Large-scale investigation of orphan genes in the human gut microbiome elucidates their evolutionary origins

Nikolaos Vakirlis[1,2*] and Anne Kupczok[3*]

Affiliations:

[1]: Institute For Fundamental Biomedical Research, B.S.R.C. "Alexander Fleming", Vari, Greece. [2]: Institute for General Microbiology, Kiel University, Kiel, Germany.
[3]: Bioinformatics Group, Wageningen University, Wageningen, The Netherlands.

*: Correspondence to vakirlis@fleming.gr , anne.kupczok@wur.nl

Abstract

Orphan genes, i.e., genes that lack homologs outside a given species, are ubiquitous in all domains of life. In prokaryotes, orphans can originate (i) in the native genome via de novo evolution from non-genic regions or alternative frames of existing genes, or by rapid divergence and remodeling, or (ii) in a foreign genome, including viruses, followed by horizontal transfer. However, strong quantitative evidence supporting either scenario is lacking. Here we performed a systematic, large-scale analysis of orphan genes from human gut prokaryotes. After exhaustive filtering, we identified more than 6 million orphans in 4,644 species, which lack similarity to other prokaryotes and have no known functional domains. We find that a given species pangenome contains on average 2.6% orphan genes, which are mostly rare within a species. Overall, orphan genes use optimal codons less frequently, and their proteins are more disordered than those of conserved (i.e., non-orphan) genes. Importantly, the GC content of orphan genes in a given genome closely matches that of conserved ones. In contrast, the 5% of orphans that share similarity to known viral sequences have distinct characteristics that set them apart from the rest of the orphans, including lower GC content. By identifying the genomic region from which they evolved in closely related species, we provide evidence for native origination for a small subset of orphan genes and find that these orphans also differ in their properties from the remaining orphans. Finally, predicting orphan function by examining functional annotations in operon-like arrangements suggests that some orphan genes are membrane-related and involved in spore germination.

Our results support that orphans emerge due to multiple routes, challenging the notion that external elements such as phages and plasmids are the primary source of prokaryotic genetic novelty. Importantly, origination in the native genome might provide a constant influx of mostly transient genes into the cloud genome of prokaryotic pangenomes, where some orphans may prove adaptive, facilitating evolutionary innovation.

Introduction

Genomes of the same prokaryotic species can vary substantially in gene content, giving rise to huge pangenomes, i.e., all genes present in a species[1]. Pangenomes were shown to generally consist of three types of genes: a small set of "core" or nearly universal genes, a larger set of "shell" or moderately conserved genes, and a huge "cloud" of rare genes[2,3]. The cloud genes are particularly intriguing, as each newly sequenced genome is adding novel genes to the pangenome. This suggests that there is an ongoing inflow of genes into pangenomes, which has been mainly attributed to horizontal gene transfer (HGT). Bacteria are constantly under selection pressure to adapt to changing conditions or to colonize new niches. Cloud genes might thus provide a continuous inflow of novel genetic material, on which selection can act. Thus, although most cloud genes are transient, some might prove to be adaptive and persist in the population[4].

Furthermore, pangenomes even contain genes that have no sequence similarity outside the species. These so-called orphan genes (short: orphans) cannot be placed in any known gene family[5]. Orphan genes cannot be explained by the insufficient sequencing of homologous sequences, because the recent increase in sequencing data also resulted in an increase in orphan genes[6,7]. The abundance of orphans is intriguing since homologous families are assumed to share a common ancestor as well as functional similarity[8,9], whereas orphans might provide novel functions. The function of a large fraction of the microbial protein universe remains unknown[10]; thus, the study of orphan genes can potentially elucidate this unknown fraction of the protein universe. In *Escherichia coli*, orphan genes have been found to be narrowly distributed within the species[7]; thus, most orphans are cloud genes and might be important contributors to the inflow of novel genetic material to pangenomes. An understanding of orphan genes could thus provide insights into the evolutionary dynamics that shape pangenomes, contributing to an ongoing debate[11].

Orphan genes have been mainly studied in eukaryotes, where they are associated with organismal novelties and species-specific traits[12], and are generally thought to be important for adaptation[13,14]. Research in eukaryotes is starting to paint a coherent picture of the evolutionary origins of orphan genes [15–17]. We now know that they can also arise entirely "from scratch", through genomic sequences that are non-coding/non-genic, a process known as de novo gene emergence[18–20]. An Orphan gene can also result from extensive sequence divergence beyond recognition[21,22] as well as a combination of divergence, de novo emergence and reuse of parts of existing protein-coding genes in alternative reading frames[17,23].

While most of our understanding about orphan genes and the mechanisms that give rise to them come from eukaryotes, much less is known about them in prokaryotes. Frequently termed "ORFans", prokaryotic orphans share some of the same properties as those in

eukaryotes, such as shorter length and higher AT content[24]. The dynamic nature of prokaryotic genomes and the known pervasiveness of HGT create additional opportunities for the evolution of orphan genes. One hypothesis is that orphan genes have originally evolved in phages or other "selfish" genetic elements and are then transferred to bacteria[24] (Figure 1). While it has been 20 years since this hypothesis was proposed, evidence to support it is limited[25,26] and other studies failed to find equally convincing traces of the viral origin of orphans[27,28].

An alternative explanation to orphans being on foreign origin is that they are of native origin[24]. As for eukaryotic orphan genes, rapid coding sequence divergence, frameshifting mutations, rearrangements or a combination of the above could explain some of the genes lacking similarity, including entirely species-specific ones[7,29]. Alternatively, orphans could emerge de novo either from non-genic regions or from already coding ones but on an alternative reading frame, the latter known as overprinting[30,31]. As in eukaryotes, pervasive transcription and translation occurs in prokaryotes and may generate raw material that natural selection can shape into a functional protein[32,33]. Artificial random peptides have been shown to readily evolve rudimentary functionality after a few rounds of selection[34–36], highlighting the potential of functional de novo genes to evolve in natural prokaryotic populations as well. However, the constrained genome size in prokaryotes, as well as the limited non-genic space impose different constraints compared to eukaryotes[37]. Nevertheless, there is recent support for de novo gene origination in prokaryotes: An analysis of taxonomically restricted genes from the genus *Bacillus* could identify homologous, noncoding regions in a genome of another genus for almost one third of them, supporting abundant de novo origination[6]. Furthermore, a recent study of *Escherichia coli* showed that "remodeling events" – i.e., fusions of alternative reading-frames of segments of existing genes – are a potent source of entirely new protein sequences[38].

As orphans are defined here as species-specific genes, genome data on related species must be abundant to confidently assign a gene as orphan. Recent sequencing efforts, in particular using metagenomics, where the whole community of a sample is sequenced, now provide a dense coverage of the sequence space[39–41]. These advancements also highlighted that many genes fall into few large protein families and that there is a long tail of singleton families that have low diversity[41]. One particularly well sampled environment is the human gut, where pangenomes of the residing species have already been constructed[42]. Those pangenomes contain the information whether a gene family is found in one or multiple species, allowing direct orphan assignment. We build on this data – implementing further filtering steps to remove families with homology to other species – to systematically identify and study orphan genes in the human gut microbiome. An open question and our primary goal here is to understand whether orphans originate mostly natively or mostly externally (see Figure 1) by looking for patterns in large-scale comparisons. This is critical to delineate the evolutionary context and selection pressures under which orphan genes evolved and hence what they evolved to do.

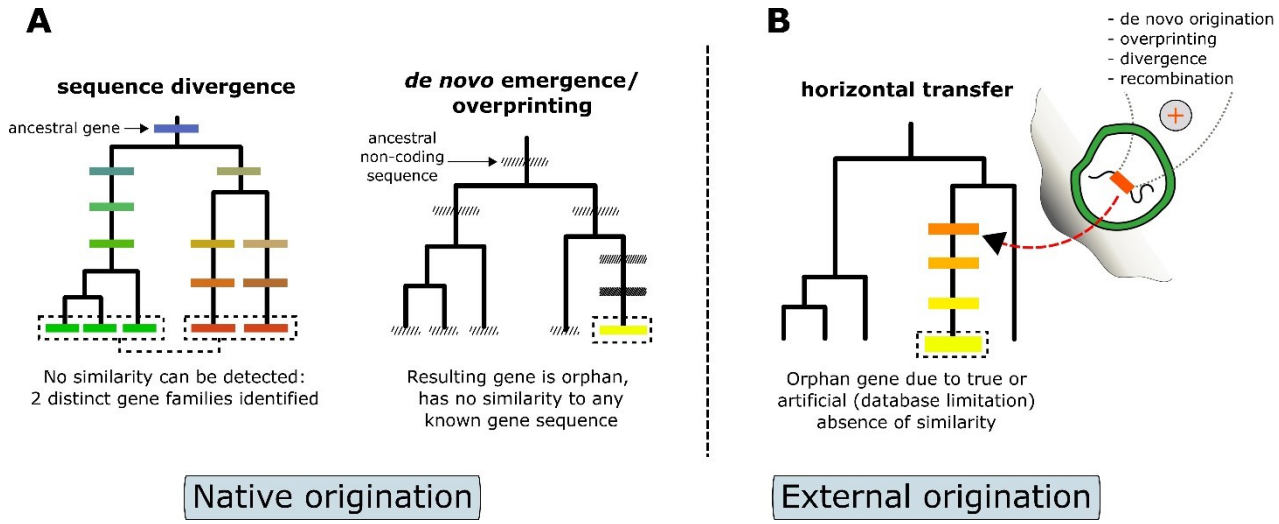


Figure 1: Two distinct evolutionary routes to prokaryotic orphan genes. **A**: Orphans can originate natively through sequence divergence potentially coupled with gene duplication or rearrangements, as well as de novo emergence from entirely non-coding (shown here) or coding regions in alternative reading-frames, the latter also known as overprinting. Darker shading represents transition from non-coding state to a coding one. **B:** Orphans can originate in phages, plasmids or integrative elements by various mechanisms such as de novo birth or overprinting, divergence and/or recombination, and then be imported into prokaryotic genomes. Either due to technical limitations or rapid divergence, similarity to the source of such externally originated genes might be lost which would result in them being perceived as orphans.

Results

Orphan genes are widespread in pangenomes of human gut prokaryotes

Our first goal was to establish a conservatively defined catalogue of prokaryotic species-specific genes. We chose the human gut environment, an extensively studied niche harboring thousands of known prokaryotic species. Recently, a comprehensive catalogue of genomes from prokaryotes of the human gut was released, where the genomes were annotated and analyzed with a common pipeline[42]. In that study, the human gut prokaryotic protein catalogue was clustered into homologous families at different cut-offs of sequence similarity. We used the most loosely defined protein families as the starting point of our analysis (version 1.0 UHGP-50, built using a 50% identity cut-off) to identify homology with a high sensitivity. By applying a stringent pipeline of sequence-similarity searches, we filtered the initial dataset consisting of more than 600 million proteins grouped in 4.7 million families, down to 3.3 million species-specific proteins (i.e., orphans), grouped into 630,000 families (see Figure 2A and Methods).

152 We find an average of 4.7 orphans per genome, and an average of 135 orphan genes per
153 species. The latter represents an average of 2.6% of a given species' pangenome, that is, the
154 entire gene repertoire of the species (Figure 2B). The number of orphans in a given
155 pangenome correlates strongly with the number of genomes available for that species
156 (Spearman's Rho=0.61, $P<2.2*10^{-16}$; Figure 2C), as well as with pangenome size (Rho=0.66,
157 $P<2.2*10^{-16}$; Supp. Figure 1A), suggesting that each genome introduces novel orphan genes
158 into the pangenome. Indeed, we find that orphan genes are rare within their respective
159 species. Most of them are shared by at most 10% of genomes, i.e., belonging to the cloud
160 genome of prokaryotic pangenomes (Figure 2D). Orphans are strongly overrepresented
161 among accessory genes compared to conserved, i.e., non-orphan, genes (Supp. Figure 2).
162 While this has been reported before in *E. coli*[7], our analysis demonstrates that this is a general
163 feature of the human gut microbiome and potentially of prokaryotic orphans in general. So,
164 while species-specific genes are indeed widespread and present in 94% of all gut species (and
165 in all species with at least 10 available genomes), they represent only a small fraction of the
166 gene repertoire and they tend to be rare within pangenomes.

167 We observe variability across taxonomic groups: some bacterial taxonomic classes have
168 markedly higher percentages of orphan genes than others, with Coriobacteriia representing
169 the lower (avg. 0.7%) and Alphaproteobacteria the higher (avg. 5.4%) extremes (see also Supp.
170 Figure 3). Species with a high proportion of metagenome assembled genomes (MAGs) have
171 comparable percentages of orphans as species with many isolate genomes (Supp. Figure 1B),
172 and species with lower quality assemblies show no trend towards more orphans (Rho=-0.04,
173 P=0.0024; Supp. Figure 1C). This suggests that orphan numbers are not driven by assembly
174 artefacts.

175

176

177

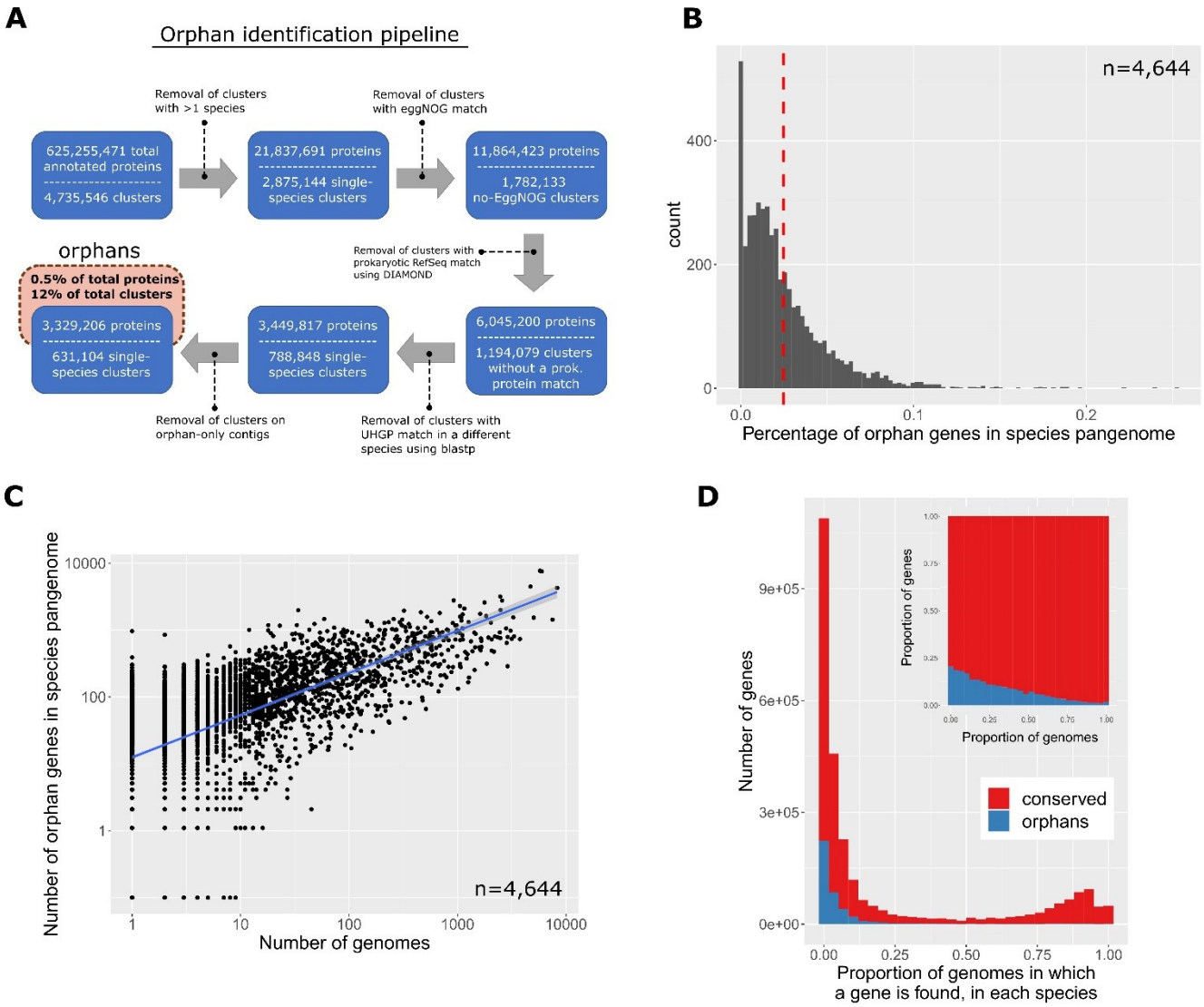


Figure 2: Large scale-identification of orphans in the human gut microbiome. **A:** The main steps of the computational workflow of similarity searches. **B**: Distribution of the percentage of orphan genes in a given species' pangenome. Red dashed line denotes distribution mean. **C**: Number of orphan genes strongly correlates with the number of genomes available for a given species. **D**: Distribution of percentage of genomes where a gene is present, for orphans and conserved genes. Only species with at least 10 genomes are included (n=1,369). The inset plot shows the same data but as relative percentages in each bin.

