Widespread synchronization of codon usage in functionally related genes

José Luis López[1,*], Aristeidis Litos[1], Malte Siemers[1], Meine D. Boer[2,3], Erica Padial[2], Swapnil Prakash Doijad[1], Bas E. Dutilh[1,2,*]

[1]Institute of Biodiversity, Ecology, and Evolution; Faculty of Biological Sciences; Cluster of Excellence Balance of the Microverse; Friedrich Schiller University Jena; Jena, Germany
[2]Theoretical Biology and Bioinformatics; Department of Biology; Science for Life; Utrecht University; Utrecht, The Netherlands
[3]Dept. Marine Microbiology and Biogeochemistry; NIOZ Royal Netherlands Institute for Sea Research; Texel, The Netherlands
*Corresponding authors. Email: joselopezarcondo@gmail.com, b.e.dutilh@uni-jena.de

Abstract:

The usage of synonymous codons varies along the genome, with strong biases in conserved and highly expressed genes that are optimized for translation. The extent of codon usage adaptation across genes and co-adaptation between genes, as well as the influence of gene function on these patterns, remain important open questions. Here, we show that codon usage is highly non-random in most bacterial genes, with at least ~20-46% of the gene families presenting synchronized codon usage evolution. We show that co-adapting genes are co-expressed, co-regulated, metabolically connected, and functionally associated. Codon usage adaptations of key marker genes highlight differences in their expression context between microbes with alternative ecological strategies. This underappreciated regulatory dimension has important implications for function discovery and engineering.

Main text:

Codons that translate to the same amino acid (synonymous codons) are used at different frequencies depending on the expression level, conservation, and genomic location of the encoded genes. In fast-growing bacteria, selection for efficient translation has led to a strong codon usage bias (CUB) in ribosomal proteins. Codon usage regulates expression by affecting the efficiency and accuracy of protein elongation *(1–3)*, and enables fine-tuning of expression across different physiological conditions *(4–6)*. CUB is also correlated with subunit stoichiometry in polycistronic genes encoding protein complexes *(7)*. While CUB has been shown to co-evolve among genes that are functionally or metabolically related and in physically interacting proteins *(8, 9)*, these studies were limited in scope, focusing on relatively few functions and organisms. The breadth of CUB co-adaptation across functional categories and

throughout the bacterial domain remains largely unexplored. Here, we hypothesized that codon usage fine-tuning is a widespread mechanism that affects all genes as they evolve, reflecting their adaptation to the expression requirements in different bacterial lineages. To test this, we used comparative genomics of thousands of bacterial genomes to investigate the link between protein function and codon usage adaptation. By analysing co-adaptation patterns between pairs of genes, we show that codon usage is strongly shaped by gene function and that its co-adaptation extends across gene families and phyla, revealing a finely tuned code that regulates gene expression and can be leveraged for protein interaction discovery and understanding bacterial physiology.

Codon usage bias is broadly conserved within gene families

Previous research revealed a spectrum of CUB values ranging from highly expressed, conserved genes to lowly expressed, accessory ones *(10–12)*. As CUB can be calculated in different ways, we first benchmarked sixteen commonly used CUB indices and found that gene codon bias (GCB) best correlated with expression data under a range of conditions, where highly expressed genes with optimized codon usage have high GCB values *(13, 14)* (figs. S1 and S2, supplementary text S1). Next, we analyzed genome sequences from 2,978 representative species in the five most abundant bacterial phyla, including *Actinomycetota*, *Bacillota*, *Bacillota A*, *Bacteroidota*, and *Pseudomonadota* (table S1). We used GCB to calculate CUB in all encoded gene families *(15)*, and assessed the CUB per functional category (Fig. 1, A and B). As expected, genes involved in protein processing and energy metabolism are highly optimized, categories that include critical functions for fast-growing bacteria *(16, 17)*. In contrast, unannotated genes were the least optimized (Fig. 1A), illustrating annotation bias: widespread, highly expressed genes are more likely to be studied and thus functionally annotated *(18)*. To test if orthologs have similar CUB values in different bacteria, we compared single-copy gene families (PGFams) shared between genomes and observed consistent overall GCB patterns (Fig. 1C). Correlation values were lower for the subset of unannotated proteins than for annotated ones (Fig. 1D), in line with the higher expression, conservation, and annotation rates of core genes, and the evolutionary variability of accessory, often hypothetical, genes *(12, 19)*.

We then explored CUB patterns in paralogous genes within the same genome *(15)* (supplementary text S2) and found that co-adaptation was stronger in colocalized paralogs within the same genomic region, as well as in remote paralogs with similar neighboring genes, compared to remote paralogs without shared neighbors (Fig. 1E, table S2, and fig. S3). Together, these results confirm that protein function drives codon usage, show that CUB is conserved between orthologs, and reveal CUB adaptation on a genome-wide scale.

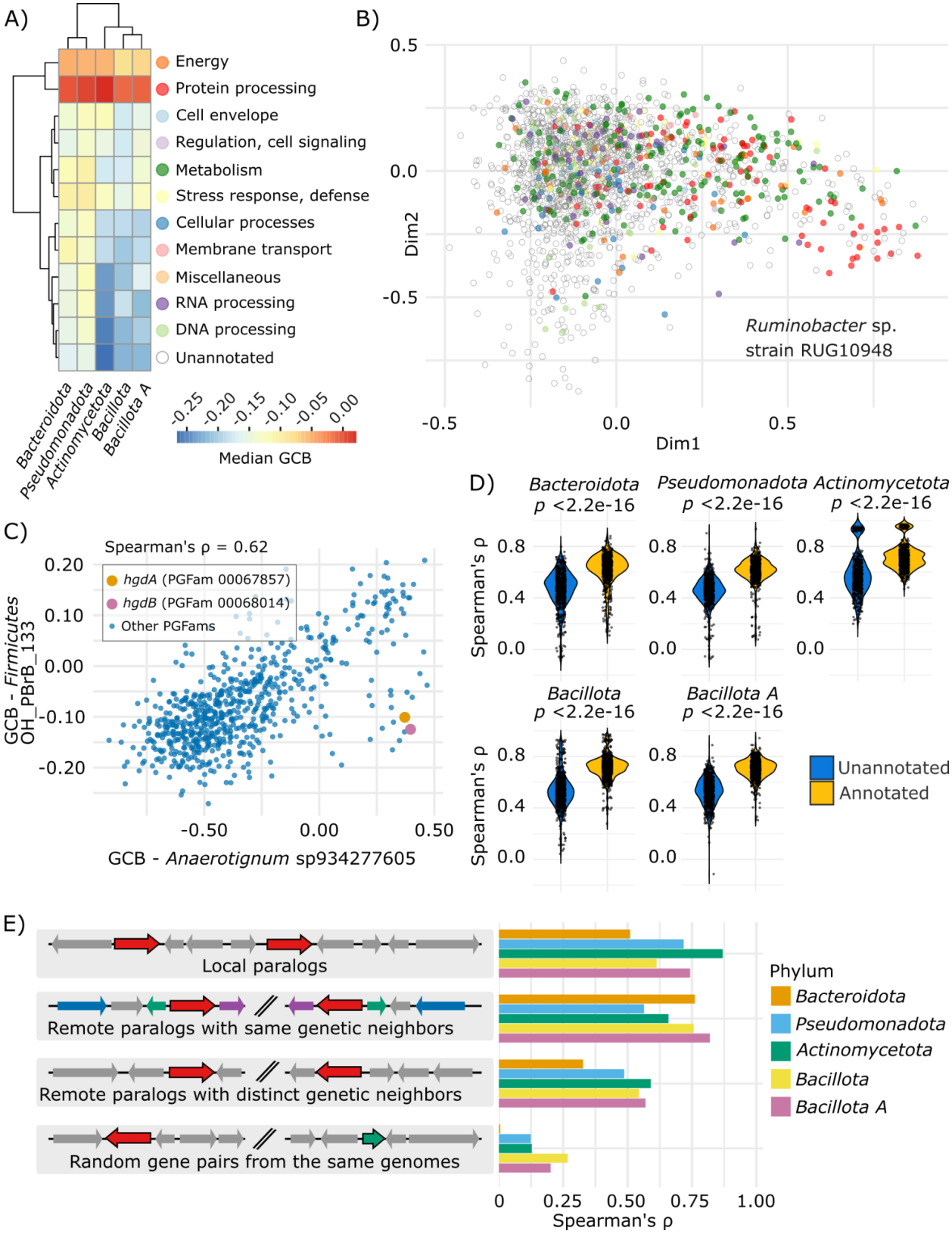

A)
Energy
Protein processing
Cell envelope
Regulation, cell signaling
Metabolism
Stress response, defense
Cellular processes
Membrane transport
Miscellaneous
RNA processing
DNA processing
Unannotated
Bacteroidota
Pseudomonadota
Actinomycetota
Bacillota
Bacillota A
-0.25 -0.20 -0.15 -0.10 -0.05 0.00
Median GCB
B)
Dim2
Dim1
Ruminobacter sp. strain RUG10948
C)
Spearman's ρ = 0.62
hgdA (PGFam 00067857)
hgdB (PGFam 00068014)
Other PGFams
GCB - Firmicutes OH_PBrB_133
GCB - Anaerotignum sp934277605
D)
Bacteroidota p <2.2e-16
Pseudomonadota p <2.2e-16
Actinomycetota p <2.2e-16
Bacillota p <2.2e-16
Bacillota A p <2.2e-16
Spearman's ρ
Unannotated
Annotated
E)
Local paralogs
Remote paralogs with same genetic neighbors
Remote paralogs with distinct genetic neighbors
Random gene pairs from the same genomes
Phylum
Bacteroidota
Pseudomonadota
Actinomycetota
Bacillota
Bacillota A
Spearman's ρ

Fig. 1. Overall codon usage optimization across bacterial functions.
(**A**) Median GCB values of genes by functional category *(28)*. The heatmap represents the median of the per-genome median GCB values for each phylum. (**B**) Example case showing the first two dimensions of a factorial correspondence analysis of per-gene codon usage vectors in *Ruminobacter* strain RUG10948 (*Pseudomonadota*). Genes are colored by functional category as in panel A. (**C**) CUB values of shared orthologous genes between *Anaerotignum sp934277605* and *Firmicutes* strain OH_PBrB_133 are positively correlated (Spearman's $\rho = 0.62$). The *hdgA* and *hdgB* genes are highlighted, these are further analyzed in Fig. 2. (**D**) Spearman correlations between GCB values of orthologous genes (single-copy PGFams) between genome pairs. Violin plots are split by annotation status of the genes. Statistics were performed with Wilcoxon rank-sum tests. (**E**) Spearman correlations between the CUB values of paralogs (PGFams with two copies per genome, highlighted red) across phyla *(15)* (table S2 and supplementary text). Genetic neighbors are defined as PGFams within ±20 genes up- or downstream of a target gene. Local paralogs have <40 genes between them, others are remote. Random gene pairs show weak genome-wide CUB correlations.

Enzymatic subunits co-adapt their codon usage

Next, we used our genome collection to investigate potential codon usage (CU) co-adaptation patterns between functionally related genes. As an example, we focus on *hgdA* and *hgdB*, which encode two subunits of 4-hydroxybenzoyl-CoA reductase, an oxygen-sensitive enzyme involved in the anaerobic degradation of aromatic compounds *(20)*. As shown above (Fig. 1C), PGFams generally had similar CUB values in different genomes, e.g. *Firmicutes* strain OH_PBrB_133 and *Anaerotignum* sp934277605, but *hdgA* and *hdgB* formed an exception, showing increased CUB in *Anaerotignum* sp934277605 (Fig. 1C, and Fig. 2, A and B). Across species, we observed strong CU co-adaptation between these genes, with *Firmicutes* OH_PBrB_133 and *Anaerotignum* sp934277605 at opposite ends of the optimization spectrum (Spearman's $\rho = 0.958$, $p < 2.2e\text{-}16$, Fig. 2C), indicative of divergent selection pressures. In *Anaerotignaceae*, which inhabit strictly anaerobic environments, aromatic compound degradation is likely essential for energy metabolism *(21–24)*, explaining the CU optimization in *hgdA*/*hgdB*. In contrast, *Lachnospiraceae*, common gut bacteria that rely primarily on polysaccharide fermentation *(24, 25)*, may need reduced expression and weaker codon optimization of these genes.

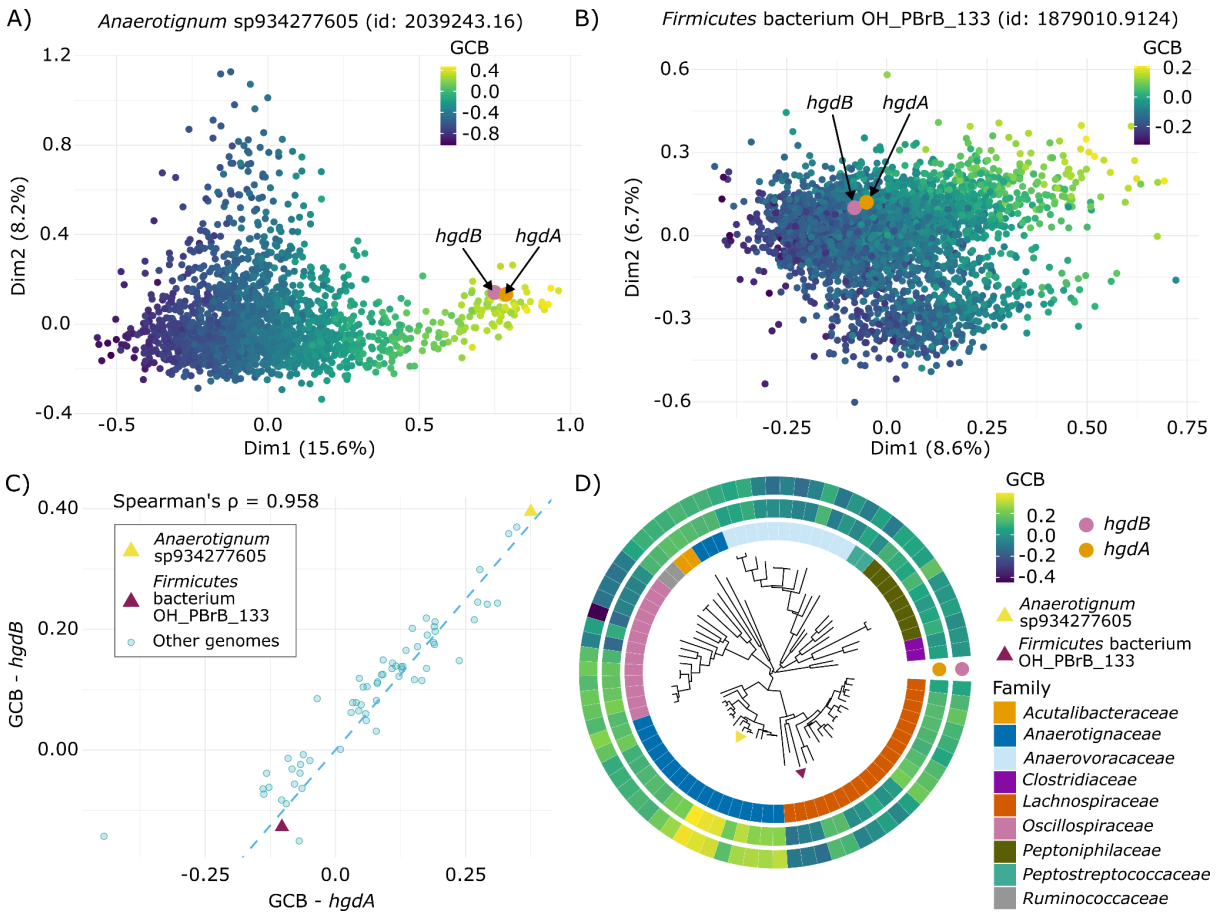


Fig. 2. CUB patterns in subunits of 4-hydroxybenzoyl-CoA reductase.
(**A and B**) First two dimensions of a factorial correspondence analysis of per-gene codon usage vectors for two *Bacillota A* genomes: (A) *Anaerotignum* sp934277605 and (B) *Firmicutes* strain OH_PBrB_133. Genes are colored according to their GCB values, with *hdgA* and *hdgB* genes being highly optimized in *Anaerotignum* sp934277605. (**C**) GCB values of *hdgA* and *hdgB* genes across reference species in *Bacillota A* are strongly correlated (Spearman's ρ = 0.958), indicating co-optimization of the enzyme subunits across genomes. The genomes shown in panels A and B are indicated. (**D**) The phylogenetic tree of genomes from *Bacillota A* coding for *hdgA* and *hdgB* was subsetted from the bac120_r220.tree, obtained from GTDB *(48)*. Color strips indicate taxonomic families and GCB values, example genomes are highlighted.

