



HAL
open science

Unlinked rRNA genes are widespread among Bacteria and Archaea

Tess E Brewer, Mads Albertsen, Arwyn Edwards, Rasmus H Kirkegaard, Eduardo P C Rocha, Noah Fierer

► **To cite this version:**

Tess E Brewer, Mads Albertsen, Arwyn Edwards, Rasmus H Kirkegaard, Eduardo P C Rocha, et al.. Unlinked rRNA genes are widespread among Bacteria and Archaea. The International Society of Microbiological Ecology Journal, 2020, 10.1038/s41396-019-0552-3 . hal-02363135

HAL Id: hal-02363135

<https://hal.science/hal-02363135>

Submitted on 14 Nov 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution - NonCommercial 4.0 International License

1 Unlinked rRNA genes are widespread among Bacteria and Archaea

2

3 Authors: Tess E. Brewer^{a,b,*}, Mads Albertsen^c, Arwyn Edwards^d, Rasmus H. Kirkegaard^c,

4 Eduardo P. C. Rocha^e, Noah Fierer^{a,f}

5

6 ^a Cooperative Institute for Research in Environmental Sciences, University of Colorado,
7 Boulder, CO 80309 USA

8 ^b Current Address: Department of Evolutionary Biology and Environmental Studies,
9 University of Zürich, Zürich, Switzerland

10 ^c Department of Chemistry and Bioscience, Aalborg University, 9220 Aalborg, Denmark

11 ^d Institute of Biological, Environmental and Rural Sciences, Aberystwyth University SY23
12 3DA UK

13 ^e Microbial Evolutionary Genomics, Institut Pasteur, CNRS, UMR3525, Paris, 75015, France.

14 ^f Department of Ecology and Evolutionary Biology, University of Colorado, Boulder, CO
15 80309 USA

16

17 * Corresponding author:

18 Tess Brewer

19 University of Zürich

20 Building Y27-J-54

21 Winterthurerstrasse 190

22 Zürich, Switzerland 8057

23 Email: tess.brewer@uzh.ch

24

25 Abstract

26 Ribosomes are essential to cellular life and the genes for their RNA components are
27 the most conserved and transcribed genes in Bacteria and Archaea. Ribosomal rRNA genes
28 are typically organized into a single operon, an arrangement thought to facilitate gene
29 regulation. In reality, some Bacteria and Archaea do not share this canonical rRNA
30 arrangement - their 16S and 23S rRNA genes are separated across the genome and referred
31 to as "unlinked". This rearrangement has previously been treated as an anomaly or a
32 byproduct of genome degradation in intracellular bacteria. Here, we leverage complete
33 genome and long-read metagenomic data to show that unlinked 16S and 23S rRNA genes
34 are more common than previously thought. Unlinked rRNA genes occur in many phyla,
35 most significantly within Deinococcus-Thermus, Chloroflexi, and Planctomycetes, and
36 occur in differential frequencies across natural environments. We found that up to 41% of

37 rRNA genes in soil were unlinked, in contrast to the human gut, where all sequenced rRNA
38 genes were linked. The frequency of unlinked rRNA genes may reflect meaningful life
39 history traits, as they tend to be associated with a mix of slow-growing free-living species
40 and intracellular species. We speculate that unlinked rRNA genes may confer selective
41 advantages in some environments, though the specific nature of these advantages remains
42 undetermined and worthy of further investigation. More generally, the prevalence of
43 unlinked rRNA genes in poorly-studied taxa serves as a reminder that paradigms derived
44 from model organisms do not necessarily extend to the broader diversity of Bacteria and
45 Archaea.

46

47 Introduction

48 Ribosomes are the archetypal “essential proteins”, so much so that they are a key
49 criteria in the division between cellular and viral life (1). In Bacteria and Archaea, the genes
50 encoding the RNA components of the ribosome are traditionally arranged in a single
51 operon in the order 16S - 23S - 5S. The rRNA operon is transcribed into a single RNA
52 precursor called the pre-rRNA 30S, which is separated and processed by a number of
53 RNases (2). This arrangement of rRNA genes within a single operon is thought to allow
54 rapid responses to changing growth conditions - the production of rRNA under a single
55 promoter allows consistent regulation and conservation of stoichiometry between all
56 three, essential components (3). Indeed, the production of rRNA is the rate-limiting step of
57 ribosome synthesis (4), and fast-growing Bacteria and Archaea accelerate ribosome
58 synthesis by encoding multiple rRNA operons (5).

59 Some Bacteria and Archaea have “unlinked” rRNA genes, where the 16S and 23S
60 rRNA genes are separated by large swaths of genomic space (Figure 1). This unlinked rRNA
61 gene arrangement was first discovered in the thermophilic bacterium *Thermus*
62 *thermophilus* (6). Reports of unlinked rRNA genes soon followed in additional Bacteria,
63 including the planctomycete *Pirellula marina* (7), the aphid endosymbiont *Buchnera*
64 *aphidicola* (8), and the intracellular pathogen *Rickettsia prowazekii* (9). Though unlinked
65 rRNA genes were first discovered in a free-living environmental bacterium, their ubiquity
66 among the order Rickettsiales has led to suggestions that unlinked rRNA genes are a result
67 of the genome degradation typical of obligate intracellular lifestyles (10-12).

68 With this study we sought to determine the frequency of unlinked rRNA genes
69 across Bacteria and Archaea and whether this unique genomic feature is largely confined to
70 those Bacteria and Archaea with an obligate intracellular lifestyle. We examined the rRNA
71 genes of over 10,000 publicly available complete bacterial and archaeal genomes to
72 identify which taxa have unlinked rRNA genes and to determine if there are any genomic
73 characteristics shared across taxa with this feature. As complete genomes are not typically
74 available for the broader diversity of Bacteria and Archaea found in environmental samples
75 (13), we also characterized rRNA gene arrangements using long-read metagenomic
76 datasets obtained from a range of environmental samples, which together encompassed
77 over 17 million sequences (≥ 1000 bp). With these long-read metagenomic datasets, we
78 were able to determine whether unlinked rRNA genes are common in environmental
79 populations and how the distributions of unlinked rRNA genes differ across prokaryotic
80 lineages and across distinct microbial habitats.

81

82 Methods

83 Analyses of complete genomes

84 We downloaded all bacterial and archaeal genomes in the RefSeq genome database
85 (14) classified with the assembly level “Complete Genome” from NCBI in January 2019
86 (12539 genomes). We removed genomes from consideration that had rRNA genes that
87 were split across the genome start and end (96 genomes), >20 reported rRNA genes (2
88 genomes), or an unequal number of 16S and 23S rRNA genes (219 genomes). This left us
89 with a set of 12222 genomes. We used gene ranges associated with each open reading
90 frame (ORF) to pair the 16S and 23S rRNA genes that were closest to each other in each
91 genome. We then checked for gene directionality (sense/antisense) and calculated the
92 distance between each pair, taking directionality into account (see Supplemental Figure S1
93 for more detail and a visual representation). rRNA pairs were classified as ‘unlinked’ if the
94 distance between each gene was greater than 1500 bp, ‘linked’ if the distance was less than
95 or equal to 1500 bp. We separated genomes that had a 16S or 23S rRNA gene that started
96 or ended within 1500 bp of the beginning or end of its genome and classified these 226
97 genomes independently to account for the circular nature of bacterial and archaeal
98 genomes. For this subset of genomes, we iteratively adjusted the start and end position of

99 those “edge-case” rRNA genes with respect to genome size and selected the smallest
100 distance between the 16S and 23S rRNA genes as the true distance, using the same formula
101 presented in Supplemental Figure 1. Each genome was classified as 'unlinked', 'linked', or
102 'mixed' depending on the status of their rRNA genes with 'mixed' genomes having multiple
103 rRNA copies with a combination of linked and unlinked rRNA genes. We re-assigned
104 taxonomy to each genome using the SILVA 132 SSU database (clustered at 99%) to
105 maintain a consistent taxonomy between our two datasets. All analyses were done in R
106 version 3.5.1 (15). Information on all genomes included in these analyses (including
107 classification of rRNA genes) is available in Supplemental Dataset S1.

108

109 Long-read metagenomic analyses

110 To investigate the prevalence of unlinked rRNA genes among those Bacteria and
111 Archaea found in environmental samples (including many taxa for which genomes are not
112 yet available), we analyzed long-read metagenomic datasets generated from soil, sediment,
113 activated sludge, anaerobic digesters, and human gut samples. These metagenomic
114 datasets were generated using either the Oxford Nanopore MinION/PromethION (6
115 samples) or the Illumina synthetic long-read sequencing technology (also known as
116 Moleculo, first described in (16), 9 samples). The Moleculo sequences originated from four
117 previously published studies covering: the human gut (17), prairie soil (18), sediment
118 (19), and grassland soils (MG-RAST project mgp14596, (20)). The Nanopore sequences
119 originated from four unpublished studies spanning a diverse range of environment types:
120 anaerobic digesters, activated sludge, sediment, and lawn soil. For these samples, DNA was
121 extracted using DNeasy PowerSoil Kits (Qiagen, DE) and libraries were prepared for
122 sequencing using the LSK108 kit (Oxford Nanopore Technologies, UK) following the
123 manufacturers protocol. The libraries were sequenced on either the MinION or the
124 PromethION sequencing platforms (Oxford Nanopore Technologies, UK). Base calling was
125 conducted using Albacore v. 2.1.10 for the lawn soil sample (VCsoil) and Albacore v. 2.3.1
126 for all other samples (Oxford Nanopore Technologies, UK). Across these 15 samples, we
127 compiled 16,870,533 Nanopore sequences and 846,437 Moleculo sequences with a
128 minimum read length of 1000 bp.

129 We trimmed the first 250 bp of each Nanopore sequence to remove low quality

130 regions, but performed no other quality filtering as not all samples included information on
131 sequence quality (some sequences were fasta format). Instead, we relied on our
132 downstream filtering steps to remove sequences of poor quality. Metaxa2 version 2.1 (21)
133 was run on all sequences with default settings to search for SSU (16S rRNA) and LSU (23S
134 rRNA) gene fragments. Taxonomy was assigned to the partial rRNA sequences using the
135 RDP classifier (22) and the SILVA 132 SSU and LSU databases (both clustered at 99%
136 sequence identity, 23). If a sequence contained both 16S and 23S rRNA genes we used the
137 taxonomy with the highest resolution (if the 16S was annotated to family level while the
138 23S was genus level, we used the 23S taxonomy for both rRNAs). Details on each sample,
139 including number of reads and median read lengths, are available in Supplemental Table
140 S1.

