Quantitatively Partitioning Microbial Genomic Traits among Taxonomic Ranks across the Microbial Tree of Life

Taylor M. Royalty[1]

Andrew D. Steen[1,2,*]

[1]Department of Earth and Planetary Sciences, University of Tennessee, Knoxville

[2]Department of Microbiology, University of Tennessee, Knoxville

*asteen1@utk.edu

Abstract

Widely used microbial taxonomies, such as the NCBI taxonomy, are based on a combination of sequence homology among conserved genes and historically accepted taxonomies, which were developed based on observable traits such as morphology and physiology. A recently-proposed alternative taxonomy, the Genome Taxonomy Database (GTDB), incorporates only sequence homology of conserved genes and attempts to partition taxonomic ranks such that each rank implies the same amount of evolutionary distance, regardless of its position on the phylogenetic tree. This provides the first opportunity to completely separate taxonomy from traits, and therefore to quantify how taxonomic rank corresponds to traits across the microbial tree of life. We quantified the enrichment of clusters of orthologous gene functional categories (COG-FCs) as a proxy for traits within the lineages of 13,735 cultured and uncultured microbial lineages from a custom-curated genome database. On average, 41.4% of the variation in COG-FC enrichment is explained by taxonomic rank, with domain, phylum, class, order, family, and genus explaining, on average, 3.2%, 14.6%, 4.1%, 9.2%, 4.8%, and 5.5% of the variance, respectively ($p<0.001$ for all). To our knowledge, this is the first work to quantify the variance in metabolic potential contributed by individual taxonomic ranks. A qualitative comparison between the COG-FC enrichments and genus-level phylogenies, generated from published concatenated protein sequence alignments, further supports the idea that metabolic potential is taxonomically coherent at higher taxonomic ranks. The quantitative analyses presented here constrain the integral relationship between diversification of microbial lineages and the metabolisms which they host.

Importance

Recently there has been great progress in defining a complete taxonomy of bacteria and archaea, which has been enabled by improvements in DNA sequencing technology and new bioinformatic techniques. A new, algorithmically-defined microbial tree of life describes those linkages relying solely

on genetic data, which raises the question of how microbial traits relate to taxonomy. Here, we adopted cluster of orthologous group functional categories as a scheme to describe the genomic contents of microbes, which can be applied to any microbial lineage for which genomes are available. This simple approach allows quantitative comparisons between microbial genomes with different gene composition from across the microbial tree of life. Our observations demonstrate statistically significant patterns in cluster of orthologous group functional categories at the taxonomic levels spanning from domain to genus.

Introduction

The relationship between microbial taxonomy and function is a longstanding problem in microbiology (1–3). Prior to the identification of the 16S rRNA gene as a taxonomic marker, microbial phylogenetic relationships were defined by traits such as morphology, behavior, and metabolic capacity. Cheap DNA sequencing has provided the ability to fortify those phenotype-based taxonomies with quantitative determinations of differences between marker genes, but canonical taxonomies such as the NCBI taxonomy continue to "reflect the current consensus in the systematic literature," which ultimately derives from trait-based taxonomies (4). Recently, Parks et al. (5) formalized the genome taxonomy database (GTDB), a phylogeny in which taxonomic ranks are defined by "relative evolutionary divergence" in order to create taxonomic ranks that have uniform evolutionary meaning across the microbial tree of life (5). This approach removes phenotype or traits entirely from taxonomic assignment. Thus, an investigation of the relationship between traits and phylogeny has not been possible until the recent publication of a microbial tree of life that is based solely on evolutionary distance.

Comparing phenotypic characteristics of microorganisms across the tree of life is not currently

59 possible, because most organisms and lineages currently lack cultured representatives (6, 7). We
60 therefore used the abundance of different Clusters of Orthologous Groups (COGs) in microbial
61 genomes, a proxy for phenotype which is available for all microorganisms for which genomes are
62 available. Clusters of orthologous groups (COGs) are a classification scheme that defines protein
63 domains based on groups of proteins sharing high sequence homology (8). More than ~5,700 COGs
64 have been identified to date. COGs are placed into one of 25 metabolic functional categories (COG-
65 FCs), which represents a generalized metabolic function. Our analyses quantify the degree to which
66 taxonomic rank (genus through domain) predicts the COG-FC content of genomes, and illustrate which
67 lineages are relatively enriched or depleted in specific COG-FCs. These analyses constitute a step
68 towards being able to probabilistically predict the metabolic or functional similarity of microbes given
69 their taxonomic classification.

70 **Results**

71 The genomes analyzed in this work were compiled from a variety of different sources, including
72 RefSeq v92, JGI IMG/M, and Genbank, in order to include genomes created using diverse sequencing
73 and assembly techniques. The integration of RefSeq v92, JGI IMG/M, and Genbank databases resulted
74 in a total of 119,852 genomes within the custom-curated database. Raw data, GTDB taxonomy, and
75 associated accessions are provided in Dataset S1. Of these genomes, we included only those that
76 satisfied a set of criteria designed to ensure that each genus contained enough genomes to allow
77 statistically robust analysis (see Methods). This resulted in a set of 13,735 lineages, representing 22
78 bacterial phyla, and 4 archaeal phyla, of which 67% have been grown in culture (Table 1).

79 Most predicted open reading frames for most lineages could be assigned to a COG-FC. Across
80 all phyla, an average of 84.3% $\pm$ 7.8% of open reading frames were assigned to a COG-FC (Fig S1).

