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, 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 3 million orphans in 4,644 species which lack similarity to other prokaryotes. 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 inflow of mostly transient genes into the cloud fraction 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 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 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. Genes without detectable similarity to genes outside of a particular group have been termed lineage-specific genes, taxonomically restricted genes, or orphan genes (short: orphans)[5–7]. Although, orphan genes can be defined at different taxonomic levels, we are here interested in genes that are restricted to a particular species and will use the terms species-specific gene and orphan interchangeably. Orphan genes cannot be explained by the insufficient sequencing of homologous sequences, since each genome typically contains additional genes that cannot be found in other genomes, as for example observed in *Escherichia coli*[8]. Also in *E. coli*, orphan genes have been found to be narrowly distributed within the species[8]; 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[9].

Orphan genes have been mainly studied in eukaryotes, where they are associated with organismal novelties and species-specific traits[10], and are generally thought to be important for adaptation[11,12]. Research in eukaryotes is starting to paint a coherent picture of the evolutionary origins of orphan genes[6,13,14]. We now know that they can arise entirely "de novo", from genomic sequences that are non-coding/non-genic[15–17] or from extensive sequence divergence beyond recognition[18,19]. Orphan genes can also result from a combination of divergence, de novo emergence, and reuse of parts of existing protein-coding genes in alternative reading frames[14,20].

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[21]. 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[21] (Figure 1). While it has been 20 years since this hypothesis was proposed, evidence to support it is limited[22,23] and other studies failed to find equally convincing traces of the viral origin of orphans[24,25].

An alternative explanation to orphans being of foreign origin is that they are of native origin[21]. 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[8,26]. Additionally, 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[27,28]. As in eukaryotes, pervasive transcription and translation occurs in prokaryotes and may generate raw material that natural selection can shape into a functional protein[29–31]. Artificial random peptides have been shown to readily evolve rudimentary functionality after a few rounds of selection[32–34], highlighting the potential of functional de novo genes to evolve in natural prokaryotic populations as well. Compared to eukaryotes, the constrained genome size, and the limited non-genic space of prokaryotes impose different constraints[35]. 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[7]. 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[36].

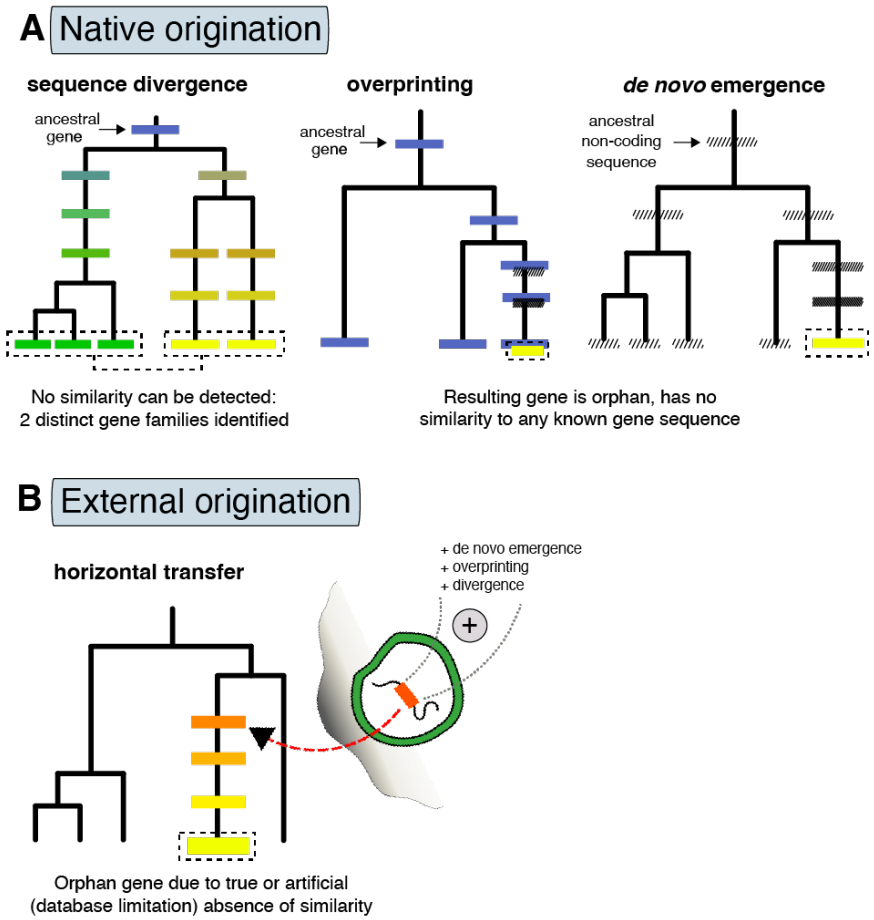
Finally, orphans can also result from annotation errors, since they are typically detected by comparative genomics approaches that can and do produce artefacts. These approaches include the automatic prediction of open reading frames (ORFs) and sequence similarity searches between species and aim to assign true orphans as orphans (i.e., a low false negative rate) while avoiding assigning non-orphans or even non-existing genes as orphans (i.e., a low false positive rate). Careful choices of the computational tools and their parameters are thus critical for a trade-off between false positive and false negative predictions.

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[37–40]. 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[39]. Furthermore, metagenomic data has been used to detect novel protein families with yet unknown functions that occur in multiple species[41]. The function of a large fraction of the microbial protein universe remains unknown, where unknown sequences are often species-specific[42], thus abundant novelty can be discovered among orphan genes. Notably, the term "orphan genes" is occasionally used differently, e.g., as novel uncharacterized protein families that are restricted to a certain environment such as the human gut[43]. In contrast, we here use an evolutionary definition of orphans, i.e., species-specific genes. Whereas evolutionary orphans are also mostly functionally uncharacterized, the restriction to only occur in one single microbial species allows us to study the evolutionary origin of novel functions.

127 One particularly well sampled environment is the human gut, where pangenomes of the
128 residing species have already been constructed[44]. Those pangenomes contain the information
129 whether a gene family is found in one or multiple species, allowing direct orphan assignment.
130 We build on this data – implementing further filtering steps to remove families with homology
131 to other species – to systematically identify and study orphan genes in the human gut
132 microbiome. By looking for patterns in large-scale comparisons, we attempt to disentangle
133 the evolutionary origin of orphan genes (Figure 1). In particular, we aim to understand
134 whether orphans originate mostly externally, as suggested before. Additionally, we aim to find
135 evidence for de novo evolution in prokaryotes. Delineation of the evolutionary context will
136 help us to understand the selection pressures under which orphan genes evolved and hence
137 what they evolved to do.

138

139



140

141

142 **Figure 1:** Evolutionary routes to prokaryotic orphan genes. **A**: Orphans can originate natively
143 through sequence divergence, as well as de novo emergence from entirely non-coding regions
144 (known as de novo emergence) or from coding regions in alternative reading-frames (known
145 as overprinting). Overprinting can also be succeeded by gene duplication, such that both
146 genes are then encoded in different regions. Darker shading represents transition from non-
147 coding state to a coding one. **B:** Orphans can originate in phages, plasmids or integrative

elements by various mechanisms such as de novo birth, overprinting, or divergence, and then be transferred to 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. Additionally, remodeling of existing genes, often in combination with overprinting and sequence divergence, can lead to orphan genes (not shown here).

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[44]. In that study, the Unified Human Gastrointestinal Protein (UHGP) catalog was clustered into homologous families at different sequence similarity cut-offs. 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. We find 3.3 million species-specific proteins (i.e., orphans), grouped into 630,000 families (see Figure 2A and Methods). Unless stated otherwise, our analyses use one representative per cluster, as defined in the UHGP-50 catalog.

We find an average of 4.7 orphans per genome, and an average of 135 orphan genes per species. The latter represents an average of 2.6% of a given species' pangenome, that is, the entire gene repertoire of the species (Figure 2B). Interestingly, the data set contains only 28 Archaea species and they harbor almost twice as many orphans as bacterial ones: 224 orphan families or 6.5% orphans per species, 10.9 orphans per genome. For all species, the number of orphans in a given pangenome correlates strongly with the number of genomes available for that species within the entire 286,997 genome dataset provided by Almeida *et al*. (Spearman's Rho=0.61, $P<2.2*10^{-16}$; Figure 2C), as well as with pangenome size (Rho=0.66, $P<2.2*10^{-16}$, partial correlation controlling for number of genomes Rho=0.39, $P<2.2*10^{-16}$ Supp. Figure 1A), suggesting that each genome introduces novel orphan genes into the pangenome. Indeed, we find that orphan genes are rare within their respective species. Most of them are shared by at most 10% of genomes, i.e., belonging to the cloud fraction of prokaryotic pangenomes (Figure 2D). Orphans are strongly overrepresented among accessory genes compared to conserved, i.e., non-orphan, genes (Supp. Figure 2). While this has been reported before in *E. coli*[8], our analysis demonstrates that this is a general feature of the human gut microbiome and potentially of prokaryotic orphans in general. So, while species-specific genes are indeed widespread and present in 94% of all gut species (and in all species with at least 10 available genomes), they represent only a small fraction of the gene repertoire and they tend to be rare within pangenomes.

We observe variability across taxonomic groups: some bacterial taxonomic classes have markedly higher percentages of orphan genes than others, with Coriobacteriia representing

192 the lower (avg. 0.7%) and Alphaproteobacteria the higher (avg. 5.4%) extremes (see also Supp.
193 Figure 3A). Many factors might influence the difference in orphan numbers across species. For
194 example, the deletion rate was shown to negatively correlate with intergenic space and
195 deletions might also result in the gene loss of orphans; thus, variation across deletion rates
196 might result in differences in orphan numbers[45]. Next, we investigated the impact of
197 metagenome assembled genomes (MAGs) on orphan numbers. Species with a high proportion
198 of MAGs have comparable percentages of orphans as species with many isolate genomes
199 (Supp. Figure 1B), and species with lower quality assemblies show no trend towards more
200 orphans (Rho=-0.04, P=0.0024; Supp. Figure 1C). This suggests that the estimated orphan
201 numbers are not driven by assembly artefacts.

