An estimate of the deepest branches of the tree of life from ancient vertically-evolving genes

Edmund R. R. Moody[1], Tara A. Mahendrarajah[2], Nina Dombrowski[2], James W. Clark[1], Celine Petitjean[1], Pierre Offre[2], Gergely J. Szöllősi[3,4,5], Anja Spang[2,6*], Tom A. Williams[1*]

1. School of Biological Sciences, University of Bristol, Bristol BS8 1TH, UK.
2. NIOZ, Royal Netherlands Institute for Sea Research, Department of Marine Microbiology and Biogeochemistry; AB Den Burg, The Netherlands
3. Dept. of Biological Physics, Eötvös Loránd University, 1117 Budapest, Hungary
4. MTA-ELTE "Lendület" Evolutionary Genomics Research Group, 1117 Budapest, Hungary;
5. Institute of Evolution, Centre for Ecological Research, 1121 Budapest, Hungary
6. Department of Cell- and Molecular Biology, Science for Life Laboratory, Uppsala University, SE-75123, Uppsala, Sweden

*Co-corresponding authors: tom.a.williams@bristol.ac.uk, anja.spang@nioz.nl

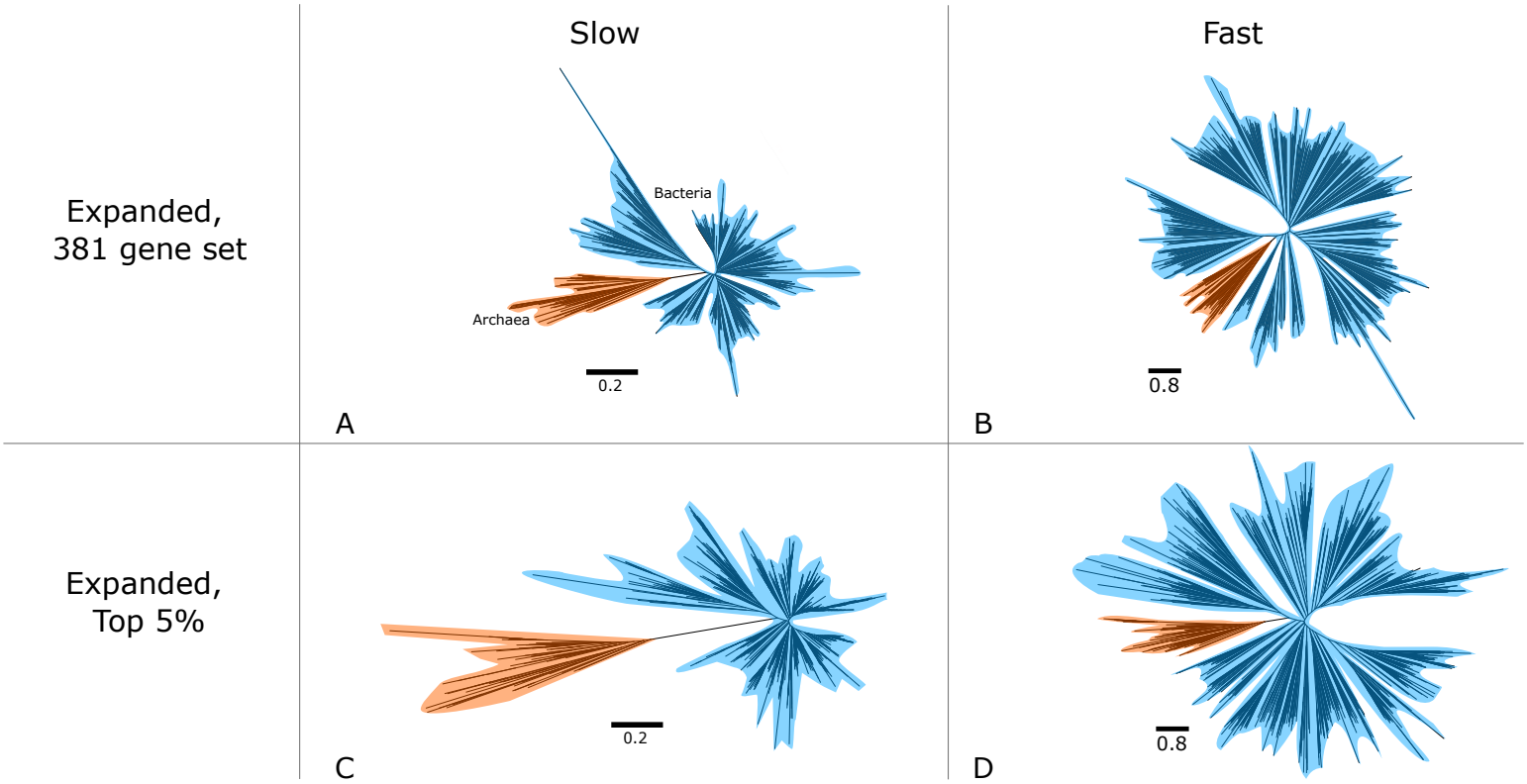

Figure S1. Vertically-evolving genes and slow-evolving sites support a longer relative AB branch length. *We estimated site-specific evolutionary rates for all marker genes in the expanded dataset (A-B), as well as for the 20 genes with the smallest ΔLL (top 5%) in that dataset (C-D). Concatenations based on the 25% slowest sites (A,C) and on the top 5% vertical genes (C,D) support a longer AB branch. This suggests that the inference of a short AB branch is impacted by both substitutional saturation and unmodelled inter-domain transfer of marker genes. Phylogenies were inferred under the LG+G4+F model in IQ-TREE 2. Branch lengths are the expected number of substitutions per site, as indicated by the scale bars. Alignment lengths in amino acids: A: 36797, B: 67274, C: 2736, D: 3884.*