81 Genomes of the same phylum tended to group together in an initial principal component analysis (PCA)
82 of raw COG-FC abundance (Fig. 1A). Since this analysis was based on absolute abundance of COG-FCs
83 in genomes, rather than relative abundance, we hypothesized that the relationship between COG-FC
84 abundance and phylum was largely a consequence of genome size, which is phylogenetically conserved
85 (9). Consistent with this possibility, position on PC 1 correlated closely with genome size (R^2=0.88; Fig
86 1B). We therefore normalized each COG-FC abundance, for each genome, to a prediction of COG-FC
87 abundance as a function of genome size derived from a generalized additive model (GAM; Fig S2;
88 summary statistics in table S1). Each GAM model was statistically significant ($p < 0.001$), and all but
89 five COG-FCs had deviance explained (analogous to adjusted R^2) of more than 50%. We interpret
90 analyses of these genome size-normalized data sets as reflecting the enrichment or depletion of COG-FC
91 abundance, relative to that expected for a given genome size. PCA of these COG-FC enrichments
92 showed that species-level lineages still tended to group by phylum, even though the inter-phyla
93 gradients in genome size were no longer apparent (Fig 1C, D).

94 To quantify the degree to which taxonomic rank explains the distribution of COG-FC
95 enrichments among individual genomes, we performed a permutation multivariate ANOVA
96 (PERMANOVA) using the following taxonomic ranks: domain, phylum, class, order, family, and genus,
97 as well as culture-status (cultured versus uncultured lineage). The rank of species was excluded from the
98 analysis as every lineage was unique, and thus, species would explain 100% of the data. Every rank
99 significantly influenced the distribution of COG-FC enrichments ($p < 0.001$), but the fraction of variance
100 that each rank explained differed substantially: phylum explained the most variance (14.6%), followed
101 by order (9.2%), genus (5.5%), family (4.8%), and class (4.1%). Domain explained only 3.1% of
102 variance in COG-FC enrichment, the least of any taxonomic rank. Culture-status was a significant
103 correlate of COG-FC abundance ($p < 0.001$) but had virtually no explanatory power, with variance

104 explained <0.001%. This observation is consistent with no particular COG-FC being systematically
105 enriched or depleted in uncultured microbes relative to cultured microbes.

106 The variability in COG-FC enrichment across different phyla was explored in addition to mean
107 COG-FC composition for individual phyla (Fig 3). Here, the distance in PCA space between each
108 lineage and the centroid of all lineages in its phyla is a measure of how different the genome is, in terms
109 of COG-FC distribution, from the typical lineage in that phylum. Among all phyla, COG-FC distribution
110 of Crenarchaeota was the most variable, followed by Patescibacteria and Cyanobacterota. The least
111 variable phyla were the Synergistota, Marinisomatota, and Fibrobacterorta, respectively (Fig 3A). We
112 explored the possibility that variance in COG-FC distribution was a function of the number of lineages
113 in the phylum. In other words, did COG-FC content of some genomes simply seem less variable because
114 they had been under-sampled? A plot of the average distance of lineages from their phylum's centroid
115 versus the number of lineages in the phylum reveals an apparent increase in variability between 10 and
116 roughly 100 genomes, at which point average distance being approaching an asymptote (Fig 3B).
117 Comparison of Akaike Information Criterion (AIC) of a saturating model of the data (Eq. 1), fit using
118 nonlinear least squares, to a linear regression indicated that the saturating model described the
119 relationship substantially better (ΔAIC = 10.5). Coefficient *A* of the saturating model, which represents
120 the value of the asymptote, was estimated to be 0.75 +/- 0.15 ($p < 0.001$). Coefficient B, which
121 represents how quickly the function approaches the asymptote, was 0.43 +/- 0.30 ($p = 0.17$). Coefficient
122 C, an offset to handle the fact that all the log-transformed distances have negative values, was -1.63 +/-
123 0.14 ($p < 0.001$).

124 Finally, we quantified the average enrichment or depletion of each COG-FC in each genus These
125 values were then visualized on a genus-level phylogenetic tree (Fig 4) built from concatenated ribosomal
126 protein sequences published by Parks et al. (5). Data underlying Fig. 4 are presented in Supplemental

Data Set 2. Several notable features appear in COG-FC enrichments at the phylum level. For example, among the four Archaea phyla represented here, Thermoplasmatota appears unique, with substantial enrichments in cell motility and depletion in every other category relative to <<relative to what? Could probably just end the sentence with “every other category”.>> . In general, the bacteria appeared more variable at the phylum level than the archaea. The clade consisting of Bacteroidota, Spirochaetota, and Verrucomicrobiota were notably depleted in the less-variable COG-FCs, including energy production and conversion, amino acid transport and metabolism, and carbohydrate transport and metabolism, among others. Another prominent feature is the near-ubiquitous enrichment of cell motility, secondary metabolites biosynthesis, transport, and catabolism, lipid transport and metabolism, and intracellular trafficking, secretion, and vesicular transport COGs in Proteobacteria. A notable dichotomy in the enrichment of RNA processing and modification within the Proteobacteria mirrors the division of the two largest clades within the proteobacteria. Overall, enrichment data qualitatively appears consistent with phylogenetic relationships, albeit, occurring on different taxonomic levels.

To gain a sense of “notable” COG-FCs associated with different phyla, we calculated the mean COG-FC across all lineages in a given phyla and compared these values against the 85th and 15th percentiles for all lineages in our custom-curated database. All COG-FCs which were significantly ($p < 0.05$; based on a 10^5-iteration bootstrap analysis) greater or less than the 85th and 15th percentiles, respectively, are shown in Table 2. Most noteworthy, was the Archaea were regularly enriched or depleted in a suite of COG-FCs. Most Bacteria phyla were enriched or depleted in in 3-4 COG-FCs. A few exceptions arose such as Fibrobacterota was deplete in 8 COG-FCs, Nitrospirota A was enriched in 4 and deplete in 5, and Proteobacteria was the only phyla not heavily enriched or deplete in any COG-FCs. Enrichment data, along with associated GTDB taxonomic assignments, used for generating Fig 4 is available in Dataset S2.