Orphan genes and their proteins have some distinct and some common properties compared to conserved genes

In both prokaryotes and eukaryotes, orphan genes are known to exhibit certain characteristics, such as shorter length, which set them apart from conserved genes[15,28,43]. These gene and protein properties might provide clues about the evolutionary origins of orphans. To compare orphans to conserved genes, we next compare properties of genes and proteins across species.

As expected, orphan genes across every taxonomic class are significantly shorter than conserved genes, at about a third of their average length (Figure 3A). There is only a weak correlation between average orphan and conserved length per species (Rho=0.14, $P<2.2*10^{-16}$).

Average GC content of orphan genes in a given species is strikingly similar (avg. 0.45 vs. 0.46) to that of conserved genes (Figure 3B), exhibiting an almost one-to-one correlation (Rho=0.96, $P<2.2*10^{-16}$). The trend of similar GC is also seen when comparing orphan and conserved genes within individual species (Supp. Figure 4A). The same is true for GC in the 3rd synonymous position of codons (GC3s; see Supp. Figure 4B) and for CpG content (Supp. Figure 5). This observation fits the native origination scenario, in which orphan GC content simply reflects the genome's GC content, but it is also consistent with external transfer followed by rapid adaptation of the incoming gene's GC content to that of the host genome[44].

In certain scenarios for orphan origination, including de novo evolution and recent transfer from external origin, orphans are expected to show suboptimal codon usage relative to the rest of the genome. We thus calculated the Codon Adaptation Index (CAI) for the representative genome of each species using the *CAIJava* tool, (which does not require a reference gene set) and compared CAI of orphan and conserved genes (see Methods). Across taxonomic classes, orphan genes have consistently lower CAI values than conserved genes (Figure 3C, all Wilcoxon test P-value < 10^{-4}). This is also true when controlling for the length differential between orphans and conserved genes by subsampling conserved genes to obtain a set of comparable mean length to that of orphan genes (Supp. Figure 6; Methods). For *E. coli*, we also calculated CAI based on codons used in reference sets of highly expressed genes and found similar results (Supp. Figure 7; Methods).

At the protein level too, orphans exhibit distinguishing characteristics (Figure 3D). Compared to conserved genes, they are, on average, richer in intrinsically disordered regions, a smaller percentage of their sequence is found in helical regions, a smaller percentage of them contains transmembrane (TM) domains, they are more aggregation prone, have higher biosynthetic cost, and a higher isoelectric point (the latter was also observed for de novo genes in yeast and fly[45,46]). All these differences hold when considering each taxonomic class individually (Supp. Figure 8). However, when controlling for the length differential between orphans and conserved genes, intrinsic disorder and TM domains are no longer significantly different on average (Supp. Figure 9). Note that novel genes in membrane-related roles have previously been suggested for eukaryotes such as budding yeast[47,48]and for prokaryotes[49]. Here we find that, even though orphans have similar nucleotide composition to conserved genes on average (and even more so when controlling for length differences) they differ in key properties, such as their length, codon usage, and some protein structural features.

Since we observed before that most orphans are rare, cloud genes (Figure 2D), we next asked if cloud orphans differ from non-cloud orphans. Indeed, we find that some properties differ, e.g., cloud orphans have lower CAI, higher GC, and fewer helices than remaining orphans (Figure 3E). Surprisingly, when comparing cloud genes to remaining genes, the observed trend for GC differs between orphans and conserved genes: whereas conserved cloud genes have a lower GC than remaining conserved genes, orphan cloud genes have higher GC than the remaining cloud genes. Furthermore, the GC of orphan cloud genes is higher than the GC of conserved cloud genes. These trends also hold when analyzing taxonomic classes separately (Supp. Figure 10) and when zooming in individual species (Supp. Figure 4A). Although cloud genes are still quite enigmatic, they could be explained by a continuous influx of genes by

horizontal gene transfer. However, HGT from other prokaryotes would result in conserved cloud genes. Due to the striking difference in the trends for GC, we believe that orphan cloud genes have different evolutionary origins than conserved cloud genes. Interestingly, orphan non-cloud genes have low GC. Potential explanations are that they originate de novo from noncoding regions that typically have lower GC than coding regions or due to rapid divergence, which could lead to the accumulation of A and T[50]. In a previous analysis of Salmonella, it has been found that genes with conserved gene order in *E. coli* have higher GC than the remaining genes, where the genes without homologs have even lower GC than the ones with homologs in another genomic context[51]. This trend has been attributed to rapid divergence and transcriptional silencing by H-NS, which might also apply to the orphan genes with low GC.

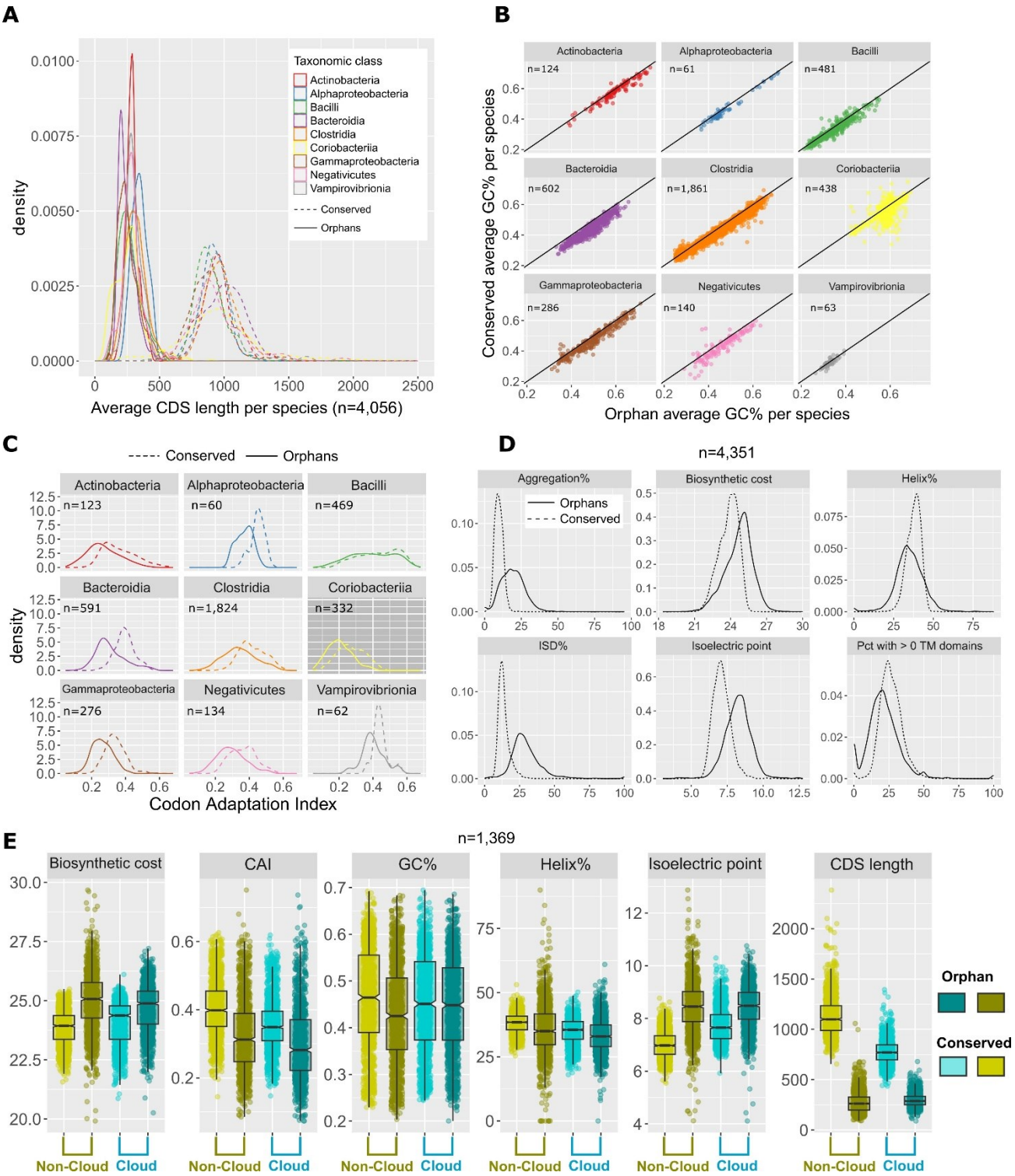


Figure 3: Properties of orphans compared to conserved genes. **A:** Density plots of the average length of orphan and conserved CDSs in each species, grouped by each of the 9 best represented taxonomic classes of our dataset. Only classes with at least 50 species are included. **B**: Average GC% of orphan and conserved genes in each species, grouped by taxonomic class (same classes as in A). **C**: Same comparison as in A (average per species), but with Codon Adaptation Index (CAI) values as predicted by the *CAIJava* tool. All Wilcoxon test

265 P-values < 10^{-4}. **D**: Comparison of proteins encoded by conserved and orphan genes, in terms
266 of average percentage of protein predicted to self-aggregate, biosynthetic cost, percentage of
267 protein predicted to be helical, percentage of protein predicted to be disordered (ISD),
268 isoelectric point of a protein, and percentage of proteins with at least 1 transmembrane
269 domain. All Wilcoxon test P-values < $2.2*10^{-16}$. Distributions of average values per species are
270 shown. **E**: Comparison of some of the properties found in panels A-D in 1,369 species with at
271 least 10 genomes where data orphan and conserved gene categories are split in cloud (present
272 in 10% of species or less) and non-cloud. Each point represents an average value per species.

273

274

275 Orphans of likely phage origin have distinct characteristics

276

277 Orphans are defined here as species-specific genes with respect to other prokaryotes,
278 however it is possible that they are the result of recent horizontal gene transfer from
279 bacteriophages. To identify such transferred genes, we conducted similarity searches against
280 two recently published phage protein databases, MGV[52] and GPD[53], that are partially
281 overlapping but mostly complementary in terms of protein sequence content.

282 Orphan genes are strongly depleted in similarity matches to phage proteins relative to
283 conserved genes, ranging from 3 to 6-fold using different parameters and cut-offs (Figure 4A;
284 Methods). At low stringency, 5.1% of orphans have similarity to viral sequences, compared to
285 63% of conserved genes. Orphans are also depleted, when separately searching against either
286 MGV or GPD (Supp. Figure 11A) and when excluding short sequences (<800nt) for which
287 similarity searches often fail (Supp. Figure 11B). Interestingly, binning conserved genes
288 according to the number of species in which they are found (a proxy of their evolutionary age)
289 reveals a pattern of increasing proportion of genes with phage similarity (Figure 4B),
290 regardless of the stringency criteria applied (Supp. Figure 11C). This pattern is inconsistent
291 with a scenario under which most orphan genes arrive through HGT from bacteriophages,
292 which we would expect to either produce the inverse trend, if only a fraction of HGT is
293 ultimately retained, or no trend at all in the case of a constant rate of transfer and retention.

294 As both GPD and MGV databases have been assembled with the use of automated annotation
295 pipelines, it is possible that they may contain bacterial contigs that are falsely classified as
296 coming from phages. To control for such a bias, we confirmed our findings by searching a third
297 phage genome database[54] which is smaller but manually curated and was built from samples
298 taken from infant guts. We observed similar patterns as for the larger databases confirming
299 that the proportion of orphans (and more generally, less conserved genes) with similarity to
300 phage proteins is considerably lower than that of widely conserved genes (Supp. Figures 11D
301 and 11E). The overall proportion of matches in all gene groups is lower compared to GPD and
302 MGV, but that is expected given the difference in size (10K phage OTUs vs. 190K and 142K in
303 MGV and GPD respectively).

304 The trend of increasing percentage of phage matches with increasing conservation across
305 species prompted us to investigate further. To dissect the evolutionary dynamics, we plotted,
306 across age categories, the query coverage and percent identity of the best phage match per
307 protein as defined by E-value, without filtering any high-identity hits (Figure 4C). This revealed
308 that matches to orphan genes are very close to 100% identity and coverage with both

309 averages decreasing with age. Yet, while the coverage distribution shows a clear downward
310 trend, the identity distribution becomes bimodal with age, with one peak staying near 100%
311 and another one initially appearing at around 35%, then increasing. Furthermore, while
312 identity and coverage are strongly correlated for viral hits of orphans, this correlation
313 gradually decreases with age (Figure 4D). One explanation for these trends is that a subset of
314 genes has only recently been transferred between phages and bacteria, while another subset
315 consists of older transfers that have undergone divergence, truncation, or both.

316 We previously established that orphans mirror conserved genes' averages in crucial aspects
317 such as nucleotide composition. Given the evidence that a small percentage of orphans
318 appears to have been recently transferred from phages, we might expect that they exhibit
319 distinguishing signatures at the level of their nucleotide composition. Indeed, we find that
320 overall, orphans with similarity to phage proteins have markedly lower GC content than those
321 without (Figure 4E), congruent with what is known for phage GC% relative to their hosts[55].
322 Most strikingly, the effect of this difference is practically negligible when looking at conserved
323 genes, albeit staying statistically significant due to the large sample size. The findings also hold
324 when only genes >800nt are considered (Supp. Figure 11F). We thus conclude that there is a
325 low proportion of orphans with hits to phages that appear distinct from the remaining
326 orphans.

327 Additionally, we computationally detected prophage regions in all genomes using Phigaro[56].
328 We find that a very low proportion of orphan genes reside within prophages
329 (13,174/3,329,206; 0.3957%), which is slightly higher than the overall proportion of prophage
330 genes among all genes (2,029,792/625,255,471; 0.3246%). Thus, although phage origin might
331 play a role, this can only explain a small fraction of orphans.

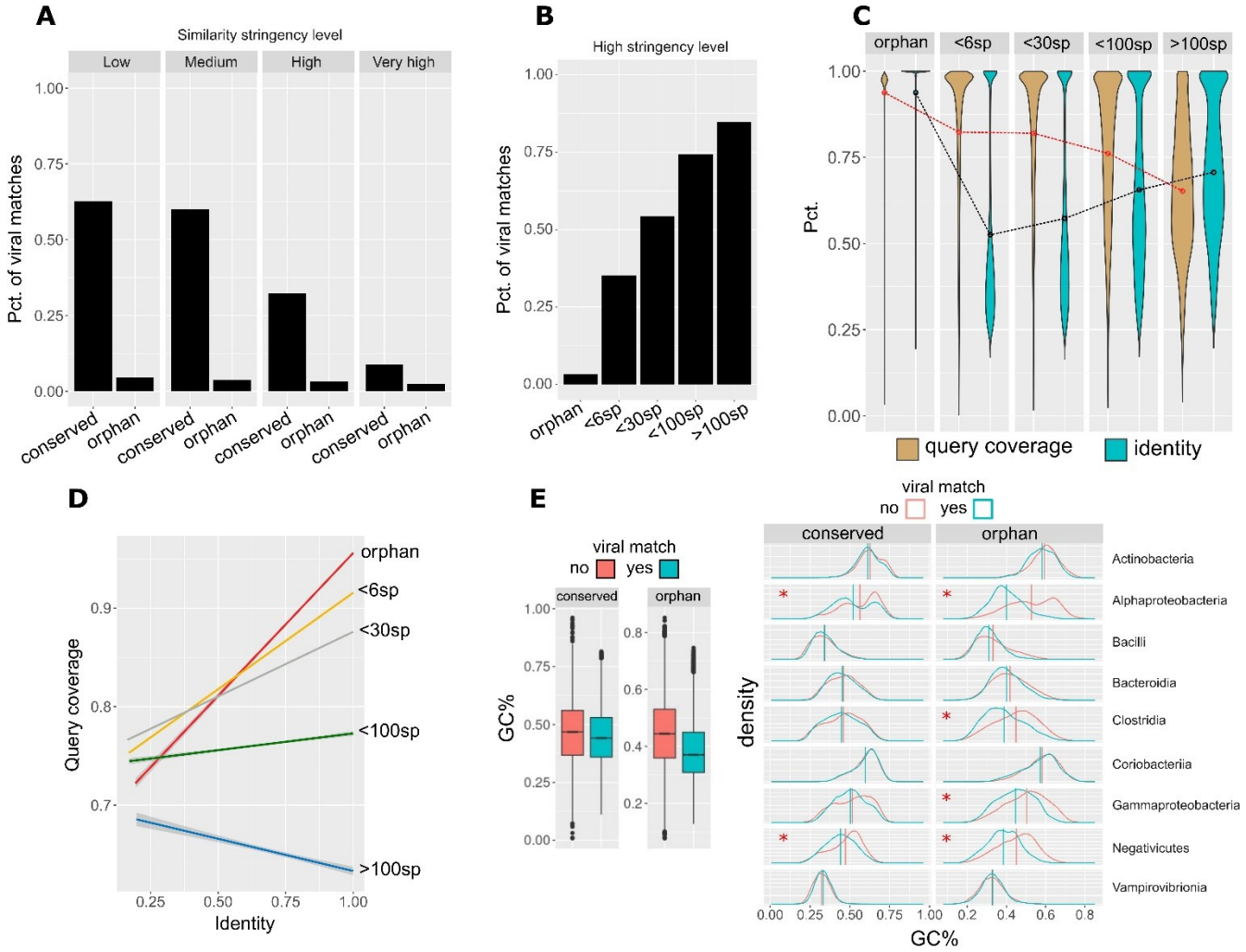


Figure 4: Properties of genes with viral matches. **A**: Percentages of conserved and orphan proteins with statistically significant similarity matches to viral proteins at four different significance thresholds. **B**: Percentage of proteins with statistically significant similarity matches to viral proteins, binned according to the number of prokaryotic species in which a homologue can been found (size of the prokaryotic protein family). **C**: Query (prokaryotic protein) coverage and percent sequence identity of top statistically significant matches (all hits with E-value < 10^{-5} are included), binned according to the number of species in which a homologue can be found. Lines connect distribution means. **D**: Correlation between query coverage and identity in sequence matches involving genes belonging to different sizes of protein families, same data as in C. **E**: Distributions of GC% of genes with and without a statistically significant viral match (high stringency), among all orphan and conserved genes (boxplot) and among the nine best represented taxonomic classes (density plots). A red asterisk denotes a non-negligible effect size in difference of means calculated by Cliff's Delta (Delta estimate > 0.15). Note that in Alphaproteobacteria and Negativicutes the difference exists also in conserved, but the effect size there is much weaker compared to orphans (-0.21 vs. -0.62 and -0.2 vs. -0.42).