141 We next used a number of criteria to filter the reads included in downstream analyses
142 and to identify taxa with unlinked rRNA genes. We only included those reads in our final
143 dataset that met the following criteria:

- 144 1) Contained a 16S rRNA gene (to avoid potentially double counting organisms with
145 unlinked 16S and 23S rRNA genes),
- 146 2) Included the last two domains of the 16S rRNA gene (V8|V9) (Metaxa2 uses
147 multiple Hidden Markov Model (HMM) profiles targeting conserved regions of
148 rRNA genes, each of these regions is referred to as a domain),
- 149 3) The length of the 16S rRNA gene was ≤ 4000 bp and the length of the 23S rRNA
150 gene (if present) was ≤ 6800 bp. These thresholds were chosen to remove
151 erroneously long rRNA genes while accommodating insertions within rRNA genes
152 such as those that occur in Candidate Phyla Radiation (CPR) taxa (24), *Nostoc*,
153 *Salmonella*, and others (25),
- 154 4) Could be classified to at least the phylum level of taxonomic resolution.

155
156 Of the subset of reads that met these criteria (112 - 878 per Moleculo sample, 3817 - 28056
157 per Nanopore sample, see Supplemental Table S1 for details), we classified reads as
158 containing unlinked rRNA genes if there was >1500 bp between the 16S and 23S rRNA
159 genes, or if there was no 23S domain found 1500 bp after the end of the 16S rRNA. We note
160 that, unlike the NCBI gene ranges, Metaxa2 takes strand information into account and

161 translates start and stop locations into sense orientation for SSU and LSU. For our final
162 analyses, we removed reads that could not be classified as linked or unlinked rRNA genes
163 (for instance a sequence with only 300 bp after the 3' end of the 16S rRNA gene). All
164 analyses were done in R version 3.5.1 (15). Information on all long-read sequences
165 included in these analyses (including classification of rRNA genes) is available in
166 Supplemental Dataset S2.

167

168 Phylogenetic tree combining long-read and NCBI datasets

169 A phylogenetic tree was created from full-length 16S rRNA gene sequences by
170 combining both the NCBI complete genomes and representatives of the long-read
171 metagenomic datasets. For the NCBI genome sequences, we selected one 16S rRNA gene
172 sequence per unique species. For the long-read datasets, we first matched the partial 16S
173 rRNA genes recovered by metaxa2 (21) to full-length 16S rRNA gene sequences in the
174 SILVA 132 SSU database (23) using the usearch10 version 10.0.240 command
175 usearch_global (settings: -id 0.95 -strand both -maxaccepts 0 -maxrejects 0; 26). The full-
176 length SILVA 16S rRNA genes sequences that matched to the long-read sequences $\geq 95\%$
177 percent identity and ≥ 500 bp alignment length were used as representatives of their long-
178 read sequence match. We used 95% percent identity as our cutoff as we found unlinked
179 rRNA gene status to generally be conserved within genera (see below and Supplemental
180 Figure S2). The NCBI and SILVA sequences were then aligned with PyNAST version 0.1
181 (27) and the phylogenetic tree was constructed using FastTree version 2.1.10 SSE3 (28),
182 and plotted with iTOL (29).

183

184 Genomic attributes associated with unlinked rRNA genes

185 All tests for genomic attributes were done with a subset of our complete genome
186 dataset - we reduced the dataset to include only one representative genome per unique
187 species and operon status. For example, if a species had 24 genomes with linked rRNA
188 genes and 3 genomes with unlinked rRNA genes, we retained two genomes total, one linked
189 and one unlinked. Species with heterogeneous rRNA gene status accounted for only 0.71%
190 of species and we found that the presence of unlinked rRNA genes was strongly conserved
191 at the species and genus level (Supplemental Figure S2).

192 With this set of reduced genomes (3967 genomes in total), we first calculated Pagel's
193 lambda (30) to determine whether there was a phylogenetic signal associated with
194 unlinked rRNA genes using the phylosig function of the phytools package version 0.6.60
195 (31). The results of this test indicated there was a strong phylogenetic signal (lambda =
196 0.96, $p < 0.0001$), so we controlled for phylogeny in all of our subsequent tests by using a
197 Phylogenetic Generalized Linear Model for continuous variables (with the function
198 phyloglm in the phylolm package version 2.6; 32).

199 To determine if taxa with unlinked rRNA genes have a lower predicted growth rate,
200 we calculated the codon usage proxy $\Delta\text{ENC}'$ (33,34), which provides an estimate of
201 minimum generation times (35). We calculated $\Delta\text{ENC}'$ with the program ENCprime (33)
202 with default options, on both the concatenated ORF sequences and concatenated ribosomal
203 protein sequences for each genome following Vieira-Silva and Rocha (2009). To determine
204 if RNaseIII was present in each genome, we used HMMER version 3.1b2 (36) to search for
205 three RNaseIII pfams (bacterial PF00636, PF14622, and archaeal PF11469) in the
206 translated protein files of each genome. We used the gathering thresholds (GA) associated
207 with each of these pfams to set all cutoffs and reduce the likelihood of false positives (--
208 cut_ga).

209

210 Results

211 Unlinked rRNA genes occur frequently in complete genomes

212 We used a set of 12222 "complete" bacterial and archaeal genomes extracted from
213 NCBI in January 2019 to determine how frequently unlinked 16S and 23S rRNA genes
214 occur. We analyzed the distribution of distances between the closest edges of the closest
215 pairs of 16S and 23S rRNA genes (known as the Internally Transcribed Spacer - ITS) in
216 each genome and found that the vast majority of 16S and 23S rRNA gene pairs (98.7%) had
217 an ITS ≤ 1500 bp with an average ITS length of 418.7 bp (± 169.7 bp, Figure 2A). However,
218 pairs with ITS lengths > 1500 bp showed a scattered distribution of distances, with an
219 average ITS length of 410374 bp (± 521792 bp). Hence, for this classification scheme we
220 called rRNA genes "unlinked" if the ITS was greater than 1500 bp in length. This 1500 bp
221 threshold is in some ways conservative, as the distance between genes in an operon is

222 usually quite low - peaking between 20 and 30 bp in most genomes (37). Additionally,
223 tRNA are the most common genes found in the space between the 16S and 23S rRNA genes,
224 and range from only 75 to 90 bp in length (38).

225 After classifying each rRNA gene pair as linked or unlinked based on the distance
226 between the 16S and 23S rRNA genes, we found that 3.65% of the genomes in our dataset
227 had exclusively unlinked rRNA genes, 0.62% had mixed rRNA gene status (i.e. genomes
228 with multiple rRNA copies that had at least one set of unlinked rRNA genes and at least one
229 canonical, linked rRNA operon), and 95.73% had exclusively linked operons (these
230 numbers do not match up with the per rRNA gene dataset as each genome has a variable
231 rRNA copy number). We found unlinked genomes to be relatively common (present in $\geq 5\%$
232 of members) in taxa characterized as having an obligate intracellular lifestyle within the
233 phyla Spirochaetes (genus *Borrelia*), Epsilonbacteraeota (family Helicobacteraceae),
234 Alphaproteobacteria (order Rickettsiales), and Tenericutes (species *Mycoplasma*
235 *gallisepticum*). However, we also found high proportions of unlinked rRNA genes in phyla
236 that are generally considered to be free-living, such as Deinococcus-Thermus (families
237 Thermaceae and Deinococcaceae), Chloroflexi (family Dehalococcoidaceae),
238 Planctomycetes (families Phycisphaeraceae and Planctomycetaceae), and Euryarchaeota
239 (class Thermoplasmata). Phyla with at least 5% of genomes having exclusively unlinked
240 rRNA genes are shown in Figure 2C.

241

242 Unlinked rRNA genes are widespread in environmental metagenomic data

243 While the results from our complete genome dataset demonstrate that unlinked
244 rRNA genes are common in some putatively free-living phyla, databases featuring complete
245 genomes do not capture the full breadth of microbial diversity and are heavily biased
246 towards cultivated organisms relevant to human health (13). Just three phyla
247 (Proteobacteria, Firmicutes, Actinobacteria) accounted for $>83\%$ of the genomes in our
248 NCBI dataset - even though recent estimates of bacterial diversity total at least 99 unique
249 phyla (39). To investigate the ubiquity of unlinked rRNA genes among those taxa
250 underrepresented in 'complete' genome databases, we analyzed long-read metagenomic
251 data from a range of distinct sample types. Focusing exclusively on long-read sequences

252 allowed us to span the 1500 bp distance required for classification of rRNA genes without
253 the need for assembly. This is important as the repetitive structure of rRNA genes makes it
254 difficult to assemble a mix of non-identical rRNA genes from the short reads typical of most
255 current metagenomic sequencing projects (40).

256 From our initial long-read dataset encompassing 15 unique samples (~890,000
257 Illumina synthetic long reads (also known as Molecuro) and ~19 million Nanopore reads,
258 with median read lengths of 8858 bp and 5398 bp, respectively), only 15855 sequences
259 contained rRNA genes and met the criteria we established for the classification of rRNA
260 genes as linked or unlinked (see Methods). Of these reads, we classified 1607 as unlinked,
261 or 10.1% of the dataset (Figure 2B). These long-read metagenomic analyses showed that
262 unlinked rRNA genes are not equally distributed across environments - we found that up to
263 41% of the taxa in soil had unlinked rRNA genes, whereas other environments had much
264 lower proportions, most notably the human gut, where all sequenced rRNA genes were
265 linked (Figure 3).

266 The results from our analyses of the long-read dataset generally mirrored the
267 corresponding results from the complete genome dataset, in that many of the long reads
268 classified as unlinked belonged to the same phyla where unlinked rRNA genes were
269 prevalent in the complete genome dataset (Figure 2). The long-read metagenomic dataset
270 confirmed that members of the phyla Deinococcus-Thermus, Planctomycetes, Chloroflexi,
271 Spirochaetes, and Euryarchaeota frequently have unlinked rRNA genes (Figure 2B). The
272 long-read dataset also allowed us to provide additional evidence for unlinked rRNA genes
273 in poorly studied phyla that were represented by only a handful of genomes in our
274 complete genome dataset, such as candidate phyla Acetothermia (1 genome and 64 long-
275 read sequences) and Patescibacteria (3 genomes and 330 long-read sequences).

276 Using the long-read dataset, we identified 18 additional phyla where unlinked rRNA
277 genes are common, including several candidate phyla (BRC1, GAL15, WS1, WS2) and
278 members of the Candidate Phyla Radiation (Patescibacteria, Figure 2). We also found
279 several clades with high proportions of unlinked rRNA genes that had no representation in
280 our complete genome dataset, including Rikenellaceae RC9 gut group (334/624),
281 Verrucomicrobia genus *Candidatus Udaeobacter* (80/80), Atribacteria order
282 Caldatribacteriales (37/37), Cyanobacteria order Obscuribacterales (4/4), Acidobacteria

283 Subgroup 2 (27/27), Planctomycetes order MSBL9 (40/40), and Chloroflexi class GIF9
284 (7/7). Overall, we found that 52% of the phyla covered in our combined datasets (37/71)
285 have at least one representative with unlinked rRNA genes.