150 **Discussion**

151 We find that the abundance of each COG-FC in genomes scales with the genome size, so PCA
152 analysis of COG-FC enrichments or depletions successfully distinguished phyla only once COG-FC
153 abundance was normalized to the abundance that was predicted based on genome size (Fig 1C, D). The
154 relationship between COG-FC abundance and genome size was clearly nonlinear, however, with higher
155 slopes at lower genome sizes, indicating that a simple linear model would not accurately capture that
156 relationship (Fig S2).

157 The PERMANOVA (Fig 2) and analysis of diversity of genomic composition within phyla (Fig
158 3) showed that microbial lineages exhibit characteristic enrichments of COG-FC, and that the extent of
159 variation varies among taxonomic ranks. Of all the taxonomic ranks, phylum was the most powerful
160 predictor of COG-FC enrichment, which is consistent with observations that phylum can be informative
161 of microbial function (e.g., 10–12). Surprisingly, lower taxonomic ranks such as genus and family had
162 little explanatory power. Many studies focus on metabolic coherence of individual traits and regularly
163 find traits conserved on the family level (2, 3, 13). The discrepancy between previous observations and
164 our observation likely relates to how we characterize patterns in metabolic potential. The tradeoff of the
165 approach used here is that, by analyzing COG-FCs, we lose information about specific genes or
166 potential metabolic functions but gain the ability to apply a consistent analysis across an entire genome
167 and across the entire microbial tree of life. Thus, the extent that observed patterns (Fig 1) reflect
168 phenotypically-expressed differences among lineages is unknown. Nonetheless, the statistical robustness
169 of the relationship between all taxonomic ranks and COG-FC patterns suggests that evolutionary
170 processes (e.g., horizontal gene transfer, vertical gene transfer, duplications, deletions, etc.) control the
171 preponderance of different COG-FCs across lineages.

172 The role that individual evolutionary processes play in influencing COG-FC enrichments at a

173 given taxonomic rank is likely variable. For instance, horizontal gene transfer is more common among
174 more closely related lineages (14) and thus, likely promotes increased levels of similarity at lower
175 taxonomic ranks. At higher taxonomic ranks, vertical processes may be more important. The asymptote
176 in the mean $\log_{10}$-distance from the centroid as function of lineages in a phylum suggests that identifying
177 more lineages for more poorly represented lineages should expand the diversity of COG-FCs that are
178 found, whereas phyla that were adequately sampled (at least ~1000 lineages) exhibited comparable
179 variability in COG-FC distributions (Fig 3B). Since many more than ~1000 distinct lineages of each
180 phylum are likely to exist (15), we propose that the taxonomic rank of phylum implies a fairly consistent
181 degree of diversity in COG-FC distribution. To the extent that phenotype matches genotype at the level
182 of COG-FC distributions, therefore, we expect that typical phyla exhibit similar phenotypic diversity. A
183 notable exception is the phylum Crenarchaeota, which were far more diverse than would be expected
184 based on the number of lineages sampled. The Crenarchaeota, as defined in the GTDB, collapsed
185 members of several phyla that had been designated separately under previous taxonomies, including
186 lineages that had previously been assigned as Crenarchaeota, Thaumarchaeota, Euryarchaeota,
187 Verstraetearchaeota, Korarchaeota, and Bathyarchaeota (5). It is possible that the relationship between
188 marker genes used in the GTDB and the rest of the genome is unusual for this clade, compared to other
189 phyla, or that the GTDB classification of Crenarchaeota is lacking in some other way.

190 Although the ranks, genus and family, explained relatively little of the variance in COG-FC
191 distribution, blocks of consistent colors were evident in Fig 4, indicating that enrichments or depletions
192 of specific COG-FCs were conserved across each taxonomic rank in some parts of the phylogenetic tree.
193 This is explained by 'distantly' (i.e., non-sister clades) related clades occupying similar COG-FC trait-
194 space. The coherence in metabolic potential at higher taxonomic ranks may help explain the distribution
195 of microbial clades across ecological niches. Analyses of habitat associations (9, 16) and single traits

196 (1–3) also support this idea. Our analysis provides quantitative evidence to this idea by demonstrating
197 coherence in metabolic potential with broad-scale patterns in genomic data (Fig 1-4). The question
198 remains: how well do the observed COG-FC enrichments reflect expressed functional traits (i.e.,
199 phenotype) across these lineages? It is difficult to address this question systematically, but some of the
200 enrichments and depletions here appear consistent with known physiologies of clades. For instance,
201 Rickettsiales were depleted in nucleotide metabolism and transport, consistent with previously observed
202 lack of a metabolic pathway for purine synthesis among five example Rickettsiales (17). Another
203 example is the depletion in the COG-FCs for energy production and conversion, amino acid transport
204 and metabolism, and carbohydrate transport and metabolism within the Bacteroidetes, Spirochaetes, and
205 Chlamydiales clade. This clade is known to contain many host-dependent pathogens and symbionts (18–
206 20), which are often depleted in these COG-FCs (21).

207 The GTDB classification is the first fully algorithmic and quantitatively self-consistent microbial
208 taxonomy that can be applied across the tree of life (5). By standardizing the meaning of taxonomic
209 ranks, it creates an objective basis on which to compare microbial functionality to phylogeny. The data
210 presented here support the idea that clades in this taxonomy are functionally coherent, as well as the idea
211 that all taxonomic ranks provide information about the metabolic and ecological niches that individual
212 microbes occupy.