202 In addition to assembly artefacts, annotation artefacts might contribute to the orphan
203 numbers. Some orphans could represent entirely spurious open reading frames (ORFs),
204 because they are annotated as protein-coding on the sole basis of computational predictions,
205 because they lack homologues in other species and because they are rare within their
206 respective species. To test this, we looked for evidence of protein-coding negative selection,
207 the most robust evidence for functionality. We estimated ratios of non-synonymous (dN) and
208 synonymous (dS) polymorphisms from intraspecific multiple sequence alignments for each of
209 the 154,650 orphan families with at least two non-identical sequences. Due to the high degree
210 of genetic similarity, a meaningful dN/dS ratio could be obtained for only 28,659 of them (see
211 Methods). The distribution shows that the peak is at 1, with a visible second peak at 2 (Fig.
212 2E). Importantly, the left side of the distribution, encompassing negatively selected genes,
213 falls of much slower than the right side, resulting in a tail of negatively selected genes (3,928
214 orphans with dN/dS <0.3) which is not observed for positively selected genes. Note that the
215 power to detect selection is limited within the same species[46], thus our analysis might even
216 miss to detect genes under protein-coding negative selection. Similar results were obtained
217 when calculating a similar metric (Supp. Figure 3B). Notably, we found only negligible
218 correlation between dN/dS and CDS length (Rho=0.115, $P<2.2*10^{-16}$). Orphans with a dN/dS <
219 0.3 (n=3,928) had almost exactly the same median and average CDS length as orphans with a
220 dN/dS between 0.9 and 1.1 (n=2,278): median 345nt vs 342nt, average 433nt vs. 422nt.

221

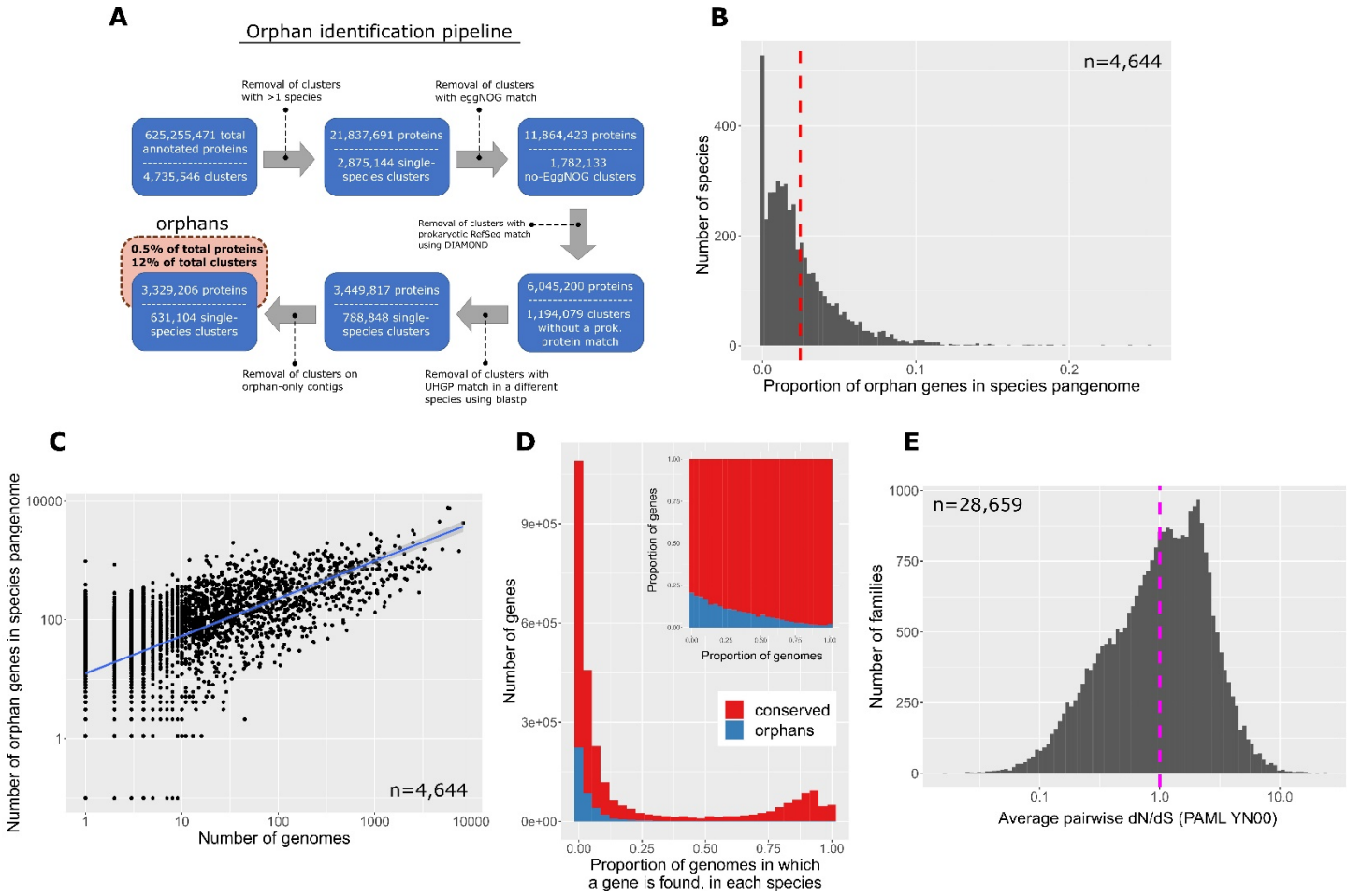


Figure 2: Large scale-identification of orphans in the human gut microbiome. **A:** The main steps of the computational workflow of similarity searches. **B**: Distribution of the proportion of orphan genes in a given species' pangenome. Red dashed line denotes distribution mean. **C**: Number of orphan 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 proportion in each bin. **E:** Distribution of average pairwise omega (dN/dS) per orphan family (see Methods for details). Dashed line marks dN/dS=1 which corresponds to neutral evolution.

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[6,25,47]. 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.

242 As expected, orphan genes across every taxonomic class are significantly shorter than
243 conserved genes, at about a third of their average length (Figure 3A). There is only a weak
244 correlation between average orphan and conserved length per species (Rho=0.14, $P<2.2*10^{-}$
245 ^{16}). Average GC content of orphan genes in a given species is strikingly similar (avg. 0.45 vs.
246 0.46) to that of conserved genes (Figure 3B), exhibiting an almost one-to-one correlation
247 (Rho=0.96, $P<2.2*10^{-16}$). This also holds when only considering orphans longer than 800nt
248 (Supp. Figure 4A) demonstrating that this similarity is not explained by the existence of
249 spurious ORFs. We see the same trend when restricting the analysis to Archaea (Supp. Figure
250 4B), as well as when comparing orphan and conserved genes within individual species (Supp.
251 Figure 4C). The same is true for GC in the 3rd synonymous position of codons (GC3s; see Supp.
252 Figure 4D) and for CpG content (Supp. Figure 5). This observation fits the native origination
253 scenario, in which orphan GC content simply reflects the genome's GC content, but, in theory
254 at least, it would also be consistent with external transfer followed by rapid adaptation of the
255 incoming gene's GC content to that of the host genome[48].

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

268 At the protein level too, orphans exhibit distinguishing characteristics (Figure 3D). Compared
269 to conserved genes, they are, on average, richer in intrinsically disordered regions, a smaller
270 percentage of their sequence is found in helical regions, a smaller percentage of them contains
271 transmembrane (TM) domains, they are more aggregation prone, have higher biosynthetic
272 cost, and a higher isoelectric point (the latter was also observed for de novo genes in yeast
273 and fly[49,50]). All these differences hold when considering each taxonomic class individually
274 (Supp. Figure 8). However, when controlling for the length differential between orphans and
275 conserved genes, intrinsic disorder and TM domains are no longer significantly different
276 (Supp. Figure 9). Note that novel genes in membrane-related roles have previously been
277 suggested for eukaryotes such as budding yeast[51,52]and for prokaryotes[53]. Here we find that,
278 even though orphans have similar nucleotide composition to conserved genes on average
279 (and even more so when controlling for length differences) they differ in key properties, such
280 as their length, codon usage, and some protein structural features.

281 Since we observed before that most orphans are cloud genes, i.e., rare in the species (Figure
282 2D), we next asked if cloud orphans differ from non-cloud orphans. Indeed, we find that some
283 properties differ, e.g., cloud orphans have lower CAI, higher GC, and fewer helices than
284 remaining orphans (Figure 3E). Surprisingly, when comparing cloud genes to remaining genes,
285 the observed trend for GC differs between orphans and conserved genes: whereas conserved
286 cloud genes have a lower GC than remaining conserved genes, orphan cloud genes have
287 higher GC than the remaining cloud genes. Furthermore, the GC of orphan cloud genes is

288 higher than the GC of conserved cloud genes. These trends also hold when analyzing
289 taxonomic classes separately (Supp. Figure 10) and when zooming in individual species (Supp.
290 Figure 4C). Although cloud genes are still quite enigmatic, they could be explained by a
291 continuous inflow of genes by horizontal gene transfer. However, HGT from other prokaryotes
292 would result in conserved cloud genes. Due to the striking difference in the trends for GC, we
293 believe that orphan cloud genes have different evolutionary origins than conserved cloud
294 genes. Interestingly, orphan non-cloud genes have low GC. Potential explanations are that
295 they originate de novo from noncoding regions that typically have lower GC than coding
296 regions or due to rapid divergence, which could lead to the accumulation of A and T[54]. In a
297 previous analysis of Salmonella, it has been found that genes with conserved gene order in *E.*
298 *coli* have higher GC than the remaining genes, where the genes without homologs have even
299 lower GC than the ones with homologs in another genomic context[55]. This trend has been
300 attributed to rapid divergence and transcriptional silencing by H-NS, which might also apply
301 to the orphan genes with low GC.