A subset of orphans shows evidence of native origination

For some orphans, it might be possible to trace the origin within their native genome, either to existing genes or to non-coding regions. We first searched for orphan genes that might have originated via utilization of an alternative frame following duplication or horizontal gene transfer[57] from closely related genomes. To this end, we performed similarity searches of six-frame translations of all orphan coding sequences against six-frame translations of all coding-sequences of their native genome. We found that only a small percentage, 1.34% of orphans (8,477 representative orphans), have at least one significant match to a conserved gene (i.e. non-orphan) excluding gene pairs with overlap. 75% of matches are on the reverse strand of the conserved gene, suggesting a more frequent exploration of the reverse complement sequence space rather than the alternative reading frames of the same strand. Out of all orphans matching conserved genes, 64% (5,397) match a single conserved gene, 26% (2,238) match two conserved genes and the rest (10%) match three or more. Orphans matching more than one conserved gene correspond to chimeras similar to what was found for *E. coli*[58].

Approximately 10% of orphans with hits (892) match conserved sequences on the forward strand and the same frame as the conserved gene. Thus, these should be considered as false positives, that is, non-orphans. The remaining 7,585 orphans with matches are strong candidates for having evolved out of alternative reading frames of preexisting native proteins. For orphans with matches on the same contig (n=1,515, 20%), the genomic distances between the two matching genes show a unimodal distribution with a mean of 131kb (Figure 5B).

We next turned to evidence for native origination that can be found in closely related species. A robust approach for the detection of de novo gene emergence is the identification of the orthologous genomic region in outgroup genomes. By examining such regions, it is possible to establish that a gene has evolved out of previously non-coding sequences, or out of an already coding locus, either coexisting with or replacing an ancestral gene. To detect both such cases, we identified a given orphan gene's orthologous region in outgroup genomes of the same species and those of its closest relative species using strict orthology criteria (see Methods) and checked for the presence of missed orthologous open reading frames (ORFs) that might have eluded the automatic annotation pipeline. We thus identified 1,075 orphans that can be considered candidates for de novo emergence from non-coding regions (Supp. Table 1).

Closely related species are more likely to share genomic synteny and hence to satisfy our conservative criteria. The de novo candidates originate from 718 different species, and these have shorter terminal branches in the tree compared to the species of the remaining orphans (Figure 5C; means of 0.03 and 0.071 substitutions per site; Wilcoxon test P-value $< 2.2*10^{-16}$; Cliff's Delta 0.28). Thus, they represent the evolutionarily younger end of the spectrum of all de novo gene candidates, as expected. Comparisons of gene and protein properties support this view (Figure 5C). De novo candidates are overall shorter (Delta=0.28; $P<2.2*10^{-16}$), have lower CAI (Delta=0.18; $P=3.64*10^{-11}$) and GC% (Delta=0.1; $P=8.5*10^{-9}$), have slightly higher isoelectric point (Delta=0.15; $P=10^{-13}$) and biosynthetic cost (Delta=0.13; $P<2.2*10^{-16}$). Since mutation is biased towards AT, most bacteria have lower GC in noncoding regions[50], which

can explain the lower GC in de novo candidates compared to remaining orphans. Additionally, de novo candidates are depleted in matches to viral proteins compared to the rest of orphans (Fisher's test odds-ratio: 40.9, P-value = $3.5*10^{-16}$). Two examples of alignments of de novo candidates to their orthologous regions in outgroup genomes can be found in Figure 5D and E. The translated sequences can be found in Supp. Figure 12. In both these cases, based on the multiple stop codons present in the orthologous sequences, we can cautiously infer that the ancestral sequence did not contain an ORF and thus lacked coding potential.

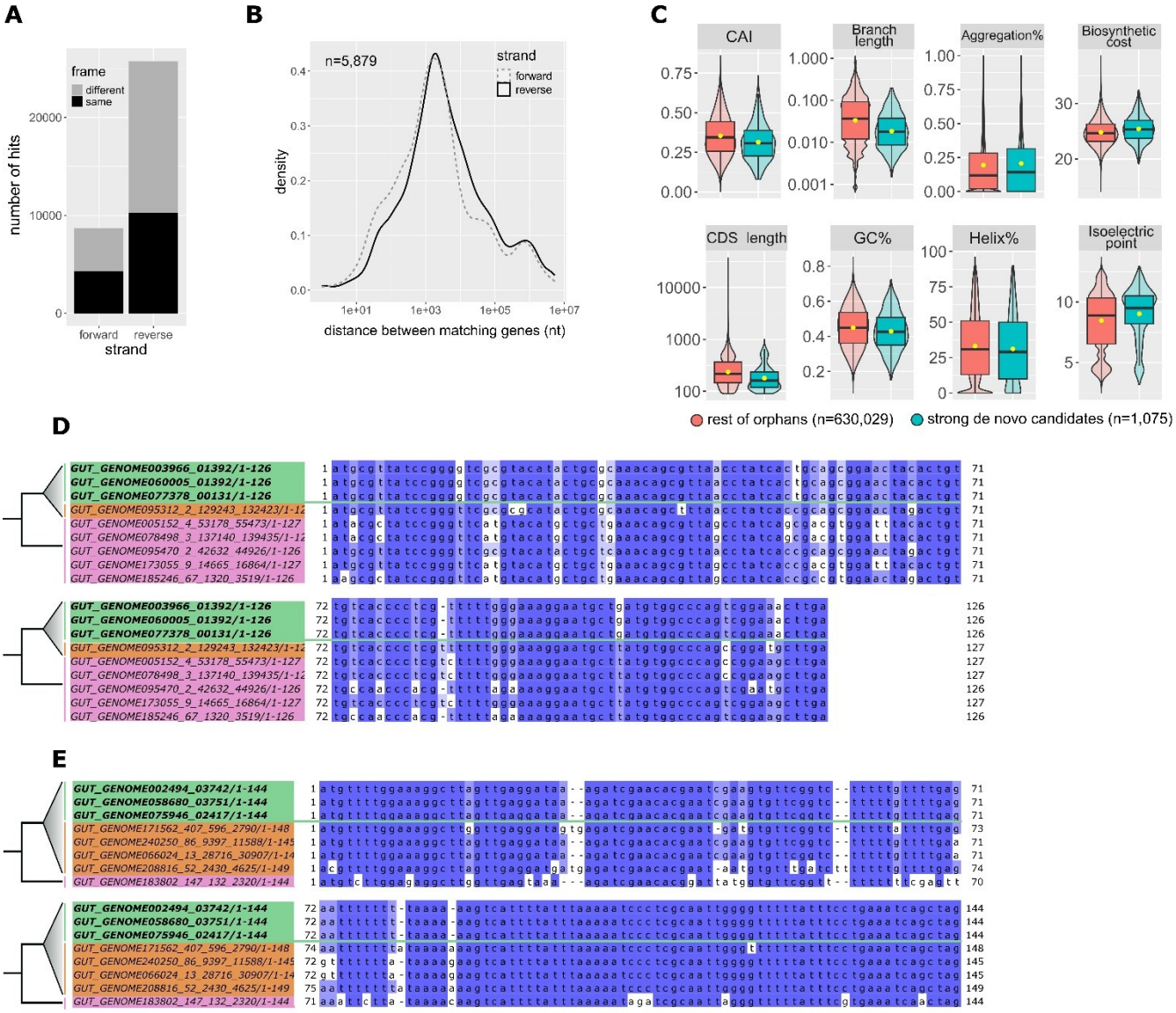


Figure 5. Evidence for native origination of orphan genes. **A**: distribution of frame and strand of similarity matches between orphan genes and conserved genes within each genome. **B**: Distribution of genomic distances between matching orphan and conserved genes, when these are found on the same contig. **C**: Comparisons of de novo candidates to remaining orphans: Distributions of Codon Adaptation Index (CAI), terminal species-level branch length, percentage of protein sequence predicted to be aggregation-prone, biosynthetic cost, CDS length, GC%, protein percentage estimated to fold into a helix, and isoelectric

410 point. Details of statistical comparisons can be found in the main text. Helix% comparison
411 Wilcoxon test P-value = 0.0064. Aggregation% P-value = 0.26. Branch length and CDS length
412 are shown on log scale for visibility. Outliers are not shown as points but are represented by
413 the violin plot range. **D**: Alignment of a de novo candidate gene from an unnamed species of
414 genus *Pauljensenia* (green) to its orthologous regions in genomes of the same species that
415 do not have annotated homologues (orange) and genomes of its closest outgroup species
416 (pink). Identical sequences have been removed from the alignment. **E**: Same as D, but for
417 another candidate from *Enterococcus raffinosus*. Alignments were generated with MAFFT[59]
418 and visualized with JalView[60].

419

420

421 Operon-like arrangements provide functional hints for some orphans

422

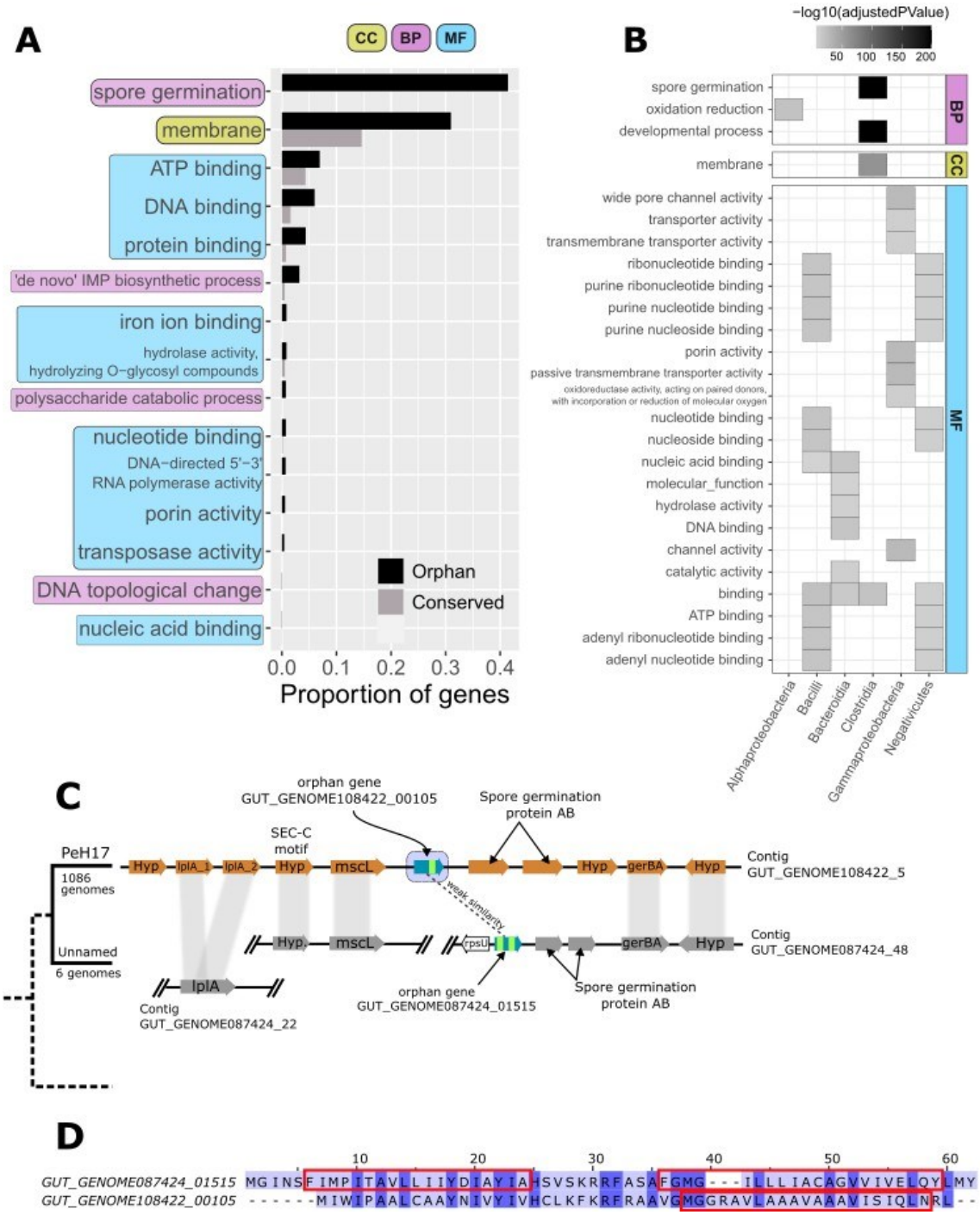
423 Assigning a function to orphan genes is notoriously difficult. For instance, we searched a
424 database from a previous study that established mutant phenotypes for thousands of
425 prokaryotic genes coming from 46 strains from 36 different species[61] including many from our
426 dataset (see Methods), but found no phenotypes for any of our orphan genes.

427 Even if no experimentally derived functional information has been determined for an orphan
428 gene, its genomic context can still offer some clues as to what its role might be in the cell[41,62].
429 We thus examined the neighbors of orphan genes in all genomes harboring the orphan gene
430 to identify cases of operon-like arrangements where one functional term is present in multiple
431 neighbors (see Methods). We identified 1,905 orphans in operon-like arrangements,
432 belonging to 405 distinct protein families. For these, at least 3 out of the 6 closest neighbors
433 (i.e. 3 downstream and 3 upstream) have a common Gene Ontology (GO) term associated. We
434 then looked for enrichment of specific GO terms within the orphans against the background
435 of all genes in operon-like arrangements (n= 4,362,138). The most overrepresented specific
436 terms by gene counts are shown in Figure 6A. The results of the full GO term enrichment
437 analysis can be found in Supp. Table 2.

438 The term "membrane" was the only cellular component that was found in at least 3 neighbors
439 of orphans and it was one of the top enriched terms overall (GO:0016020; Figure 6A, Supp.
440 Table 2). This is intriguing given previous findings linking TM domains to novel genes[47,49]. A
441 closer look showed that the membrane enrichment was due almost exclusively to orphans
442 from the class Clostridia. Thus, to account for taxonomic specificities, we repeated the analysis
443 separately for each of the 9 best represented taxonomic classes and found limited overlap
444 between the GO terms enriched within each class (Figure 6B; Supp. Table 3).

445 At the level of biological processes, one of the strongest enrichments was "spore germination"
446 and this enrichment too is specific to Clostridia (Figure 6B). Of the seven Clostridia orphans
447 related to this term, four had at least one TM domain (57%) including three that are also
448 annotated as "membrane", which is expected given that the cell membrane is central to the
449 process of germination. We examined in detail one large family with "spore germination"
450 annotation (see Methods). A closer look at the wider genomic region of this family showed
451 clear evidence of syntenic and functional conservation with its closest neighboring species
452 (unnamed species, see Figure 6C). Furthermore, another orphan gene, also encoding TM
453 domains, is found at the predicted orthologous locus in the closest species. Alignment of these

two orphan protein sequences of similar size reveals traces of similarity and an almost complete overlap of the TM domains in each sequence (Figure 6D). This intriguing finding suggests that these two genes are homologous and likely functionally similar but have diverged beyond detectable similarity, as observed before[21,22]. When their common ancestor might have first evolved is unclear. Overall, this analysis suggests that orphans might evolve different functions in different lineages and point towards spore germination as a particularly relevant process in Clostridia.



