A method for achieving complete microbial genomes and improving bins from metagenomics data

Authors:

Lauren M. Lui[1], Torben N. Nielsen[1], Adam P. Arkin[1,2,3]*

Affiliations:

[1]Environmental Genomics and Systems Biology Division, Lawrence Berkeley National Laboratory, Berkeley, CA, USA.

[2]Department of Bioengineering, University of California, Berkeley, CA, USA

[3]Innovative Genomics Institute, Berkeley, CA, USA

*Correspondence: aparkin@lbl.gov

Abstract

Metagenomics facilitates the study of the genetic information from uncultured microbes and complex microbial communities. Assembling complete microbial genomes (*i.e.*, circular with no misassemblies) from metagenomics data is difficult because most samples have high organismal complexity and strain diversity. Less than 100 circularized bacterial and archaeal genomes have been assembled from metagenomics data despite the thousands of datasets that are available. Circularized genomes are important for (1) building a reference collection as scaffolds for future assemblies, (2) providing complete gene content of a genome, (3) confirming little or no contamination of a genome, (4) studying the genomic context and synteny of genes, and (5) linking protein coding genes to ribosomal RNA genes to aid metabolic inference in 16S rRNA gene sequencing studies. We developed a method to achieve circularized genomes using iterative assembly, binning, and read mapping. In addition, this method exposes potential misassemblies from k-mer based assemblies. We chose species of the Candidate Phyla Radiation (CPR) to focus our initial efforts because they have small genomes and are only known to have one ribosomal RNA operon. We present 34 circular CPR genomes, one circular Margulisbacteria genome, and two circular megaphage genomes from 19 public and published datasets. We demonstrate findings that would likely be difficult without circularizing genomes, including that ribosomal genes are likely not operonic in the majority of CPR, and that some CPR harbor diverged forms of RNase P RNA. Code and a tutorial for this method is available at https://github.com/lmlui/Jorg.

Introduction

Shotgun metagenomics and marker gene sequencing are powerful tools to survey and study organisms that we cannot yet isolate and culture in the laboratory. This is especially true for environmental samples where culturability estimates for bacterial and archaeal communities range from ~22-53% for soil, ~10-70% for ocean and lakes, and ~8-32% for ocean sediment [1]. Scientists have turned to shotgun metagenomics to provide genome-resolved analysis of complex samples, but assembling genomes from shotgun metagenomics data is inherently more difficult than assembling those from cultured isolates. Challenges in metagenomics assembly arise from the heterogeneity of samples, available sequencing technology, and the limitations of bioinformatics algorithms we use for assembly and genome binning. Metagenomes contain uneven amounts of an unknown number of genomes, which creates a compounded computational problem in terms of simplifying assumptions, time, and computer memory.

In the 1990s when the first genomes were sequenced and assembled, scientists used long reads from Sanger sequencing and overlap layout consensus (OLC) methods for assembly [2]. With the development of next-generation sequencing technologies, we gained the ability to sequence millions of reads at a massively reduced cost, but using traditional OLC algorithms became too computationally intensive. The computational complexity of OLC algorithms scale as the square of the number of input reads (because each read is compared to every other read), so they are impractical for datasets of millions of reads, compared to the thousands of reads generated from Sanger sequencing. To handle the deluge of sequencing data (in terms of the volume of reads and projects) de Bruijn graph assembly methods were developed. The time and memory complexity of de Bruijn based assembly algorithms typically scale with the size of the metagenome instead of the number of reads. However, de Bruijn graph methods can introduce misassemblies; due to the decomposition of reads into k-mers, context is lost and it is

69 possible for the graphs to contain paths that do not correspond to real genomic sequence [3,4].
70 Traditional OLC assemblers such as the Celera Assembler [5], SGA [6] and MIRA [7] ensure
71 that only contigs consistent with actual genome sequence are produced (this is sometimes
72 referred to as maintaining read coherence). Using long-read sequencing can overcome some of
73 the issues with k-mer based assembly and newer assemblers for this type of data have started
74 using OLC assembly methods again [3]. However, long-read sequencing requires much larger
75 amounts of DNA (micrograms) that is high quality and high molecular weight (10-50kb) as
76 compared to short read technologies (as little as 1 ng). Extracting enough high molecular weight
77 DNA can be limited by sampling costs and available biomass, especially for certain types of
78 environmental samples, so short read sequencing may sometimes be the only option.
79 Beyond assembly, a key challenge with metagenomics is grouping contigs into genome
80 bins. We use “contig” in the way it was originally defined by Rodger Staden, where a contig is a
81 set of overlapping segments of DNA from shotgun sequencing [8]. It is rare for a complete
82 genome to be assembled into a single piece *de novo* from short reads, so contigs are grouped
83 into “bins,” often based on coverage and tetranucleotide frequencies. If two contigs belong to
84 the same genome, they are expected to have similar coverage and tetranucleotide profiles [9].
85 However, coverage has problems for multiple reasons. If a particular microbe is growing rapidly,
86 some regions may have higher coverage than the rest of the genome [10]. In addition, for
87 organisms where the copy number of ribosomal RNA (rRNA) operons exceeds unity, the
88 contig(s) with the rRNA genes will not have the same coverage as the rest of the contigs in the
89 genome. This is also true of other multi-copy genes and other repetitive elements.
90 Tetranucleotide frequencies are problematic because horizontally transferred regions may have
91 different frequencies than the rest of the genome [11] and this can result in such pieces being
92 put into different bins by the binning algorithm. Despite these issues, binning is helpful in
93 identifying potential genomes in metagenomics data, especially when using short read
94 sequencing technologies.

95 To evaluate the quality of a bin, the metrics of “contamination” and “completeness” are
96 often used. Completeness and contamination are detected generally by looking for violations of
97 conserved features of complete isolate genomes. Such features include having complete sets of
98 universally (or at least phylogenetically) conserved single-copy protein genes without any
99 duplication or excessive variation in tetranucleotide frequency. Other measures of completeness
100 have been suggested, such as establishment of a core conserved set of ubiquitous genes.
101 Tools such as RefineM [12] and CheckM [13] apply these rules to assemblies to determine
102 completeness and contamination. However, these tools are not always accurate for species
103 that are not well studied. Candidate Phyla Radiation (CPR) species are often classified as
104 having 60-80% completeness by these tools, even for circular genomes. To overcome
105 challenges of binning, scientists have started to assemble circular, complete genomes from
106 metagenomes [14–23], which are also called CMAGs (complete metagenomic-assembled
107 genomes) [24]. In comparison to genome bins, a high quality reference collection that controls
108 for misassemblies and is composed of circular genomes (1) provides more accurate inference
109 of identity and estimation of capabilities of uncultured microbes within complex microbiomes, (2)
110 allows more accurate taxonomic assessment of the composition of these microbiomes through
111 better linkage of marker genes in single organisms, (3) provides high-quality scaffolds on which
112 reads can be assembled, both to allow measures of strain variation within a microbiome study
113 and to aid in better assembly of reads across many microbiome samples, and (4) affords the
114 ability to study synteny and genomic context of genes in these organisms. In addition, while
115 there are existing methods for generating high-quality MAGs, there is evidence that these MAGs
116 still contain significant contamination by exogenous sequence and have misassemblies
117 triggered by lack of read coherence. Circularization of genomes helps assure that there is likely
118 no contamination in the assemblies. Despite the advantages of circularizing genomes, very few
119 metagenomics studies to date (<30) have published circular genomes [14,15,24,16–23].

We describe a semi-automated method that facilitates recovery of circular archaeal and bacterial genomes from metagenomics data and that also provides checks for misassemblies. In this study we do not focus on how many circularized genomes we can get from a dataset, but rather the method itself to help circularize bins of interest and to ensure that they are high quality. To assist with the travails of circularizing genomes, our method overcomes issues from using k-mer based assembly and automates iterative extension of contigs. Our general approach is to produce a “standard” metagenomic assembly, bin using a “standard” binning tool, extract reads based on k-mer similarity and reassemble these using a “standard” isolate-focused assembler. To demonstrate this method, we have obtained 34 circular CPR genomes, one circular Margulisbacteria genome, and two circular megaphage genomes from 19 public and published metagenomics datasets. To our knowledge, only 41 other CPR circularized genomes have been published from 9 studies [14–22], so we believe this to be the largest presentation of circularized CPR genomes in a single study. With this set, we demonstrate findings that would likely be difficult without a large number of unique circularized genomes, including that ribosomal genes are likely not operonic in the majority of CPR and finding diverged forms of RNase P RNA in CPR species.

Results

Circularization Method

To have confidence that the genomes we generate match real organisms, we looked for criteria that would indicate that a genome is circular and complete. The literature is replete with techniques and proposals for measuring the completeness of genomes and to what level they are complete [13,16], but these often have difficulty when encountering novel genomes because their criteria are based on known isolate genomes. We focus on evidence for incompleteness in terms of missing essential genes that are found across the tree of life. That is, we are more

concerned with ensuring that anything we label a circularized genome meets basic criteria that indicates that it is not incomplete. We posit that a complete, circularized genome must satisfy at least the following conditions:

1. The genome is either circular or there is solid evidence that it is linear. While rare, linear bacterial genomes exist [25].
2. The genome has a full complement of rRNAs (16S, 23S, 5S), transfer RNAs (all amino acids represented), and RNase P RNA (since this is nearly universally necessary to process tRNA transcripts). Absence of any of these genes must be explained. We advocate using these as a check for a complete genome instead of single copy marker protein genes because checks for single copy marker protein genes can vary by clade; in only rare instances would these noncoding RNA genes be missing [26,27].
3. There is significant read coverage across the entire genome. Assemblies that rely on single reads for continuity are prone to error. With some exceptions for very high coverage organisms, we generally require minimum coverage no lower than 30% of the average coverage.

To develop and test this method, we mined the Sequence Read Archive (SRA) hosted at the National Center for Biotechnology Information (NCBI) [28] for metagenomic sequencing data generated from groundwater samples, where CPR are prevalent (see Table 1 for a list of the datasets used in this study. We focused on assembling CPR genomes because they are (1) small and thus easier to assemble, (2) to the best of our knowledge only have one set of rRNA genes. These two criteria gave us the easiest targets for circularization.

170 **Table 1: Description of Metagenomes in this Study.**

171 We focused on groundwater datasets because they have a higher fraction of CPR. Many of the datasets

172 are from studies of anthropologically contaminated sites. All identifiers are SRA except for SURF_D.

