Partitioning of microbial function among taxonomic ranks across the tree of life

Taylor M. Royalty
Andrew D. Steen

Department of Earth and Planetary Sciences, University of Tennessee, Knoxville

Concatenated protein sequence alignments are a reliable tool to determine phylogenetic relationships among microbes[1] but do not necessarily imply microbial function[2]. Microbiomes influence their environments via the collective action of microbial metabolisms, so it is important to develop a quantitative understanding of the degree to which taxonomy provides information about the function of microbes. However, the overwhelming majority of microbial lineages are resistant to culturing[3], leading to difficulties characterizing their metabolisms. Here we quantify the enrichment of clusters of orthologous gene functional categories (COG-FCs)[4], a proxy for metabolic potential, within the genomes of 3207 cultured and uncultured microbe lineages from a custom-curated genome database. Results show that, on average, 36.7% of the variation in COG-FC enrichment is explained by taxonomic rank, with domain, phylum, class, order, family, and genus explaining, on average, 2.3%, 12.9%, 2.6%, 9.4%, 6.1%, and 3.5% of the variance, respectively (p<0.001 for all). Furthermore, taxonomic rank explained 81.5% of the variance when a only 18 of the 23 COG-FC categories were considered for analysis. To our knowledge, this is the first work which quantifies the variance in metabolic potential from individual taxonomic ranks. A qualitative comparison between the COG-FC enrichments and order-level phylogenies, generated from concatenated protein sequence alignments, further supports the idea that metabolic potential is taxonomically coherent at higher taxonomic ranks. The coherence in metabolic potential at higher taxonomic ranks provides support to hypotheses which suggest that microbial lineages may be biased to occupy specific ecological or metabolic niches[5].

High-performance sequencing and related bioinformatic techniques have led to many newly identified microbial lineages and a reorganization of the tree of life[1,3]. However, understanding of the distribution of functional traits among lineages still lags[7]. The challenge of identifying novel genes and experimentally determining the function of their products is partially to blame. Analyses measuring the conservation of functional traits across lineages often trace single genes through phylogenies[2,8]. This approach works well for carefully-chosen genes within specific lineages but cannot be applied across

34 the tree of life. As such, the role that phylogeny has in influencing the distribution of functional traits
35 across microbial lineages is at best, poorly unconstrained, and at worst, incompletely constrained.
36 Here we explored the distribution of metabolic potential across 3207 microbial lineages from a
37 custom-curated database. The databased was built with genomes from RefSeq[9] and metagenome-
38 assembled genomes published by Parks et al.[1,10] (Table 1; Supplementary Dataset 1). Variations in the
39 abundance of COG-FCs were used as a proxy for trait variation. COG-FCs have previously been used
40 to characterize broad-scale trends across microbial lineages, for instance, identifying last universal
41 common ancestor's metabolism[11], the rapid emergence of metabolic diversity during the Archaean
42 Explosion[12], and the construction of "functional phylogenies"[13]. In the dataset described here, the
43 abundance of COG-FC within a genome correlated positively with genome size (Supplementary Figure
44 1). The inter-relationship between COG-FC abundance, genome size, and phylogeny was also clear in
45 non-metric multidimensional scaling (NMDS) plots of the scaled COG-FC abundances (Figure 1a,b).
46 The observation that genome size is conserved at higher taxonomic ranks is not novel[5]. Our objective
47 was to measure how metabolic potential partitions among taxonomic ranks, so for downstream
48 analyses, COG-FC abundances required genome size normalization. Variations in the explanatory
49 power of genome size for COG-FC abundances implied that other extrinsic variables influence the
50 preponderance of COG-FCs, so normalizing by genome size may increase noise rather than reducing it.
51 To avoid this problem, we normalized each scaled COG-FC abundance, for each genome, by a
52 generalized additive model (GAM) prediction of annotation abundance as a function of genome size
53 (see methods). This generated COG-FC enrichments relative to the expected COG-FC abundance for
54 given a genome size. A second NMDS of COG-FC enrichments, opposed to abundances, leads to the
55 loss of the strong, inter-phyla gradient in genome size while conserving the spatial relationship of
56 phylogeny (Figure 1 c,d). The complete removal of a gradient in genome size is not expected, as
57 genome size depends on phylogeny[5]. Still, the removal of the inter-phyla genome size gradient
58 indicates that our normalization removed the first-order effect of genome size.