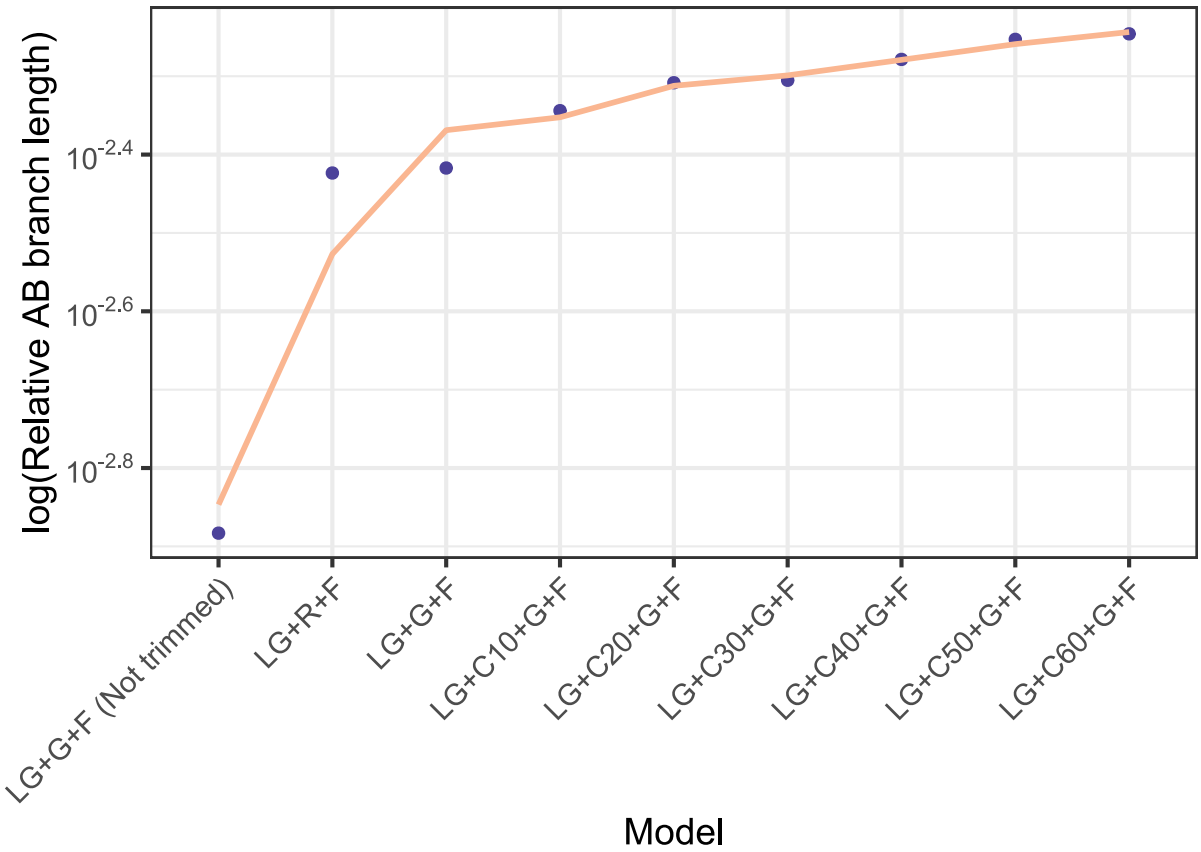

Figure S2. The effect of modelling site heterogeneity on AB branch length. *Increasing the number of protein mixture profiles, as well as trimming is associated with a change in AB branch length on the expanded marker set. All analyses used LG exchangeabilities, four rate categories (Gamma-distributed or freely estimated), and included a general composition vector containing the empirical amino acid frequencies (+F). Modelling of site heterogeneity with the C10-C60 models increases the inferred AB branch length ~2-fold. Trimming poorly-aligned sites slightly increases the AB branch estimation whereas relaxing the gamma rate categories slightly decreases estimation of AB branch length. LG (LG substitution matrix), G (four gamma rate categories), F (empirical site frequencies estimated from the data), C10-60 (number of protein mixture profiles used) R (four free rate categories which relax the assumption of a gamma distribution for rates, BMGE (trimming using Block Mapping and Gathering with Entropy). The trendline is estimated using a LOESS regression.*

