Unlinked rRNA genes are widespread among Bacteria and Archaea

Authors: Tess E. Brewer[a], Mads Albertsen[b], Arwyn Edwards[c], Rasmus H. Kirkegaard[b], Eduardo P. C. Rocha[d], Noah Fierer[a,e]

[a] Cooperative Institute for Research in Environmental Sciences, University of Colorado, Boulder, CO 80309 USA
[b] Department of Chemistry and Bioscience, Aalborg University, 9220 Aalborg, Denmark
[c] Institute of Biological, Environmental and Rural Sciences, Aberystwyth University SY23 3DA UK
[d] Microbial Evolutionary Genomics, Institut Pasteur, CNRS, UMR3525, Paris, 75015, France.
[e] Department of Ecology and Evolutionary Biology, University of Colorado, Boulder, CO 80309 USA

Abstract

Ribosomes are essential to cellular life and the genes for their RNA components are the most conserved and transcribed genes in Bacteria and Archaea. These ribosomal rRNA genes are typically organized into a single operon, an arrangement that is thought to facilitate gene regulation. In reality, some Bacteria and Archaea do not share this canonical rRNA arrangement - their 16S and 23S rRNA genes are not co-located, but are instead separated across the genome and referred to as "unlinked". This rearrangement has previously been treated as a rare exception or a byproduct of genome degradation in obligate intracellular bacteria. Here, we leverage complete genome and long-read metagenomic data to show that unlinked 16S and 23S rRNA genes are much more common than previously thought. Unlinked rRNA genes occur in many phyla, most significantly within Deinococcus-Thermus, Chloroflexi, Planctomycetes, and Euryarchaeota, and occur in differential frequencies across natural environments. We found that up to 41% of the taxa in soil, including dominant taxa, had unlinked rRNA genes, in contrast to the human gut, where all sequenced rRNA genes were linked. The frequency of unlinked rRNA genes may reflect meaningful life history traits, as they tend to be associated with a mix of slow-growing free-living species and obligatory intracellular species. Unlinked rRNA genes are also associated with changes in RNA metabolism, notably the loss of RNaseIII. We propose that unlinked rRNA genes may confer selective advantages in some environments, though the specific nature of these advantages remains undetermined and worthy of further investigation.

Introduction

Ribosomes are the archetypal "essential proteins", so much so that they are a key criteria in the division between cellular and viral life (Raoult and Forterre, 2008). In Bacteria and Archaea, the genes encoding the RNA components of the ribosome are traditionally arranged in a single operon in the order 16S - 23S - 5S. The rRNA operon is transcribed into a single RNA precursor called the pre-rRNA 30S, which is separated and processed by a number of RNases (Srivastava and Schlessinger, 1990). This arrangement of rRNA genes within a single operon is thought to allow rapid responses to changing growth conditions - the production of rRNA under a single promoter allows consistent regulation and conservation of stoichiometry between all three, essential components (Condon et al., 1995). Indeed, the production of rRNA is the rate-limiting step of ribosome synthesis (Gourse et al., 1996), and fast-growing Bacteria and Archaea accelerate ribosome synthesis by encoding multiple rRNA operons (Klappenbach et al., 2000).

Some Bacteria and Archaea have "unlinked" rRNA genes, where the 16S and 23S rRNA genes are separated by large swaths of genomic space (Figure 1). This unlinked rRNA gene arrangement was first discovered in the thermophilic bacterium *Thermus thermophilus* (Hartmann et al., 1987). Reports of unlinked rRNA genes soon followed in additional Bacteria, including the planctomycete *Pirellula marina* (Liesack and Stackebrandt, 1989), the aphid endosymbiont *Buchnera aphidicola* (Munson et al., 1993), and the intracellular pathogen *Rickettsia prowazekii* (Andersson et al., 1995). Though unlinked rRNA genes were first discovered in a free-living environmental bacteria, their ubiquity among the order Rickettsiales has led to suggestions that unlinked rRNA genes are a result of the genome degradation typical of obligate intracellular lifestyles (Rurangirwa et al., 2002; Merhej et al., 2009; Andersson and Andersson, 1999).

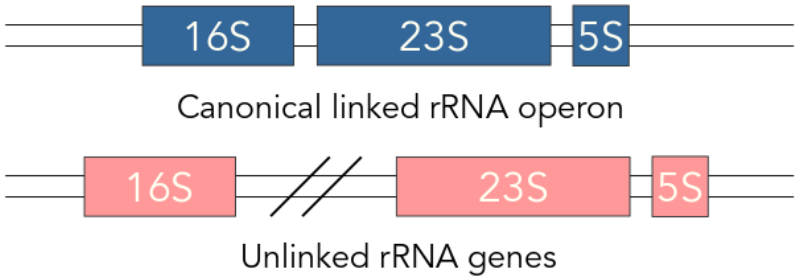


Figure 1: In most Bacteria and Archaea, the rRNA genes are arranged in the order 16S - 23S - 5S, and are transcribed and regulated as a single unit. However, in some cases, the 16S is separated from the 23S and 5S, and is referred to as "unlinked".

With this study, we sought to determine the frequency of unlinked rRNA genes across Bacteria and Archaea and whether this unique genomic feature is largely confined to those Bacteria and Archaea with an obligate intracellular lifestyle. We examined the rRNA genes of over 10,000 publicly available complete bacterial and archaeal genomes to identify which taxa have unlinked rRNA genes and to determine if there are any genomic characteristics shared across taxa with this feature. As complete genomes are not typically available for the broader diversity of Bacteria and Archaea found in environmental samples (Zhi et al., 2012), we also characterized rRNA gene arrangements using long-read metagenomic datasets obtained from a range of environmental samples, which together encompassed over 17 million sequences (≥1000bp). With these long-read metagenomic datasets, we were able to determine whether unlinked rRNA genes are common in environmental populations and how the distributions of unlinked rRNA genes differ across prokaryotic lineages and across distinct microbial habitats.

Methods

Analyses of complete genomes

We downloaded all bacterial and archaeal genomes in the RefSeq genome database (O'Leary et al., 2016) classified with the assembly level "Complete Genome" from NCBI in January 2019 (12539 genomes). We removed genomes from consideration that had non-numeric gene ranges (96 genomes), >20 reported rRNA genes (2 genomes), or an unequal number of 16S and 23S rRNA genes (219 genomes). This left us with a set of 12222 genomes. We used gene ranges associated with each open reading frame (ORF) to pair the 16S and 23S rRNA genes that were closest to each other in each genome. We then checked for gene directionality (sense/antisense) and calculated the distance between each pair, taking directionality into account (see Supplemental Figure S1 for more detail and a visual representation). rRNA pairs were classified as 'unlinked' if the distance between each gene was greater than 1500bp, 'linked' if the distance was less than or equal to 1500bp. We separated genomes that had a 16S or 23S rRNA gene that started or ended within 1500bp of the beginning or end of its genome and classified these 226 genomes independently to account for the circular nature of bacterial and archaeal genomes. For this subset of genomes, we iteratively adjusted the start and end position of those "edge-case" rRNA genes with respect to genome size and selected the smallest distance between the 16S and 23S rRNA genes as the true distance, using the same formula presented in

Supplemental Figure 1. Each genome was classified as 'unlinked', 'linked', or 'mixed' depending
94 on the status of their rRNA genes with 'mixed' genomes having multiple rRNA copies with a combination of linked and unlinked rRNA genes. We re-assigned taxonomy to each genome
96 using the SILVA 132 SSU database (clustered at 99%) to maintain a consistent taxonomy between our two datasets. All analyses were done in R version 3.5.1 (R Core Team, 2018).
98 Information on all genomes included in these analyses (including classification of rRNA genes) is available in Supplemental Dataset S1.
100