CU co-adaptation across functionally related proteins

To assess the scope of CU co-adaptation between protein-coding genes on a large scale, we correlated GCB and GCB z-scores between all PGFam pairs across genomes *(15)* (supplementary text S1). We hypothesized that additional factors beyond PGFam identity, such as phylogenetic signal, genomic mean GCB, and genomic GC content, could influence our co-adaptation values, given the strong associations observed between individual PGFam CUB and these factors (fig. S4). To assess these potential confounders, we took all 15,003 high-quality genomes from the five major phyla (table S1) *(15)* and constructed phylogenetically informed regression models for each PGFam pair, incorporating the confounders as covariates

and considering potential interactions *(15)*. Based on these codon usage co-adaptation models (table S3), we generated graphs where edges represent pairwise relationships with *t* values between PGFams. We next analysed the functional connections by defining co-adaptation clusters. First, we applied iterative Louvain clustering *(26, 15)*, and then assessed the STRING association score *(27)* of PGFam pairs within these clusters as a proxy for known functional interactions (table S3).

We benchmarked the different approaches used here to optimize the mean STRING combined interaction score between gene pairs within the same co-adaptation cluster. First, the STRING score increased significantly when phylogenetic signal, genomic mean GCB, and GC content were included as covariates (fig. S5, left panels). Furthermore, using normalized GCB z-scores instead of raw GCB values further improved functional linkages (higher mean STRING scores, Fig. 3A, and fig. S5, right panels). Finally, using the *t* values (co-adaptation estimates normalized by their standard errors) *(15)* increased the number of detected clusters while maintaining comparable STRING scores. Accordingly, we selected the phylogenetic linear models that used GCB z-score as input and incorporated the covariates and interactions; the resulting *t* values were used for all subsequent analyses.

These results indicate that genomic factors, major confounders in assessing CU co-adaptation among genes, can be effectively accounted for by using phylogenetic multiple regression with interaction effects. Using GCB z-scores as input already corrected for biases including genomic GCB, GC content, and potential other unknown confounders. Subsequent correction for the strong phylogenetic signal (Fig. 2D, and fig. S6) further improved model performance (fig. S5).

Subgraphs consisting of the top 200 strongest co-adapted PGFam pairs (Bonferroni-adjusted $p < 0.01$, table S2) *(15)* formed small clusters of 2-5 highly co-adapted PGFams, many of which included protein families with strong functional links based on SEED annotations *(28)* (Fig. 3B, table S2). These links included subunits of the same enzyme, components of enzyme complexes or cellular machinery (red clusters in Fig. 3B), or enzymes from the same pathway or metabolic process (pink clusters). We also identified several novel interactions (blue clusters), for which low or no STRING scores were observed (table S3).

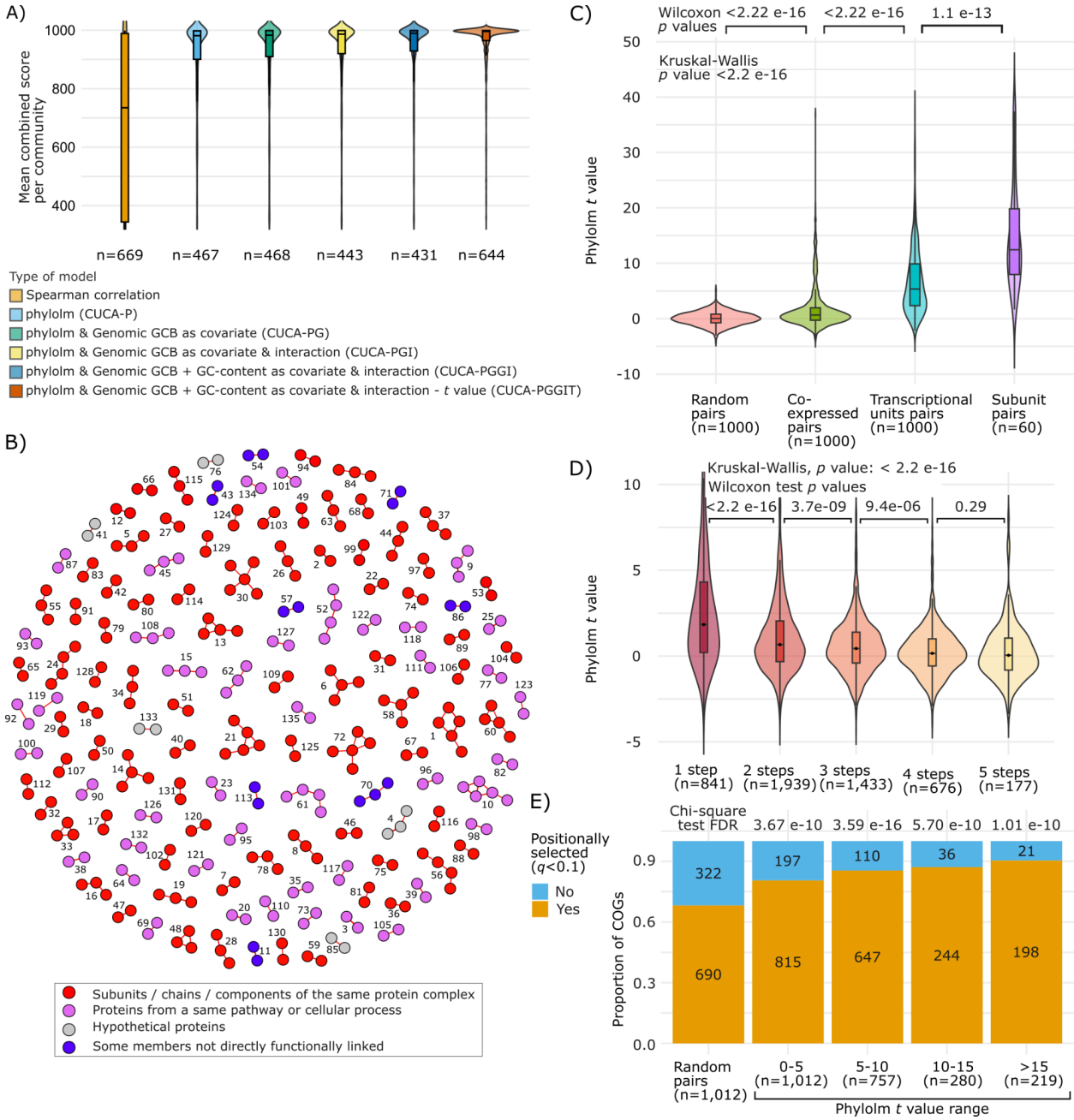


Fig. 3. CU co-adaptation between PGFam pairs.

(**A**) STRING combined scores for clusters obtained using Spearman correlation and different codon usage co-adaptation (CUCA) models that account for phylogenetic signal, genomic GCB, and GC content as covariates with interaction terms *(15)*; n represents the number of clusters with STRING interactions (>2 members). (**B**) The top 200 positive co-adapting genes from the optimal *Pseudomonadota* model are visualized as a network. Colors represent different types of functional relatedness, and numbers denote the cluster identifier (table S2). (**C**) Phylolm *t* value distributions increased from random gene pairs, to co-expressed pairs, to co-transcribed genes, to genes encoding subunits of the same protein complex. (**D**) Phylolm *t* values of enzymes that are closer in metabolic pathways are higher than more distant ones (1-5 steps apart in the metabolic network). (**E**) Positionally selected functions are more prevalent among highly co-adapted pairs *(15)*. The proportion of positionally selected PGFams in pairs with Phylolm *t* values within ranges: 0-5, 5-10, 10-15, and >15, are compared with random pairs, using chi-squared tests.

Many clusters contained genes with closely related functions, suggesting that CU co-adaptation reflects co-expression of functionally linked genes. To test this hypothesis, we compared co-adaptation values between random gene pairs with those between enzyme subunits, genes within the same transcriptional unit, and co-expressed genes (Fig. 3C) *(15)*. Subunits of enzymes or protein complexes showed significantly higher co-adaptation values than genes within the same transcriptional unit (Wilcoxon test *p* value = 1.1e-13, Fig. 3C). Genes from the same transcriptional unit, in turn,showed significantly higher co-adaptation than co-expressed pairs, while co-expressed pairs had significantly higher values than random pairs (Wilcoxon test, *p* value < 2.22e-16, Fig. 3C). These results support a model where CU co-adaptation reflects functional association. The weakest association is between co-expressed PGFams, which may span multiple transcriptional units *(29)*. Next, genes within a single transcriptional unit are more tightly co-regulated, co-expressed, and functionally associated. Finally, this effect is strongest for subunits or components of protein complexes, whose genes require highly coordinated expression to maintain balanced cellular dosage *(7–9, 30)*.

To investigate the co-adaptation between metabolically associated genes, we quantified the distance between enzymes in metabolic networks *(15)*. Enzymes separated by a single metabolic step showed significantly higher co-adaptation (Wilcoxon test, *p* value < 2.22e-16), followed closely by those separated by two and three steps (3.7e-09 and 9.4e-06, respectively, Fig. 3D), although co-adaptation rapidly declined with increasing metabolic distance. The effect was stronger for enzymes within the same subsystem subclass and was consistently observed across all five phyla (fig. S7).

A recent study reported that two-thirds of functions are positionally selected on the bacterial genome, revealing an evolutionary mechanism that links gene expression to cellular growth rate *(31)*. We hypothesized that co-adapting pairs may also be positionally selected to further optimize co-expression. To test this, we compared the proportions of positionally selected genes among co-adapting pairs across different ranges of co-adaptation values with those of random gene pairs. Indeed, co-adapting pairs showed a significant signature of positional selection (Chi-square test, *p* value < 1.01e-10, Fig. 3E), supporting CU adaptation and positional selection as complementary evolutionary layers that fine-tune gene expression.

CU co-adaptation in half of the protein families across bacterial phyla

Above, we focused on the five most species-rich phyla because they contain the most genomes in the database, and thus likely contain the strongest CUB signal. To generalize these observations, we extended our models to include ten phyla with >100 available representative species genome sequences, including *Acidobacteriota*, *Bacillota C*, *Bdellovibrionota*,

Campylobacterota, *Chloroflexota*, *Cyanobacteriota*, *Desulfobacterota*, *Planctomycetota*, *Spirochaetota*, and *Verrucomicrobiota*. The co-adapting gene families identified within these additional phyla had clear functional connections, although fewer significant pairs were detected (fig. S8, and table S4). Thus, CU co-adaptation is widespread in the bacterial domain. Using stringent multiple-testing correction (Bonferroni $p < 0.01$), approximately 77% of gene pairs with significant CU co-adaptation also had high STRING combined scores (>700, table S4). For the remainder of cases, i.e. 23% of the gene families (n=1,051), this suggests that CU co-adaptation provides new evidence about their function.

With the stringent Bonferroni criterion used above, only 20% of all PGFams had at least one significant co-adapting partner (table S4). To reduce the stringency of multiple-testing correction and reveal more functional associations, we applied a Benjamini-Hochberg (BH) procedure. To identify a good cutoff, we selected co-adapting pairs at a range of different BH cutoff values, and determined when the median STRING combined score of the significant pairs exceeded 700 (fig. S9, and table S4). Using this approach, we found that approximately 46% of all PGFams (ranging from 14% to 60% for individual phyla) had at least one significantly co-adapting partner, and about 30% of all PGFams (ranging from 8% to 43% across phyla) had a significantly co-adapting partner with a high STRING interaction score (combined score >700, table S4). These results indicate that CU is under significant adaptive selection in a substantial fraction of bacterial functions. The remaining PGFams without a detected co-adapting partner may correspond to genes with less variation in CUB values across genomes, suggesting that alternative modelling strategies may be necessary to capture their co-adaptation signal. Alternatively, these genes may not co-adapt with others, e.g. because they represent independently functioning enzymes.

Co-adapting gene pairs have similar CUB values, codons, and amino acids

Above, we tested whether gene pairs had correlating CUB patterns across genomes. To test whether they also have similar CUB values, we calculated the median absolute difference in GCB z-scores (|ΔGCBz|) within their genomes, and compared these values between non-significantly and significantly co-adapting pairs (fig. S10A). The latter gene pairs typically fell within the lowest 5-10% of median |ΔGCBz| values for pairs of a given PGFam and had significantly lower median |ΔGCBz| values, indicating that co-adapting genes tend to have similar codon usage bias (fig. S10, B to D). These results reveal that co-adapting gene pairs closely track one another as their CU evolves across genomes, and suggest that median |ΔGCBz| could serve as an additional criterion to enhance detection of functionally linked pairs.

To determine whether similar CUB values reflect similar codon composition, we calculated codon composition distances between gene pairs corresponding to all PGFams included in our models *(15)*. We found that co-adapting pairs had significantly smaller codon composition

distances than non-significant pairs (fig. S11A), consistent with their more similar CUB values (fig. S10). Interestingly, co-adapting pairs also had significantly similar amino acid composition (fig. S11B). This not only contributes to their shared codon composition, but might also constitute an additional layer of selection. Indeed, the positive correlation between codon and amino acid distances was stronger in significant than in non-significant pairs (Spearman's ρ values of 0.78 and 0.62, respectively, fig. S11C). Mirroring the patterns observed for median |ΔGCBz| values, the most significant co-adapting pairs typically fell within the lowest 10–15% of codon composition distances and the lowest 15–20% of amino acid composition distances (fig. S11D). This suggests an evolutionary coupling between codon and amino acid co-adaptation where both contribute to protein expression optimization.

CU co-adaptation identifies candidate novel functional interactions

CU co-adaptation complements existing evidence of functional interactions between genes and can be leveraged to identify novel interacting partners, prioritising targets for experimental validation (tables S3 and S4). We will present four cases where proteins were not previously recognized to interact. First, we will present the computational evidence, and then summarize experimental results that provide additional support for their functional association.

First, the histone-like *Bacillus subtilis* protein HBsu, an essential DNA-binding protein linked to replication initiation in *B. subtilis (32, 33)*, strongly clusters with ribosomal proteins based on CU co-adaptation (table S3), with high PhyloIm *t* values and low median |ΔGCBz| values (Fig. 4, A and B, and fig. S12, A and B) in *Bacillota* (cluster 14, table S3) and *Bacillota A* (cluster 137, table S3) phyla. As STRING currently provides limited evidence for interactions between HBsu and ribosomal proteins, mainly through co-expression and genomic neighbourhood (only two ribosomal proteins have a combined score >700, table S3, and Fig. 4B), our findings point to a complementary role of HBsu and ribosomes in optimizing cell division. This is further supported by regulatory links between HBsu mRNA stability and ribosomal gene expression *(34)*, and evidence of a role in chromosome organization *(35)*. A similar co-adaptation with ribosomal proteins (cluster 550, table S3) could be detected for DNA-binding protein HUβ (i.e., Heat-unstable beta subunit) in *Pseudomonadota*, which belongs to the same class of histone-like protein families with similar roles binding DNA *(32, 36)*. We hypothesize that these proteins could have a role in coupling DNA replication and translation, where HBsu and HUβ optimization may be required to sustain fast growth.

Whereas the example above is closely associated with ribosomal proteins and has high CUB and expression in fast-growing bacteria, there are also interesting examples of the converse. We found that in *Bacteroidota*, CU of the energy-dependent translational throttle protein EttA, an ATPase from the ATP-binding cassette subfamily F (ABC-F) that regulates the translation elongation cycle *(37)*, significantly co-adapts with a series of proteins involved in different

translation-related functions (all *t* values > 4.5, FDR < 8 e-04, Fig. 4, C and D, fig. S12, C and D, and table S5). Interestingly, most of these proteins consume either ATP or GTP and are involved in ribosome functioning, yet have no strong known interactions with EttA (low STRING combined scores with weak evidence of co-expression in *Bacteroidota*, table S5). By binding to the ribosome, EttA restricts ribosome entry into the translation elongation cycle at a high ADP:ATP ratio *(37, 38)*. EttA restriction on translation is reflected by its negative *t* values with ribosomal proteins (Fig. 4C, and table S5), suggesting that slow-growing bacteria (i.e., those with lower CUB in their ribosomal proteins) evolved a higher CUB for this protein. Although the physiological function of EttA has not been fully uncovered, we hypothesize that it has a critical role in modulating ribosomal function, especially in slow-growing bacteria that thrive in low nutrient conditions and must tightly couple translation to energy availability. This hypothesis is consistent with the fact that co-expression evidence of EttA with ribosomal proteins is only found in slow-growing bacteria (table S5), and with the loss of EttA ortholog in a beetle symbiont, *Burkholderia* sp. Lv-StB, which may reflect an adaptation to a more steady nutritional niche *(39)*. Moreover, EttA protects against antibiotics in the slow-growing *Mycobacterium tuberculosis (40–42)* but not in fast-growing bacteria such as *Staphylococcus aureus (43)* or *Escherichia coli (37)*, where it has a role in entering the stationary phase *(37, 38)*. Thus, this suggests EttA may be more essential in slow than in fast growers and could hold potential for targeting slow-growing pathogens.

