bioRxiv preprint doi: https://doi.org/10.1101/159798; this version posted July 14, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

riboSeed: leveraging prokaryotic genomic architecture to assemble across ribosomal regions

Nicholas R Waters,[1,2*] Florence Abram,[1] Ashleigh Holmes,[2] Fiona Brennan,[1,3] and Leighton Pritchard[2]

[1] *National University of Ireland, Galway*
[2] *The James Hutton Institute, Dundee, Scotland*
[3] *Teagasc, Johnstown Castle, Wexford*

*To whom correspondence should be addressed; E-mail: nickp60@gmail.com

Compiled: 2017/07/14 15:56:00

The vast majority of bacterial genome sequencing has been performed using Illumina short reads. Because of the inherent difficulty of resolving repeated regions with short reads alone, only ~10% of sequencing projects have resulted in a closed genome. The most common repeated regions are those coding for ribosomal operons (rDNAs), which occur in a bacterial genome between 1 and 15 times and are typically used as sequence markers to classify and identify bacteria. Here, we show that by utilizing the conserved nature of rDNAs across taxa and the uniqueness of their flanking regions, it is possible to improve assembly of these regions relative to *de novo* sequencing. We describe a method which constructs targeted pseudocontigs generated by iteratively assembling reads that map to a reference genome's rDNAs. These pseudocontigs are then used to more accurately assemble the newly-sequenced chromosome. We show that this method, implemented as riboSeed, correctly bridges across adjacent contigs in bacterial genome assembly and, when used in conjunction with other genome polishing tools, can result in closure of a genome.

Keywords: genome assembly, ribosome, benchmarking, scaffolding, *de fere novo*

Background

Sequencing bacterial genomes has become much more cost effective and convenient, but the number of complete, closed bacterial genomes remains a small fraction of the total number sequenced (Table 1). Even with the advent of new technologies for long-read sequencing and improvements to short read platforms, assemblies typically remain in draft status due to the computational bottleneck of genome closure [6, 32]. Although draft genomes are often of very high quality and suited for many types of analysis, researchers must choose between working with these draft genomes (and the inherent potential loss of data), or spending time and resources polishing the genome with some combination of

Table 1: NCBI Genome Assemblies of Bacteria

Date	Total	Complete	Chromosome	Scaffold	Contig
January 4th, 2017	85799	6255	1143	39972	38429
May 17th, 2017	96849	7212	1254	42839	43899

Source: https://www.ncbi.nlm.nih.gov/genome/browse/

bioRxiv preprint doi: https://doi.org/10.1101/159798; this version posted July 14, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

in silico tools, PCR, optical mapping, re-sequencing, or hybrid sequencing [32, 45]. Many *in silico* genome finishing tools are available, and we summarise several of these in Table 2.

Table 2: A sample of available *in silico* genome polishing tools

Tool	Method Summary
GapFiller [4]	iteratively utilizes paired-end reads to close contig junctions
GapCloser [25]	uses paired-end reads to close contig junctions
IMAGE [44]	iteratively uses local assemblies of reads belonging to assembly gaps
CloG [52]	uses trimmed *de novo* contigs in hybrid assembly followed by a stitching algorithm
FGap [15, 35]	uses BLAST to find potential gap closures from alternate assemblies, libraries or references.
GFinisher [15]	uses GC-skew to refine assemblies
GapFiller [31]	produces "long-reads" from paired-end sequencing data using a local assembler, which can then be used in a *de novo* assembly.
CONTIGuator [13]	uses contigs from a *de novo* assembly along with one or more reference sequences to generate a contig map and PCR primer sets to validate in the lab.
Konnector [46]	uses paired-end reads to make long reads to be used in a Bloom filter representation of a de Bruijn graph
MapRepeat [27]	uses a directed scaffolding method to fill in rDNA gaps, but limited to Ion Torrent reads, and affected by inversions between rDNAs [28]
GRabB [5]	selectively assembles tandem rDNAs and mitochondria

The Illumina entries in NCBI's Sequence Read Archive (SRA) [21] outnumber all other technologies combined by about an order of magnitude (Table S2). Draft assemblies from these datasets have systematic problems common to short read datasets, including gaps in the scaffolds due to the difficulty of resolving assemblies of repeated regions [43, 50]. By resolving repeated regions in assemblies, it may be possible to improve on existing assemblies, and therefore obtain additional sequence information from existing short read datasets in the SRA.