Long-read metagenomic analyses

102 To investigate the prevalence of unlinked rRNA genes among those Bacteria and Archaea found in environmental samples (including many taxa for which genomes are not yet available),
104 we analyzed long-read metagenomic datasets generated from soil, sediment, activated sludge, anaerobic digesters, and human gut samples. These metagenomic datasets were generated using
106 either the Oxford Nanopore MinION/PromethION (6 samples) or the Illumina synthetic long-read sequencing technology (also known as Moleculo, first described in (Kuleshov et al., 2014), 9
108 samples). The Moleculo sequences originated from four previously published studies covering: the human gut (Kuleshov et al., 2016), prairie soil (White et al., 2016), sediment (Sharon et al.,
110 2015), and grassland soils (MG-RAST project mgp14596, Flynn et al., 2017). The Nanopore sequences originated from four unpublished studies that studied anaerobic digesters, activated
112 sludge, sediment, and lawn soil. For these samples, DNA was extracted using DNeasy PowerSoil Kits (Qiagen, DE) and libraries were prepared for sequencing using the LSK108 kit (Oxford
114 Nanopore Technologies, UK) following the manufacturers protocol. The libraries were sequenced on either the MinION or the PromethION sequencing platforms (Oxford Nanopore
116 Technologies, UK). Base calling was conducted using Albacore v. 2.1.10 for the lawn soil sample (VCsoil) and Albacore v. 2.3.1 for all other samples (Oxford Nanopore Technologies, UK).
118 Across these 15 samples, we compiled 16,870,533 Nanopore sequences and 846,437 Moleculo sequences with a minimum read length of 1000bp.
120 We trimmed the first 250bp of each Nanopore sequence to remove low quality regions, but performed no other quality filtering as not all samples included information on sequence quality
122 (some sequences were fasta format). Instead, we relied on our downstream filtering steps to remove sequences of poor quality. Metaxa2 version 2.1 (Bengtsson-Palme et al., 2015) was run

124 on all sequences with default settings to search for SSU (16S rRNA) and LSU (23S rRNA) gene
fragments. Taxonomy was assigned to the partial rRNA sequences using the RDP classifier
126 (Wang et al., 2007) and the SILVA 132 SSU and LSU databases (both clustered at 99%
sequence identity, Quast et al., 2012). If a sequence contained both 16S and 23S rRNA genes we
128 used the taxonomy with the highest resolution (if the 16S was annotated to family level while the
23S was genus level, we used the 23S taxonomy for both rRNA). Details on each sample,
130 including number of reads and median read lengths, are available in Supplemental Table S1.

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

134 1) Included at least 2 domains of the 16S or 23S rRNA genes (Metaxa2 uses multiple
HMM profiles targeting conserved regions of the 16S and 23S rRNA genes, each of
136 these regions is referred to as a domain),
2) Included either the last two domains of the 16S rRNA gene (V8|V9) or the first two
138 domains of the 23S rRNA gene (C01|C02),
3) Were ≤4000bp if a 16S rRNA gene and ≤ 6800bp if a 23S rRNA gene (these limits were
140 chosen to accommodate insertions within rRNA genes such as those that occur in
Candidate Phyla Radiation (CPR) taxa (Brown et al., 2015), *Nostoc*, *Salmonella*, and others
142 (Pei et al., 2009)),
4) Could be classified to at least the phylum level of taxonomic resolution.

144

Of the subset of reads that met these criteria (112 - 878 per Moleculo sample, 3817 - 28056 per
146 Nanopore sample, see Supplemental Table S1 for details), we classified reads as containing
unlinked rRNA genes if there was >1500bp between the 16S and 23S rRNA genes, or if there
148 was no 23S domain found 1500bp after the end of the 16S rRNA. We note that, unlike the
NCBI gene ranges, Metaxa2 takes strand information into account and translates start and stop
150 locations into sense orientation for SSU and LSU. For our final analyses, we removed reads that
could not be classified as linked or unlinked rRNA genes (for instance a sequence with only
152 300bp after the 3' end of the 16S rRNA gene) and included only reads that contained a 16S
rRNA gene to avoid potentially double counting organisms with unlinked 16S and 23S rRNA
154 genes. All analyses were done in R version 3.5.1 (R Core Team, 2018). Information on all long-

read sequences included in these analyses (including classification of rRNA genes) is available in
156 Supplemental Dataset S2.

158 **Phylogenetic tree combining long-read and NCBI datasets**

A phylogenetic tree was created from full-length 16S rRNA sequences by combining both
160 the NCBI complete genomes and representatives of the long-read metagenomic datasets. For the NCBI genome sequences, we selected one 16S rRNA gene sequence per unique species. For the
162 long-read datasets, we first matched the partial 16S rRNA genes recovered by metaxa2 (Bengtsson-Palme et al., 2015) to full-length 16S rRNA gene sequences in the SILVA 132 SSU
164 database (Quast et al., 2012) using the usearch10 version 10.0.240 command usearch_global (settings: -id 0.95 -strand both -maxaccepts 0 -maxrejects 0; Edgar, 2010). The full-length SILVA
166 16S rRNA genes sequences that matched to the long-read sequences ≥95% percent identity and ≥500bp alignment length were added to the complete genome sequences as representatives of
168 their long-read sequence match. We used 95% percent identity as our cutoff as we found unlinked rRNA gene status to generally be conserved within genera (see below and Supplemental
170 Figure S2). The NCBI and SILVA sequences were then aligned with PyNAST version 0.1 (Caporaso et al., 2010) and the phylogenetic tree was constructed using FastTree version 2.1.10
172 SSE3 (Price et al., 2009), and plotted with iTOL (Letunic and Bork, 2016).

174 **Genomic attributes associated with unlinked rRNA genes**

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

182 With this set of reduced genomes (3967 genomes in total), we first calculated Pagel's lambda (Pagel, 1999) to determine whether there was a phylogenetic signal associated with
184 unlinked rRNA genes using the phylosig function of the phytools package version 0.6.60 (Revell, 2011). The results of this test indicated there was a strong phylogenetic signal (lambda = 0.96, p

186 < 0.0001), so we controlled for phylogeny in all of our subsequent tests by using a Phylogenetic Generalized Linear Model for continuous variables (with the function phyloglm in the phylolm
188 package version 2.6, Tung Ho and Ané, 2014).

To determine if taxa with unlinked rRNA genes have a lower predicted growth rate, we
190 calculated the codon usage proxy ΔENC' (Novembre, 2002; Rocha, 2004), which provides an estimate of minimum generation times (Vieira-Silva and Rocha, 2009). We calculated ΔENC'
192 with the program ENCprime (Novembre, 2002) with default options, on both the concatenated ORF sequences and concatenated ribosomal protein sequences for each genome following
194 Vieira-Silva and Rocha (2009). To determine if RNaseIII was present in each genome, we used HMMER version 3.1b2 (Eddy, 2011) to search for three RNaseIII pfams (bacterial PF00636,
196 PF14622, and archaeal PF11469) in the translated protein files of each genome. We used the GA gathering cutoffs profile associated with each of these pfams to set all thresholding (--cut_ga).
198

200 **Results**

Unlinked rRNA genes occur frequently in complete genomes

202 We used a set of 12222 "complete" bacterial and archaeal genomes extracted from NCBI in January 2019 to determine how frequently unlinked 16S and 23S rRNA genes occur. We
204 analyzed the distribution of distances between the closest edges of the closest pairs of 16S and 23S rRNA genes (known as the Internally Transcribed Spacer - ITS) in each genome and found
206 that the vast majority of 16S and 23S rRNA gene pairs (98.7%) had an ITS ≤ 1500bp with an average ITS length of 418.7bp (±169.7bp, Figure 2A). However, pairs with ITS lengths >
208 1500bp showed a scattered distribution of distances, with an average ITS length of 410374bp (±521792bp). Hence, for this classification scheme we called rRNA genes "unlinked" if the ITS
210 was greater than 1500bp. This 1500bp cutoff is in some ways conservative, as the distance between genes in an operon is usually quite low (peaking between -20 and 30bp in most
212 genomes; Moreno-Hagelsieb and Collado-Vides, 2002) and the genes most frequently located between the 16S and 23S rRNA genes encode tRNA, which range from 75 to 90bp in length
214 (Shepherd and Ibba, 2015).

After classifying each rRNA gene pair as linked or unlinked based on the distance
216 between the 16S and 23S rRNA genes, we found that 3.65% of the genomes in our dataset had

exclusively unlinked rRNA genes, 0.62% had mixed rRNA gene status (i.e. genomes with
218 multiple rRNA copies that had at least one set of unlinked rRNA genes and at least one canonical, linked rRNA operon), and 95.73% had exclusively linked operons (these numbers do
220 not match up with the per rRNA gene dataset as each genome has a variable rRNA copy number). We found unlinked genomes to be relatively common (present in ≥5% of members) in
222 taxa characterized as having an obligate intracellular lifestyle within the phyla Spirochaetes (genus *Borrelia*), Epsilonbacteraeota (family Helicobacteraceae), Alphaproteobacteria (order
224 Rickettsiales), and Tenericutes (species *Mycoplasma gallisepticum*). However, we also found high proportions of unlinked rRNA genes in phyla that are generally considered to be free-living, such
226 as Deinococcus-Thermus (families Thermaceae and Deinococcaceae), Chloroflexi (family Dehalococcoidaceae), Planctomycetes (families Phycisphaeraceae and Planctomycetaceae), and
228 Euryarchaeota (class Thermoplasmata). Phyla with at least 5% of genomes having exclusively unlinked rRNA genes are shown in Figure 2C.