59 To quantify the degree to which taxonomic rank explains the distribution of COG-FC
60 enrichments among individual genomes, we performed a permutation multivariate ANOVA
61 (PERMANOVA) using the following taxonomic ranks: domain, phylum, class, order, family, and
62 genus, as well as culture-status (cultured versus uncultured). The rank of species was excluded from the
63 analysis as every lineage was unique, and thus, species would explain 100% of the data. Every rank
64 significantly ($p<0.001$) influenced the distribution of COG-FC enrichments, but surprisingly, phylum
65 explained the most variance (12.9%), followed by order (9.4%). In total, an organism's full taxonomy
66 to the rank of genus explained 36.7% of the variance in COG-FC enrichment (Figure 2). In contrast,
67 culture-status failed to reject the null hypothesis ($p\leq0.05$), consistent with "unculturedness" not being
68 associated with enrichment of COG-FCs. The underrepresentation of uncultured lineages may have
69 contributed to this observation (3089 cultured lineages versus 128 uncultured lineages after data
70 filtering and averaging COG-FC content of strains of the same species; see methods). In order to
71 determine whether the explanatory power of taxonomic rank was biased by COG-FCs with low
72 genome size dependency, we performed a second PERMANOVA excluding five COG-FCs for which
73 the GAM explained <50% of COG-FC abundance. According to this analysis, a microbes's full
74 taxonomy explains ~81.5% variance based on 18 COG-FC categories. The relative importance of
75 taxonomic ranks did not change in this analysis, and all the additional explanatory power was
76 partitioned to the ranks of phylum, order, family, and genus. This analysis indicates that domain and
77 class offer essentially no information about an organism's metabolic potential while a full taxonomy
78 can predict the relative abundance of COG-FCs within a genome relatively accurately.
79 Of all the taxonomic ranks, phylum was the most powerful predictor of COG-FC abundance,
80 consistent with the observation that phylum is typically informative of microbial function. We therefore
81 quantified the average enrichment for all COG-FCs across all taxonomic orders to visualize this effect.
82 These enrichments were then qualitatively compared to an order-level phylogenetic tree (Figure 3) built
83 from concatenated ribosomal protein sequences published by Parks et al.[1]. Numerous coherent features

84 appear in COG-FC enrichments. For example, a prominent feature for the clade consisting of
85 Bacteroidetes, Spirochaetes, and Verrucomicrobia is a paucity of the less variable COG-FCs, including
86 energy production and conversion, amino acid transport and metabolism, and carbohydrate transport
87 and metabolism, among others. Another prominent feature is the near-ubiquitous enrichment of cell
88 motility, secondary metabolites biosynthesis, transport, and catabolism, lipid transport and metabolism,
89 and intracellular trafficking, secretion, and vesicular transport COGs in Proteobacteria. Interestingly, a
90 notable dichotomy in the enrichment of RNA processing and modification mirrors the division of the
91 two largest clades within the proteobacteria. Not surprisingly, the three Archaea orders have similar
92 COG-FC enrichments. Overall, enrichment data qualitatively appears consistent with phylogenetic
93 relationships. The coherence in metabolic potential at higher taxonomic ranks may help explain the
94 distribution of microbial clades across ecological niches. Analyses of habitat associations[5,15] and single
95 traits[2,7,8] also support this idea. Our analysis provides concrete evidence to this idea by demonstrating
96 coherence in metabolic potential with broad-scale patterns in genomic data (Figure 1-3).
97 The question remains: how well do the observed COG-FC enrichments (Figure 3) reflect
98 expressed functional traits (i.e., phenotype) across these lineages? Broad assertions are difficult to
99 validate based on the literature due to the recent development of the Genome Taxonomy Database
100 (GTDB) taxonomic classification[1] (adopted here) and the sheer magnitude of the data in Figure 3.
101 Nonetheless, we investigated some anomalies in our data. For instance, Rickettsiales were depleted in
102 nucleotide metabolism and transport, consistent with previously observed lack of a metabolic pathway
103 for purine synthesis among five example Rickettsiales[16]. Another example is the depletion in the COG-
104 FCs for energy production and conversion, amino acid transport and metabolism, and carbohydrate
105 transport and metabolism within the Bacteroidetes, Spirochaetes, and Chlamydiales clade. This clade is
106 known to contain many host-dependent pathogens and symbionts[17–19], which are often deplete in these
107 COG-FCs[20].

108 Lastly, the adoption of a new taxonomic classification system presents an opportunity for
109 recharacterizing the metabolic tendencies of taxonomic groups. The qualitative and quantitative data
110 presented here supports coherence in metabolic potential across higher taxonomic orders using the
111 GTDB taxonomic classification while also implicating a possible bias in niche occupation by certain
112 clades.

113 **Methods Summary:**

114 *COG Functional Category Identification, Enumeration, and Normalization*

115 Genes were predicted from individual genomes and translated into protein sequences using Prodigal
116 v.2.6.3[21]. The resulting protein sequences were analyzed for COGs[4]. COG position-specific scoring
117 matrices (PSSMs) were downloaded from NCBI's Conserved Domain Database (27 March 2017
118 definitions). COG PSSMs were BLASTed against protein sequences with the Reverse Position
119 Specific-BLAST (RPS-BLAST) algorithm[22]. Following previously a reported protocol[23], we used an E-
120 value cutoff of 0.01 to assign COGs with RPS-BLAST. The retrieved COGs were assigned to their
121 respective COG functional categories (COG-FCs; 25 in total) and the abundance of each functional
122 category was tabulated using cdd2cog[24] for each genome. The abundance for individual COG-FCs was
123 normalized by the respective COG-FC standard deviation across all lineages. For the COG-FCs,
124 extracellular structures and nuclear structures, the standard deviation was 0. Consequently, data could
125 not be normalized, and thus, these two categories were discarded from all analyses.

126 COG-FC abundances were normalized by their respective regression slopes of COG-FC
127 abundance for a given genome as a function of genome size. COG-FC abundances were modelled as a
128 function of genome size for individual categories using a generalized additive model (GAM) with a
129 smoothing term due to the pairwise response to genome size (Sup. Figure 1). We used the gam function
130 from the R package, mgcv[25]. All regressions were significant with $p<0.001$.