286

287 Unlinked rRNA genes are strongly conserved

288 We found that taxa with unlinked rRNA genes are not randomly distributed across
289 bacterial and archaeal lineages - rather, we observed a strong phylogenetic signal for this
290 trait, which we confirmed by calculating Pagel's lambda ($\lambda = 0.96$, $p > 0.001$). To
291 highlight this point, we assembled a phylogenetic tree from full-length 16S rRNA gene
292 sequences representing both the complete genome dataset and the long-read metagenomic
293 dataset. We found clusters of related taxa with exclusively unlinked rRNA genes (Figure 4)
294 including: Euryarchaeota class Thermoplasmata, the vast majority of Deinococcus-
295 Thermus, CPR division Patescibacteria, Verrucomicrobia DA101 group, Chloroflexi class
296 Dehalococcoidia, and Alphaproteobacteria class Rickettsiales.

297

298 Genomic attributes associated with unlinked rRNA genes

299 Given that there are numerous bacterial and archaeal lineages where unlinked rRNA
300 genes are commonly observed, we next sought to determine what other genomic features
301 may be associated with this non-standard rRNA gene arrangement. We treated the
302 presence of unlinked rRNA genes as a binary trait - if a genome had at least one unlinked
303 rRNA gene we counted the genome as "unlinked". In our NCBI complete genome dataset,
304 we found rRNA gene status to be conserved strongly at the species level - meaning that the
305 majority of species had either exclusively linked or unlinked rRNA genes among their
306 members (Supplemental Figure S2). Therefore, for the following tests, we used a subset of
307 our NCBI complete genome dataset - retaining only a single representative of each species,
308 unless the species had heterogeneous rRNA gene status (0.71% of species), in which case
309 we retained one genome of each rRNA gene status. The analyses were corrected in order to
310 account for the effect of phylogenetic structure in the data (see Methods).

311 Historically, unlinked rRNA genes have been strongly associated with the reduced
312 genomes of obligate intracellular bacteria, implying that this trait may merely be a side
313 effect of the strong genetic drift and weak selection these taxa experience (10-12). To test

314 this hypothesis, we compared the genome sizes of species with linked and unlinked rRNA
315 genes using Phylogenetic Generalized Linear Models (phyloglm). While we found that
316 genomes with unlinked rRNA genes had smaller genomes on average, this difference was
317 not significant (Figure 5, phyloglm $p=0.12$, means of groups: 4.15 Mbp linked, 2.72 Mbp
318 unlinked).

319 The organization of rRNA genes within the same operon facilitates their joint
320 regulation and co-expression at precise stoichiometric ratios. Selection for this trait is
321 expected to be stronger in faster growing Bacteria and Archaea, where, at maximum
322 growth rates, synthesis of the ribosome is the cell's chief energy expenditure (4). To test
323 this hypothesis, we analyzed the association between the linkage of rRNA genes and traits
324 related to rapid growth in Bacteria and Archaea. On average, genomes with unlinked rRNA
325 genes had significantly fewer rRNA copies (Figure 5, phyloglm $p < 0.0001$, means of
326 groups: 4.25 copies linked, 2.72 copies unlinked). We also calculated $\Delta\text{ENC}'$ for each
327 complete genome - a measure of codon usage bias that is negatively correlated with
328 minimum generation time in Bacteria and Archaea (35). Interestingly, genomes with
329 unlinked rRNA genes were predicted to have significantly longer minimal generation times
330 (Figure 5, phyloglm $p=0.028$, means of groups: 0.23 linked, 0.18 unlinked). Additionally, in
331 our long-read dataset we found that unlinked rRNA genes were more common in
332 environments typified by slow growth rates; soil and sediment samples had higher
333 proportions of unlinked rRNA genes than samples from anaerobic digesters and the human
334 gut (Figure 3).

335 RNaseIII separates the precursors of the 16S and 23S rRNA from their common
336 transcript for subsequent maturation and inclusion in the ribosome (2). RNaseIII is not an
337 essential protein in most Bacteria and Archaea, and several phyla in which unlinked rRNA
338 genes are common do not encode RNaseIII (e.g. *Deinococcus-Thermus* and Euryarchaeota;
339 41). Therefore, we checked if there was a significant association between unlinked rRNA
340 genes and the presence of RNaseIII genes. Interestingly, we found that genomes with
341 unlinked rRNA genes were significantly less likely to encode the bacterial form of RNaseIII
342 genes (Figure 5 and Supplemental Figure S3, PF00636: phyloglm $p < 0.001$, means of
343 groups: 1.0 linked, 0.71 unlinked; PF14622: phyloglm $p = 0.007$, means of groups: 0.86
344 linked, 0.66 unlinked). We were unable to check this relationship for archaeal RNaseIII, due

345 to the size of our archaeal dataset (phyloglm failed to converge, only 39 genomes in our
346 dataset had this gene). However, we note that the archaeal RNaseIII PF11469 was found in
347 only two clades that feature exclusively linked rRNA genes (Euryarchaeota family
348 Thermococcaceae and Crenarchaeota family Thermofilaceae).

349

350 Discussion

351 While unlinked rRNA genes have been documented previously, we have
352 demonstrated that they are far more widespread among Bacteria and Archaea than
353 expected. We found that unlinked rRNA genes consistently occur in 12 phyla using a
354 dataset of complete genomes (Figure 2C), and 18 additional phyla using a dataset of long-
355 read metagenomic sequences obtained from environmental samples (Figure 2D).

356 Interestingly, some phyla were classified as exclusively linked in our complete genome
357 dataset, yet had many members with unlinked rRNA genes in our long-read dataset. For
358 example, while there were no complete genomes in the phylum Verrucomicrobia with
359 unlinked rRNA genes (0/32), 38% of verrucomicrobial rRNA sequences were unlinked in
360 our long-read dataset (82/217), with the majority of this group closely related to the
361 bacterium *Ca. Udaeobacter copiosus* from the DA101 soil group (42). This imbalance is
362 likely due to the strong bias towards faster-growing organisms when using traditional
363 cultivation methods (43), and the fact that cultivated Bacteria and Archaea still make up
364 the majority of high-quality genomes in public databases (13). Our results highlight the
365 importance of using a combination of complete genomes, where genetic organization and
366 traits can be assessed rigorously, with metagenomic data that allows us to sample the
367 diversity found in selected environments in an unbiased manner. Together, these
368 independent datasets show that unlinked rRNA genes occur across many bacterial and
369 archaeal phyla.

370 The widespread prevalence of unlinked rRNA genes in many environmental samples
371 has important implications for the use of community analysis methods that require the 16S
372 and 23S rRNA genes to be in close proximity. For instance, before 16S rRNA gene
373 sequencing became common practice, the ITS region of the 16S and 23S rRNA operon was
374 routinely used to fingerprint microbial communities (44). Likewise, the increasing
375 popularity of long-read sequencing technologies has led to bacterial genotyping methods

376 that target the full rRNA operon. While sequencing from the 16S rRNA gene into the 23S
377 rRNA gene (thus including the ITS region of the rRNA operon) can increase taxonomic
378 resolution and allow strain level identification (45), our work shows that amplicon-based
379 studies dependent on 16S and 23S rRNA genes being located in close proximity may miss a
380 large portion of bacterial and archaeal diversity. We found the average distance between
381 unlinked 16S and 23S rRNA genes in our complete genome dataset to be ~410 Kbp, a
382 rather impractical distance to amplify by PCR. While strategies which use reads spanning
383 the 16S and 23S rRNA genes to improve taxonomic resolution (e.g. 45,46) are less likely to
384 introduce biases in some environments (e.g. human gut), they will miss many phylogenetic
385 groups in other environments like soil and sediment, where a significant fraction of taxa
386 have unlinked rRNA genes (Figure 3).

387 We used our long-read metagenomic dataset to not only bypass the cultivation bias
388 of our complete genome dataset, but to also estimate the abundance of unlinked rRNA
389 genes in a range of microbial community types. Our analyses of the long-read metagenomic
390 dataset show that taxa with unlinked rRNA genes are far more abundant in some
391 environments than others. Most notably, unlinked rRNA genes were far more common in
392 soil (where as many as 41% of rRNA genes detected were unlinked) than the human gut
393 (where no unlinked rRNA genes were detected, Figure 3). The environments with higher
394 proportions of unlinked rRNA genes (soil and sediment) are generally thought to be
395 populated by slower growing taxa (35,47). Likewise, we found that genomes with unlinked
396 rRNA genes have significantly fewer rRNA copies than genomes with exclusively linked
397 rRNA genes, a trait which is correlated with maximum potential growth rate (35,48). We
398 also found that genomes with unlinked rRNA genes are predicted to have significantly
399 longer generation times (using codon usage bias in ribosomal proteins as a proxy for
400 maximal growth rates) compared to genomes with exclusively linked rRNA genes. These
401 lines of evidence suggest that unlinked rRNA genes are more common in the genomes of
402 taxa with slower potential growth rates.

403 The existence of numerous genomes that have unlinked 16S and 23S rRNA genes
404 and the differential frequency of these genomes across environments raise the question of
405 the role and implications of this genetic organization. Upon first consideration, having
406 unlinked 16S and 23S rRNA genes would seem to be disadvantageous given that both rRNA

407 molecules are needed in equal proportions to yield a functioning ribosome. The importance
408 of linkage for identical expression of both rRNA genes should be greater in faster growing
409 taxa, where a higher rate of ribosome synthesis is key to rapid growth and accounts for a
410 large proportion of the cell energy budget (4). Studies in the fast-growing species *E.coli*
411 have shown that, while unbalanced rRNA gene dosage has a slight negative effect on
412 doubling times, balanced synthesis of ribosomal proteins still occurs in most cases (49). If
413 unequal expression of rRNA subunits is associated with unlinked rRNA genes, it may not
414 confer a selective disadvantage in many environments (like soils and sediments) where
415 longer generation times are the norm, not the exception. For slower-growing taxa, the
416 selection coefficient associated with the effect of linked rRNA genes on growth may be
417 small, because rRNAs are less expressed and rapid growth is a trait under weaker selection.
418 Under these circumstances, unlinked rRNA genes may become fixed in populations by
419 genetic drift. This is more likely to occur in species with small effective population sizes, i.e.
420 few effectively reproducing individuals, where natural selection is not efficient enough to
421 avoid the loss of genes or the degradation of genome organizational traits that are under
422 weak selection (50). This is the most common explanation for the occurrence of unlinked
423 16S and 23S rRNA genes (10-12). It fits our observations that many of the taxa we
424 identified with unlinked rRNA genes are restricted to obligate intracellular lifestyles
425 (including members of the phyla Spirochaetes, Epsilonbacteraeota, Alphaproteobacteria,
426 and Tenericutes) or contain signatures of symbiotic lifestyles (CPR phyla; 51,52).

