PREDICTED PROTEOMES

STARTING DATABASE GENETIC CODE ANALYSIS (optional)

(single-gene sets)

PRELIMINARY DATASETS

- (candidate seqs from new taxa included)
- phylogenetically aware (optional)
- can keep more seqs per gene
- obvious contaminations removed

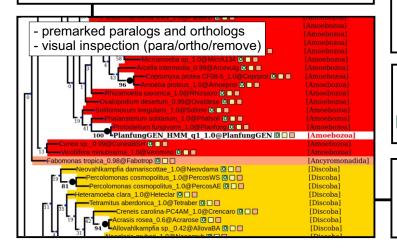
3x

- gene occupancy
- organism statistics
- gene statistics

FILTERED DATASETS

(selected taxa and genes) (optional: include paralogs for certain taxa)

MULTIPLE SEQUENCE ALIGNMENTS TRIMMING (PREQUAL, DIVVIER, TRIMAL) SINGLE GENE TREES





PHYLOFISHER

SITES REMOVAL

PHYLOGENOMIC MATRIX (CONCATENATION)

MSA TRIMMING SINGLE GENE TREES

SINGLE GENE TREES

FINAL DATASETS (selected taxa and genes)

CSV

- gene occupancyorganism statistics
- gene statistics

UPDATED DATABASE

(selected taxa and genes) (optional: include paralogs for certain taxa) COMPOSITIONAL BIAS TEST

FAST-EVOLVING

RANDOM SUBSAMPLING