230

Unlinked rRNA genes are widespread in environmental metagenomic data

232 While the results from our complete genome dataset demonstrate that unlinked rRNA genes are common in some putatively free-living phyla, databases featuring complete genomes do
234 not capture the full breadth of microbial diversity and are heavily biased towards cultivated organisms relevant to human health (Zhi et al., 2012). Just three phyla (Proteobacteria,
236 Firmicutes, Actinobacteria) accounted for >83% of the genomes in our NCBI dataset - even though recent estimates of bacterial diversity total at least 99 unique phyla (Parks et al., 2018). To
238 investigate the ubiquity of unlinked rRNA genes among those taxa underrepresented in 'complete' genome databases, we analyzed long-read metagenomic data from a range of distinct
240 sample types. Focusing exclusively on long-read sequences allowed us to span the 1500bp distance required for classification of rRNA genes without the need for assembly. This is
242 important as the repetitive structure of rRNA genes makes it difficult to assemble a mix of non-identical rRNA genes from the short reads typical of most current metagenomic sequencing
244 projects (Yuan et al., 2015).

246

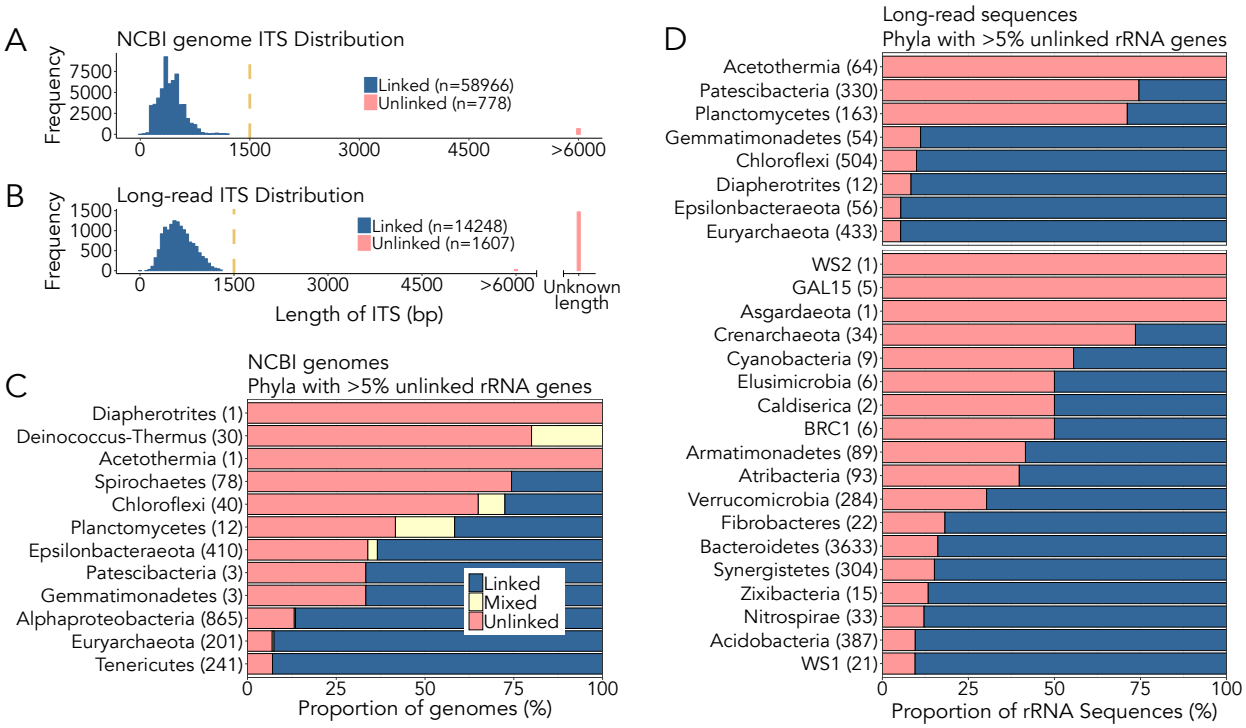


Figure 2: Unlinked rRNA genes can be found in 30 phyla. **A**) Distribution of ITS lengths in complete genomes from NCBI. 98.7% of NCBI rRNA genes have an ITS region ≤ 1500bp in length. The majority of unlinked rRNA genes have an ITS of > 6000bp (682/778) with a mean length of 410374bp (±521792bp). **B**) Distribution of ITS lengths in the long-read sequence dataset. 10.1% of rRNA genes have an ITS > 1500bp. The majority of the unlinked genes have an ITS of unknown length due to sequence length constraints in the long-read dataset (1470/1607). **C**) Within our set of complete genomes from NCBI, 12 phyla had genomes containing at least one set of unlinked rRNA genes in >5% of members. Linked refers to genomes with exclusively linked rRNA genes, unlinked refers to genomes with exclusively unlinked rRNA genes, and mixed refers to genomes with at least one set each linked and unlinked rRNA gene. **D**) By analyzing long-read metagenomic datasets, we confirmed that 8 of the phyla with complete genomes also had unlinked rRNA genes in our environmental or host-associated samples (top portion), and added an additional 18 phyla in which >5% of reads that met our criteria for inclusion in downstream analyses (see Methods) encoded unlinked rRNA genes.

From our initial long-read dataset encompassing 15 unique samples (~890,000 Illumina synthetic long reads (also known as Moleculo) and ~19 million Nanopore reads, with median read lengths of 8858bp and 5398bp, respectively), only 15855 sequences contained rRNA genes and met the criteria we established for the classification of rRNA genes as linked or unlinked (see Methods). Of these reads, we classified 1607 as unlinked, or 10.1% of the dataset (Figure 2B). These long-read metagenomic analyses showed that unlinked rRNA genes are not equally

distributed across environments - we found that up to 41% of the taxa in soil had unlinked rRNA
272 genes, whereas other environments had much lower proportions, most notably the human gut, where all sequenced rRNA genes were linked (Figure 3).

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

284 Using the long-read dataset, we identified 18 additional phyla where unlinked rRNA genes are common, including several candidate phyla (BRC1, GAL15, WS1, WS2) and members
286 of the Candidate Phyla Radiation (Patescibacteria, Figure 2). We also found several clades with high proportions of unlinked rRNA genes that had no representation in our complete genome
288 dataset, including Rikenellaceae RC9 gut group (334/624), Verrucomicrobia genus *Candidatus Udaeobacter* (80/80), Atribacteria order Caldatribacteriales (37/37), Cyanobacteria order
290 Obscuribacterales (4/4), Acidobacteria Subgroup 2 (27/27), Planctomycetes order MSBL9 (40/40), and Chloroflexi class GIF9 (7/7). Overall, we found that 52% of the phyla covered in
292 our combined datasets (37/71) have at least one representative with unlinked rRNA genes.