The most common repeated regions are those coding for ribosomal RNA operons (rDNAs), as ribosomes are essential for cell function. Sequencing of the 16S ribosomal region is widely used to identify bacteria and explore microbial community dynamics [7, 8, 49, 51], as the region is conserved within taxa, yet retains enough variability to act as a bacterial "fingerprint" to separate clades informatively. However, the 16S, 23S, and 5S ribosomal subunit coding regions are often present multiple times in a single prokaryotic genome, and commonly exhibit polymorphism [9, 24, 30, 47]. These long, inexactly repeated regions [2] are problematic for short-read genome assembly. As rDNAs are frequently used as a sequence marker for taxonomic classification, resolving their copy number and sequence diversity from short read collections where the assembled genome has collapsed several repeats into a single region could help improve reference databases, increasing the accuracy of community analysis. We present here an *in silico* method, riboSeed, that capitalizes on the genomic conservation of rDNA regions within a taxon to improve resolution of these difficult regions and provide a means to benefit from unexploited information in the SRA/ENA short read archives.

riboSeed is most similar in concept to GRabB, the method of Brankovics et al. [5] for assembling mitochondrial and rDNA regions in eukaryotes, as both use targeted assembly. However, GRabB does not make inferences about the number of rDNA clusters present in the genome, or take advantage of their genomic context. In riboSeed, genomic context is resolved by exploiting both the rDNA sequences and their flanking regions, harnessing unique characteristics of the broader rDNA region within a single genome to improve assembly.

bioRxiv preprint doi: https://doi.org/10.1101/159798; this version posted July 14, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

The riboSeed algorithm proceeds from two observations: (1) although repeated rRNA coding sequences within a single genome are nearly identical, their flanking regions (that is, the neighboring locations within the genome) are distinct in that genome, and (2) the genomic contexts of equivalent rDNA sequences are also conserved within a taxonomic grouping. riboSeed uses only reads that map to rDNA regions from a reference genome, and is not affected by chromosomal rearrangements that occur outside the flanking regions immediately adjacent to each rRNA.

Briefly, riboSeed uses rDNA regions from a closely-related organism's genome to generate rDNA cluster-specific "pseudocontigs" that are seeded into the raw short reads to generate a final assembly. We refer to this process in this work as *de fere novo* (meaning "starting from almost nothing") assembly.