Identifiers	Study Description	Reference
ERX2165959	Groundwater from monitoring wells from naphthalene contaminated surface sediments, where effluent from the coal-tar contaminated groundwater surfaces	[29]
SRX1085364	Terrestrial subsurface C, N, S and H cycles cross-linked by metabolic handoffs	[16]
SRX1775573 SRX1775577 SRX1775579	Potential for microbial H2 and metal transformations associated with novel bacteria and archaea in deep terrestrial subsurface sediments	[21]
SRX1990955	Groundwater microbial communities from Rifle, Colorado - Rifle Oxygen_injection A2 metagenome	[16]
SRX2838984	Coupling Microbial Communities to Carbon and Contaminant Biogeochemistry in the Groundwater-Surface Water Interaction Zone	[30–32]
SRX3024504 SRX3024507 SRX3024508	DNA from groundwater after nitrate injection, filter size 0.2μm and 0.1μm	[33]
SRX3307784	Subsurface groundwater microbial communities from S. Glens Falls, New York, USA - GMW37 contaminated, 5.8 m metagenome	[29]
SRX3348993	Development of a pipeline for high-throughput recovery of near-complete and complete microbial genomes from complex metagenomic datasets: Groundwater sample from aquifer - Crystal Geyser CG19_WC_8/21/14_NA	[34]
SRX3574179	Investigating microbial roles in methane emission, contaminant degradation, and biogeochemical cycles in an aquifer near a municipal landfill	Laura Hug Lab; https://uwaterloo.ca/hug-research-group/
SRX3602289 SRX3602720	Groundwater microbial communities from the Aspo Hard Rock Laboratory (HRL) deep subsurface site, Sweden	Mark Dopson Lab; https://lnu.se/en/staff/mark.dopson/
SURF_D	Groundwater samples from the Sanford Underground Research Facility (SURF)	[35]

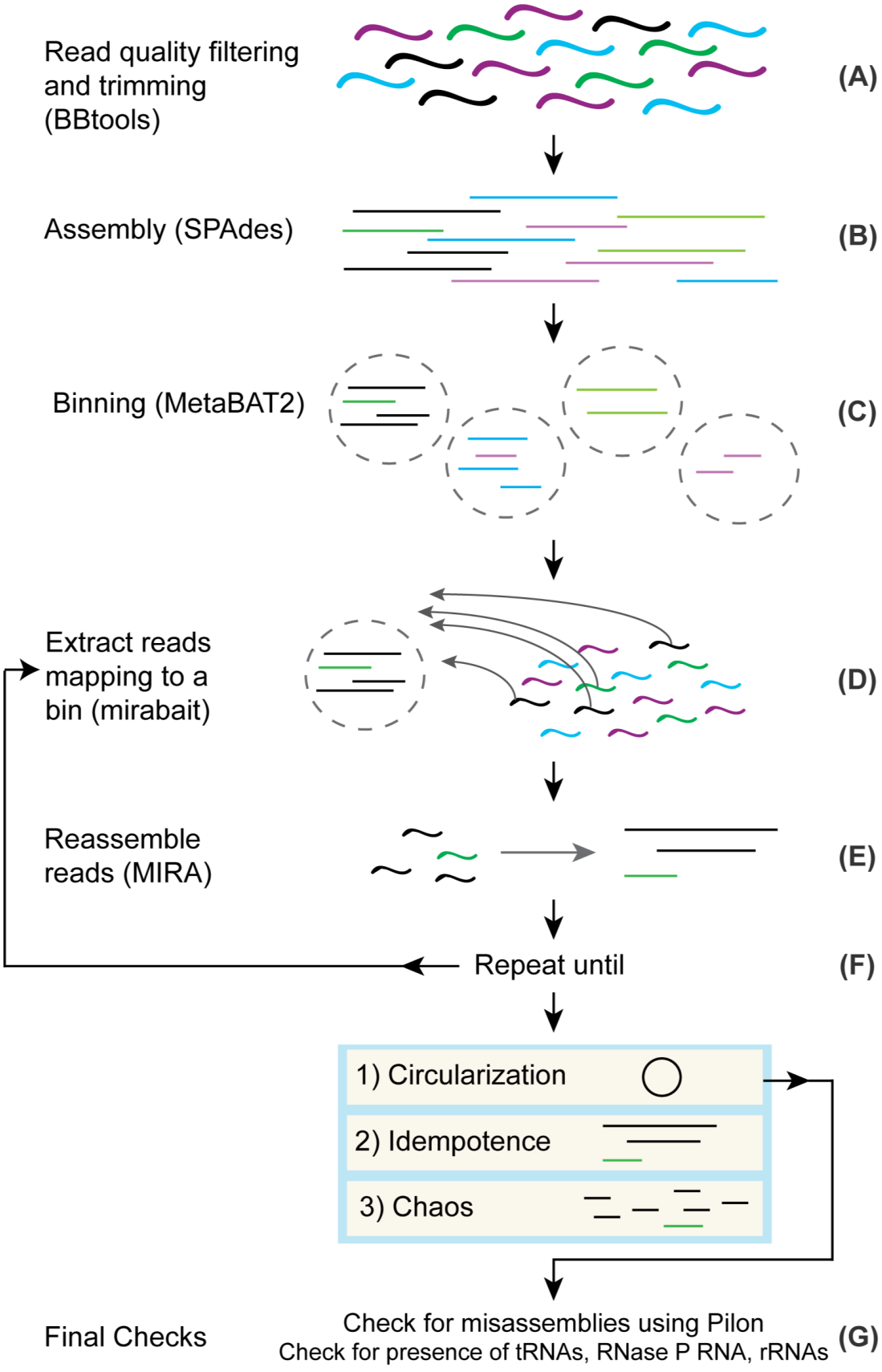
173

174 The first steps of the method are standard to regular metagenomics assembly pipelines.

175 For each metagenome, we trimmed the reads to remove any remaining Illumina adapter

176 fragments and low quality ends, as well as whole reads that weren't of sufficient quality, using

177 BBtools (Fig 1A). Next, we assembled the processed reads using SPAdes [36] (Fig 1B). We
178 proceeded with successful assemblies and used MetaBAT 2 [37] (Fig 1C) to produce a
179 collection of bins for each. We went through 188 assembled metagenomes and picked bins with
180 5 or fewer contigs and coverage above 40X, although we made exceptions for bins that looked
181 promising, such as a bin with many contigs, but with one or two large contigs that comprise
182 most of the bin's sequence length (Table 2). We used GTDB-Tk [38] to classify the bins and
183 picked a set of CPR bins. We used these bins as "bait" to select read pairs for use with the
184 isolate-focused assembler MIRA (Fig 1D).
185
186 **Fig 1. General method for circularizing genomes from metagenomes.**
187 (A) Reads have adapters trimmed and low quality reads are filtered using BBtools. (B)
188 Processed reads are assembled into contigs using SPAdes. (C) Contigs are grouped into bins
189 using MetaBAT 2. (D) After choosing a bin for circularization, reads mapping to the bin are
190 extracted from the original processed reads and used as input into (E) where they are
191 assembled into contigs using MIRA. (F) Steps D and E are repeated as necessary until the bin
192 is deemed to be in a Circularization, Idempotence, or Chaos state. (G) If a bin is deemed
193 circular, we do final checks for misassemblies using Pilon and for the presence of rRNA, tRNAs,
194 and RNase P RNA before officially calling the bin a circularized genome.
195


Read quality filtering and trimming (BBtools)
(A)
Assembly (SPAdes)
(B)
Binning (MetaBAT2)
(C)
Extract reads mapping to a bin (mirabait)
(D)
Reassemble reads (MIRA)
(E)
Repeat until
(F)
1) Circularization
2) Idempotence
3) Chaos
Final Checks
Check for misassemblies using Pilon
Check for presence of tRNAs, RNase P RNA, rRNAs
(G)

196

197 **Table 2: List of 35 circularized bacterial genomes in this study.**

198 Thirty-four of the genomes are classified as CPR and one is classified as a Margulisbacteria by GTDB-Tk.

199 The coverage, original number of contigs, and length of the original bin is also included.

Genome ID	GTDB-TK Taxonomy (Class)	Genome Size (bp)	Original Bin Stats		
			Num of Contigs	Coverage	Size of Bin (bp)
ERX2165959_bin_184	Paceibacteria	523910	2	126X	531828
ERX2165959_bin_23	Microgenomatia	1147419	1	134X	1146985
ERX2165959_bin_53	Microgenomatia	646630	3	116X	780569
ERX2165959_bin_80	Microgenomatia	1014979	1	216X	1024366
SRX1085364_bin_95	Microgenomatia	819458	4	109X	826172
SRX1775573_bin_5	Gracilibacteria	998919	1	104X	999239
SRX1775577_bin_36	Gracilibacteria	999108	1	74X	998727
SRX1775579_bin_0	Dojkabacteria	732899	1	51X	733907
SRX1990955_bin_0	Margulisbacteria (phylum); WOR-1	1676518	1	56X	1673447
SRX1990959_bin_38	Paceibacteria	585024	2	35X	582583
SRX2838984_bin_5	Paceibacteria	1030062	7	274X	1030337
SRX3024504_bin_47	Paceibacteria	672946	1	41X	673298
SRX3024507_bin_14	ABY1	1064268	4	48X	1106392
SRX3024507_bin_96	Paceibacteria	672946	1	29X	673073
SRX3024508_bin_27	Paceibacteria	672946	1	91X	673184
SRX3307784_bin_186	Paceibacteria	581622	4	174X	578873
SRX3307784_bin_197	Paceibacteria	822324	8	203X	962091
SRX3307784_bin_224	Microgenomatia	646579	3	118X	780569
SRX3307784_bin_45	UBA1384	872881	1	93X	872947
SRX3307784_bin_80	Microgenomatia	1013439	1	220X	1024366
SRX3307784_bin_91	Paceibacteria;	523446	3	129X	531688
SRX3348993_bin_93	Saccharimonadia	1005778	1	64X	1005352
SRX3574179_bin_116	Saccharimonadia	949592	4	37X	1000785
SRX3574179_bin_12	ABY1	1027227	1	135X	1028393
SRX3574179_bin_242	Microgenomatia	1127729	3	48X	1130277
SRX3574179_bin_244	Paceibacteria	707009	1	106X	708095
SRX3574179_bin_38	Paceibacteria	682291	5	124X	680516
SRX3574179_bin_63	ABY1	1209998	2	88X	1209921
SRX3574179_bin_75	ABY1	954552	3	152X	979051
SRX3602289_bin_51	Microgenomatia	852061	2	94X	853126
SRX3602720_bin_127	Paceibacteria	750123	18	160X	826132
SRX3602720_bin_74	ABY1	1035066	3	32X	1041342
SRX5650846_bin_20	Microgenomatia	947771	9	53X	949620
SURF_D_bin_21	Microgenomatia	978848	3	193X	978553
SURF_D_bin_31	Microgenomatia	1496927	5	159X	1501138

200

201 The purpose for assembling contigs first with SPAdes and then switching to MIRA with a
202 subset of reads is that the computational requirements of MIRA make it impractical as a
203 metagenome assembler. This is in part because MIRA does full alignment of the reads during
204 assembly. We would like to note that OLC metagenomics assemblers exist [3], but their memory
205 and time requirements are high compared to SPAdes or are not appropriate for assembling
206 bacterial genomes with paired-end sequencing data. MIRA has been used to extract
207 mitochondrial genomes [39] from eukaryotic sequencing projects. Our approach is very similar;
208 instead of providing seed sequences to separate the mitochondrial genomes from the
209 eukaryotic DNA, we use bins as seeds to separate genomes from the entire metagenomics
210 dataset. In our experience, MIRA produces superior results for isolates, and it also provides
211 additional features that benefit our method. MIRA comes with the tool mirabait, which provides
212 support for extracting read pairs based on k-mer content. MIRA also has a variety of features
213 that help expose problematic parts of assemblies. For example, MIRA sets tags to indicate parts
214 of the assembly that may require manual intervention, based on changes in coverage, GC, and
215 other anomalies. These tags are extremely useful in conjunction with traditional assembly
216 finishing tools such as Gap4/Gap5 [40].
217 Perhaps even more critical to this method than MIRA's utilities is the fact that MIRA also
218 ensures read coherence as an overlap-based assembler, unlike k-mer based assemblers like
219 SPAdes. SPAdes is commonly used for metagenomics assembly, and in our experience,
220 produces results that are as good as any other metagenomics assembler that is typically
221 recommended [4]. However, there are often misassemblies caused by running SPAdes on a
222 large and heterogeneous collection of metagenomes with the same set of k-mers. Ideally, the
223 user would conduct tests to find the optimal collection of k-mers for each individual
224 metagenome, but this step is time consuming. Thus, many users - us included - pick a canonical
225 set like 21, 33, 55, 77, 99 and 127 that in most cases give the greatest contiguity in the
226 assembly. Unfortunately, this practice can produce illusionary contiguity if the read coverage

227 cannot support all of the k-mer sizes [3]. Larger k-mers increase contiguity, but the read
228 coverage may not support them. By using MIRA, contigs that do not have read coherence may
229 be exposed.