302

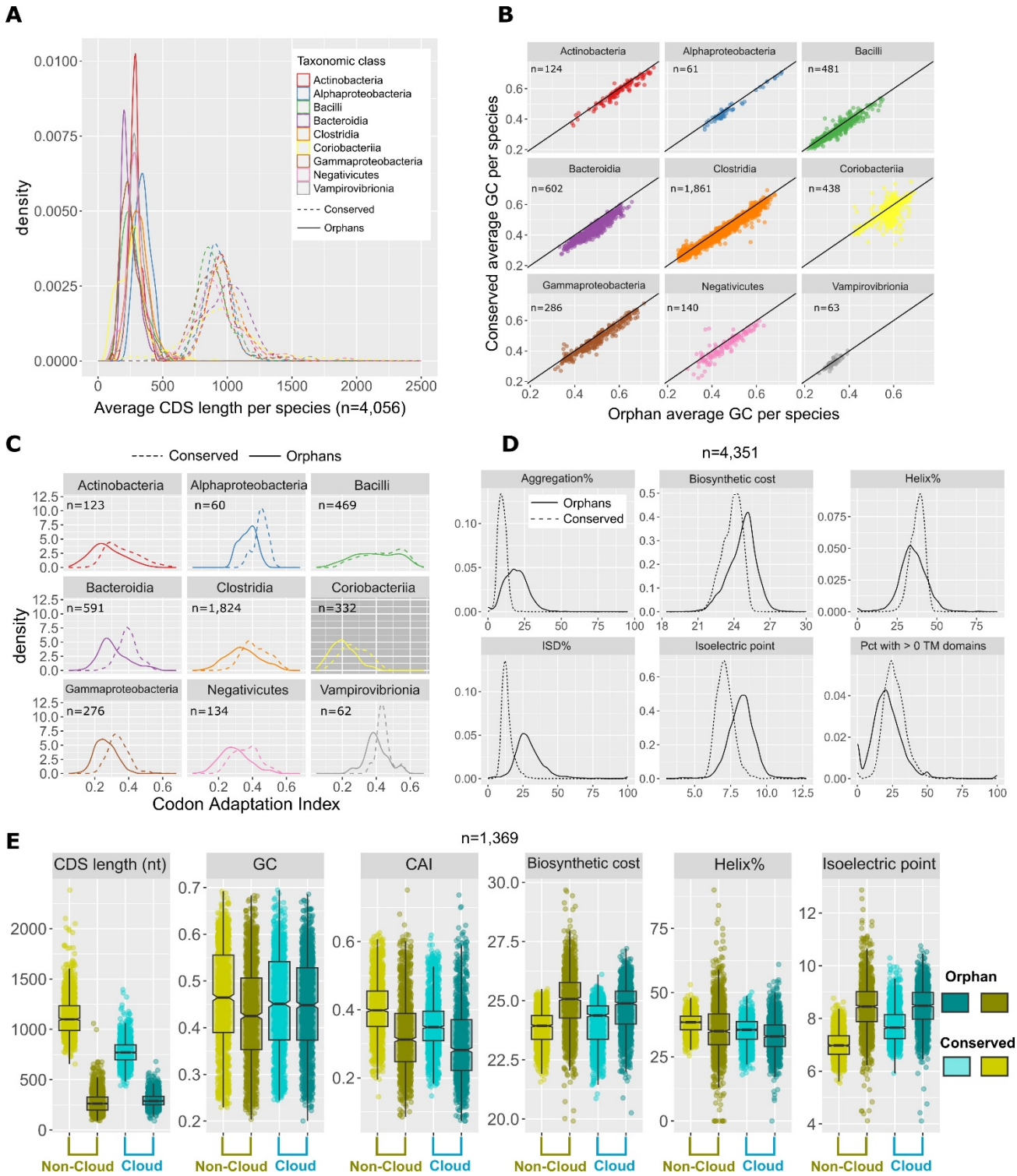


Figure 3: Properties of orphans compared to conserved genes. **A:** Density plots of the average length of orphan and conserved CDSs in each species, grouped by each of the 9 best represented taxonomic classes of our dataset. Only classes with at least 50 species are included. **B**: Average GC of orphan 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

311 P-values < 10^{-4}. **D**: Comparison of proteins encoded by conserved and orphan genes, in terms
312 of average percentage of protein predicted to self-aggregate, biosynthetic cost, percentage of
313 protein predicted to be helical, percentage of protein predicted to be disordered (ISD),
314 isoelectric point of a protein, and percentage of proteins with at least 1 transmembrane
315 domain. All Wilcoxon test P-values < $2.2*10^{-16}$. **E**: Comparison of some of the properties found
316 in panels A-D in 1,369 species with at least 10 genomes where data orphan and conserved
317 gene categories are split in cloud (present in 10% of species or less) and non-cloud. All plots
318 show points or distributions of average values per species.

319

320 Orphans of likely phage origin have distinct characteristics

321

322 Orphans are defined here as species-specific genes with respect to other prokaryotes,
323 however it is possible that they are the result of recent horizontal gene transfer from
324 bacteriophages. To identify such transferred genes, we conducted similarity searches against
325 two recently published phage protein databases, MGV[56] and GPD[57], that are partially
326 overlapping but mostly complementary in terms of protein sequence content.

327 Orphan genes are strongly depleted in similarity matches to phage proteins relative to
328 conserved genes, ranging from 3 to 6-fold using different parameters and cut-offs (Figure 4A;
329 Methods). At low stringency, 5.1% of orphans have similarity to viral sequences, compared to
330 63% of conserved genes. Orphans are also depleted, when separately searching against either
331 MGV or GPD (Supp. Figure 11A) and when excluding short sequences (<800nt) for which
332 similarity searches often fail (Supp. Figure 11B). Interestingly, binning conserved genes
333 according to the number of species in which they are found (a proxy of their evolutionary age)
334 reveals a pattern of increasing proportion of genes with phage similarity (Figure 4B),
335 regardless of the stringency criteria applied (Supp. Figure 11C). This pattern is inconsistent
336 with a scenario under which most orphan genes arrive through HGT from bacteriophages,
337 which we would expect to either produce the inverse trend, if only a fraction of HGT is
338 ultimately retained, or no trend at all in the case of a constant rate of transfer and retention.

339 As both GPD and MGV databases have been assembled with the use of automated annotation
340 pipelines, it is possible that they may contain bacterial contigs that are falsely classified as
341 coming from phages. To control for such a bias, we confirmed our findings by searching a third
342 phage genome database[58] which is smaller but manually curated and was built from samples
343 taken from infant guts. We observed similar patterns as for the larger databases confirming
344 that the proportion of orphans (and more generally, less conserved genes) with similarity to
345 phage proteins is considerably lower than that of widely conserved genes (Supp. Figures 11D
346 and 11E). The overall proportion of matches in all gene groups is lower compared to GPD and
347 MGV, but that is expected given the difference in size (10K phage OTUs vs. 190K and 142K in
348 MGV and GPD respectively).

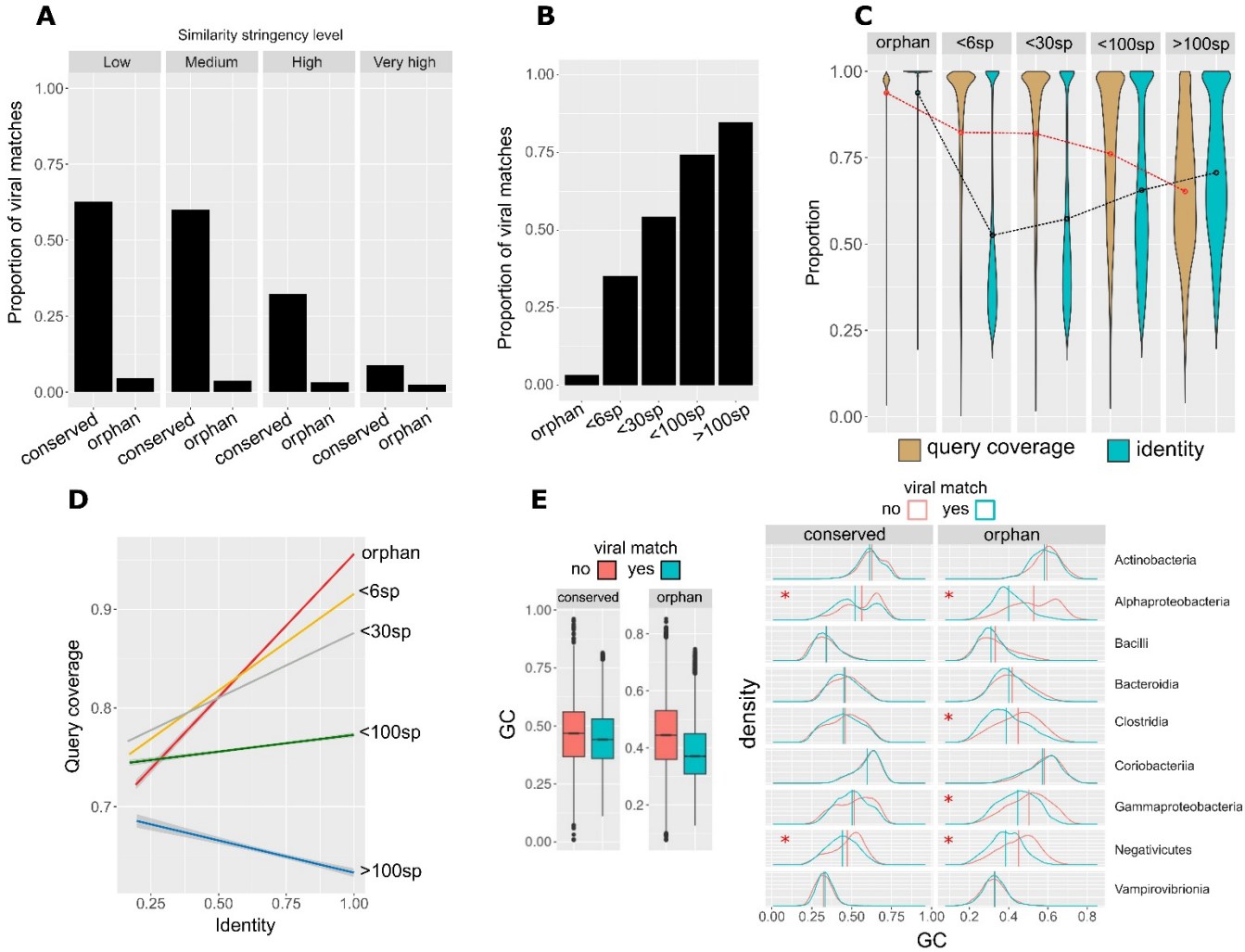
349 The trend of increasing percentage of phage matches with increasing conservation across
350 species prompted us to investigate further. To dissect the evolutionary dynamics, we plotted,
351 across age categories, the query coverage and percent identity of the best phage match per
352 protein as defined by E-value, without filtering any high-identity hits (Figure 4C). This revealed
353 that matches to orphan genes are very close to 100% identity and coverage with both
354 averages decreasing with age. Yet, while the coverage distribution shows a clear downward

355 trend, the identity distribution becomes bimodal with age, with one peak staying near 100%
356 and another one initially appearing at around 35%, then increasing. Furthermore, while
357 identity and coverage are strongly correlated for viral hits of orphans, this correlation
358 gradually decreases with age (Figure 4D). One explanation for these trends is that a subset of
359 genes has only recently been transferred between phages and bacteria, while another subset
360 consists of older transfers that have undergone divergence, truncation, or both.

361 We previously established that orphans mirror conserved genes' averages in crucial aspects
362 such as nucleotide composition. Given the evidence that a small percentage of orphans
363 appears to have been recently transferred from phages, we might expect that they exhibit
364 distinguishing signatures at the level of their nucleotide composition. Indeed, we find that
365 overall, orphans with similarity to phage proteins have markedly lower GC content than those
366 without (Figure 4E), congruent with what is known for phage GC% relative to their hosts[59].
367 Most strikingly, the effect of this difference is practically negligible when looking at conserved
368 genes, albeit staying statistically significant due to the large sample size. The findings also hold
369 when only genes >800nt are considered (Supp. Figure 11F). We thus conclude that there is a
370 low proportion of orphans with hits to phages that appear distinct from the remaining
371 orphans.