213 Materials and Methods

214 Genome Database Curation

215 All bacterial and archaeal genomes from the RefSeq database v92 (22), uncultured bacterial and
216 archaeal (UBA) metagenome-assembled genomes (MAGs) reported in Parks et al. (5, 23), bacterial and
217 archaeal MAGs from Integrated Microbial Genomes and Microbiomes (IMG/M), and bacterial and

218 archaeal single amplified genomes (SAGs) from IMG/M were curated into a single database. All
219 genomic content within the curated database is referred to as "genome(s)" for simplicity. Genomes were
220 assigned taxonomy consistent with the Genome Taxonomy Database (GTDB) using the GTDB toolkit
221 (GTDB-Tk) v0.2.1 (5). The GTDB-Tk taxonomic assignments were consistent with reference package
222 GTDB r86. Lineages which did not receive a genus classification, due to the absence of a reference
223 lineage were excluded from analyses. Due to bias in the abundance of strains in specific clades (e.g., *E.*
224 *coli*), the lowest taxonomic rank considered during our analysis was species. The COG-FC enrichments
225 (see below) were averaged together for all strains within a given species. An exception was made for
226 lineages which shared a genus classification but lacked a species classification. In this scenario, each
227 genome was treated as an independent lineage. Lastly, only genomes belonging to genera with at least
228 ten unique species in the database were retained. This criterion ensured enough data to generate
229 meaningful statistics during our PERMANOVA. The final database is summarized in Table 1. The
230 genus-level phylogenetic tree was generated from concatenated protein sequence alignments published
231 in Parks et al. (5).

232 COG FUNCTIONAL CATEGORY IDENTIFICATION, ENUMERATION, AND NORMALIZATION

233 Genes were predicted from individual genomes and translated into protein sequences using
234 Prodigal v.2.6.3 (24). The resulting protein sequences were analyzed for COGs (8). COG position-
235 specific scoring matrices (PSSMs) were downloaded from NCBI's Conserved Domain Database (27
236 March 2017 definitions). COG PSSMs were BLASTed against protein sequences with the Reverse
237 Position Specific-BLAST (RPS-BLAST) algorithm (25). Following previously a reported protocol (25),
238 we used an E-value cutoff of 0.01 to assign COGs with RPS-BLAST. The retrieved COGs were
239 assigned to their respective COG functional categories (COG-FCs; 25 in total) and the abundance of
240 each functional category was tabulated using cdd2cog (26) for each genome. The abundance for

241 individual COG-FCs was normalized by the respective COG-FC standard deviation across all lineages.
242 For the COG-FCs, extracellular structures and nuclear structures, the standard deviation was 0.
243 Consequently, data could not be normalized, and thus, these two categories were discarded from all
244 analyses.

245 COG-FC abundances were normalized by their respective regression slopes of COG-FC
246 abundance for a given genome as a function of genome size. COG-FC abundances were modelled as a
247 function of genome size for individual categories using a generalized additive model (GAM) with a
248 smoothing term due to the pairwise response to genome size (Sup. Figure 1). We used the gam function
249 from the R package, mgcv (27). In some instances, regression fits were visibly skewed by high-leverage
250 data points. High-leverage data were filtered using the influence.gam function in the mgcv package.
251 Data in the 99.5% percentile for influence were excluded when performing regression analysis but were
252 included in all downstream analyses. All regressions were significant with $p<0.001$.

253 PRINCIPAL COMPONENT ANALYSIS (PCA)

254 We performed PCA on the normalized COG-FC abundances and enrichments. Prior to PCA,
255 assumptions of normality were achieved by performing a boxcox transformation on individual COG-FC
256 abundance and enrichment distributions with the boxcox function from the R Package, MASS (28). The
257 resulting distributions were then scaled by the respective COG-FC standard deviation calculated from all
258 genomes. PCA was performed using the princomp function from the R package, stats (29).

259 QUANTIFYING COG-FC VARIANCE EXPLAINED BY TAXONOMIC RANK

260 We performed permutational multivariate analysis of variance (PERMANOVA) using the adonis
261 function from the R package, vegan (30). The taxonomic ranks domain, phylum, class, order, family,
262 and genus as well as cultured-status were used as test categorical variables for quantifying variance in

263 COG-FC enrichments explained by the mean taxonomic rank centroids. The default, 999 permutations
264 test, was performed using each categorical variable. Distances were calculated between mean phyla
265 COG-FC enrichment centroids and the respective genomes within that phyla by performing an analysis
266 of multivariate homogeneity of groups dispersions with the betadisper function from the R package,
267 vegan (30). The distance matrix used for both the adonis and betadisper analyses was generated
268 calculating Euclidean distance on the normalized COG-FC enrichments.

269 The mean $\log_{10}$-distance from phylum centroid for each phylum and modeled with the following
270 equation, which represents a hyperbola shifted on the x-axis to ensure that mean distance is zero when n
271 = 1:

272 $$log_{10}(Mean\ Distance) = \frac{A(log_{10}(n)-1)}{B+log_{10}(n)-1} + C \quad (1$$

273 where A, B, and C are fit coefficients and n is the total number of lineages in the given phylum. The
274 Akaike Information Criterion was calculated with the fit from eq 1 using the AIC function from the R
275 package, stats (29).

276 DATA AVAILABILITY

277 The genomes analyzed for the current study are available in NCBI's RefSeq database
278 (ftp://ftp.ncbi.nlm.nih.gov/refseq/release/). UBA MAGs used for the current study are available under
279 NCBI's BioProject PRJNA417962 and PRJNA348753. JGI IMG/M acquired from Chad Burdyshaw.
280 Associated genome accessions are available in Dataset S1.

281

282

Acknowledgments

Funding for this project was provided by a C-DEBI graduate fellowship to TR and an-kind grant of resources from the University of Tennessee / Oak Ridge National Lab Joint Institute for Computational Sciences (JICS). We thank Chad Burdyshaw for help obtaining the genomes used in this project. This is C-DEBI contribution number [to be assigned upon manuscript acceptance].

References

1. Martiny JBH, Jones SE, Lennon JT, Martiny AC. 2015. Microbiomes in light of traits: A phylogenetic perspective. Science (80-) 350:aac9323.

2. Martiny AC, Treseder K, Pusch G. 2013. Phylogenetic conservatism of functional traits in microorganisms. ISME J 7:830–838.