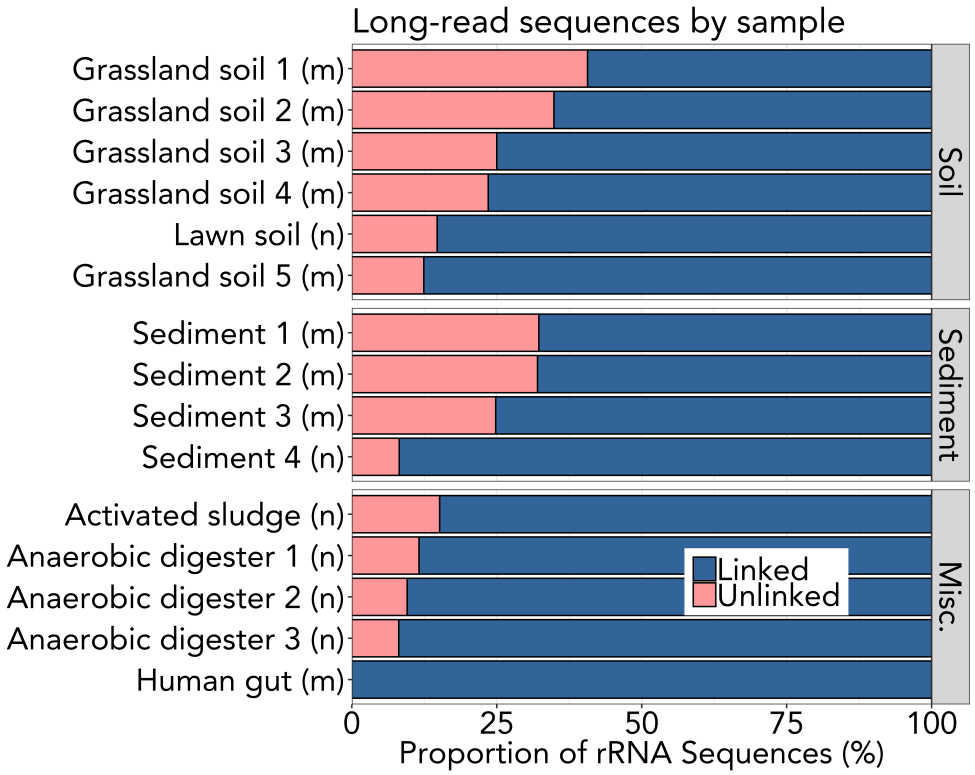


Figure 3: Unlinked rRNA genes have differential frequencies across environments. We found that soils (13-41% unlinked) and sediments (7.7-29%) have more unlinked rRNA genes on average than anaerobic digesters (8.1-8.8%) and the human gut (0%). Results obtained from analyses of Moleculo and Nanopore metagenomic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

Unlinked rRNA genes are strongly conserved

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

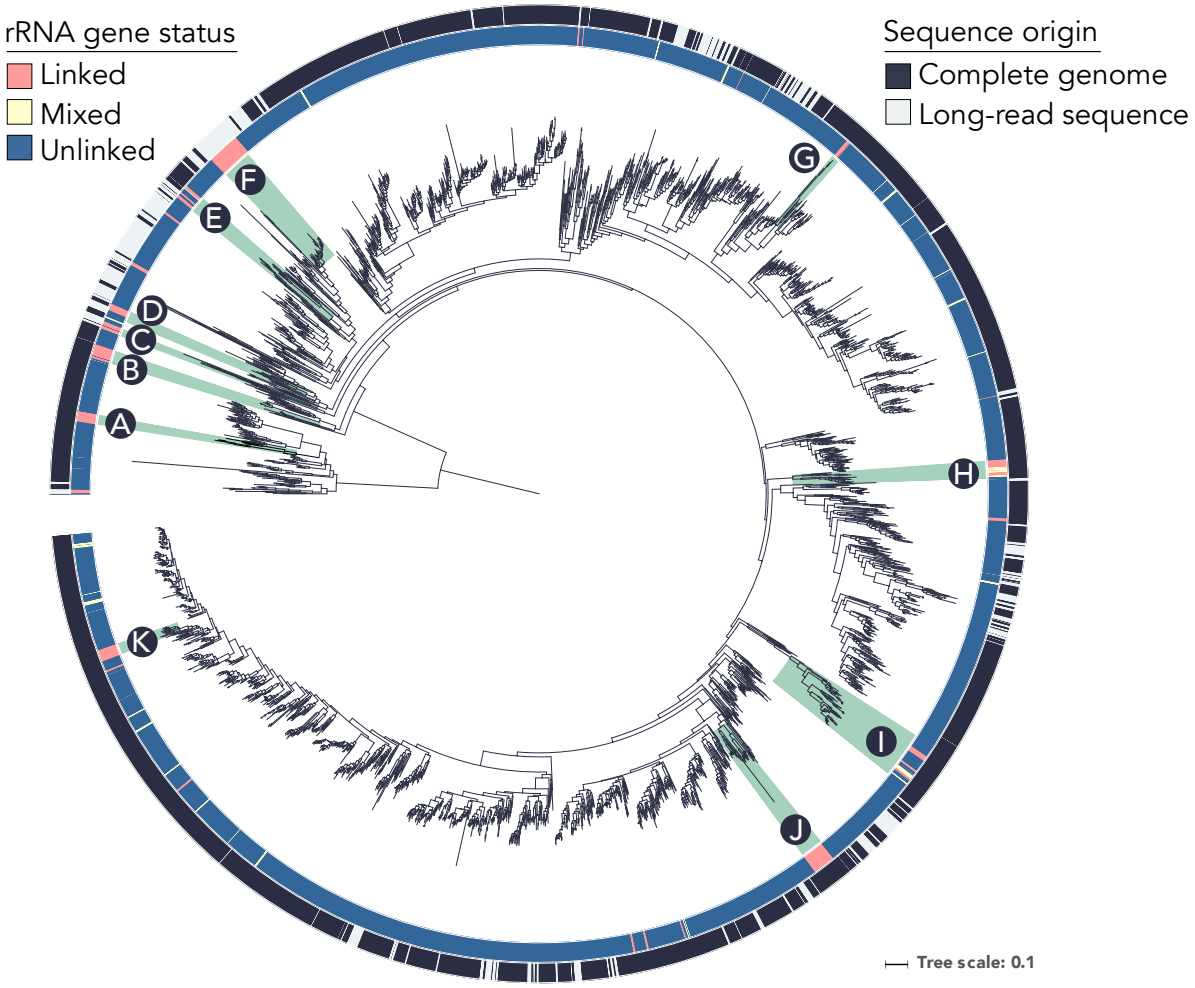


Figure 4: Unlinked rRNA genes occur in coherent phylogenetic clusters. This phylogenetic tree was created from full-length 16S rRNA sequences by combining both the NCBI complete genome and long-read metagenomic datasets (details in Methods). The outer ring indicates which dataset each sequence originated from (complete genomes from NCBI versus long-read sequences from metagenomes), while the inner ring indicates the status of rRNA genes as either linked, mixed, or unlinked. Sequence representatives of the long-read dataset cannot be mixed, as we could not distinguish multi-copy rRNA genes. Clades with high proportions of unlinked members *and* good representation in the tree are indicated in green: A) Euryarchaeota class Thermoplasmata, B) Spirochaetae classes Leptospirae and Spirochaetia, C) Patescibacteria, D) Chlorflexi class Dehalococcoidia, E) Planctomycetes classes Phycisphaerae and Planctomycetacia, F) Verrucomicrobia genus *Candidatus* Udaeobacter, G) Tenericutes genus *Mycoplasma*, H) Deinococcus-Thermus, I) Epsilonbacteraeota genera *Helicobacter* and *Campylobacter*, J) Alphaproteobacteria order Rickettsiales and K) Gammaproteobacteria genus *Buchnera*.

332 **Genomic attributes associated with unlinked rRNA genes**

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

Historically, unlinked rRNA genes have been strongly associated with the reduced
346 genomes of obligate intracellular bacteria, implying that this trait may merely be a side effect of the strong genetic drift and weak selection these taxa experience. To test this hypothesis, we
348 compared the genome sizes of species with linked and unlinked rRNA genes using Phylogenetic Generalized Linear Models (phyloglm). While we found that genomes with unlinked rRNA genes
350 had smaller genomes on average, this difference was not significant (Figure 5, phyloglm $p = 0.12$, means of groups: 4.15Mbp linked, 2.72Mbp unlinked).

352 The organization of rRNA genes within the same operon facilitates their joint regulation and co-expression at precise stoichiometric ratios. Selection for this trait is expected to be
354 stronger in faster growing Bacteria and Archaea, where, at maximum growth rates, synthesis of the ribosome is the cell's chief energy expenditure (Gourse et al., 1996). To test this hypothesis,
356 we analyzed the association between the linkage of rRNA genes and traits related to rapid growth in Bacteria and Archaea. On average, genomes with unlinked rRNA genes had
358 significantly fewer rRNA copies (Figure 5, phyloglm $p < 0.0001$, means of groups: 4.25 copies linked, 2.72 copies unlinked). We also calculated ΔENC' for each complete genome - a measure
360 of codon usage bias that is negatively correlated with minimum generation time in Bacteria and Archaea (Vieira-Silva and Rocha, 2009). Interestingly, genomes with unlinked rRNA genes were
362 predicted to have significantly longer minimal generation times (Figure 5, phyloglm $p=0.028$,

means of groups: 0.23 linked, 0.18 unlinked). Additionally, in our long-read dataset we found
364 that unlinked rRNA genes were more common in environments typified by slow growth rates;
soil and sediment samples had higher proportions of unlinked rRNA genes than samples from
366 anaerobic digesters and the human gut (Figure 3).