463

464 **Figure 6.** Genomic context offers functional clues of orphans. **A**: Proportion of orphan and
465 conserved genes annotated with a given GO term based on genomic context. These GO
466 terms are those with the highest proportion of orphan genes and are all significantly
467 overrepresented with an FDR adjusted P-value < 10^{-5}. The complete results of the GO
468 enrichment analysis can be found in Supp. Table 2. CC: Cellular Component. BP: Biological
469 Process. MF: Molecular Function. **B**: Significantly overrepresented GO terms among orphans,
470 when analyzing separately genomes of species belonging to each of the nine best
471 represented taxonomic classes. For visualization purposes, only GO terms significantly
472 overrepresented with an FDR adjusted P-value < 10^{-5} and found in at least 10% of orphans in
473 a class are shown. The complete lists, including for two of the nine classes (Actinobacteria,
474 Vampirovibrionia) for which no enrichment was strong enough to be included in this figure,
475 can be found in Supp. Table 3. (no enrichment was found for Coriobacteriia). **C**: Genomic
476 region around Clostridia species PeH17 orphan gene GUT_GENOME108422_00105 and
477 syntenic conservation in its sister species (Unnamed, representative genome
478 GENOME056772). Gene names are included in each gene shape. Vertical grey bands connect
479 genes found within the same homologous family. Blue arrows denote orphan genes, and
480 green bands within them represent TM domain encoding segments. lplA: lipoate-protein
481 ligase A. mscL: Large Conductance Mechanosensitive Ion Channel protein. gerBA: spore
482 germination protein B1. Hyp: hypothetical protein. No annotation was present in the GFF
483 files for the two proteins to the right of the orphan genes, but all four had the same match in
484 InterPro (Spore germination protein AB; IPR004761). **D:** Pairwise alignment generated by
485 MAFFT, of the protein sequences of the two orphan genes highlighted in C. Rectangles mark
486 the positions of predicted TM domains in the two protein sequences.

487

488

489 Discussion

490

491 Orphan genes are ubiquitous. Across the tree of life, species-specific genes may be seen as
492 one end of a taxonomic gradient, with universally conserved genes at the other end and genes
493 restricted to smaller taxonomic groups as intermediates[15]. Some recent large-scale studies
494 have focused on novel prokaryotic gene families present in more than a single species and
495 have firmly demonstrated their functional relevance and hinted towards crucial habitat-
496 specific roles[41,49,63]. The question of the evolutionary origin of novel prokaryotic families
497 however was mostly left unexplored. This question is best approached by focusing on orphan
498 genes, which are species-specific and thus represent the most-recently evolved form of
499 genetic novelty.

500 Our results show that orphans emerge due to multiple routes. We find that a low proportion
501 of orphans has matches in viral proteins. Thanks to the extensive, high-quality sampling of
502 both the prokaryotic and the viral diversity of the human gut, we can be relatively confident
503 that our findings are not significantly biased by limitations of the dataset. Additionally,
504 orphans with viral matches have a distinct composition compared to the remaining orphans
505 (Figure 4E), which is inconsistent with the external origination scenario for most orphans.

506 Conversely, there is a low proportion of genes, where we can recover evidence for de novo
507 origination *from non-coding sequences* (not overprinting). These orphans also differ from the
508 remaining orphans by being shorter, having lower GC%, and lower CAI (Figure 5C), although
509 these differences likely also reflect the fact that we can only detect the very youngest of the
510 de novo originated ones. Nonetheless, these results are also inconsistent with de novo
511 origination from non-coding sequences as a dominant source behind orphan genes. This
512 leaves the native processes of sequence divergence, overprinting, and remodeling as the next
513 best candidates to explain orphan origination.

514 Our analyses go significantly further and deeper than previous efforts to understand the origin
515 of orphan and novel genes in prokaryotes. Our results are consistent with some previous
516 studies where low fractions of viral matches have been found[27], but there are noteworthy
517 cases of discrepancy. Some of the earliest inquiries into the origin of bacterial orphans came
518 from Daubin and Ochman[24,25]. Based on compositional similarity between orphans in *E. coli*
519 and bacteriophage sequences known at the time, they proposed acquisition from
520 bacteriophages as a reasonable explanation. Our analysis of some of the same features (GC
521 content, CpG content), across thousands of species, for which data was not available at the
522 time, shows instead that orphans closely match their native genomes composition. Cortez et
523 al.[26] proposed that integrative elements are the major source of ORFans in an analysis of 119
524 prokaryotic genomes, based on their identification of clusters of genes of atypical
525 composition. But their evidence is tangential and relies on the fact that some of the identified
526 ORFans (39% of 8,428 genes) are found in such clusters of genes with atypical composition,
527 likely to be derived from integrative elements. Furthermore, results on the crucial point of the
528 composition of ORFans themselves are missing from this study. As there is little overlap
529 between the species surveyed by Cortez et al. (entire prokaryotic diversity) and here (human
530 microbiome), it might be the case that external origination of orphans is more prevalent in
531 species found in different environments.

532 If most of orphan genes analyzed here were indeed "grown locally", by what mechanisms did
533 that happen? Extensive sequence divergence is surely one of them. In a previous study,
534 extensive remote homology searches of metagenomic orphans using Hidden Markov Model
535 (HMM) profiles retrieved significant similarity for ~15% of them[29]. Even more sensitive
536 methods might reveal the true percentage to be even higher. In eukaryotes, an approach that
537 harnesses conserved micro-synteny found that on average about one third of genes without
538 similarity might have simply diverged beyond recognition[21]. While low syntenic conservation
539 prohibits meaningful application of this method to our dataset, our example of Figure 6C and
540 D clearly shows that this can occur. Divergence of gene duplicates within the same genome
541 can result in additional orphans, either by repurposing the same reading frame, or alternative
542 ones. The percentage of orphans sharing similarity to conserved genes from the same genome
543 that we report here is likely an underestimate, as divergence beyond detectable similarity is
544 potentially at play at this level as well. A crucial question, outside the scope of this study, is
545 how often this absence of statistically significant similarity should be taken as absence of
546 functional similarity. The alternatives to divergence are de novo gene birth and overprinting.
547 Have these mechanisms been neglected in prokaryotes? In our view, at least within the
548 environment of the human gut, this seems to be the case. We cannot generalize to other
549 ecological habitats but future studies that leverage the ever-increasing global metagenomic
550 sequences will be useful. Additionally, experimental evidence shows that the emergence of
551 functional proteins via de novo gene evolution from noncoding sequences can provide

552 adaptation in *Escherichia coli*[64,65]. Thus, de novo evolution may be more plausible, and
553 prevalent, than usually assumed.

554 A mostly native orphan origin does not diminish the role of HGT in prokaryotic evolution.
555 Indeed, when one looks at how the percentage of viral matches increases as the number of
556 prokaryotic species in a family increases (Figure 4), it is evident that novel genes can spread
557 between phages and bacteria. It has been known for a long time that phages also invent novel
558 genes[66–68]. An intriguing question for future studies then is whether there is any qualitative or
559 quantitative difference between the processes that surely occur in bacteria and phages, or
560 whether both simply serve as input to a common pool of novel sequences that is accessible
561 via horizontal gene transfer and is maintained under selective pressures or neutrally[69,70].

562 Another important but difficult question is whether our estimates of the number of orphan
563 genes is a good approximation of the true number of entirely novel, *functional* coding
564 sequences unique to each species. We have relied on preexisting annotation that has certainly
565 omitted many novel small ORFs or overlapping reading frames with important physiological
566 effects. Note that using consistent, robust annotations is an important measure against
567 spurious results when it comes to estimating species-specific genes[71]. Furthermore, we have
568 considered any sequence with a Pfam or EggNog match as conserved, but it is possible that
569 an evolutionarily constrained gene might converge towards an existing domain or motif.
570 Finally, we have excluded any recently emerged gene from orphans that was transferred to
571 another species shortly after having emerged. For all these reasons we might have
572 underestimated the true number of orphans. On the other hand, given that orphans are
573 overwhelmingly short, it is reasonable to assume that some of them could be prediction
574 artefacts[7]. For example, conserved gene families have been built which were found to be
575 composed of spurious open reading frames (ORFs) located on the noncoding strand of another
576 conserved gene family[72]. Nevertheless, functional evidence for such "antisense proteins"
577 shows that the noncoding strand also provides potential for evolutionary innovation[73], making
578 it difficult to distinguish functional and non-functional antisense proteins. Future studies
579 based on transcriptomics, ribosome profiling, or functional assays will provide further insights
580 into the expression and function of the orphan genes described here.

581 We have identified more than half a million orphan genes and even if a large part of them
582 turns out to be due to annotation artefacts, we would still be left with thousands of candidates
583 that could be expressed as entirely novel proteins within the boundaries of the human body.
584 It is more and more acknowledged that pangenome diversity is important in the human gut
585 microbiome and might have consequences for human health[74]. Orphan genes within strains
586 of human gut bacteria could perhaps interact with the host, contributing to immune or
587 inflammatory responses and potentially even disease.

588 While such a link of prokaryotic orphan genes to human health can only be speculative, their
589 overall importance for niches-specific adaptations is far from it. Although the cloud genome
590 of rare genes has been found to be an integral part of all pangenomes, the evolutionary
591 processes that generate and maintain the cloud genomes are not well understood and
592 explanations range from niche-specific genes to a constant influx of transient genes[11]. Our
593 results highlight that orphan genes are an important component of prokaryotic pangenomes
594 and especially of the cloud genome. To gain a comprehensive picture of prokaryotic evolution,
595 we thus need to understand where orphan genes come from, what specific functions they
596 assume and how they evolve to be maintained in the population.

Materials and Methods

Identification of orphan genes

Species-specific orphan genes were identified as follows: Initially, the UHGP50 v.1 catalogue which is clustered into protein families at 50% identity by Almeida et al.[42] was parsed and families with members in more than one species were removed. Next, we removed families with a precomputed EggNOG[75] match (based on Almeida et al. data for their representative sequence, i.e. the first in order sequence in the family file). We then performed two similarity searches against the entire NCBI prokaryotic RefSeq database (downloaded January 2022), using DIAMOND[76] v2.0 *blastp* with the representative protein sequence of each family as query: first using DIAMOND "fast" mode, and removing any sequence that had a significant match (E-value<0.001) , except for sequences with >90% coverage and >95% identity that can reasonably be expected to be derived from the same species as the query sequence, present in RefSeq. The same search was then performed for the remaining sequences using the "ultra-sensitive" mode and the same criteria were applied. A final search was performed using BLASTp[77] against the entire UHGP50 protein catalogue (consisting of the representative protein sequences) with an E-value cut-off of 10^{-5} and *"-max_target_seqs 1000"* and any orphan sequence matching with a protein from a different species was filtered out. The remaining sequences constitute the set of species-specific genes, i.e., orphan genes. To mitigate contamination, we also filtered out orphans found on contigs that only contain orphans (only the representative sequence was tested for this). All the non-orphan families (i.e. those removed in the previous steps) were henceforth considered "conserved" and we counted the number of unique species represented within it.

Prediction of CDS and protein properties

Protein secondary structure and intrinsically disordered regions were predicted using the RaptorX Predict_Property package v1.01[78] in default mode. Protein self-aggregation was predicted using PASTA[79] v2.0 in default mode with an energy threshold of -5, and the percentage of the protein expected to form aggregate-prone regions was used. Transmembrane domains were predicted using Phobius[80]. Codon Adaptation Index (CAI) was calculated for all sequences using the CAIJava[81] tool with arguments *"-s -i 15 -k 3 -g"*, and for *E. coli* sequences additionally using the *codonw*[82] tool and the EMBOSS *cai* tool, both using their respective *E. coli* set of reference genes. GC% and GC% in the 3rd synonymous codon position was calculated with *codonw*[82]. CpG dinucleotide frequency was calculated in *R* with the SeqinR[83] package. Biosynthetic cost was calculated by averaging the Akashi and Gojobori[84,85] amino-acid scores for each protein sequence. Isoelectric point was calculated with the *R* package *peptides*[86].

638 ## Similarity searches

639 We downloaded two gut phage protein sequence databases in FASTA format, MGV[52] and
640 GPD[53]. Additionally, protein sequences were extracted from phage genomes assembled by
641 Shah et al.[54] and downloaded from http://copsac.com/earlyvir/f1y/gbks/ in May 2022.
642 Similarity searches against viral protein databases were conducted using DIAMOND *blastp* in
643 'very sensitive' mode. The criteria applied to define significant hits, in increasing order of
644 stringency, were the following: identity <95% & E-value < 10^{-3} (low), identity <95% & E-value
645 < 10^{-5} (medium), 95%>identity >40% & query coverage % >40% & E-value < 10^{-5} (high), 95% >
646 identity > 60% & query coverage >70% & E-value < 10^{-5} (very high). To identify orphans
647 potentially derived from conserved native genes, we also conducted similarity searches of
648 orphan genes against all annotated CDS in their respective genomes using NCBI's TBLASTX
649 with 80% identity, 70% query coverage and E-value 10^{-5}.

650

651 ## Genomic and functional features

652 We predicted the presence of prophages in all 286,997 prokaryotic genomes by running
653 Phigaro[56] in default mode. We then counted the number of orphans and non-orphans that are
654 situated within the predicted prophage genomes.

655 All general genomic data such as assembly quality, as well as taxonomic information were
656 retrieved from the 'genomes-all_metadata.tsv' table provided by Almeida et al. The species-
657 level phylogenetic tree used was taken from the file 'bac120_iqtree.nwk' provided by the
658 authors. All pangenome related statistics were calculated using the pangenome files already
659 available from UHGP, and accessed through links such as
660 ftp://ftp.ebi.ac.uk/pub/databases/metagenomics/mgnify_genomes/human-
661 gut/v1.0/uhgg_catalogue/MGYG-HGUT-000/MGYG-HGUT-00001/pan-genome/ where
662 "MGYG-HGUT-00001" is a species code. Protein/CDS IDs (e.g. GUT_GENOME255189_00514)
663 were matched to unique gene locus IDs (e.g. epsF_2) used within pan-genomes by Almeida et
664 al. and then statistics such as number of genomes where a gene is present were extracted by
665 the relevant tables provided by the authors ('genes_presence-absence_locus.csv' files). The
666 categorization of core (defined at 90% presence) and accessory genes was also retrieved from
667 these directories ('accessory_genes.faa' files). To count the number of orphans per
668 pangenome, we counted the number of unique gene locus IDs, in each species, that we had
669 classified as orphans (one name identifies a single locus and is present only once, even if found
670 in multiple genomes; provided by Almeida et al.).

671 We sought to identify orphan genes among those prokaryotic genes included in the fitness
672 browser[61]. We downloaded protein sequences of fitness browser proteins as well as raw t-
673 score tables for each species from the fitness browser website. We then used DIAMOND to
674 search for similarity between our orphan proteins and those fitness browser proteins. We
675 considered a hit as significant if it had E<10^{-5}, % identity > 70%, and 80% query coverage. We
676 then filtered matches to retain only those to fitness browser proteins with statistically
677 significant fitness effects, defined as t-score > 4 in at least one condition (as detailed in the
678 original publication). To only consider unambiguous cases, we looked only for orphan genes
679 matching to a unique fitness browser protein.

680

681 ## Identification of orthologous regions and definition of de novo candidates

682 For each orphan gene, we searched for each orthologous region in two sets of genomes. In
683 every genome of the same species without the orphan gene, and in every genome of its closest
684 neighboring species, as defined by the species-level phylogeny provided by Almeida et al[42]. To
685 identify the orthologous region in each genome, we took the immediate upstream and
686 downstream neighbors of the orphan gene and searched whether a homologue, that is, a
687 member of the same family, exists in that genome. We only considered cases where a single
688 homologue existed, that is, cases with paralogues were discarded. If both a homologue of the
689 downstream and the upstream neighbor existed in a given outgroup genome, and these were
690 separated by at most 1 gene, we considered the region as conserved in synteny. In order for
691 an orphan gene to be characterized as a de novo candidate, such a conserved syntenic region
692 had to be identified in at least one genome of the same species and at least one genome from
693 the closest species. Additionally, we discarded regions where an unannotated ORF with at
694 least 80% query coverage and 60% similarity to the orphan gene existed in the syntenic region.
695 Finally, we discarded 23 candidates which had similarity matches to conserved genes from the
696 same genome (see 'Similarity searches' subsection above).

697

698 ## Identification of operon-like genomic regions

699 To identify operon-like gene arrangements, we scanned all UHGG genomes and took
700 advantage of the existing Gene Ontology (GO) annotations available from the InterPro
701 searches performed by the authors. For every gene of every genome, we examined its three
702 immediate downstream and three immediate upstream neighbors. If all seven genes were on
703 the same contig, had the same orientation, and at least three of the neighbors shared at least
704 one GO term annotation, then that gene was considered to be in an operon-like arrangement,
705 and the GO term was associated to that gene. Note that we merged the terms GO:0016020
706 and GO:16021 to be consistent with the current version of GO, after we noticed that terms
707 related to the membrane that came up later in our analyses had been tagged as obsolete and
708 collapsed into the general term "membrane". All associations (one gene can have more than
709 one if more than one GO term is found more than three times) for all genes were formatted
710 into a custom annotation file. This file was used as the background annotation, together with
711 a list of orphan genes in BiNGO v3.0.3[87] to perform the overrepresentation analysis. The
712 options used were: hypergeometric test, Benjamini-Hochberg FDR correction and a 0.05
713 threshold. For analysis restricted to each taxonomic class, the same was performed but only
714 including genes from species of a given class.