3. Zimmerman AE, Martiny AC, Allison SD. 2013. Microdiversity of extracellular enzyme genes among sequenced prokaryotic genomes. ISME J 7:1187–1199.

4. Federhen S. 2012. The NCBI Taxonomy database. Nucleic Acids Res 40:136–143.

5. Parks DH, Chuvochina M, Waite DW, Rinke C, Skarshewski A, Chaumeil P-A, Hugenholtz P. 2018. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. Nat Biotechnol 36:996–1004.

6. Lloyd KG, Steen AD, Ladau J, Yin J, Crosby L. 2018. Phylogenetically Novel Uncultured Microbial Cells Dominate Earth Microbiomes. mSystems 3:1–12.

7. Hug LA, Baker BJ, Anantharaman K, Brown CT, Probst AJ, Castelle CJ, Butterfield CN, Hernsdorf AW, Amano Y, Ise K, Suzuki Y, Dudek N, Relman DA, Finstad KM, Amundson R,

304 Thomas BC, Banfield JF. 2016. A new view of the tree of life. Nat Microbiol 1:1–6.

305 8. Tatusov RL, Galperin MY, Natale DA, Koonin E V. 2000. The COG database: a tool for genome-
306 scale analysis of protein functions and evolution. Nucleic Acids Res 28:33–36.

307 9. Philippot L, Andersson SGE, Battin TJ, Prosser JI, Schimel JP, Whitman WB, Hallin S. 2010.
308 The ecological coherence of high bacterial taxonomic ranks. Nat Rev Microbiol 8:523–529.

309 10. Danczak RE, Johnston MD, Kenah C, Slattery M, Wrighton KC, Wilkins MJ. 2017. Members of
310 the Candidate Phyla Radiation are functionally differentiated by carbon- and nitrogen-cycling
311 capabilities. Microbiome 5:112.

312 11. Emerson D, Fleming EJ, McBeth JM. 2010. Iron-Oxidizing Bacteria: An Environmental and
313 Genomic Perspective. Annu Rev Microbiol 64:561–583.

314 12. Singh AH, Doerks T, Letunic I, Raes J, Bork P. 2009. Discovering functional novelty in
315 metagenomes: Examples from light-mediated processes. J Bacteriol 91:32–41.

316 13. Mendler K, Chen H, Parks DH, Lobb B, Hug LA, Doxey AC. 2019. AnnoTree: visualization and
317 exploration of a functionally annotated microbial tree of life. Nucleic Acids Res 4567:1–7.

318 14. Soucy SM, Huang J, Gogarten JP. 2015. Horizontal gene transfer: Building the web of life. Nat
319 Rev Genet 16:472–482.

320 15. Schloter M, Lebuhn M, Heulin T, Hartmann A. 2000. Ecology and evolution of bacterial
321 microdiversity. FEMS Microbiol Rev 24:647–660.

322 16. Koeppel AF, Wu M. 2012. Lineage-dependent ecological coherence in bacteria. FEMS Microbiol
323 Ecol 81:574–582.

324 17. Min CK, Yang JS, Kim S, Choi MS, Kim IS, Cho NH. 2008. Genome-based construction of the

325 metabolic pathways of Orientia tsutsugamushi and comparative analysis within the Rickettsiales
326 order. Comp Funct Genomics 2008.

327 18. Wolgemuth CW. 2015. Flagellar motility of the pathogenic spirochetes. Semin Cell Dev Biol
328 46:104–112.

329 19. Dreyer M, Aeby S, Oevermann A, Greub G. 2015. Prevalence and diversity of Chlamydiales in
330 Swiss ruminant farms. Pathog Dis 73:1–4.

331 20. Wexler HM. 2007. Bacteroides: The good, the bad, and the nitty-gritty. Clin Microbiol Rev
332 20:593–621.

333 21. Moran NA. 2002. Microbial Minimalism: Genome Reduction in Bacterial Pathogens. Cell
334 108:583–586.

335 22. O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, Rajput B, Robbertse B,
336 Smith-White B, Ako-Adjei D, Astashyn A, Badretdin A, Bao Y, Blinkova O, Brover V,
337 Chetvernin V, Choi J, Cox E, Ermolaeva O, Farrell CM, Goldfarb T, Gupta T, Haft D, Hatcher E,
338 Hlavina W, Joardar VS, Kodali VK, Li W, Maglott D, Masterson P, McGarvey KM, Murphy
339 MR, O'Neill K, Pujar S, Rangwala SH, Rausch D, Riddick LD, Schoch C, Shkeda A, Storz SS,
340 Sun H, Thibaud-Nissen F, Tolstoy I, Tully RE, Vatsan AR, Wallin C, Webb D, Wu W, Landrum
341 MJ, Kimchi A, Tatusova T, DiCuccio M, Kitts P, Murphy TD, Pruitt KD. 2016. Reference
342 sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional
343 annotation. Nucleic Acids Res 44:D733–D745.

344 23. Parks DH, Rinke C, Chuvochina M, Chaumeil P-A, Woodcroft BJ, Evans PN, Hugenholtz P,
345 Tyson GW. 2017. Recovery of nearly 8,000 metagenome-assembled genomes substantially
346 expands the tree of life. Nat Microbiol 2:1533–1542.

347 24. Hyatt D, Chen G-L, LoCascio PF, Land ML, Larimer FW, Hauser LJ. 2010. Prodigal: prokaryotic
348 gene recognition and translation initiation site identification. BMC Bioinformatics 11:119.

349 25. Marchler-Bauer A, Bryant SH. 2004. CD-Search: Protein domain annotations on the fly. Nucleic
350 Acids Res 32:W327–W331.