230 After we used mirabait to extract read pairs that mapped to selected bins (Fig 1D) and
231 reassembled them using MIRA (Fig 1E), we iterated these two steps (Fig 1F). This iterative
232 process results in “digital primer walking” to extend the contigs of the bin, similar to primer or
233 genome walking that was initially used to sequence genomes in the late 1980s to early 1990s
234 [41]. At each iteration, reads with a portion mapping to any part of a contig will be included and
235 can lead to extension or fusion of contigs. We specifically chose to reassemble all of the reads
236 during each iteration to provide a more robust handling of repeats. On occasion, the extension
237 of the contigs resulted in overlap with contigs from other bins and unbinned contigs. Manually
238 including these contigs as part of the bait can speed up the process significantly. However, we
239 also routinely examined intermediate results and, in some cases, we saw anomalous coverage
240 values for different contigs indicating possible chimerism. If we saw the bin containing contigs
241 with significantly different coverage values (>10% difference), we removed the offending contigs
242 and restarted. We iterated this process until one of these outcomes occurred:

243 1. **Circularization.** For us to decide that this had occurred, we looked for a single contig
244 with a significant - and exact - repeat at the ends. In addition, we required that the repeat
245 be at least 100 nt in length, was longer than any other repeat in the contig, and did not
246 match any of the other repeats.
247 2. **Idempotence.** In some cases, we observed no change in the assembled contigs after a
248 round of read pair extraction and reassembly with MIRA. We examined some of these
249 instances in detail and we believe the change in coverage causes MIRA to refuse to
250 continue extending contigs. It is possible to adjust MIRA’s thresholds of what constitutes
251 low and high coverage to allow contig extension to continue. However, this modification
252 increases the risk of collapsing repeats or creating chimeric assemblies.

253 3. **Chaos.** There are cases where a bin is shattered into a multitude of pieces. We are not
254 certain as to the exact cause, but this result is likely due to misassemblies from the initial
255 SPAdes assembly (discussed in more depth in a later section). Chaos appears strongly
256 correlated with GC and tends to occur more often when the GC content is high. We have
257 investigated a few in more detail and for some found that the contigs that shatter have
258 low 127-mer coverage as reported by SPAdes. We believe that Chaos bins are caused
259 by lack of read coherence in the contigs and if that is indeed the case, there is little we
260 can do. Once we see Chaos set in, it appears to be permanent.
261
262 After circularizing a contig, we did final checks for misassemblies with Pilon (Fig 1G). We
263 used Pilon [42] on the contig and then we rotated it by half the length to ensure that the ends
264 were in the middle and applied Pilon again (Fig 1G). We rotate the genomes because Pilon is
265 not capable of covering the ends of a contig. While Pilon found minor insertions/deletions due to
266 the circularization, it did not find any other issues in the genomes.
267 We next searched the genomes for a full complement of ribosomal RNAs (16S, 23S,
268 5S), tRNAs (all amino acids represented) and RNase P RNA to check that the genome was
269 correctly circularized and was not missing regions. For RNase P RNA, we needed to manually
270 reduce score thresholds to find all RNase P RNAs (discussed in more detail in a later section).
271 We were able to find tRNAs for all amino acids, although some tRNAs had Group I introns,
272 making them difficult to detect. Structural RNAs are sometimes invaded by Group I introns,
273 which is particularly true for tRNAs [Patricia Chan, private communication]. When a genome
274 passed the final check with the detection of the set of non-coding RNAs, we considered the
275 genome to be accurately circularized.
276 SRX3307784_bin_197 is an example of a bin that appeared to be circular, but did not
277 pass the check of having RNase P RNA. The assembled contig had a solid 414 base pairs of
278 overlap at the ends, ribosomal proteins present and tRNAs for all amino acids. However, we did

279 not find a copy of RNase P RNA even when we lowered the detection threshold. This caused us
280 to look closer and we discovered that there was another contig in the assembly which we had
281 thought was contamination after the initial circularization. This contig has a copy of RNase P
282 RNA and we were able to incorporate it into the assembly after we discovered a repeat that was
283 too long for the reads to span and that Pilon did not detect. We came to the conclusion that this
284 was a case of false circularization. To address the misassembly, we put the bin through more
285 iterations with mirabait and MIRA, which resulted in a larger genome which passed all of the
286 final checks.

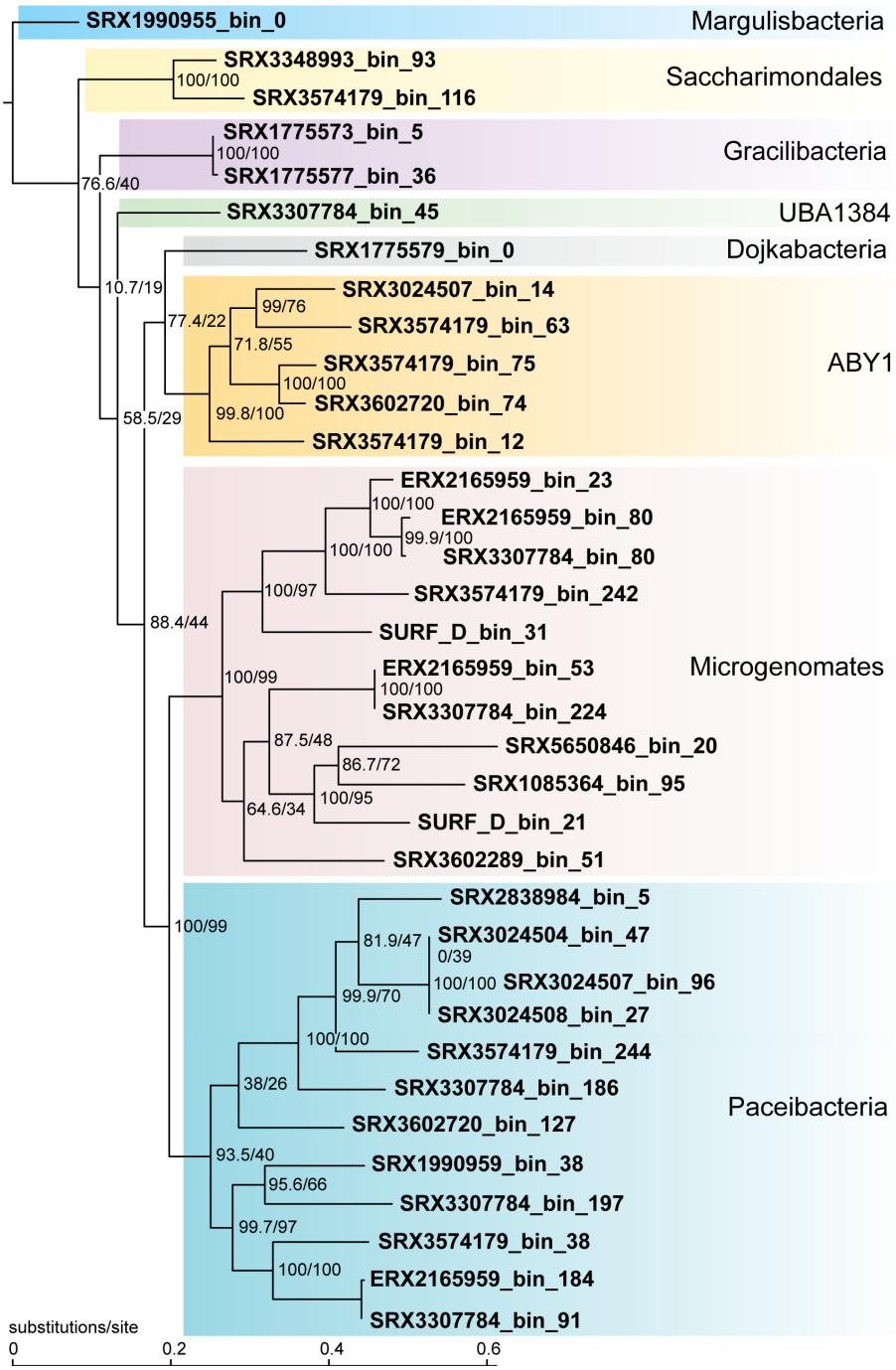
287 In approximately 10% of the cases we attempted, we succeeded in creating a
288 circularized genome out of a bin. However, our selection of bins was not random as we were
289 heavily biased in favor of bins that we judged easiest to circularize, such as selecting bins
290 classified as CPR, had relatively few contigs, and had solid coverage. We are confident that we
291 can assemble more genomes from these datasets because we have also been able to
292 circularize genomes from archaea and other bacteria; these genomes will be published in future
293 papers. We intend to make these genomes generally accessible as we finish them. The code for
294 iterating to pull reads mapping to a bin and reassembly with MIRA are available on Github
295 (https://github.com/lmlui/Jorg) and we are currently making an app in the Department of Energy
296 KnowledgeBase [43].

297 Description of circularized CPR genomes

298 Using our method, we circularized 34 CPR genomes and one Margulisbacteria genome
299 (Table 2 and Fig 2). To create a phylogenetic tree, we used a structural alignment 16S rRNA
300 genes. During this process we found that many of the 16S included large introns with
301 LAGLIDADG homing endonucleases, an observation that has been noted in other CPR studies
302 [17].
303

Fig 2. Phylogenetic Tree based on SSU structural alignment.

At the base of the branches, in the fraction the top value is SH-aLRT which is a branch test [44] and the bottom value is the bootstrap value. Class is listed for the CPR based on GTDB-TK taxonomy of that genome. The class UBA1384 is also known as Berkelbacteria. We specified the Margulisbacteria as the outgroup when creating the tree using IQ-TREE [45].



310 In general, these genomes are novel, but in one case, SRX1085364_bin_95, we found
311 that the genome is 100% identical to a previously circularized genome (INSDC Accession
312 CP011214.1) from that dataset [17]. Assembling the same genome as another group helps
313 validate our findings and that with careful manual curation two different groups can come to the
314 same assembly result despite differences with assemblers. Four of the other genomes had 16S
315 genes that had 100% hits in NCBI. Some of the 16S genes only had percent similarity in the low
316 80s to other sequences in Genbank.