372 Additionally, we computationally detected prophage regions in all genomes using Phigaro[60].
373 We find that a very low proportion of orphan genes reside within prophages
374 (13,174/3,329,206; 0.3957%), which is slightly higher than the overall proportion of prophage
375 genes among all genes (2,029,792/625,255,471; 0.3246%). Thus, although phage origin might
376 play a role, this can only explain a small fraction of orphans.

377



378

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

395

396 ## A subset of orphans shows evidence of native origination
397

398 For some orphans, it might be possible to trace the origin within their native genome, either
399 to existing genes or to non-coding regions. We first searched for orphan genes that might have
400 originated via utilization of an alternative frame following duplication or horizontal gene
401 transfer[61] from closely related genomes. To this end, we performed similarity searches of six-
402 frame translations of all orphan coding sequences against six-frame translations of all coding-
403 sequences of their native genome. We found that only a small percentage, 1.34% of orphans
404 (8,477 representative orphans), have at least one significant match to a conserved gene (i.e.
405 non-orphan) excluding gene pairs with overlap. 75% of matches are on the reverse strand of
406 the conserved gene, suggesting a more frequent exploration of the reverse complement
407 sequence space rather than the alternative reading frames of the same strand. Out of all
408 orphans matching conserved genes, 64% (5,397) match a single conserved gene, 26% (2,238)
409 match two conserved genes and the rest (10%) match three or more. Orphans matching more
410 than one conserved gene correspond to chimeras similar to what was found for *E. coli*[62].

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

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

427 Closely related species are more likely to share genomic synteny and hence to satisfy our
428 conservative criteria. The de novo candidates originate from 718 different species, and these
429 have shorter terminal branches in the tree compared to the species of the remaining orphans
430 (Figure 5C; means of 0.03 and 0.071 substitutions per site; Wilcoxon test P-value < $2.2*10^{-16}$;
431 Cliff's Delta 0.28). Thus, they represent the evolutionarily younger end of the spectrum of all
432 de novo gene candidates, as expected. Comparisons of gene and protein properties support
433 this view (Figure 5C). De novo candidates are overall shorter (Delta=0.28; P<$2.2*10^{-16}$), have
434 lower CAI (Delta=0.18; P=$3.64*10^{-11}$) and GC% (Delta=0.1; P=$8.5*10^{-9}$), have slightly higher
435 isoelectric point (Delta=0.15; P=10^{-13}) and biosynthetic cost (Delta=0.13; P<$2.2*10^{-16}$). Since
436 mutation is biased towards AT, most bacteria have lower GC in noncoding regions[54], which
437 can explain the lower GC in de novo candidates compared to remaining orphans. Additionally,
438 de novo candidates are depleted in matches to viral proteins compared to the rest of orphans
439 (Fisher's test odds-ratio: 40.9, P-value = $3.5*10^{-16}$). Two examples of alignments of de novo
440 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, together with two additional examples and their translations (selected for visualization based on their short alignment length). 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.

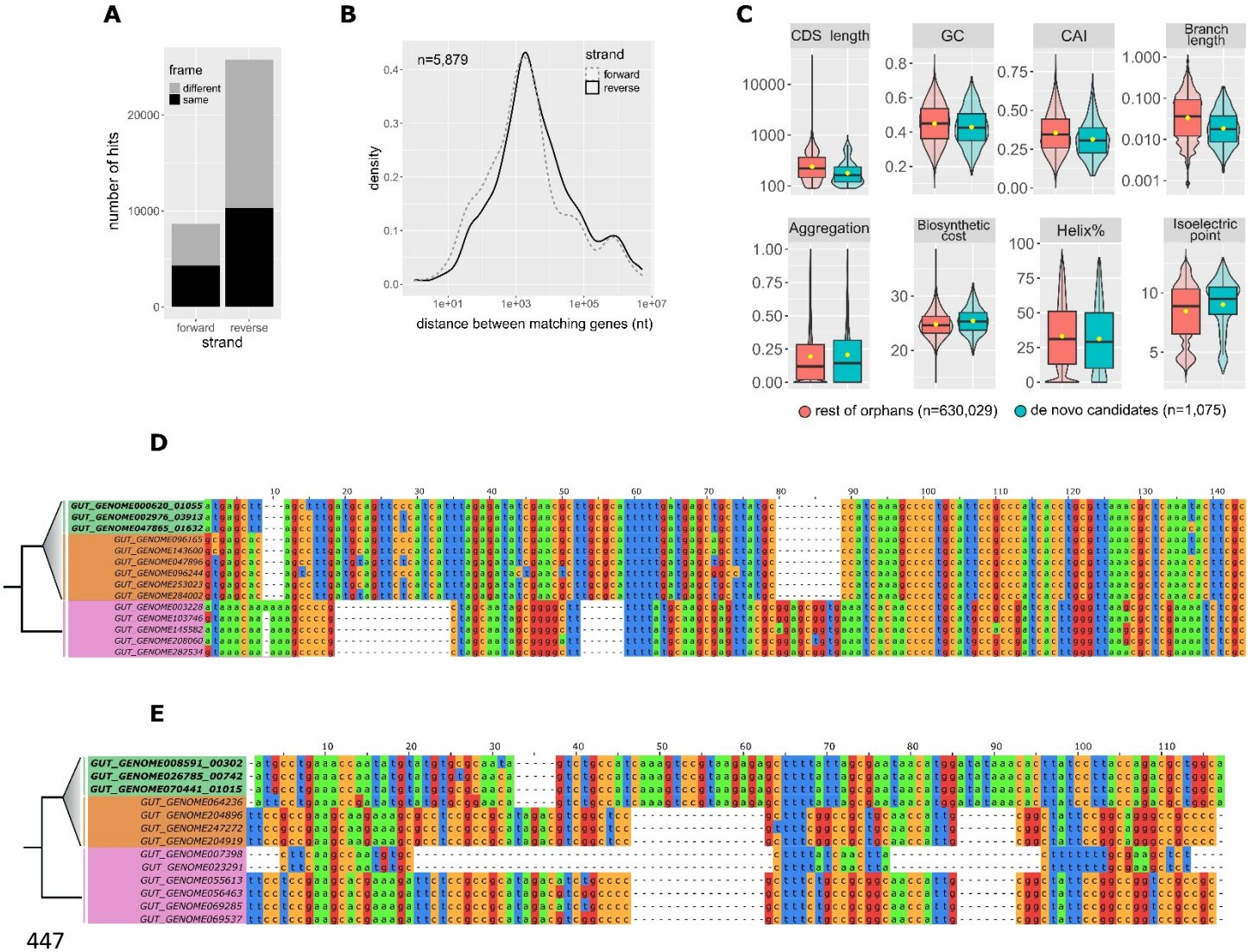


Figure 5. Evidence for native origination of orphan genes. **A**: distribution of frame and strand of similarity matches between orphan genes and conserved genes within each genome. **B**: Distribution of genomic distances between matching orphan and conserved genes, when these are found on the same contig. **C**: Comparisons of de novo candidates to remaining orphans: Distributions of Codon Adaptation Index (CAI), terminal species-level branch length, proportion of protein sequence predicted to be aggregation-prone, biosynthetic cost, CDS length, GC, protein percentage estimated to fold into a helix, and isoelectric point. Details of statistical comparisons can be found in the main text. Helix% comparison Wilcoxon test P-value = 0.0064. Aggregation% P-value = 0.26. Branch length and CDS length are shown on log scale for visibility. Outliers are not shown as points but are represented by the violin plot range. **D**: Alignment of a de novo candidate gene from *Hafnia paralvei* (green, three non-identical members of an eight-member family are shown) to its orthologous regions in

460 genomes of the same species that do not have annotated homologues (orange) and
461 genomes of its closest outgroup species (pink). Identical sequences have been removed from
462 the alignment. The average pairwise dN/dS for this family is 2.5. **E**: Same as D, but for
463 another candidate from an unnamed species of the order Opitutales. Three non-identical
464 members of a family of six members are shown. Five orthologous loci (i.e. from five
465 genomes) that contain a long insertion have been removed for visual purposes, and an
466 additional 37 orthologous loci are not shown again for visual purposes (their sequences are
467 almost identical to those shown in the figure). The average pairwise dN/dS for this family is
468 3.4. Alignments were generated with MAFFT[63] and visualized with JalView[64].

469

470 Operon-like arrangements provide functional hints for some orphans

471

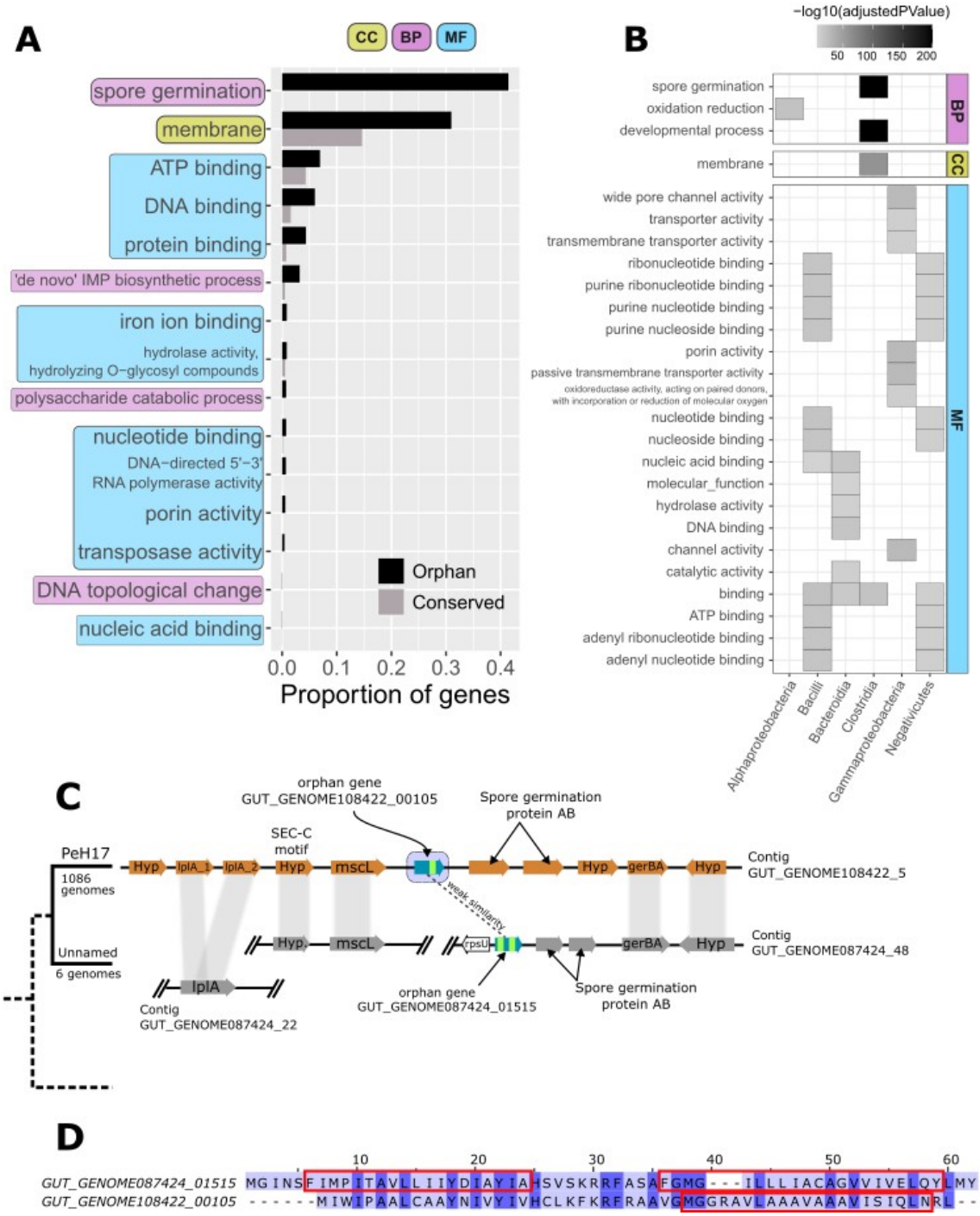
472 Assigning a function to orphan genes is notoriously difficult. For instance, we searched a
473 database from a previous study that established mutant phenotypes for thousands of
474 prokaryotic genes coming from 46 strains from 36 different species[65] including many from our
475 dataset (see Methods), but found no phenotypes for any of our orphan genes.