131 *COG-FC Enrichments Significance*

132 We tested the COG-FC enrichment significance by comparing the 95% confidence interval for a given
133 COG-FC enrichment of a given taxonomic order against the 95% confidence interval for the same
134 COG-FC enrichment across all lineages. The 95% confidence intervals were generated by a 10^5
135 iteration bootstrap analysis. If 95% confidence intervals overlapped for a given COG-FC, then the
136 COG-FC enrichment was not considered significant.

137 *Non-metric Multidimensional Scaling (NMDS)*

138 We performed NMDS on the normalized COG-FC abundances and enrichments. NMDS was
139 performed using the metaMDS function from the R package, vegan[26]. Euclidean distance was selected
140 as the distance metric for NMDS. Three dimensions were considered for NMDS analysis (only NMDS
141 1 and NMDS 2 shown) as to reduce stress to ~0.1.

142 *Permutational Multivariate Analysis of Variance (PERMANOVA)*

143 We performed PERMANOVA using the adonis function from the R package, vegan[26]. The taxonomic
144 ranks domain, phylum, class, order, family, and genus as well as cultured-status were used as test
145 categorial variables for quantifying variance based on a default, 999, permutation test for each variable.
146 A distance matrix was generated calculating Euclidean distance on the normalized COG-FC
147 enrichments.

148 *Genome Database Curation*

149 All bacterial and archaeal genomes from the RefSeq database v86[9] and uncultured bacterial and
150 archaeal (UBA) metagenome-assembled genomes (MAGs) reported in Parks et al.[1,10] were curated into
151 a single database. UBA MAGs were filtered to only include genomes with >90% completion. All
152 genomic content within the curated database is referred to as "genomes." Genomes were assigned
153 taxonomy consistent with the Genome Taxonomy Database[1]. Due to bias in the abundance of strains in
154 specific clades, the lowest taxonomic rank considered during our analysis was species. The COG-FC
155 enrichments were averaged together for all strains within a given species. We filtered all genera with
156 <10 unique species. This criterion ensured enough data to generate meaningful statistics during our

157 PERMANOVA. The final database is summarized in Table 1. The order-level phylogenetic tree was

158 generated from concatenated protein sequence alignments published in Parks et al.[1] and graphically

159 generated with Interactive Tree of Life v2[27].

160

References

1. Parks, D. H. *et al.* A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nat. Biotechnol.* **36,** 996–1004 (2018).
2. Zimmerman, A. E., Martiny, A. C. & Allison, S. D. Microdiversity of extracellular enzyme genes among sequenced prokaryotic genomes. *ISME J.* **7,** 1187–1199 (2013).
3. Hug, L. A. *et al.* A new view of the tree of life. *Nat. Microbiol.* **1,** 16048 (2016).
4. Tatusov, R. L., Galperin, M. Y., Natale, D. A. & Koonin, E. V. The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res.* **28,** 33–36 (2000).
5. Philippot, L. *et al.* The ecological coherence of high bacterial taxonomic ranks. *Nat. Rev. Microbiol.* **8,** 523–529 (2010).
6. Quince, C., Walker, A. W., Simpson, J. T., Loman, N. J. & Segata, N. Shotgun metagenomics, from sampling to analysis. *Nat. Biotechnol.* **35,** 833–844 (2017).
7. Martiny, J. B. H., Jones, S. E., Lennon, J. T. & Martiny, A. C. Microbiomes in light of traits: A phylogenetic perspective. *Science (80-.).* **350,** aac9323 (2015).
8. Martiny, A. C., Treseder, K. & Pusch, G. Phylogenetic conservatism of functional traits in microorganisms. *ISME J.* **7,** 830–838 (2013).
9. O'Leary, N. A. *et al.* Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* **44,** D733–D745 (2016).
10. Parks, D. H. *et al.* Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nat. Microbiol.* **2,** 1533–1542 (2017).
11. Weiss, M. C. *et al.* The physiology and habitat of the last universal common ancestor. *Nat. Microbiol.* **1,** 1–8 (2016).
12. David, L. A. & Alm, E. J. Rapid evolutionary innovation during an Archaean genetic expansion. *Nature* **469,** 93–96 (2011).
13. Segata, N. & Huttenhower, C. Toward an efficient method of identifying core genes for evolutionary and functional microbial phylogenies. *PLoS One* **6,** e24704 (2011).
14. Morris, J. J., Lenski, R. E. & Zinserc, E. R. The Black Queen Hypothesis: Evolution of Dependencies through Adaptive Gene Loss. *MBio* **3,** 1–7 (2012).
15. Koeppel, A. F. & Wu, M. Lineage-dependent ecological coherence in bacteria. *FEMS Microbiol. Ecol.* **81,** 574–582 (2012).
16. Min, C. K. *et al.* Genome-based construction of the metabolic pathways of Orientia tsutsugamushi and comparative analysis within the Rickettsiales order. *Comp. Funct. Genomics* **2008,** (2008).
17. Wolgemuth, C. W. Flagellar motility of the pathogenic spirochetes. *Semin Cell Dev Biol.* **46,** 104–112 (2015).
18. Dreyer, M., Aeby, S., Oevermann, A. & Greub, G. Prevalence and diversity of Chlamydiales in Swiss ruminant farms. *Pathog. Dis.* **73,** 1–4 (2015).
19. Wexler, H. M. Bacteroides: The good, the bad, and the nitty-gritty. *Clin. Microbiol. Rev.* **20,** 593–621 (2007).
20. Moran, N. A. Microbial Minimalism: Genome Reduction in Bacterial Pathogens. *Cell* **108,** 583–586 (2002).
21. Hyatt, D. *et al.* Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11,** 119 (2010).
22. Marchler-Bauer, A. & Bryant, S. H. CD-Search: Protein domain annotations on the fly. *Nucleic Acids Res.* **32,** W327–W331 (2004).
23. Mavromatis, K. *et al.* The DOE-JGI Standard Operating Procedure for the Annotations of Microbial Genomes. *SIGS* **1,** 63–67 (2009).
24. Leimbach, A. *et al.* Whole-Genome Draft Sequences of Six Commensal Fecal and Six Mastitis-