715 In the case of orphan gene GUT_GENOME108422_00105 (representative of a family of 135
716 proteins) our analysis of the syntenic region showed the gene to be present at the exact
717 orthologous locus as orphan gene GUT_GENOME123088_00522 (representative of a family of
718 866 members) both belonging to species PeH17 (GUT_GENOME014725). Aligning the two
719 representative sequences gave results consistent with the criteria used by Almeida et al. for
720 protein clustering: the sequences align with 100% identity but only over their first 25 residues
721 (46% of the length of the shortest sequence), while the rest of the sequences align very poorly.
722 However, all other protein sequences of the GUT_GENOME123088_00522 family align nearly
723 perfectly with any protein from the GUT_GENOME108422_00105 family, with no gaps and
724 with 94% identity. Indeed, aligning all CDS sequences from both families together results in

725 an alignment with average pairwise percent identity of 97% (calculated using the *goalign* tool
726 and its *compute distance* method with the *pdist* metric). Furthermore, a nucleotide alignment
727 of the representative sequences showed that their difference at the protein level is due to a
728 frameshift. Overall, these two families, with the exception of the representative sequence of
729 GUT_GENOME123088_00522, would have been expected to cluster together given the
730 parameters used by Almeida et al. (*linclust* with 80% coverage and 50% identity).

731

732 Quantification and statistical analysis

733 All statistics were done in R v3.6.2. Plots were generated using *ggplot2*[88]. All statistical details
734 including the type of statistical test performed and exact value of n (n represents either
735 number of genomes or number of genes) can be found in the Results and figure legends.
736 Boxplots show median (horizontal line inside the box), first and third quartiles of data (lower
737 and upper hinges) and values no further or lower than 1.5*distance between the first and
738 third quartiles (upper and lower whisker). No methods were used to determine whether the
739 data met assumptions of the statistical approaches. Subsampling of conserved genes to
740 control for length difference to orphans was performed with replacement, using a customized
741 version of inverse transform sampling. We sampled conserved genes belonging to each
742 species separately, and only species with at least 100 conserved and at least 10 orphans were
743 included (n=3,769). The length median of the sampled conserved was 228nt and the mean
744 was 305nt, compared to 225nt and 300nt in orphans respectively.

745

746

747 Acknowledgements

748

749 We thank Daniel Tamarit and Franz Baumdicker for valuable comments on an earlier version
750 of the manuscript. We acknowledge funding from the DFG in the framework of the SPP2141
751 to A.K. (KU3610/2-1). The research project was supported by the Hellenic Foundation for
752 Research and Innovation (H.F.R.I.) under the "3rd Call for H.F.R.I. Research Projects to
753 support Post-Doctoral Researchers" to N.V. (Project Number:7330). This research was
754 supported in part through high-performance computing resources available at the Kiel
755 University Computing Centre.

756

757

758 Author information

759

760 Contributions

761 AK and NV conceived the study. AK and NV wrote the manuscript. AK supervised the study.
762 NV performed the analyses.

763

Affiliations

Anne Kupczok

Bioinformatics Group, Wageningen University, Wageningen, The Netherlands

Nikolaos Vakirlis

Institute For Fundamental Biomedical Research, B.S.R.C. "Alexander Fleming", Vari, Greece (Current)

Institute for General Microbiology, Kiel University, Kiel, Germany

Corresponding authors

Correspondence to anne.kupczok@wur.nl and vakirlis@fleming.gr

References

1. Brockhurst, M. A. *et al.* The Ecology and Evolution of Pangenomes. *Current Biology* **29**, R1094–R1103 (2019).
2. Koonin, E. V. & Wolf, Y. I. Genomics of bacteria and archaea: the emerging dynamic view of the prokaryotic world. *Nucleic Acids Research* **36**, 6688–6719 (2008).
3. Koonin, E. V., Makarova, K. S. & Wolf, Y. I. Evolution of Microbial Genomics: Conceptual Shifts over a Quarter Century. *Trends in Microbiology* **29**, 582–592 (2021).
4. Conrad, R. E. *et al.* Toward quantifying the adaptive role of bacterial pangenomes during environmental perturbations. *ISME J* 1–13 (2021) doi:10.1038/s41396-021-01149-9.
5. Dujon, B. The yeast genome project: what did we learn? *Trends in Genetics* **12**, 263–270 (1996).

791 6. Karlowski, W. M., Varshney, D. & Zielezinski, A. Taxonomically Restricted Genes in
792 Bacillus may Form Clusters of Homologs and Can be Traced to a Large Reservoir of
793 Noncoding Sequences. *Genome Biology and Evolution* **15**, evad023 (2023).
794 7. Yu, G. & Stoltzfus, A. Population diversity of ORFan genes in Escherichia coli. *Genome*
795 *Biol Evol* **4**, 1176–1187 (2012).
796 8. Fitch, W. M. Homology: a personal view on some of the problems. *Trends in Genetics* **16**,
797 227–231 (2000).
798 9. Koonin, E. V. Orthologs, Paralogs, and Evolutionary Genomics. *Annual Review of*
799 *Genetics* **39**, 309–338 (2005).
800 10. Vanni, C. *et al.* Unifying the known and unknown microbial coding sequence space. *eLife*
801 **11**, e67667 (2022).
802 11. Baumdicker, F. & Kupczok, A. Tackling the Pangenome Dilemma Requires the Concerted
803 Analysis of Multiple Population Genetic Processes. *Genome Biology and Evolution* **15**,
804 evad067 (2023).
805 12. Light, S., Basile, W. & Elofsson, A. Orphans and new gene origination, a structural and
806 evolutionary perspective. *Current Opinion in Structural Biology* **26**, 73–83 (2014).
807 13. Khalturin, K., Hemmrich, G., Fraune, S., Augustin, R. & Bosch, T. C. G. More than just
808 orphans: are taxonomically-restricted genes important in evolution? *Trends in Genetics*
809 **25**, 404–413 (2009).
810 14. Santos, M. E., Bouquin, A. L., Crumière, A. J. J. & Khila, A. Taxon-restricted genes at the
811 origin of a novel trait allowing access to a new environment. *Science* **358**, 386–390
812 (2017).
813 15. Tautz, D. & Domazet-Lošo, T. The evolutionary origin of orphan genes. *Nat Rev Genet*
814 **12**, 692–702 (2011).
815 16. Andersson, D. I., Jerlström-Hultqvist, J. & Näsvall, J. Evolution of New Functions De Novo
816 and from Preexisting Genes. *Cold Spring Harb Perspect Biol* **7**, a017996 (2015).

817 17. Prabh, N. & Rödelsperger, C. De Novo, Divergence, and Mixed Origin Contribute to the
818 Emergence of Orphan Genes in Pristionchus Nematodes. *G3: Genes, Genomes, Genetics*
819 **9**, 2277–2286 (2019).
820 18. McLysaght, A. & Guerzoni, D. New genes from non-coding sequence: the role of de novo
821 protein-coding genes in eukaryotic evolutionary innovation. *Phil. Trans. R. Soc. B* **370**,
822 20140332 (2015).
823 19. Oss, S. B. V. & Carvunis, A.-R. De novo gene birth. *PLOS Genetics* **15**, e1008160 (2019).
824 20. Bornberg-Bauer, E., Hlouchova, K. & Lange, A. Structure and function of naturally
825 evolved de novo proteins. *Current Opinion in Structural Biology* **68**, 175–183 (2021).
826 21. Vakirlis, N., Carvunis, A.-R. & McLysaght, A. Synteny-based analyses indicate that
827 sequence divergence is not the main source of orphan genes. *eLife* **9**, e53500 (2020).
828 22. Weisman, C. M., Murray, A. W. & Eddy, S. R. Many, but not all, lineage-specific genes
829 can be explained by homology detection failure. *PLOS Biology* **18**, e3000862 (2020).
830 23. McLysaght, A. & Hurst, L. D. Open questions in the study of de novo genes: what, how
831 and why. *Nat Rev Genet* **17**, 567–578 (2016).
832 24. Daubin, V. & Ochman, H. Start-up entities in the origin of new genes. *Current Opinion in*
833 *Genetics & Development* **14**, 616–619 (2004).
834 25. Daubin, V. & Ochman, H. Bacterial Genomes as New Gene Homes: The Genealogy of
835 ORFans in E. coli. *Genome Res.* **14**, 1036–1042 (2004).
836 26. Cortez, D., Forterre, P. & Gribaldo, S. A hidden reservoir of integrative elements is the
837 major source of recently acquired foreign genes and ORFans in archaeal and bacterial
838 genomes. *Genome Biology* **10**, R65 (2009).
839 27. Yin, Y. & Fischer, D. On the origin of microbial ORFans: quantifying the strength of the
840 evidence for viral lateral transfer. *BMC Evolutionary Biology* **6**, 63 (2006).
841 28. Yomtovian, I., Teerakulkittipong, N., Lee, B., Moult, J. & Unger, R. Composition bias and
842 the origin of ORFan genes. *Bioinformatics* **26**, 996–999 (2010).

843 29. Lobb, B., Kurtz, D. A., Moreno-Hagelsieb, G. & Doxey, A. C. Remote homology and the
844 functions of metagenomic dark matter. *Front Genet* **6**, (2015).
845 30. Delaye, L., DeLuna, A., Lazcano, A. & Becerra, A. The origin of a novel gene through
846 overprinting in Escherichia coli. *BMC Evolutionary Biology* **8**, 31 (2008).
847 31. Ardern, Z. Alternative Reading Frames are an Underappreciated Source of Protein
848 Sequence Novelty. *Journal of Molecular Evolution* (2023) doi:10.1007/s00239-023-
849 10122-3.
850 32. Smith, C. *et al.* Pervasive translation in Mycobacterium tuberculosis. *eLife* **11**, e73980
851 (2022).
852 33. Wade, J. T. & Grainger, D. C. Pervasive transcription: illuminating the dark matter of
853 bacterial transcriptomes. *Nat Rev Microbiol* **12**, 647–653 (2014).
854 34. Knopp, M. *et al.* De Novo Emergence of Peptides That Confer Antibiotic Resistance.
855 *mBio* **10**, e00837-19 (2019).
856 35. Knopp, M. *et al.* A novel type of colistin resistance genes selected from random
857 sequence space. *PLOS Genetics* **17**, e1009227 (2021).
858 36. Tenson, T., Xiong, L., Kloss, P. & Mankin, A. S. Erythromycin Resistance Peptides Selected
859 from Random Peptide Libraries. *J. Biol. Chem.* **272**, 17425–17430 (1997).
860 37. Kirchberger, P. C., Schmidt, M. L. & Ochman, H. The Ingenuity of Bacterial Genomes.
861 *Annual Review of Microbiology* **74**, 815–834 (2020).
862 38. Watson, A. K., Lopez, P. & Bapteste, E. Hundreds of Out-of-Frame Remodeled Gene
863 Families in the Escherichia coli Pangenome. *Molecular Biology and Evolution* **39**,
864 msab329 (2022).
865 39. Chibani, C. M. *et al.* A catalogue of 1,167 genomes from the human gut archaeome. *Nat*
866 *Microbiol* **7**, 48–61 (2022).
867 40. Nayfach, S. *et al.* A genomic catalog of Earth's microbiomes. *Nat Biotechnol* **39**, 499–509
868 (2021).

869 41. Coelho, L. P. *et al.* Towards the biogeography of prokaryotic genes. *Nature* **601**, 252–256
870 (2022).
871 42. Almeida, A. *et al.* A unified catalog of 204,938 reference genomes from the human gut
872 microbiome. *Nat Biotechnol* **39**, 105–114 (2021).
873 43. Vakirlis, N. *et al.* A Molecular Portrait of De Novo Genes in Yeasts. *Mol Biol Evol* **35**, 631–
874 645 (2018).
875 44. Lawrence, J. G. & Ochman, H. Amelioration of bacterial genomes: rates of change and
876 exchange. *J Mol Evol* **44**, 383–397 (1997).
877 45. Montañés, J. C., Huertas, M., Messeguer, X. & Albà, M. M. Evolutionary Trajectories of
878 New Duplicated and Putative De Novo Genes. *Molecular Biology and Evolution* **40**,
879 msad098 (2023).
880 46. Blevins, W. R. *et al.* Uncovering de novo gene birth in yeast using deep transcriptomics.
881 *Nat Commun* **12**, 604 (2021).
882 47. Vakirlis, N. *et al.* De novo emergence of adaptive membrane proteins from thymine-rich
883 genomic sequences. *Nature Communications* **11**, 1–18 (2020).
884 48. Tassios, E., Nikolaou, C. & Vakirlis, N. Intergenic regions of Saccharomycotina yeasts are
885 enriched in potential to encode transmembrane domains. *Molecular Biology and*
886 *Evolution* msad059 (2023) doi:10.1093/molbev/msad059.
887 49. Sberro, H. *et al.* Large-Scale Analyses of Human Microbiomes Reveal Thousands of Small,
888 Novel Genes. *Cell* **178**, 1245-1259.e14 (2019).
889 50. Hershberg, R. & Petrov, D. A. Evidence That Mutation Is Universally Biased towards AT in
890 Bacteria. *PLOS Genetics* **6**, e1001115 (2010).
891 51. Papanikolaou, N., Trachana, K., Theodosiou, T., Promponas, V. J. & Iliopoulos, I. Gene
892 socialization: gene order, GC content and gene silencing in Salmonella. *BMC Genomics*
893 **10**, 597 (2009).

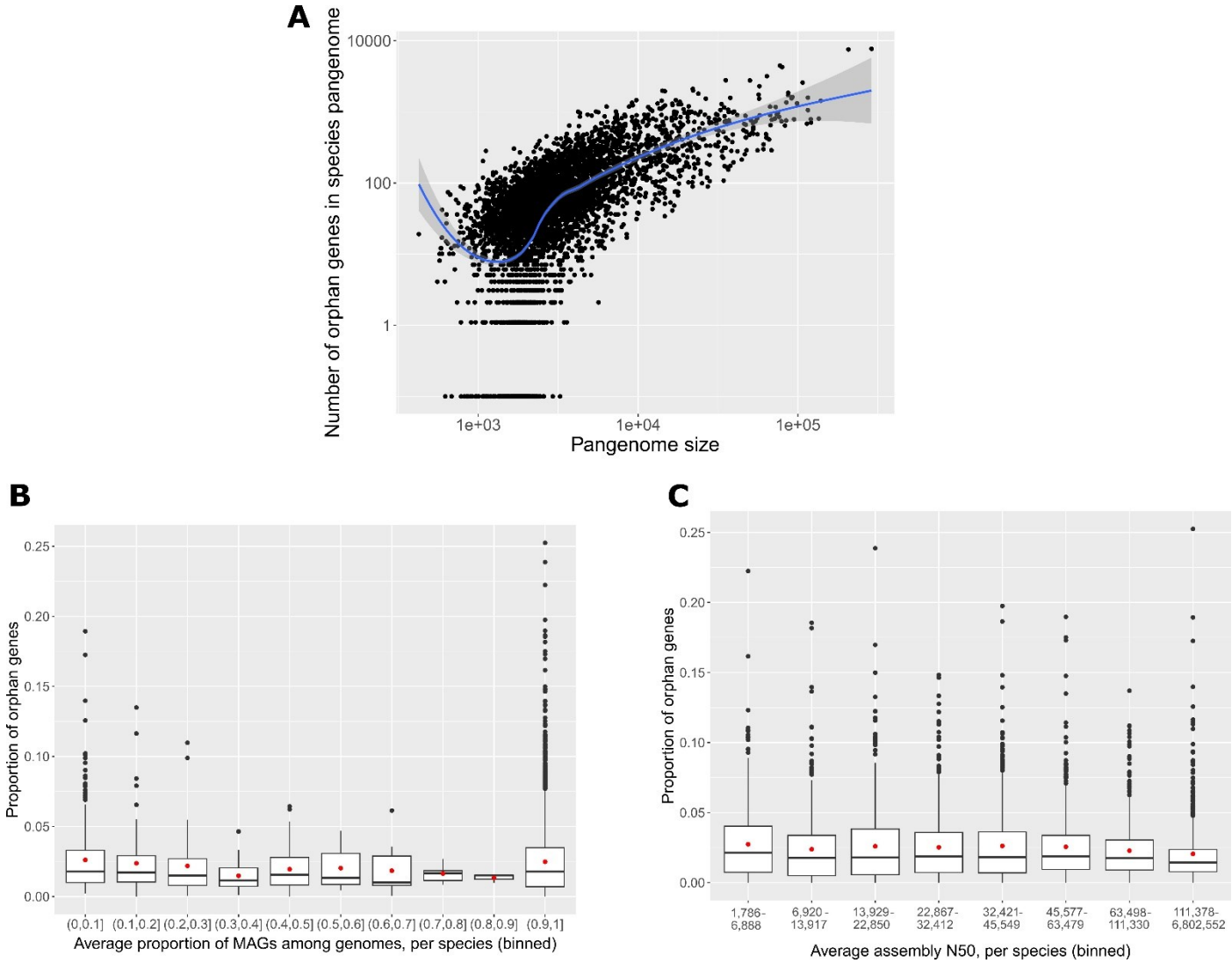
894 52. Nayfach, S. *et al.* Metagenomic compendium of 189,680 DNA viruses from the human
895 gut microbiome. *Nat Microbiol* **6**, 960–970 (2021).
896 53. Camarillo-Guerrero, L. F., Almeida, A., Rangel-Pineros, G., Finn, R. D. & Lawley, T. D.
897 Massive expansion of human gut bacteriophage diversity. *Cell* **184**, 1098-1109.e9
898 (2021).
899 54. Shah, S. A. *et al.* Expanding known viral diversity in the healthy infant gut. *Nat Microbiol*
900 **8**, 986–998 (2023).
901 55. Almpanis, A., Swain, M., Gatherer, D. & McEwan, N. Correlation between bacterial G+C
902 content, genome size and the G+C content of associated plasmids and bacteriophages.
903 *Microb Genom* **4**, e000168 (2018).
904 56. Starikova, E. V. *et al.* Phigaro: high-throughput prophage sequence annotation.
905 *Bioinformatics* **36**, 3882–3884 (2020).
906 57. Tria, F. D. K. & Martin, W. F. Gene Duplications Are At Least 50 Times Less Frequent than
907 Gene Transfers in Prokaryotic Genomes. *Genome Biology and Evolution* **13**, evab224
908 (2021).
909 58. Watson, A. K., Lopez, P. & Bapteste, E. Hundreds of out-of-frame remodelled gene
910 families in the E. coli pangenome. *Molecular Biology and Evolution* (2021)
911 doi:10.1093/molbev/msab329.
912 59. Katoh, K. & Standley, D. M. MAFFT Multiple Sequence Alignment Software Version 7:
913 Improvements in Performance and Usability. *Mol Biol Evol* **30**, 772–780 (2013).
914 60. Waterhouse, A. M., Procter, J. B., Martin, D. M. A., Clamp, M. & Barton, G. J. Jalview
915 Version 2—a multiple sequence alignment editor and analysis workbench.
916 *Bioinformatics* **25**, 1189–1191 (2009).
917 61. Price, M. N. *et al.* Mutant phenotypes for thousands of bacterial genes of unknown
918 function. *Nature* **557**, 503–509 (2018).
919 62. Osbourn, A. E. & Field, B. Operons. *Cell. Mol. Life Sci.* **66**, 3755–3775 (2009).