317 We compared the genome sizes before and after circularization, and in most cases the
318 size of the genome decreased after circularization compared to the original bin. Typically the
319 genome shrank from a few hundred bases to a few thousand, but in some cases the genome
320 shrank by more than 130kbp (Table 2). This shrinkage may be a result of SPAdes artifacts that
321 MIRA determines to be lacking in read coherence. We are also aware of cases where SPAdes
322 generates contigs that are effectively duplicates of each other apart from short stretches at the
323 ends, and MIRA is able to resolve these into one contig.

324 Misassembled contigs can be found with MIRA, *i.e.* Chaos

325 In some cases, when we attempt the reassembly step of a bin with MIRA, we end up
326 with many more short contigs than what was in the original bin. SRX3024505_bin_48 started
327 with just 7 contigs with coverage 21X and a GC content of 59%. Superficially, it looks like a
328 reasonable bin. GTDB-Tk classifies it as a CPR in the Gracilibacteria class. However, after
329 going through 5 rounds of our method, we ended up with 136 contigs, *i.e.*, this a Chaos bin.

330 We do not know exactly what happened in the case of SRX3024505_bin_48, but we see
331 Chaos routinely during reassembly with MIRA. In some cases, we have been able to conclude
332 that Chaos results from insufficient read support for the largest k-mer used in the original
333 SPAdes assembly. Put differently, the assembly graph wasn't sufficiently well connected at the
334 highest k-mer used. To determine if the Chaos of SRX3024505_bin_48 was solely a result of

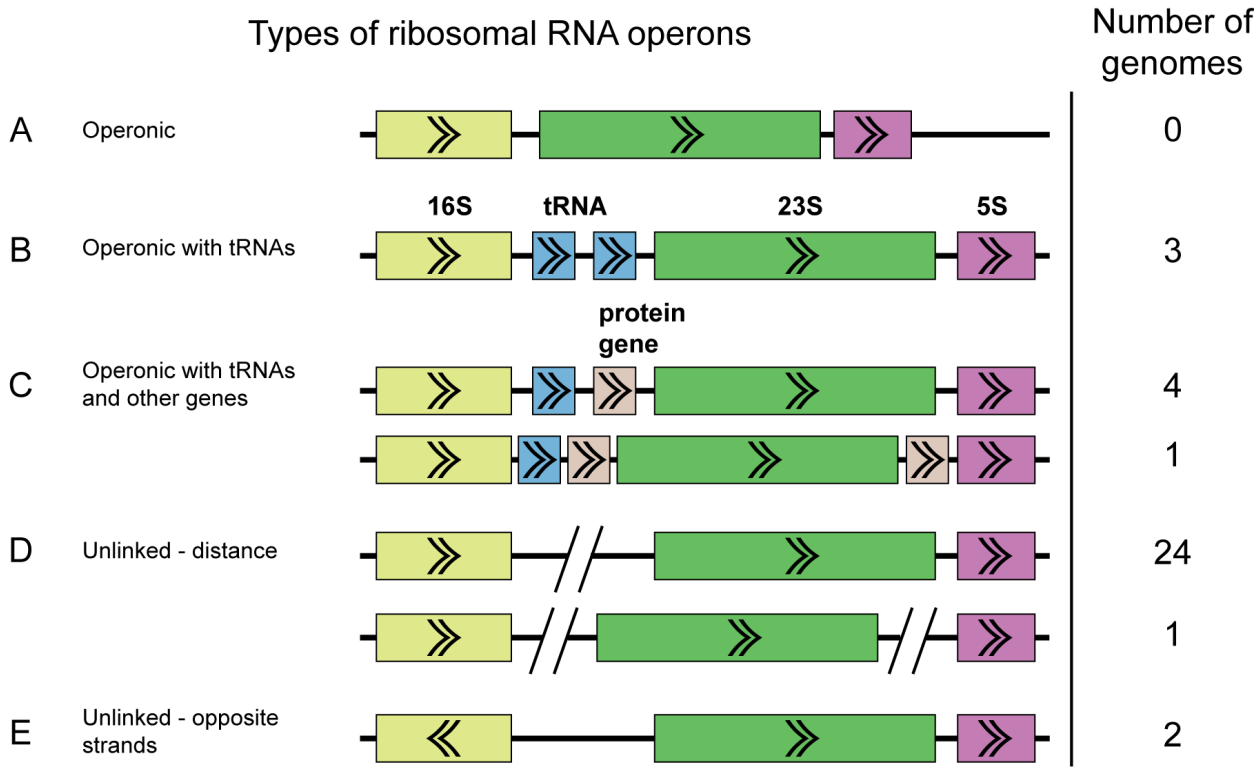
335 using MIRA, we used the same reads that we gave to MIRA as input into a SPAdes assembly.
336 We ended up with 47 contigs, which was still significantly worse than the original bin. It is worth
337 noting that the size of SRX3024505_bin_48 remained relatively constant during the testing and
338 reassembly process. Although Chaos is a disappointing result in assembly, knowing that a bin
339 likely has misassembled contigs is valuable.

340 Chaos predominantly occurs when the coverage is less than ~30X. Most of the genomes
341 we successfully circularized have much higher coverage. Based on our experience, we believe
342 that coverage requirements for successful circularization of genomes from metagenomes are
343 significantly higher than for isolates.

344 Likely nearly all CPR have unlinked ribosomal operons

345 Typically in bacteria and archaea, the 16S, 23S, and 5S ribosomal RNA genes are found
346 in an operon in the order 16S-23S-5S [46] (Fig 3A). In contrast, we noted that in the CPR
347 genomes that we circularized, nearly 80% of them had unlinked 16S and 23S genes and
348 sometimes unlinked 23S and 5S genes (27/34 genomes). We observed the following types of
349 ribosomal operons in our circularized genomes: (1) operonic, but the 16S and 23S are
350 separated by tRNAs on the same strand (Fig 3B), (2) operonic, but the 16S and 23S (or 23S
351 and 5S) are separated by tRNAs and/or protein coding genes on the same strand (Fig 3C), (3)
352 unlinked by distance, all three ribosomal rRNA genes are on the same strand but the 16S is
353 separated from the 23S-5S or all three are separated by more than 2000bp and there are no
354 possible intervening genes in the spacer regions that could connect the ribosomal genes in an
355 operon (Fig 3D), and (4) unlinked because the 16S is on the opposite strand from the 23S and
356 5S (Fig 3E). In three cases, tRNA genes and/or protein coding genes on the same strand were
357 located between the 16S and 23S or between the 23S and 5S, but there are 300-500bp regions
358 between the genes, so in these cases the ribosomal genes may be uncoupled, but
359 conservatively we counted them as operonic. In-depth analysis of gene spacing in operons of

360 these genomes would be required to determine if these cases are operonic or not. In the
361 SRX1085364_bin_95 genome, we noted that there is a homing endonuclease between both the
362 16S and 23S, and the 23S and 5S. Given that our genomes span a large part of the CPR
363 phylogeny, we infer that most of the CPR likely have unlinked ribosomal operons.
364
365 **Fig 3. Diagram of placement types of ribosomal RNA genes.**
366 Number of genomes in this study for each category are indicated in the rightmost column. (A)
367 Operonic. The 16S (yellow), 23S (green), and 5S (purple) ribosomal RNA genes are in an
368 operon. (B) Operonic with tRNAs (blue). The three ribosomal RNA genes are still in an operon,
369 but one or more tRNAs are located in the spacer between the 16S and 23S genes. (C) Operonic
370 with tRNAs and protein coding genes (beige). The three ribosomal RNA genes are still in an
371 operon, but one or more tRNAs or protein coding genes are located in the spacer between the
372 16S and 23S genes or 23S and 5S genes. It is not unusual to find that the protein coding gene
373 is a homing endonuclease. (D) Unlinked ribosomal RNA genes by distance. The 16S gene is
374 unlinked from the 23S and 5S genes, or the 23S is also unlinked from the 5S gene, by enough
375 distance (>2000bp) and intervening genes on the opposite strand where it is not possible for
376 them to be transcribed from the same promoter. (E) Unlinked ribosomal RNA genes that are on
377 opposite strands. The 16S is on the opposite strand from the 23S and 5S genes.



The most common type of bacterial rRNA operons are those where 16S-23S-5S are transcribed together (Fig 3A), sometimes with tRNAs between the 16S and 23S (Fig 3B), so it is notable that most of the genomes in this study have unlinked rRNA operons. Instances where 16S and 23S are decoupled are unusual, although not unknown [46,47]. Separation of 23S and 5S is very unusual in bacteria but typical in archaea [47]. Decoupling between the 16S and 23S is known to occur especially in bacteria and archaea with reduced genomes (<2Mb) [46] such as *Mycoplasma gallisepticum* [48], *Borrelia burgdorferi* [49], *Ferroplasma acidarmanus* [50], as well as obligate symbionts with small genomes such as *Buchnera aphidicola* [51], *Wolbachia pipientis* [52], and *Nanoarchaeaum equitans* [47]. In a recent study of isolate genomes and pairing long reads with metagenomics data, others have also noted that a large percentage of the CPR likely have unlinked ribosomal operons based on analyzing the distance between the 16S and 23S genes [53]. However, to our knowledge, no one else has checked for tRNAs and protein coding genes comprising the operon in this type of analysis. We also do not know of

392 other studies of CPR that have documented possible separation of 23S and 5S genes, proteins
393 in the spacer regions between ribosomal RNA genes, and 16S and 23S on opposite strands.

394 Diverged forms of RNase P RNA in CPR

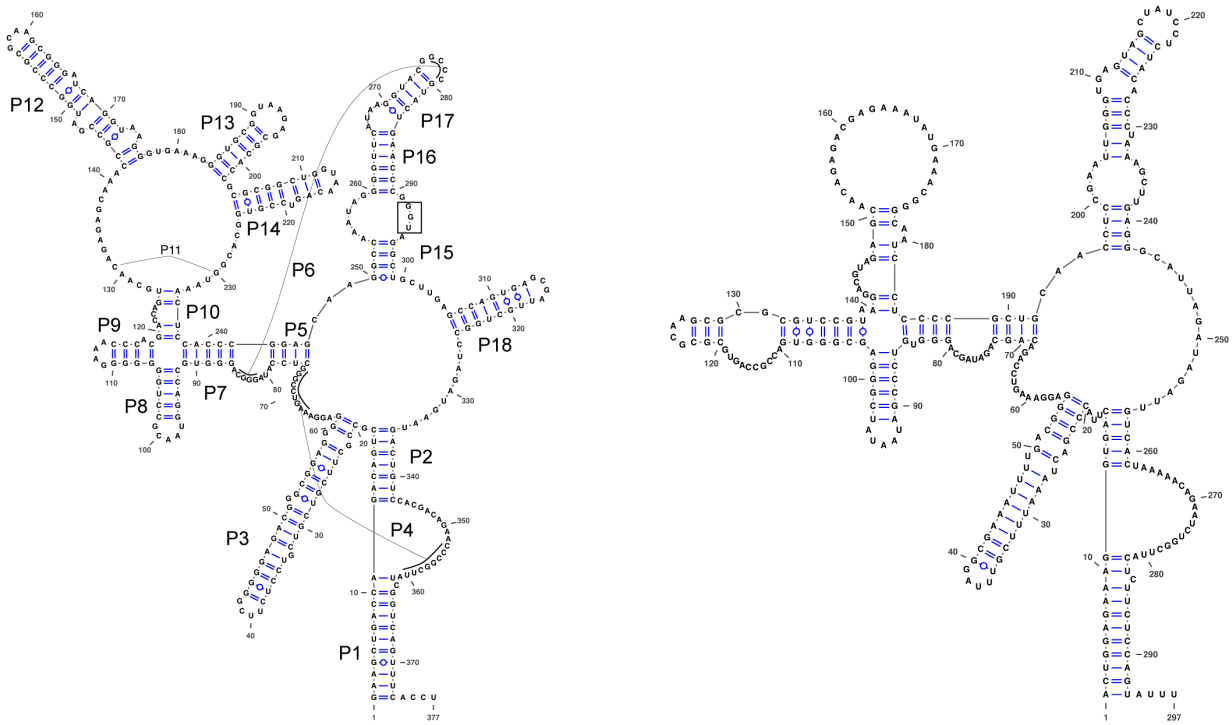
395 RNase P is an RNA-protein endonuclease involved in the maturation of tRNAs by
396 trimming the 5' leader of pre-tRNAs. The RNA component of this complex is considered
397 essential for all organisms except for species of the Aquificaceae family, which contain a protein
398 that does not require the RNA component for tRNA trimming [27], and *Nanoarchaeum equitans*,
399 an obligate symbiont that does not have any detectable RNase P RNA in its reduced genome,
400 nor any detectable RNase P activity [26].