Associated *Escherichia coli* Strains of Bovine Origin: TABLE 1. *Genome Announc.* **4,** e00753-16 (2016).

25. Wood, S. mgcv: Mixed GAM Computation Vehicle with GCV/AIC/REML Smoothness Estimation. (2017).
26. Philip, D. Computer program review VEGAN , a package of R functions for community ecology. *J. Veg. Sci.* **14,** 927–930 (2003).
27. Letunic, I. & Bork, P. Interactive Tree of Life v2: Online annotation and display of phylogenetic trees made easy. *Nucleic Acids Res.* **39,** W475–W478 (2011).

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 Species	Cultured Lineages	Uncultured Lineages
Bacteria	Actinobacteria	1	5	11	19	889	880	9
	Bacteroidetes	1	5	9	13	338	306	32
	Deinococcota	1	1	2	2	35	35	0
	Epsilonbacterota	1	1	2	2	26	25	1
	Firmicutes	2	5	13	25	492	486	6
	Firmicutes A	2	4	4	4	58	56	2
	Firmicutes B	1	1	1	1	11	9	2
	Fusobacteria	1	1	1	1	11	11	0
	Proteobacteria	2	12	25	55	1214	1147	67
	Spirochaetes	2	3	3	3	39	31	8
	Verrucomicrobia	1	1	1	1	10	10	0
Archaea	Euryarchaeota	1	1	1	1	20	20	0
	Halobacterota	2	2	3	5	74	73	1
Total	13	18	42	76	132	3207	3089	128

Figures

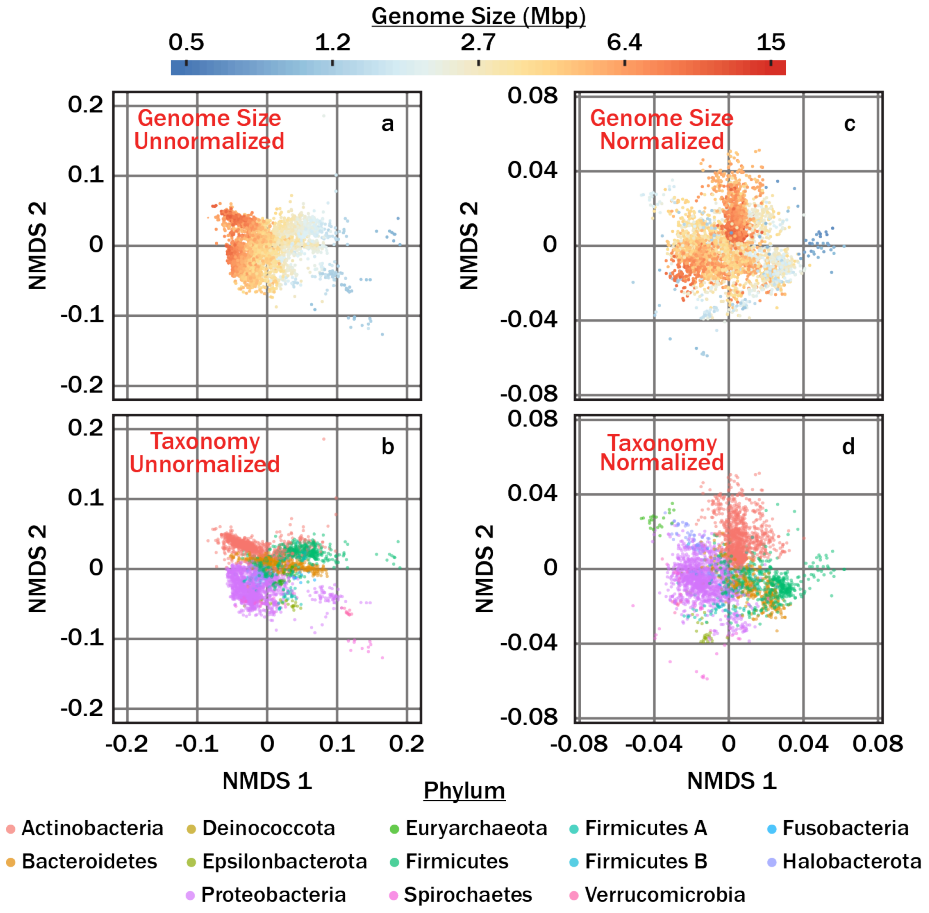


Figure 1. NMDS plots of COG-FC abundance (a,b) and enrichment (c,d) of the lineages summarized in Table 1. Individual data points are colored by genome size (a,c) and phylum (b, d). Panels a,b were not normalized by genome size while panels c,d were normalized by genome size. Note the difference in scaling for the axes.

