIFIED FIGTREE / TABLE-ETE3]

[PREQUAL]
[MAFFT]

FINAL OUTPUT OF THE PIPELINE

FINAL GENES AND TAXA SELECTION
BUILDING CLEANED SINGLE GENE ALIGNMENTS
CONCATENATION TO PHYLOGENOMIC MATRIX
SUMMARY STATISTICS

[BMGE /?DIVVIER-TRIMAL? / NO TRIMMING]

ADD MARKED PARALOGS AND ORTHOLOGS FROM NEW TAXA TO THE STARTING DATASET