RNaseIII separates the precursors of the 16S and 23S rRNA from their common
368 transcript for subsequent maturation and inclusion in the ribosome (Srivastava and Schlessinger,
1990). RNaseIII is not an essential protein in most Bacteria and Archaea and several phyla in
370 which unlinked rRNA genes are common have been reported to not encode RNaseIII (e.g.
Deinococcus-Thermus and Euryarchaeota; Durand et al., 2012). Therefore, we checked if there
372 was a significant association between unlinked rRNA genes and the presence of RNaseIII genes.
Interestingly, we found that genomes with unlinked rRNA genes were significantly less likely to
374 encode the bacterial form of RNaseIII genes (Figure 5 and Supplemental Figure S3, PF00636:
phyloglm $p < 0.001$, means of groups: 1.0 linked, 0.71 unlinked; PF14622: phyloglm $p = 0.007$,
376 means of groups: 0.86 linked, 0.66 unlinked). We were unable to check this relationship for
archaeal RNaseIII, due to the size of our archaeal dataset (phyloglm failed to converge, only 39
378 genomes in our dataset had this gene). However, we note that the archaeal RNaseIII PF11469
was found in only two clades that feature exclusively linked rRNA genes (Euryarcheaota family
380 Thermococcaceae and Crenarchaeota family Thermofilaceae).

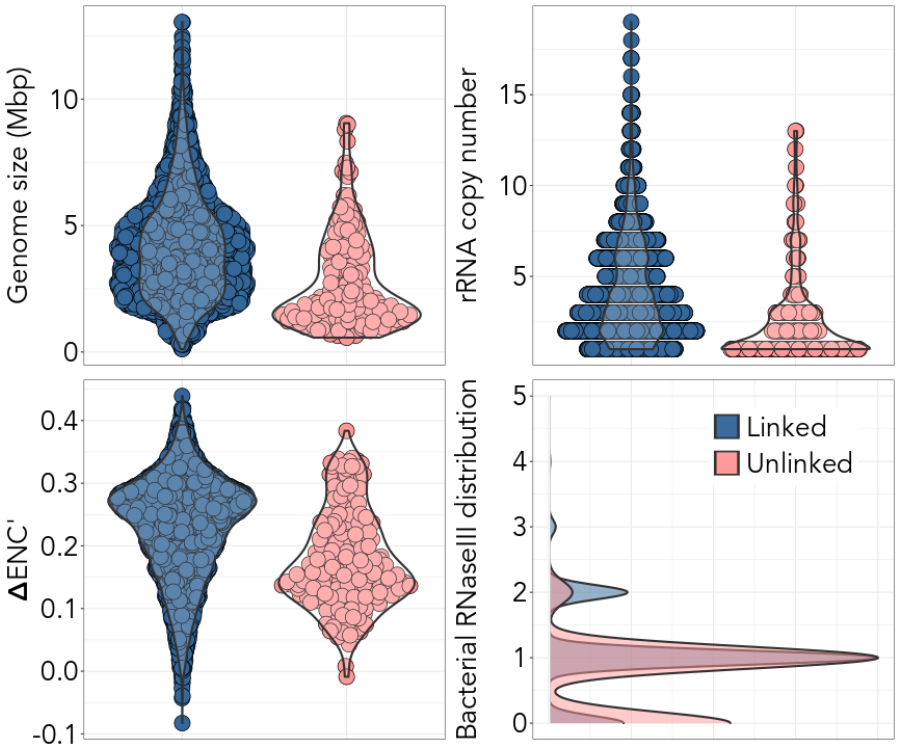


Figure 5: Genomic attributes of NCBI complete genomes based on their rRNA gene status. Linked genomes feature exclusively linked rRNA genes; unlinked genomes have at least one set of unlinked rRNA genes. We calculated these statistics using a subset of our complete genomes with one genome per unique species and rRNA gene status. **A**) Genomes with unlinked rRNA genes have smaller genomes on average, but this difference was not significant (phyloglm p =0.12, means of groups: 4.15Mbp linked, 2.72Mbp unlinked). **B**) On average, genomes with unlinked rRNA genes had significantly fewer rRNA copies (phyloglm $p < 0.0001$, means of groups: 4.25 copies linked, 2.72 copies unlinked). **C**) Genomes with exclusively unlinked rRNA genes are predicted to have longer average generation times (phyloglm p=0.028, means of groups: 0.23 linked, 0.18 unlinked; as reference E. *coli* has an average ΔENC' of 0.3). **D**) We found that there were significantly fewer RNaseIII genes in genomes with unlinked rRNA genes (only PF00636 shown, for more detail see Supplemental Figure S3: phyloglm $p<0.001$, means of groups: 1.0 linked, 0.71 unlinked). This panel is a density plot, which shows the proportional distribution of PF00636 hits for our genome dataset.

Discussion

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

404 in our complete genome dataset, yet had many members with unlinked rRNA genes in our long-
read dataset. For example, while there were no complete genomes in the phylum
406 Verrucomicrobia with unlinked rRNA genes (0/32), 38% of verrucomicrobial rRNA sequences
were unlinked (82/217) in our long-read dataset, with the majority of this group closely related to
408 the bacterium *Ca. Udaeobacter copiosus* from the DA101 soil group (Brewer et al., 2016). This
highlights the importance of using a combination of complete genomes, where genetic
410 organization and traits can be assessed rigorously, with metagenomic data that allows us to
sample the diversity found in selected environments in an unbiased manner. Together, these
412 independent datasets show that unlinked rRNA genes are present across many bacterial and
archaeal phyla.

414 One obvious ramification of the prevalence of unlinked rRNA genes in environmental
samples relates to bacterial genotyping using the full rRNA operon. While sequencing from the
416 16S rRNA gene into the 23S rRNA gene (thus including the ITS region of the rRNA operon)
can increase taxonomic resolution and allow strain level identification (Zeng et al., 2012), our
418 work shows that amplicon-based studies dependent on 16S and 23S rRNA genes being located in
close proximity may miss a large portion of bacterial and archaeal diversity. We found the
420 average distance between unlinked 16S and 23S rRNA genes in our complete genome dataset to
be ~410kbp, a rather impractical distance to PCR amplify. While strategies which use reads
422 spanning the 16S and 23S rRNA genes to improve taxonomic resolution (e.g. Zeng et al., 2012;
Cuscó et al., 2018) are less likely to introduce biases in some environments (e.g. human gut), they
424 will miss many phylogenetic groups in other environments like soil and sediment, where a
significant fraction of taxa lack 16S and 23S rRNA genes located in sufficient proximity to be
426 detected with such approaches (Figure 3).

We used our long-read metagenomic dataset to not only bypass the cultivation bias of our
428 complete genome dataset, but to also estimate the abundance of unlinked rRNA genes in a range
of microbial community types. Our analyses of the long-read metagenomic dataset show that
430 taxa with unlinked rRNA genes are far more abundant in some environments than others. Most
notably, unlinked rRNA genes were far more common in soil (where as many as 41% of rRNA
432 genes detected were unlinked) than the human gut (where no unlinked rRNA genes were
detected, Figure 3). The environments with higher proportions of unlinked rRNA genes (soil and
434 sediment) are generally thought to be populated by slower growing taxa (Brown et al., 2016;

Vieira-Silva and Rocha, 2009). Likewise, we found that genomes with unlinked rRNA genes
436 have significantly fewer rRNA copies than genomes with exclusively linked rRNA genes, a trait
which is correlated with maximum potential growth rate (Vieira-Silva and Rocha, 2009). We also
438 found that genomes with unlinked rRNA genes are predicted to have significantly longer
generation times (using codon usage bias in ribosomal proteins as a proxy for maximal growth
440 rates) compared to genomes with exclusively linked rRNA genes. These lines of evidence suggest
that unlinked rRNA genes are more common in the genomes of taxa with slower potential
442 growth rates.

