Pseudogenes as a neutral reference for detecting selection in prokaryotic pangenomes

Gavin M. Douglas[a,b,*], W. Ford Doolittle[c], B. Jesse Shapiro[a,b,d,*]

[a]*Department of Microbiology and Immunology, McGill University, Montréal, QC, Canada*

[b]*McGill Genome Centre, McGill University, Montréal, QC, Canada*

[c]*Department of Biochemistry and Molecular Biology, Dalhousie University, Halifax, NS, Canada*

[d]*McGill Centre for Microbiome Research, McGill University, Montréal, QC, Canada*

[*]Emails for correspondence: gavin.douglas@mcgill.ca and jesse.shapiro@mcgill.ca

Keywords: Pseudogenes, pangenome, evolution, horizontal gene transfer, lateral gene transfer, mobile genes, mobilome, adaptation

Abstract

A long-standing question is to what degree genetic drift vs. selection drives the divergence in rare accessory gene content between closely related bacteria. Rare genes, including singletons, make up a large proportion of pangenomes (the set of all genes in a set of genomes), but it remains unclear how many such genes are adaptive, deleterious, or neutral to their host genome. Estimates of species' effective population sizes (N_e) are positively associated with pangenome size and fluidity, which has independently been interpreted as evidence for both neutral and adaptive pangenome models. We hypothesised that these models could be distinguished if measures of pangenome diversity were normalized by pseudogene diversity as a proxy for neutral genic diversity. To this end, we defined the ratio of singleton intact genes to singleton pseudogenes (s_i/s_p) within a pangenome, which shows a signal across prokaryotic species consistent with the relative adaptive value of many rare accessory genes. We also identified differences in functional annotations between intact genes and pseudogenes. For instance, transposons are highly enriched among pseudogenes, while most other functional categories are more often intact. Our work demonstrates that including pseudogenes as a neutral reference leads to improved inferences of the evolutionary forces driving pangenome variation.

Main

A long-standing question is to what degree genetic drift vs. selection drives the divergence in rare accessory gene content between closely related bacteria[1–4]. Rare genes, including singletons, make up a large proportion of pangenomes (the set of all genes in a set of genomes), but it remains unclear how many such genes are adaptive, deleterious, or neutral to their host genome. Estimates of species' effective population sizes (N_e) are positively associated with pangenome size and fluidity[5–7], which has independently been interpreted as evidence for both neutral[6] and adaptive[5,7] pangenome models. We hypothesised that these models could be distinguished if measures of pangenome diversity were normalized by pseudogene diversity as a proxy for neutral genic diversity. To this end, we defined the ratio of singleton intact genes to singleton pseudogenes (s_i/s_p) within a pangenome, which shows a signal across prokaryotic species consistent with the relative adaptive value of many rare accessory genes. We also identified differences in functional annotations between intact genes and pseudogenes. For instance, transposons are highly enriched among pseudogenes, while most other functional categories are more often intact. Our work demonstrates that including pseudogenes as a neutral reference leads to improved inferences of the evolutionary forces driving pangenome variation. (Please note that the first paragraph was duplicated as the above abstract to allow bioRxiv to correctly parse it: the original version of this preprint did not have a separate abstract).

These evolutionary forces have been investigated through several approaches, such as analysing gene frequency distributions,[8,9] gene co-occurrence[10], and patterns of nucleotide variation within transferred genes[11]. This work has primarily provided insight into the higher frequency accessory genes, rather than rare genes that make up the largest fraction of pangenomes[12]. These rare genes (often singletons sequenced in just one genome) are frequently mobile genes with high turnover rates and dubious adaptive value to their bacterial hosts. Nonetheless, rare genes have been hypothesised to provide adaptative benefits in rare ecological niches[13–15], although this hypothesis remains largely untested. Here, we propose a new metric of selection on rare accessory genes, which we apply to a dataset of >600 prokaryotic species. We then analyse a subset of well-sampled bacterial species to identify functional categories that are enriched in intact genes compared to pseudogenes. Our results provide strong evidence that an identifiable subset of rare accessory genes likely provide adaptive value to their hosts.

Pseudogenization – gene degeneration through the introduction of mutations, such as premature stop codons, insertions, and deletions – can occur when genetic drift overcomes purifying selection to retain a gene[16], or through positive selection to eliminate a deleterious gene[17]. We reasoned that rare accessory gene families that tend to remain intact are under stronger positive selection than those that tend to be pseudogenized. We expressed this by calculating the mean percentages of intact singleton genes (s_i) and pseudogenes (s_p) within a species' pangenome. We analysed 668 named prokaryotic species represented by at least nine genomes in the Genome Taxonomy Database[18] and found that the mean values of s_i and s_p were correlated (Spearman's ρ=0.57; $P < 0.001$), with deviations suggesting species-specific differences in selection on rare accessory genes (**Figure 1**). For example, *Escherichia coli* has a high s_i/s_p ratio, consistent with selection to retain rare accessory genes, while the obligate intracellular bacteria *Chlamydia trachomatis* and *Rickettsia prowazekii* exhibit the lowest ratios, which could indicate less selective constraint on their rare genes.

The analysed species span substantial prokaryotic diversity (**Extended Data Table 1**) but were biased towards Gammaproteobacteria (286 species) and Bacilli (161 species). We identified pseudogenes with Pseudofinder[19], which identifies several classes of potential pseudogenes. We focused on intergenic pseudogenes, which represent significant matches to database sequences outside of gene calls, as this class is more likely to represent degenerating gene sequences compared to the other candidate pseudogene classes (see Online Methods). We filtered out pseudogenes based on several criteria, including restricting analysed pseudogenes to those >= 100 bp and <= 5000 bp in length. Based on all criteria, a mean of 11.90% (standard deviation [SD]: 5.78%) of pseudogenes were excluded per species. After this filtering, intergenic pseudogenes represented a mean of 4.52% (SD: 2.96%) of called elements per genome (range: 0.30-19.81%). These elements comprised an even smaller portion of overall genome size (mean: 1.42%, SD: 0.99%) compared to intact genes (mean: 87.34%, SD: 2.77%) because pseudogenes are generally shorter than intact genes.