427 However, fixation of mutations due to genetic drift is much less likely to explain the
428 presence of unlinked rRNA genes among the large proportion of free-living taxa that we
429 have identified (including members of the phyla Deinococcus-Thermus, Euryarchaeota,
430 Chloroflexi, Planctomycetes, and Verrucomicrobia). Some of these taxa are abundant and
431 ubiquitous in their respective environments, e.g. the Verrucomicrobia *Ca. U. copiosus* (42)
432 and members of the Rikenellaceae RC9 gut group (53). These genomes do not show traits
433 typically associated with genome reduction caused by small effective population sizes, i.e.
434 abundant pseudogenes, transposable elements, or small genomes. While we found that, on
435 average, the genomes of taxa with unlinked rRNA genes were smaller than those with
436 linked rRNA genes, this difference was not significant after accounting for phylogeny. Thus,
437 there is little evidence that the highly conserved trait of unlinked rRNA genes is caused

438 exclusively by genetic drift - especially in free-living taxa.

439 Unlinked rRNA genes could provide a selective advantage in certain circumstances,
440 which may explain their existence in free-living taxa. Transcribing the 16S and 23S rRNA
441 genes separately may eliminate or reduce the need for RNaseIII, which we found to occur in
442 lower frequencies in taxa with unlinked rRNA genes (Supplemental Figure S3). We also
443 found RNaseIII to be completely absent in the phyla Deinococcus-Thermus and
444 Gemmatimonadetes, both phyla with high proportions of unlinked rRNA genes.
445 Interestingly, two recent studies have investigated the function of RNaseIII in *Borrelia*
446 *burgdorferi* (54) and *Helicobacter pylori* (55), two intracellular bacteria with exclusively
447 unlinked rRNA genes. When RNaseIII was knocked out both bacteria remained viable, but
448 accumulated unprocessed rRNA intermediates and exhibited decreased growth rates
449 (54,55). On the other hand, some bacteriophages hijack host RNaseIII to process their own
450 mRNA (56) - in some cases, host RNaseIII can stimulate the translation of infecting phage
451 mRNA by several orders of magnitude (57) (although other phage appear indifferent to the
452 presence of RNaseIII; 58). Regardless, increased resistance to predation at the cost of
453 reduced maximum potential growth rates is a widely observed ecological trade-off (59).
454 Lastly, recent work has shown that some rRNA loci specialize in the translation of genes
455 involved in adaption to temperature and nutrient shifts (60). It is thus tempting to
456 speculate that unlinked rRNA genes could facilitate the production of heterogeneous
457 ribosomes with a diverse range of characteristics.

458

459 Conclusions

460 Unlinked rRNA genes are far more prevalent than expected, especially among those
461 Bacteria and Archaea found in environmental samples for which complete genomes are not
462 yet available. While this rearrangement appears to occur more frequently in slower-
463 growing taxa and may be related to the presence of RNaseIII, it remains to be determined if
464 unlinked rRNA genes confer any specific advantages. Regardless, we have shown that 52%
465 of the phyla included in our combined datasets (37/71) have at least one member with
466 unlinked rRNA genes, that unlinked rRNA genes occur in taxa that are abundant and
467 ubiquitous, and that up to 41% of rRNA genes in some environments are unlinked -
468 meaning unlinked rRNA genes are far from atypical anomalies. Indeed, unlinked rRNA

469 genes function as a reminder that the metabolisms of poorly-studied environmental
470 Bacteria and Archaea sometimes differ from conventions derived from model organisms.
471 We have developed hypotheses about the potential advantages of unlinked rRNA genes,
472 hypotheses which could be tested experimentally and represent a promising direction for
473 future research - especially as some taxa with unlinked rRNA genes are relatively easy to
474 manipulate in culture (61,62).

475

476 Acknowledgements

477 This research was supported in part by the Chateaubriand Fellowship awarded to
478 T.E.B. from the Office for Science & Technology of the Embassy of France in the United
479 States and a grant to N.F. from the U.S. National Science Foundation (EAR1331828). M.A.
480 was supported by a research grant (15510) from Villum Fonden. A.E. gratefully
481 acknowledges the support of a Leverhulme Trust Research Fellowship (RF-2017-652\2).
482 E.R. was supported by the INCEPTION project (PIA/ANR-16-CONV-0005). We thank Will
483 Trimble for assistance tracking down publicly available MolecuLo sequences, Michael Engel
484 for figure design input, and Eric Johnston for introducing the lead author to unlinked rRNA
485 genes.

486

487 Author contributions

488 TEB, ER, and NF conceived and designed the project and wrote the paper with input
489 from all co-authors. AE, MA, and RK performed the Nanopore sequencing. TEB performed
490 all analyses.

491

492 Conflict of interest statement

493 MA and RK own a portion of the company DNASense.

494

495 Data availability

496 All genomes used in this study were downloaded from NCBI, with assembly IDs
497 listed in Supplemental Dataset S1. All Nanopore data is available at the Sequence Read
498 Archive (SRA) under Bioproject ID PRJNA553237 or the European Nucleotide Archive
499 (ENA) under PRJEB33278. All MolecuLo data has been published previously, with

500 publications listed in methods. Classifications and details of both the complete genome and
501 long-read datasets are included in Supplemental Dataset S1 and S2, respectively.

502

503 References

- 504 1. Raoult D, Forterre P. Redefining viruses: lessons from Mimivirus. *Nat Rev Micro.*
505 2008 Mar 4;6:315–9.
- 506 2. Srivastava AK, Schlessinger D. Mechanism and Regulation of Bacterial Ribosomal
507 RNA Processing. *Annu Rev Microbiol.* 1990;44:105–29.
- 508 3. Condon C, Squires C, Squires CL. Control of rRNA Transcription in *Escherichia coli*.
509 *Microbiological Reviews.* 1995 Dec;59:623–45.
- 510 4. Gourse RL, Gaal T, Bartlett MS, Appleman JA, Ross W. rRNA Transcription and
511 Growth Rate–Dependent Regulation of Ribosome Synthesis in *Escherichia coli*. *Annu*
512 *Rev Microbiol.* 1996;50:645–77.
- 513 5. Klappenbach JA, Dunbar JM, Schmidt TM. rRNA Operon Copy Number Reflects
514 Ecological Strategies of Bacteria. *Applied and Environmental Microbiology.* 2000
515 Apr;66:1328–33.
- 516 6. Hartmann RK, Ulbrich N, Erdmann VA. An unusual rRNA operon constellation: in
517 *Thermus thermophilus* HB8 the 23S/5S rRNA operon is a separate entity from the
518 16S rRNA operon. *Biochimie.* 1987;69:1097–104.
- 519 7. Liesack W, Stackebrandt E. Evidence for Unlinked rrn Operons in the Planctomycete
520 *Pirellula marina*. *Journal of Bacteriology.* 1989;171:5025–30.
- 521 8. Munson MA, Baumann L, Baumann P. *Buchnera aphidicola* (a prokaryotic
522 endosymbiont of aphids) contains a putative 16S rRNA operon unlinked to the 23s
523 rRNA-encoding gene: sequence determination, and promoter and terminator
524 analysis. *Gene.* 1993;137:171–8.
- 525 9. Andersson SGE, Zomorodipour A, Winkler HH, Kurland CG. Unusual Organization of
526 the rRNA Genes in *Rickettsia prowazekii*. *Journal of Bacteriology.* 1995;177:4171–5.
- 527 10. Rurangirwa FR, Brayton KA, McGuire TC, Knowles DP, Palmer GH. Conservation of
528 the unique rickettsial rRNA gene arrangement in *Anaplasma*. *International Journal of*
529 *Systemic and Evolutionary Microbiology.* 2002 Jul 1;52(4):1405–9.
- 530 11. Merhej V, Royer-Carenzi M, Pontarotti P, Raoult D. Massive comparative genomic
531 analysis reveals convergent evolution of specialized bacteria. *Biol Direct. BioMed*
532 *Central;* 2009;4(13):13–25.
- 533 12. Andersson JO, Andersson SGE. Genome Degradation is an Ongoing Process in

- 534 *Rickettsia*. Molecular Biology and Evolution. 1999;16(9):1178–91.
- 535 13. Zhi X-Y, Zhao W, Li W-J, Zhao G-P. Prokaryotic systematics in the genomics era.
536 Antonie van Leeuwenhoek. Springer Netherlands; 2012 Nov 25;101(1):21–34.
- 537 14. O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, et al. Reference
538 sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and
539 functional annotation. Nucleic Acids Research. 2016 Jan 3;44(D1):D733–45.
- 540 15. Team RC. R: A language and environment for statistical computing. 2018.
- 541 16. Kuleshov V, Xie D, Chen R, Pushkarev D, Ma Z, Blauwkamp T, et al. Whole-genome
542 haplotyping using long reads and statistical methods. Nat Biotechnol. Nature
543 Publishing Group; 2014 Feb 23;32(3):261–6.
- 544 17. Kuleshov V, Jiang C, Zhou W, Jahanbani F, Batzoglou S, Snyder M. Synthetic long-read
545 sequencing reveals intraspecies diversity in the human microbiome. Nat Biotechnol.
546 Nature Publishing Group; 2016 Jan;34(1):64–9.
- 547 18. White RA, Bottos EM, Roy Chowdhury T, Zucker JD, Brislawn CJ, Nicora CD, et al.
548 Moleculo Long-Read Sequencing Facilitates Assembly and Genomic Binning from
549 Complex Soil Metagenomes. mSystems. American Society for Microbiology Journals;
550 2016 Jun;1(3):309–15.
- 551 19. Sharon I, Kertesz M, Hug LA, Pushkarev D, Blauwkamp TA, Castelle CJ, et al. Accurate,
552 multi-kb reads resolve complex populations and detect rare microorganisms.
553 Genome Res. Cold Spring Harbor Lab; 2015 Apr;25(4):534–43.
- 554 20. Flynn TM, Koval JC, Greenwald SM, Owens SM, Kemner KM, Antonopoulos DA.
555 Parallelized, Aerobic, Single Carbon-Source Enrichments from Different Natural
556 Environments Contain Divergent Microbial Communities. Front Microbiol. Frontiers;
557 2017 Nov 28;8:1540–14.
- 558 21. Bengtsson-Palme J, Hartmann M, Eriksson KM, Pal C, Thorell K, Larsson DGJ, et al.
559 metaxa2: improved identification and taxonomic classification of small and large
560 subunit rRNA in metagenomic data. Mol Ecol Resour. 2015 Mar 23;15(6):1403–14.
- 561 22. Wang Q, Garrity GM, Tiedje JM, Cole JR. Naive Bayesian Classifier for Rapid
562 Assignment of rRNA Sequences into the New Bacterial Taxonomy. Applied and
563 Environmental Microbiology. 2007 Aug 10;73(16):5261–7.
- 564 23. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, et al. The SILVA ribosomal
565 RNA gene database project: improved data processing and web-based tools. Nucleic
566 Acids Research. Oxford University Press; 2012;41(D1):D590–6.
- 567 24. Brown CT, Hug LA, Thomas BC, Sharon I, Castelle CJ, Singh A, et al. Unusual biology
568 across a group comprising more than 15% of domain Bacteria. Nature. 2015 Jun