The existence of numerous genomes that have unlinked 16S and 23S rRNA genes and
444 the differential frequency of these genomes across environments raise the question of the role and
implications of this genetic organization. Upon first consideration, having unlinked 16S and 23S
446 rRNA genes would seem to be disadvantageous given that both rRNA molecules are needed in
equal proportions to yield a functioning ribosome. The importance of linkage for identical
448 expression of both rRNA genes should be greater in faster growing taxa, where a higher rate of
ribosome synthesis is key to rapid growth and accounts for a large proportion of the cell energy
450 budget (Gourse et al., 1996). Studies in the fast-growing species *E.coli* have shown that, while
unbalanced rRNA gene dosage has a slight negative effect on doubling times, balanced synthesis
452 of ribosomal proteins still occurs in most cases (Siehnel and Morgan, 1985). If unequal expression
of rRNA subunits is associated with unlinked rRNA genes, it may not confer a selective
454 disadvantage in many environments (like soils and sediments) where longer generation times are
the norm, not the exception. For slower-growing taxa, the selection coefficient associated with
456 the effect of linked rRNA genes on growth may be small, because rRNAs are less expressed and
rapid growth is a trait under weaker selection. Under these circumstances, unlinked rRNA genes
458 may become fixed in populations by genetic drift. This is more likely to occur in species with
small effective population sizes, i.e. few effectively reproducing individuals, where natural
460 selection is not efficient enough to avoid the loss of genes or the degradation of genome
organizational traits that are under weak selection (Moran, 2002). This is the most frequent
462 explanation for the occurrence of unlinked 16S and 23S rRNA genes (Rurangirwa et al., 2002;
Merhej et al., 2009; Andersson and Andersson, 1999). It fits our observations that many of the
464 taxa we identified with unlinked rRNA genes are restricted to obligate intracellular lifestyles
(including members of the phyla Spirochaetes, Epsilonbacteraeota, Alphaproteobacteria, and

466 Tenericutes) or contain signatures of symbiotic lifestyles (CPR phyla; Nelson and Stegen, 2015;
Burstein et al., 2016).

468 However, fixation of mutations due to genetic drift is much less likely to explain the
presence of unlinked rRNA genes among the large proportion of free-living taxa that we have
470 identified (including members of the phyla Deinococcus-Thermus, Euryarchaeota, Chloroflexi,
Planctomycetes, and Verrucomicrobia). Some of these taxa are abundant and ubiquitous in their
472 respective environments, e.g. the Verrucomicrobia *Ca. U. copiosus* (Brewer et al., 2016) and
members of the Rikenellaceae RC9 gut group (Holman et al., 2017). These genomes do not show
474 traits typically associated with genome reduction caused by small effective population sizes, i.e.
abundant pseudogenes, transposable elements, or small genomes. Indeed, we found that
476 differences in genome size between species with linked and unlinked rRNA genes were not
significant, when accounting for phylogeny. Thus, there is little evidence that the highly
478 conserved trait of unlinked rRNA genes is caused by genetic drift in free-living taxa.

Unlinked rRNA genes could provide a selective advantage in certain circumstances,
480 which could explain their existence in free-living taxa. Transcribing the 16S and 23S rRNA
genes separately may eliminate or reduce the need for RNaseIII, which we showed to occur in
482 lower frequencies in taxa with unlinked rRNA genes (Supplemental Figure S3). We also found
RNaseIII to be completely absent in the phyla Deinococcus-Thermus and Gemmatimonadetes,
484 both phyla with high proportions of unlinked rRNA genes. Interestingly, there is evidence that
the loss of RNaseIII has negative consequences for growth in organisms with unlinked rRNA
486 genes. Recent work has shown that knocking out RNaseIII in *Borrelia burgdorferi* (a spirochete with
unlinked rRNA genes) results in a decreased growth rate (Anacker et al., 2018). On the other
488 hand, it is known that some bacteriophages hijack host RNaseIII to process their own mRNA
(Gone et al., 2016). In some cases, host RNaseIII can stimulate the translation of infecting phage
490 mRNA by several orders of magnitude (Wilcon et al., 2002), (although other phage appear
indifferent to the presence of RNaseIII; Hagen and Young, 1978). Regardless, increased
492 resistance to predation (phage attack) at the cost of reduced maximum potential growth rates is a
widely observed ecological trade-off (Bohannan and Lenski, 2000). Finally, recent work has
494 showed that some rRNA loci specialize in the translation of certain types of genes in *Vibrio*
vulnificus (Song et al., 2019). It is thus tempting to speculate that unlinked rRNA genes could
496 facilitate the production of heterogeneous ribosomes with a diverse range of characteristics.

Conclusions

While we do not know why unlinked rRNA genes are so prevalent (especially for those Bacteria and Archaea found in environmental samples for which complete genomes are not yet available), this rearrangement appears to occur more frequently in slower-growing taxa and may be related to the presence of RNaseIII. Regardless, we have shown that 52% of the phyla included in our combined datasets (37/71) have at least one member with unlinked rRNA genes, and that up to 41% of rRNA genes in some environments are unlinked - meaning unlinked rRNA genes are far from atypical anomalies. We have developed hypotheses about potential advantages of unlinked rRNA operons that could be tested experimentally - especially as a number of taxa with unlinked rRNA operons are relatively easy to manipulate in culture (Holland et al., 2006; Devos, 2013).

Acknowledgements

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

Author contributions

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

Conflict of interest statement

MA and RK own a portion of the company DNASense.

Data availability

All genomes used in this study were downloaded from NCBI, with assembly IDs listed in Supplemental Dataset S1. All Nanopore data is available at the Sequence Read Archive (SRA) under Bioproject ID PRJNA553237 or the European Nucleotide Archive (ENA) under PRJEB33278. All Moleculo data has been published previously, with publications listed in methods. Classifications and details of both the complete genome and long-read datasets are included in Supplemental Dataset S1 and S2, respectively.

References

Anacker ML, Drecktrah D, LeCoultre RD, Lybecker M, Samuels DS. (2018). RNase III Processing of rRNA in the Lyme Disease Spirochete *Borrelia burgdorferi*. *Journal of Bacteriology* **200**: 1–11.

Andersson JO, Andersson SGE. (1999). Genome Degradation is an Ongoing Process in *Rickettsia*. *Molecular Biology and Evolution* **16**: 1178–1191.

Andersson SGE, Zomorodipour A, Winkler HH, Kurland CG. (1995). Unusual Organization of the rRNA Genes in *Rickettsia prowazekii*. *Journal of Bacteriology* **177**: 4171–4175.

Bengtsson-Palme J, Hartmann M, Eriksson KM, Pal C, Thorell K, Larsson DGJ, *et al.* (2015). metaxa2: improved identification and taxonomic classification of small and large subunit rRNA in metagenomic data. *Mol Ecol Resour* **15**: 1403–1414.

Bohannan BJM, Lenski RE. (2000). Linking genetic change to community evolution: insights from studies of bacteria and bacteriophage. *Ecology Letters* **3**: 362–377.

Brewer TE, Handley KM, Carini P, Gilbert JA, Fierer N. (2016). Genome reduction in an abundant and ubiquitous soil bacterium '*Candidatus* Udaeobacter copiosus'. *Nature Microbiology* **2**: 16198.

Brown CT, Hug LA, Thomas BC, Sharon I, Castelle CJ, Singh A, *et al.* (2015). Unusual biology across a group comprising more than 15% of domain Bacteria. *Nature* **523**: 208–211.

Brown CT, Olm MR, Thomas BC, Banfield JF. (2016). Measurement of bacterial replication rates in microbial communities. *Nat Biotechnol* **34**: 1256–1263.

Burstein D, Sun CL, Brown CT, Sharon I, Anantharaman K, Probst AJ, *et al.* (2016). Major bacterial lineages are essentially devoid of CRISPR-Cas viral defence systems. *Nature Communications* **7**: 1–8.

Caporaso JG, Bittinger K, Bushman FD, DeSantis TZ, Andersen GL, Knight R. (2010). PyNAST: a flexible tool for aligning sequences to a template alignment. *Bioinformatics* **26**: 266–267.

Condon C, Squires C, Squires CL. (1995). Control of rRNA Transcription in *Escherichia coli*. *Microbiological Reviews* **59**: 623–645.

Cuscó A, Catozzi C, Viñes J, Sanchez A, Francino O. (2018). Microbiota profiling with long amplicons using Nanopore sequencing: full-length 16S rRNA gene and whole rrn operon. *F1000Res* **7**: 1755–25.

Devos DP. (2013). *Gemmata obscuriglobus*. *CURBIO* **23**: R705–R707.

Durand S, Gilet L, Condon C. (2012). The Essential Function of *B. subtilis* RNase III Is to Silence Foreign Toxin Genes. *PLOS Genetics* **8**: e1003181–11.

Eddy SR. (2011). Accelerated Profile HMM Searches. *PLoS Comput Biol* **7**: e1002195–16.

Edgar RC. (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**: 2460–2461.

Flynn TM, Koval JC, Greenwald SM, Owens SM, Kemner KM, Antonopoulos DA. (2017). Parallelized, Aerobic, Single Carbon-Source Enrichments from Different Natural Environments Contain Divergent Microbial Communities. *Front Microbiol* **8**: 1540–14.