351 26. Leimbach A, Poehlein A, Witten A, Wellnitz O, Shpigel N, Petzl W, Zerbe H, Daniel R, Dobrindt
352 U. 2016. Whole-Genome Draft Sequences of Six Commensal Fecal and Six Mastitis-Associated
353 *Escherichia coli* Strains of Bovine Origin: TABLE 1. Genome Announc 4:e00753-16.

354 27. Wood S. 2017. mgcv: Mixed GAM Computation Vehicle with GCV/AIC/REML Smoothness
355 Estimation.

356 28. Venables WN, Ripley BD. 2002. Modern Applied Statistics with S. 7.3.49. Springer, New York.

357 29. R Core Team. 2018. R: A Language and Environmen for Statistical Computing. Vienna, Austria.

358 30. Philip D. 2003. Computer program review VEGAN , a package of R functions for community
359 ecology. J Veg Sci 14:927–930.

360

361

Tables

Table 1. A summary of the custom-curated genome database used in this work.

Unique Domains	Unique Phyla	Unique Classes	Unique Orders	Unique Families	Unique Genera	Unique Lineages	Cultured Lineages	Uncultured Lineages
Bacteria	Actinobacteriota	3	9	22	50	2286	2115	171
	Bacteroidota	3	7	19	50	1606	741	865
	Campylobacterota	1	1	6	8	270	203	67
	Cyanobacteriota	2	3	4	7	119	84	35
	Deinococcota	1	1	2	2	44	44	0
	Desulfobacterota	2	2	2	4	43	23	20
	Elusimicrobiota	1	1	1	1	22	0	22
	Fibrobacterota	1	1	1	1	34	22	12
	Firmicutes	3	10	23	48	1543	1356	187
	Firmicutes A	2	7	10	31	600	304	296
	Firmicutes B	1	1	1	1	22	11	11
	Firmicutes C	1	2	3	4	53	32	21
	Fusobacteriota	1	1	2	2	40	40	0
	Marinisomatota	1	1	1	1	10	0	10
	Nitrospirota	2	2	2	2	30	6	24
	Nitrospirota A	1	1	1	1	14	2	12
	Patescibacteria	6	16	26	36	707	0	707
	Proteobacteria	3	25	59	163	5589	3952	1637
	Spirochaetota	3	4	4	6	153	89	64
	Synergistota	1	1	1	1	19	2	17
	Thermotogota	1	1	2	2	23	15	8
	Verrucomicrobiota	2	4	5	7	84	16	68
Archaea	Crenarchaeota	1	1	1	2	68	7	61
	Euryarchaeota	2	2	2	2	45	26	19
	Halobacterota	4	5	7	9	164	97	67
	Thermoplasmatota	1	1	2	9	147	0	147
Total	26	50	110	209	450	13,735	9187	4548

365 **Table 2.** Phylum highly enriched (>85th percentile) or depleted (<15th percentile) in COG-FCs and
366 depletion. All reported categories are statistically significant.

Phylum	Enriched COG-FC	Depleted COG-FC
Actinobacteriota	--	4,6
Bacteroidota	--	11,18,22
Campylobacterota	4,6,10,15,19	8,12
Cyanobacteriota	19	12
Deinococcota	--	15
Desulfobacterota	13	7
Elusimicrobiota	3,10,15	14,16
Fibrobacterota	--	7,11,12,13,16, 18,20,22
Firmicutes	9,12,21	--
Firmicutes A	9	5,7,16,20
Firmicutes B	13,17,19	--
Firmicutes C	19	--
Fusobacteriota	--	10,21
Marinisomatota	--	12,14
Nitrospirota	6,10,17	8
Nitrospirota A	4,6,10,15	12,17,22,23
Patescibacteria	--	7,11,16,19
Proteobacteria	--	--
Spirochaetota	4	5,16,19
Synergistota	3,4,11,16	--
Thermotogota	3,4,8,10	--
Verrucomicrobiota	1	10,12,14,21, 23
Crenarchaeota	3,13,11,19,7, 12,16,5,22	9,10,14,15,17
Euryarchaeota	2,3,13,18,22	5,7,10,14,15
Halobacterota	3,19,22	15,16
Thermoplasmatota	1,2,7,13	4,6,8,9,10,14, 15,16

COG-FC	Table Key
Cytoskeleton	1
RNA Processing and Modification	2
Chromatin Structure and Dynamics	3
Cell Motility	4
Secondary Metabolites Biosynthesis, Transport, and Catabolism	5
Intracellular Trafficking, Secretion, and Vesicular Transport	6
Lipid Transport and Metabolism	7
Carbohydrate Transport and Metabolism	8
Defense Mechanism	9
Signal Transduction Mechanisms	10
Amino Acid Transport and Metabolism	11
Transcription	12
Energy Production and Conversion	13
Replication, Recombination, and Repair	14
Cell Wall/Membrane/Envelope Biogenesis	15
Inorganic Ion Transport and Metabolism	16
Cell Cycle Control, Cell Division, and Chromosome Partitioning	17
Function Unknown	18
Coenzyme Transport and Metabolism	19
Post-translational Modification, Protein Turnover, and Chaperone	20
Nucleotide Transport and Metabolism	21
General Function Prediction Only	22
Translation Ribosomal Structure and Biogenesis	23