476 Even if no experimentally derived functional information has been determined for an orphan
477 gene, its genomic context can still offer some clues as to what its role might be in the cell[39,66].
478 We thus examined the neighbors of all orphan genes (not only representative ones) in all
479 genomes harboring orphan genes to identify cases of operon-like arrangements where one
480 functional term is present in multiple neighbors (see Methods). We identified 1,905 orphans
481 in operon-like arrangements, belonging to 405 distinct protein families. For these, at least 3
482 out of the 6 closest neighbors (i.e. 3 downstream and 3 upstream) have a common Gene
483 Ontology (GO) term associated. We then looked for enrichment of specific GO terms within
484 the orphans against the background of all genes in operon-like arrangements (n= 4,362,138).
485 The most overrepresented specific terms by gene counts are shown in Figure 6A. The results
486 of the full GO term enrichment analysis can be found in Supp. Table 2.

487 The term "membrane" was the only cellular component that was found in at least 3 neighbors
488 of orphans and it was one of the top enriched terms overall (GO:0016020; Figure 6A, Supp.
489 Table 2). This is intriguing given previous findings linking TM domains to novel genes[51,53]. A
490 closer look showed that the membrane enrichment was due almost exclusively to orphans
491 from the class Clostridia. Thus, to account for taxonomic specificities, we repeated the analysis
492 separately for each of the 9 best represented taxonomic classes and found limited overlap
493 between the GO terms enriched within each class (Figure 6B; Supp. Table 3).

494 At the level of biological processes, one of the strongest enrichments was “spore germination”
495 and this enrichment too is specific to Clostridia (Figure 6B). Of the seven Clostridia orphans
496 related to this term, four had at least one TM domain (57%) including three that are also
497 annotated as “membrane”, which is expected given that the cell membrane is central to the
498 process of germination. We examined in detail one large family with “spore germination”
499 annotation (see Methods). A closer look at the wider genomic region of this family showed
500 clear evidence of syntenic and functional conservation with its closest neighboring species
501 (unnamed species, see Figure 6C). Furthermore, another orphan gene, also encoding TM
502 domains, is found at the predicted orthologous locus in the closest species. Alignment of these
503 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[18,19]. 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.



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

536

537 # Discussion

538

539 Orphan genes are ubiquitous. Across the tree of life, species-specific genes may be seen as
540 one end of a taxonomic gradient, with universally conserved genes at the other end and genes
541 restricted to smaller taxonomic groups as intermediates[6]. Some recent large-scale studies
542 have focused on novel prokaryotic gene families present in more than a single species and
543 have firmly demonstrated their functional relevance and hinted towards crucial habitat-
544 specific roles[39–41,53]. The question of the evolutionary origin of novel prokaryotic families
545 however was mostly left unexplored. This question is best approached by focusing on orphan
546 genes, which are species-specific and thus represent the most-recently evolved form of
547 genetic novelty.

548 Our results show that prokaryotic orphans from the human gut microbiome emerge due to
549 multiple routes. We find that a low proportion of orphans has matches in viral proteins.
550 Thanks to the extensive, high-quality sampling of both the prokaryotic and the viral diversity
551 of the human gut, we can be relatively confident that our findings are not significantly biased
552 by limitations of the dataset. Additionally, orphans with viral matches have a distinct
553 composition compared to the remaining orphans (Figure 4E), which is inconsistent with the
554 external origination scenario for most orphans. Conversely, there is a low proportion of genes,
555 where we can recover evidence for de novo origination from non-coding sequences. These
556 orphans also differ from the remaining orphans by being shorter, having lower GC%, and lower

557 CAI (Figure 5C), although these differences likely also reflect the fact that we can only detect
558 the very youngest of the de novo originated orphans. Nonetheless, these results are also
559 inconsistent with de novo origination from non-coding sequences as a dominant source
560 behind orphan genes. This leaves the native processes of sequence divergence, overprinting,
561 and remodeling as the next best candidates to explain orphan origination.

562 Our analyses go deeper than previous efforts to understand the origin of orphan genes in
563 prokaryotes. Our results are consistent with a previous study where a low fraction of viral
564 matches have been found[24]. In contrast, earlier inquiries into the origin of bacterial orphans
565 from Daubin and Ochman[21,22] investigated compositional similarity between orphans in *E. coli*
566 and bacteriophage sequences known at the time and proposed acquisition from
567 bacteriophages as a reasonable explanation for orphan origin. Our analysis of some of the
568 same features (GC content, CpG content), across thousands of species, for which data was not
569 available at the time, shows instead that orphans closely match their native genomes'
570 composition. Cortez et al.[23] proposed in 2009 that integrative elements are the major source
571 of ORFans in an analysis of 119 prokaryotic genomes, based on their identification of clusters
572 of genes of atypical composition. But their evidence is tangential and relies on the fact that
573 some of the identified ORFans (39% of 8,428 genes) are found in such clusters of genes with
574 atypical composition, likely to be derived from integrative elements. Furthermore, results on
575 the crucial point of the composition of ORFans themselves are missing from this study. As
576 there is little overlap between the 119 species surveyed by Cortez et al. (entire prokaryotic
577 diversity) and here (>4,000 species of the human gut microbiome), it might be the case that
578 external origination of orphans is more prevalent in species found in different environments.

579 If most of orphan genes analyzed here were indeed "grown locally", by what mechanisms did
580 that happen? Extensive sequence divergence is surely one of them. In a previous study,
581 extensive remote homology searches of novel genes in metagenomes using Hidden Markov
582 Model profiles retrieved significant similarity for ~15% of them[26]. Even more sensitive
583 methods might reveal the true percentage to be even higher. In eukaryotes, an approach that
584 harnesses conserved micro-synteny found that on average about one third of genes without
585 similarity might have simply diverged beyond recognition[18]. While low syntenic conservation
586 prohibits meaningful application of this method to our dataset, our example in Figure 6C and
587 D clearly shows that syntenic conservation can occur. Divergence of gene duplicates within
588 the same genome can result in additional orphans, either by repurposing the same reading
589 frame, or alternative ones. The percentage of orphans sharing similarity to conserved genes
590 from the same genome that we report here is likely an underestimate, as divergence beyond
591 detectable similarity is potentially at play at this level as well. A crucial question, outside the
592 scope of this study, is how often this absence of statistically significant similarity should be
593 taken as absence of functional similarity. The alternatives to divergence are de novo gene
594 birth and overprinting. Have these mechanisms been neglected in prokaryotes? In our view,
595 at least within the environment of the human gut, this seems to be the case. Notably, the
596 human gut environment is well-sampled and allows us to detect orphan genes with
597 confidence. Nevertheless, this does not give us a good representation of the entire prokaryotic
598 diversity. For example, gut bacteria are generally anaerobic and strongly associated with the
599 eukaryotic hosts. Thus, their ecology is very different from environmental bacteria and the
600 extent to which our findings generalize to other environments is currently unclear. Future
601 studies that leverage the ever-increasing global metagenomic sequences will be useful to
602 tackle that question. Additionally, experimental evidence shows that the emergence of
603 functional proteins via de novo gene evolution from noncoding sequences can provide

604 adaptation in *Escherichia coli*[67,68]. Thus, de novo evolution may be more plausible, and
605 prevalent, than usually assumed.

606 A mostly native orphan origin does not diminish the role of HGT in prokaryotic evolution.
607 Indeed, when one looks at how the percentage of viral matches increases as the number of
608 prokaryotic species in a family increases (Figure 4), it is evident that novel genes can spread
609 between phages and bacteria. It has been known for a long time that phages also invent novel
610 genes[69–71]. An intriguing question for future studies then is whether there is any qualitative or
611 quantitative difference between the processes that surely occur in bacteria and phages, or
612 whether both simply serve as input to a common pool of novel sequences that is accessible
613 via horizontal gene transfer and is maintained under selective pressures or neutrally[72,73].

614 Another important but difficult question is whether our estimates of the number of orphan
615 genes is a good approximation of the true number of entirely novel, *functional* coding
616 sequences unique to each species. Given that orphans are overwhelmingly short, it is
617 reasonable to assume that some of them are prediction artefacts[8]. This is in our view the most
618 important limitation of the present study. Several steps have been carried out to mitigate this
619 limitation: We have relied on preexisting annotation consistent for all genomes. Note that
620 using consistent, robust annotations is an important measure against spurious results when it
621 comes to estimating species-specific genes[74]. Furthermore, we have considered any sequence
622 with a EggNog match as conserved, but it is possible that an evolutionarily constrained gene
623 might converge towards an existing domain or motif. Finally, we have excluded any recently
624 emerged gene from orphans that was transferred to another species shortly after having
625 emerged. Notably, novel small ORFs and overlapping reading frames have been omitted in the
626 automatic annotation. Nevertheless, it is plausible that some homologues of the shorter
627 orphan genes could exist in other species but might have been missed by the annotation
628 pipeline due to their short length, which would push orphan numbers higher. For all these
629 reasons, the presented orphans are expected to contain both false positives and false
630 negatives.