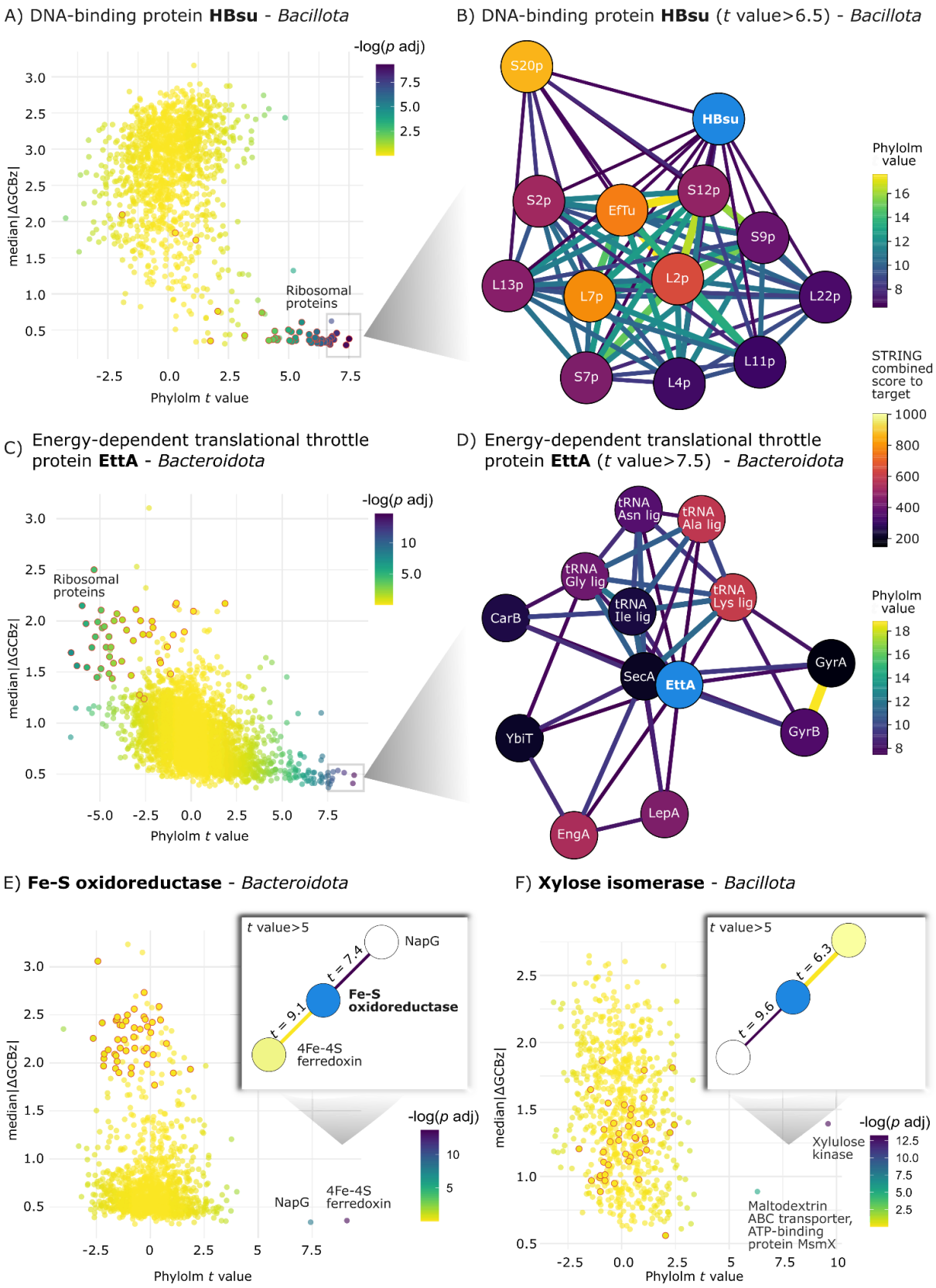

A) DNA-binding protein **HBsu** - *Bacillota*
-log(*p* adj)
7.5
5.0
2.5
median|ΔGCBz|
Ribosomal proteins
Phylolm *t* value
B) DNA-binding protein **HBsu** (*t* value>6.5) - *Bacillota*
S20p
HBsu
S2p
EfTu
S12p
S9p
L13p
L7p
L2p
L22p
S7p
L4p
L11p
Phylolm value
STRING combined score to target
C) Energy-dependent translational throttle protein **EttA** - *Bacteroidota*
D) Energy-dependent translational throttle protein **EttA** (*t* value>7.5) - *Bacteroidota*
tRNA Asn lig
tRNA Ala lig
tRNA Gly lig
tRNA Ile lig
tRNA Lys lig
CarB
SecA
EttA
GyrA
GyrB
YbiT
LepA
EngA
E) **Fe-S oxidoreductase** - *Bacteroidota*
t value>5
NapG
t = 7.4
Fe-S oxidoreductase
t = 9.1
4Fe-4S ferredoxin
F) **Xylose isomerase** - *Bacillota*
t = 6.3
t = 9.6
Xylulose kinase
Maltodextrin ABC transporter, ATP-binding protein MsmX

Fig. 4. Co-adaptation partners represent known and putative functional interactions.
(**A**, **C**, **E**, **F**) PhyloIm *t* values and median |ΔGCBz| values to the target PGFam are shown to represent the space of interactions with all other PGFam pairs. Color indicates the significance of CU co-adaptation according to the CUCA-PGGIT model. Ribosomal proteins are outlined in red. (**B** and **D**) Subsets of significantly co-adapting PGFams are represented by nodes with their color representing the highest known STRING combined score in interactions with the target PGFam (in light blue). Edge thickness is proportional to the PhyloIm *t* values. Target PGFam and their co-adaptation partners include (A and B) DNA-binding protein HBsu, (C and D) energy-dependent translational throttle protein EttA, (E) Fe-S oxidoreductase, and (F) xylose isomerase.

Other co-adapting pairs showed no prior evidence of known functional interaction (STRING interactions are absent across genomes). In these cases, CU co-adaptation suggests previously unrecognized interactions based on shared components or processes. For instance, cluster 242 in *Bacteroidota* (Fig. 4E, and table S5) includes Fe-S oxidoreductase and 4Fe-4S ferredoxin iron-sulfur binding domain proteins, which are known to interact (STRING combined score = 972) but co-adapt with a ferredoxin-type protein NapG (periplasmic nitrate reductase, Fig. 4E), with absent STRING interactions. In *Pseudomonadota*, NapG is involved in transferring electrons from quinol molecules to the nitrate reductase during anaerobiosis, together with other proteins encoded in the *nap* operon *(44)*. In *Bacteroidota*, only NapG ortholog is present and, although this electron shuttling function is known to interact with other oxidoreductases such as NADH-ubiquinone oxidoreductase (table S5), the co-adapting Fe-S oxidoreductase (PGF_00004164) might be an unknown electron shuttling partner, or alternatively they co-adapt because they both utilize 4Fe-4S clusters, typical of bacterial ferredoxins *(45)*.

Another co-adaptation pair brings together xylose isomerase (EC 5.3.1.5), xylulose kinase (EC 2.7.1.17), and the maltodextrin ABC transporter ATP-binding protein MsmX (Fig. 4F, cluster 86 in *Bacillota*). Although MsmX is connected with many other sugar transporters (table S5), its specific connection to xylose is not known at STRING so we speculate a role in transporting specifically this sugar.

We provide all the putative novel interactions (table S5) and a database of PGFam pairs displaying CU co-adaptation, along with STRING combined scores and median |ΔGCBz| values for the phyla analyzed here, at https://coadaptation.shinyapps.io/CUCA_PGGI/. Future improvements of this web app and of our models may include incorporating multi-copy genes, lowering the minimum number of representatives in the calculation of co-adaptation values to increase the coverage of accessory functions, using genome sets at different taxonomic ranks, and implementing alternative or multiple CUB indices and normalization methods, together with different modelling strategies that identify and control for additional confounding factors.

Conclusions

In this study, we have shown that codon usage co-adaptation in bacteria extends to many different protein families and reflects an overlooked layer of genomic adaptive regulation.

Mining this signal offers the possibility to obtain novel biological insights into bacterial physiology, life-history traits, and expression adaptation for genes in many functional categories. Codon usage patterns have recently been leveraged in diverse applications, including deep learning approaches to improve species classification, predict protein and transcript abundances, and melting point estimations *(46)*, and we hypothesize that these models may have captured some of the co-adaptation signals that we described above *(47)*. Detecting co-adaptation will enhance protein-protein interaction predictions and help assign new functional annotations to otherwise uncharacterized proteins. This opens up the possibility of applying our models to the discovery of novel biosynthetic gene clusters, anti-defense systems, and other uncharacterized genomic modules. We anticipate that these patterns will be valuable for metabolic modeling and metabolic engineering as well, because codon optimization might reflect proteome partitioning and potentially metabolic fluxes. Additionally, co-adaptation signatures may aid in predicting biome or environmental preferences and in broadly describing life-history strategies across bacteria, archaea, and eukaryotes. Overall, we reveal with unprecedented scope and scale that codon usage is under strong selective pressure, fine-tuning gene expression throughout the bacterial domain.

References and Notes

1. J. B. Plotkin, G. Kudla, Synonymous but not the same: The causes and consequences of codon bias. *Nature Reviews Genetics* **12**(1), 32–42 (2010).

2. T. E. F. Quax, N. J. Claassens, D. Söll, J. van der Oost, Codon bias as a means to fine-tune gene expression. *Molecular Cell* **59**(2), 149–161 (2015).

3. G. Hanson, J. Coller, Codon optimality, bias and usage in translation and mRNA decay. *Nature Reviews Molecular Cell Biology* **19**, 20–30 (2018).

4. H. S. Najafabadi, H. Goodarzi, R. Salavati, Universal function-specificity of codon usage. *Nucleic Acids Research* **37**(21), 7014–7023 (2009).

5. Y. Xu, P. Ma, P. Shah, A. Rokas, Y. Liu, C. H. Johnson, Non-optimal codon usage is a mechanism to achieve circadian clock conditionality. *Nature* **495**, 116–120 (2013).

6. M. Zhou, J. Guo, J. Cha, M. Chae, S. Chen, J. M. Barral, M. S. Sachs, Y. Liu, Non-optimal codon usage affects expression, structure and function of clock protein FRQ. *Nature* **495**, 111–115 (2013).

7. T. E. F. Quax, Y. I. Wolf, J. J. Koehorst, O. Wurtzel, R. van der Oost, W. Ran, F. Blombach, K. S. Makarova, S. J. J. Brouns, A. C. Forster, E. G. H. Wagner, R. Sorek, E. V. Koonin, J. van der Oost, Differential translation tunes uneven production of operon-encoded proteins. *Cell Reports* **4**(5), 938–944 (2013).

8. H. B. Fraser, A. E. Hirsh, D. P. Wall, M. B. Eisen, Coevolution of gene expression among interacting proteins. *Proceedings of the National Academy of Sciences U.S.A.* **101**, 9033–9038 (2004).

9. G. Lithwick, Relative predicted protein levels of functionally associated proteins are conserved across organisms. *Nucleic Acids Research* **33**, 1051–1057 (2005).

10. J. J. Davis, G. J. Olsen, Modal codon usage: Assessing the typical codon usage of a genome. *Molecular Biology and Evolution* **27**, (2010).

11. J. L. López, M. J. Lozano, A. Lagares Jr., M. L. Fabre, W. O. Draghi, M. F. Del Papa, M. Pistorio, A. Becker, D. Wibberg, A. Schlüter, A. Pühler, J. Blom, A. Goesmann, A. Lagares, Codon usage heterogeneity in the multipartite prokaryote genome: Selection-based coding bias associated with gene location, expression level, and ancestry. *mBio* **10**(3), e00505-19 (2019).

12. J. L. López, M. J. Lozano, M. L. Fabre, A. Lagares, Codon usage optimization in the prokaryotic tree of life: How synonymous codons are differentially selected in sequence domains with different expression levels and degrees of conservation. *mBio* **11**, e00766-20 (2020).

13. R. Merkl, A survey of codon and amino acid frequency bias in microbial genomes focusing on translational efficiency. *Journal of Molecular Evolution* **57**(4), 453–466 (2003).

14. E. A. Catoiu, J. Krishnan, G. Li, X. A. Lou, K. Rychel, Y. Yuan, H. Bajpe, A. Patel, D. Choe, J. Shin, J. Burrows, P. V. Phaneuf, D. C. Zielinski, B. O. Palsson, iModulonDB 2.0: Dynamic tools to facilitate knowledge-mining and user-enabled analyses of curated transcriptomic datasets. *Nucleic Acids Research* **53**, D99–D106 (2024).

15. Materials and methods are available as supplementary materials.

16. S. Karlin, J. Mrázek, A. Campbell, D. Kaiser, Characterizations of highly expressed genes of four fast-growing bacteria. *Journal of Bacteriology* **183**, 5025–5040 (2001).

17. S. Vieira-Silva, M. Touchon, S. S. Abby, E. P. C. Rocha, Investment in rapid growth shapes the evolutionary rates of essential proteins. *Proceedings of the National Academy of Sciences* **108**(50), 20030–20035 (2011).

18. S. R. Grigson, G. Bouras, B. E. Dutilh, R. D. Olson, R. A. Edwards, Computational function prediction of bacteria and phage proteins. *Microbiology and Molecular Biology Reviews* **89**, e00022-25 (2025).

19. J. J. Davis, G. J. Olsen, Characterizing the native codon usages of a genome: An axis projection approach. *Molecular Biology and Evolution* **28**, 211–221 (2011).

20. R. Glöckler, A. Tschech, G. Fuchs, Reductive dehydroxylation of 4-hydroxybenzoyl-CoA to benzoyl-CoA in a denitrifying, phenol-degrading *Pseudomonas* species. *FEBS Letters* **251**, 237–240 (1989).

21. A. Ueki, K. Goto, Y. Ohtaki, N. Kaku, K. Ueki, Description of *Anaerotignum aminivorans* gen. nov., sp. nov., a strictly anaerobic, amino-acid-decomposing bacterium isolated from a methanogenic reactor, and reclassification of *Clostridium propionicum*, *Clostridium neopropionicum* and *Clostridium lactatifermentans* as species of the genus *Anaerotignum*. *International Journal of Systematic and Evolutionary Microbiology* **67**(10), 4146–4153 (2017).

22. J. Gibson, M. Dispensa, C. S. Harwood, 4-Hydroxybenzoyl coenzyme A reductase (dehydroxylating) is required for anaerobic degradation of 4-hydroxybenzoate by *Rhodopseudomonas palustris* and shares features with molybdenum-containing hydroxylases. *Journal of Bacteriology* **179**, 634–642 (1997).

23. R. Brackmann, G. Fuchs, Enzymes of anaerobic metabolism of phenolic compounds. *European Journal of Biochemistry* **213**, 563–571 (1993).

24. C. J. Meehan, R. G. Beiko, A phylogenomic view of ecological specialization in the *Lachnospiraceae*, a family of digestive tract-associated bacteria. *Genome Biology and Evolution* **6**(3), 703–713 (2014).

25. M. T. Sorbara, E. R. Littmann, E. Fontana, T. U. Moody, C. E. Kohout, M. Gjonbalaj, V. Eaton, R. Seok, I. M. Leiner, E. G. Pamer, Functional and genomic variation between human-derived isolates of *Lachnospiraceae* reveals inter- and intra-species diversity. *Cell Host & Microbe* **28**(1), 134–146.e4 (2020).

26. V. D. Blondel, J.-L. Guillaume, R. Lambiotte, R. Lefebvre, Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment* **2008**, P10008 (2008).

27. D. Szklarczyk, R. Kirsch, M. Koutrouli, K. Nastou, F. Mehryary, R. Hachilif, A. L. Gable, T. Fang, N. T. Doncheva, S. Pyysalo, P. Bork, L. J. Jensen, C. von Mering, The STRING database in 2023: Protein–protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic Acids Research* **51**(D1), D638–D646 (2022).

28. R. Overbeek, R. Olson, G. D. Pusch, G. J. Olsen, J. J. Davis, T. Disz, R. A. Edwards, S. Gerdes, B. Parrello, M. Shukla, V. Vonstein, A. R. Wattam, F. Xia, R. Stevens, The SEED and the Rapid Annotation of Microbial Genomes using Subsystems Technology (RAST). *Nucleic Acids Research* **42**(D1), D206–D214 (2013).

29. A. V. Sastry, Y. Gao, R. Szubin, Y. Hefner, S. Xu, D. Kim, K. S. Choudhary, L. Yang, Z. A. King, B. O. Palsson, The *Escherichia coli* transcriptome mostly consists of independently regulated modules. *Nature Communications* **10**(1), (2019).

30. B. Papp, C. Pál, L. D. Hurst, Dosage sensitivity and the evolution of gene families in yeast. *Nature* **424**(6945), 194–197 (2003).

31. X.-P. Hu, B. Brahmantio, K. Bartoszek, M. J. Lercher, Most bacterial gene families are biased toward specific chromosomal positions. *Science* **388**, 186–191 (2025).

32. B. Micka, N. Groch, U. Heinemann, M. A. Marahiel, Molecular cloning, nucleotide sequence, and characterization of the *Bacillus subtilis* gene encoding the DNA-binding protein HBsu. *Journal of Bacteriology* **173**(10), 3191–3198 (1991).

33. X. Karaboja, X. Wang, HBsu is required for the initiation of DNA replication in *Bacillus subtilis*. *Journal of Bacteriology* **204**, e00119-22 (2022).