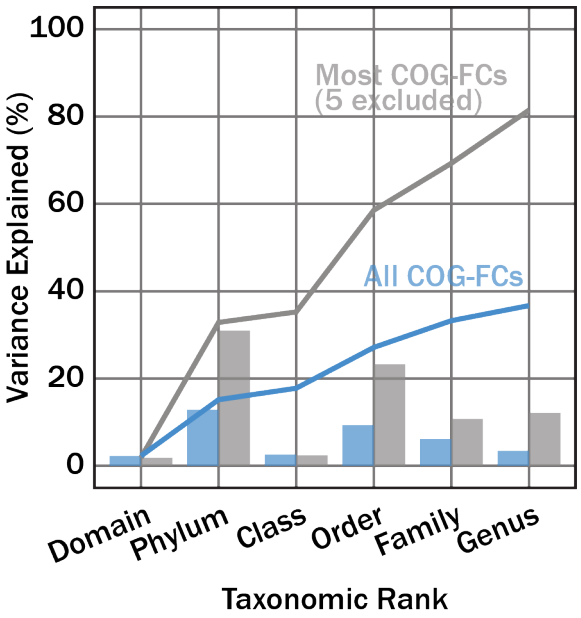


Figure 2. The average variance in COG-FC enrichment explained by different taxonomic ranks (bars) and the cumulative variance explained by taxonomic ranks (lines). Blue corresponds to PERMANOVA performed on all COG-FCs while grey corresponds to PERMANOVA performed on all COG-FCs excluding: cytoskeleton, RNA processing and modification, chromatin structure and dynamics, cell motility, and intracellular trafficking, secretion, and vesicular transport (see text for details). All variance explained by taxonomic ranks was significant ($p<0.001$). The F-value (blue) for domain, phylum, class, order, family, and genus, was 75.1, 47.7, 40.7, 28.8, 20.9, and 13.7, respectively. The F-value (grey) for domain, phylum, class, order, family, and genus, was 61.7, 130.6, 102.44, 109.3, 94.5, and 103.4, respectively.