- 569 15;523(7559):208–11.
- 570 25. Pei A, Nossa CW, Chokshi P, Blaser MJ, Yang L, Rosmarin DM, et al. Diversity of 23S
571 rRNA Genes within Individual Prokaryotic Genomes. PLoS ONE. Public Library of
572 Science; 2009 May 5;4(5):1–9.
- 573 26. Edgar RC. Search and clustering orders of magnitude faster than BLAST.
574 Bioinformatics. 2010 Aug 12;26(19):2460–1.
- 575 27. Caporaso JG, Bittinger K, Bushman FD, DeSantis TZ, Andersen GL, Knight R. PyNAST:
576 a flexible tool for aligning sequences to a template alignment. Bioinformatics. 2010
577 Jan 11;26(2):266–7.
- 578 28. Price MN, Dehal PS, Arkin AP. FastTree: Computing Large Minimum Evolution Trees
579 with Profiles instead of a Distance Matrix. Molecular Biology and Evolution. 2009 Jun
580 9;26(7):1641–50.
- 581 29. Letunic I, Bork P. Interactive tree of life (iTOL) v3: an online tool for the display and
582 annotation of phylogenetic and other trees. Nucleic Acids Research. 2016 Jul
583 4;44(W1):W242–5.
- 584 30. Pagel M. Inferring the historical patterns of biological evolution. Nature. 1999 Oct
585 20;401:877–84.
- 586 31. Revell LJ. phytools: an R package for phylogenetic comparative biology (and other
587 things). Methods Ecol Evol. John Wiley & Sons, Ltd (10.1111); 2011 Dec
588 15;3(2):217–23.
- 589 32. Tung Ho LS, Ané C. A Linear-Time Algorithm for Gaussian and Non-Gaussian Trait
590 Evolution Models. Systematic Biology. 2014 Feb 4;63(3):397–408.
- 591 33. Novembre JA. Accounting for Background Nucleotide Composition When Measuring
592 Codon Usage Bias. Molecular Biology and Evolution. 2002 Jul 18;19(8):1390–4.
- 593 34. Rocha E. Codon usage bias from tRNA's point of view: Redundancy, specialization,
594 and efficient decoding for translation optimization. Genome Res. 2004 Oct
595 17;14:2279–86.
- 596 35. Vieira-Silva S, Rocha E. The Systemic Imprint of Growth and Its Uses in Ecological
597 (Meta)Genomics. PLOS Genetics. 2009 Dec 22;6(1):1–15.
- 598 36. Eddy SR. Accelerated Profile HMM Searches. PLoS Comput Biol. 2011 Oct
599 20;7(10):e1002195–16.
- 600 37. Moreno-Hagelsieb G, Collado-Vides J. A powerful non-homology method for the
601 prediction of operons in prokaryotes. Bioinformatics. 2002;18:S329–36.
- 602 38. Shepherd J, Ibba M. Bacterial transfer RNAs. FEMS Microbiol Rev. 2015 Mar

- 603 20;39(3):280–300.
- 604 39. Parks DH, Chuvochina M, Waite DW, Rinke C, Skarszewski A, Chaumeil P-A, et al. A
605 standardized bacterial taxonomy based on genome phylogeny substantially revises
606 the tree of life. *Nat Biotechnol.* Nature Publishing Group; 2018 Aug 27;36(10):996–
607 1004.
- 608 40. Yuan C, Lei J, Cole J, Sun Y. Reconstructing 16S rRNA genes in metagenomic data.
609 *Bioinformatics.* 2015 Jun 13;31(12):i35–i43.
- 610 41. Durand S, Gilet L, Condon C. The Essential Function of *B. subtilis* RNase III Is to
611 Silence Foreign Toxin Genes. *PLoS Genetics.* 2012 Dec 27;8(12):e1003181–11.
- 612 42. Brewer TE, Handley KM, Carini P, Gilbert JA, Fierer N. Genome reduction in an
613 abundant and ubiquitous soil bacterium “*Candidatus Udaeobacter copiosus*.” *Nature*
614 *Microbiology.* The Author(s) SN ; 2016 Oct 31;2:16198.
- 615 43. Vartoukian SR, Palmer RM, Wade WG. Strategies for culture of “unculturable”
616 bacteria. *FEMS Microbiology Letters.* 2010 Apr 27;309:1–7.
- 617 44. Garcia-Martinez J, Acinas SG, Anton AI, Rodriguez-Valera F. Use of the 16S-23S
618 ribosomal genes spacer region in studies of prokaryotic diversity. *J Microbiol*
619 *Methods.* 1999 Apr 22;36:55–64.
- 620 45. Zeng YH, Koblížek M, Li YX, Liu YP, Feng FY, Ji JD, et al. Long PCR-RFLP of 16S-ITS-
621 23S rRNA genes: a high-resolution molecular tool for bacterial genotyping. *J Appl*
622 *Microbiol.* John Wiley & Sons, Ltd (10.1111); 2012 Dec 20;114(2):433–47.
- 623 46. Cuscó A, Catozzi C, Viñes J, Sanchez A, Francino O. Microbiota profiling with long
624 amplicons using Nanopore sequencing: full-length 16S rRNA gene and whole *rrn*
625 operon. *F1000Res.* 2018 Nov 6;7:1755–25.
- 626 47. Brown CT, Olm MR, Thomas BC, Banfield JF. Measurement of bacterial replication
627 rates in microbial communities. *Nat Biotechnol.* Nature Publishing Group; 2016
628 Dec;34(12):1256–63.
- 629 48. Roller BRK, Stoddard SF, Schmidt TM. Exploiting rRNA operon copy number to
630 investigate bacterial reproductive strategies. *Nature Microbiology.* 2016 Sep 12;1:1–
631 7.
- 632 49. Siehnel RJ, Morgan EA. Unbalanced rRNA Gene Dosage and its Effects on rRNA and
633 Ribosomal-Protein Synthesis. *Journal of Bacteriology.* 1985 Aug;163(2):476–86.
- 634 50. Moran NA. Microbial Minimalism: Genome Reduction in Bacterial Pathogens. *Cell.*
635 2002 Mar 1;108:583–6.
- 636 51. Nelson WC, Stegen JC. The reduced genomes of *Parcubacteria* (OD1) contain

- 637 signatures of a symbiotic lifestyle. *Front Microbiol.* 2015 Jul 21;6(110):693–14.
- 638 52. Burstein D, Sun CL, Brown CT, Sharon I, Anantharaman K, Probst AJ, et al. Major
639 bacterial lineages are essentially devoid of CRISPR-Cas viral defence systems. *Nature*
640 *Communications.* Nature Publishing Group; 2016 Jan 27;7:1–8.
- 641 53. Holman DB, Brunelle BW, Trachsel J, Allen HK. Meta-analysis To Define a Core
642 Microbiota in the Swine Gut. *mSystems.* 11 ed. 2017 May 23;2(3):676–14.
- 643 54. Anacker ML, Drecktrah D, LeCoutre RD, Lybecker M, Samuels DS. RNase III
644 Processing of rRNA in the Lyme Disease Spirochete *Borrelia burgdorferi.* *Journal of*
645 *Bacteriology.* American Society for Microbiology Journals; 2018 Jul 1;200(13):1–11.
- 646 55. Iost I, Chabas S, Darfeuille F. Maturation of atypical ribosomal RNA precursors in
647 *Helicobacter pylori.* *Nucleic Acids Research.* Oxford University Press; 2019 Apr
648 22;47(11):5906–21.
- 649 56. Gone S, Alfonso-Prieto M, Paudyal S, Nicholson AW. Mechanism of Ribonuclease III
650 Catalytic Regulation by Serine Phosphorylation. *Nature.* Nature Publishing Group;
651 2016 Apr 26;6(25448):1–9.
- 652 57. Wilcon HR, Yu D, Peters HK III, Zhou J-G, Court DL. The global regulator RNase III
653 modulates translation repression by the transcription elongation factor N. *The EMBO*
654 *Journal.* 2002;21:4154–61.
- 655 58. Hagen FS, Young ET. Effect of RNase III on Efficiency of Translation of Bacteriophage
656 T7 Lysozyme mRNA. *Journal of Virology.* 1978 Feb 26;26:793–804.
- 657 59. Bohannan BJM, Lenski RE. Linking genetic change to community evolution: insights
658 from studies of bacteria and bacteriophage. *Ecology Letters.* 2000 Jul 7;3:362–77.
- 659 60. Song W, Joo M, Yeom J-H, Shin E, Lee M, Choi H-K, et al. Divergent rRNAs as
660 regulators of gene expression at the ribosome level. *Nature Microbiology.* Springer
661 US; 2019;4:515–26.
- 662 61. Holland AD, Rothfuss HM, Lidstrom ME. Development of a defined medium
663 supporting rapid growth for *Deinococcus radiodurans* and analysis of metabolic
664 capacities. *Appl Microbiol Biotechnol.* 2006 Mar 31;72(5):1074–82.
- 665 62. Devos DP. *Gemmata obscuriglobus.* *Current Biology.* Elsevier; 2013 Sep
666 9;23(17):R705–7.

667

668 Figure Captions

669

670 Figure 1: In most Bacteria and Archaea, rRNA genes are arranged in the order 16S - 23S -
671 5S, and are transcribed and regulated as a single unit. However, in some cases, the 16S is
672 separated from the 23S and 5S, and is referred to as “unlinked”.