631 Additionally, one can assume that such mis-assignments are even higher for metagenome-
632 assembled genomes (MAGs) which also contribute to the UHGP. First, assembly and binning
633 might miss to reconstruct particular species, especially when they are highly diverse. Such
634 missing species could result in false positives orphans in related species. Nevertheless, on
635 average 86% of the reads from a human gut metagenome map to the genomes that are the
636 raw material of the UHGP[44], suggesting that the species diversity in the human gut is well
637 covered. Second, errors at the binning step could lead to the inclusion of wrong genes into
638 MAGs, which then appear to be orphan. To mitigate this effect, we have excluded contigs
639 without conserved genes. Third, MAGs can be incomplete and are known to be biased against
640 mobile elements[75]. Note that completeness and contamination cutoffs have been applied to
641 the MAGs included in the UHGP[44]; however these estimates are based on marker genes and
642 can underestimate the actual values[76,77]. The bias against mobile elements particularly
643 impacts the accessory genome, where also orphan genes can be found, potentially leading to
644 a bias against orphans. Fourth, misassembly or incomplete contigs can lead to the prediction
645 of spurious ORFs. We examined the impact of this bias by comparing the orphan numbers for
646 species with different assembly qualities and cannot find an association (Supplementary figure
647 1C). Taken together, the estimated orphan numbers do not seem to be driven by the inclusion
648 of MAGs in the data set.

649 To estimate if orphans are indeed conserved at the protein level, we estimated selection for
650 orphans with some diversity and we find an encouraging number of orphans under negative
651 selection. Those that do not should be viewed with suspicion but drawing strong conclusions
652 on them is inappropriate, first because of the technical limitations of dN/dS at the intraspecific
653 level[46] but also because a recently originated gene would not have had the time to accumulate
654 enough mutations for selection to be detectable. Furthermore, it is difficult to extrapolate the
655 percentage calculated among the small subset of orphans with enough genetic diversity to all
656 the orphans, given that it leads to a bias: orphans that have existed long enough to have the
657 necessary diversity are more likely to be functional than those restricted to only one genome.
658 Yet, conservation across multiple genomes, or even species is not always infallible: for
659 example, conserved gene families have been built which were found to be composed of
660 spurious open reading frames (ORFs) located on the noncoding strand of another conserved
661 gene family[78]. Nevertheless, functional evidence for such "antisense proteins" shows that the
662 noncoding strand also provides potential for evolutionary innovation[79], making it difficult to
663 distinguish functional and non-functional antisense proteins.

664 Our estimate of the number of orphan genes should be viewed with caution and with all the
665 aforementioned caveats in mind. However, the primary focus of our study is not the number
666 of orphans per se, but rather what we can glean about their origins when subjecting them to
667 comparisons. Notably, using a more stringent length cut-off does not impact our main
668 conclusion suggesting that our results are not driven by false positive orphans. In any case,
669 future studies might provide further experimental insights into the expression and function of
670 the orphan genes described here. Approaches incorporating transcriptomics, ribosome
671 profiling, and high-throughput functional assays might be especially informative. To
672 distinguish functional expression from pervasive transcription and translation, analysis of
673 differential expression of orphan genes under different conditions is of particular interest.

674 We have identified more than half a million orphan genes and even if a large part of them
675 turns out to be spurious, we would still be left with thousands of candidates that could be
676 expressed as entirely novel proteins within the boundaries of the human body. It is more and
677 more acknowledged that pangenome diversity is important in the human gut microbiome and
678 might have consequences for human health[80]. Orphan genes within strains of human gut
679 bacteria could perhaps interact with the host, contributing to immune or inflammatory
680 responses and potentially even disease.

681 While such a link of prokaryotic orphan genes to human health can only be speculative, their
682 overall importance for niches-specific adaptations is far from it. Although the cloud of rare
683 genes has been found to be an integral part of all pangenomes, the evolutionary processes
684 that generate and maintain cloud genes are not well understood and explanations range from
685 niche-specific genes to a constant inflow of transient genes[9]. Particularly, a class of transient
686 genes with effectively instantaneous gene replacement rates has been suggested to
687 contribute to the cloud fraction of microbial pangenomes[72]. Here we highlight that orphan
688 genes are an important component of prokaryotic pangenomes, and especially of the cloud
689 fraction, suggesting that they are mostly transient. Thus, most orphans might be removed
690 quickly after their origination, where some of them could prove adaptive and persist in the
691 population, resulting in non-cloud orphans and – after longer evolutionary time – even in
692 conserved genes. To gain a comprehensive picture of prokaryotic evolution, we thus need to
693 understand where orphan genes come from, what specific functions they assume and how
694 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.[44] was parsed and families with members in more than one species were removed. Next, we removed families with a precomputed EggNOG[81] 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[82] 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[83] 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[84] in default mode. Protein self-aggregation was predicted using PASTA[85] 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[86]. Codon Adaptation Index (CAI) was calculated for all sequences using the CAIJava[87] tool with arguments *"-s -i 15 -k 3 -g"*, and for *E. coli* sequences additionally using the *codonw*[88] 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*[88]. CpG dinucleotide frequency was calculated in *R* with the SeqinR[89] package. Biosynthetic cost was calculated by averaging the Akashi and Gojobori[90,91] amino-acid scores for each protein sequence. Isoelectric point was calculated with the *R* package *peptides*[92].

Similarity searches

We downloaded two gut phage protein sequence databases in FASTA format, MGV[56] and GPD[57]. Additionally, protein sequences were extracted from phage genomes assembled by Shah et al.[58] and downloaded from http://copsac.com/earlyvir/f1y/gbks/ in May 2022. Similarity searches against viral protein databases were conducted using DIAMOND *blastp* in 'very sensitive' mode. The criteria applied to define significant hits, in increasing order of

739 stringency, were the following: identity <95% & E-value < 10^{-3} (low), identity <95% & E-value
740 < 10^{-5} (medium), 95%>identity >40% & query coverage % >40% & E-value < 10^{-5} (high), 95% >
741 identity > 60% & query coverage >70% & E-value < 10^{-5} (very high). To identify orphans
742 potentially derived from conserved native genes, we also conducted similarity searches of
743 orphan genes against all annotated CDS in their respective genomes using NCBI's TBLASTX
744 with 80% identity, 70% query coverage and E-value 10^{-5}.

745

746 Genomic and functional features

747 We predicted the presence of prophages in all 286,997 prokaryotic genomes by running
748 Phigaro[60] in default mode. We then counted the number of orphans and non-orphans that are
749 situated within the predicted prophage genomes.

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

766 We sought to identify orphan genes among those prokaryotic genes included in the fitness
767 browser[65]. We downloaded protein sequences of fitness browser proteins as well as raw t-
768 score tables for each species from the fitness browser website. We then used DIAMOND to
769 search for similarity between our orphan proteins and those fitness browser proteins. We
770 considered a hit as significant if it had E<10^{-5}, % identity > 70%, and 80% query coverage. We
771 then filtered matches to retain only those to fitness browser proteins with statistically
772 significant fitness effects, defined as t-score > 4 in at least one condition (as detailed in the
773 original publication). To only consider unambiguous cases, we looked only for orphan genes
774 matching to a unique fitness browser protein.

775

776 Identification of orthologous regions and definition of de novo candidates

777 For each orphan gene, we searched for each orthologous region in two sets of genomes. In
778 every genome of the same species without the orphan gene, and in every genome of its closest
779 neighboring species, as defined by the species-level phylogeny provided by Almeida et al[44]. To
780 identify the orthologous region in each genome, we took the immediate upstream and
781 downstream neighbors of the orphan gene and searched whether a homologue, that is, a
782 member of the same family, exists in that genome. We only considered cases where a single

783 homologue existed, that is, cases with paralogues were discarded. If both a homologue of the
784 downstream and the upstream neighbor existed in a given outgroup genome, and these were
785 separated by at most 1 gene, we considered the region as conserved in synteny. In order for
786 an orphan gene to be characterized as a de novo candidate, such a conserved syntenic region
787 had to be identified in at least one genome of the same species and at least one genome from
788 the closest species. Additionally, we discarded regions where an unannotated ORF with at
789 least 80% query coverage and 60% similarity to the orphan gene existed in the syntenic region
790 of the outgroup species. Finally, we discarded 23 candidates which had similarity matches to
791 conserved genes from the same genome (see 'Similarity searches' subsection above).

792

793 Identification of operon-like genomic regions

794 To identify operon-like gene arrangements, we scanned all UHGG genomes and took
795 advantage of the existing Gene Ontology (GO) annotations available from the InterPro
796 searches performed by the authors. For every gene of every genome, we examined its three
797 immediate downstream and three immediate upstream neighbors. If all seven genes were on
798 the same contig, had the same orientation, and at least three of the neighbors shared at least
799 one GO term annotation, then that gene was considered to be in an operon-like arrangement,
800 and the GO term was associated to that gene. Note that we merged the terms GO:0016020
801 and GO:16021 to be consistent with the current version of GO, after we noticed that terms
802 related to the membrane that came up later in our analyses had been tagged as obsolete and
803 collapsed into the general term "membrane". All associations (one gene can have more than
804 one if more than one GO term is found more than three times) for all genes were formatted
805 into a custom annotation file. This file was used as the background annotation, together with
806 a list of orphan genes in BiNGO v3.0.3[93] to perform the overrepresentation analysis. The
807 options used were: hypergeometric test, Benjamini-Hochberg FDR correction and a 0.05
808 threshold. For analysis restricted to each taxonomic class, the same was performed but only
809 including genes from species of a given class.

810 In the case of orphan gene GUT_GENOME108422_00105 (representative of a family of 135
811 proteins) our analysis of the syntenic region showed the gene to be present at the exact
812 orthologous locus as orphan gene GUT_GENOME123088_00522 (representative of a family of
813 866 members) both belonging to species PeH17 (GUT_GENOME014725). Aligning the two
814 representative sequences gave results consistent with the criteria used by Almeida et al. for
815 protein clustering: the sequences align with 100% identity but only over their first 25 residues
816 (46% of the length of the shortest sequence), while the rest of the sequences align very poorly.
817 However, all other protein sequences of the GUT_GENOME123088_00522 family align nearly
818 perfectly with any protein from the GUT_GENOME108422_00105 family, with no gaps and
819 with 94% identity. Indeed, aligning all CDS sequences from both families together results in
820 an alignment with average pairwise percent identity of 97% (calculated using the *goalign* tool
821 and its *compute distance* method with the *pdist* metric). Furthermore, a nucleotide alignment
822 of the representative sequences showed that their difference at the protein level is due to a
823 frameshift. Overall, these two families, with the exception of the representative sequence of
824 GUT_GENOME123088_00522, would have been expected to cluster together given the
825 parameters used by Almeida et al. (*linclust* with 80% coverage and 50% identity).