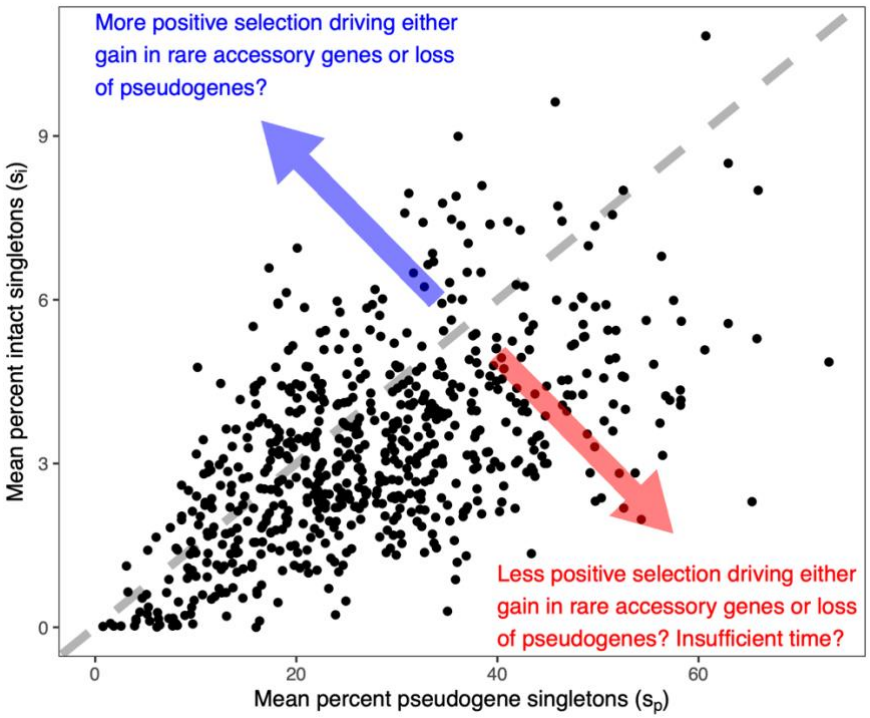

Figure 1: Mean percentage of intact genes and pseudogenes that are singletons (i.e. genome-specific) per species. Each point represents one of 668 prokaryotic species (>= nine genomes each). The mean percent singletons (for both intact genes and pseudogenes) per species was based on repeated subsampling to nine genomes (for up to 100 replicates). Possible (but non-exhaustive) drivers of higher or lower s_i/s_p ratios are indicated alongside coloured arrows.

Species' pangenome size and complexity have been characterised based on different metrics, including the mean number of genes per genome[5] and genomic fluidity[6,20]. We computed these metrics for all species based on both intact genes and pseudogenes. As we were especially interested in rare elements, we computed the mean numbers and percentages of singleton genes and pseudogenes per species (i.e. those present in a single genome per species), based on repeated subsampling to nine genomes. Larger genomes tend to encode more singletons, both in mean number and percentage (**Extended Data Fig. 1a,b**). In addition, the percentage of intact singletons (s_i) is highly correlated with genomic fluidity, but the traditional fluidity metric is sensitive to intermediate frequency accessory genes (**Extended Data Fig. 1c,d**). We therefore focused on the percentage of intact (s_i) and pseudogene (s_p) singletons for most analyses. All metrics ranged substantially across species for both intact genes (fluidity: 0.003-0.246; mean number: 836.4-8692.7; mean number of singletons: 0.00-581.29; s_i: 0.00-10.83%) and pseudogenes (fluidity: 0.014-0.513; mean number: 8.1-922.5; mean number of singletons: 0.78-325.17; s_p: 0.78-72.97%). These results highlight that, as expected[21], pseudogenes are frequently genome-specific.

We next recapitulated the previously observed association between genome-wide non-synonymous to synonymous substitution rates (dN/dS) and pangenome diversity[5,7], and then explored whether dN/dS is also associated with s_i/s_p. We computed dN/dS across the core genome of each species, based on the mean values of all pairwise strain comparisons. This metric is often taken as a proxy for selection efficacy: lower dN/dS values indicate increased efficacy of purifying selection (which is associated with higher N_e) against non-synonymous changes, which tend to be deleterious. However, within-species dN/dS values are highly dependent on strain divergence times, with recent divergences enriched in higher dN/dS due to insufficient time for purifying selection to purge deleterious non-synonymous mutations[22,23]. Using within-species dS as a proxy for the molecular clock, we also observed a time-dependence of dN/dS in our data (**Extended Data Fig. 2a**).

Due to this relationship between within-species dN/dS and dS, we included dS as a covariate when computing partial Spearman correlations between measures of pangenome diversity and dN/dS. Based on this approach, the mean number of genes, genomic fluidity, and s_i were all significantly negatively correlated with dN/dS across species (**Figure 2 a-c**; Partial Spearman correlations, $P < 0.05$). This observation agrees with past work[5,7], which has been taken as evidence for an adaptive pangenome model. However, N_e, which determines selection efficacy and thus core genome dN/dS (assuming equal selection pressure), is also associated with higher standing levels of neutral variation, due to less variation being lost through genetic drift in larger populations. Accordingly, a positive association between pangenome diversity and N_e can be explained by both an adaptive or neutral model.

To disentangle these models, we explored whether our new metric, s_i/s_p, is differently associated with dS and dN/dS compared to the unnormalized measures of pangenome diversity. Based on a partial Spearman correlation, we found s_i/s_p to be significantly associated with dN/dS (partial Spearman's ρ=0.237; $P < 0.001$; **Figure 2d**), although less so than s_i alone (partial Spearman's ρ=0.372; $P < 0.001$). This result highlights that s_i remains associated with dN/dS even after normalization by s_p. If pseudogene diversity is assumed to be a proxy for neutral genic diversity, this finding suggests that intact singleton gene prevalence is particularly associated with selection efficacy, and not simply with standing neutral variation. In other words, there is a role for natural selection in maintaining even very rare intact genes within pangenomes.

Although it is difficult to prove that most rare pseudogenes are evolving neutrally, it is possible to test for signals expected if there is positive selection for pseudogene loss. If this were true, pseudogene content would be expected to be lower in species with higher efficacy of selection. Contrary to this prediction, the mean percent of species' genomes covered by pseudogenes was not significantly associated with dN/dS (partial Spearman's ρ=0.005; P = 0.8972; **Extended Data Fig. 2b**), which is inconsistent with a model of widespread slightly deleterious pseudogenes that are purged only in species with sufficiently high N_e.

A limitation of our partial correlation analyses is that they did not control for systematic differences across taxonomic groups. In addition, they provide no insight into the relative explanatory power of dN/dS vs. dS for explaining pangenome diversity. To address these points, we conducted a complementary linear modelling analysis, where a separate model was generated with each of the four pangenome diversity measures as the response, and dS, dN/dS, and taxonomic class as predictors. Continuous variables were converted to standard units so that coefficients could be compared across models. All models were highly significant (P<0.001; **Figure 3**) and ranged in adjusted R^2 values from 0.197 to 0.420 for the s_i/s_p and s_i models, respectively.

All but one class (Bacilli) were significant predictors in at least one model, and Clostridia, Bacteroidia, and Chlamydiia were significant predictors across all four models. Similarly, dS was a significant predictor of all pangenome diversity metrics except for s_i/s_p. In contrast, dN/dS was a significant predictor for all pangenome diversity metrics except for the mean number of genes, which could indicate that gene number is an overly simplistic measure of pangenome diversity. Most pertinently, these results highlight that dN/dS, a proxy for selection efficacy, remains a significant predictor of s_i/s_p. In addition, dS, a measure that incorporates both divergence time and the species-wide level of standing neutral variation, is a predictor of s_i, but not s_i/s_p, which would be unexpected were singleton intact genes and pseudogenes both evolving neutrally. Instead, these results are consistent with s_i/s_p behaving somewhat analogously to dN/dS as a measure of the efficacy of selection. As a higher fraction of rare genes (relative to pseudogenes) are retained when selection is more effective, this is consistent with many singleton genes conferring adaptive benefits, and/or some singleton pseudogenes being slightly deleterious. As the latter effect is undetectable in our data (**Extended Data Fig. 2b**), we favour the hypothesis that rare intact genes tend to provide benefits to their host genomes.