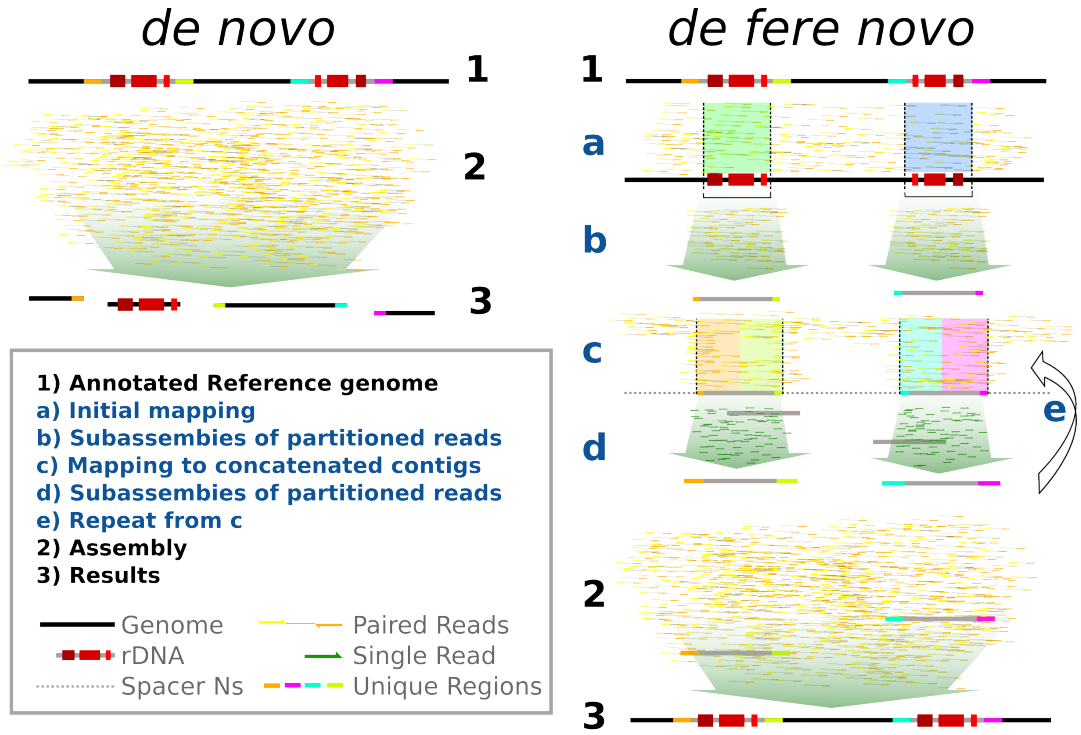


Figure 1: A comparison of *de novo* assembly to *de fere novo* assembly, as implemented in riboSeed. Reads are mapped to a reference genome, and those reads that align to rDNA and flanking regions are extracted. A subassembly for each group of reads that maps to an rDNA region is constructed to produce a "pseudocontig" for each region. These pseudocontigs are then concatenated together separated by 5kb of Ns as a spacer. Reads are then iteratively mapped to the concatenated pseudocontigs, extracted, and again subassembled to each region. After the final iteration, the pseudocontigs are included with raw reads in a standard *de novo* assembly. The subassemblies attempt to bridge proper rDNA regions by ensuring that flanking regions (represented here by colors) remain correctly paired. The *de novo* assembly collapses the rDNAs, but *de fere novo* places the rDNAs in the proper genomic context.

Implementation

We present riboSeed: a software suite that allows users to easily perform *de fere novo* assembly, given a reference genome sequence from a closely-related organism and single or paired-end short reads. The code is primarily written in Python3, with accessory shell and R scripts.

riboSeed relies on a closed reference genome assembly that is sufficiently closely-related to the isolate being assembled (which can be estimated using an alignment-free approach such as the KGCAK database [48]), in which rDNA regions are assembled and known to be in the correct context, as discussed below.

bioRxiv preprint doi: https://doi.org/10.1101/159798; this version posted July 14, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

riboSeed proceeds in three stages: preprocessing, *de fere novo* assembly, and assessment/visualization.

1. Preprocessing

riboScan.py

riboScan.py uses Barrnap (`https://github.com/tseemann/barrnap`) to annotate rRNAs in the reference genome, and EMBOSS's seqret [37] to create GenBank, FASTA, and GFF formatted versions of the reference genome. This preprocessing step unifies the annotation vocabulary for downstream processes.

riboSelect.py

riboSelect.py infers ribosomal operon structure from the genomic location of constituent 16S, 23S and 5S sequences. Jenks natural breaks algorithm is then employed to group rRNA annotations into likely operons on the basis of their genomic coordinates, using the number of 16S annotations to set the number of breaks. The output defines individual rDNA clusters and describes their component elements in a plain text file. This output can be manually adjusted before assembly if the clustering does not accurately reflect the known arrangement of operons based on visualization of the annotations in a genome browser.

2. *De Fere Novo* Assembly

riboSeed.py

riboSeed.py implements the algorithm described in Figure 2 in the current release. Short reads for the sequenced isolate are mapped to the reference genome using BWA [22]. Reads that map to each annotated rDNA and its flanking regions (default size 1kbp) are extracted into subsets (one subset per cluster). Each subset is independently assembled into a representative pseudocontig with SPAdes [3], using the reference rDNA regions as a trusted contig. The resulting pseudocontigs are evaluated for inclusion in future mapping/subassembly iterations based on their length (as discussed below), and concatenated into a pseudogenome, in which pseudocontigs are separated by 5kb of Ns as a spacer. This process is repeated in each subsequent iteration, using the previous round's pseudogenome as the reference.