826

827 ## Quantification and statistical analysis

828 All statistics were done in R v3.6.2. Plots were generated using *ggplot2*[94]. All statistical details
829 including the type of statistical test performed and exact value of n (n represents either
830 number of genomes or number of genes) can be found in the Results and figure legends.
831 Boxplots show median (horizontal line inside the box), first and third quartiles of data (lower
832 and upper hinges) and values no further or lower than 1.5*distance between the first and
833 third quartiles (upper and lower whisker). No methods were used to determine whether the
834 data met assumptions of the statistical approaches. Subsampling of conserved genes to
835 control for length difference to orphans was performed with replacement, using a customized
836 version of inverse transform sampling. We sampled conserved genes belonging to each
837 species separately, and only species with at least 100 conserved and at least 10 orphans were
838 included (n=3,769). The length median of the sampled conserved was 228nt and the mean
839 was 305nt, compared to 225nt and 300nt in orphans respectively. To detect signatures of
840 selection acting on protein-coding sequences, we first generated DNA multiple sequence
841 alignment of the CDSs of 154,650 orphan families with at least 2 non-identical sequences. We
842 then used the *yn00* executable of PAML[95] to calculate dN, dS and omega (dN/dS) with the
843 Yang & Nielsen 2000 method[96]. The program was run with default values and was given as
844 input the aforementioned MSA. The dN, dS (with their estimated standard errors) and omega
845 values were obtained for all pairwise combinations of members of each family. For each
846 family, to only take into account statistically meaningful comparisons we kept only those
847 comparisons were dN>0, dS>0 , SEdN/dN < 0.5 and SEdS/dS < 2 . We then calculated the mean
848 for a given family from the remaining comparisons. Gene-wide selection signatures were also
849 calculated using the FEL[97] method from HyPhy[98] with default options and the GTR model.

850

851 ## Acknowledgements

852

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

860

861 ## Author information

862

863 **Contributions**

864 AK and NV conceived the study. AK and NV wrote the manuscript. AK supervised the study.
865 NV performed the analyses.

866

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).
6. Tautz, D. & Domazet-Lošo, T. The evolutionary origin of orphan genes. *Nat Rev Genet* **12**, 692–702 (2011).
7. Karlowski, W. M., Varshney, D. & Zielezinski, A. Taxonomically Restricted Genes in Bacillus may Form Clusters of Homologs and Can be Traced to a Large Reservoir of Noncoding Sequences. *Genome Biology and Evolution* **15**, evad023 (2023).

895 8. Yu, G. & Stoltzfus, A. Population diversity of ORFan genes in Escherichia coli. *Genome*
896 *Biol Evol* **4**, 1176–1187 (2012).
897 9. Baumdicker, F. & Kupczok, A. Tackling the Pangenome Dilemma Requires the Concerted
898 Analysis of Multiple Population Genetic Processes. *Genome Biology and Evolution* **15**,
899 evad067 (2023).
900 10. Light, S., Basile, W. & Elofsson, A. Orphans and new gene origination, a structural and
901 evolutionary perspective. *Current Opinion in Structural Biology* **26**, 73–83 (2014).
902 11. Khalturin, K., Hemmrich, G., Fraune, S., Augustin, R. & Bosch, T. C. G. More than just
903 orphans: are taxonomically-restricted genes important in evolution? *Trends in Genetics*
904 **25**, 404–413 (2009).
905 12. Santos, M. E., Bouquin, A. L., Crumière, A. J. J. & Khila, A. Taxon-restricted genes at the
906 origin of a novel trait allowing access to a new environment. *Science* **358**, 386–390
907 (2017).
908 13. Andersson, D. I., Jerlström-Hultqvist, J. & Näsvall, J. Evolution of New Functions De Novo
909 and from Preexisting Genes. *Cold Spring Harb Perspect Biol* **7**, a017996 (2015).
910 14. Prabh, N. & Rödelsperger, C. De Novo, Divergence, and Mixed Origin Contribute to the
911 Emergence of Orphan Genes in Pristionchus Nematodes. *G3: Genes, Genomes, Genetics*
912 **9**, 2277–2286 (2019).
913 15. McLysaght, A. & Guerzoni, D. New genes from non-coding sequence: the role of de novo
914 protein-coding genes in eukaryotic evolutionary innovation. *Phil. Trans. R. Soc. B* **370**,
915 20140332 (2015).
916 16. Oss, S. B. V. & Carvunis, A.-R. De novo gene birth. *PLOS Genetics* **15**, e1008160 (2019).
917 17. Bornberg-Bauer, E., Hlouchova, K. & Lange, A. Structure and function of naturally
918 evolved de novo proteins. *Current Opinion in Structural Biology* **68**, 175–183 (2021).
919 18. Vakirlis, N., Carvunis, A.-R. & McLysaght, A. Synteny-based analyses indicate that
920 sequence divergence is not the main source of orphan genes. *eLife* **9**, e53500 (2020).

921 19. Weisman, C. M., Murray, A. W. & Eddy, S. R. Many, but not all, lineage-specific genes
922 can be explained by homology detection failure. *PLOS Biology* **18**, e3000862 (2020).
923 20. McLysaght, A. & Hurst, L. D. Open questions in the study of de novo genes: what, how
924 and why. *Nat Rev Genet* **17**, 567–578 (2016).
925 21. Daubin, V. & Ochman, H. Start-up entities in the origin of new genes. *Current Opinion in*
926 *Genetics & Development* **14**, 616–619 (2004).
927 22. Daubin, V. & Ochman, H. Bacterial Genomes as New Gene Homes: The Genealogy of
928 ORFans in E. coli. *Genome Res.* **14**, 1036–1042 (2004).
929 23. Cortez, D., Forterre, P. & Gribaldo, S. A hidden reservoir of integrative elements is the
930 major source of recently acquired foreign genes and ORFans in archaeal and bacterial
931 genomes. *Genome Biology* **10**, R65 (2009).
932 24. Yin, Y. & Fischer, D. On the origin of microbial ORFans: quantifying the strength of the
933 evidence for viral lateral transfer. *BMC Evolutionary Biology* **6**, 63 (2006).
934 25. Yomtovian, I., Teerakulkittipong, N., Lee, B., Moult, J. & Unger, R. Composition bias and
935 the origin of ORFan genes. *Bioinformatics* **26**, 996–999 (2010).
936 26. Lobb, B., Kurtz, D. A., Moreno-Hagelsieb, G. & Doxey, A. C. Remote homology and the
937 functions of metagenomic dark matter. *Front Genet* **6**, (2015).
938 27. Delaye, L., DeLuna, A., Lazcano, A. & Becerra, A. The origin of a novel gene through
939 overprinting in Escherichia coli. *BMC Evolutionary Biology* **8**, 31 (2008).
940 28. Ardern, Z. Alternative Reading Frames are an Underappreciated Source of Protein
941 Sequence Novelty. *Journal of Molecular Evolution* (2023) doi:10.1007/s00239-023-
942 10122-3.
943 29. Smith, C. *et al.* Pervasive translation in Mycobacterium tuberculosis. *eLife* **11**, e73980
944 (2022).
945 30. Wade, J. T. & Grainger, D. C. Pervasive transcription: illuminating the dark matter of
946 bacterial transcriptomes. *Nat Rev Microbiol* **12**, 647–653 (2014).

947 31. Wacholder, A. *et al.* A vast evolutionarily transient translatome contributes to
948 phenotype and fitness. *Cell Systems* **14**, 363-381.e8 (2023).
949 32. Knopp, M. *et al.* De Novo Emergence of Peptides That Confer Antibiotic Resistance.
950 *mBio* **10**, e00837-19 (2019).
951 33. Knopp, M. *et al.* A novel type of colistin resistance genes selected from random
952 sequence space. *PLOS Genetics* **17**, e1009227 (2021).
953 34. Tenson, T., Xiong, L., Kloss, P. & Mankin, A. S. Erythromycin Resistance Peptides Selected
954 from Random Peptide Libraries. *J. Biol. Chem.* **272**, 17425–17430 (1997).
955 35. Kirchberger, P. C., Schmidt, M. L. & Ochman, H. The Ingenuity of Bacterial Genomes.
956 *Annual Review of Microbiology* **74**, 815–834 (2020).
957 36. Watson, A. K., Lopez, P. & Bapteste, E. Hundreds of Out-of-Frame Remodeled Gene
958 Families in the Escherichia coli Pangenome. *Molecular Biology and Evolution* **39**,
959 msab329 (2022).
960 37. Chibani, C. M. *et al.* A catalogue of 1,167 genomes from the human gut archaeome. *Nat*
961 *Microbiol* **7**, 48–61 (2022).
962 38. Nayfach, S. *et al.* A genomic catalog of Earth's microbiomes. *Nat Biotechnol* **39**, 499–509
963 (2021).
964 39. Coelho, L. P. *et al.* Towards the biogeography of prokaryotic genes. *Nature* **601**, 252–256
965 (2022).
966 40. Pavlopoulos, G. A. *et al.* Unraveling the functional dark matter through global
967 metagenomics. *Nature* **622**, 594–602 (2023).
968 41. Río, Á. R. del *et al.* Functional and evolutionary significance of unknown genes from
969 uncultivated taxa. 2022.01.26.477801 Preprint at
970 https://doi.org/10.1101/2022.01.26.477801 (2022).
971 42. Vanni, C. *et al.* Unifying the known and unknown microbial coding sequence space. *eLife*
972 **11**, e67667 (2022).

973 43. Ellrott, K., Jaroszewski, L., Li, W., Wooley, J. C. & Godzik, A. Expansion of the Protein
974 Repertoire in Newly Explored Environments: Human Gut Microbiome Specific Protein
975 Families. *PLOS Computational Biology* **6**, e1000798 (2010).
976 44. Almeida, A. *et al.* A unified catalog of 204,938 reference genomes from the human gut
977 microbiome. *Nat Biotechnol* **39**, 105–114 (2021).
978 45. Kuo, C.-H. & Ochman, H. Deletional Bias across the Three Domains of Life. *Genome*
979 *Biology and Evolution* **1**, 145–152 (2009).
980 46. Kryazhimskiy, S. & Plotkin, J. B. The Population Genetics of dN/dS. *PLOS Genetics* **4**,
981 e1000304 (2008).
982 47. Vakirlis, N. *et al.* A Molecular Portrait of De Novo Genes in Yeasts. *Mol Biol Evol* **35**, 631–
983 645 (2018).
984 48. Lawrence, J. G. & Ochman, H. Amelioration of bacterial genomes: rates of change and
985 exchange. *J Mol Evol* **44**, 383–397 (1997).
986 49. Montañés, J. C., Huertas, M., Messeguer, X. & Albà, M. M. Evolutionary Trajectories of
987 New Duplicated and Putative De Novo Genes. *Molecular Biology and Evolution* **40**,
988 msad098 (2023).
989 50. Blevins, W. R. *et al.* Uncovering de novo gene birth in yeast using deep transcriptomics.
990 *Nat Commun* **12**, 604 (2021).
991 51. Vakirlis, N. *et al.* De novo emergence of adaptive membrane proteins from thymine-rich
992 genomic sequences. *Nature Communications* **11**, 1–18 (2020).
993 52. Tassios, E., Nikolaou, C. & Vakirlis, N. Intergenic regions of Saccharomycotina yeasts are
994 enriched in potential to encode transmembrane domains. *Molecular Biology and*
995 *Evolution* msad059 (2023) doi:10.1093/molbev/msad059.
996 53. Sberro, H. *et al.* Large-Scale Analyses of Human Microbiomes Reveal Thousands of Small,
997 Novel Genes. *Cell* **178**, 1245-1259.e14 (2019).