920 63. Río, Á. R. del *et al.* Functional and evolutionary significance of unknown genes from
921 uncultivated taxa. 2022.01.26.477801 Preprint at
922 https://doi.org/10.1101/2022.01.26.477801 (2022).
923 64. Babina, A. M. *et al.* Rescue of Escherichia coli auxotrophy by de novo small proteins.
924 *eLife* **12**, e78299 (2023).
925 65. Frumkin, I. & Laub, M. T. Selection of a de novo gene that can promote survival of E. coli
926 by modulating protein homeostasis pathways. 2023.02.07.527531 Preprint at
927 https://doi.org/10.1101/2023.02.07.527531 (2023).
928 66. Fremin, B. J. *et al.* Thousands of small, novel genes predicted in global phage genomes.
929 *Cell Reports* **39**, (2022).
930 67. Pavesi, A., Magiorkinis, G. & Karlin, D. G. Viral Proteins Originated De Novo by
931 Overprinting Can Be Identified by Codon Usage: Application to the "Gene Nursery" of
932 Deltaretroviruses. *PLoS Comput Biol* **9**, (2013).
933 68. Sabath, N., Wagner, A. & Karlin, D. Evolution of Viral Proteins Originated De Novo by
934 Overprinting. *Mol Biol Evol* **29**, 3767–3780 (2012).
935 69. Wolf, Y. I., Makarova, K. S., Lobkovsky, A. E. & Koonin, E. V. Two fundamentally different
936 classes of microbial genes. *Nat Microbiol* **2**, 1–6 (2016).
937 70. Iyengar, B. R. & Bornberg-Bauer, E. Neutral Models of De Novo Gene Emergence Suggest
938 that Gene Evolution has a Preferred Trajectory. *Molecular Biology and Evolution* **40**,
939 msad079 (2023).
940 71. Weisman, C. M., Murray, A. W. & Eddy, S. R. Mixing genome annotation methods in a
941 comparative analysis inflates the apparent number of lineage-specific genes. *Current*
942 *Biology* **32**, 2632-2639.e2 (2022).
943 72. Eberhardt, R. Y. *et al.* AntiFam: a tool to help identify spurious ORFs in protein
944 annotation. *Database (Oxford)* **2012**, bas003 (2012).

945 73. Ardern, Z., Neuhaus, K. & Scherer, S. Are Antisense Proteins in Prokaryotes Functional?
946 *Frontiers in Molecular Biosciences* **7**, (2020).
947 74. Vatanen, T. *et al.* Genomic variation and strain-specific functional adaptation in the
948 human gut microbiome during early life. *Nat Microbiol* **4**, 470–479 (2019).
949 75. Hernández-Plaza, A. *et al.* eggNOG 6.0: enabling comparative genomics across 12 535
950 organisms. *Nucleic Acids Research* **51**, D389–D394 (2023).
951 76. Buchfink, B., Reuter, K. & Drost, H.-G. Sensitive protein alignments at tree-of-life scale
952 using DIAMOND. *Nat Methods* **18**, 366–368 (2021).
953 77. Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database
954 search programs. *Nucleic Acids Res.* **25**, 3389–3402 (1997).
955 78. Wang, S., Li, W., Liu, S. & Xu, J. RaptorX-Property: a web server for protein structure
956 property prediction. *Nucleic Acids Res* **44**, W430–W435 (2016).
957 79. Walsh, I., Seno, F., Tosatto, S. C. E. & Trovato, A. PASTA 2.0: an improved server for
958 protein aggregation prediction. *Nucleic Acids Research* **42**, W301–W307 (2014).
959 80. Käll, L., Krogh, A. & Sonnhammer, E. L. L. Advantages of combined transmembrane
960 topology and signal peptide prediction—the Phobius web server. *Nucleic Acids Res* **35**,
961 W429–W432 (2007).
962 81. Carbone, A., Zinovyev, A. & Képès, F. Codon adaptation index as a measure of
963 dominating codon bias. *Bioinformatics* **19**, 2005–2015 (2003).
964 82. Correspondence Analysis of Codon Usage. http://codonw.sourceforge.net/.
965 83. Charif, D. & Lobry, J. R. SeqinR 1.0-2: A Contributed Package to the R Project for
966 Statistical Computing Devoted to Biological Sequences Retrieval and Analysis. in
967 *Structural Approaches to Sequence Evolution: Molecules, Networks, Populations* (eds.
968 Bastolla, U., Porto, M., Roman, H. E. & Vendruscolo, M.) 207–232 (Springer, 2007).
969 doi:10.1007/978-3-540-35306-5_10.

970 84. Akashi, H. & Gojobori, T. Metabolic efficiency and amino acid composition in the
971 proteomes of Escherichia coli and Bacillus subtilis. *Proc. Natl. Acad. Sci. U.S.A.* **99**, 3695–
972 3700 (2002).
973 85. Barton, M. D., Delneri, D., Oliver, S. G., Rattray, M. & Bergman, C. M. Evolutionary
974 Systems Biology of Amino Acid Biosynthetic Cost in Yeast. *PLOS ONE* **5**, e11935 (2010).
975 86. Osorio, D., Rondón-Villarreal, P. & Torres, R. Peptides: A package for data mining of
976 antimicrobial peptides. *Small* **12**, 44–444 (2015).
977 87. Maere, S., Heymans, K. & Kuiper, M. BiNGO: a Cytoscape plugin to assess
978 overrepresentation of gene ontology categories in biological networks. *Bioinformatics*
979 **21**, 3448–3449 (2005).
980 88. Wickham, H. ggplot2. *WIREs Computational Statistics* **3**, 180–185 (2011).
981
982
983
984
985
986
987
988
989
990
991
992
993
994
995
996
997

Supplementary Figures



Supplementary Figure 1. **A**: Number of orphan genes in species pangenome as a function of pangenome size. Data are log-transformed. A LOESS fit in the log-transformed data is shown. We note that some saturation is visible in values of highest pangenome sizes. **B**: Proportion of orphan genes as a fraction of the pangenome, in species binned according to their MAG proportion. **C**: Same as B, but species are binned according to their average assembly quality N50. Bins have been set to contain the same number of orphan genes.

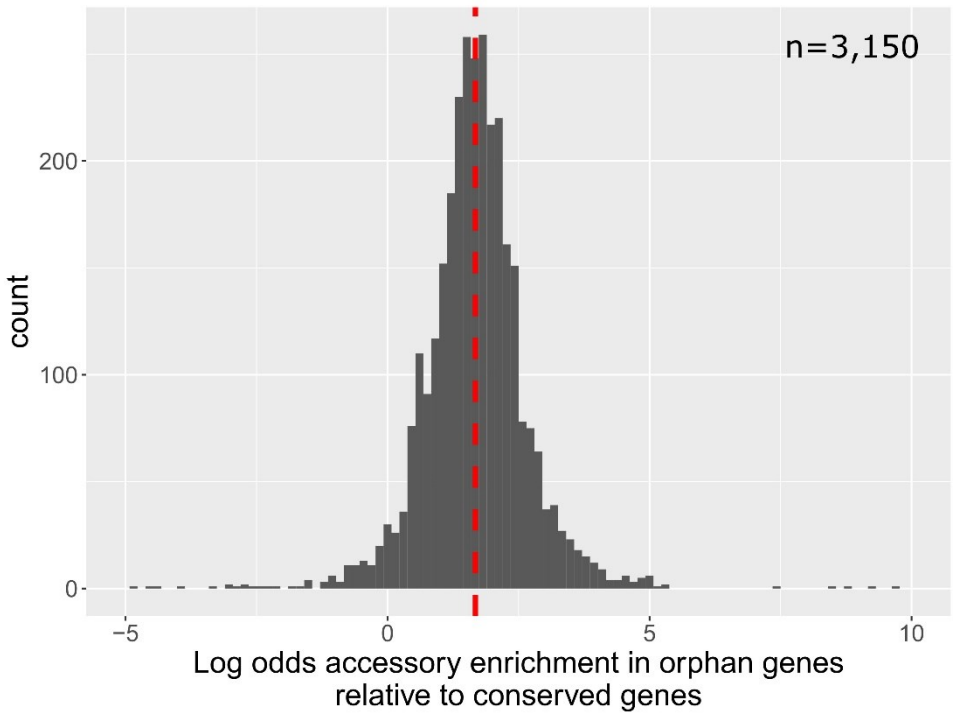
1012

1013

1014

1015

1016



1017

1018

1019 **Supplementary Figure 2**: Enrichment of accessory genes over core genes in orphans relative
1020 to conserved genes (see Methods). Range of X-axis has been reduced for visibility, 12
1021 genomes with extreme values are not shown. Core and accessory genes were defined by
1022 Almeida et al[42].

1023

1024

1025

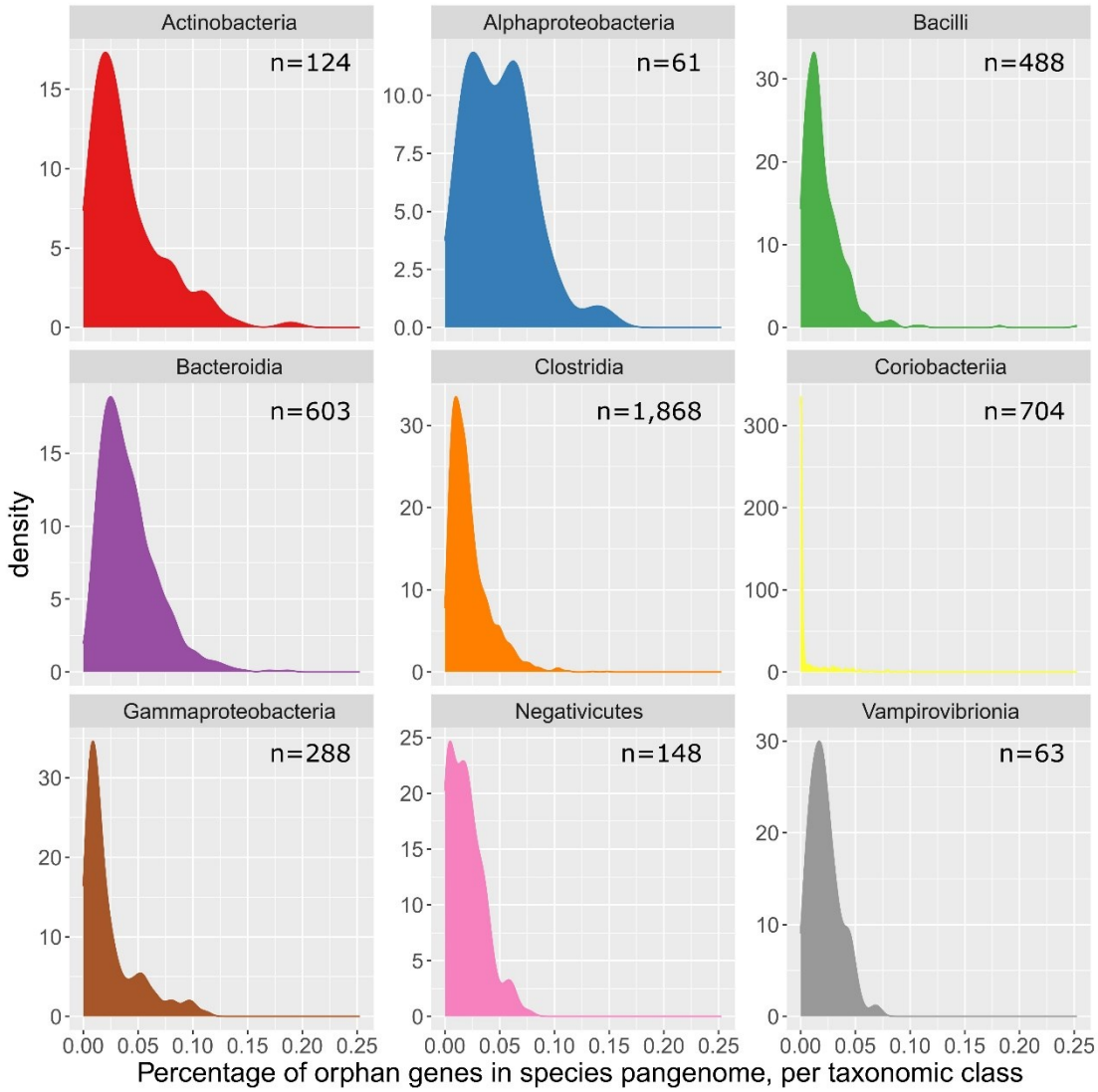
1026

1027

1028

1029

1030



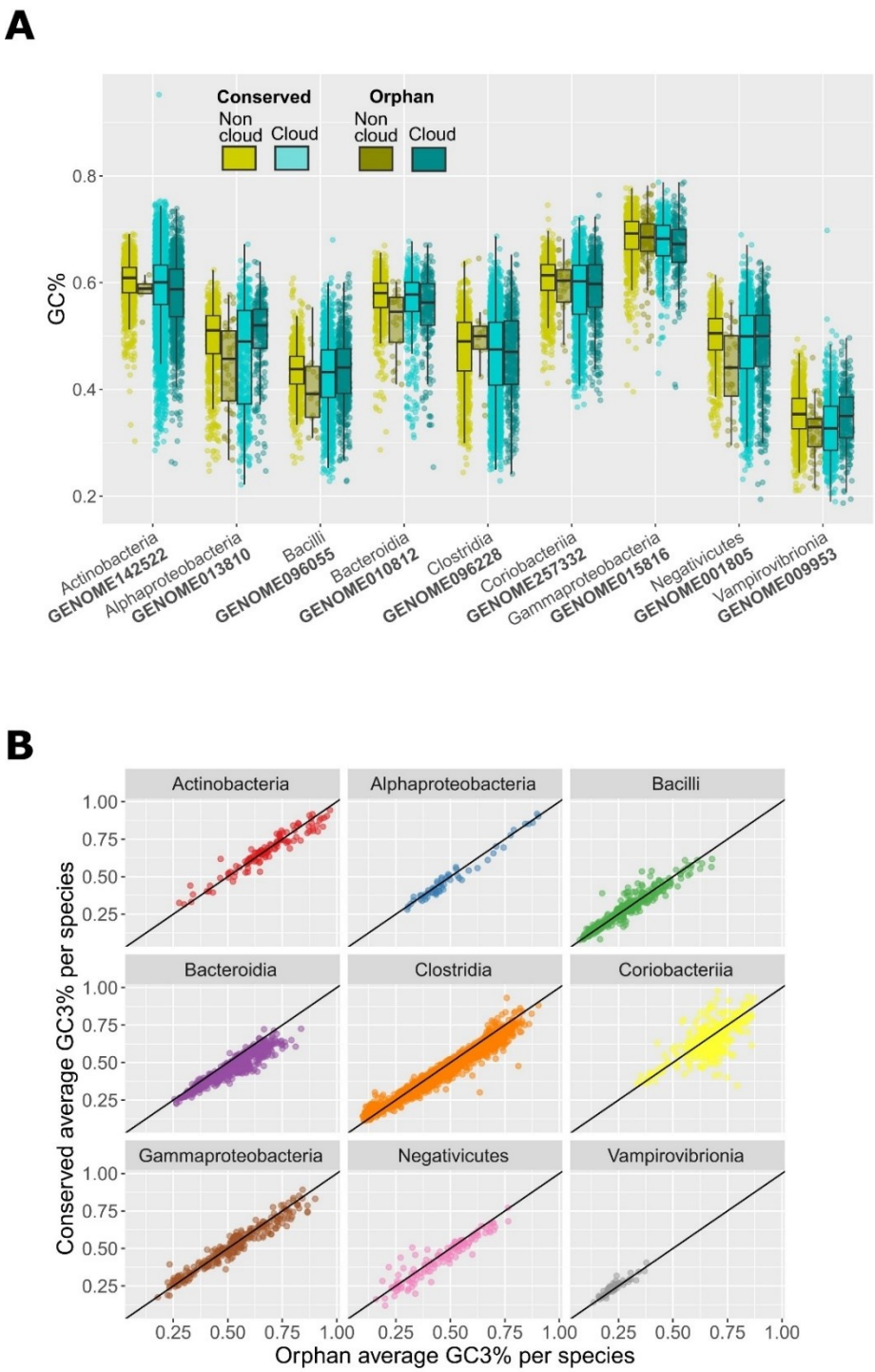
1031

1032

1033 **Supplementary Figure 3.** Distributions of percentage of orphan genes in each
1034 species pangenome split across the nine best represented taxonomic classes.