After a specified number of iterations (3 by default), SPAdes is used to assemble all short reads in a hybrid assembly that includes the pseudocontigs from the final iteration as "trusted contigs" (or as "untrusted contigs" if the mapping quality of reads to that pseudocontig falls below a threshold, defined below). As a control, the short reads are also *de novo* assembled without the pseudocontigs.

Although this implementation of riboSeed uses SPAdes to perform both the subassemblies and the final *de fere novo* assembly, the pseudocontigs can be submitted to any hybrid assembler that accepts short read libraries and contigs. After assembly, the *de fere novo* and *de novo* assemblies are assessed with QUAST [16].

3. Assessment and Visualization

riboScore.py

riboScore.py extracts the regions flanking the rDNAs in the reference and in the assemblies generated by riboSeed. The flanking regions from the assembly are matched with the reference flanking regions using BLAST, and depending on the ordering of the matches, calls a junction a correct, incorrect, or ambiguous join based on the criteria outlined

bioRxiv preprint doi: https://doi.org/10.1101/159798; this version posted July 14, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

```
riboSeed (reference, riboSelect_clusters, reads, iters, flanking_width)
    ref = reference;
    clusters = parse riboSelect_clusters;
    region = clusters + flanking_width;
    for i in iters do
        map reads to ref;
        for cluster in clusters do
            filter and extract reads region;
            subassemble;
            return pseudocontig;
        end
        assess subassembly;
        if success then
            make pseudogenome from pseudocontigs ;
            ref = pseudogenome ;
        end
    end
    run assembler with reads and pseudocontigs;
end
```

Figure 2: Pseudocode of riboSeed algorithm

below.

riboSnag.py

riboSnag.py is provided as a helper tool to produce useful diagnostics and visualisation concerning rDNA sequence in the reference genome. Using the clustering generated by riboSelect.py, sequences for the clusters can be extracted from the genome, aligned, and Shannon entropy [40] plotted with consensus depth for each position in the alignment.

riboSwap.py

In all cases, we recommend assessing the performance of the riboSeed pipeline visually using Mauve [10,11], Gingr [42], or a similar genome assembly visualizer to compare reference, *de novo*, and *de fere novo* assemblies in addition to riboScore.py. If contigs appear to be incorrectly joined, the offending *de fere novo* contig can be replaced with syntenic contigs from the *de novo* assembly using the riboSwap.py script.

riboStack.py

riboStack.py uses bedtools [36] and samtools [22] to compare the depths of coverage of reads aligning to the reference genome in the rDNA regions to randomly sampled regions elsewhere in the reference genome. riboStack.py takes output from riboScan.py, and a BAM file of reads that map to the reference. If the number of riboScan.py-annotated rDNAs matches the number of rDNAs in the sequenced isolate, the coverage depths within the rDNAs will be similar to the coverage in other locations in the genome. If the coverage of rDNA regions sufficiently exceeds the average coverage elsewhere in the genome, this may indicate that the reference strain has fewer rDNAs than the sequenced isolate. In this case, using an alternative reference genome may produce improved results.

bioRxiv preprint doi: https://doi.org/10.1101/159798; this version posted July 14, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Results

Characteristics of rDNA flanking regions

The use of rDNA flanking sequences to uniquely identify and place rDNAs in their genomic context requires their flanking sequences to be distinct within the genome for each region. This is expected to be the case for most, if not all, prokaryotic genomes. We determined that using 1kb flanking widths was sufficient to include differentiating sequence (Figure S1). To demonstrate this, rDNA and 1kb flanking regions were extracted from *E. coli Sakai* [17] (BA000007.2), in which the rDNAs have been well characterized [33]. These regions were aligned with MAFFT [20], and their consensus depth and Shannon entropy [40] calculated for each position in the alignment (Figure 3a).

Figure 3a (and Figure S3) shows that within a single genome the regions flanking rDNAs are variable between operons. This enables unique placement of reads at the edges of rDNA coding sequences in their genomic context (i.e. there is not likely to be confusion between the placements of rDNA edges within a single genome).