34. F. Braun, S. Durand, C. Condon, Initiating ribosomes and a 5′/3′-UTR interaction control ribonuclease action to tightly couple *B. subtilis* hbs mRNA stability with translation. *Nucleic Acids Research* **45**, 11386–11400 (2017).

35. R. T. Dame, The role of nucleoid-associated proteins in the organization and compaction of bacterial chromatin. *Molecular Microbiology* **56**, 858–870 (2005).

36. M. Hammel, D. Amlanjyoti, F. E. Reyes, J.-H. Chen, R. Parpana, H. Y. H. Tang, C. A. Larabell, J. A. Tainer, S. Adhya, HU multimerization shift controls nucleoid compaction. *Science Advances* **2**, e1600650 (2016).

37. G. Boël, P. C. Smith, W. Ning, M. T. Englander, B. Chen, Y. Hashem, A. J. Testa, J. J. Fischer, H.-J. Wieden, J. Frank, R. L. Gonzalez Jr, J. F. Hunt, The ABC-F protein EttA gates ribosome entry into the translation elongation cycle. *Nature Structural & Molecular Biology* **21**, 143–151 (2014).

38. B. Chen, G. Boël, Y. Hashem, W. Ning, J. Fei, C. Wang, R. L. Gonzalez Jr, J. F. Hunt, J. Frank, EttA regulates translation by binding the ribosomal E site and restricting ribosome-tRNA dynamics. *Nature Structural & Molecular Biology* **21**, 152–159 (2014).

39. S. C. Waterworth, L. V. Flórez, E. R. Rees, C. Hertweck, M. Kaltenpoth, J. C. Kwan, Horizontal gene transfer to a defensive symbiont with a reduced genome in a multipartite beetle microbiome. *mBio* **11**, e02430-19 (2020).

40. Z. Cui, X. Li, J. Shin, H. Gamper, Y.-M. Hou, J. C. Sacchettini, J. Zhang, Interplay between an ATP-binding cassette F protein and the ribosome from *Mycobacterium tuberculosis*. *Nature Communications* **13**, 28078 (2022).

41. A. K. Gupta, V. M. Katoch, D. S. Chauhan, R. Sharma, M. Singh, K. Venkatesan, V. D. Sharma, Microarray analysis of efflux pump genes in multidrug-resistant *Mycobacterium tuberculosis* during stress induced by common anti-tuberculous drugs. *Microbial Drug Resistance* **16**, 21–28 (2010).

42. S. Li, N. C. Poulton, J. S. Chang, Z. A. Azadian, M. A. DeJesus, N. Ruecker, M. D. Zimmerman, K. A. Eckartt, B. Bosch, C. A. Engelhart, D. F. Sullivan, M. Gengenbacher, V. A. Dartois, D. Schnappinger, J. M. Rock, CRISPRi chemical genetics and comparative genomics identify genes mediating drug potency in *Mycobacterium tuberculosis*. *Nat. Microbiol.* **7**, 766–779 (2022).

43. M. Meir, A. Rozenblit, S. Fliger, Y. Geffen, D. Barkan, EttA is likely non-essential in *Staphylococcus aureus* persistence, fitness or resistance to antibiotics. *BMC Microbiology* **20**(1), (2020).

44. T. H. C. Brondijk, A. Nilavongse, N. Filenko, D. J. Richardson, J. A. Cole, NapGH components of the periplasmic nitrate reductase of *Escherichia coli* K-12: Location, topology and physiological roles in quinol oxidation and redox balancing. *Biochemical Journal* **379**, 47–55 (2004).

45. C. Sparacino-Watkins, J. F. Stolz, P. Basu, Nitrate and periplasmic nitrate reductases. *Chemical Society Reviews* **43**(2), 676–706 (2014).

46. C. Outeiral, C. M. Deane, Codon language embeddings provide strong signals for use in protein engineering. *Nature Machine Intelligence* **6**(2), 170–179 (2024).

47. T. Sidi, S. Bahiri-Elitzur, T. Tuller, R. Kolodny, Predicting gene sequences with AI to study codon usage patterns. *Proceedings of the National Academy of Sciences* **122**(1), (2024).

48. D. H. Parks, M. Chuvochina, C. Rinke, A. J. Mussig, P.-A. Chaumeil, P. Hugenholtz, GTDB: An ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. *Nucleic Acids Research* **50**(D1), D785–D794 (2021).

49. J. L. López, A. Litos, M. Siemers, M. D. Boer, E. Padial, S. P. Doijad, B. E. Dutilh, Widespread synchronization of codon usage in functionally related genes, Zenodo (2025).

50. S. Poudel, H. Tsunemoto, Y. Seif, A. V. Sastry, R. Szubin, S. Xu, H. Machado, C. A. Olson, A. Anand, J. Pogliano, V. Nizet, B. O. Palsson, Revealing 29 sets of independently modulated genes in *Staphylococcus aureus*, their

regulators, and role in key physiological response. *Proceedings of the National Academy of Sciences* **117**(29), 17228–17239 (2020).

51. K. Rychel, A. V. Sastry, B. O. Palsson, Machine learning uncovers independently regulated modules in the Bacillus subtilis transcriptome. *Nature Communications* **11**(1), (2020).

52. P. M. Sharp, W.-H. Li, The codon adaptation index—a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Research* **15**(3), 1281–1295 (1987).

53. F. Supek, K. Vlahoviček, Comparison of codon usage measures and their applicability in prediction of microbial gene expressivity. *BMC Bioinformatics* **6**(1), (2005).

54. S. Karlin, J. Mrázek, Predicted highly expressed genes of diverse prokaryotic genomes. *Journal of Bacteriology* **182**, 5238–5250 (2000).

55. F. Wright, The 'effective number of codons' used in a gene. *Gene* **87**, 23–29 (1990).

56. X.-F. Wan, D. Xu, A. Kleinhofs, J. Zhou, Quantitative relationship between synonymous codon usage bias and GC composition across unicellular genomes. *BMC Evol. Biol.* **4**, 1–11 (2004).

57. T. Ikemura, Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes: A proposal for a synonymous codon choice optimal for the *E. coli* translational system. *Journal of Molecular Biology* **151**, 389–409 (1981).

58. A. Elek, coRdon: Codon usage analysis and prediction of gene expressivity. R package version 1.0.3. (2019).

59. M. Wang, C. J. Herrmann, M. Simonovic, D. Szklarczyk, C. von Mering, Version 4.0 of PaxDb: Protein abundance data, integrated across model organisms, tissues, and cell-lines. *Proteomics* **15**, 3163–3168 (2015).

60. D. H. Parks, M. Imelfort, C. T. Skennerton, P. Hugenholtz, G. W. Tyson, CheckM: Assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Research* **25**(7), 1043–1055 (2015).

61. R. D. Olson, R. Assaf, T. Brettin, N. Conrad, C. Cucinell, J. J. Davis, D. M. Dempsey, A. Dickerman, E. M. Dietrich, R. W. Kenyon, M. Kuscuoglu, E. J. Lefkowitz, J. Lu, D. Machi, C. Macken, C. Mao, A. Niewiadomska, M. Nguyen, G. J. Olsen, J. C. Overbeek, B. Parrello, V. Parrello, J. S. Porter, G. D. Pusch, M. Shukla, I. Singh, L. Stewart, G. Tan, C. Thomas, M. VanOeffelen, V. Vonstein, Z. S. Wallace, A. S. Warren, A. R. Wattam, F. Xia, H. Yoo, Y. Zhang, C. M. Zmasek, R. H. Scheuermann, R. L. Stevens, Introducing the Bacterial and Viral Bioinformatics Resource Center (BV-BRC): a resource combining PATRIC, IRD and ViPR. *Nucleic Acids Res*. **51**, D678–D689 (2023).

62. A. A. Hagberg, D. A. Schult, P. J. Swart, Exploring network structure, dynamics, and function using NetworkX. *Proceedings of the Python in Science Conference*, 11–15 (2008).

63. G. Csárdi, T. Nepusz, The igraph software package for complex network research. *InterJournal,* **Complex Systems**, 1695 (2006).

64. D. Machado, S. Andrejev, M. Tramontano, K. R. Patil, Fast automated reconstruction of genome-scale metabolic models for microbial species and communities. *Nucleic Acids Research* **46**(15), 7542–7553 (2018).

65. Z. Zhou, Y. Dang, M. Zhou, L. Li, C. Yu, J. Fu, S. Chen, Y. Liu, Codon usage is an important determinant of gene expression levels largely through its effects on transcription. *Proc. Natl. Acad. Sci. U.S.A.* **113**, E6117–E6125 (2016).

66. J. J. Davis, S. Gerdes, G. J. Olsen, R. Olson, G. D. Pusch, M. Shukla, V. Vonstein, A. R. Wattam, H. Yoo, PATtyFams: Protein families for the microbial genomes in the PATRIC database. *Frontiers in Microbiology* **7**, 118 (2016).

67. M. Pagel, Inferring evolutionary processes from phylogenies. *Zoologica Scripta* **26**(4), 331–348 (1997).

Acknowledgments

Funding:

This study was supported by NWO Groen II project number ALWGR.2017.002, the European Research Council (ERC) Consolidator grant 865694: DiversiPHI, and the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC 2051 – Project-ID 390713860, and the Alexander von Humboldt Foundation in the context of an Alexander von Humboldt-Professorship founded by the German Federal Ministry of Education and Research.

Authors contributions:

Conceptualization: JLL, BED
Methodology: JLL, MDB, AL
Investigation: JLL
Visualization: JLL
Funding acquisition: BED
Project administration: BED
Resources: BED
Supervision: BED, JLL
Writing - original draft: JLL
Writing - review & editing: JLL, BED
Software: JLL, AL
Data curation and validation: SPD, JLL, MS, MDB, AL, EP
Formal analysis: JLL, BED, MS, AL

Competing interests:

The authors declare no conflict of interest.

Data and materials availability:

All data and code generated and/or analyzed in this study are deposited at Zenodo *(49)* and GitHub at https://github.com/JoseLopezArcondo/CUCA.git

Supplementary Materials

The PDF file includes:

Materials and Methods
Supplementary Text
Figs. S1 to S13
Tables S1 to S5
References (50–66)

Materials and Methods

Codon usage bias index selection

We calculated and compared different codon usage bias (CUB) indexes to expression data available from *Escherichia coli* K-12 MG1655 *(28)*, *Staphylococcus aureus* USA 300 TCH1516 *(50)*, and *Bacillus subtilis* 168 *(51)* (Supplementary Table 1). These indices included CAI *(52)*, MELP *(53)*, E *(54)*, GCB *(13)*, MCB *(55)*, SCUO *(56)* and Fop *(57)*, and were computed using the coRdon R package *(58)*. As input, we used nucleotide coding direct sequences and a list of putatively highly expressed genes, either genes coding for ribosomal proteins or a set of manually selected highly expressed genes (PHX) based on proteome data from the PaxDB database *(59)*. Proteins shorter than 80 amino acids were excluded, as codon bias measures are generally unreliable for short sequences.
To evaluate each index's ability to predict gene expression, we calculated Spearman correlations between index and transcriptome expression data for the strains listed above, using data from the iModulon database (https://imodulondb.org/index.html) *(14)*. For GCB calculation across genomes, we required that more than half of the ribosomal genes be among the top 200 genes by GCB value, to avoid cases where GCB index iterations converge to a local optima distant from the ribosomal genes (e.g., in the accessory genome).
Finally, we included a custom modified version of GCB (GDB mod), consisting of breaking the optimization loop when the new reference set includes <50% of the ribosomal genes. This allows a few iteration steps without drifting to different local maximum reference sets away from the ribosomal genes. While GCB only allowed us to include a subset of genomes, GCB

mod allowed us to include all high quality genomes in our analyses. We selected the GCB index for the first part of our analyses, including results shown in Fig. 1, and GCB mod index, for all analyses in this study, both using ribosomal protein coding genes as the reference set.
To normalize CUB values within each genome and reduce the impact of genomic confounders, we calculated GCB z-scores by subtracting the genome-wide mean GCB and dividing by its standard deviation.

Genome selection and annotation

We downloaded genomes from the BV-BRC database (https://www.bv-brc.org/) (61) as of September 2023 that were also listed as reference species in GTDB release 207 *(48)* and affiliated with the 5 main phyla, *Bacillota*, *Bacillota A*, *Actinomycetota*, *Pseudomonadota*, and *Bacteroidota*, and with another 10 less represented phyla (Supplementary Table 1). We retained genomes meeting the following criteria: more than 30 ribosomal proteins, more than 1,000 CDS, and CheckM *(60)* completeness and contamination scores above 80% and below 10%, respectively. Finally, we only kept genomes for which ribosomal protein genes had a significantly lower Nc value (Mann-Whitney test p value<0.01) than the rest of genes in the genome, to avoid the effect of other dominating selective pressures, such as GC content and G+C strand bias, following the rationale in Lithwick *et al*. *(9)*.

Functional annotations from the BV-BRC database

We used the “features.tab”, “pathway.tab”, and “subsystem.tab” annotation files from the BV-BRC database, which provide SEED-based annotations of protein functions and their association with specific metabolic pathways, subsystems, subclasses, and classes *(28)*. Cross-genera protein families (PGFams) were used to group orthologous protein-coding genes, offering a standardized approach to represent and compare the protein content of bacterial genomes across genomes in the BV-BRC *(66)*.

Genomic mean GCB calculation

The mean GCB of all genes in a genome, here referred to as the genomic mean GCB, reflects the strength of selection on codon usage. Genomes with lower genomic mean GCB show stronger selection, as ribosomal genes are farther from the rest of the genes, resulting in most of the genes having more negative values *(13)*.
We obtained GC content of genomes from their GTDB associated metadata. To reduce collinearity between co-variates and interaction terms, both mean GCB and GC content were centered by subtracting their mean across genomes and used as co-variates in the models (see below) as centered meanGCB and centered GC content, respectively.

Codon usage co-adaptation (CUCA) in PGFam pairs across genomes

For genes belonging to single-copy PGFams per genome, we calculated pairwise Spearman correlations, as well as a series of phylogenetically informed regression models (CUCA models, described below), between their GCB and GCB z-scores. For the five main phyla, only PGFam pairs that co-occurred in at least 100 genomes were included, ensuring sufficient power to detect significant correlations widespread among different species. For the additional phyla, this threshold was reduced to 50. PGFams with more than one representative gene per genome were excluded. Since these correlations reflect CUB between pairs of functions, significantly positive values are considered here as CU co-adaptations.
CUCA models were implemented using the phylolm R package and a GTDB-based phylogenomic tree (bac120_r220.tree) subsetted to comprise the selected genomes for each phylum. For each pair, GCB values and GCB z-scores of one PGFam were used to model the values of the other. Four models were fitted including the phylogenomic tree: 1) A model only controlling for the phylogenetic signal (CUCA-P), 2) a model including genomic mean GCB as a co-variate (CUCA-PG), 3) a model including the genomic mean GCB as a co-variate but including an interaction term with the independent variable (CUCA-PGI), and 4) a full model including the genomic mean GCB, and genomic GC content, as co-variates, each with interaction terms (CUCA-PGGI). Model estimates and *t* values (estimate / SE) were then extracted to quantify co-adaptation for each PGFam pair (table S3).

iModulon, transcriptional units, subunits, and metabolic analyses

We obtained iModulon tables from *Escherichia coli* K-12 MG1655 *(29)*, and mapped the genes within each iModulon to their corresponding PGFam annotations, thereby grouping PGFams that were part of the same iModulon. Genes from the same transcriptional units in *Escherichia coli* K-12 MG1655 were obtained from EcoCyc, and their corresponding PGFam annotations were mapped in order to identify PGFams belonging to the same transcriptional unit. Subunits of the same enzyme were identified in the RAST-based functional annotations by searching the keywords “subunit” and “chain”, followed by manual verification. PGFams with Enzyme Commission (EC) annotations were selected within each phylum to assess metabolic connections. Using the networkx package v3.4.2 *(62)* and the universal model from CarveMe *(63)*, we built a metabolite–reaction network linking reactions by shared metabolites. Corresponding reactions for each PGFam were retrieved from the CarveMe general model, and the minimum number of metabolic steps between enzymes was determined by counting intermediate metabolites, excluding the 1,000 most common (e.g., water, ATP, NADH).

Clustering and STRING based analyses

Using the Spearman's ρ values and the Phylolm estimates and *t* values from co-adaptation pairs, we constructed graphs in the igraph R package version 2.1.4 *(64)*, with these values used as edge weights. We then applied the Louvain method with a custom iterative approach to detect clusters, recursively subdividing them into smaller internal clusters until no further partitioning was possible.

To link PGFams with STRING *(27)*, we retrieved all STRING interaction gene pairs from genomes included in our dataset (table S1) and that were also represented in the STRING database, using the STRINGdb R package version 2.16.4. For each phylum, we mapped genes to PGFams using the refseq_locus_tag identifier in BV-BRC annotations. Across all genomes, when a PGFam pair was present multiple times, we retained only the highest combined interaction score, since we were interested in the highest reported combined score. This process produced a non-redundant PGFam pair dataset. For benchmarking purposes, we used this dataset to select the highest STRING combined score *per* PGFam within each cluster, and the mean per cluster was compared between different models (Fig. 3A, fig. S5, and fig. S9).