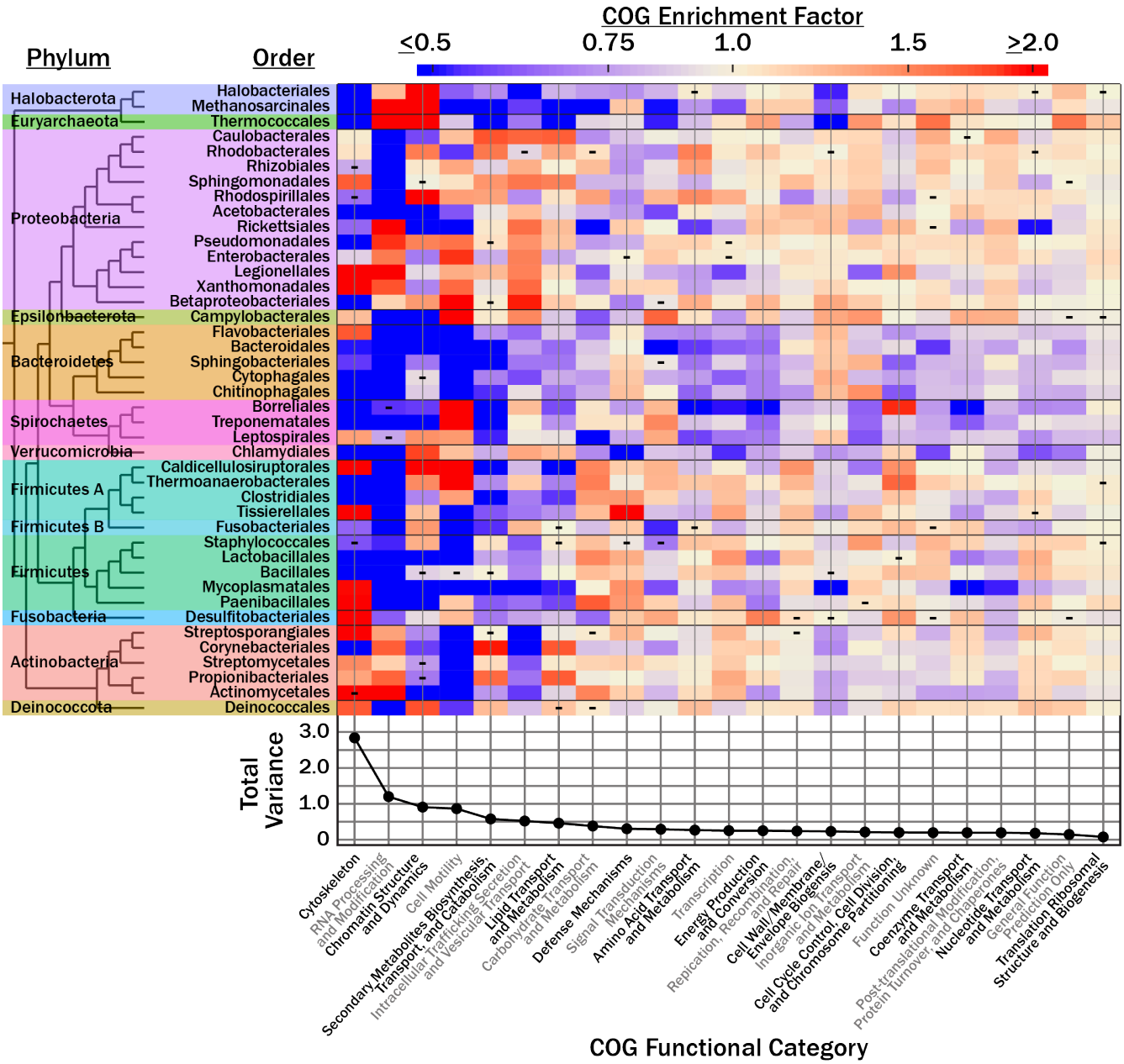
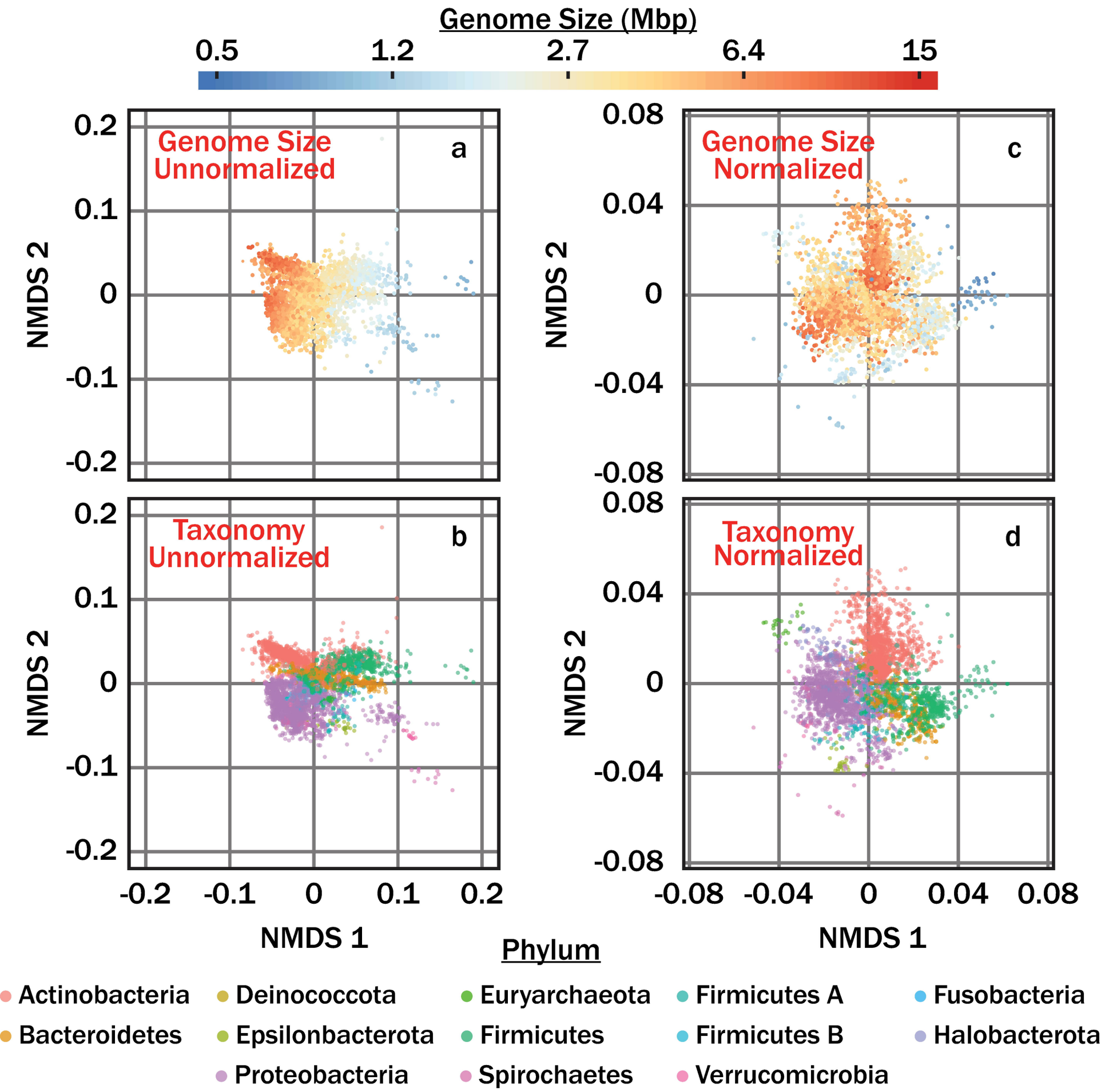