In *E. coli MG1655,* the first rDNA is located 363 bases downstream of *gmhB* (locus tag b0200). Homologous rDNA regions were extracted from 25 randomly selected complete *E. coli* chromosomes (Table S1). We identified the 20kb region surrounding *gmhB* in each of these genomes, then annotated and extracted the corresponding rDNA and flanking sequences. These sequences were aligned with MAFFT, and the Shannon entropies and consensus depth plotted (Figure 3b).

Figure 3b shows that equivilent *E. coli* rDNAs, plus their flanking regions, are well-conserved across several related genomes. Assuming that individual rDNAs are monophyletic within a taxonomic group, short reads that can be uniquely placed on a related genome's rDNA as a reference template are also likely able to be uniquely-placed in the appropriate homologous rDNA of the genome to be assembled.

Taken together, when these two properties hold, this allow for unique placement of reads from homologous rDNA regions in the appropriate genomic context. These "anchor points" effectively reduce the number of branching possibilities in de Bruijn graph assembly for each individual rDNA, and thereby permit a complete balanced path through the full rDNA region.

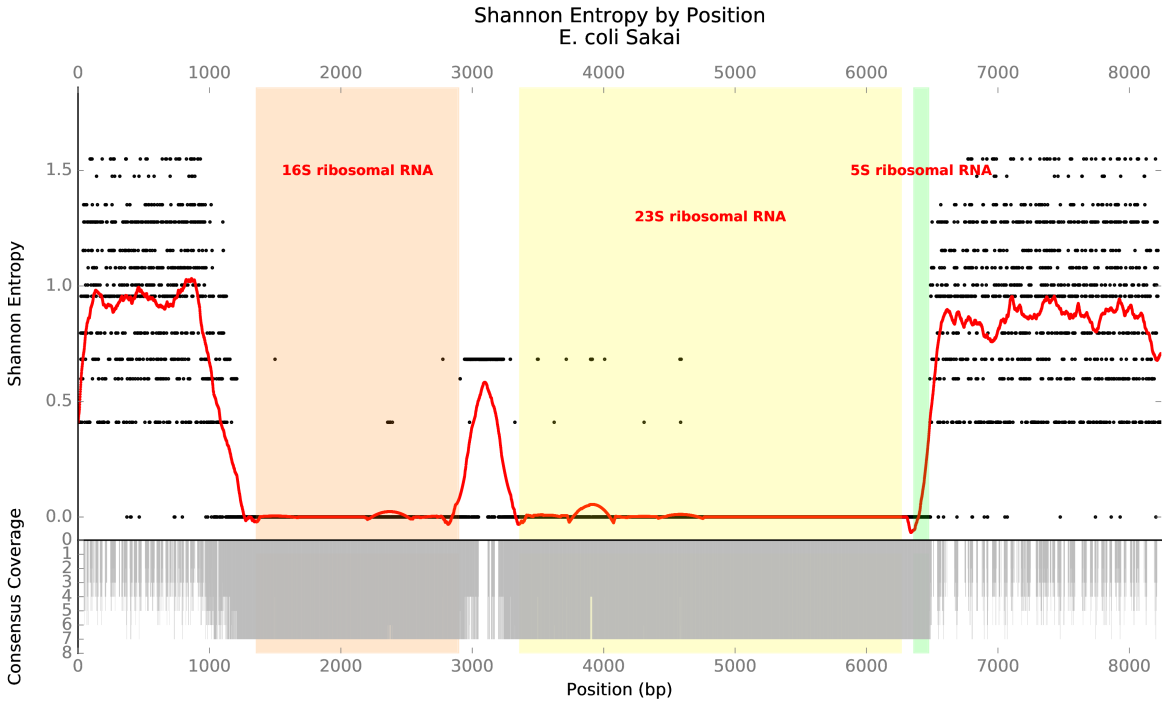
Validating Assembly across rDNA regions

Settings used for analyses in this manuscript are the defaults as of riboSeed version 0.4.09 (except where otherwise noted).