401 Given the otherwise ubiquitous nature of RNase P RNA, we require detection of this
402 gene as a final quality check of assembled isolate genomes and circularized genomes.
403 However, in the set of circularized CPR genomes in this study, we found that a significant
404 number that lacked RNase P RNA (10/35). Absent a high degree of confidence that these are
405 indeed circular genomes, we would not have noticed this anomaly. We suspected that the
406 RNase P RNA gene was not being detected by the models because the genomes that lacked
407 the gene did not fall into a specific clade and the genes that were detected still had many
408 conserved features of RNase P RNA. To find the missing genes, we reduced the bitscore
409 threshold below the model noise cutoffs when running cmsearch from the Infernal software
410 package [54]. The noise cutoff is the score generally considered to be the score of the highest
411 scoring false positive for that model (Infernal User's Guide, https://infernal.janelia.org). After
412 reducing the thresholds, we were able to detect the missing RNase P RNAs.

413 Most of the RNase P RNA genes that we found, even the ones we found initially,
414 required major manual refolding because of the diverged structures with either large extensions
415 of helices (S1 Fig) or missing helices. Many are missing P13, P14, P16, and P17, which is not
416 unusual. However, the RNase P RNA from SRX3307784_bin_224 appears to be missing P12

417 (Fig 4B), which is highly unusual because this helix is one of the most conserved across the tree
418 of life [40], and it is only known to be missing in *Mycoplasma fermentans* [55] and members of
419 the archaeal family Thermoproteaceae [56]. The closely related genome in this study,
420 ERX2165959_bin_53, is also missing P12 (S2 Fig). Another unusual feature is that
421 approximately two-thirds of the RNase P RNA (23/35) are missing the UGG motif that binds to
422 the CCA in pre-tRNAs. This motif tends to be missing from cyanobacteria and chloroplasts,
423 which may not have the CCA in their pre-tRNAs [57]. Given that cyanobacteria are one of the
424 closer lineages to the CPR in the bacterial tree, the loss of the UGG motif may be related to
425 lineage. A final example of a diverged feature is that the RNase P RNA from
426 SRX1775579_bin_0 appears to be missing nearly the entire P15 helix (S3 Fig). This helix is
427 responsible for establishing binding to pre-tRNAs in bacteria and typically contains the UGG
428 motif, although it is missing from all known RNase P RNA in eukaryotes and some archaea.
429
430 **Fig 4. RNase P RNA can have diverged forms in CPR genomes.**
431 Structures were drawn using VARNA (Visualization Applet for RNA, http://varna.lri.fr/) [58]. (A)
432 Structure of RNase P RNA from *Escherichia coli K-12* substr. MG1655. Helices P1-P18 are
433 labeled. The "UGG" sequence in the P15 loop that binds to the 3' end of the pre-tRNAs is
434 highlighted by a box. (B) Putative Structure of RNase P RNA from SRX3307784_bin_224
435 genome. Note that the P12, P13, P14, and P18 helices are missing, as well as the UGG motif.
436 Although it is not uncommon for P13, P14, and P18 to be missing in various bacteria, it is
437 unusual that P12 is missing. To compare the two structures, large regions of the RNA had to be
438 refolded manually from the original cmsearch prediction.

A E. coli RNase P RNA B SRX3307784_bin_224



Finding these diverged forms of RNase P RNA would not have been possible without having confidence that we had a complete genome. Finding these diverged structures also illustrates that we may find diversity of genes in metagenomics data when we are no longer restricted by what we can culture in the laboratory.

Detection and Assembly of Megaphage Genomes

In the process of circularizing genomes, we circularized what we first thought were two novel isolates with small genomes (~0.5 Mb). However, since one of our standard checks is to run all circular sequences against a full set of Rfam models, these immediately stood out because the only RNAs detected were tRNAs and a tmRNA. Also, GTDB-tk was unable to assign a taxonomy. Cursory BLASTX searches of large regions of the genome yielded only distant hits. Based on this, we decided that they were likely megaplasmids, but have now concluded that they are megaphage based on a recent publication [59].

452 SRX3024509_bin_4 is an example of one of these putative megaphages. It is 536,059 nt
453 long and codes for 74 tRNA sequences along with one tmRNA. We have seen more than 10
454 similar - in terms of size and RNA content - megaphages in a variety of environments. They
455 appear to be quite common and if we expanded our size limits, we believe we would find many
456 more. Our method should in theory perform even better on plasmids and viruses than on normal
457 genomes since the former are less likely to have repeats. Extraction of plasmids and viruses
458 from the metagenomes is a matter for future work.

459 Discussion

460 We believe it is crucial to have a substantial collection - on the order of hundreds per
461 phylum - of genomes that approach traditional finished genome standards as closely as
462 possible, such as having a single circular, contiguous sequence with an error rate less than 1
463 per 100,000 bp [60]. Given that we have not yet succeeded in isolating many of the species
464 found in metagenomics datasets, our focus is on extracting their genomes from environmental
465 metagenomes and enrichments. By checking that assembled genomes are circular and possess
466 all standard known components of genomes - such as RNAs without which life as we know it
467 cannot exist - are present, we gain high confidence that we had nothing but the genome and
468 that the genome was not falsely circularized. We believe that circularity is a top criterion for a
469 high quality assembly, along with checks for misassemblies. We see a clear need for an
470 ongoing curation of collections of genomes. As more circularized genomes are generated from
471 metagenomics data, comparisons will help expose misassemblies and false circularization. For
472 metagenomics data, checking for misassemblies is crucial because they can produce chimeric
473 genomes and lead to erroneous conclusions and information in public genomic databases [46].
474 During the development of our circularization method, we learned some lessons about
475 when it is the most successful and instances where it will likely fail:

476 1. Our method works well for small genomes without repeats of significant length. Exact
477 repeats longer than the fragment length remain an issue. If the fragment length is less
478 than the length of a repeat, then it cannot resolve the repeat in the assembly. Once
479 repeats get above the fragment length, the process will - and should - fail.
480 2. We noted that genomes with rRNA copy numbers greater than one will almost always
481 fail to circularize. We are aware of a single case where MIRA assembled two copies of
482 an SSU on a single contig, so it is possible to do, but it is rare. Binners almost always fail
483 to correctly bin multiple copies of rRNA operons as they end up on shorter contigs with
484 coverage that is a multiple of the single copy stretches. Because we do "digital primer
485 walking", it is possible to extend a contig to cover a portion of an unbinned contig
486 containing the ribosomal RNA genes While our method will not result in automatic
487 circularization in this case, it can set the stage for further manual curation and possible
488 eventual circularization.
489 3. Circularization of genomes from metagenomes depends heavily on coverage. All of the
490 genomes we circularized had coverage greater than 29X (Table 2), but it may be
491 possible to circularize a genome with lower coverage. However, in these cases,
492 circularization will generally require manual intervention and we do not know how it
493 would be automated.
494 4. Circularizing genomes with high GC content is more difficult. This is not particularly
495 surprising given that all of this data was Illumina sequenced and there are known biases
496 against high GC content [61]. It is also possible that genomes with high GC content are
497 larger and hence more difficult to succeed with circularization.
498
499 Even for bins that did not achieve circularity, if the bin didn't become Chaos, we suggest
500 that it likely still improved and, in some cases, significantly in terms of the length of contigs and

501 number of misassemblies. MIRA is conservative in extending contigs, so in some cases we
502 believe that with some manual intervention these bins can be circularized.
503 We are well aware that circularity is not sufficient in and of itself to account for all of the
504 genetic material of a microbe, *i.e.*, multiple chromosomes, plasmids, etc. may comprise a
505 genome and not just one chromosome. As long read technology such as PacBio and Nanopore
506 become more feasible for metagenomics, assembling microbial genomes will also become
507 easier [18,23]. In addition, long read technologies are starting to yield methylation patterns that
508 can be used to associate multiple replicons with each other, so it may be possible to resolve if
509 there are multiple chromosomes and plasmids in a genome [62].
510 We have at least another 100 CPR genomes that are close to being circularized and we
511 believe will only require only minor curation to achieve full circularization. We will release these
512 genomes in future publications. By generating a set of high quality reference genomes, we
513 intend to use these as scaffolds for future assemblies and comparative genomics. Since these
514 genomes are complete, we can also begin to study the genomic context of genes, such as
515 synteny and operons. As we demonstrated with our investigation into RNase P RNA, the study
516 of non-coding RNAs in the intergenic regions of these genomes is also now more feasible than
517 ever. Finally, we also plan to link the 16S to the functional potential that we can see in the
518 genome, allowing us to glean more information from 16S studies in terms of metabolic
519 inference. As we continue to generate high quality complete genomes from metagenomics
520 data, we will be able to more accurately analyze the functional potential of microorganisms that
521 we cannot yet culture.
522

523 # Materials and Methods

524 ## Metagenomics Datasets

525 We used datasets from the NCBI Sequence Read Archive
526 (https://www.ncbi.nlm.nih.gov/sra). Accession numbers are listed in Table 1. In addition, we
527 obtained the SURF datasets from Lily Momper [35].

528 ## Read Processing Assembly

529 Metagenomic reads were preprocessed using BBtools version 38.60 to remove Illumina
530 adapters, perform quality filtering and trimming, and remove PhiX174 spike-ins. We are not
531 aware of any published papers documenting these tools. However, it is a standard tool suite
532 developed at the Department of Energy Joint Genome Institute (JGI) and it is documented at
533 https://jgi.doe.gov/data-and-tools/bbtools/. Processing was done in two passes. First bbduk.sh
534 ran with parameters *ktrim=r k=23 mink=11 hdist=1 ref=adapters.fa tbo tpe 2*. This was to
535 remove any remaining Illumina adapters given in adapters.fa (standard Illumina adapters). Then
536 bbduk.sh was run again with parameters *bf1 k=27 hdist=1 qtrim=rl trimq=17 cardinality=t*
537 *ref=phix174_Illumina.fa*. This was to perform quality filtering and trimming as well as remove
538 Illumina PhiX174 spike ins given in the file phix174_Illumina.fa.