Analyses of duplicate genes within the same PGFam and genome

We identified cases where multiple genes from the same genome belonged to the same PGFam, focusing on cases with exactly two genes ("duplicates"). To confirm homology, we performed BLASTp comparisons and retained only pairs with DIAMOND query and subject coverages above 90%.

For all duplicates in the five taxa, we extracted 10 upstream and 10 downstream genes (neighboring regions) and analyzed their CUB patterns. We excluded overlapping regions containing at least one shared gene ("far" duplicates) and, in parallel, analyzed overlapping regions sharing at least one gene ("close" duplicates). Paralogs are considered within the same region if there are <40 genes between them, and remote paralogs otherwise. For remote paralogs, a distinct genetic neighborhood means that there are no common PGFams among the 20 upstream or 20 downstream genes, whereas similar genetic neighborhoods share at least two other PGFam in that region. For each duplicate, we also sampled a random gene pair from the same genome to control for genomic CUB patterns that might arise simply from co-occurence within a genome. The latter still show weak positive correlations due to overall genome-wide CUB patterns.

Codon and amino acid composition distances

To quantify compositional similarity between genes, we calculated codon and amino acid–based distances using 300 randomly selected genomes from *Pseudomonadota*. We quantified distances within each genome, using only gene pairs corresponding to all PGFam pairs analyzed in our models.

Codon composition distances were computed in two steps. First, for each of the 18 amino acids with synonymous codons, we calculated the Manhattan distance as the sum of absolute differences in relative codon usage frequencies between each pair of genes. For a given amino acid, this distance equals 0 when codon usage is identical and 2 when there is no overlap in codon frequencies, irrespective of the number of synonymous codons for that amino acid.

Second, the overall codon composition distance between two genes was obtained as a multidimensional Euclidean distance, calculated as the square root of the sum of squared amino acid specific distances. The resulting values range from 0 (identical codon composition) to $\sqrt{72} \simeq 8.48$ (no overlap in codon usage).

For amino acid composition, relative amino acid frequencies were computed for each gene, and Euclidean distances were calculated between all gene pairs, yielding values ranging from 0 (identical amino acid composition) to 1 (no shared amino acid usage).

Figures and code availability

All figures were generated in R using ggplot2 (v3.5.2) and, when necessary, edited in InkScape. Most of the analyses were performed with custom R scripts, available at https://github.com/JoseLopezArcondo/CUCA.

Supplementary Text

This section should provide additional information regarding control or supplemental experiments, field sites, observations, hypotheses, etc., that bear directly on the arguments of the main paper. Further discussion or development of arguments beyond those in the main text is not permitted in supplementary text. This section can be referred to in the main text as "supplementary text" (reference note not required).

S1 – GCB and GCB_mod are the CUB indices that best reflect expression

As codon usage has been proposed to enable bacteria to fine-tune gene expression under different conditions *(65)*, we benchmarked sixteen CUB indices against gene expression levels across a range of experimental conditions. Expression data was obtained from the iModulon database *(51, 15)*. Among the tested indices, the gene codon bias (GCB) index showed the strongest and most consistent correlation with gene expression across genomes compared with other commonly used indices (fig. S1 and fig. S2).

The selected GCB index initially uses the set of ribosomal protein genes as a seed reference, estimates GCB values for all genes in the genome using this reference, and then iteratively

searches for the optimal reference set, comprising the genes with the highest GCB values, until reaching a local maximum *(13)*. For our first analyses, we retained only those genomes in which at least 50% of ribosomal protein genes were included in the final optimized reference set. This filtering was necessary because, in most genomes, the iterative optimization converged to alternative local maxima driven by accessory genes whose codon usage patterns are closer to those of ribosomal genes *(12)*. Thus, the genomes used in the initial analyses (Figs. 1–2) were those meeting this criterion. Comparable results were obtained using a modified version of the index (GCB_mod; fig. S1 and fig. S2) *(15)*, which enabled the inclusion of additional reference species to increase statistical power in our co-adaptation analyses. We therefore used GCB_mod for the second part of the study (Figs. 3–4), as it allowed a larger and more phylogenetically diverse dataset, yielding more clusters and including additional phyla.

S2 – Most duplicated PGFams conserve similar CUB within the same genome

To examine how homologs within the same genome adapt their CUB, we analyzed PGFams containing multiple gene copies per genome. We found that 75-88% of PGFams included at least one genome with more than one copy. Although some PGFams showed a high number of gene copies per genome, such as those encoding transposases and other mobile genetic elements, ABC transporters, two-component system proteins, transcriptional regulators, motility and chemotaxis-related proteins, Type VI secretion system components, and RND efflux system proteins, most multiple copy PGFams were represented by two-copy genes (fig. S13).

We detected between 12,184 and 44,381 cases of duplicated genes across 267 to 1,122 genomes and 568 to 1,896 PGFams among the main five phyla, with averages ranging from 1.1 to 3.8 duplicates per genome. To focus on true isofunctional homologs, we retained only those gene pairs with $>90\%$ query and subject coverage in DIAMOND blastp searches, which represented 7.1-14.1% of all detected duplicates. When comparing GCB values among duplicates located in non-overlapping genomic regions *(15)*, we observed a positive and significant correlation across the five phyla (fig S3). To rule out potential biases arising from genome-wide CUB patterns shared by duplicated genes, we repeated the same analyses using randomized sets of gene pairs from the same genomes. These random pairs also showed positive but significantly lower correlations than the duplicates. Across all five phyla, duplicated genes had significantly higher ρ values compared to random pairs (Fisher's Z-transformation test, p value < 0.05, table S2), indicating that functional constraints, rather than genomic context alone, primarily drive the conservation of CUB optimization among duplicates. We also found that correlations were stronger for duplicates located within the same genomic region, or for those positioned distantly but sharing neighboring PGFams, likely reflecting functional conservation mediated by homologous recombination that preserved ancestral CUB patterns. Importantly, even after removing these cases and retaining only duplicates without shared

neighboring PGFams, correlations remained significantly positive (fig. S3). Altogether, these results indicate that duplicated genes generally preserve their CUB optimization patterns, suggesting that codon usage adaptation is maintained following gene duplication due to functional constraints and not only due to shared ancestry.

Figures

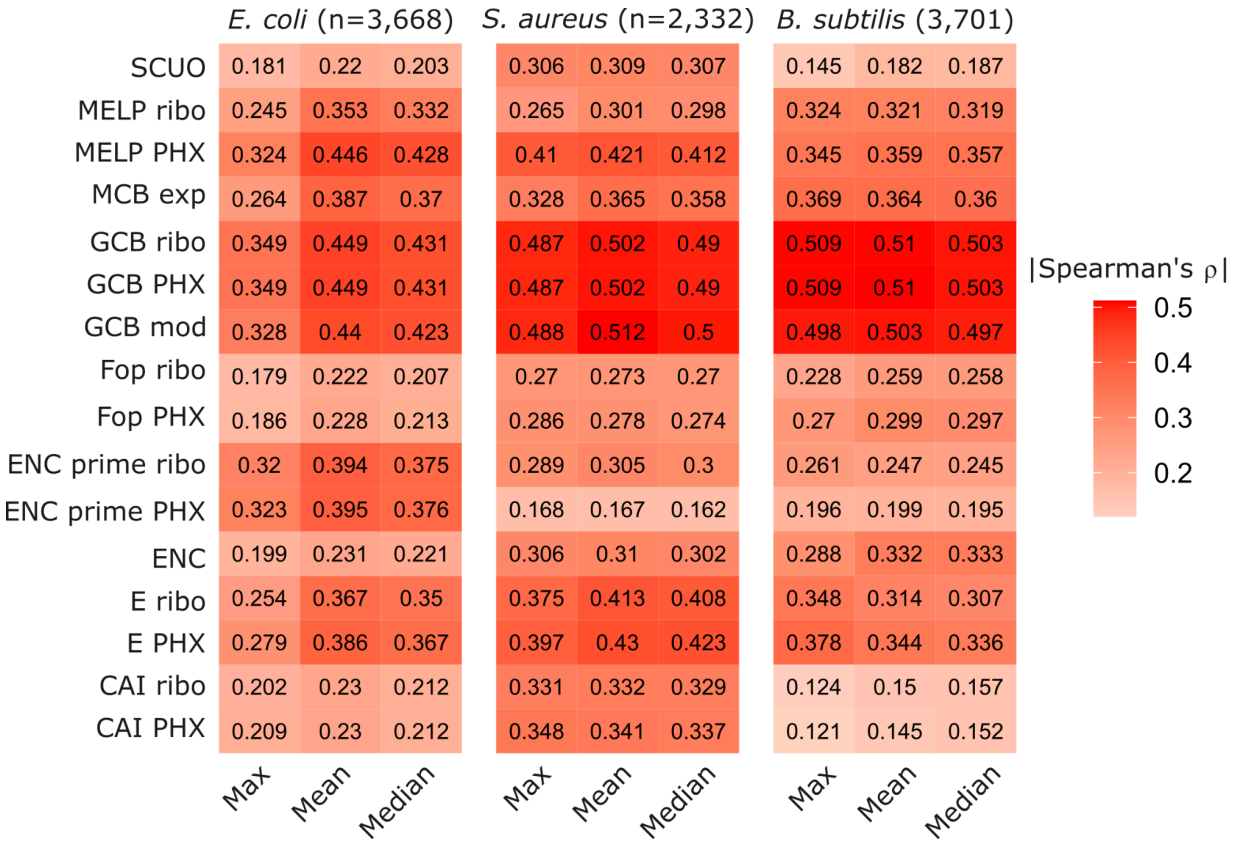


Fig. S1.

Spearman correlations between CUB indices and gene expression levels. To benchmark the CUB indexes (rows), transcriptional data from the iModulon database *(14)* were summarized using different statistical measures, and absolute Spearman correlations between CUB indexes and gene expression were calculated. CUB indices were computed with the coRdon package *(58)* using two sets of putatively highly expressed reference genes: ribosomal proteins ("ribo") and a manually curated set ("PHX") comprising genes with high protein expression in the PaxDB database *(59)* in at least two of the three species. Maximum, mean, and median expression values were calculated from TPM-normalized expression values across a wide range of conditions: (**A**) *Escherichia coli* K-12 MG1655, n = 3,668 *(29)*, (**B**) *Staphylococcus. aureus* USA300 TCH1516, n = 2,332 *(50)*, and (**C**) *Bacillus subtilis* 168, n = 3,701 *(51)*. For subsequent analyses, we selected GCB *(13)* and GCBmod *(15)*, both calculated using ribosomal proteins as the reference set, as this approach showed the strongest correlations and was simpler to implement than the PHX-based method.

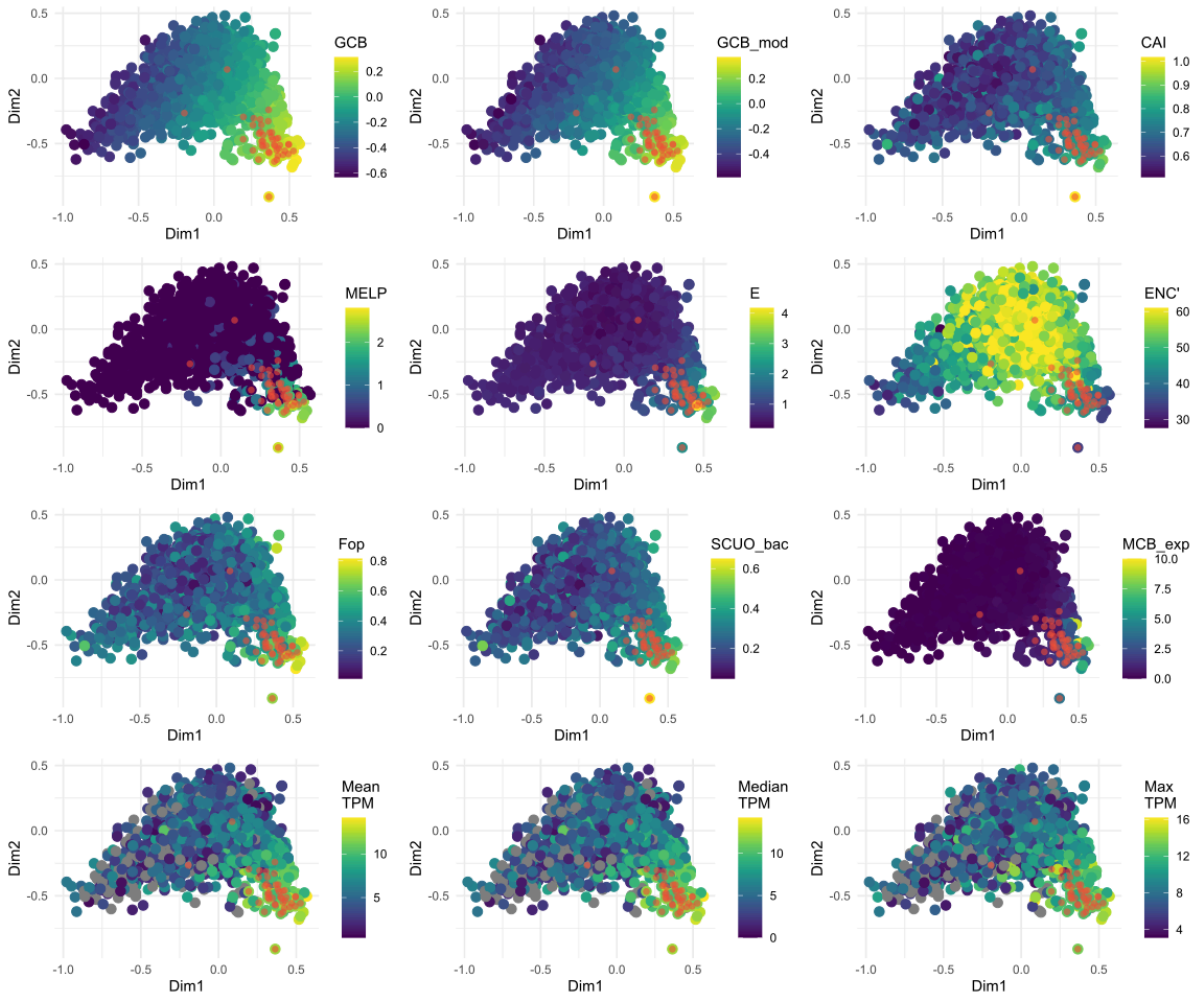


Fig. S2.

Factorial correspondence analysis of the *Escherichia coli* K-12 MG1655 genome. Different CUB indices (GCB, GCB_mod, CAI, MELP, E, ENC', Fop, SCUO, and MCB) are shown in distinct colors to illustrate how each index represents gene-level CUB. Ribosomal genes (red dots), were used as the reference set of highly expressed genes. Maximum, mean, and median expression values were calculated from TPM-normalized expression data across a wide range of conditions obtained from the iModulon database *(14)*. Among the tested indices, GCB decreases most gradually from ribosomal genes and shows the strongest correlation with expression data (fig. S1). All indices were calculated using ribosomal genes as the reference. Gray dots denote genes lacking expression data.