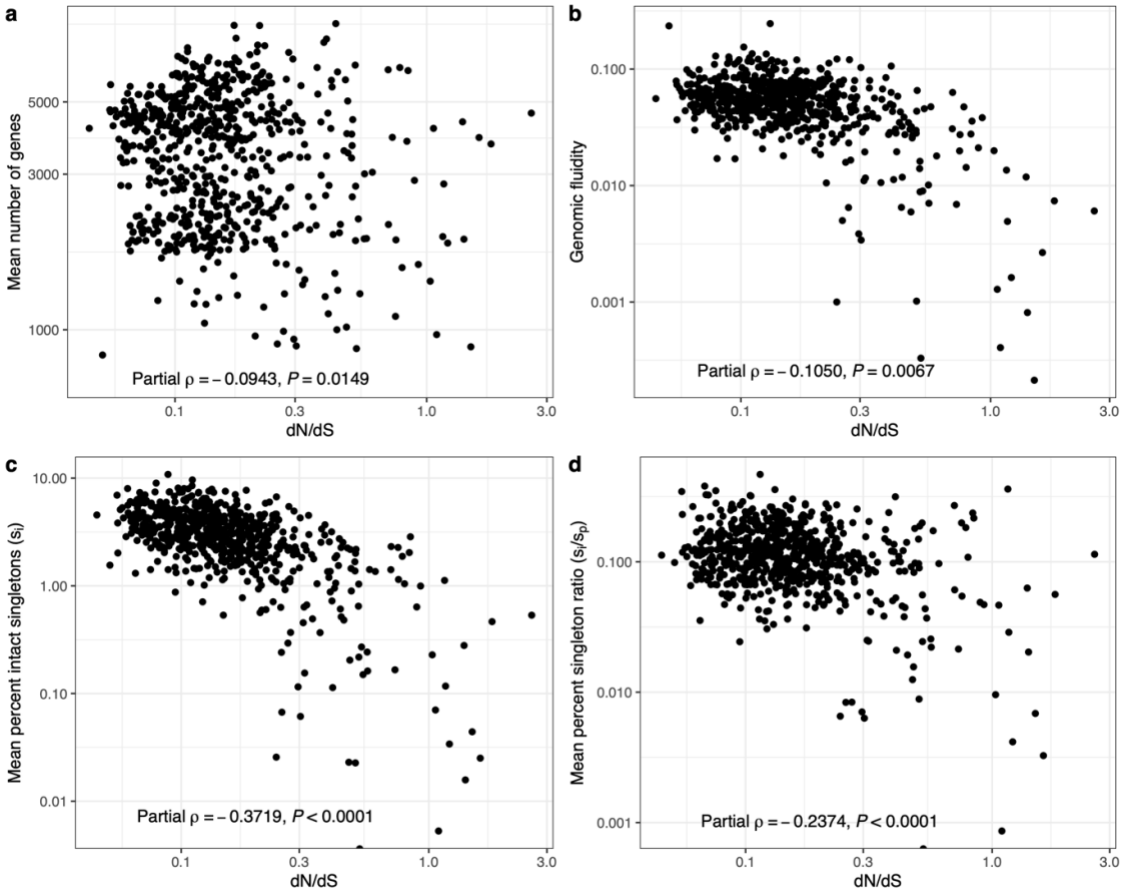

Figure 2: Associations between pangenome diversity metrics and estimated efficacy of selection (dN/dS). Each panel presents the association between the ratio of non-synonymous to synonymous substitution rates (dN/dS; across each species' core genome) and one of the following measures: (a) the mean number of genes per genome, (b) genomic fluidity, (c) the mean percent of intact singletons, and the percentage of singleton intact genes normalized by the percentage of singleton pseudogenes per species Each point is one of 668 prokaryotic species, plotted on $\log_{10}$ scales. The partial Spearman correlation coefficients (which control for dS) and *P*-values are indicated in the bottom left corners. In both panels c and d, one species (*Rickettsia prowazekii*) contained no singleton intact genes and is indicated by the point intersecting the x-axis in both panels.

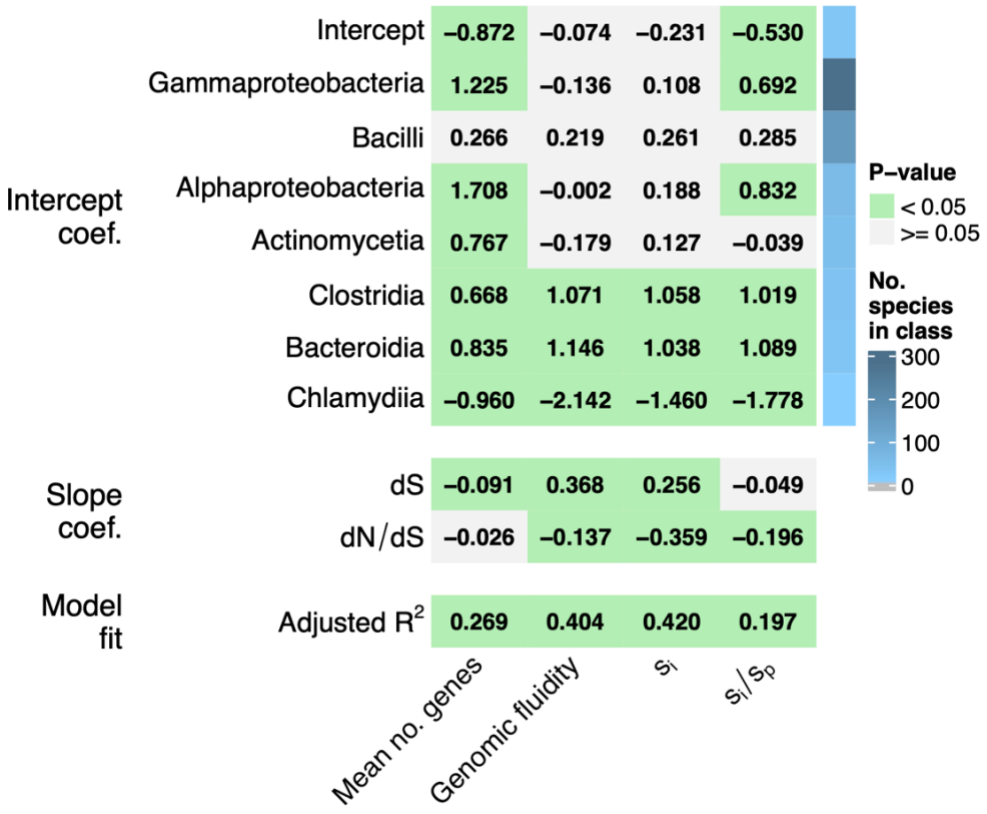

Figure 3: **The s_i/s_p metric varies across taxa and is correlated with the efficacy of selection.** Summaries of four pangenome diversity linear models are shown. One model was fit for each pangenome diversity metric: the mean number of genes, genomic fluidity, the percentage of singleton intact genes (s_i), and the ratio of the percentages of singleton intact genes vs. pseudogenes (s_i/s_p). All continuous response and predictor variables were standardized (i.e. converted to z-scores) prior to building models. Most continuous variables were also transformed to normal distributions prior to this standardization (see Online Methods). Coefficients are displayed for each model, split by those that affect the intercept vs. the slope. The adjusted R^2 is also indicated for each model, and the cell colouring indicates whether each value is statistically significant ($P < 0.05$). The number of genomes per taxonomic class is indicated by the blue bar. The category used to infer the overall intercept was based on a combination of all classes with <= 5 species present. These models were built based on 667 species, after excluding one species with no singleton intact genes.