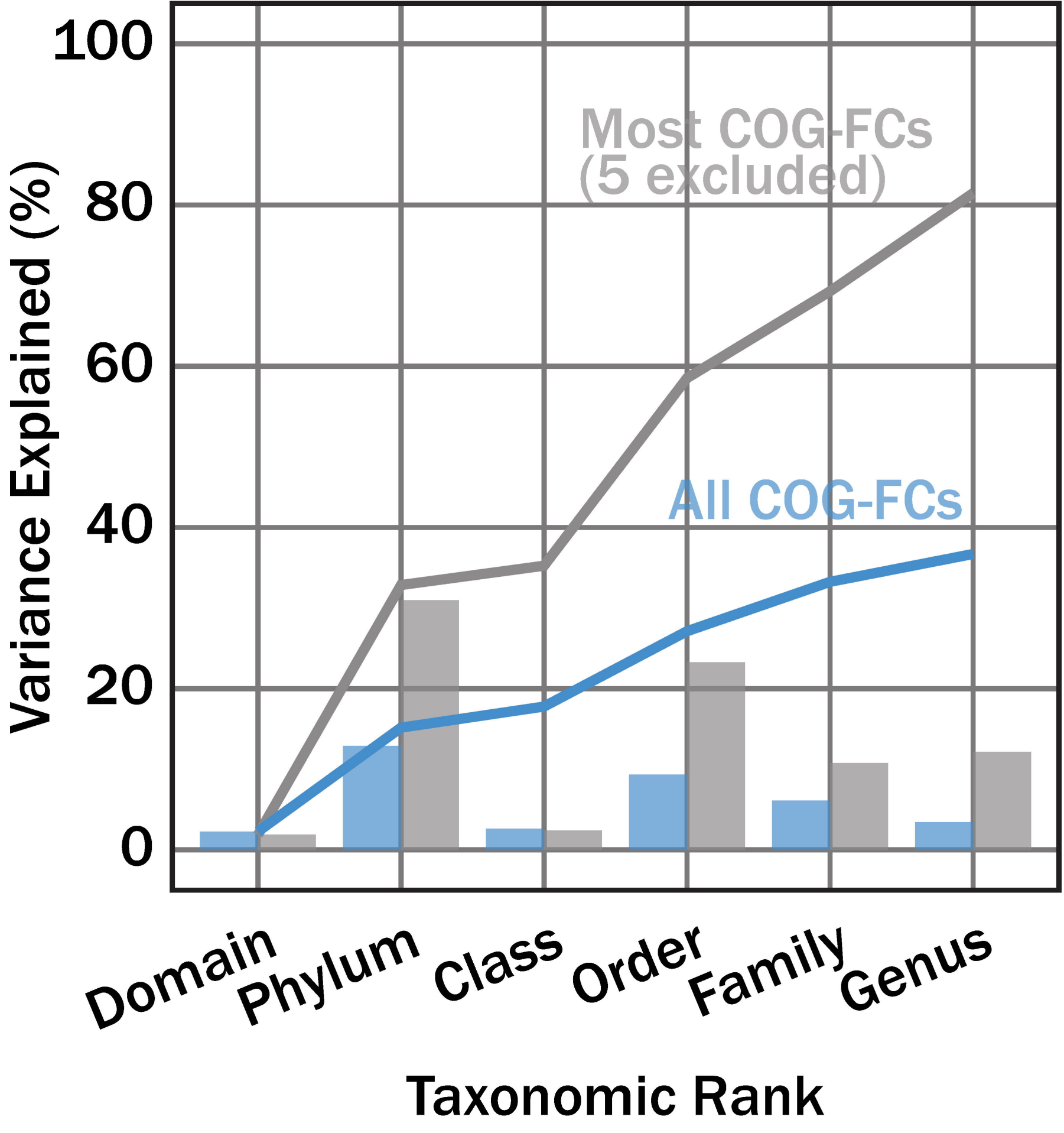
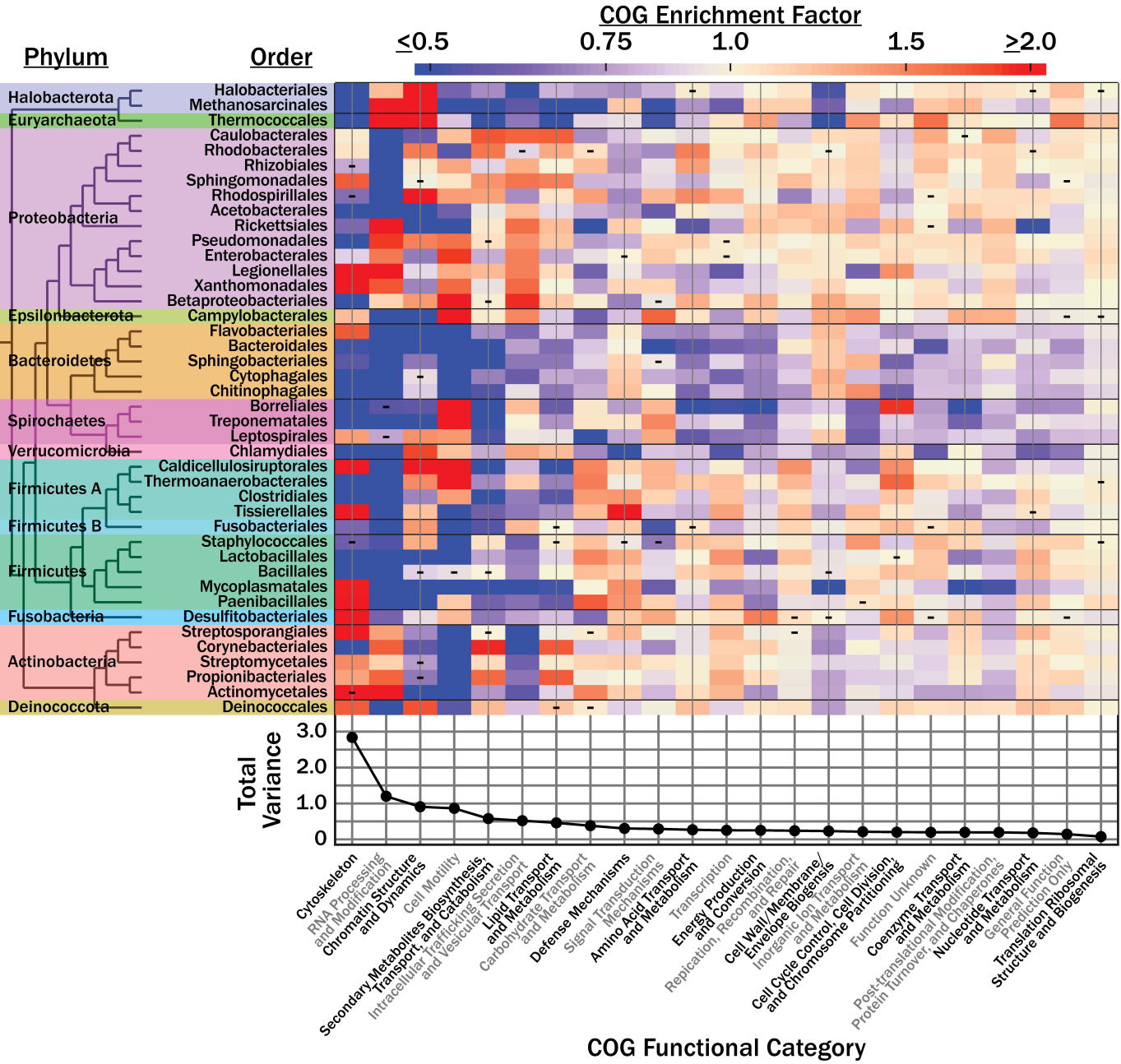