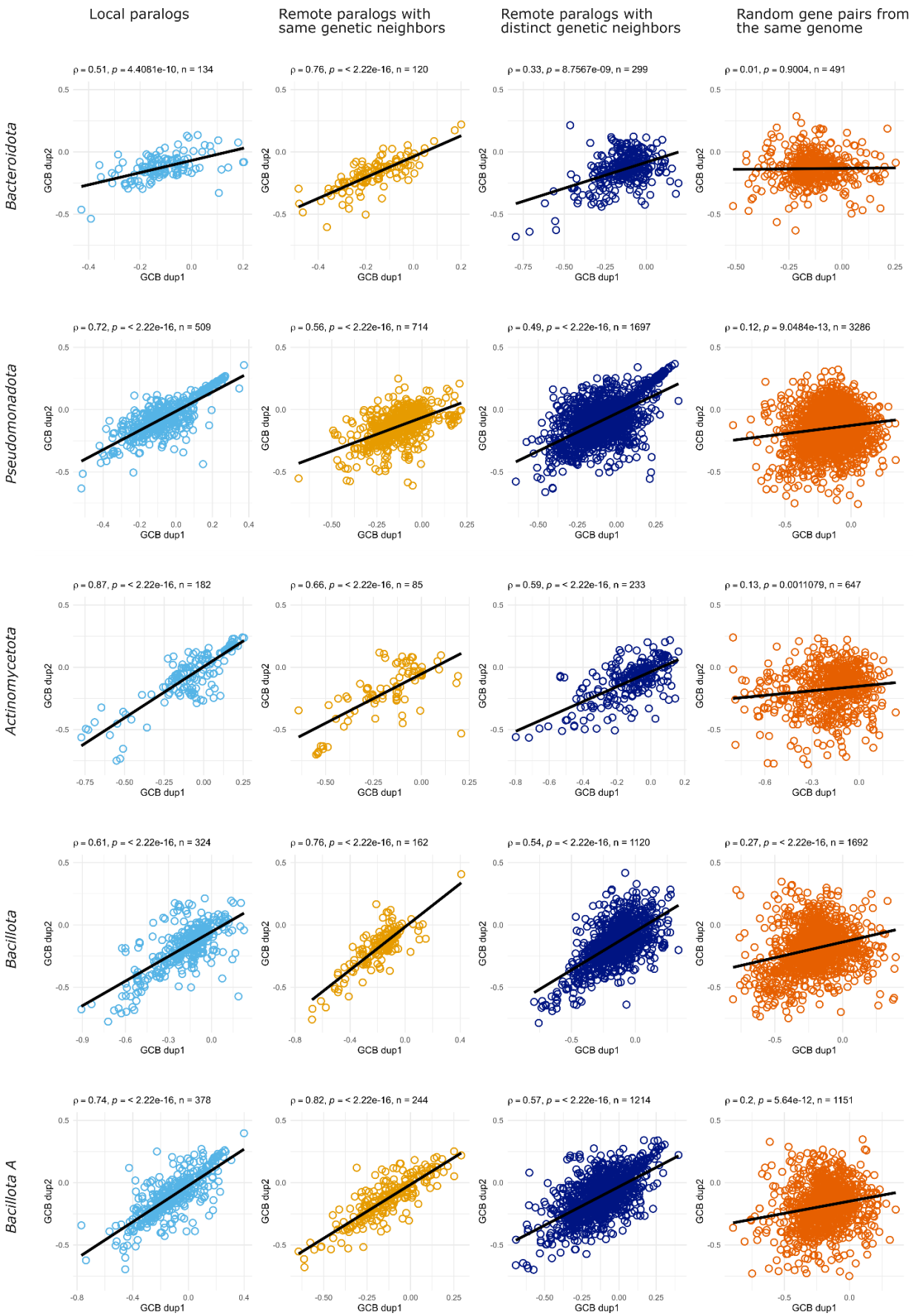

Local paralogs
Remote paralogs with same genetic neighbors
Remote paralogs with distinct genetic neighbors
Random gene pairs from the same genome
Bacteroidota
Pseudomonadota
Actinomycetota
Bacillota
Bacillota A
ρ = 0.51, p = 4.4081e-10, n = 134
ρ = 0.76, p = < 2.22e-16, n = 120
ρ = 0.33, p = 8.7567e-09, n = 299
ρ = 0.01, p = 0.9004, n = 491
ρ = 0.72, p = < 2.22e-16, n = 509
ρ = 0.56, p = < 2.22e-16, n = 714
ρ = 0.49, p = < 2.22e-16, n = 1697
ρ = 0.12, p = 9.0484e-13, n = 3286
ρ = 0.87, p = < 2.22e-16, n = 182
ρ = 0.66, p = < 2.22e-16, n = 85
ρ = 0.59, p = < 2.22e-16, n = 233
ρ = 0.13, p = 0.0011079, n = 647
ρ = 0.61, p = < 2.22e-16, n = 324
ρ = 0.76, p = < 2.22e-16, n = 162
ρ = 0.54, p = < 2.22e-16, n = 1120
ρ = 0.27, p = < 2.22e-16, n = 1692
ρ = 0.74, p = < 2.22e-16, n = 378
ρ = 0.82, p = < 2.22e-16, n = 244
ρ = 0.57, p = < 2.22e-16, n = 1214
ρ = 0.2, p = 5.64e-12, n = 1151
GCB dup1
GCB dup2

Fig. S3.

Correlation between duplicate genes from the same PGFam and genome in main phyla. Both genomically co-localized paralogs, and remote paralogs with similar genomic neighbourhoods present the highest correlation values, while distant duplicates without evident synteny present a lower but still significant positive correlation.

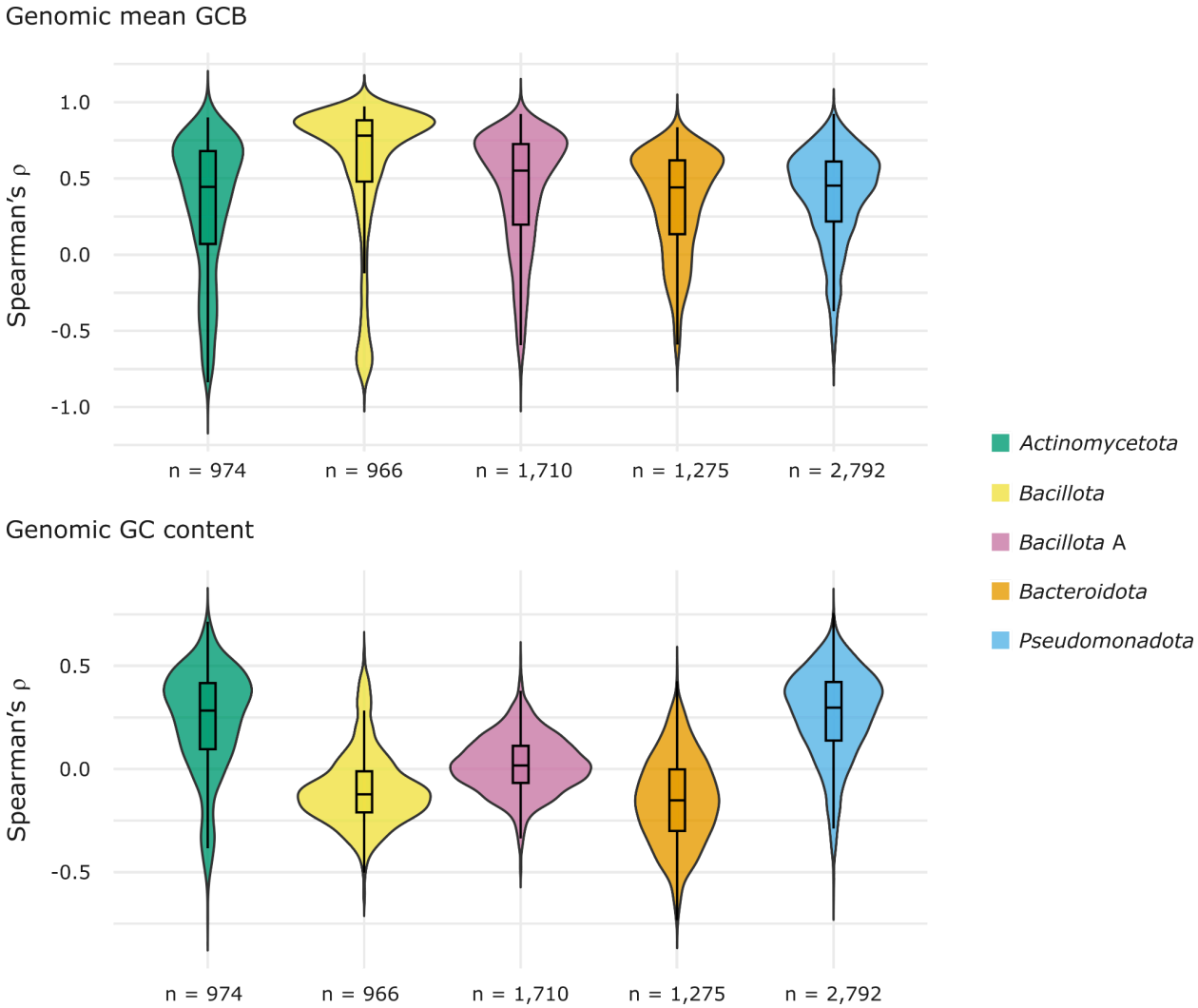


Fig. S4.

Genomic mean GCB and GC content in genomes are highly correlated with GCB and are factors that influence co-adaptation values. Spearman correlations were calculated between single-copy PGFam GCB values and the corresponding genomic mean CUB or GC content and their ρ values are shown for each phyla (table S2). Only PGFams present in more than 100 genomes from the small genome set (table S1) were used for the correlations, whose number of correlations (n) is indicated for each phylum.

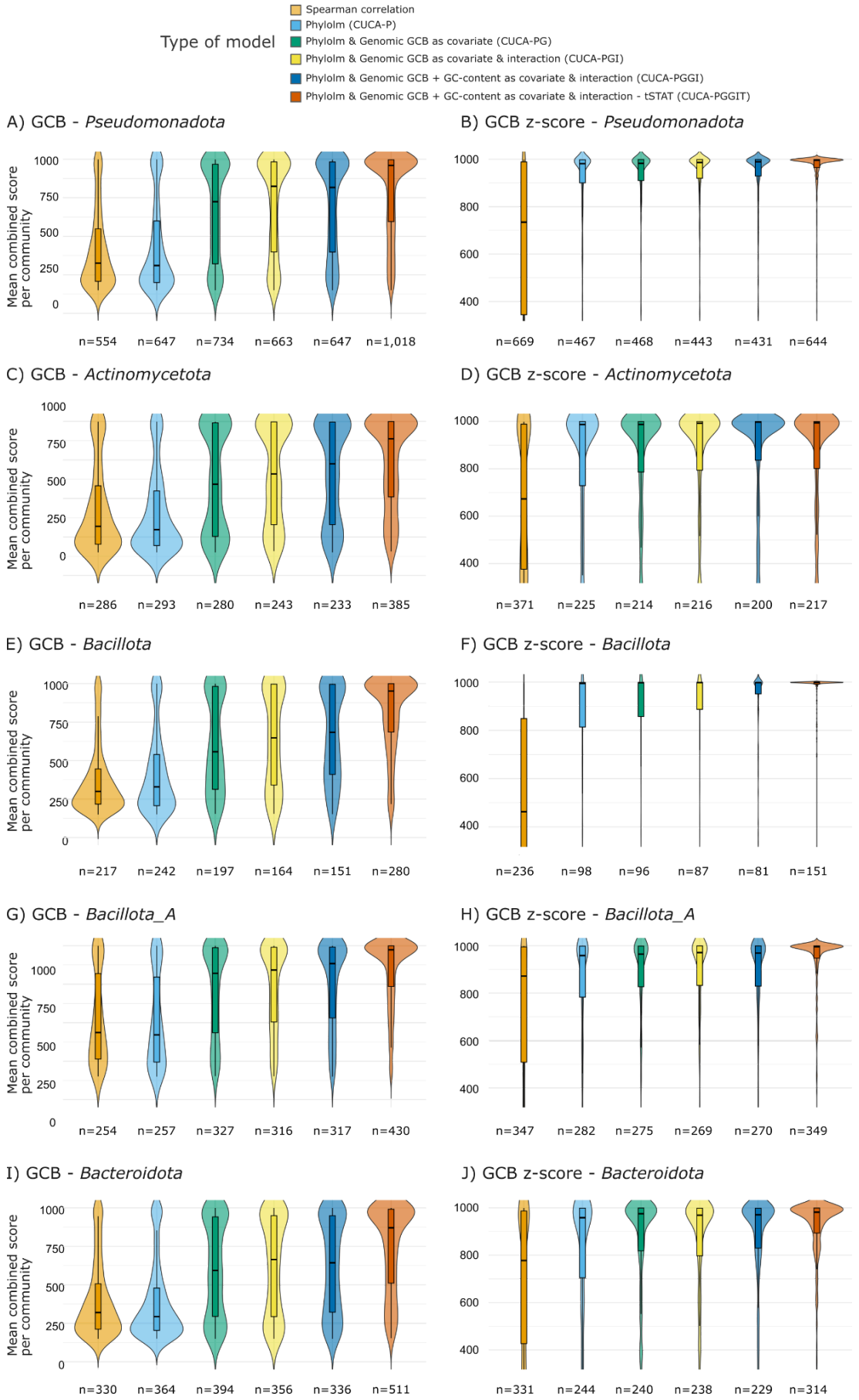

Type of model
Spearman correlation
PhyloIm (CUCA-P)
PhyloIm & Genomic GCB as covariate (CUCA-PG)
PhyloIm & Genomic GCB as covariate & interaction (CUCA-PGI)
PhyloIm & Genomic GCB + GC-content as covariate & interaction (CUCA-PGGI)
PhyloIm & Genomic GCB + GC-content as covariate & interaction - tSTAT (CUCA-PGGIT)
A) GCB - Pseudomonadota
Mean combined score per community
n=554 n=647 n=734 n=663 n=647 n=1,018
B) GCB z-score - Pseudomonadota
n=669 n=467 n=468 n=443 n=431 n=644
C) GCB - Actinomycetota
n=286 n=293 n=280 n=243 n=233 n=385
D) GCB z-score - Actinomycetota
n=371 n=225 n=214 n=216 n=200 n=217
E) GCB - Bacillota
n=217 n=242 n=197 n=164 n=151 n=280
F) GCB z-score - Bacillota
n=236 n=98 n=96 n=87 n=81 n=151
G) GCB - Bacillota_A
n=254 n=257 n=327 n=316 n=317 n=430
H) GCB z-score - Bacillota_A
n=347 n=282 n=275 n=269 n=270 n=349
I) GCB - Bacteroidota
n=330 n=364 n=394 n=356 n=336 n=511
J) GCB z-score - Bacteroidota
n=331 n=244 n=240 n=238 n=229 n=314

Fig. S5.

Clusters obtained with iterative Louvain using different correlation and CUCA models. GCB (**ACEGI**) or GCB z-score (**BDFHJ**) values for each pair of PGFams across genomes were used as input for Spearman correlation or phylolm regression models. We used phylolm models with a GTDB based tree as a phylogenetic control, and Genomic GCB of each genome as a covariate, with and without an interaction term to our independent variable. Spearman's ρ values or Phylolm estimates or *t* values from the CUCA models were used as weights to generate networks using the igraph R package. An iterative Louvain clustering approach was subsequently applied and resulting clusters were mapped to STRING interactions in these genomes *(15)*. We obtained more biologically meaningful clusters (higher mean STRING combined scores) using GCB z-scores, and including the genomic GCB and GC content as covariate and interaction. N represents the number of clusters obtained with at > 2 members having STRING interaction scores.

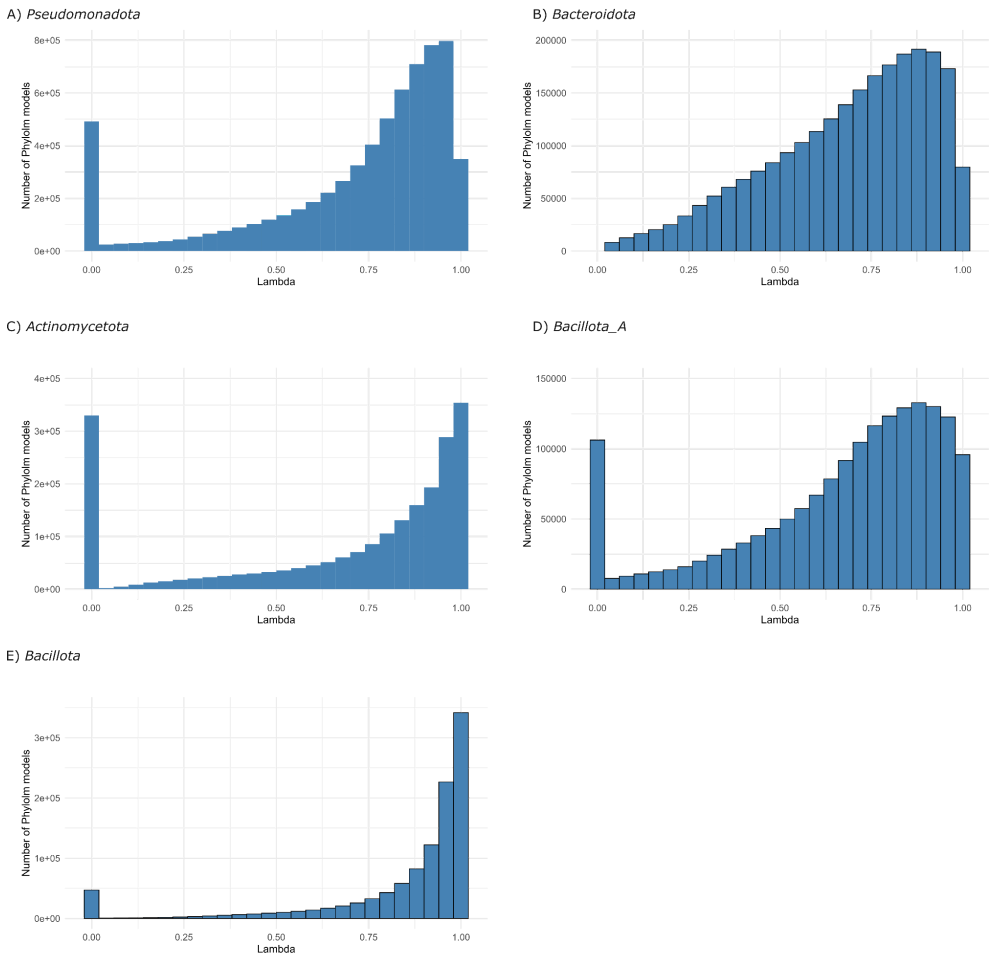
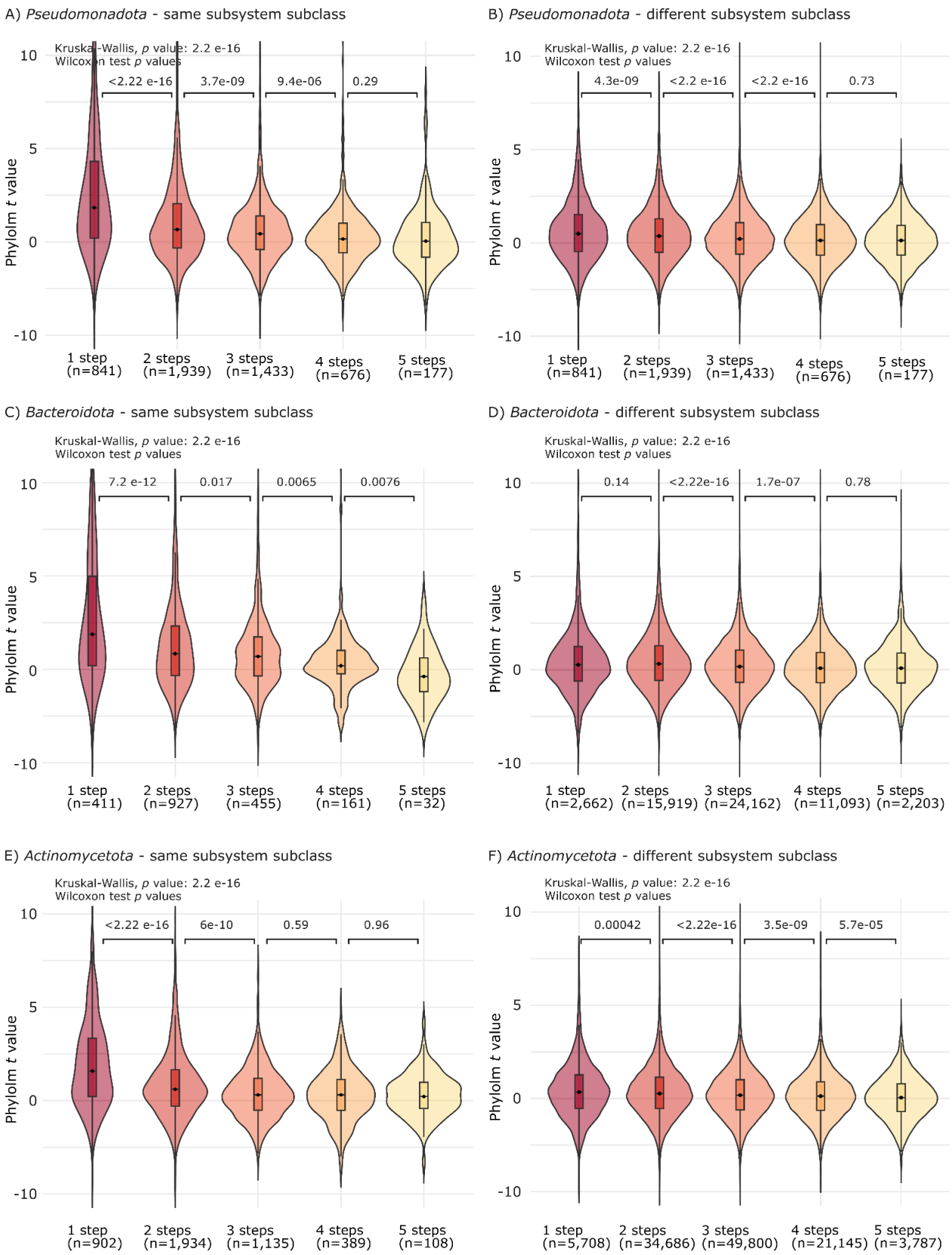