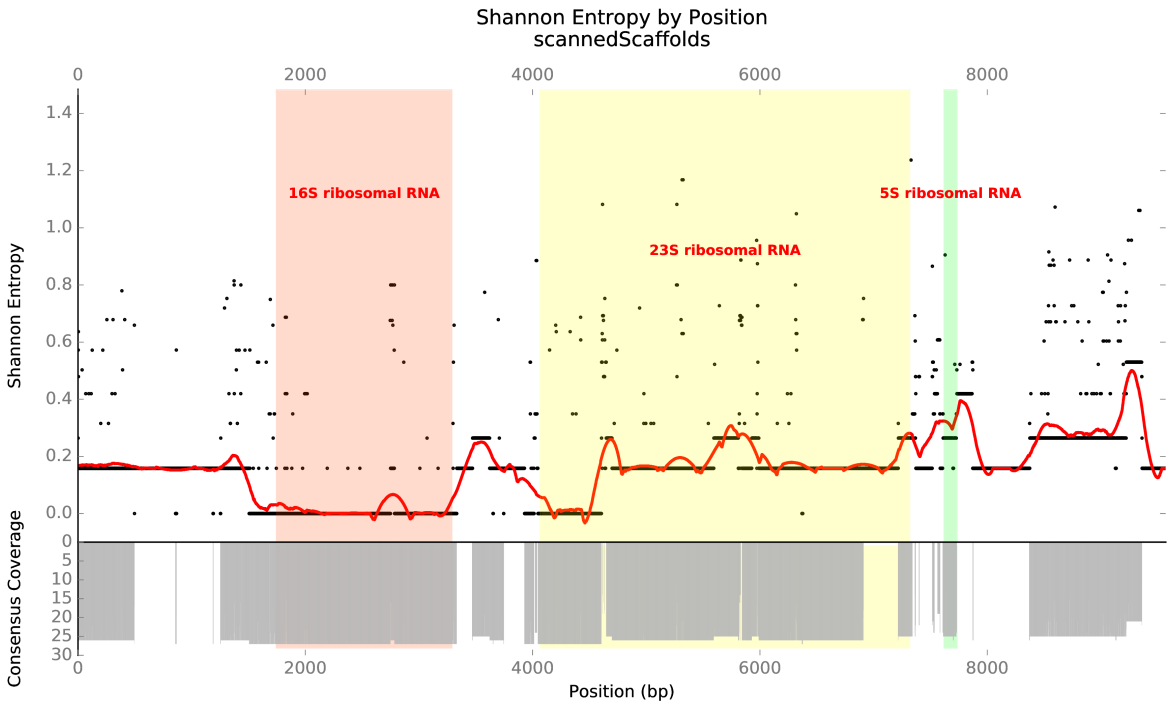
To evaluate the performance of *de fere novo* assembly compared to *de novo* assembly methods, we used Mauve to visualize syntenic regions and contig breaks of each assembly in relation to the reference genome that was used to generate pseudocontigs. We categorized each rDNA in an assembly as either a success, failure, or misassembly, as follows.

An rDNA assembly was classed as correct if two criteria were met: (i) the assembly merged two contigs across a rDNA

bioRxiv preprint doi: https://doi.org/10.1101/159798; this version posted July 14, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.



(a) rDNAs from *E. coli Sakai*



(b) Equivalent rDNAs from 25 *E. coli* genomes

Figure 3: Consensus coverage depth (grey bars) and Shannon entropy (black points, smoothed with a window size of 351bp as red line) for aligned rDNA regions. For the seven *E. coli Sakai* rDNA regions (a), entropy sharply increases moving away from the the 16S and 5S ends of the operon. In this case flanking regions would be expected to assemble uniquely withing a genome. By contrast, the rDNA regions occurring closest to homologous *gmbH* genes from 25 *E. coli* genomes (b) show greater conservation in their flanking regions. This indicates that flanking regions are more conserved for homologous rDNA than for paralogous rDNA operons, and implies that related genomes can be useful reference templates for assembling across these regions. Similar plots for each of the GAGE-B genomes used later for benchmarking can be found in Figure S3.

region such that, based on the reference, the flanking regions of the *de fere novo* assembly were syntenous with those of the reference; and (ii) the assembled contig extends at least 90% of the flanking length. An assembled cluster was

bioRxiv preprint doi: https://doi.org/10.1101/159798; this version posted July 14, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

defined as skipped if the ends of one or more contigs aligned within the rDNA or flanking regions (signalling that extension across the rDNA region was not achieved). Finally, if two contigs assembled across a rDNA region in a manner that conflicted with the orientation indicated in the reference genome, the rDNA region was deemed to be incorrect.