673
674 Figure 2: Unlinked rRNA genes can be found in 30 phyla. A) The distribution of ITS lengths
675 in complete genomes from NCBI. 1.3% of NCBI rRNA genes have an ITS region > 1500 bp in
676 length. The majority of unlinked rRNA genes have an ITS of > 6000 bp (682/778) with a
677 mean length of 410374 bp (± 521792 bp). B) The distribution of ITS lengths in the long-
678 read sequence dataset. 10.1% of rRNA genes have an ITS > 1500 bp. The majority of
679 unlinked genes have an ITS of unknown length due to sequence length constraints in the
680 long-read dataset (1470/1607). C) Within our set of complete genomes from NCBI, 12
681 phyla had genomes containing at least one set of unlinked rRNA genes in >5% of members.
682 Linked refers to genomes with exclusively linked rRNA genes, unlinked refers to genomes
683 with exclusively unlinked rRNA genes, and mixed refers to genomes with at least one set
684 each linked and unlinked rRNA genes. D) By analyzing long-read metagenomic datasets, we
685 confirmed that 8 of the phyla with unlinked rRNA genes in the complete genome dataset
686 also had unlinked rRNA genes in environmental samples (top portion), and found an
687 additional 18 phyla in which >5% of reads that met our criteria for inclusion in
688 downstream analyses (see Methods) contained unlinked rRNA genes.

689
690 Figure 3: Unlinked rRNA genes have differential frequencies across environments. We
691 found that soils (13-41% unlinked) and sediments (7.7-29%) have more unlinked rRNA
692 genes on average than anaerobic digesters (8.1-8.8%) and the human gut (0%). Results
693 obtained from analyses of MolecuLo and Nanopore metagenomic data are indicated with
694 (m) and (n), respectively.

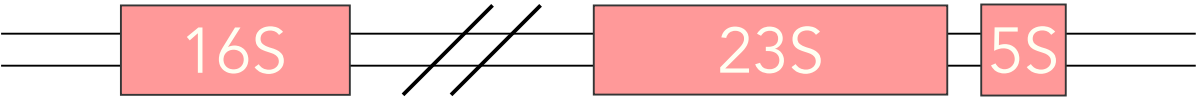
695
696 Figure 4: Unlinked rRNA genes occur in coherent phylogenetic clusters. This phylogenetic
697 tree was created from full-length 16S rRNA sequences by combining both the NCBI
698 complete genome and long-read metagenomic datasets (details in Methods). The outer ring
699 indicates which dataset each sequence originated from, while the inner ring indicates the

700 status of rRNA genes. Sequences originating from the long-read dataset cannot be mixed, as
701 we could not distinguish multi-copy rRNA genes. Clades with high proportions of unlinked
702 members *and* good representation in the tree are indicated in green: A) Euryarchaeota
703 class Thermoplasmata, B) Spirochaetae classes Leptospirae and Spirochaetia, C)
704 Patescibacteria, D) Chlorflexi class Dehalococcoidia, E) Planctomycetes classes
705 Phycisphaerae and Planctomycetacia, F) Verrucomicrobia genus *Candidatus* Udaeobacter,
706 G) Tenericutes genus *Mycoplasma*, H) Deinococcus-Thermus, I) Epsilonbacteraeota genera
707 *Helicobacter* and *Campylobacter*, J) Alphaproteobacteria order Rickettsiales and K)
708 Gammaproteobacteria genus *Buchnera*.

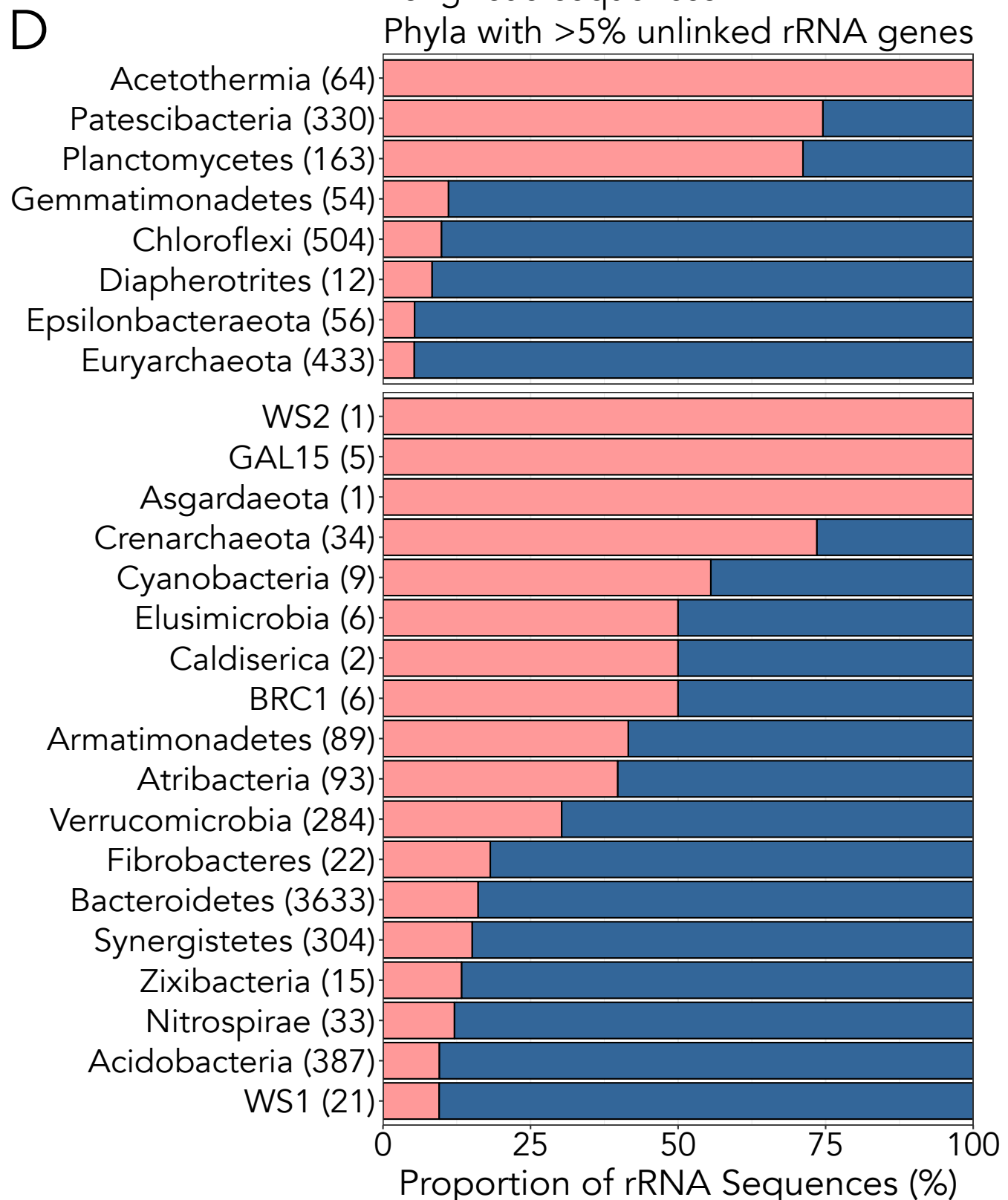
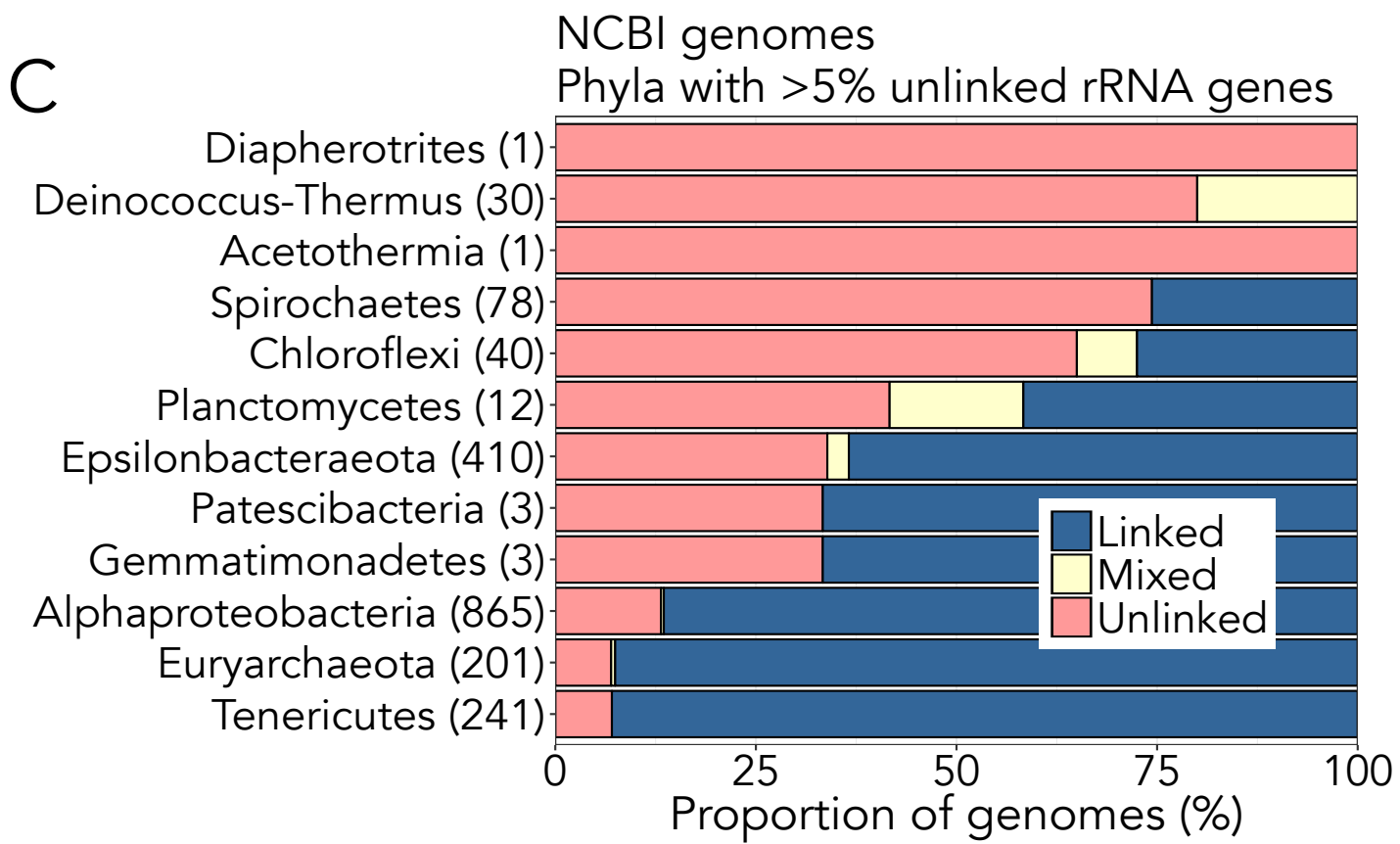
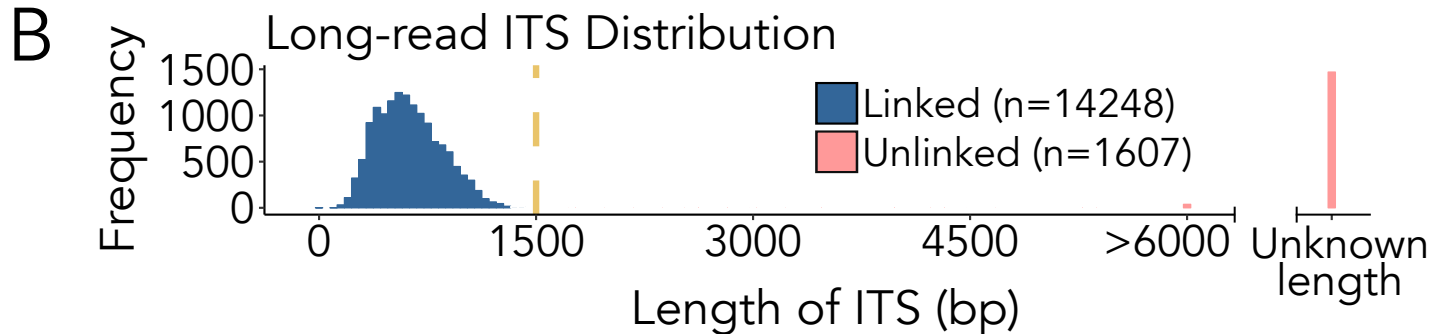
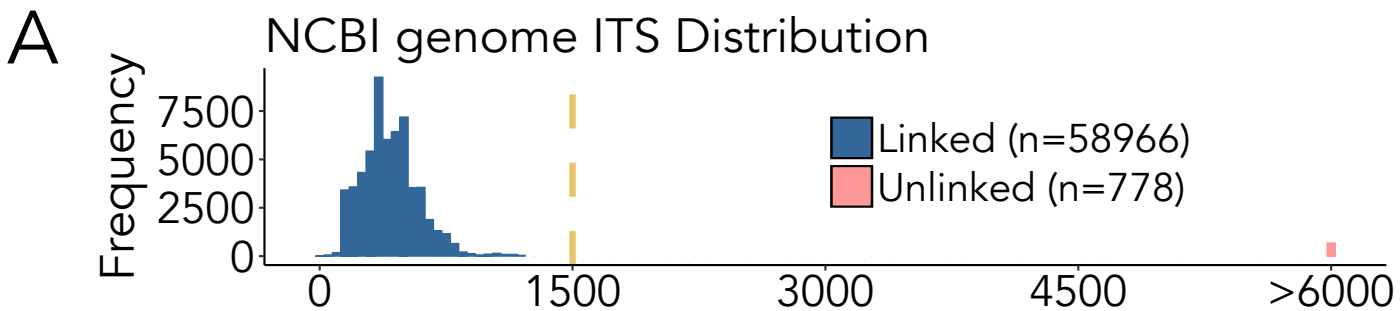
709
710 Figure 5: Genomic attributes of NCBI complete genomes based on their rRNA gene status.
711 Linked genomes feature exclusively linked rRNA genes; unlinked genomes have at least one
712 set of unlinked rRNA genes. We calculated these statistics using a subset of our complete
713 genomes, including one genome per unique species and rRNA gene status. A) Genomes
714 with unlinked rRNA genes have smaller genomes on average, but this difference was not
715 significant after accounting for phylogeny (phyloglm $p = 0.12$, means of groups: 4.15 Mbp
716 linked, 2.72 Mbp unlinked). B) On average, genomes with unlinked rRNA genes had
717 significantly fewer rRNA copies (phyloglm $p < 0.0001$, means of groups: 4.25 copies linked,
718 2.72 copies unlinked). C) Genomes with unlinked rRNA genes are predicted to have longer
719 average generation times (phyloglm $p = 0.028$, means of groups: 0.23 linked, 0.18 unlinked;
720 as a reference *E. coli* has an average $\Delta ENC'$ of 0.3). D) We found that there were
721 significantly fewer RNaseIII genes in genomes with unlinked rRNA genes (only PF00636
722 shown, for more detail see Supplemental Figure S3: phyloglm $p < 0.001$, means of groups:
723 1.0 linked, 0.71 unlinked).