Fig. S6.

Histogram of lambda values in CUCA-P models. Pagel's λ values *(67)* were extracted from the output of each CUCA-P model and are displayed as histograms. Values approaching 1 indicate a strong phylogenetic signal in the models.


A) *Pseudomonadota* - same subsystem subclass
Kruskal-Wallis, *p* value: 2.2 e-16
Wilcoxon test *p* values
<2.22 e-16
3.7e-09
9.4e-06
0.29
Phylolm *t* value
10
5
0
-10
1 step (n=841)
2 steps (n=1,939)
3 steps (n=1,433)
4 steps (n=676)
5 steps (n=177)
B) *Pseudomonadota* - different subsystem subclass
Kruskal-Wallis, *p* value: 2.2 e-16
Wilcoxon test *p* values
4.3e-09
<2.2 e-16
<2.2 e-16
0.73
Phylolm *t* value
10
5
0
-10
1 step (n=841)
2 steps (n=1,939)
3 steps (n=1,433)
4 steps (n=676)
5 steps (n=177)
C) *Bacteroidota* - same subsystem subclass
Kruskal-Wallis, *p* value: 2.2 e-16
Wilcoxon test *p* values
7.2 e-12
0.017
0.0065
0.0076
Phylolm *t* value
10
5
0
-10
1 step (n=411)
2 steps (n=927)
3 steps (n=455)
4 steps (n=161)
5 steps (n=32)
D) *Bacteroidota* - different subsystem subclass
Kruskal-Wallis, *p* value: 2.2 e-16
Wilcoxon test *p* values
0.14
<2.22e-16
1.7e-07
0.78
Phylolm *t* value
10
5
0
-10
1 step (n=2,662)
2 steps (n=15,919)
3 steps (n=24,162)
4 steps (n=11,093)
5 steps (n=2,203)
E) *Actinomycetota* - same subsystem subclass
Kruskal-Wallis, *p* value: 2.2 e-16
Wilcoxon test *p* values
<2.22 e-16
6e-10
0.59
0.96
Phylolm *t* value
10
5
0
-10
1 step (n=902)
2 steps (n=1,934)
3 steps (n=1,135)
4 steps (n=389)
5 steps (n=108)
F) *Actinomycetota* - different subsystem subclass
Kruskal-Wallis, *p* value: 2.2 e-16
Wilcoxon test *p* values
0.00042
<2.22e-16
3.5e-09
5.7e-05
Phylolm *t* value
10
5
0
-10
1 step (n=5,708)
2 steps (n=34,686)
3 steps (n=49,800)
4 steps (n=21,145)
5 steps (n=3,787)

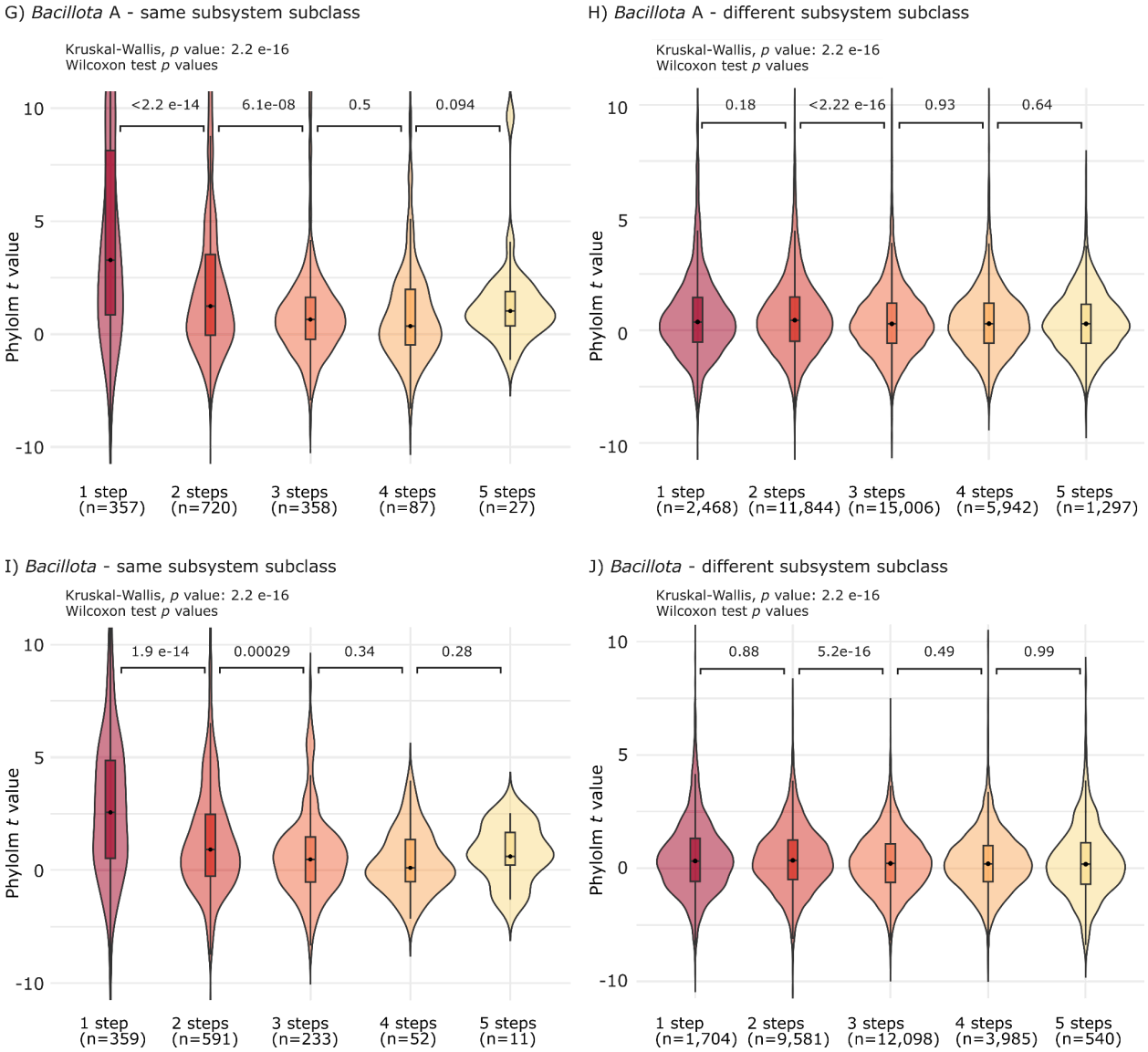


Fig. S7.

Metabolic co-adaptation is higher in closer enzymes in metabolism, especially when classified to the same subsystem. CUCA-PGGI derived Phylolm *t* values increase when genes from the same (**A,C,E,G,I**) and different (**B,D,F,H,J**) subsystem subclass are closer in metabolic pathways (enzymes 1 step apart in metabolic networks) compared to more distant genes (5 steps apart).

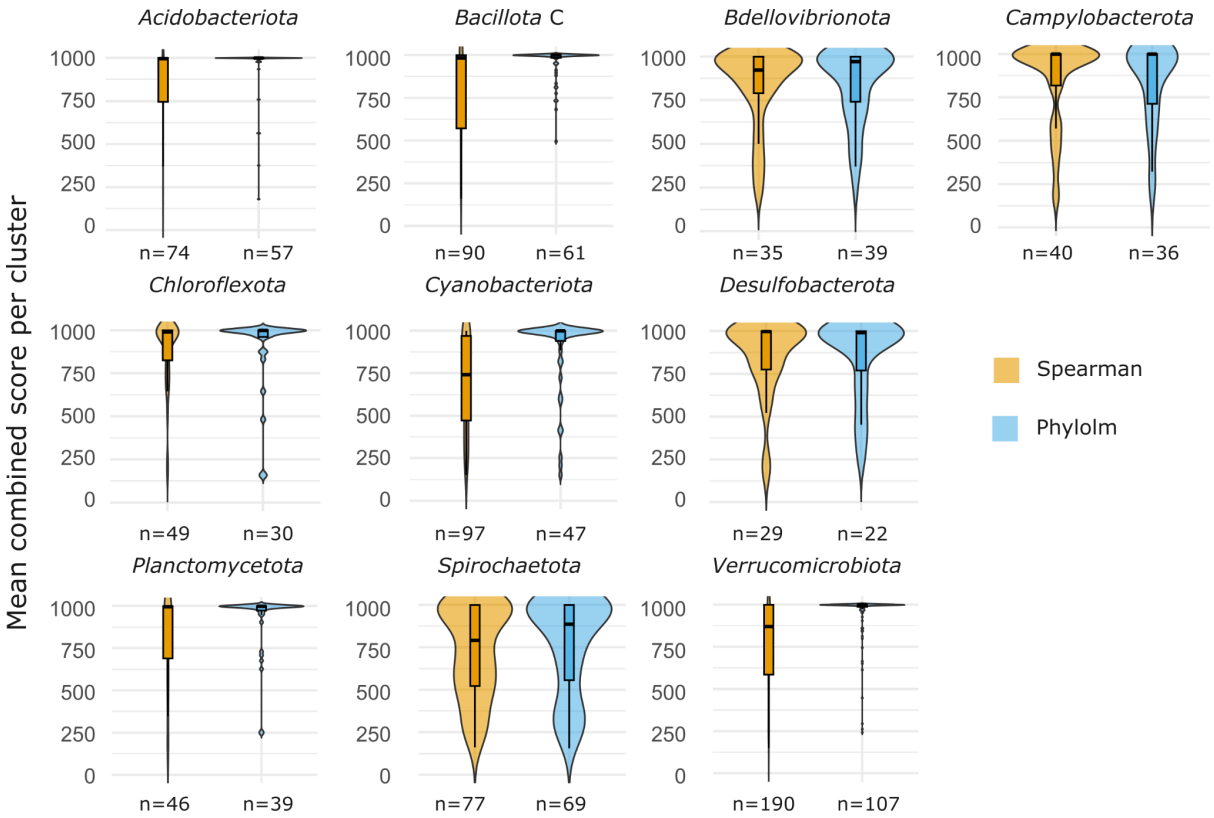


Fig. S8.

Clusters obtained with iterative Louvain across an extended set of phyla. GCB z-score values for each pair of PGFam across genomes were used as input for Spearman correlation, or CUCA-P models. We used a GTDB based tree as a phylogenetic control. Spearman's ρ values or CUCA-P *t* values were used as weights to generate networks using the igraph R package. An iterative Louvain clustering approach was subsequently applied and resulting clusters were mapped to STRING interactions in these genomes *(15)*. N represents the number of clusters obtained with members having known interactions within clusters (n of members > 2).

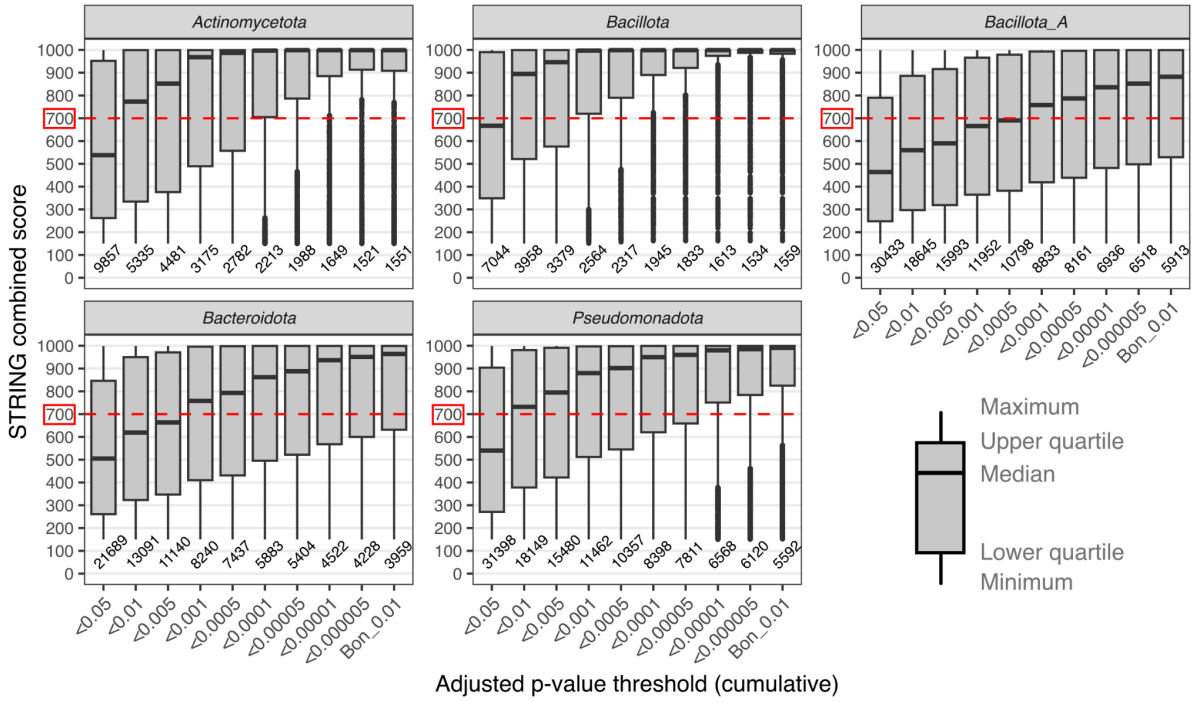


Fig. S9.

Comparison of *p* adjusted thresholds in CUCA-PGGI *t* values. Different Benjamini-Hochberg (BH) p adjusted cutoffs were applied to co-adaptation pairs which were subsequently mapped to their STRING combined scores. X axis denote the BH cutoff values and the Bonferroni stringent cutoff (Bon_0.01 for Bonferroni p adjusted <0.01). Each phylum results in a different optimal cutoff where its median combined score reaches a value of ca. 700 (i.e., Actinomycetota cutoff < 0.01, table S4). The box represents the interquartile range (IQR: 25th–75th percentile), the horizontal line inside the box is the median, and the whiskers extend to the minimum and maximum non-outlier values.