In all cases, SPAdes was used with the same parameters for both *de fere novo* assembly and *de novo* assembly, apart from the addition of pseudocontigs in the *de fere novo* assembly.

Simulated Reads with Artificial Genome

To create a small dataset for testing, we extracted all 7 distinct rDNA regions from the *E. coli Sakai* genome (BA000007.2), including 5kb upstream and downstream flanking sequence, using the tools riboScan.py, riboSelect.py and riboSnag.py. Those regions were concatenated to produce a ~100kb artificial test chromosome (see supplementary methods). pIRS [19] was used to generate simulated reads (100bp, 300bp inserts, stdev 10, 30-fold coverage, built-in error profile) from this test chromosome. These reads were assembled using riboSeed, using the *E. coli MG1655* genome (NC_000913.3) as a reference. The simulation was run 8 times.

The *de fere novo* assembly bridged 4 of the 7 rDNA regions in the artificial genome, while the *de novo* assembly method failed to bridge any (Table S2). To explain how the choice of reference sequence determines the ability to assemble correctly through rDNA regions, we ran riboSeed with the same *E. coli* reads using pseudocontigs derived from the *Klebsiella pneumoniae HS11286* (CP003200.1) reference genome [23]. The *de fere novo* assembly with pseudocontigs from *K. pneumoniae* bridged between 1 and 2 rDNAs, but also misassembled several rDNA gaps (Figure 4).

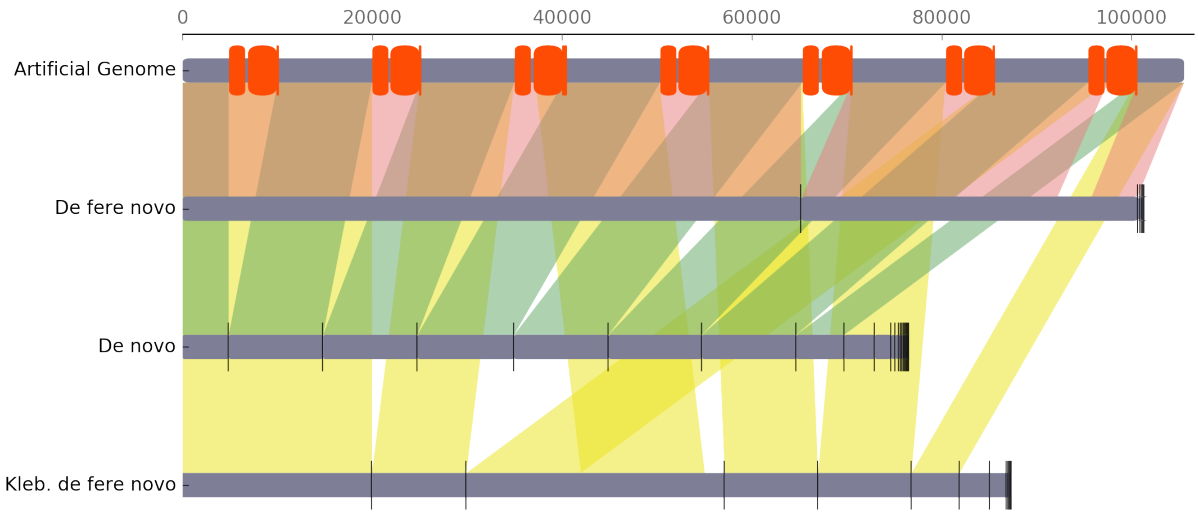


Figure 4: Representative Mauve output from a single similuation describing the results of riboSeed assemblies of simulated reads generated by pIRS from the concatenated *E. coli Sakai* artificial genome. From top to bottom: artificial reference chromosome; rRNA annotations (red bars); *de fere novo* assembly (*E. coli reference), de novo* assembly (*E. coli reference),* and *de fere novo* assembly (*K. pneumoniae* reference). riboSeed with *E. coli reference* assembles 4 of 7 rDNA regions, but the *de novo* assembly recovers no rDNA regions correctly. riboSeed using a *K. pneumoniae* reference resolves a single rDNA region, but misassembles clusters.