Having established s_i/s_p as a measure of selection on rare accessory genes, we asked how selection varies across different functional categories of rare genes. To answer this question, we used a dataset of 10 species with a relatively high number of genomes, including highly sampled human pathogens and bacteria with other lifestyles: *Agrobacterium tumefaciens* (223 genomes), *Enterococcus faecalis* (1,298 genomes), *Escherichia coli* (2,955 genomes), *Lactococcus lactis* (135 genomes), *Pseudomonas aeruginosa* (4,115 genomes), *Sinorhizobium meliloti* (166 genomes), *Staphylococcus epidermidis* (447 genomes), *Streptococcus pneumoniae* (6,845 genomes), *Wolbachia pipientis* (716 genomes), and *Xanthomonas oryzae* (326 genomes). We

called intact genes and intergenic pseudogenes across these genomes as described above, but performed joint clustering of intact genes and pseudogenes, to ensure that differences in how sequence clusters are defined do not influence the results. These 10 species substantially varied in genome content and characteristics (**Extended Data Table 2**); for example, *Wolbachia pipientis* genomes encoded a mean of 897.0 intact genes (SD: 25.1) and 55.4 pseudogenes (SD: 20.8), while *Sinorhizobium meliloti* genomes encoded a mean of 6032.8 intact genes (SD: 205.7) and 489.7 pseudogenes (SD: 53.4).

We annotated each sequence cluster using eggNOG-mapper[24] to identify Clusters of Orthologous Genes (COG) annotations[25]. This tool annotates protein sequences, which is problematic for most pseudogenes as the protein-coding information is generally lost. Instead, we annotated all proteins (i.e. those from a larger database used to define pseudogenes originally) that matched each pseudogene sequence. We identified a mean of 57.94% (SD: 7.06%) of intact gene clusters and 49.46% (SD: 7.09%) of pseudogene clusters as COG-annotated. The ratio of the percent COG-annotated intact genes vs. pseudogenes was significantly higher than one in 6/10 of species and lower than one in 2/10 species (Fisher's exact tests, P <0.05). We separated all clusters into three pangenome partitions, based on their frequency within a species: cloud (<=15%), shell (>15% and <95%), and soft-core (>=95%). We also further partitioned cloud clusters into ultra-rare, including clusters found in only one or two genomes (singletons and doubletons), and other-rare, referring to higher-frequency cloud clusters. As expected, most pseudogene clusters were within the cloud partitions: mean of 95.46% (SD: 3.78%) vs. a mean of 84.01% (SD: 8.34%) for intact genes (**Extended Data Figure 3a**). Some pseudogene clusters were in the soft-core partition (mean: 0.54%, SD: 0.66%), which primarily lacked COG annotations (**Extended Data Figure 3b**). For the subsequent analyses we proceeded with COG-annotated clusters only (**Extended Data Figure 4**).

We applied generalized linear mixed models, for each pangenome partition separately (excluding soft-core elements), to investigate which factors best explain whether an element is intact or a pseudogene. These models included 213,912, 3,650,010, and 12,234,597 elements for the ultra-rare, other-rare, and shell partitions, respectively. The fixed effects included each element's COG category and whether the element was redundant with an intact gene with the same COG ID in the same genome. We included the 'redundancy' effect because adaptive genes might neutrally degenerate if they are complemented by an intact copy of the same gene family

in the genome. The interaction between COG category and functional redundancy was also included as a fixed effect. Last, we also included species names, the interaction between COG category and species, and the interaction between functional redundancy and species random effects. All variables added significant information to these models, but there were some slight differences in their relative contributions. For instance, species identity and element functional redundancy were particularly informative in the ultra-rare model compared to the more frequent categories of genes (**Extended Data Figure 5**), and certain species displayed different associations with pseudogenization by pangenome partition (**Extended Data Figure 6**).

We identified significant coefficients in the ultra-rare model (**Figure 4**), which provided insight into what factors were most associated with pseudogene status ($P < 0.05$). These coefficients represent decreased log-odds (logit) probabilities of an element being a pseudogene. Five COG categories were positively associated with pseudogenization: 'energy production and conversion' (C), 'nucleotide transport and metabolism' (F), 'translation, ribosomal structure and biogenesis' (J), 'function unknown' (S), and – most strongly – 'mobilome: prophages, transposons' (X). 'Cell cycle control, cell division, chromosome partitioning' (D), was the sole COG category specifically associated with decreased pseudogenization. Non-redundant elements were highly associated with decreased pseudogenization, over most COG categories. This indicates that even very rare accessory genes are often under selection to maintain a functional copy in the genome. Non-redundant elements were also depleted for pseudogenes in the other-rare and shell models, but different COG categories were associated with pseudogenization overall (**Extended Data Figure 7**). The exception was an enrichment of pseudogenes in mobilome-associated elements in the other-rare partition.

In the study of pangenome evolution, a key question is what proportion of rare genes are under selection or subject to genetic drift. This question is challenging to answer precisely; yet our models yield estimates of the percentage of genes found in functional groupings depleted for pseudogenes, providing a lower bound for the percentage of adaptive genes. For instance, genes in COG category D and non-redundant genes in COG category E are two such pseudogene-depleted groupings. Based on these definitions, a mean of 19.41% (SD: 5.27%), 20.32% (SD: 6.84%), and 26.02% (SD: 7.05%) of intact genes are found in pseudogene-depleted groupings across the ultra-rare, other-rare, and shell partitions, respectively. The increasing percentage of genes classified as pseudogene-depleted as gene frequency increases from ultra-rare to shell is

consistent with more frequent genes being more likely adaptive to their host. Nevertheless, an appreciable percentage (>19%) of ultra-rare genes are likely adaptive according to this estimate. Note that although element COG non-redundancy was highly negatively associated with pseudogenization, only 24.39% of elements were non-redundant, which accounts for why only a minority of intact genes were categorized into pseudogene-depleted groupings. Conversely, 18.68% (SD: 5.62%), 13.29% (SD: 7.69%), and 3.65% (SD: 0.74%) of intact genes are found in groupings enriched for pseudogenes across these three partitions. The decreasing percentages as gene frequency increases is consistent with rarer genes being more likely deleterious to their host. Therefore, although rare accessory genes may on average be adaptive to their host genomes, a substantial fraction may also be deleterious. Most intact genes do not fall cleanly into either the pseudogene-enriched or -depleted category, meaning that these estimates represent rough lower bounds of how many genes are likely adaptive or deleterious.