998 54. Hershberg, R. & Petrov, D. A. Evidence That Mutation Is Universally Biased towards AT in
999 Bacteria. *PLOS Genetics* **6**, e1001115 (2010).
1000 55. Papanikolaou, N., Trachana, K., Theodosiou, T., Promponas, V. J. & Iliopoulos, I. Gene
1001 socialization: gene order, GC content and gene silencing in Salmonella. *BMC Genomics*
1002 **10**, 597 (2009).
1003 56. Nayfach, S. *et al.* Metagenomic compendium of 189,680 DNA viruses from the human
1004 gut microbiome. *Nat Microbiol* **6**, 960–970 (2021).
1005 57. Camarillo-Guerrero, L. F., Almeida, A., Rangel-Pineros, G., Finn, R. D. & Lawley, T. D.
1006 Massive expansion of human gut bacteriophage diversity. *Cell* **184**, 1098-1109.e9
1007 (2021).
1008 58. Shah, S. A. *et al.* Expanding known viral diversity in the healthy infant gut. *Nat Microbiol*
1009 **8**, 986–998 (2023).
1010 59. Almpanis, A., Swain, M., Gatherer, D. & McEwan, N. Correlation between bacterial G+C
1011 content, genome size and the G+C content of associated plasmids and bacteriophages.
1012 *Microb Genom* **4**, e000168 (2018).
1013 60. Starikova, E. V. *et al.* Phigaro: high-throughput prophage sequence annotation.
1014 *Bioinformatics* **36**, 3882–3884 (2020).
1015 61. Tria, F. D. K. & Martin, W. F. Gene Duplications Are At Least 50 Times Less Frequent than
1016 Gene Transfers in Prokaryotic Genomes. *Genome Biology and Evolution* **13**, evab224
1017 (2021).
1018 62. Watson, A. K., Lopez, P. & Bapteste, E. Hundreds of out-of-frame remodelled gene
1019 families in the E. coli pangenome. *Molecular Biology and Evolution* (2021)
1020 doi:10.1093/molbev/msab329.
1021 63. Katoh, K. & Standley, D. M. MAFFT Multiple Sequence Alignment Software Version 7:
1022 Improvements in Performance and Usability. *Mol Biol Evol* **30**, 772–780 (2013).

1023 64. Waterhouse, A. M., Procter, J. B., Martin, D. M. A., Clamp, M. & Barton, G. J. Jalview
1024 Version 2—a multiple sequence alignment editor and analysis workbench.
1025 *Bioinformatics* **25**, 1189–1191 (2009).
1026 65. Price, M. N. *et al.* Mutant phenotypes for thousands of bacterial genes of unknown
1027 function. *Nature* **557**, 503–509 (2018).
1028 66. Osbourn, A. E. & Field, B. Operons. *Cell. Mol. Life Sci.* **66**, 3755–3775 (2009).
1029 67. Babina, A. M. *et al.* Rescue of Escherichia coli auxotrophy by de novo small proteins.
1030 *eLife* **12**, e78299 (2023).
1031 68. Frumkin, I. & Laub, M. T. Selection of a de novo gene that can promote survival of E. coli
1032 by modulating protein homeostasis pathways. 2023.02.07.527531 Preprint at
1033 https://doi.org/10.1101/2023.02.07.527531 (2023).
1034 69. Fremin, B. J. *et al.* Thousands of small, novel genes predicted in global phage genomes.
1035 *Cell Reports* **39**, (2022).
1036 70. Pavesi, A., Magiorkinis, G. & Karlin, D. G. Viral Proteins Originated De Novo by
1037 Overprinting Can Be Identified by Codon Usage: Application to the “Gene Nursery” of
1038 Deltaretroviruses. *PLoS Comput Biol* **9**, (2013).
1039 71. Sabath, N., Wagner, A. & Karlin, D. Evolution of Viral Proteins Originated De Novo by
1040 Overprinting. *Mol Biol Evol* **29**, 3767–3780 (2012).
1041 72. Wolf, Y. I., Makarova, K. S., Lobkovsky, A. E. & Koonin, E. V. Two fundamentally different
1042 classes of microbial genes. *Nat Microbiol* **2**, 1–6 (2016).
1043 73. Iyengar, B. R. & Bornberg-Bauer, E. Neutral Models of De Novo Gene Emergence Suggest
1044 that Gene Evolution has a Preferred Trajectory. *Molecular Biology and Evolution* **40**,
1045 msad079 (2023).
1046 74. Weisman, C. M., Murray, A. W. & Eddy, S. R. Mixing genome annotation methods in a
1047 comparative analysis inflates the apparent number of lineage-specific genes. *Current*
1048 *Biology* **32**, 2632-2639.e2 (2022).

1049 75. Maguire, F. *et al.* Metagenome-assembled genome binning methods with short reads
1050 disproportionately fail for plasmids and genomic Islands. *Microbial Genomics* **6**, e000436
1051 (2020).
1052 76. Meziti, A. *et al.* The Reliability of Metagenome-Assembled Genomes (MAGs) in
1053 Representing Natural Populations: Insights from Comparing MAGs against Isolate
1054 Genomes Derived from the Same Fecal Sample. *Applied and Environmental*
1055 *Microbiology* **87**, e02593-20 (2021).
1056 77. Chen, L.-X., Anantharaman, K., Shaiber, A., Eren, A. M. & Banfield, J. F. Accurate and
1057 complete genomes from metagenomes. *Genome Res.* **30**, 315–333 (2020).
1058 78. Eberhardt, R. Y. *et al.* AntiFam: a tool to help identify spurious ORFs in protein
1059 annotation. *Database (Oxford)* **2012**, bas003 (2012).
1060 79. Ardern, Z., Neuhaus, K. & Scherer, S. Are Antisense Proteins in Prokaryotes Functional?
1061 *Frontiers in Molecular Biosciences* **7**, (2020).
1062 80. Vatanen, T. *et al.* Genomic variation and strain-specific functional adaptation in the
1063 human gut microbiome during early life. *Nat Microbiol* **4**, 470–479 (2019).
1064 81. Hernández-Plaza, A. *et al.* eggNOG 6.0: enabling comparative genomics across 12 535
1065 organisms. *Nucleic Acids Research* **51**, D389–D394 (2023).
1066 82. Buchfink, B., Reuter, K. & Drost, H.-G. Sensitive protein alignments at tree-of-life scale
1067 using DIAMOND. *Nat Methods* **18**, 366–368 (2021).
1068 83. Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database
1069 search programs. *Nucleic Acids Res.* **25**, 3389–3402 (1997).
1070 84. Wang, S., Li, W., Liu, S. & Xu, J. RaptorX-Property: a web server for protein structure
1071 property prediction. *Nucleic Acids Res* **44**, W430–W435 (2016).
1072 85. Walsh, I., Seno, F., Tosatto, S. C. E. & Trovato, A. PASTA 2.0: an improved server for
1073 protein aggregation prediction. *Nucleic Acids Research* **42**, W301–W307 (2014).

1074 86. Käll, L., Krogh, A. & Sonnhammer, E. L. L. Advantages of combined transmembrane
1075 topology and signal peptide prediction—the Phobius web server. *Nucleic Acids Res* **35**,
1076 W429–W432 (2007).
1077 87. Carbone, A., Zinovyev, A. & Képès, F. Codon adaptation index as a measure of
1078 dominating codon bias. *Bioinformatics* **19**, 2005–2015 (2003).
1079 88. Correspondence Analysis of Codon Usage. http://codonw.sourceforge.net/.
1080 89. Charif, D. & Lobry, J. R. SeqinR 1.0-2: A Contributed Package to the R Project for
1081 Statistical Computing Devoted to Biological Sequences Retrieval and Analysis. in
1082 *Structural Approaches to Sequence Evolution: Molecules, Networks, Populations* (eds.
1083 Bastolla, U., Porto, M., Roman, H. E. & Vendruscolo, M.) 207–232 (Springer, 2007).
1084 doi:10.1007/978-3-540-35306-5_10.
1085 90. Akashi, H. & Gojobori, T. Metabolic efficiency and amino acid composition in the
1086 proteomes of Escherichia coli and Bacillus subtilis. *Proc. Natl. Acad. Sci. U.S.A.* **99**, 3695–
1087 3700 (2002).
1088 91. Barton, M. D., Delneri, D., Oliver, S. G., Rattray, M. & Bergman, C. M. Evolutionary
1089 Systems Biology of Amino Acid Biosynthetic Cost in Yeast. *PLOS ONE* **5**, e11935 (2010).
1090 92. Osorio, D., Rondón-Villarreal, P. & Torres, R. Peptides: A package for data mining of
1091 antimicrobial peptides. *Small* **12**, 44–444 (2015).
1092 93. Maere, S., Heymans, K. & Kuiper, M. BiNGO: a Cytoscape plugin to assess
1093 overrepresentation of gene ontology categories in biological networks. *Bioinformatics*
1094 **21**, 3448–3449 (2005).
1095 94. Wickham, H. ggplot2. *WIREs Computational Statistics* **3**, 180–185 (2011).
1096 95. Yang, Z. PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Mol Biol Evol* **24**, 1586–
1097 1591 (2007).
1098 96. Yang, Z. & Nielsen, R. Estimating Synonymous and Nonsynonymous Substitution Rates
1099 Under Realistic Evolutionary Models. *Mol Biol Evol* **17**, 32–43 (2000).

1100 97. Kosakovsky Pond, S. L. & Frost, S. D. W. Not So Different After All: A Comparison of
1101 Methods for Detecting Amino Acid Sites Under Selection. *Molecular Biology and*
1102 *Evolution* **22**, 1208–1222 (2005).
1103 98. HyPhy: Hypothesis Testing Using Phylogenies | SpringerLink.
1104 https://link.springer.com/chapter/10.1007%2F0-387-27733-1_6.
1105