539 ## Genome Assembly and Classification

540 Assembly was performed using SPAdes version 3.13.0 [36] with parameters *--meta -k*
541 *21,33,55,77,99,127*. Following assembly, from BWA version 0.7.17-r1188 [63], we used the
542 BWA-MEM algorithm with default parameters to map the reads to the set of contigs produced by
543 the assembly. We did this to obtain the BAM file required by MetaBAT 2 version 2.0 [37]. We
544 used MetaBAT 2 with parameters *--unbinned --minContig 1500 --maxEdges 500* to bin the
545 contigs. The iterative assemblies were performed using MIRA 5.0rc1 [7]. The parameters set
546 were *-NW:cac=warn, -CO:fnic=yes -AS:nop=6:sdlpo=no -KS:fenn=0.3*. We used Pilon version
547 1.23 [42] with default parameters to run final read coherence checks and clean up issues

548 created by the circularization. Taxonomic classification was generated using GTDB-Tk version
549 0.3.3 [38].
550 We carried out all of the work using standard Haswell architectures with 20 cores and
551 256 GB of main memory. SPAdes is generally memory limited and that is where the high point
552 of memory use occurred. Most of the iterative binning work is possible on a standard desktop or
553 even a laptop with 32 GB of memory as long as the coverage of the candidate genomes doesn't
554 exceed ~100X.

555 **Gene Annotation**

556 All of the RNA annotations were generated by Infernal 1.1.2 [54] using *cmsearch* with
557 parameters *--notextw --cut_tc*. We also used in-house scripts to handle RNA clan processing
558 [64]. We used RFAM version 14.1 for the models except when we used SSU-ALIGN (see
559 Phylogenetic Tree section below) which uses built-in custom models. For RNase P RNA, we
560 reduced the required bit score threshold 5 using the bacterial Class A model (RF00010) to find
561 the diverged forms. Gene calling was done using Prodigal version 2.6.3 [65]. We used *prodigal*
562 with parameters *-n -p single*.

563 **Phylogenetic Tree**

564 A tree was constructed from a structural alignment of the 16S genes generated by SSU-
565 ALIGN version 0.1.1 with default parameters [66,67]. Some 16S genes required manual folding
566 and adjustments to correct the structural alignment. We used IQ-TREE version 2.0-rc1 [45] to
567 generate the tree via the web server at Los Alamos National Laboratory.

568 **Circularization code**

569 The code and a tutorial of this method is available on Github
570 (https://github.com/lmlui/Jorg).

571 ## Data Availability

572 Assembled genomes from this study were deposited at NCBI under BioProject
573 accession numbers PRJNA416593, PRJEB22302, PRJNA640834, PRJNA640835,
574 PRJNA640837, PRJNA640839, PRJNA640841, PRJNA640844-PRJNA640846, PRJNA640848,
575 PRJNA640849, PRJNA640851-PRJNA640855, PRJNA640857, PRJNA640860-PRJNA640863,
576 PRJNA640871-PRJNA640876, PRJNA640878, PRJNA640925, PRJNA640926, PRJNA640929,
577 PRJNA641932, PRJNA641935, and PRJNA641936, and PRJNA641940.
578 .

579 ## Acknowledgements

580 This work was funded by ENIGMA, a Science Focus Area Program supported by the U.
581 S. Department of Energy, Office of Science, Office of Biological and Environmental Research,
582 Genomics Sciences Program: Foundational Science; this research also used resources of the
583 Joint Genome Institute (JGI) and the National Energy Research Scientific Computing Center
584 (NERSC), U.S. Department of Energy Office of Science User Facilities; all managed by
585 Lawrence Berkeley National Laboratory under Contract No. DE-AC02-05CH11231. The authors
586 would also like to thank members of the ENIGMA Science Focus Area as well as Heidi Smith,
587 Patricia Chan, and Norman Pace for helpful feedback and suggestions. We would also like to
588 thank Sean Jungbluth and Lily Momper for giving us access to their datasets.
589

590 ## References

591 1. Martiny AC. The ‘1% culturability paradigm’ needs to be carefully defined. ISME J.
592 2020;14: 10–11. doi:10.1038/s41396-019-0507-8

593 2. Myers Jr EW. A history of DNA sequence assembly. it - Inf Technol. 2016;58.
594 doi:10.1515/itit-2015-0047

595 3. Ayling M, Clark MD, Leggett RM. New approaches for metagenome assembly with short
596 reads. Brief Bioinform. 2020;21: 584–594. doi:10.1093/bib/bbz020

597 4. Vollmers J, Wiegand S, Kaster A-K. Comparing and Evaluating Metagenome Assembly
598 Tools from a Microbiologist's Perspective - Not Only Size Matters! Rodriguez-Valera F,
599 editor. PLoS One. 2017;12: e0169662. doi:10.1371/journal.pone.0169662

600 5. Myers EW, Sutton GG, Delcher AL, Dew IM, Fasulo DP, Flanigan MJ, et al. A whole-
601 genome assembly of Drosophila. Science (80-). 2000;287: 2196–2204.
602 doi:10.1126/science.287.5461.2196

603 6. Simpson JT, Durbin R. Efficient de novo assembly of large genomes using compressed
604 data structures. Genome Res. 2012;22: 549–556. doi:10.1101/gr.126953.111

605 7. Chevreux B, Pfisterer T, Drescher B, Driesel AJ, Müller WEG, Wetter T, et al. Using the
606 miraEST assembler for reliable and automated mRNA transcript assembly and SNP
607 detection in sequenced ESTs. Genome Res. 2004;14: 1147–1159.
608 doi:10.1101/gr.1917404

609 8. Staden R. A new computer method for the storage and manipulation of DNA gel reading
610 data. Nucleic Acids Res. 1980;8: 3673–3694. Available:
611 https://www.ncbi.nlm.nih.gov/pmc/articles/PMC324183/

612 9. Teeling H, Meyerdierks A, Bauer M, Amann R, Glöckner FO. Application of
613 tetranucleotide frequencies for the assignment of genomic fragments. Environ Microbiol.
614 2004;6: 938–947. doi:10.1111/j.1462-2920.2004.00624.x

615 10. Korem T, Zeevi D, Suez J, Weinberger A, Avnit-Sagi T, Pompan-Lotan M, et al. Growth
616 dynamics of gut microbiota in health and disease inferred from single metagenomic
617 samples. Science (80-). 2015;349: 1101–1106. doi:10.1126/science.aac4812

618 11. Tang K, Lu YY, Sun F. Background adjusted alignment-free dissimilarity measures
619 improve the detection of horizontal gene transfer. Front Microbiol. 2018;9.
620 doi:10.3389/fmicb.2018.00711

621 12. Parks DH, Rinke C, Chuvochina M, Chaumeil PA, Woodcroft BJ, Evans PN, et al.
622 Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the
623 tree of life. Nat Microbiol. 2017;2: 1533–1542. doi:10.1038/s41564-017-0012-7

624 13. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. CheckM: Assessing the
625 quality of microbial genomes recovered from isolates, single cells, and metagenomes.
626 Genome Res. 2015;25: 1043–1055. doi:10.1101/gr.186072.114

627 14. Chen LX, Anantharaman K, Shaiber A, Murat Eren A, Banfield JF. Accurate and
628 complete genomes from metagenomes. Genome Res. 2020;30: 315–333.
629 doi:10.1101/gr.258640.119

630 15. Kantor RS, Wrighton KC, Handley KM, Sharon I, Hug LA, Castelle CJ, et al. Small
631 genomes and sparse metabolisms of sediment-associated bacteria from four candidate
632 phyla. MBio. 2013;4. doi:10.1128/mBio.00708-13

633 16. Anantharaman K, Brown CT, Hug LA, Sharon I, Castelle CJ, Probst AJ, et al. Thousands
634 of microbial genomes shed light on interconnected biogeochemical processes in an
635 aquifer system. Nat Commun. 2016;7: 1–11. doi:10.1038/ncomms13219

636 17. Brown CT, Hug LA, Thomas BC, Sharon I, Castelle CJ, Singh A, et al. Unusual biology
637 across a group comprising more than 15% of domain Bacteria. Nature. 2015;523: 208–
638 211. doi:10.1038/nature14486

639 18. Kadnikov V V., Mardanov A V., Beletsky A V., Karnachuk O V., Ravin N V. Complete
640 genome of a member of a new bacterial lineage in the microgenomates group reveals an

641 unusual nucleotide composition disparity between two strands of dna and limited
642 metabolic potential. Microorganisms. 2020;8. doi:10.3390/microorganisms8030320

643 19. Albertsen M, Hugenholtz P, Skarshewski A, Nielsen KL, Tyson GW, Nielsen PH. Genome
644 sequences of rare, uncultured bacteria obtained by differential coverage binning of
645 multiple metagenomes. Nat Biotechnol. 2013;31: 533–538. doi:10.1038/nbt.2579

646 20. Starr EP, Shi S, Blazewicz SJ, Probst AJ, Herman DJ, Firestone MK, et al. Stable isotope
647 informed genome-resolved metagenomics reveals that Saccharibacteria utilize
648 microbially-processed plant-derived carbon. Microbiome. 2018;6: 122.
649 doi:10.1186/s40168-018-0499-z

650 21. Hernsdorf AW, Amano Y, Miyakawa K, Ise K, Suzuki Y, Anantharaman K, et al. Potential
651 for microbial H 2 and metal transformations associated with novel bacteria and archaea in
652 deep terrestrial subsurface sediments. ISME J. 2017;11: 1915–1929.
653 doi:10.1038/ismej.2017.39

654 22. Sieber CMK, Probst AJ, Sharrar A, Thomas BC, Hess M, Tringe SG, et al. Recovery of
655 genomes from metagenomes via a dereplication, aggregation and scoring strategy. Nat
656 Microbiol. 2018;3: 836–843. doi:10.1038/s41564-018-0171-1

657 23. Moss EL, Maghini DG, Bhatt AS. Complete, closed bacterial genomes from microbiomes
658 using nanopore sequencing. Nat Biotechnol. 2020;38: 701–707. doi:10.1038/s41587-
659 020-0422-6

660 24. Chivian D, Brodie EL, Alm EJ, Culley DE, Dehal PS, DeSantis TZ, et al. Environmental
661 genomics reveals a single-species ecosystem deep within earth. Science (80-).
662 2008;322: 275–278. doi:10.1126/science.1155495

663 25. Galperin MY. Linear chromosomes in bacteria: No straight edge advantage?