Several COG categories were significant in our models, but these are broad groupings that can be difficult to biologically interpret. We investigated which individual COG IDs within significant COG categories were driving the overall signals in the ultra-rare model (see Online Methods). The clearest signal was of transposase-associated COGs being highly enriched among pseudogenes (mean of significant odds ratios: 5.10, SD: 6.86), which contrasted with other mobilome-associated COGs (**Extended Data Fig. 8**). We also identified several COGs highly associated with pseudogenization in specific species. For instance, anaerobic selenocysteine-containing dehydrogenases (COG0243, category C), were highly enriched for pseudogenes across multiple species, particularly in *Agrobacterium tumefaciens* (odds ratio: 103.6, $P < 0.001$). In addition, several COGs in category D involved in cell division and chromosome segregation were significantly depleted for pseudogenes, including BcsQ (COG1192), a ParA-like ATPase, which was significantly depleted for pseudogenes in six species (false discovery rate < 0.05).

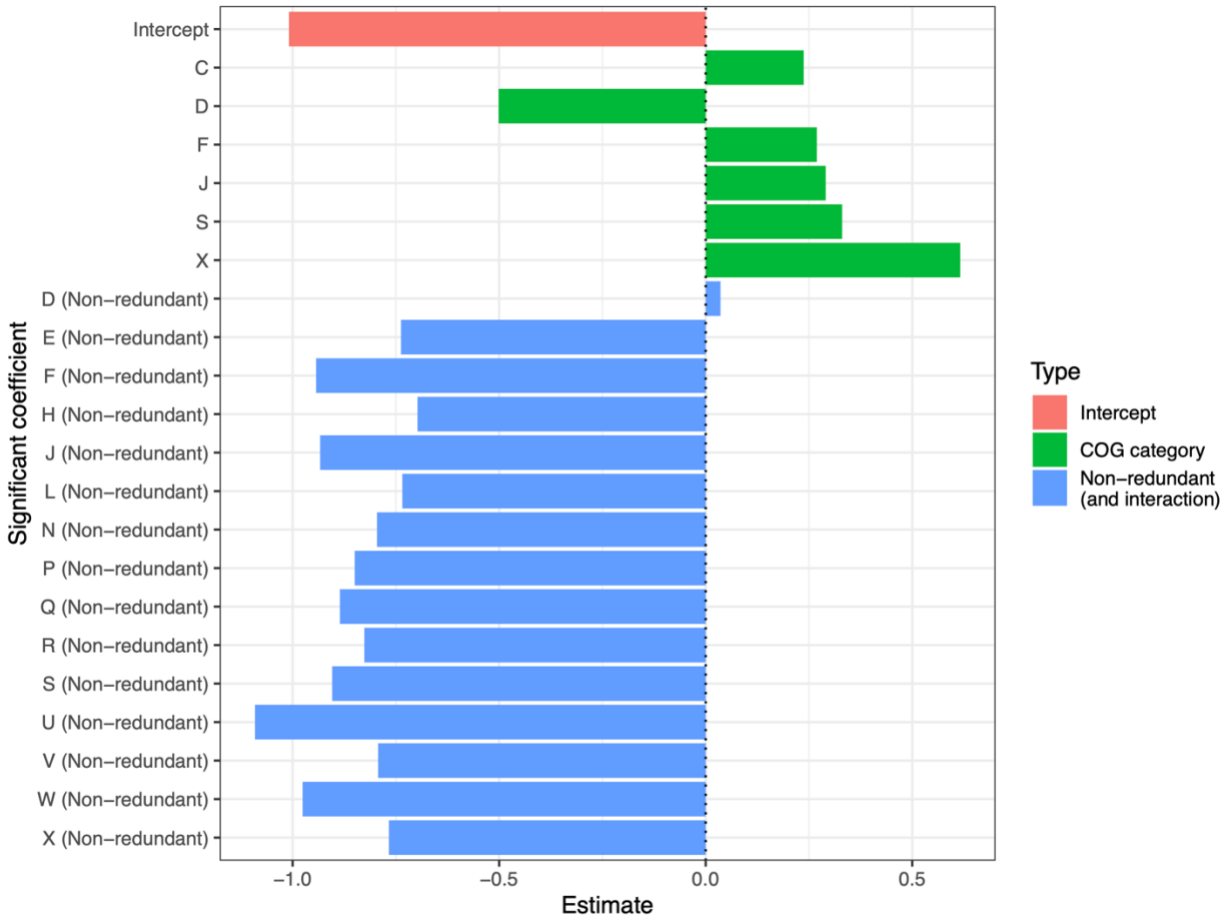

Figure 4: Summary of significant coefficients ($P < 0.05$) in generalized linear mixed model with singleton and doubleton (ultra-rare) element state (intact or pseudogene) as the response. The predictors were each element's annotated COG category (indicated by single-letter codes), whether the element is redundant with an intact gene of the same COG ID (i.e. gene family, not COG category) in the same genome, and the interaction between these variables. The non-redundant coefficients represent the sum of the overall non-redundant coefficient and the interaction of non-redundancy and each COG category. Estimates correspond to logit (log-odds) values: estimates > 0 indicate an increased probability of an element being classified as a pseudogene. Significant COG categories (excluding those significant when non-redundant) include: energy production and conversion (C), cell cycle control, cell division, chromosome partitioning (D), nucleotide transport and metabolism (F), translation, ribosomal structure and biogenesis (J), function unknown (S), and mobilome: prophages, transposons (X).

The ability to distinguish neutral and adaptive models of pangenome evolution has been hindered by a lack of tools to test for selection acting on gene content. This contrasts with an established toolkit of tests for selection at the nucleotide or protein level, including dN/dS and its extensions. Here we propose pseudogene diversity as a reference for distinguishing neutral and adaptive forces acting on pangenomes – particularly rare genes. We showed that the association between pangenome diversity and synonymous-site variation disappears after correcting for pseudogene diversity with the s_i/s_p metric, while the association with dN/dS is maintained. This

indicates that a higher proportion of intact singleton genes (relative to singleton pseudogenes) are present when selection is more effective. This would be unexpected if all rare intact genes were evolving neutrally, and so is strong evidence against a fully neutral model of prokaryotic pangenome diversity. Instead, it is consistent with a model where rare intact genes confer slightly adaptive functions, which are more likely to be preserved by selection given higher selection efficacy[7] (such as in *E. coli*), but that may degenerate neutrally and become pseudogenes in species with lower N_e (such as obligate intracellular bacteria). It would also be consistent with a model where there are widespread slightly deleterious rare pseudogenes, which can be purged only in species with high N_e, but we did not detect a significant association between dN/dS and pseudogene content, making this less likely.