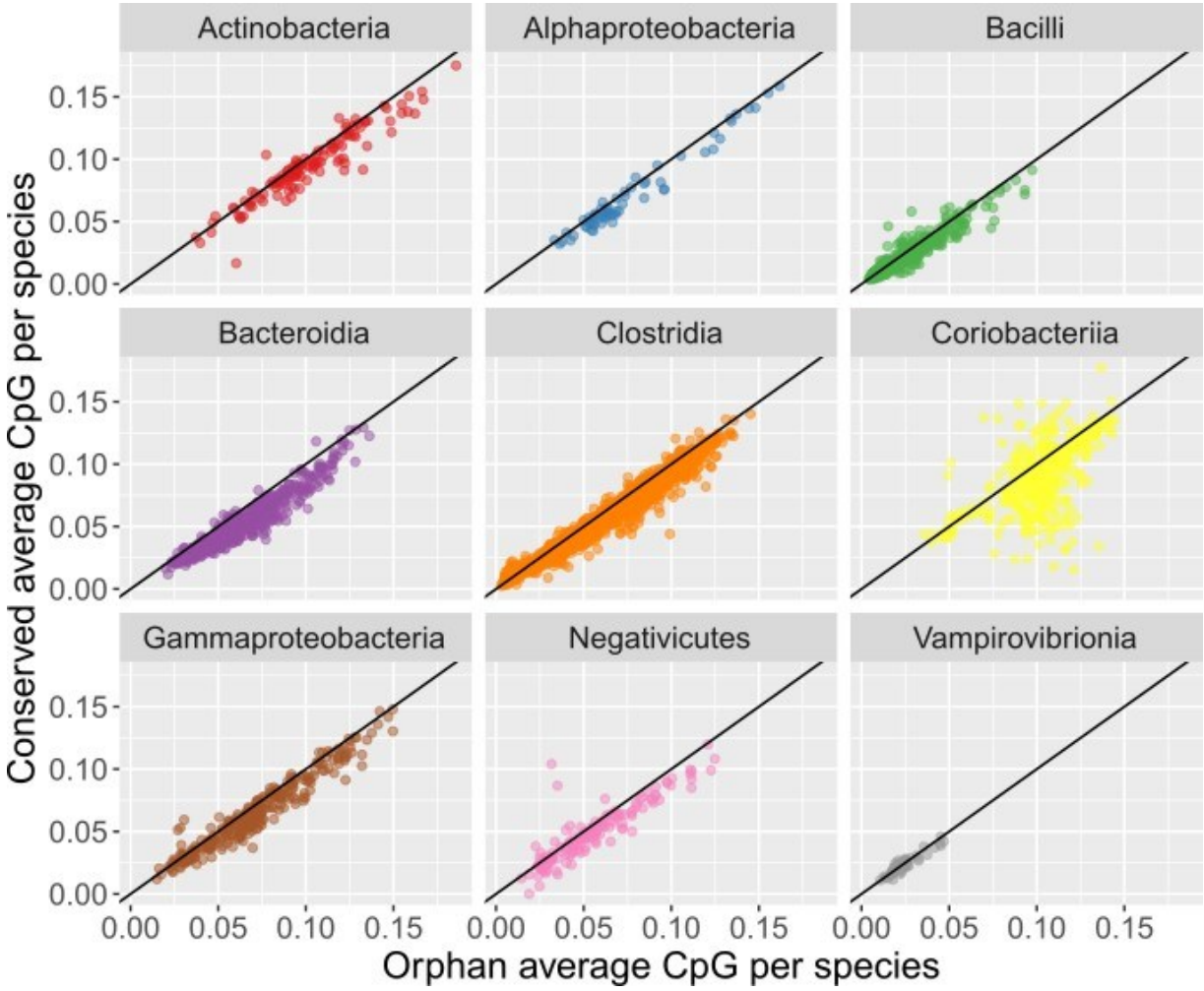
1035

1036



Supplementary Figure 4: **A:** Intraspecific comparison of GC% in orphan and conserved genes that are found in 10% or less of genomes in their species (cloud) or not (non-cloud). Data are shown for a sample of 10 species with at least 200 orphans and at least 10 genomes, drawn randomly from each of the nine best represented taxonomic classes. Each point represents one gene, that is data here are not averaged as in Figure 3. **B**: Comparison of per species average GC3s% between orphan and conserved genes in the nine best represented taxonomic classes.

1046



1047

1048

1049 **Supplementary Figure 5**: Comparison of CpG% between orphan and conserved genes in the
1050 nine best represented taxonomic classes.

1051

1052

1053

1054

1055

1056

1057

1058

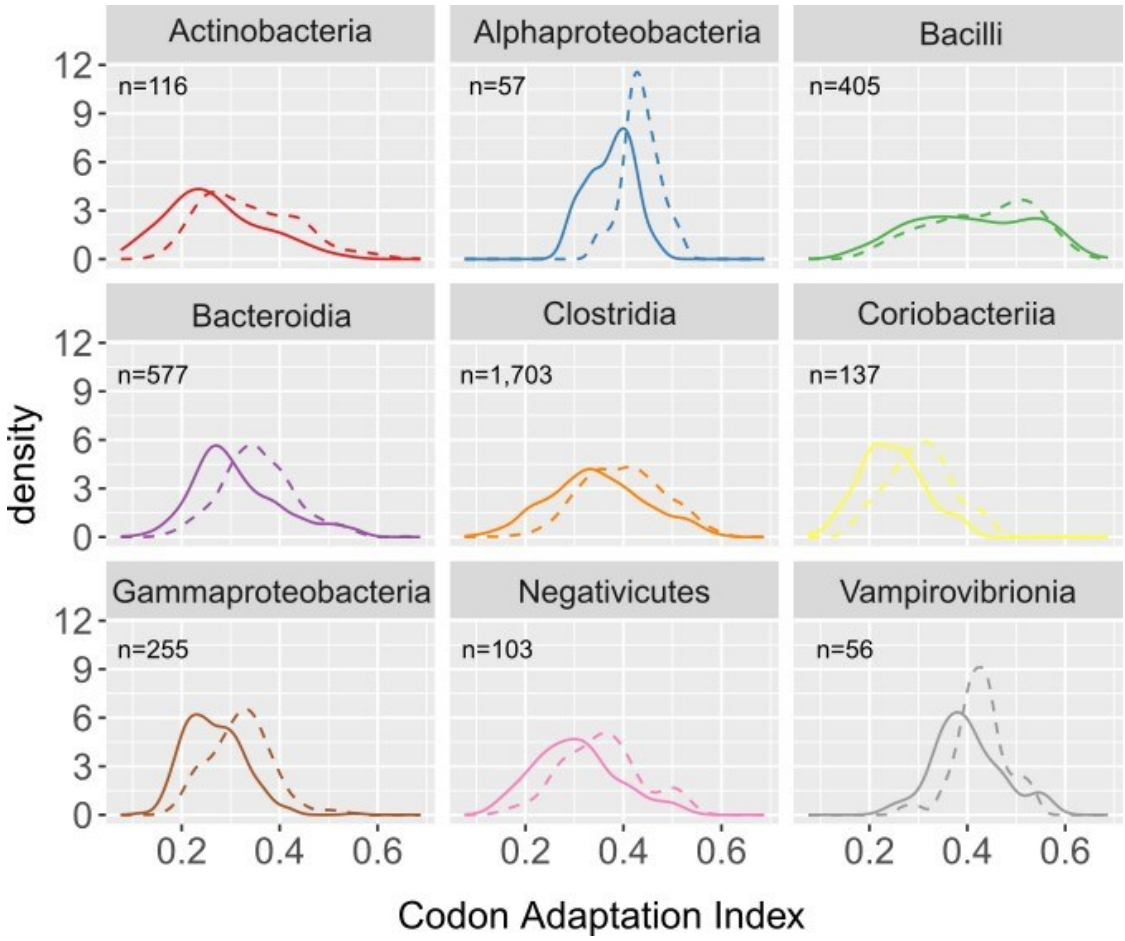
1059

1060

1061

1062

1063



1064

1065 **Supplementary Figure 6:** Comparison of CAI value distributions between orphans (solid line)
1066 and conserved genes (dotted line), subsampling conserved genes to match the length
1067 distribution of orphan. All differences are statistically significant with Wilcoxon test P-value <
1068 10^{-4} except for Bacilli (P=0.02) and Vampirovibrionia (P=0.005).

1069

1070

1071

1072

1073

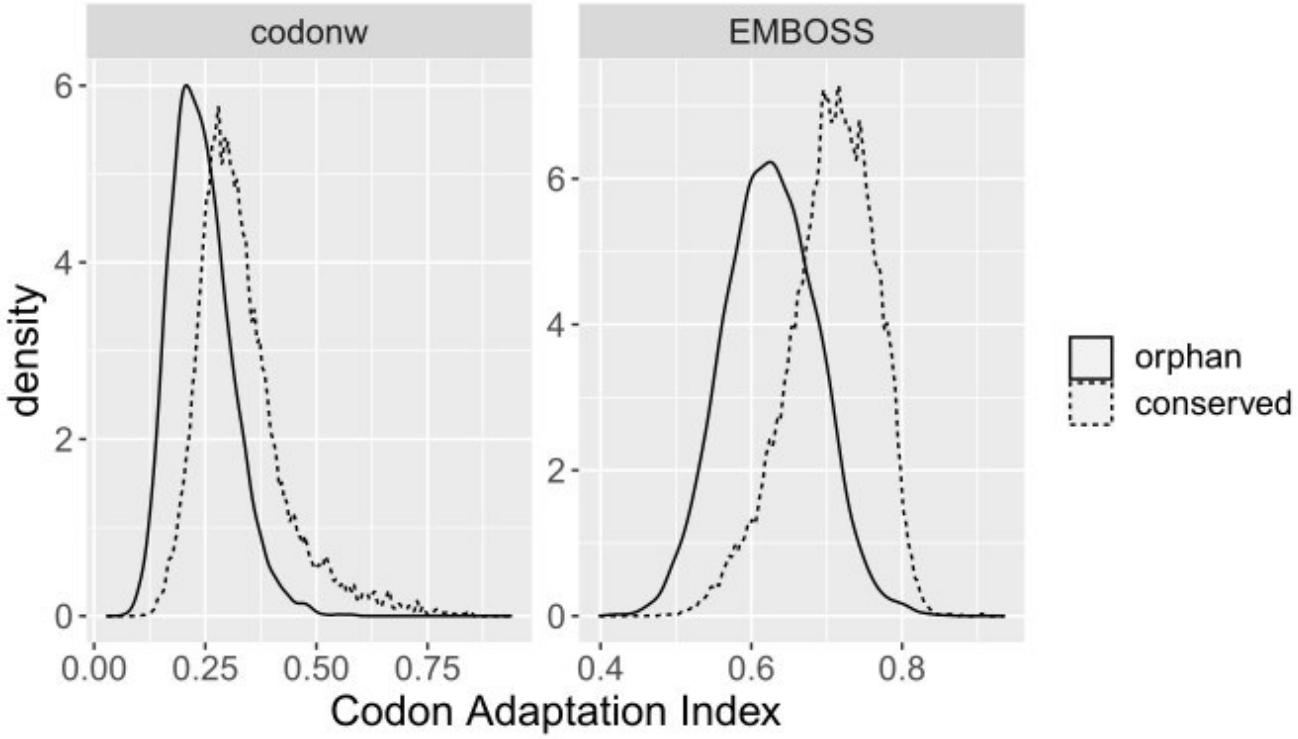
1074

1075

1076

1077

1078



1079

1080 **Supplementary Figure 7**: CAI comparison of all orphan (n=4,703) and all conserved genes
1081 (n=35,300,683) from 8,288 *E. coli* genomes included in our dataset, calculated using two
1082 different tools/reference gene sets. Both Wilcoxon test P-values < $2.2*10^{-16}$.