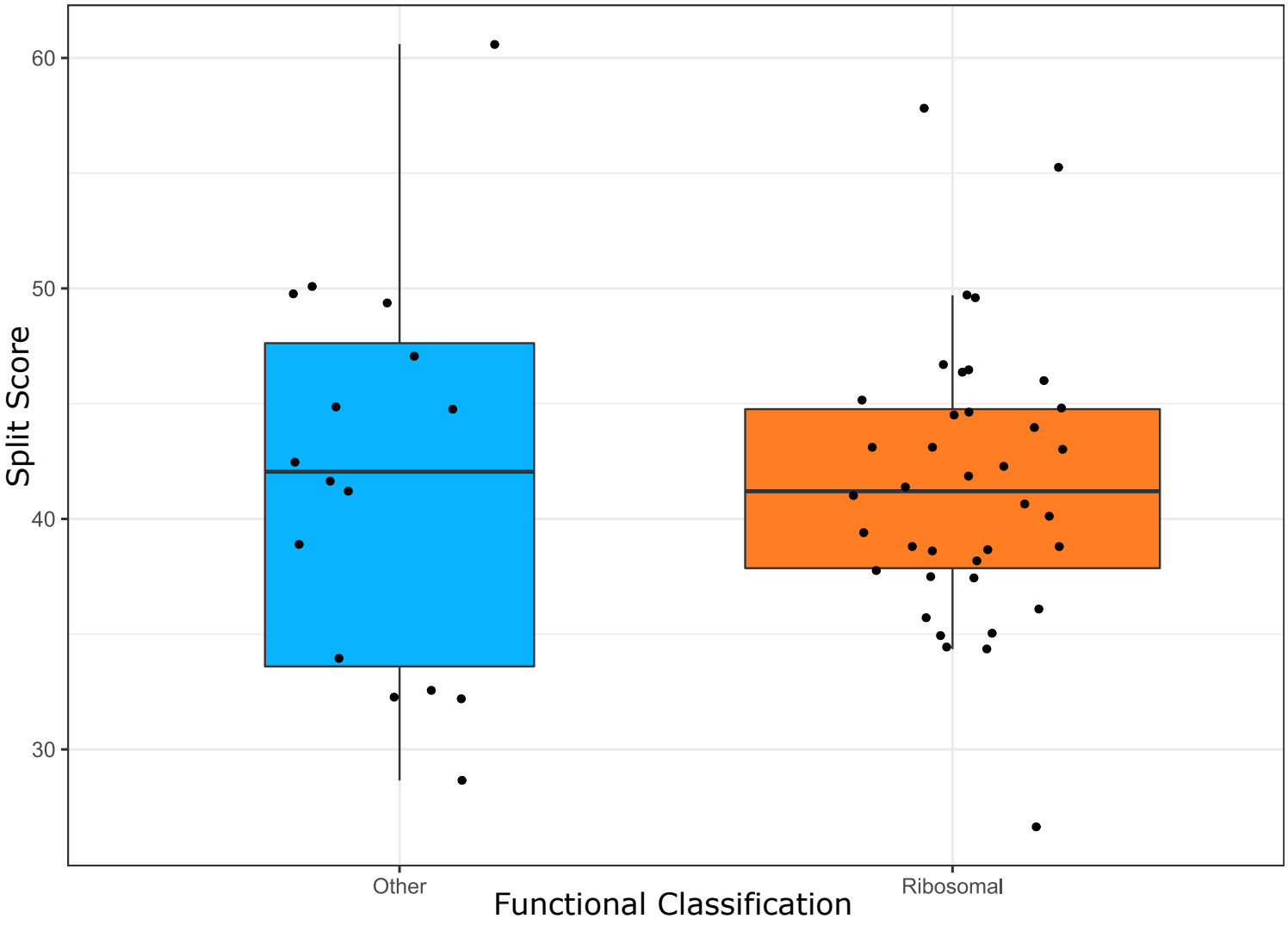

Figure S3. Among vertically-evolving marker genes, ribosomal genes are no better or worse as phylogenetic markers. *The plot shows functional classification of markers (ribosomal markers or other) against the split score (a higher split score denotes a worse recovery of established within-domain relationships) using 54 markers from the new analysis. After removing genes that do not appear to have been vertically inherited since the divergence of Archaea and Bacteria, split scores of ribosomal and non-ribosomal markers were statistically indistinguishable (P = 0.7836).*

Tables:

S1_Data_Supplement_Table.xlsx – Table containing marker metadata, KO and Pfam annotations and descriptions, and manual inspection notes for reciprocal monophyly and presence of paralogues, for 381 marker genes used in Zhu et al., 2019 (Methods).

S2_Markermetadata.xlsx – Table containing marker metadata, KO and Pfam annotations and descriptions, and manual inspection notes for 95 markers in the Core, Bacterial, and Non-Ribosomal marker gene sets (Methods).

S3_700ArcBac_species_list.xlsx – Table containing NCBI taxonomic information for 350 Archaeal and 350 Bacterial.

S4_split_score_stats.xlsx – Table containing clade definitions for quantifying taxonomic splits and split-score statistical summaries for ranking of the Core, Bacterial, and Non-Ribosomal marker genes.