A common explanation for widespread selection on rare accessory genes is adaptation to highly specialized niches[13–15]. While genes recently acquired through horizontal gene transfer are often hypothesised to be niche-specific adaptations[26], it is challenging to make high-confidence inferences without knowing the background of all recently transferred genes that were not retained – and are thus unobservable by definition. By focusing on pseudogenes, which are observable but likely to evolve mostly by drift, we can establish a (nearly) neutral background against which to discern potentially niche-specific adaptations.

We relied on the assumption that any selection pressures acting upon pseudogenes overall are of much lower magnitude compared to intact genes. In other words, we assumed that, overall, the pseudogenization instances we identified do not reflect adaptive gene loss[27] (which is unlikely to substantially increase with selection efficacy, as described above), nor do they represent adaptive regulatory informative transferred between bacteria through HGT[28]. This second possibility would be inconsistent with the positive association we observed between s_i/s_p and selection efficacy. Instead, our results are consistent with rare pseudogenes evolving under a regime closer to neutrality relative to rare intact genes.

Our enrichment test results highlight that a significant proportion of rare accessory genes are under selection. Notably, 19% of ultra-rare intact genes are in COG categories significantly depleted for pseudogenes. We hypothesise that many such genes are under purifying selection, while relaxed purifying selection could account for the observed enrichment of transposons among pseudogenes. The clear enrichment of selenocysteine-containing dehydrogenases could similarly reflect relaxed, or sporadic, purifying selection on these elements, which is interesting

as selenium, selenocysteine's defining component, is sporadically used across the prokaryotic tree[29].

Gene-level selection could also account for certain observations. For instance, the DNA partitioning protein highly enriched in intact ultra-rare genes, COG1192, is a known plasmid-encoded element predicted to be involved with plasmid partitioning[30]. It is possible that there is an ascertainment bias in identifying such genes as intact, because were they pseudogenized or lost the entire plasmid might not be transferred to daughter cells. Similar biases could also account for why prophage and plasmid-associated elements in the mobilome more generally are depleted among pseudogenes, although these elements are also more likely to be adaptive to the host genome[31,32].

Another caveat is that pseudogene diversity can be influenced by many factors, including life history. For instance, obligate intracellular bacteria are characterized by widespread degeneration of their genome, followed by streamlining[33]. Depending on a species' stage in this evolutionary process, its genome could be enriched or depleted for pseudogenes relative to other bacteria. This likely accounts for certain s_i/s_p outliers we observed, such as the obligate intracellular bacteria *Rickettsia prowazekii*, which had the lowest s_i/s_p ratio. Accordingly, our framework could be improved by incorporating per-species parameters of pseudogene gain and loss dynamics.

Despite these caveats, our work highlights the value of using pseudogene diversity as a neutral null[34] for evaluating the evolutionary forces acting upon intact accessory genes. Establishing true neutrality in microbial genomes is challenging[35], but the clear association we identified between dN/dS and s_i/s_p suggests that pseudogene diversity can provide insight into how rare accessory genes evolve. Using this approach, we show that a purely neutral pangenome model can be rejected and identify which types of rare genes, based on their functional annotation and what species encodes them, are more likely to be retained by selection.

Code and data availability

The code used for the analyses in this manuscript is located at https://github.com/gavinmdouglas/pangenome_pseudogene_null and the key datafiles are available on Zenodo (DOI: 10.5281/zenodo.7942837). All analysed genomes are publicly available as part of NCBI RefSeq/GenBank.

Acknowledgements

We would like to thank Louis-Marie Bobay for reading a draft of this manuscript and providing feedback. GMD is supported by a Natural Sciences and Engineering Research Council of Canada (NSERC) Postdoctoral Fellowship. WFD is funded by the Gordon and Betty Moore Foundation. BJS is supported by an NSERC Discovery Grant.

Ethics declarations

The authors declare that they have no competing interests related to the content of this article.

Online Methods

Dataset processing – broad pangenome analysis

We downloaded all genomes used in this study from the Genome Taxonomy Database[18] release 202. We identified all species in this database with at least ten high quality genomes, based on these criteria: (1) marked as passing the minimum information about a metagenome-assembled genome[36] check; (2) CheckM[37] completeness > 98% and contamination < 1%; (3) fewer than 1000 contigs; (4) contig N50 > 5000; (6) fewer than 100,000 ambiguous bases. We also restricted our analyses to genomes in RefSeq (rather than those in GenBank only), except for *Wolbachia pipientis* genomes, which were numerous but primarily limited to GenBank. For species with more than twenty genomes, we randomly sampled down to twenty genomes. We identified 670 species that fit these criteria and downloaded the corresponding genomes. Certain genomes had been relabelled or removed from NCBI since the release of Genome Taxonomy Database release 202, which resulted in a minimum of nine genomes per species (we eliminated two species with fewer than nine genomes). We annotated all genomes with Prokka[38] version 1.14.5 with the –kingdom, --compliant, and –rfam options. We also specified the —metagenome flag for all genomes with 50 or more contigs. We ran Panaroo[39] version 1.3.0 on all output GFFs, with the –remove-invalid-genes and --clean-mode strict options. We then ran Pseudofinder[19] on the Prokka-output GenBank files to identify all putative pseudogenes, using protein sequences from the UniRef90 database[40] (UniProt KB release 2022_01) as a reference database. We restricted the output to intergenic pseudogenes specifically, as the other pseudogene types identified by Pseudofinder correspond to divergent intact coding sequences (in length or modularity), which are difficult to interpret as truly degenerating sequences, and could simply

represent functionally divergent proteins. We performed three filtering steps on the output intergenic pseudogenes. Specifically, we excluded all (1) pseudogene calls within 500 bp of contig ends, (2) pseudogenes of called length < 100 bp or > 5000 bp, and (3) pseudogenes that substantially differed from the mean size of all matching database hits (mean database size – pseudogene size was inclusively required to be between -500 bp and 2000 bp). Pseudogenes were clustered with cd-hit[41] version 4.8.1 with an identity cut-off of 95% over at least 90% of both compared sequences. The mean numbers of genes and singletons per species were identified by repeated subsampling to nine strains per species and then comparing Panaroo gene sets. This procedure was repeated for up to 100 replicates (or until the maximum number of strain combinations was reached) and the mean number of genes and singletons per genome was computed across all replicates. This same procedure was repeated for computing the pseudogene statistics, and the mean percentage of singletons per species was calculated by dividing the mean number of singletons by the mean number of genes per species (and multiplying by 100). To be clear, this computation means that the s_i/s_p metric corresponds to a comparison of the percentage of singleton intact and pseudogene calls overall per species, rather than of calls within each individual genome. Where possible, these commands were parallelized with GNU Parallel[42] version 20161222.

Metric computation

We performed codon-aware multiple-sequence alignment of all ubiquitous and single-copy genes sequences per-species with muscle[43] version 3.8.1551, based on the HyPhy[44] version 2.5.36 codon-aware workflow (https://github.com/veg/hyphy-analyses/tree/master/codon-msa). We then concatenated the core gene alignments per species with a Python script (cat_core_genome_msa.py) and computed pairwise dN/dS and dS for each combination of strain pairs per species with an additional script (mean_pairwise_dnds.py). Both scripts, and the bash commands for running the codon-aware alignments, are available in v1.1.0 of this repository: https://github.com/gavinmdouglas/handy_pop_gen. The latter script identifies potential non-synonymous and synonymous mutation sites between each sequence pair using the NG86 approach[45]. We computed the mean values across all pairwise strain comparisons, resulting in a single measure of dN/dS and dS per species.