Environmental Microbiology. Wiley-Blackwell; 2007. pp. 1357–1362. doi:10.1111/j.1462-2920.2007.01328.x

26. Randau L, Schröder I, Söll D. Life without RNase P. Nature. 2008;453: 120–123. doi:10.1038/nature06833

27. Nickel AI, Wäber NB, Gößringer M, Lechner M, Linne U, Toth U, et al. Minimal and RNA-free RNase P in Aquifex aeolicus. Proc Natl Acad Sci U S A. 2017;114: 11121–11126. doi:10.1073/pnas.1707862114

28. Kodama Y, Shumway M, Leinonen R. The sequence read archive: explosive growth of sequencing data on behalf of the International Nucleotide Sequence Database Collaboration. Nucleic Acids Res. 2012;40. doi:10.1093/nar/gkr854

29. Wilhelm RC, Hanson BT, Chandra S, Madsen E. Community dynamics and functional characteristics of naphthalene-degrading populations in contaminated surface sediments and hypoxic/anoxic groundwater. Environ Microbiol. 2018;20: 3543–3559. doi:10.1111/1462-2920.14309

30. Nelson WC, Graham EB, Crump AR, Fansler SJ, Arntzen E V., Kennedy DW, et al. Distinct temporal diversity profiles for nitrogen cycling genes in a hyporheic microbiome. Liu W, editor. PLoS One. 2020;15: e0228165. doi:10.1371/journal.pone.0228165

31. Graham EB, Crump AR, Kennedy DW, Arntzen E, Fansler S, Purvine SO, et al. Multi ’omics comparison reveals metabolome biochemistry, not microbiome composition or gene expression, corresponds to elevated biogeochemical function in the hyporheic zone. Sci Total Environ. 2018;642: 742–753. doi:10.1016/j.scitotenv.2018.05.256

32. Graham EB, Tfaily MM, Crump AR, Goldman AE, Bramer LM, Arntzen E, et al. Carbon Inputs From Riparian Vegetation Limit Oxidation of Physically Bound Organic Carbon Via

687 Biochemical and Thermodynamic Processes. J Geophys Res Biogeosciences. 2017;122:
688 3188–3205. doi:10.1002/2017JG003967

689 33. Jewell TNM, Karaoz U, Brodie EL, Williams KH, Beller HR. Metatranscriptomic evidence
690 of pervasive and diverse chemolithoautotrophy relevant to C, S, N and Fe cycling in a
691 shallow alluvial aquifer. ISME J. 2016;10: 2106–2117. doi:10.1038/ismej.2016.25

692 34. Probst AJ, Ladd B, Jarett JK, Geller-Mcgrath DE, Sieber CMK, Emerson JB, et al.
693 Differential depth distribution of microbial function and putative symbionts through
694 sediment-hosted aquifers in the deep terrestrial subsurface. Nat Microbiol. 2018;3: 328–
695 336. doi:10.1038/s41564-017-0098-y

696 35. Momper L, Jungbluth SP, Lee MD, Amend JP. Energy and carbon metabolisms in a deep
697 terrestrial subsurface fluid microbial community. ISME J. 2017;11: 2319–2333.
698 doi:10.1038/ismej.2017.94

699 36. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes: A
700 new genome assembly algorithm and its applications to single-cell sequencing. J Comput
701 Biol. 2012;19: 455–477. doi:10.1089/cmb.2012.0021

702 37. Kang DD, Li F, Kirton E, Thomas A, Egan R, An H, et al. MetaBAT 2: An adaptive binning
703 algorithm for robust and efficient genome reconstruction from metagenome assemblies.
704 PeerJ. 2019;2019. doi:10.7717/peerj.7359

705 38. Chaumeil P-A, Mussig AJ, Hugenholtz P, Parks DH. GTDB-Tk: a toolkit to classify
706 genomes with the Genome Taxonomy Database. Bioinformatics. 2020;36: 1925–1927.
707 doi:10.1093/bioinformatics/btz848

708 39. Hahn C, Bachmann L, Chevreux B. Reconstructing mitochondrial genomes directly from
709 genomic next-generation sequencing reads-a baiting and iterative mapping approach.

710 Nucleic Acids Res. 2013;41. doi:10.1093/nar/gkt371

711 40. Bonfield JK, Whitwham A. Gap5—editing the billion fragment sequence assembly.
712 Bioinformatics. 2010;26: 1699. doi:10.1093/bioinformatics/btq268

713 41. Shapter FM, Waters DLE. Genome walking. In: Henry R, Furtado A, editors. Cereal
714 Genomics Methods in Molecular Biology (Methods and Protocols). Totowa, NJ: Humana
715 Press Inc.; 2014. pp. 133–146. doi:10.1007/978-1-62703-715-0_12

716 42. Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, et al. Pilon: An
717 Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly
718 Improvement. Wang J, editor. PLoS One. 2014;9: e112963.
719 doi:10.1371/journal.pone.0112963

720 43. Arkin AP, Cottingham RW, Henry CS, Harris NL, Stevens RL, Maslov S, et al. KBase:
721 The United States Department of Energy Systems Biology Knowledgebase. Nat
722 Biotechnol. 2018;36: 566–569. doi:10.1038/nbt.4163

723 44. Guindon S, Dufayard J-FF, Lefort V, Anisimova M, Hordijk W, Gascuel O. New algorithms
724 and methods to estimate maximum-likelihood phylogenies: Assessing the performance of
725 PhyML 3.0. Syst Biol. 2010;59: 307–321. doi:10.1093/sysbio/syq010

726 45. Trifinopoulos J, Nguyen L, von Haeseler A, Minh B. W-IQ-TREE: A Fast Online
727 Phylogenetic Tool for Maximum Likelihood Analysis - PubMed. Nucleic Acids Res.
728 2016;44: W232–W235. Available: https://pubmed.ncbi.nlm.nih.gov/27084950/

729 46. Andersson SG, Kurland CG. Genomic evolution drives the evolution of the translation
730 system. Biochem cell Biol. 1995;73: 775–787. doi:10.1139/o95-086

731 47. Londei P. Translational Mechanisms and Protein Synthesis. In: Garrett RA, Klenk H-P,

732 editors. Archaea. Malden, MA, USA: Blackwell Publishing Ltd; 2007. pp. 217–228.
733 doi:10.1002/9780470750865.ch19

734 48. Chen X, Finch LR. Novel arrangement of rRNA genes in Mycoplasma gallisepticum:
735 Separation of the 16S gene of one set from the 23S and 5S genes. J Bacteriol. 1989;171:
736 2876–2878. doi:10.1128/jb.171.5.2876-2878.1989

737 49. Davidson MM, Evans R, Ling CL, Wiseman AD, Joss AWL, Ho-Yen DO. Isolation of
738 Borrelia burgdorferi from ticks in the Highlands of Scotland. J Med Microbiol. 1999;48:
739 59–65. doi:10.1099/00222615-48-1-59

740 50. Allen EE, Tyson GW, Whitaker RJ, Detter JC, Richardson PM, Banfield JF. Genome
741 dynamics in a natural archaeal population. Proc Natl Acad Sci U S A. 2007;104: 1883–
742 1888. doi:10.1073/pnas.0604851104

743 51. Rouhbakhsh D, Baumann P. Characterization of a putative 23S-5S rRNA operon of
744 buchnera aphidicola (endosymbiont of aphids) unlinked to the 16S rRNA-encoding gene.
745 Gene. 1995;155: 107–112. doi:10.1016/0378-1119(94)00910-K

746 52. Bensaadi-Merchermek N, Salvado JC, Cagnon C, Karama S, Mouchès C.
747 Characterization of the unlinked 16S rDNA and 23S-5S rRNA operon of Wolbachia
748 pipientis, a prokaryotic parasite of insect gonads. Gene. 1995;165: 81–86.
749 doi:10.1016/0378-1119(95)00539-I

750 53. Brewer TE, Albertsen M, Edwards A, Kirkegaard RH, Rocha EPC, Fierer N. Unlinked
751 rRNA genes are widespread among bacteria and archaea. ISME J. 2020;14: 597–608.
752 doi:10.1038/s41396-019-0552-3

753 54. Nawrocki EP, Eddy SR. Infernal 1.1: 100-fold faster RNA homology searches.
754 Bioinformatics. 2013;29: 2933–2935. doi:10.1093/bioinformatics/btt509

755 55. Siegel RW, Banta AB, Haas ES, Brown JW, Pace NR. Mycoplasma fermentans simplifies
756 our view of the catalytic core of ribonuclease P RNA. RNA. 1996;2: 452.

757 56. Lai LB, Chan PP, Cozen AE, Bernick DL, Brown JW, Gopalan V, et al. Discovery of a
758 minimal form of RNase P in Pyrobaculum. Proc Natl Acad Sci U S A. 2010;107: 22493–8.
759 doi:10.1073/pnas.1013969107

760 57. Schön A. Ribonuclease P: the diversity of a ubiquitous RNA processing enzyme. FEMS
761 Microbiol Rev. 1999;23: 391–406. doi:10.1111/j.1574-6976.1999.tb00406.x

762 58. Darty K, Denise A, Ponty Y. VARNA: Interactive drawing and editing of the RNA
763 secondary structure. Bioinformatics. 2009;25: 1974–1975.
764 doi:10.1093/bioinformatics/btp250

765 59. Al-Shayeb B, Sachdeva R, Chen LX, Ward F, Munk P, Devoto A, et al. Clades of huge
766 phages from across Earth's ecosystems. Nature. 2020;578: 425–431.
767 doi:10.1038/s41586-020-2007-4

768 60. Chain PSG, Grafham D V., Fulton RS, Fitzgerald MG, Hostetler J, Muzny D, et al.
769 Genome project standards in a new era of sequencing. Science. NIH Public Access;
770 2009. pp. 236–237. doi:10.1126/science.1180614

771 61. Chen Y-C, Liu T, Yu C-H, Chiang T-Y, Hwang C-C. Effects of GC Bias in Next-
772 Generation-Sequencing Data on De Novo Genome Assembly. Xu Y, editor. PLoS One.
773 2013;8: e62856. doi:10.1371/journal.pone.0062856

774 62. Beaulaurier J, Zhu S, Deikus G, Mogno I, Zhang XS, Davis-Richardson A, et al.
775 Metagenomic binning and association of plasmids with bacterial host genomes using
776 DNA methylation. Nat Biotechnol. 2018;36: 61–69. doi:10.1038/nbt.4037

63. Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics. 2009;25: 1754–1760. doi:10.1093/bioinformatics/btp324

64. Gardner PP, Daub J, Tate J, Moore BL, Osuch IH, Griffiths-Jones S, et al. Rfam: Wikipedia, clans and the ``decimal'' release. Nucleic Acids Res. 2011;39: D141-5. doi:10.1093/nar/gkq1129

65. Hyatt D, Chen GL, LoCascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: Prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics. 2010;11: 119. doi:10.1186/1471-2105-11-119