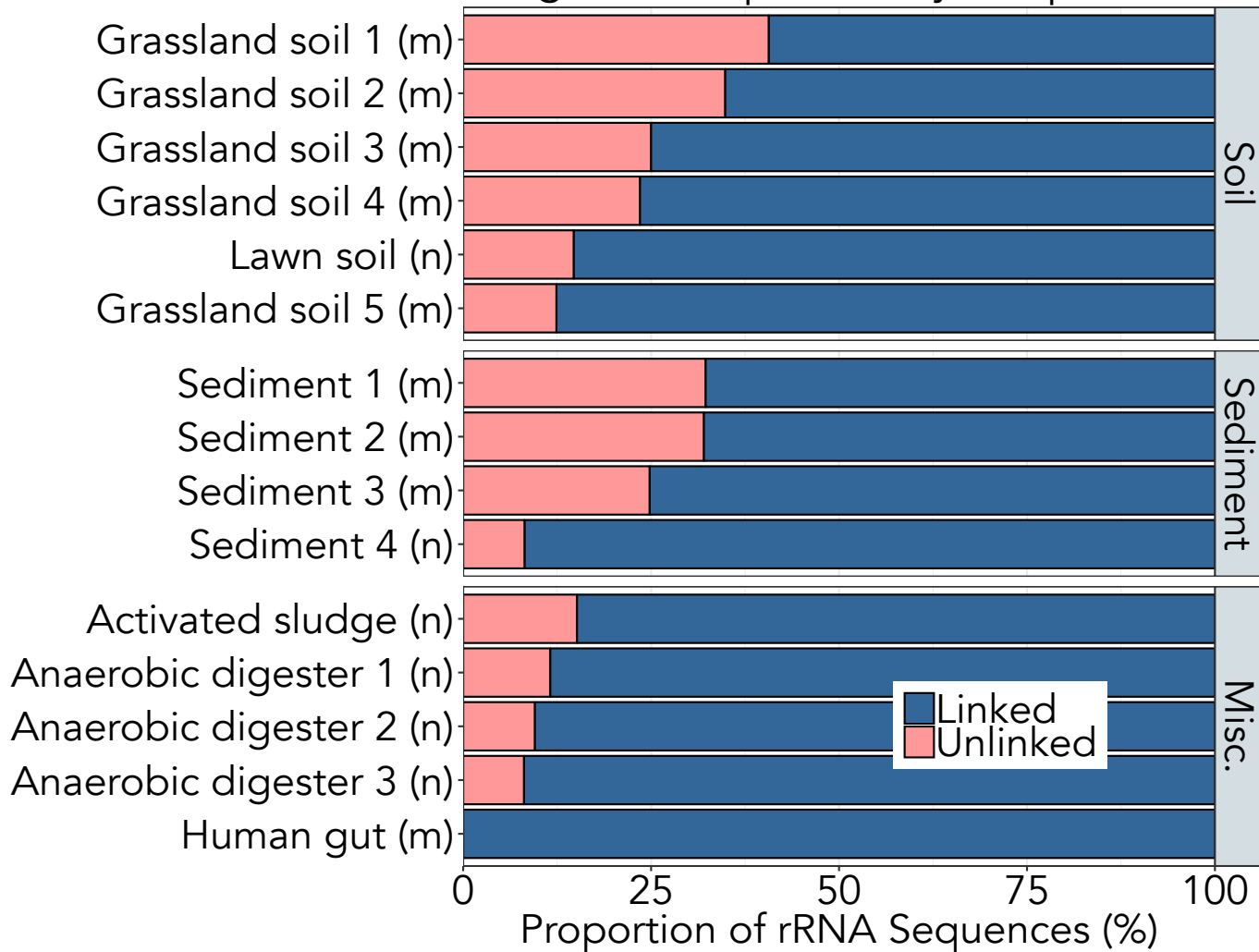
Canonical linked rRNA operon



Unlinked rRNA genes



Long-read sequences by sample

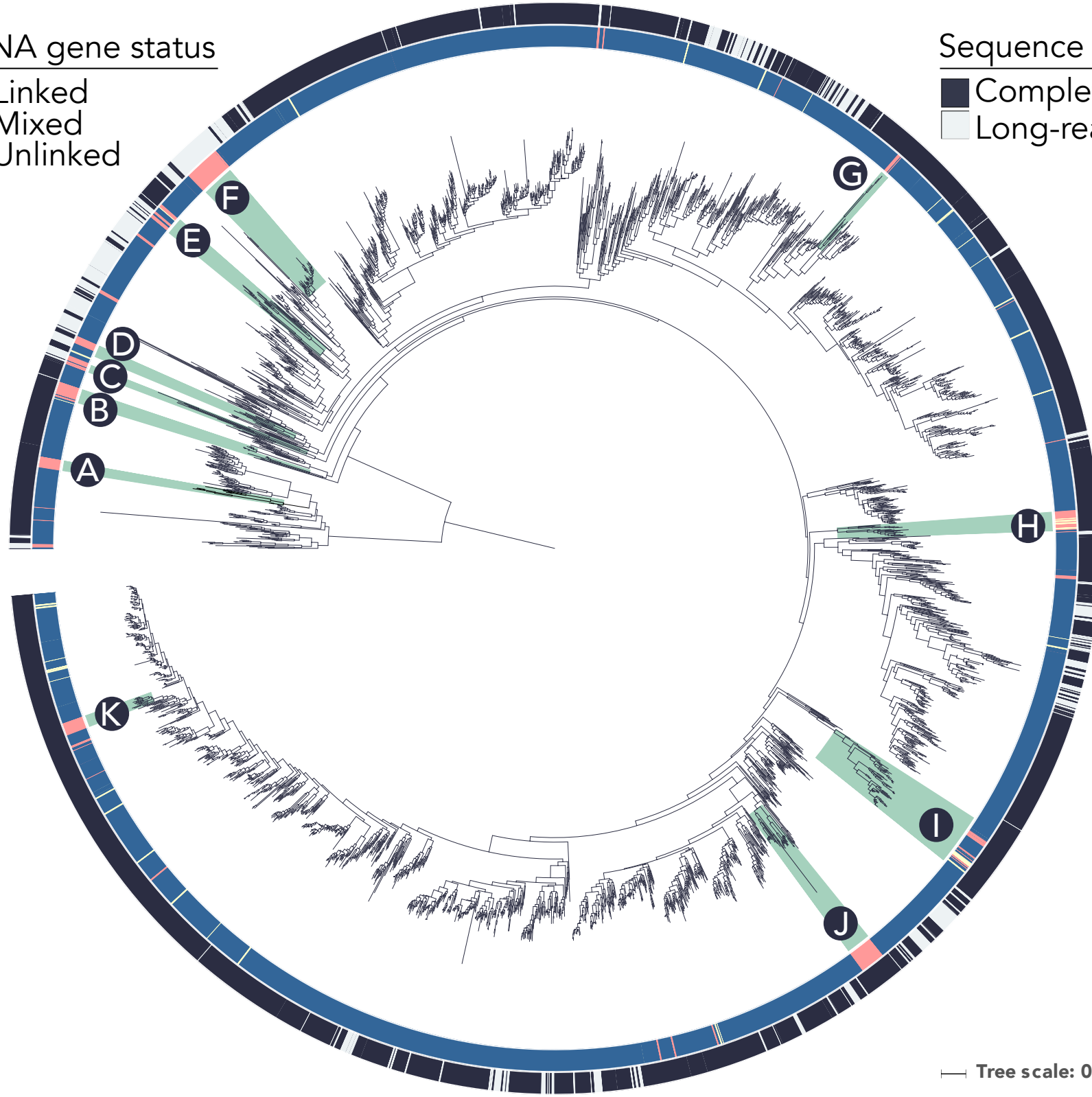


rRNA gene status

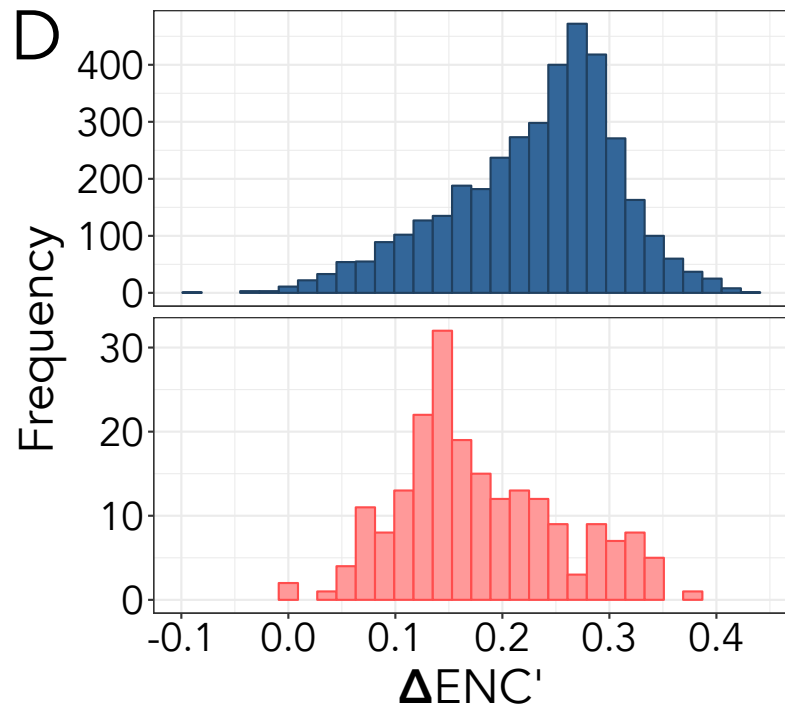
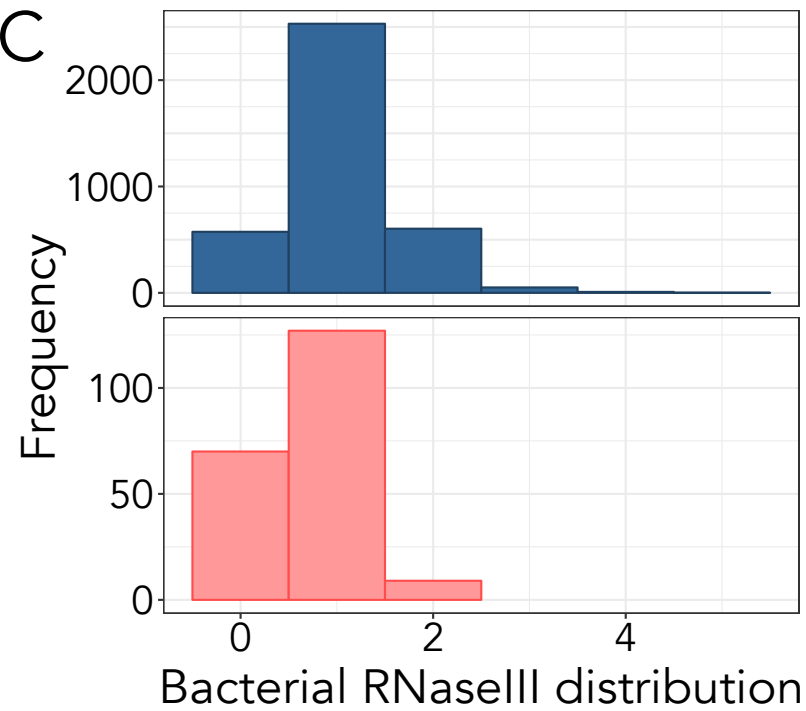
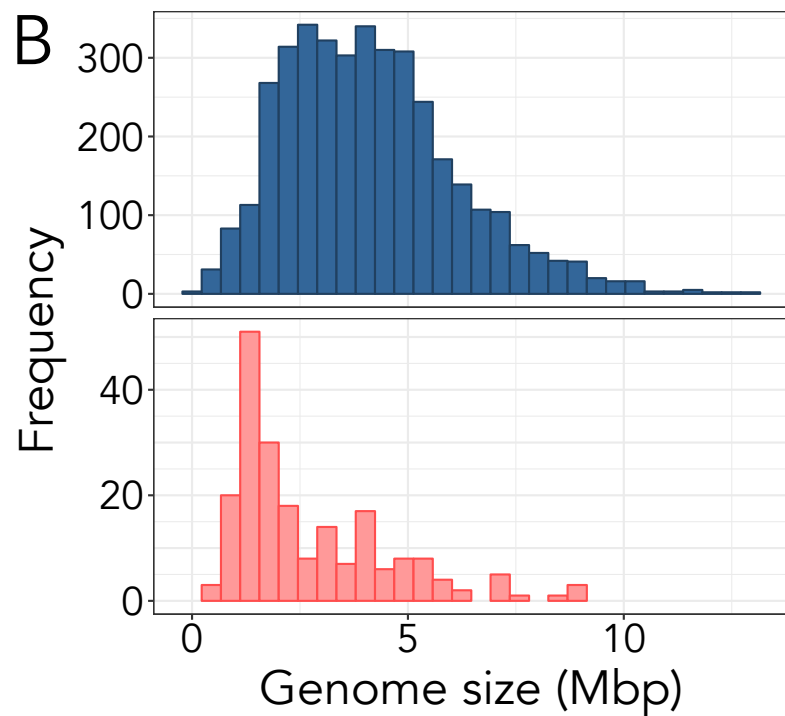
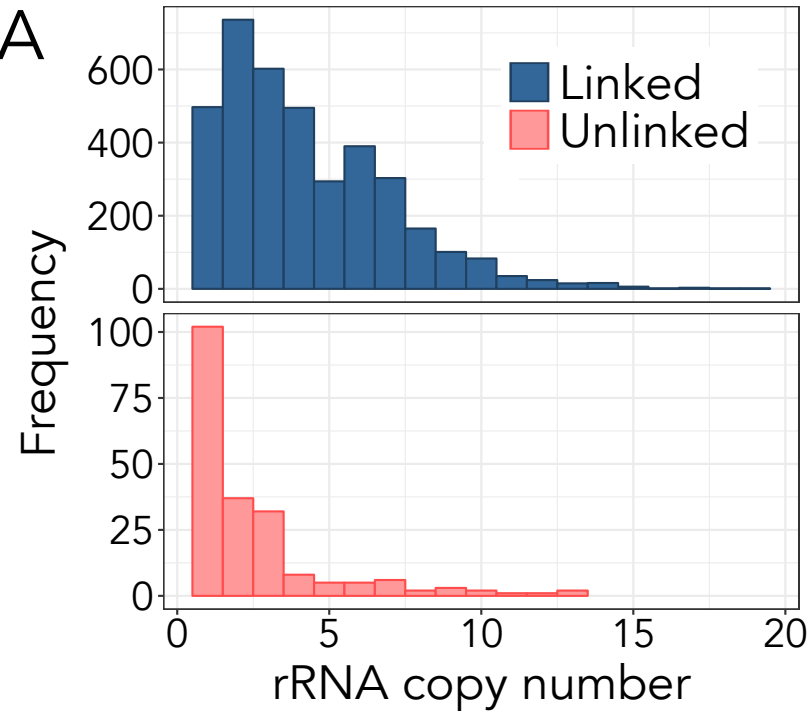
- Linked
- Mixed
- Unlinked

Sequence origin

- Complete genome
- Long-read sequence



Tree scale: 0.1



sample	type	file_type	total_sequences	sequences > 1000	median_length
LIB-RHK-1851	nanopore	fastq	6194277	4953661	3255
20180216_SMK_J3	nanopore	fastq	3747204	3747204	6023
LIB-RHK-1848	nanopore	fasta	3362711	2653517	3228
JMJ	nanopore	fastq	2775301	2114004	4212
MHA-58	nanopore	fastq	1784659	1650522	5375
VCsoil	nanopore	fasta	1751625	1751625	2456
SRR3505613	moleculo	fastq	247328	247328	7197
SRR2822456	moleculo	fastq	130702	130702	7808
KA3UB14	moleculo	fasta	115256	93161	8850