Linear models

We built linear models using the lm function in R to predict pangenome diversity, based on (per species) either the mean number of genes, the genomic fluidity, s_i, or s_i/s_p. The predictors included dS, dN/dS, and taxonomic class. Classes with <= 5 member species were collapsed into the "Other" category, which was set as the intercept for the models. One species, *Rickettsia prowazekii*, was excluded from this analysis due to values of zero for s_i and s_i/s_p. We transformed all continuous variables to be normally distributed, except for the mean number of genes, which was already normally distributed. We performed a square-root transformation of the genomic fluidity, s_i, s_i/s_p , and dS values. The dN/dS values were especially right skewed and required a negative inverse transformation (-1 * 1/(x), where x is each dN/dS value) to be normalized. We then converted each continuous variable to standardized units, by mean-centring and dividing by the standard deviation. This step means that the model outputs refer to units of standard deviation per variable, which makes it possible to compare the magnitude of coefficients across models with different response variables.

Dataset processing – In-depth pangenome analysis

We conducted a subsequent analysis on 10 bacterial species with a relatively high number of genomes (ranging from 135-6,916). We selected these species from our original set as those with > 100 genomes that were not phylogenetically redundant. For these data, we clustered both intact genes and pseudogenes with cd-hit, using the same settings as above. This clustering was performed on all genes and pseudogenes across all ten species. We functionally annotated each resulting cluster with COG IDs and categories[25] using eggNOG-mapper[24] version 2.1.6 (based on eggNOG orthology data[46] version 5.0.2) with DIAMOND[47] version 2.0.14 and these parameter options: --score 60, --pident 40, --query_cover 20, --subject_cover 20, --tax_scope auto, and --target_orthologs all. This was performed for individual elements separately (i.e. the original sequences rather than the cluster representatives), and for database sequence matches to pseudogene hits. We used majority rule of all member sequences per cluster to assign individual COG IDs and categories, and the same approach for assigning functions to individual pseudogene sequences based on database sequence annotations. We manually assigned COG categories based on a mapping of COG IDs from the COG 2020 database release. This was

performed as the raw output COG categories were based on an earlier version of the database that did not include mobilome (category X) annotations.

Generalized linear mixed models

Generalized linear mixed models were fit in R using the glmmTMB[48] package v1.1.5, one for the ultra-rare, other-rare, and shell pangenome partitions, respectively. Only COG-annotated elements were included in these models, excluding those annotated by the (rare) A, B, Y, and Z COG categories only. We used the binomial family and nlminb optimization algorithm with 1000 set for both iter.max and eval.max. The full R-style formula for each model was:

pseudogene ~ COG-category + non-redundant-status + COG-category: non-redundant-status + (1 | species) + (1 | COG-category: species) + (1 | non-redundant-status:species)

In this formula, random effects are specified as those in parentheses including "1|" and interaction terms are indicated with ":". The response was a Boolean variable indicating whether each element is a pseudogene. The COG-category variable is categorical indicating the one-letter COG category code that each element belongs to. In cases where elements were members of multiple categories, duplicate rows were created for each category. The Transcription category (K) was selected as the first level, to be used for the intercept, as it was the most consistently abundant COG category across all three partitions (third in the other-rare and shell, and fourth in ultra-rare). The non-redundant-status variable was a Boolean variable indicating whether each element was not redundant with another intact element of the same COG ID (gene family, not category) in the same genome. This negative formulation of redundancy (i.e. whether an element is not redundant, rather than whether it is redundant) was chosen as most elements were redundant, and so we decided to set the default level in each model (False) to be more representative. The species variable corresponded to the name of the species encoding each element.

We also fit simpler models with subsets of these variables and computed Akaike Information Criterion (AIC) values for each model, that allowed us to compare across models and investigate whether more complex models provide significantly more information. We visualized the AICs

per model based on normalized scores that transformed the minimum model AIC per partition to be 0 and the maximum model AIC per partition to be 1.

Finally, for each significant COG category in the ultra-rare generalized linear model (excluding those interacting with non-redundancy), we systematically tested whether individual COG IDs were enriched for pseudogenes based on Fisher's exact tests comparing the number of pseudogene and intact genes within each COG ID (and with the same redundancy status and in the same species) compared to the background of all other elements with the same redundancy status in the same species.

General analyses

No tests for statistical power were conducted to determine the sample sizes required for this study, but we used genomes from all available species in the Genome Taxonomy Database of sufficient quality. All analyses were conducted in R v4.2.2. Figures were generated with ggplot2[49] v3.4.0, with the exception of the heatmaps, which were created with the ComplexHeatmap[50] package v2.14.0.

References

1. Tettelin, H. *et al.* Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: Implications for the microbial 'pan-genome'. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 3950–13955 (2005).
2. Vos, M., Hesselman, M. C., te Beek, T. A., van Passel, M. W. J. & Eyre-Walker, A. Rates of Lateral Gene Transfer in Prokaryotes: High but Why? *Trends Microbiol.* **23**, 598–605 (2015).
3. Innamorati, K. A., Earl, J. P., Aggarwal, S. D., Ehrlich, G. D. & Hiller, N. L. The Bacterial Guide to Designing a Diversified Gene Portfolio. in *The Pangenome: Diversity, Dynamics and Evolution of Genomes* (eds. Tettelin, H. & Medini, D.) 51–87 (Springer, 2020). doi:10.1007/978-3-030-38281-0_3.

4. Novick, A. & Doolittle, W. F. Horizontal persistence and the complexity hypothesis. *Biol. Philos.* **35**, 2 (2020).
5. Sela, I., Wolf, Y. I. & Koonin, E. V. Theory of prokaryotic genome evolution. *Proc. Natl. Acad. Sci. U. S. A.* **113**, 11399–11407 (2016).
6. Andreani, N. A., Hesse, E. & Vos, M. Prokaryote genome fluidity is dependent on effective population size. *ISME J.* **11**, 1719–1721 (2017).
7. Bobay, L. M. & Ochman, H. Factors driving effective population size and pan-genome evolution in bacteria. *BMC Evol. Biol.* **18**, 153 (2018).
8. Haegeman, B. & Weitz, J. S. A neutral theory of genome evolution and the frequency distribution of genes. *BMC Genomics* **13**, 196 (2012).
9. Lobkovsky, A. E., Wolf, Y. I. & Koonin, E. V. Gene frequency distributions reject a neutral model of genome evolution. *Genome Biol. Evol.* **5**, 233–242 (2013).
10. Fiona J Whelan, Rebecca J Hall, & James O McInerney. Evidence for Selection in the Abundant Accessory Gene Content of a Prokaryote Pangenome. *Mol. Biol. Evol.* **38**, 3697–3708 (2021).
11. N'Guessan, A., Brito, I. L., Serohijos, A. W. R. & Shapiro, J. Mobile Gene Sequence Evolution within Individual Human Gut Microbiomes Is Better Explained by Gene-Specific Than Host-Specific Selective Pressures. *Genome Biol. Evol.* **13**, (2021).
12. Wolf, Y. I., Makarova, K. S., Lobkovsky, A. E. & Koonin, E. V. Two fundamentally different classes of microbial genes. *Nat. Microbiol.* **2**, 1–6 (2016).
13. Boucher, Y. *et al.* Local Mobile Gene Pools Rapidly Cross Species Boundaries To Create Endemicity within Global *Vibrio cholerae* Populations. *mBio* **2**, e00335-10 (2011).