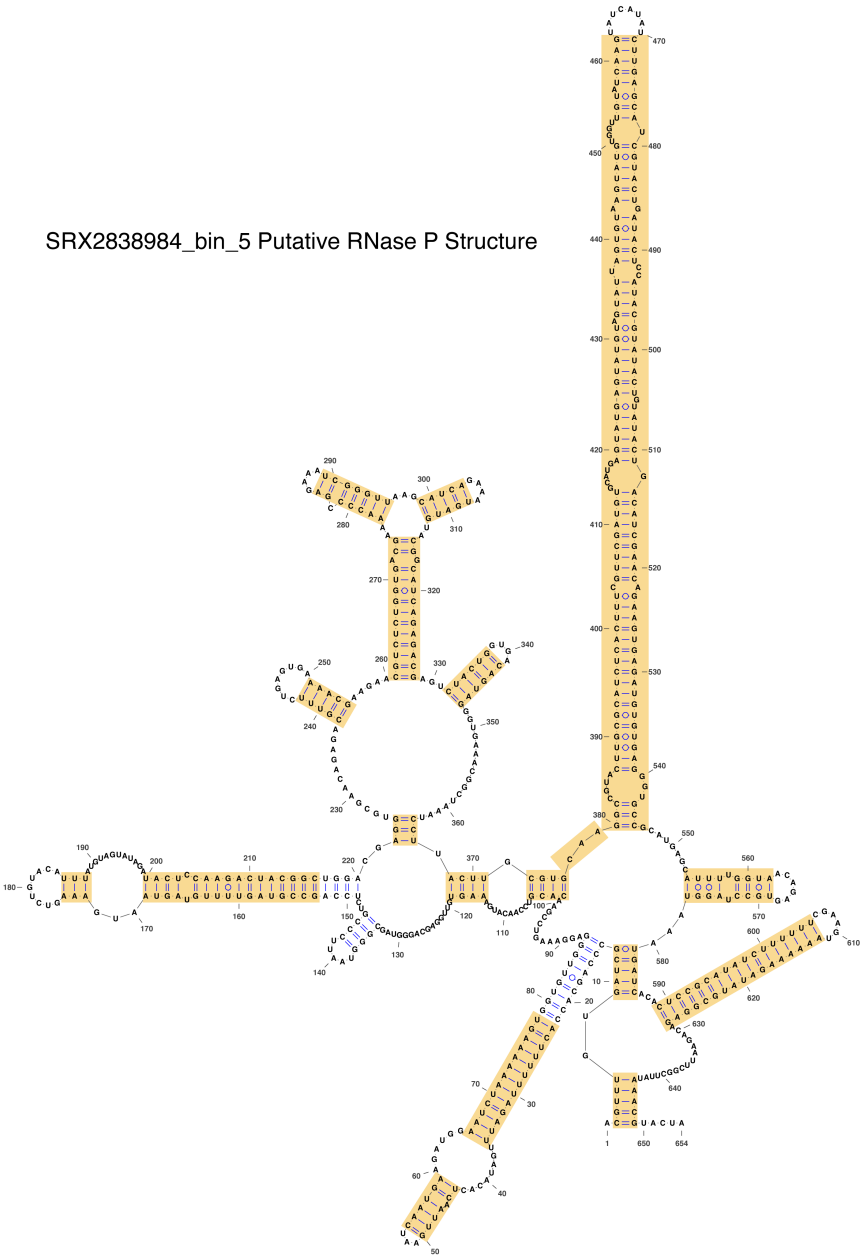
66. Cannone JJ, Gutell RR. RNA Structure Prediction (Comparative Sequence Analysis). Dictionary of Bioinformatics and Computational Biology. Chichester, UK: John Wiley & Sons, Ltd; 2004. doi:10.1002/9780471650126.dob0634.pub2

67. Nawrocki EP. Structural RNA Homology Search and Alignment using Covariance Models. Washington University in Saint Louis. 2009. doi:http://openscholarship.wustl.edu/etd/256

Supplementary figures

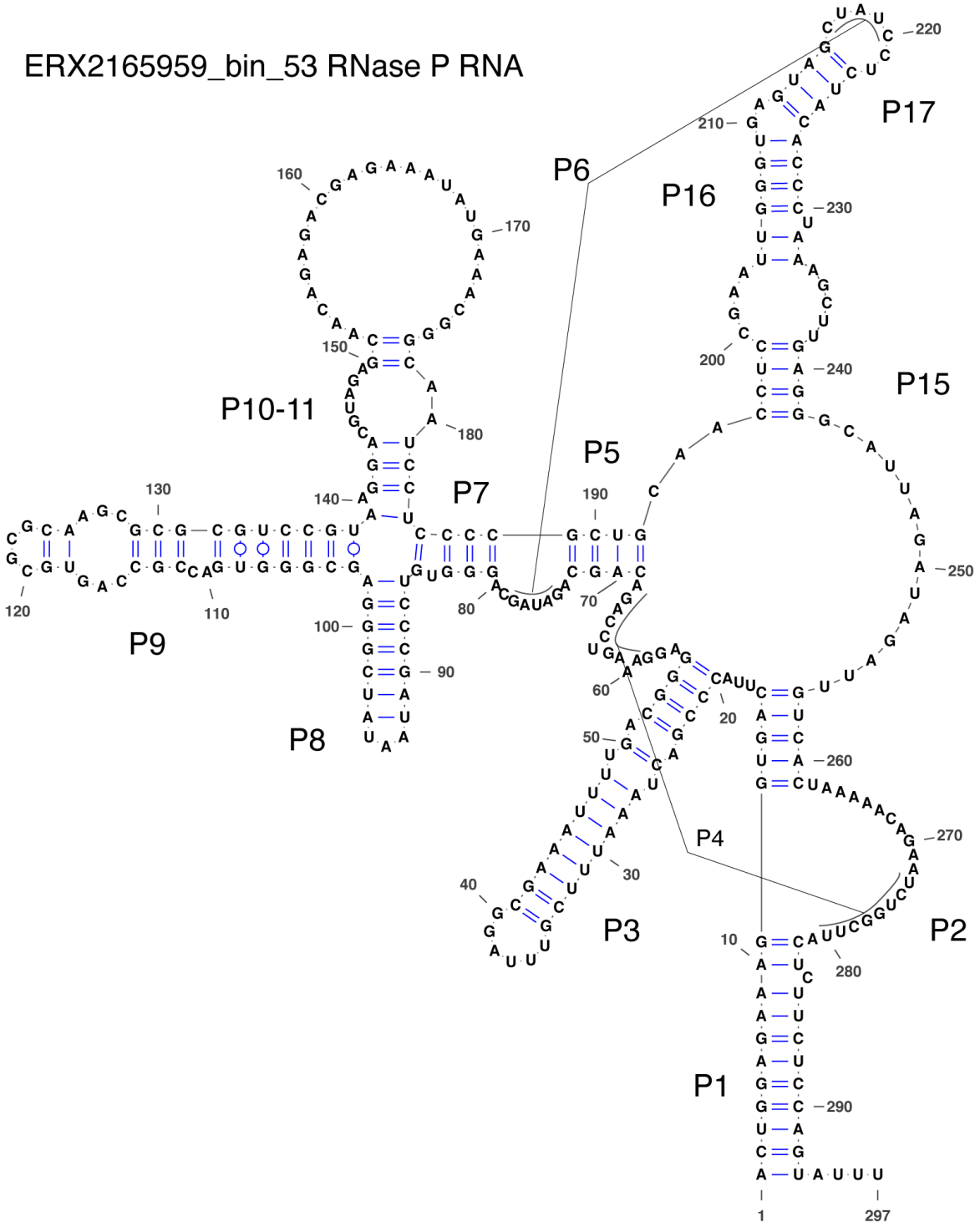
S1 Fig. Putative RNase P RNA structure of SRX2838984_bin_5.

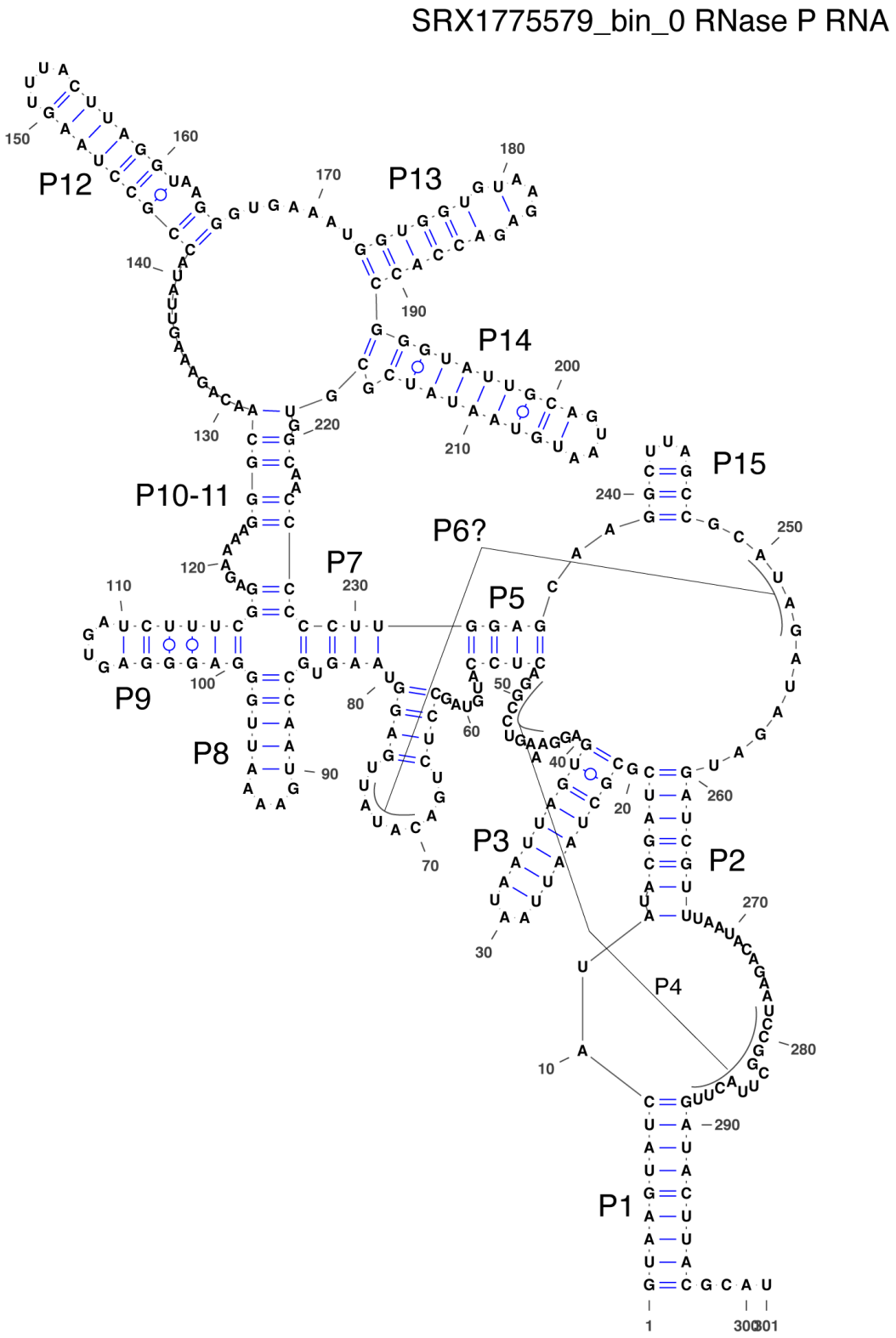
This RNase P RNA appears to have an extended P15 helix compared to typical RNase P RNA (see the *E. coli* RNase P RNA structure in Fig 4A of the main text). Yellow highlights indicate the portions of the RNA that had to be refolded manually. This amount of refolding was not unusual for the RNase P RNAs found in this study.



S2 Fig. Putative RNase P RNA structure of ERX2165959_bin_53.

This structure is missing P12, P13, P14, and P18. It is not unusual to be missing these helices, except for P12 which is found in nearly all RNase P RNA structures.



S3 Fig. Putative RNase P RNA structure of SRX1775579_bin_0.

This structure appears to be missing most of the P15 helix.