367

368

Figures

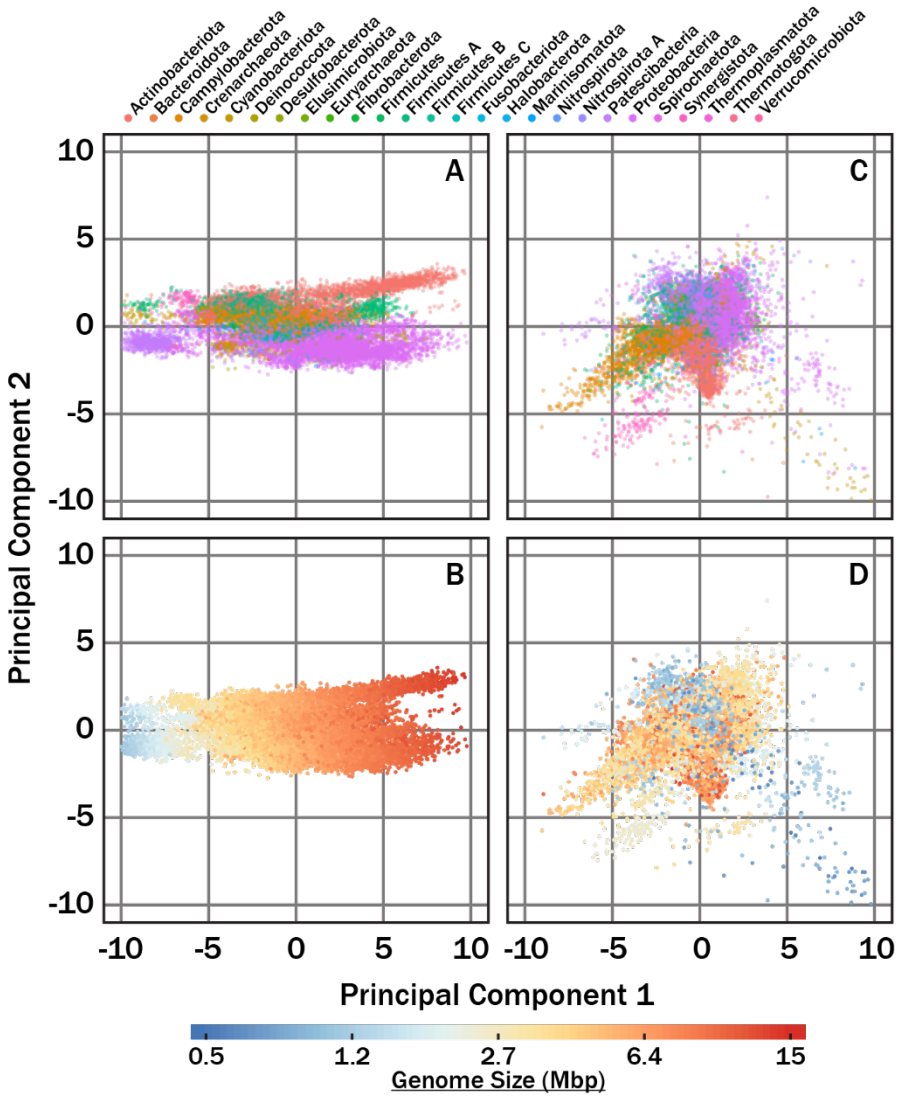

Fig 1. PCA plots of COG-FC abundance (A,B) and enrichment (C,D) of the lineages summarized in Table 1. Individual data points are colored by genome size (A,C) and phylum (B, D). Panels A and B were not normalized by genome size while panels C and D were normalized by genome size. For panels A and B, PC1 explained 71% and PC2 explained 7.0% of variance. For panels C and D, PC1 explained 21% and PC2 explained 16% of variance.

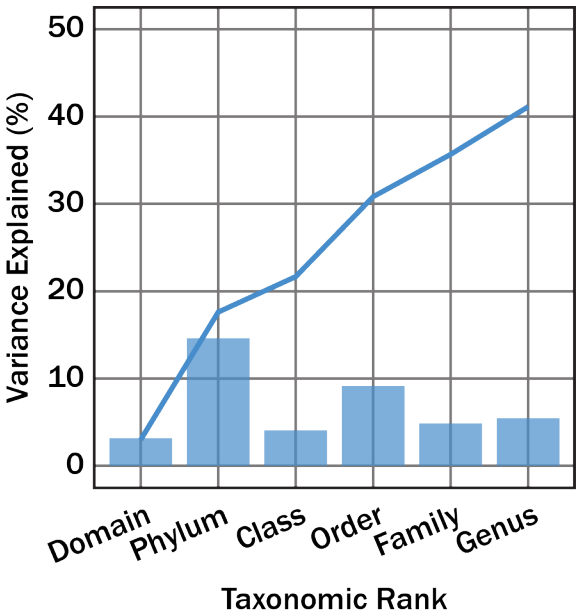

Fig 2. The average variance in COG-FC enrichment explained by different taxonomic ranks (bars) and the cumulative variance explained by taxonomic ranks (lines). All variance explained by taxonomic ranks was significant ($p<0.001$). The F-value for domain, phylum, class, order, family, and genus, was 726.0, 128.8, 38.76, 34.4, 11.2, and 5.1, respectively.