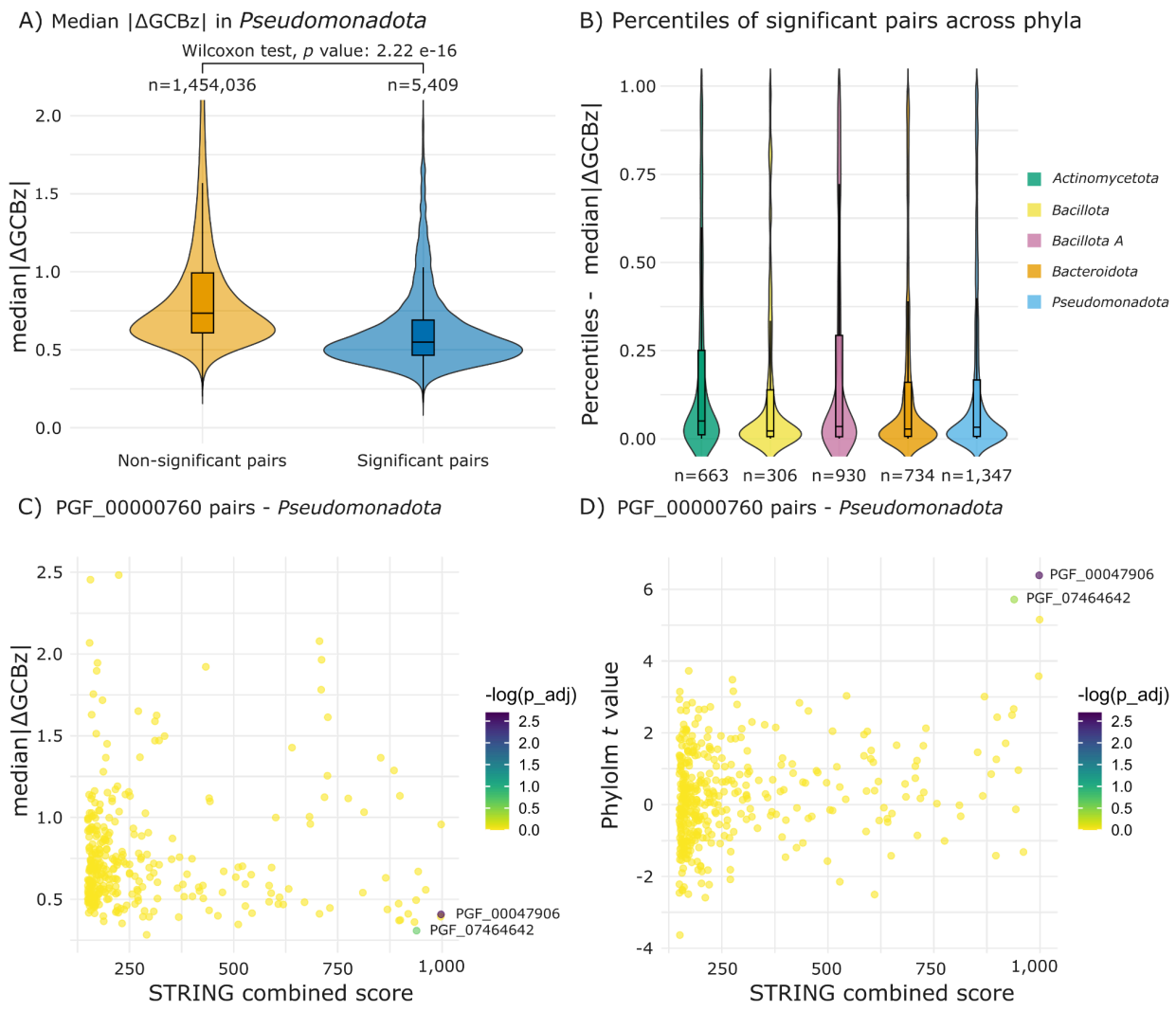


Fig. S10.

Low median differences in GCB z-scores (i.e., |ΔGCBz|) serves as an additional indicator of co-adaptation. (**A**) Median |ΔGCBz| values are significantly lower in significantly co-adapting pairs, indicating that they have more similar CUB across genomes than pairs that do not co-evolve. (**B**) For each PGFam containing significantly co-adapting pairs, the most significant pair typically ranks among the lowest percentiles of median |ΔGCBz| compared to all other pairs from the same PGFam. The distribution of these percentiles across the five main phyla is shown. n indicates the number of significant pairs analyzed. (**C-D**) Example of co-adaptation pairs associated with PGF_00000760 (α-glucoside ABC transporter, permease protein AglF). The two top-scoring pairs are PGF_00047906 (α-glucoside ABC transporter, permease protein AglG) and PGF_07464642 (maltodextrin glucosidase, EC 3.2.1.20), both showing low median |ΔGCBz| (**C**) and high Phylolm *t* values (**D**) compared to all other pairs. While PGF_07464642 is non-significant under Bonferroni correction (*p* adjusted <0.01), it becomes significant when using a more relaxed FDR correction (*p* adjusted <0.05) .

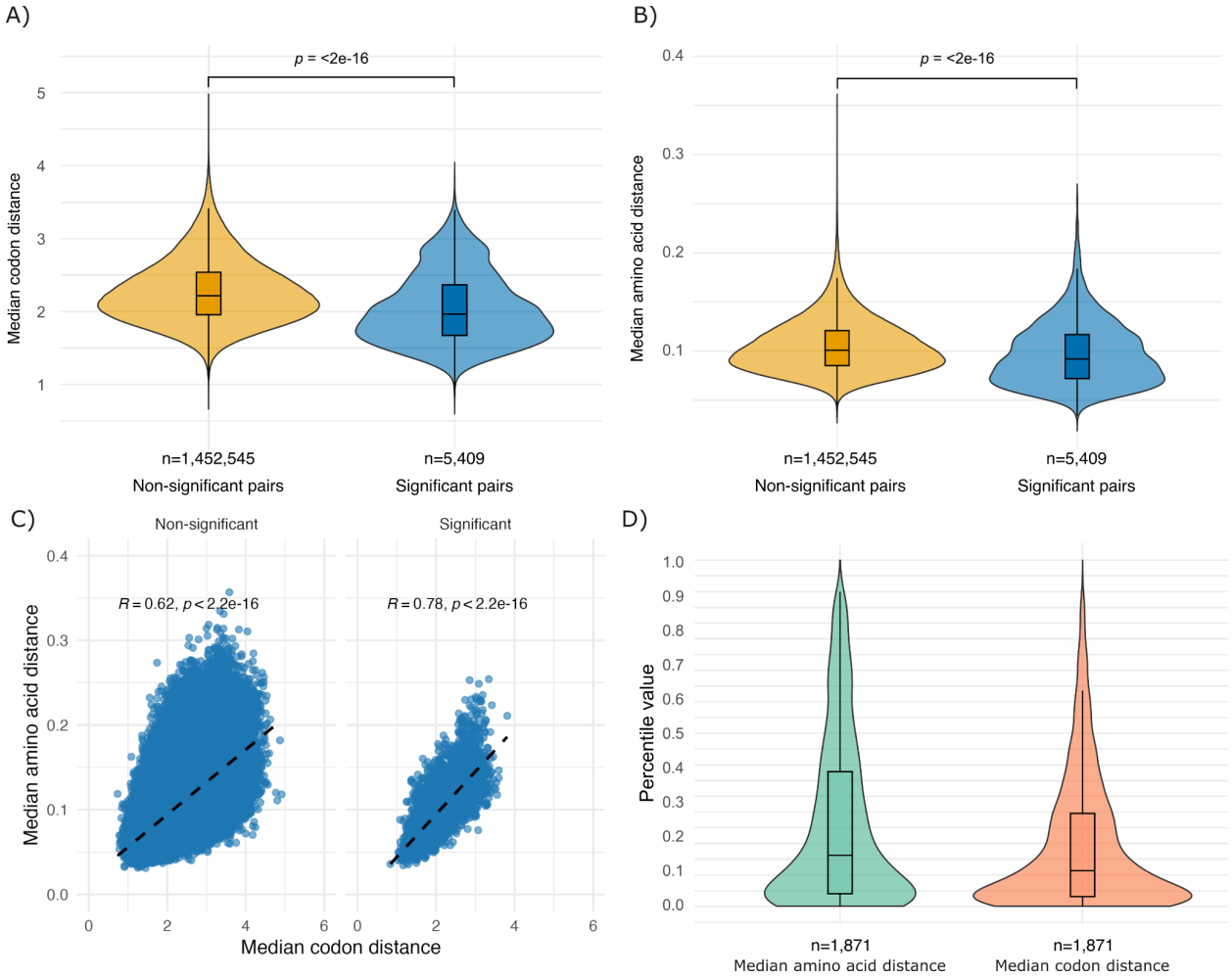


Fig. S11.

Codon and amino acid distances in co-adapting pairs. Median (**A**) codon and (**B**) amino acid distances are significantly lower in significant (co-adapting) compared to non-significant gene pairs. Distances were calculated using gene pairs corresponding to PGFams included in our models that presented at least one significant pair (n=1,826), from 300 randomly selected genomes affiliated with *Pseudomonadota*. Significant pairs were considered those with *p* adjusted (Bonferroni) values < 0.01. The Wilcoxon test was used to compare the distances between significant and non-significant gene pairs. N represents the number of paired distances in each group, derived from the 300 genomes. C) Spearman correlation plots between median codon distances and median amino acid distances for significant and non-significant pairs. Significant pairs show a higher correlation (R=0.77 *vs* 0.6) between codon and amino acid distances. D) For each PGFam containing significantly co-evolving pairs, the most significant pair typically ranks among the lowest percentiles of codon and amino acid distances compared to all other pairs from the same PGFam. The distribution of these percentiles within genomes affiliated with *Pseudomonadota* are shown. n indicates the number of significant pairs analyzed.

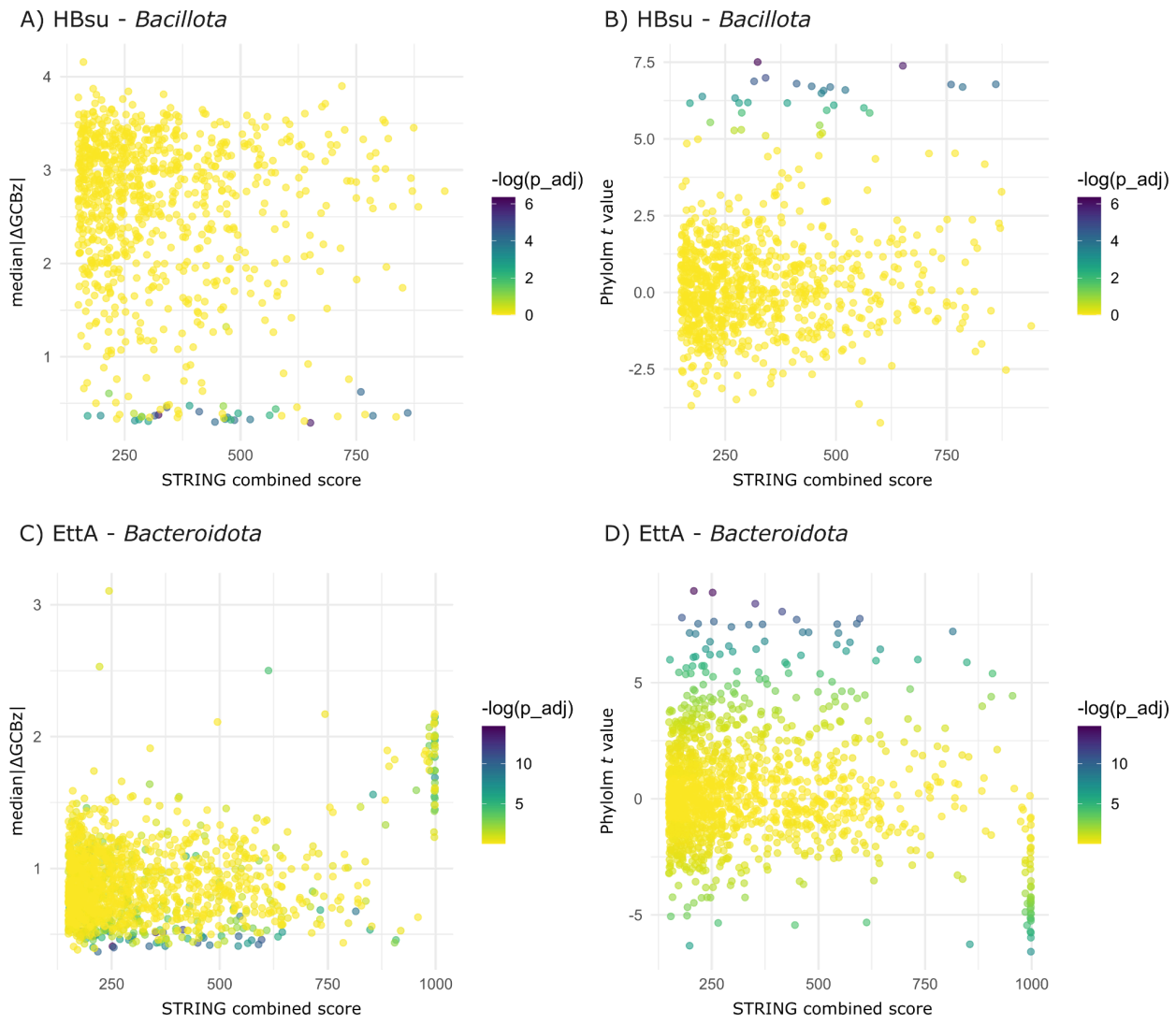


Fig. S12.

Median |ΔGCBz| and phylolm *t* values vs. STRING combined scores are shown for PGFam pairs corresponding to HBsu (PGF_01175502) in *Bacillota* and EttA (PGF_02059020) in *Bacteroidota*. Genes corresponding to PGFam pairs across genomes were selected and their |ΔGCBz| was calculated per genome *(15)*. Median values across all genomes were calculated. A lower median |ΔGCBz| indicates genes have similar GCB values. Phylolm *t* values were obtained from CUCA-PGGI models for each PGFam pair.

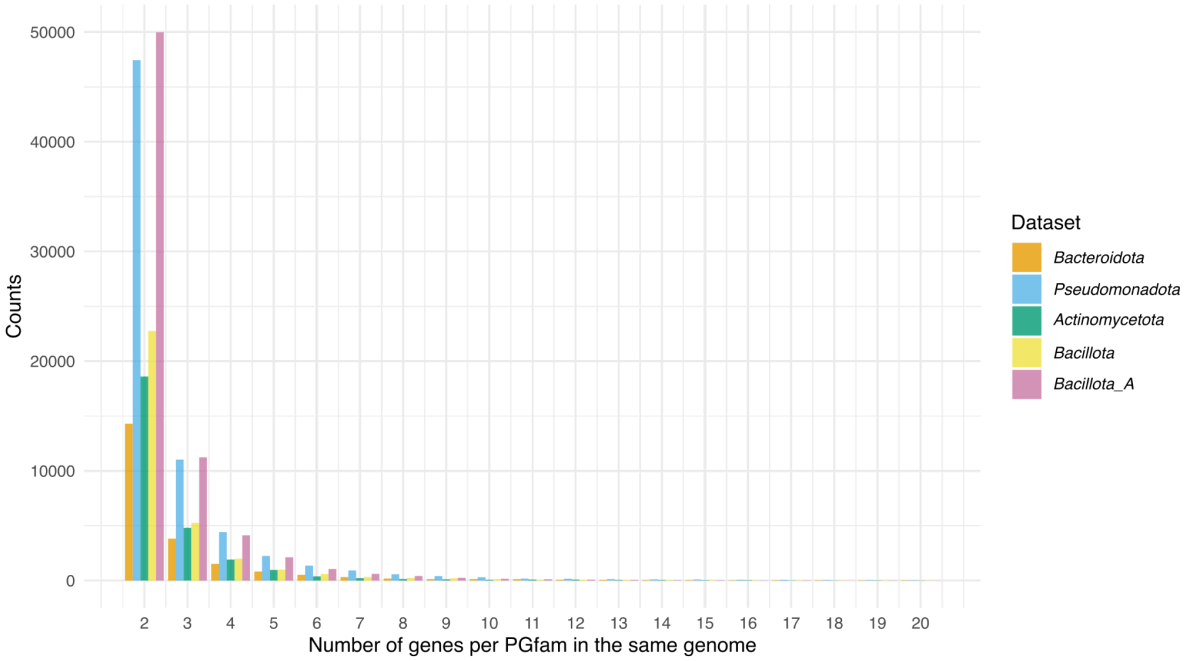


Fig. S13

Cases where PGFams have multiple representatives in the same genome. Significant changes in CUB across PGFam trees were selected where the same genome had multiple representatives in the same PGFam, with at least one representative being significantly different compared to the others, and then classified according to the number of representatives in the same genome. Counts represent the number of cases across PGFams and genomes. Note that one PGFam may include multiple genomes with > 1 representative, and each genome may include multiple PGFams with > 1 representative.