14. Niehus, R., Mitri, S., Fletcher, A. G. & Foster, K. R. Migration and horizontal gene transfer divide microbial genomes into multiple niches. *Nat. Commun.* **6**, 8924 (2015).
15. Smillie, C. S. *et al.* Ecology drives a global network of gene exchange connecting the human microbiome. *Nature* **480**, 241–244 (2011).
16. Danneels, B., Pinto-Carbó, M. & Carlier, A. Patterns of nucleotide deletion and insertion inferred from bacterial pseudogenes. *Genome Biol. Evol.* **10**, 1792–1802 (2018).
17. Kuo, C. H. & Ochman, H. The extinction dynamics of bacterial pseudogenes. *PLoS Genet.* **6**, e1001050 (2010).
18. Parks, D. H. *et al.* A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nat. Biotechnol.* **36**, 996–1004 (2018).
19. Syberg-Olsen, M. J., Garber, A. I., Keeling, P. J., McCutcheon, J. P. & Husnik, F. Pseudofinder: Detection of Pseudogenes in Prokaryotic Genomes. *Mol. Biol. Evol.* **39**, msac153 (2022).
20. Kislyuk, A. O., Haegeman, B., Bergman, N. H. & Weitz, J. S. Genomic fluidity: An integrative view of gene diversity within microbial populations. *BMC Genomics* **12**, (2011).
21. Lerat, E. & Ochman, H. Ψ-Φ: Exploring the outer limits of bacterial pseudogenes. *Genome Res.* **14**, 2273–2278 (2004).
22. Rocha, E. P. C. *et al.* Comparisons of dN/dS are time dependent for closely related bacterial genomes. *J. Theor. Biol.* **239**, 226–235 (2006).
23. Kryazhimskiy, S. & Plotkin, J. B. The Population Genetics of dN/dS. *PLoS Genet.* **4**, e1000304 (2008).

571 24. Cantalapiedra, C. P., Hernández-Plaza, A., Letunic, I., Bork, P. & Huerta-Cepas, J. eggNOG-
572 mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the
573 Metagenomic Scale. *Mol. Biol. Evol.* **38**, 5825–5829 (2021).
574 25. Galperin, M. Y. *et al.* COG database update: focus on microbial diversity, model organisms,
575 and widespread pathogens. *Nucleic Acids Res.* **49**, D274–D281 (2021).
576 26. McInerney, J. O., McNally, A. & O'Connell, M. J. Why prokaryotes have pangenomes. *Nat.*
577 *Microbiol.* **2**, 170402 (2017).
578 27. Hottes, A. K. *et al.* Bacterial Adaptation through Loss of Function. *PLoS Genet.* **9**, e1003617
579 (2013).
580 28. Oren, Y. *et al.* Transfer of noncoding DNA drives regulatory rewiring in bacteria. *Proc. Natl.*
581 *Acad. Sci. U. S. A.* **111**, 16112–16117 (2014).
582 29. Peng, T., Lin, J., Xu, Y.-Z. & Zhang, Y. Comparative genomics reveals new evolutionary
583 and ecological patterns of selenium utilization in bacteria. *ISME J.* **10**, 2048–2059 (2016).
584 30. A Schlüter *et al.* Erythromycin Resistance-Conferring Plasmid pRSB105, Isolated from a
585 Sewage Treatment Plant, Harbors a New Macrolide Resistance Determinant, an Integron-
586 Containing Tn402-Like Element, and a Large Region of Unknown Function. *Appl. Environ.*
587 *Microbiol.* **73**, (2007).
588 31. Bobay, L. M., Rocha, E. P. C. & Touchon, M. The adaptation of temperate bacteriophages to
589 their host genomes. *Mol. Biol. Evol.* **30**, 737–751 (2013).
590 32. McKerral, J. C. *et al.* The Promise and Pitfalls of Prophages. *bioRxiv* 2023.04.20.537752
591 (2023) doi:10.1101/2023.04.20.537752.
592 33. Giovannoni, S. J., Cameron Thrash, J. & Temperton, B. Implications of streamlining theory
593 for microbial ecology. *ISME J.* **8**, 1553–1565 (2014).

34. Koonin, E. V. Splendor and misery of adaptation, or the importance of neutral null for understanding evolution. *BMC Biol.* **14**, 114 (2016).

35. Rocha, E. P. C. Neutral Theory, Microbial Practice: Challenges in Bacterial Population Genetics. *Mol. Biol. Evol.* **35**, 1338–1347 (2018).

36. Bowers, R. M. *et al.* Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nat. Biotechnol.* **35**, 725–731 (2017).

37. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* **25**, 1043–55 (2015).

38. Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068–2069 (2014).

39. Tonkin-Hill, G. *et al.* Producing polished prokaryotic pangenomes with the Panaroo pipeline. *Genome Biol.* **21**, 180 (2020).

40. The UniProt Consortium. The Universal Protein Resource. *Nucleic Acids Res.* **36**, D190–D195 (2008).

41. Li, W. & Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**, 1658–1659 (2006).

42. Tange, O. GNU Parallel: the command-line power tool. *Login USENIX Mag.* **36**, 42–47 (2011).

43. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).

616 44. Kosakovsky Pond, S. L. *et al.* HyPhy 2.5—A Customizable Platform for Evolutionary
617 Hypothesis Testing Using Phylogenies. *Mol. Biol. Evol.* **37**, 295–299 (2020).
618 45. Nei, M. & Gojobori, T. Simple methods for estimating the numbers of synonymous and
619 nonsynonymous nucleotide substitutions. *Mol. Biol. Evol.* **3**, 418–426 (1986).
620 46. Huerta-Cepas, J. *et al.* eggNOG 5.0: A hierarchical, functionally and phylogenetically
621 annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res.*
622 **47**, D309–D314 (2019).
623 47. Buchfink, B., Reuter, K. & Drost, H.-G. Sensitive protein alignments at tree-of-life scale
624 using DIAMOND. *Nat. Methods* **18**, 366–368 (2021).
625 48. Brooks, M., E. *et al.* glmmTMB Balances Speed and Flexibility Among Packages for Zero-
626 inflated Generalized Linear Mixed Modeling. *R J.* **9**, 378 (2017).
627 49. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis*. (Springer-Verlag New York,
628 2016).
629 50. Zuguang Gu. Complex heatmap visualization. *iMeta* **1**, e43 (2022).
630