Gone S, Alfonso-Prieto M, Paudyal S, Nicholson AW. (2016). Mechanism of Ribonuclease III Catalytic Regulation by Serine Phosphorylation. *Nature* **6**: 1–9.

Gourse RL, Gaal T, Bartlett MS, Appleman JA, Ross W. (1996). rRNA Transcription and Growth Rate–Dependent Regulation of Ribosome Synthesis in *Escherichia coli*. *Annu Rev Microbiol* **50**: 645–677.

Hagen FS, Young ET. (1978). Effect of RNase III on Efficiency of Translation of Bacteriophage T7 Lysozyme mRNA. *Journal of Virology* **26**: 793–804.

Hartmann RK, Ulbrich N, Erdmann VA. (1987). An unusual rRNA operon constellation: in *Thermus thermophilus* HB8 the 23S/5S rRNA operon is a separate entity from the 16S rRNA operon. *Biochimie* **69**: 1097–1104.

Holland AD, Rothfuss HM, Lidstrom ME. (2006). Development of a defined medium supporting rapid growth for *Deinococcus radiodurans* and analysis of metabolic capacities. *Appl Microbiol Biotechnol* **72**: 1074–1082.

Holman DB, Brunelle BW, Trachsel J, Allen HK. (2017). Meta-analysis To Define a Core Microbiota in the Swine Gut. *mSystems* **2**: 676–14.

Klappenbach JA, Dunbar JM, Schmidt TM. (2000). rRNA Operon Copy Number Reflects Ecological Strategies of Bacteria. *Applied and Environmental Microbiology* **66**: 1328–1333.

Kuleshov V, Jiang C, Zhou W, Jahanbani F, Batzoglou S, Snyder M. (2016). Synthetic long-read sequencing reveals intraspecies diversity in the human microbiome. *Nat Biotechnol* **34**: 64–69.

596 Kuleshov V, Xie D, Chen R, Pushkarev D, Ma Z, Blauwkamp T, *et al.* (2014). Whole-genome haplotyping using long reads and statistical methods. *Nat Biotechnol* **32**: 261–266.

598 Letunic I, Bork P. (2016). Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucleic Acids Research* **44**: W242–W245.

600 Liesack W, Stackebrandt E. (1989). Evidence for Unlinked rrn Operons in the Planctomycete *Pirellula marina*. *Journal of Bacteriology* **171**: 5025–5030.

602 Merhej V, Royer-Carenzi M, Pontarotti P, Raoult D. (2009). Massive comparative genomic analysis reveals convergent evolution of specialized bacteria. *Biol Direct* **4**: 13–25.

604 Moran NA. (2002). Microbial Minimalism: Genome Reduction in Bacterial Pathogens. *Cell* **108**: 583–586.

606 Moreno-Hagelsieb G, Collado-Vides J. (2002). A powerful non-homology method for the prediction of operons in prokaryotes. *Bioinformatics* **18**: S329–S336.

608 Munson MA, Baumann L, Baumann P. (1993). *Buchnera aphidicola* (a prokaryotic endosymbiont of aphids) contains a putative 16S rRNA operon unlinked to the 23s rRNA-encoding gene:
610 sequence determination, and promoter and terminator analysis. *Gene* **137**: 171–178.

Nelson WC, Stegen JC. (2015). The reduced genomes of Parcubacteria (OD1) contain signatures
612 of a symbiotic lifestyle. *Front Microbiol* **6**: 693–14.

Novembre JA. (2002). Accounting for Background Nucleotide Composition When Measuring
614 Codon Usage Bias. *Molecular Biology and Evolution* **19**: 1390–1394.

O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, *et al.* (2016). Reference
616 sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Research* **44**: D733–D745.

618 Pagel M. (1999). Inferring the historical patterns of biological evolution. *Nature* **401**: 877–884.

Parks DH, Chuvochina M, Waite DW, Rinke C, Skarshewski A, Chaumeil P-A, *et al.* (2018). A
620 standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nat Biotechnol* **36**: 996–1004.

622 Pei A, Nossa CW, Chokshi P, Blaser MJ, Yang L, Rosmarin DM, *et al.* (2009). Diversity of 23S rRNA Genes within Individual Prokaryotic Genomes. *PLoS ONE* **4**: 1–9.

624 Price MN, Dehal PS, Arkin AP. (2009). FastTree: Computing Large Minimum Evolution Trees with Profiles instead of a Distance Matrix. *Molecular Biology and Evolution* **26**: 1641–1650.

626 Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, *et al.* (2012). The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic*
628 *Acids Research* **41**: D590–6.

R Core Team. (2018). R: A language and environment for statistical computing. *R-project.org*.

630 Raoult D, Forterre P. (2008). Redefining viruses: lessons from Mimivirus. *Nat Rev Micro* **6**: 315–
319.

632 Revell LJ. (2011). phytools: an R package for phylogenetic comparative biology (and other
things). *Methods Ecol Evol* **3**: 217–223.

634 Rocha E. (2004). Codon usage bias from tRNA's point of view: Redundancy, specialization, and
efficient decoding for translation optimization. *Genome Res* **14**: 2279–2286.

636 Rurangirwa FR, Brayton KA, McGuire TC, Knowles DP, Palmer GH. (2002). Conservation of
the unique rickettsial rRNA gene arrangement in *Anaplasma*. *International Journal of Systemic and*
638 *Evolutionary Microbiology* **52**: 1405–1409.

Sharon I, Kertesz M, Hug LA, Pushkarev D, Blauwkamp TA, Castelle CJ, *et al.* (2015). Accurate,
640 multi-kb reads resolve complex populations and detect rare microorganisms. *Genome Res* **25**: 534–
543.

642 Shepherd J, Ibba M. (2015). Bacterial transfer RNAs. *FEMS Microbiol Rev* **39**: 280–300.

Siehnel RJ, Morgan EA. (1985). Unbalanced rRNA Gene Dosage and its Effects on rRNA and
644 Ribosomal-Protein Synthesis. *Journal of Bacteriology* **163**: 476–486.

Song W, Joo M, Yeom J-H, Shin E, Lee M, Choi H-K, *et al.* (2019). Divergent rRNAs as
646 regulators of gene expression at the ribosome level. *Nature Microbiology* **4**: 515–526.

Srivastava AK, Schlessinger D. (1990). Mechanism and Regulation of Bacterial Ribosomal RNA
648 Processing. *Annu Rev Microbiol* **44**: 105–129.

Tung Ho LS, Ané C. (2014). A Linear-Time Algorithm for Gaussian and Non-Gaussian Trait
650 Evolution Models. *Systematic Biology* **63**: 397–408.

Vieira-Silva S, Rocha E. (2009). The Systemic Imprint of Growth and Its Uses in Ecological
652 (Meta)Genomics. *PLOS Genetics* **6**: 1–15.

Wang Q, Garrity GM, Tiedje JM, Cole JR. (2007). Naive Bayesian Classifier for Rapid
654 Assignment of rRNA Sequences into the New Bacterial Taxonomy. *Applied and Environmental*
Microbiology **73**: 5261–5267.

656 White RA, Bottos EM, Roy Chowdhury T, Zucker JD, Brislawn CJ, Nicora CD, *et al.* (2016).
Moleculo Long-Read Sequencing Facilitates Assembly and Genomic Binning from Complex Soil
658 Metagenomes. *mSystems* **1**: 309–15.

Wilcon HR, Yu D, Peters HK III, Zhou J-G, Court DL. (2002). The global regulator RNase III
660 modulates translation repression by the transcription elongation factor N. *The EMBO Journal* **21**:
4154–4161.

662 Yuan C, Lei J, Cole J, Sun Y. (2015). Reconstructing 16S rRNA genes in metagenomic data. *Bioinformatics* **31**: i35–i43.

664 Zeng YH, Koblížek M, Li YX, Liu YP, Feng FY, Ji JD, *et al.* (2012). Long PCR-RFLP of 16S-ITS-23S rRNA genes: a high-resolution molecular tool for bacterial genotyping. *J Appl Microbiol*
666 **114**: 433–447.

Zhi X-Y, Zhao W, Li W-J, Zhao G-P. (2012). Prokaryotic systematics in the genomics era. *Antonie*
668 *van Leeuwenhoek* **101**: 21–34.