SRR1605785_sedin	moleculo	fastq	95045	95045	7317
SRR1605725_sedin	moleculo	fastq	76499	76499	7863
SRR1605797_sedin	moleculo	fastq	73515	73515	7859
KA3FB3	moleculo	fasta	67177	60415	9774
KA3FB14	moleculo	fasta	50850	40895	4548
KA3UB3	moleculo	fasta	34170	28877	8527

total_lsu_hits	total_ssu_hits	sequences_passing_filters	environment	sample_name_fig3
21463	17761	28056	Misc.	Anaerobic digester 3 (n)
6049	4906	7858	Sediment	Sediment 4 (n)
11955	9842	15672	Misc.	Anaerobic digester 2 (n)
6970	5777	9172	Misc.	Anaerobic digester 1 (n)
4273	3473	5577	Misc.	Activated sludge (n)
2658	1976	3817	Soil	Lawn soil (n)
248	213	328	Soil	Grassland soil 5 (m)
692	534	878	Misc.	Human gut (m)
229	213	367	Soil	Grassland soil 2 (m)
274	253	405	Sediment	Sediment 2 (m)
232	196	325	Sediment	Sediment 1 (m)
258	187	325	Sediment	Sediment 3 (m)
135	129	207	Soil	Grassland soil 1 (m)
88	57	124	Soil	Grassland soil 4 (m)
69	69	112	Soil	Grassland soil 3 (m)