1083

1084

1085

1086

1087

1088

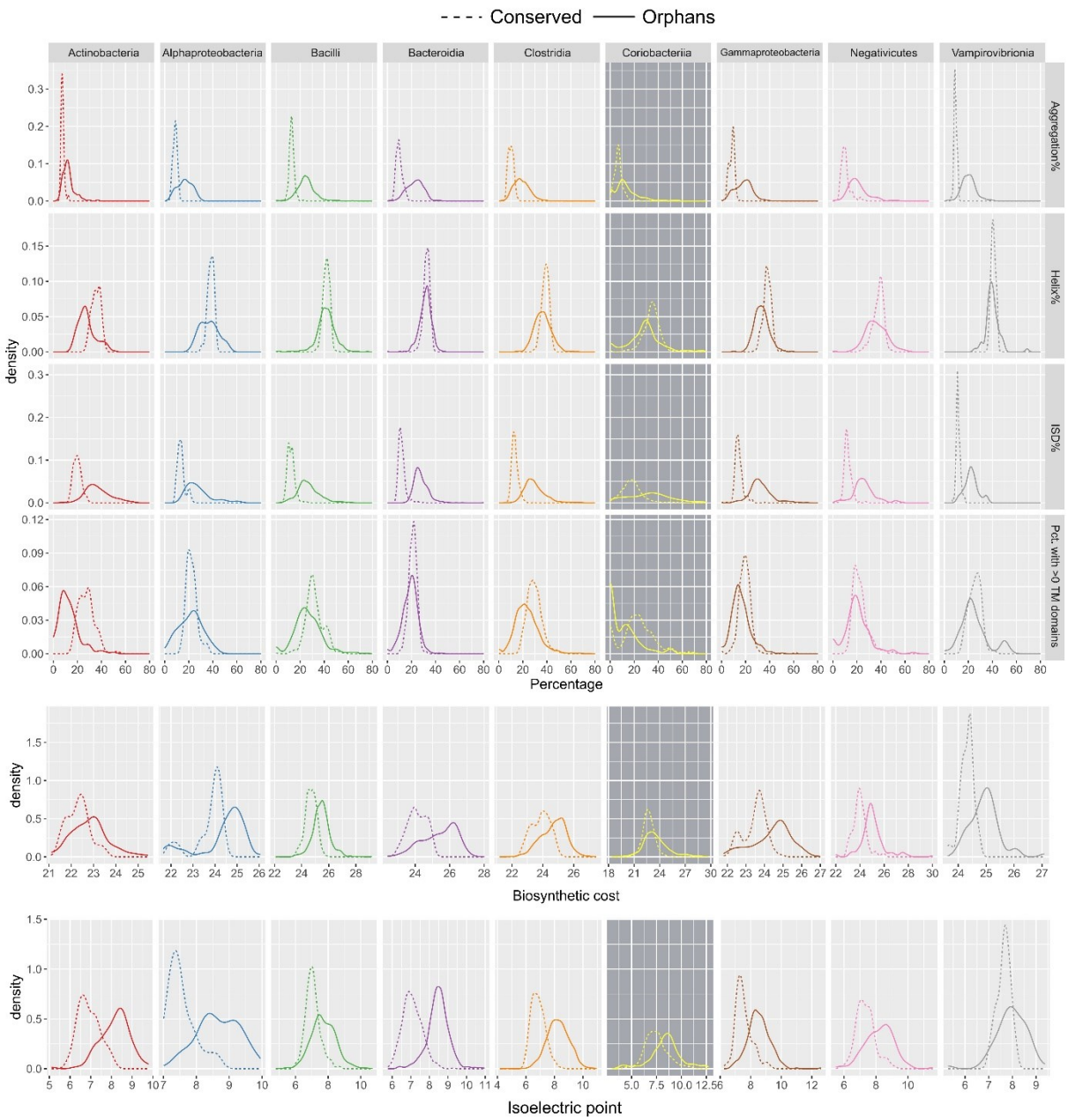
1089

1090

1091

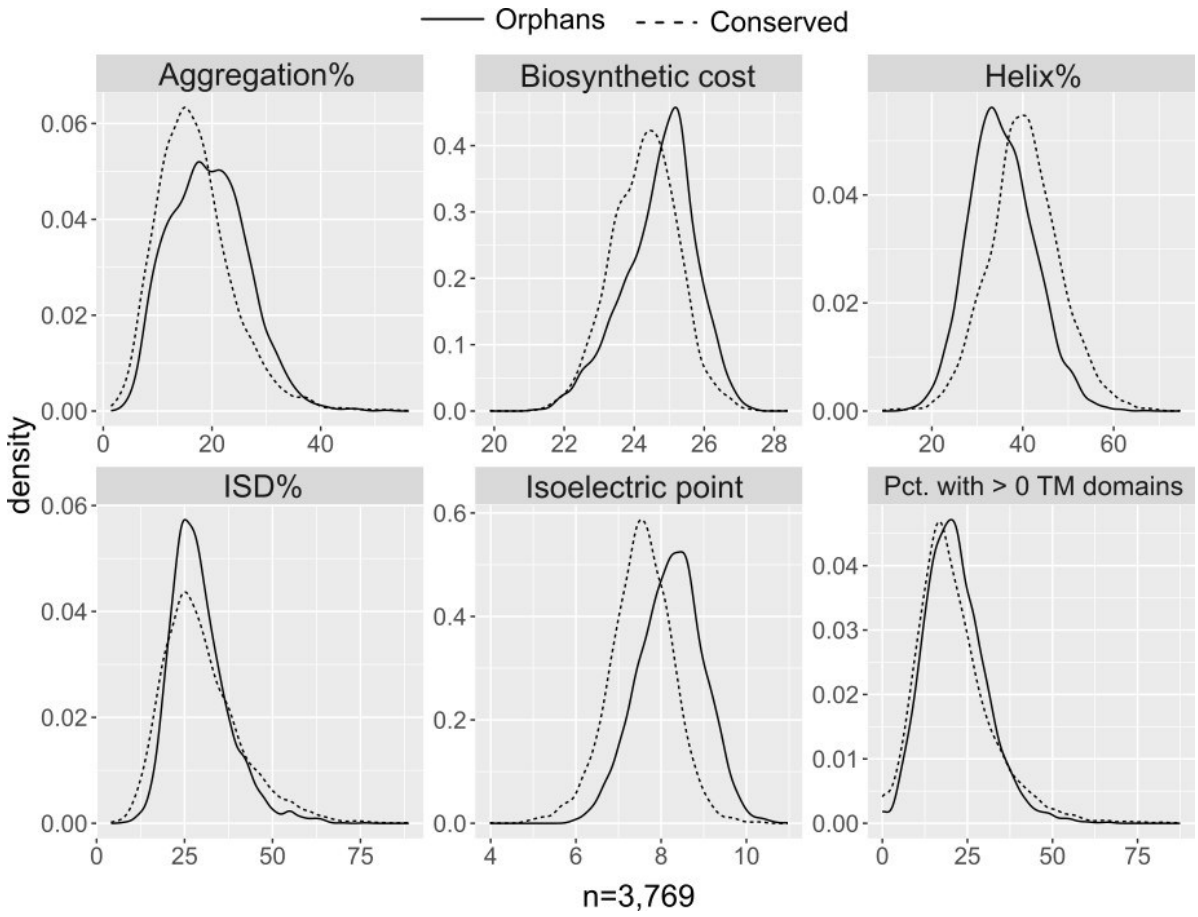
1092

1093



Supplementary Figure 8. Comparison of proteins encoded by conserved and orphan genes, in each taxonomic class, in terms of average percentage of proteins predicted to self-aggregate, percentage of protein predicted to be helical, percentage of protein predicted to be disordered, percentage of proteins with at least 1 transmembrane domain, average isoelectric point and average biosynthetic cost.

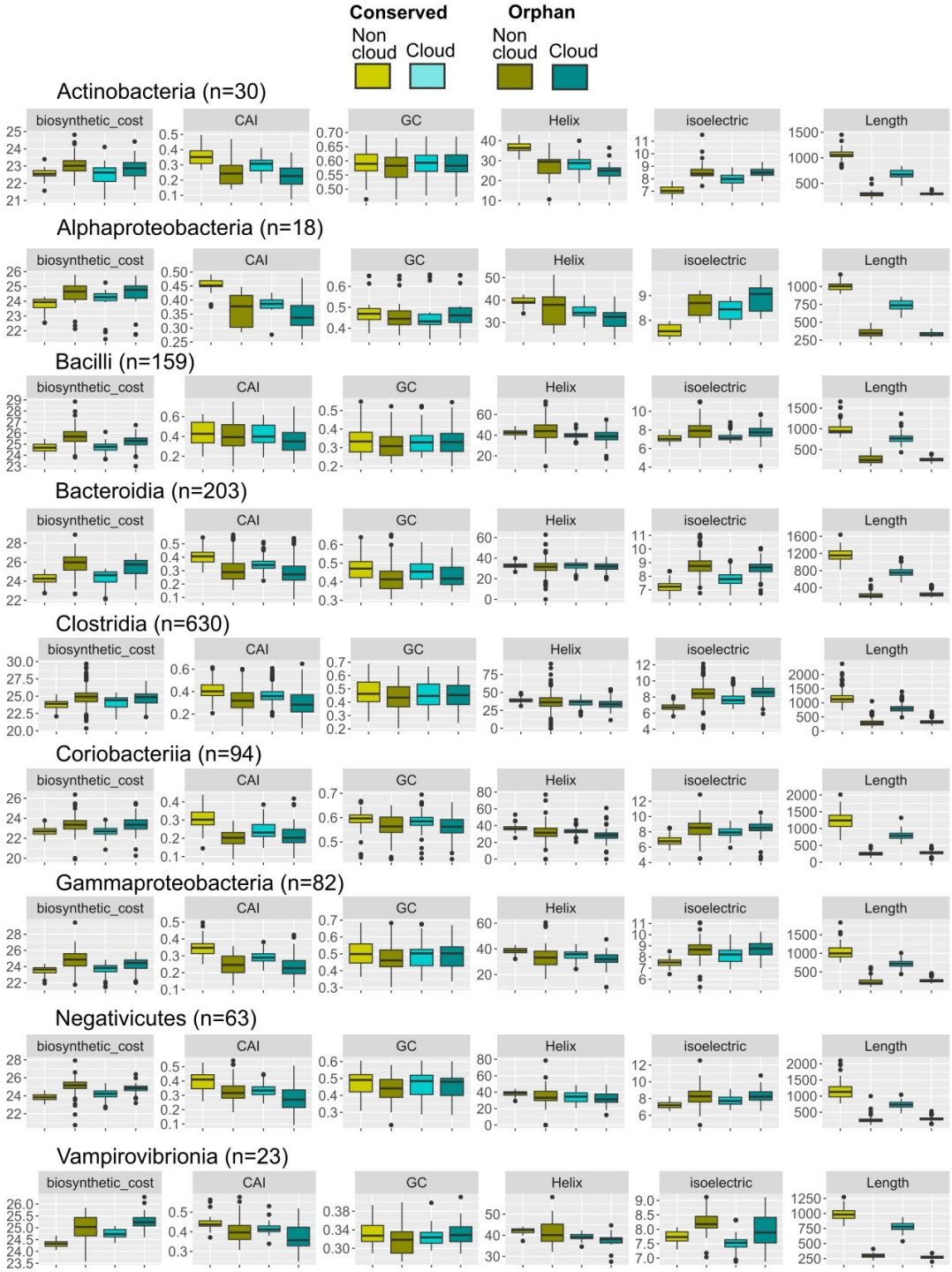
1105



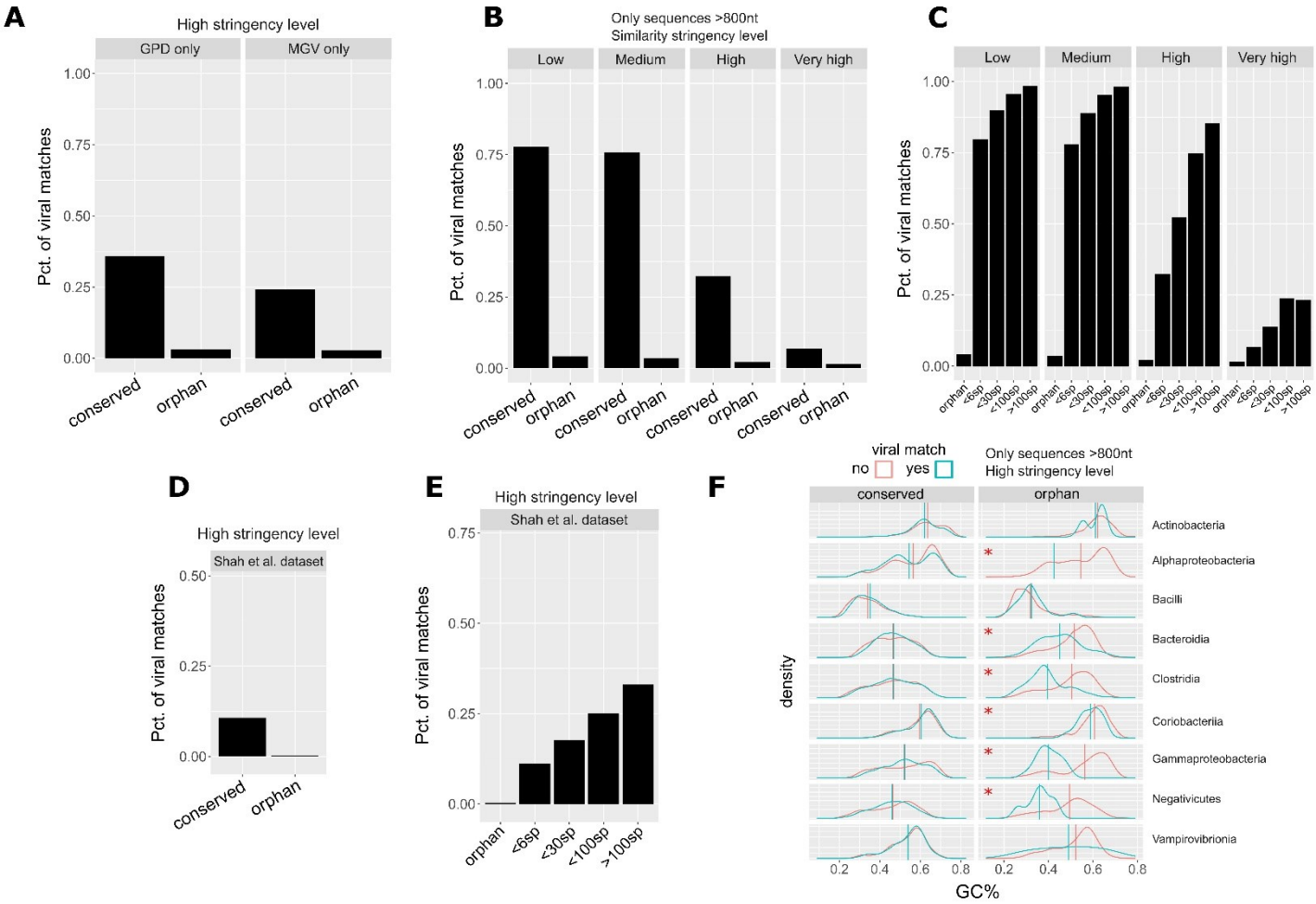
1106

1107 **Supplementary Figure 9**. Comparison of proteins, where conserved proteins have been
1108 subsampled in each species to have similar average lengths as orphan proteins. Comparison
1109 of proteins in terms of average percentage of protein predicted to self-aggregate,
1110 percentage of protein predicted to be helical, percentage of protein predicted to be
1111 disordered, percentage of proteins with at least one transmembrane domain, isoelectric
1112 point and biosynthetic cost. No statistically significant difference exists in intrinsic protein
1113 disorder (Wilcoxon test P=0.6). Difference in percentage with at least one transmembrane
1114 domain is statistically significant ($P=3.7*10^{-7}$) but with a negligible effect size (Cliff's delta
1115 0.07). All other P-value$<2.2*10^{-16}$

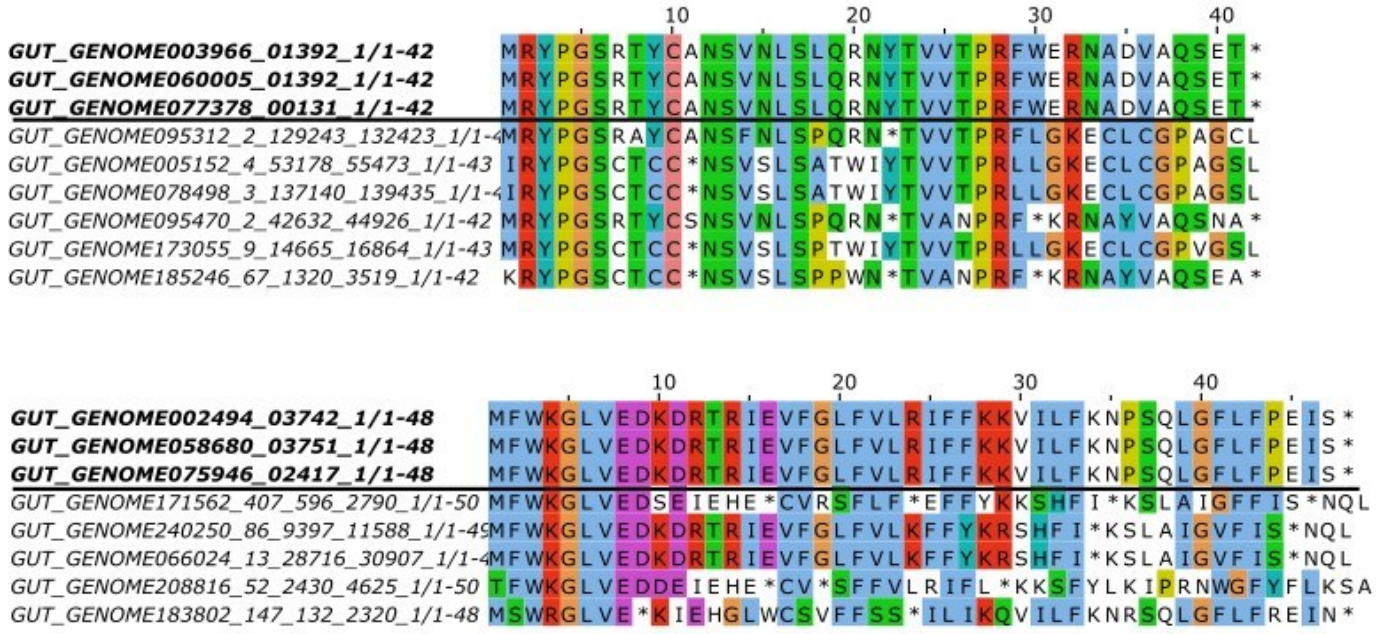
1116



Supplementary Figure 10. Comparison of protein biosynthetic cost, CAI, GC%, percentage of protein predicted to be helical, isoelectric point of a protein, and CDS length in 1,302 species belonging to the nine best represented taxonomic classes, and where each species has at least 10 genomes. Data are shown for orphan and conserved genes split in cloud (present in 10% of species or less) and non-cloud. Each data point represents an average value per species.



Supplementary Figure 11. A: Percentages of orphan and conserved genes with viral matches, at high stringency level, in two separate human gut viral protein databases. **B:** Percentages of orphan and conserved genes longer than 800nt with viral protein matches at different levels of stringency. **C**: Same percentages as B but including all genes and grouped in bins according to the number of species present. **D**: Same as A but using the Shah et al. viral protein dataset. **E**: Same as C but using the Shah et al. dataset at high stringency. **F**: Differences in GC% between genes longer than 800nt, with and without viral match, in orphan and conserved genes. Vertical lines show distribution means. Red asterisk denotes a non-negligible effect size as measured by Cliff's Delta.



Supplementary Figure 12. Protein translations of two de novo candidate examples and their orthologous sequences in outgroup genomes. De novo protein sequences shown in bold. Note that the sequences are not aligned.

1165 # Supplementary Tables
1166

1167 **Supplementary Table 1**

1168 List of representative de novo candidates.

1169 **Supplementary Table 2**

1170 Results of GO term enrichment analysis over all genes in operon-like arrangements.

1171 **Supplementary Table 3**

1172 Results of GO term enrichment analysis per taxonomic class. Note that rows are not sorted.

1173 **Supplementary Table 4**

1174 Representative Clostridia orphans in spore germination-associated operon-like
1175 arrangements.