670

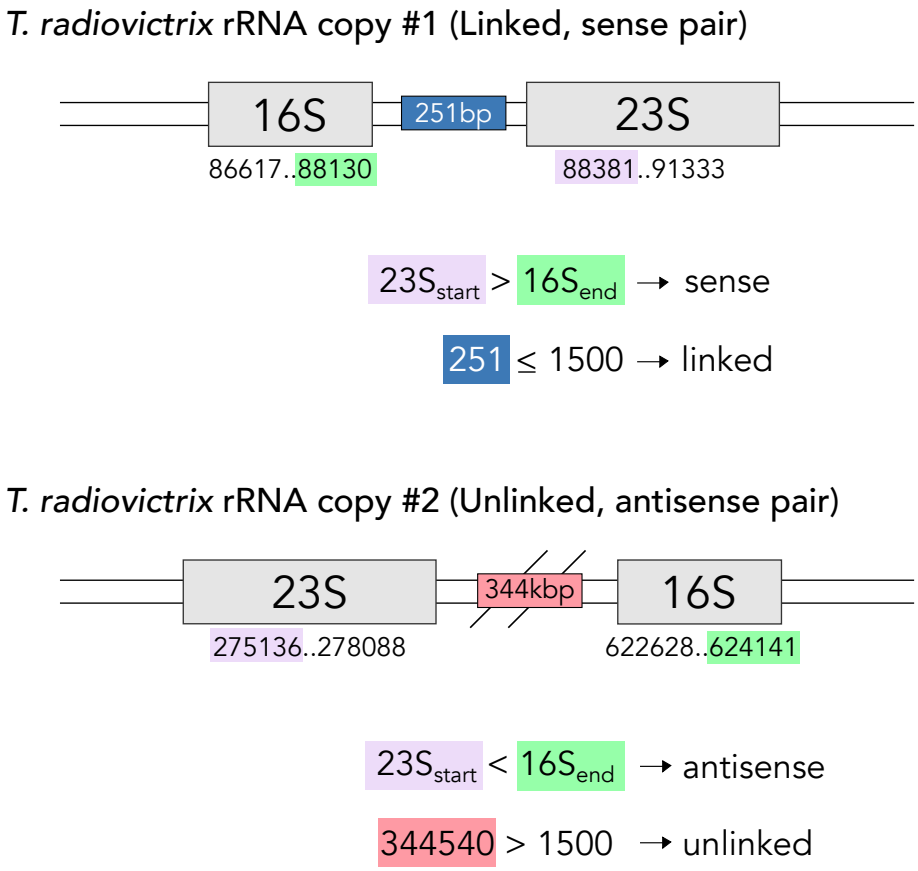
672

674

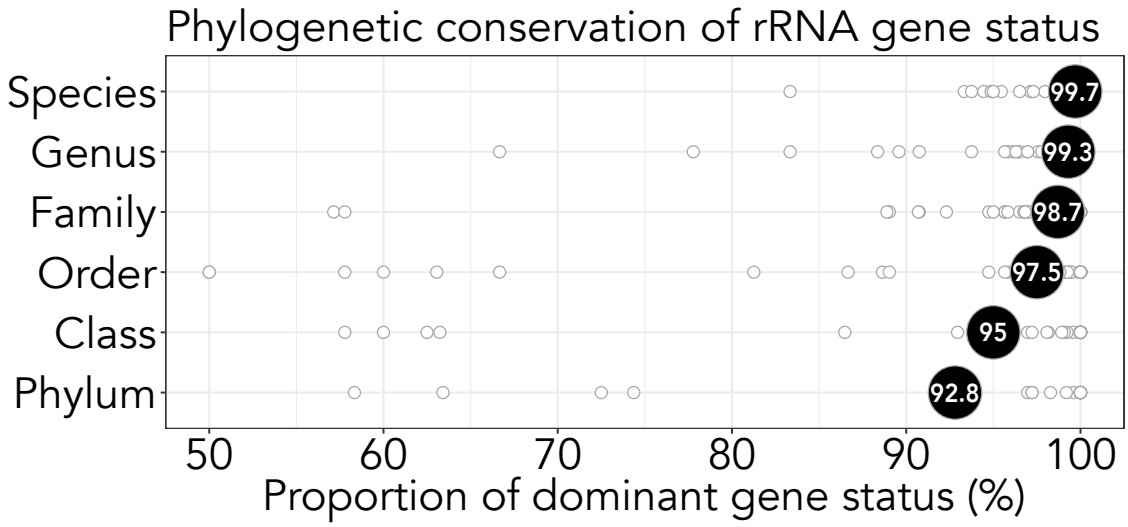
676

678

680



682 **Supplemental Figure S1:** Example of ITS length calculation in *Truepera radiovictrix* DSM17093. To classify the rRNA genes of NCBI genomes, we first used the gene ranges
684 associated with each ORF to pair the 16S and 23S rRNA genes that were closest to each other in each genome. Next, we checked for gene directionality. If the $23S_{start} > 16S_{end}$ the pair is sense,
686 otherwise antisense. We then calculated the distance between the closest edges of the 16S and 23S (the ITS); if this distance was less than or equal to 1500bp the pair was linked, otherwise
688 unlinked. *Truepera radiovictrix* has an rRNA copy number of two - one pair is linked and sense, the other is unlinked and antisense.
690

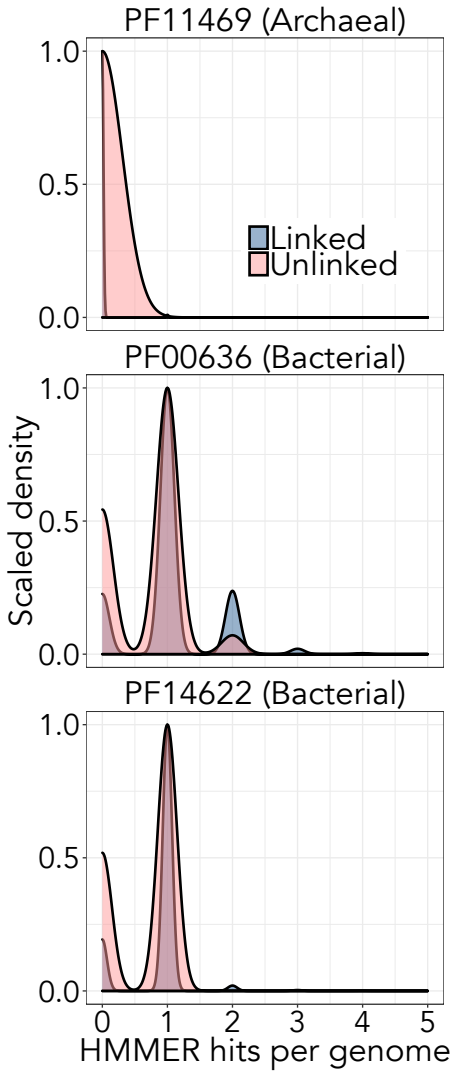


692 **Supplemental Figure S2:** rRNA gene status is phylogenetically conserved, most dramatically
at the species and genus levels. Each point represents a unique taxonomic group (for instance,
694 one point at the phylum level corresponds to Chloroflexi). We calculated the proportion of the
dominant rRNA gene status for each unique taxonomic group as a measure of trait heterogeneity
696 (for instance, 29/40 Chloroflexi genomes contain an unlinked rRNA gene, making the
proportion of the dominant gene status in this phylum 72.5%). The labeled dot corresponds to
698 the average for all taxonomic groups at that specific level. Genera and families with a dominant
gene status proportion < 70% are: *Sodalis* (genus), Cellvibrionaceae (family), Spirochaetaceae
700 (family). We also calculated Pagel's lambda to show that rRNA gene status has a strong
phylogenetic signal (lambda = 0.96, $p < 0.0001$).
702

704

706

708



Supplemental Figure S3: Genomes with unlinked rRNA genes encode fewer bacterial RNaseIII genes. We found that there were significantly fewer bacterial RNaseIII genes in genomes with unlinked rRNA genes (PF00636: phyloglm $p<0.001$, means of groups: 1.0 linked, 0.71 unlinked; PF14622: phyloglm $p = 0.007$, means of groups: 0.86 linked, 0.66 unlinked). We were unable to check this relationship for archaeal RNaseIII due to the size of our archaeal dataset. However, the archaeal RNaseIII PF11469 was found in only two clades that featured exclusively linked rRNA genes (Euryarcheaota family Thermococcaceae and Crenarchaeota family Thermofilaceae). We calculated these statistics using a subset of our complete genomes with only one genome per unique species and operon status. These figures are density plots and are scaled to be proportional between the linked and unlinked groups.