Figure 3. A heat map showing the average COG-FC enrichment for all orders in Table 1. Categories were arranged along the x axis based on the total variance in enrichment across all lineages. Clades were organized along the y axis using phylogenetic relatedness based on the reported concatenated protein sequence alignments in Parks et al.[1]. Minus signs correspond to enrichments not consider significantly different than 1 (see methods).


Genome Size (Mbp)
0.5
1.2
2.7
6.4
15
Genome Size Unnormalized
a
Genome Size Normalized
c
Taxonomy Unnormalized
b
Taxonomy Normalized
d
NMDS 2
NMDS 1
0.2
0.1
0
-0.1
-0.2
0.08
0.04
-0.04
-0.08
Phylum
Actinobacteria
Deinococcota
Euryarchaeota
Firmicutes A
Fusobacteria
Bacteroidetes
Epsilonbacterota
Firmicutes
Firmicutes B
Halobacterota
Proteobacteria
Spirochaetes
Verrucomicrobia


Variance Explained (%)
100
80
60
40
20
0
Most COG-FCs
(5 excluded)
All COG-FCs
Domain
Phylum
Class
Order
Family
Genus
Taxonomic Rank


COG Enrichment Factor
≤0.5
0.75
1.0
1.5
≥2.0
Phylum
Order
Halobacterota
Euryarchaeota
Proteobacteria
Epsilonbacterota
Bacteroidetes
Spirochaetes
Verrucomicrobia
Firmicutes A
Firmicutes B
Firmicutes
Fusobacteria
Actinobacteria
Deinococcota
Halobacteriales
Methanosarcinales
Thermococcales
Caulobacterales
Rhodobacterales
Rhizobiales
Sphingomonadales
Rhodospirillales
Acetobacterales
Rickettsiales
Pseudomonadales
Enterobacterales
Legionellales
Xanthomonadales
Betaproteobacteriales
Campylobacterales
Flavobacteriales
Bacteroidales
Sphingobacteriales
Cytophagales
Chitinophagales
Borreliales
Treponematales
Leptospirales
Chlamydiales
Caldicellulosiruptorales
Thermoanaerobacterales
Clostridiales
Tissierellales
Fusobacteriales
Staphylococcales
Lactobacillales
Bacillales
Mycoplasmatales
Paenibacillales
Desulfitobacteriales
Streptosporangiales
Corynebacteriales
Streptomycetales
Propionibacteriales
Actinomycetales
Deinococcales
Total Variance
3.0
2.0
1.0
0
Cytoskeleton
RNA Processing and Modification
Chromatin Structure and Dynamics
Cell Motility
Secondary Metabolites Biosynthesis, Transport, and Catabolism
Intracellular Trafficking, Secretion and Vesicular Transport
Lipid Transport and Metabolism
Carbohydrate Transport and Metabolism
Defense Mechanisms
Signal Transduction Mechanisms
Amino Acid Transport and Metabolism
Transcription
Energy Production and Conversion
Replication, Recombination and Repair
Cell Wall/Membrane/ Envelope Biogenesis
Inorganic Ion Transport and Metabolism
Cell Cycle Control, Cell Division, and Chromosome Partitioning
Function Unknown
Coenzyme Transport and Metabolism
Post-translational Modification, Protein Turnover, and Chaperones
Nucleotide Transport and Metabolism
General Function Prediction Only
Translation, Ribosomal Structure and Biogenesis
COG Functional Category