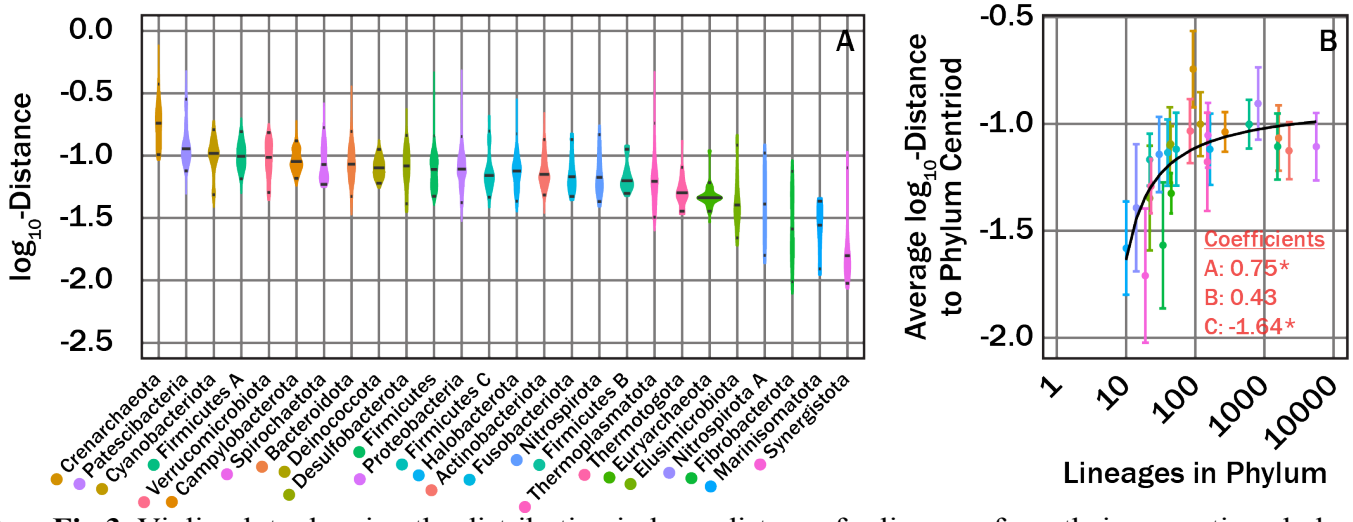

Fig 3. Violin plots showing the distribution in $\log_{10}$-distance for lineages from their respective phyla mean COG-FC centroid (A) and average $\log_{10}$-distance to phyla mean COG-FC centroid as a function of number of lineages in a given phylum (B). Coefficients in panel B correspond to fit parameters from eq 1. The * symbols denote significance as defined in the text. We note three outliers: the Crenarhaeota are characterized by unusually high diversity of COG-FCs distributions, and the Synergistota and Fibrobacterota are characterized by an unusually low diversity of COG-FC distributions.

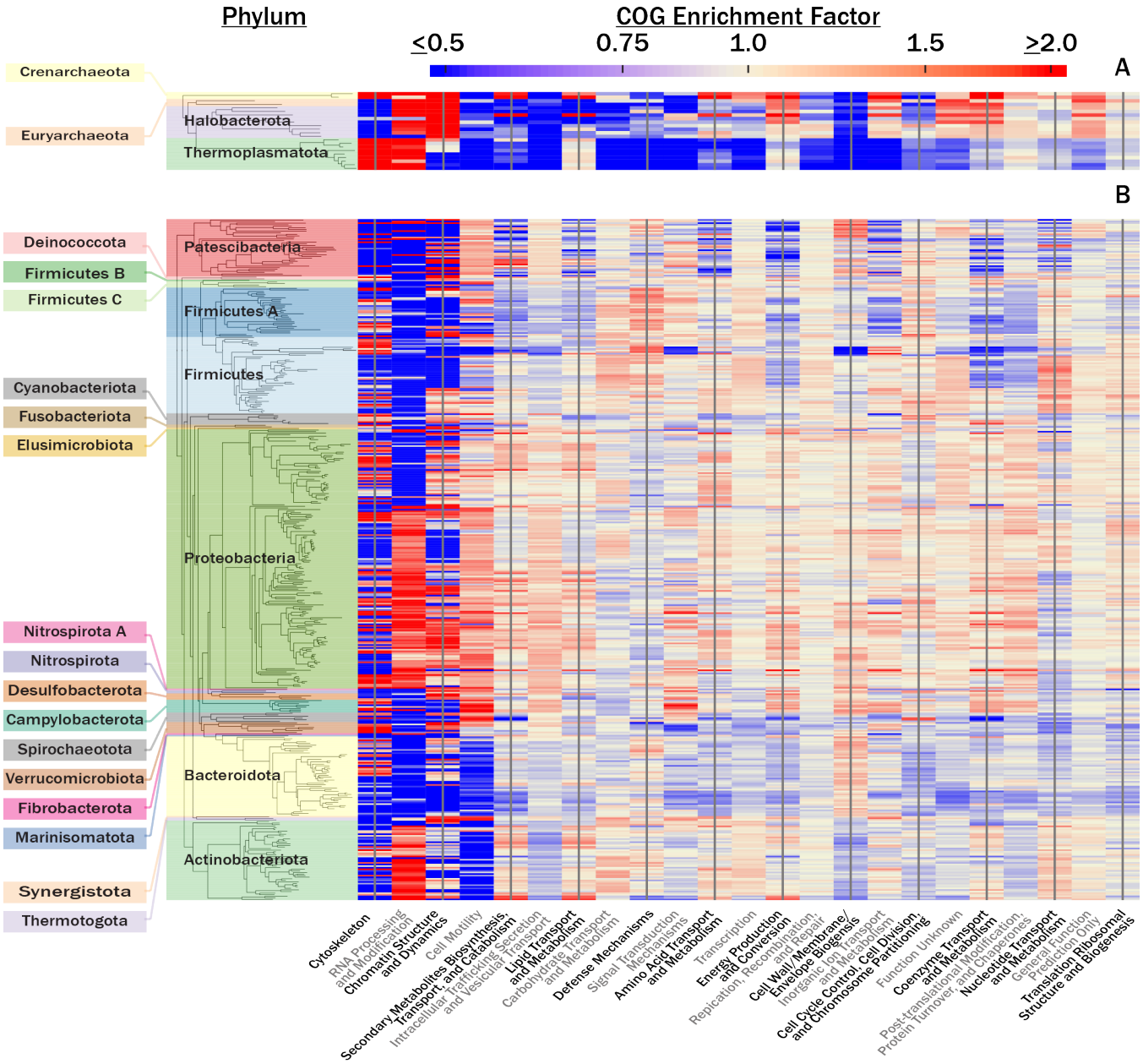

Fig 4. A heat map showing the average COG-FC enrichment for all Archaea genera (A) and Bacteria genera (B). Categories were arranged from left to right along the x axis in order of decreasing total variance in enrichment across all lineages. Clades were organized along the y axis using phylogenetic relatedness based on the reported concatenated protein sequence alignments in Parks et al. (1).