Effect of reference sequence identity on riboSeed performance

To investigate how riboSeed assembly is affected by choice of reference strain, we implemented a simple mutation model to generate reference sequence variants of the artificial chromosome described above, with a specified rate of

bioRxiv preprint doi: https://doi.org/10.1101/159798; this version posted July 14, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

substitution. A simple substitution rate was applied across all bases uniformly does not address the disparity of conservation between rDNAs and their flanking region, but a second model was also applied wherein substitutions were allowed only to the rDNA flanking regions. We then assembled the artificial genome's reads using the mutated artificial genome as a referece, using these models (Figure 5).

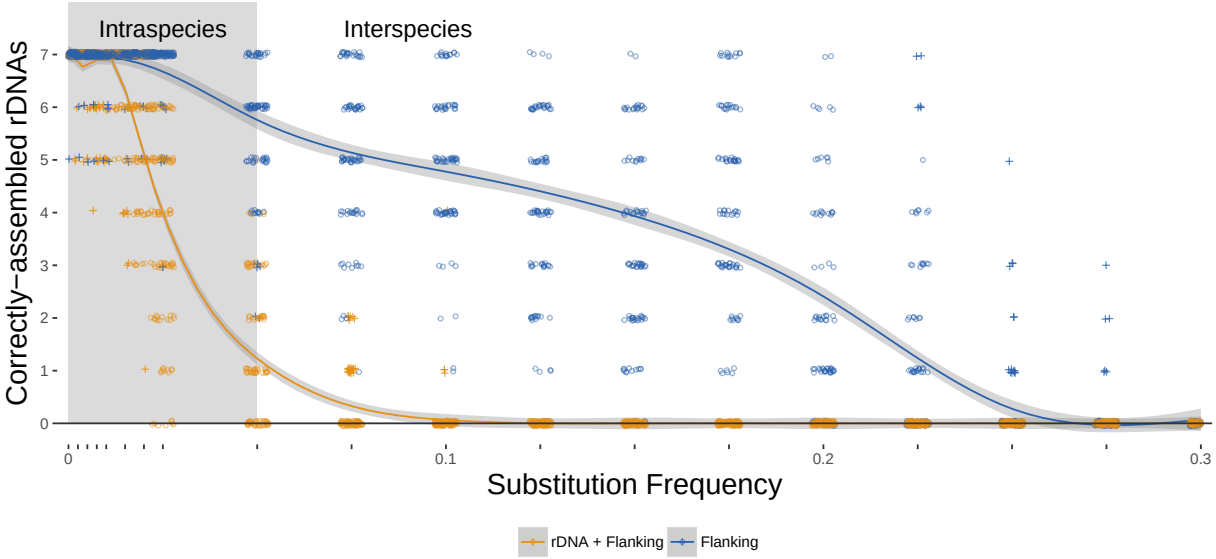


Figure 5: Variants of the artificial genome with substitutions frequencies between 0 and 0.3 (i.e. 300 substitutions per kbp). Correctly-assembled rDNAs were counted, and the distribution of results shown against the appropriate substitution frequency. Results are shown for models where substitutions are permitted throughout the chromosome (orange), and only in the flanking regions (blue), the latter approximating the relative rate of substitution in rDNA and flanking regions. Grey area corresponds to substitution frequencies resulting in average sequence identity over 95%. Outliers are shown as +'s. N=100.

To obtain an estimate of substitution rate for the *E. coli* data used above, Parsnp [42] and Gingr [42] were used to identify SNPs in the 25 genomes used above (Figure 3), with respect to the same region in *E. coli Sakai*. An average substitution rate of 0.0062 was observed. Compared to the results from the simulated genomes, we could expect successful performance under the model of mutated flanking regions, and partial success under the model of substitutions throughout the region.

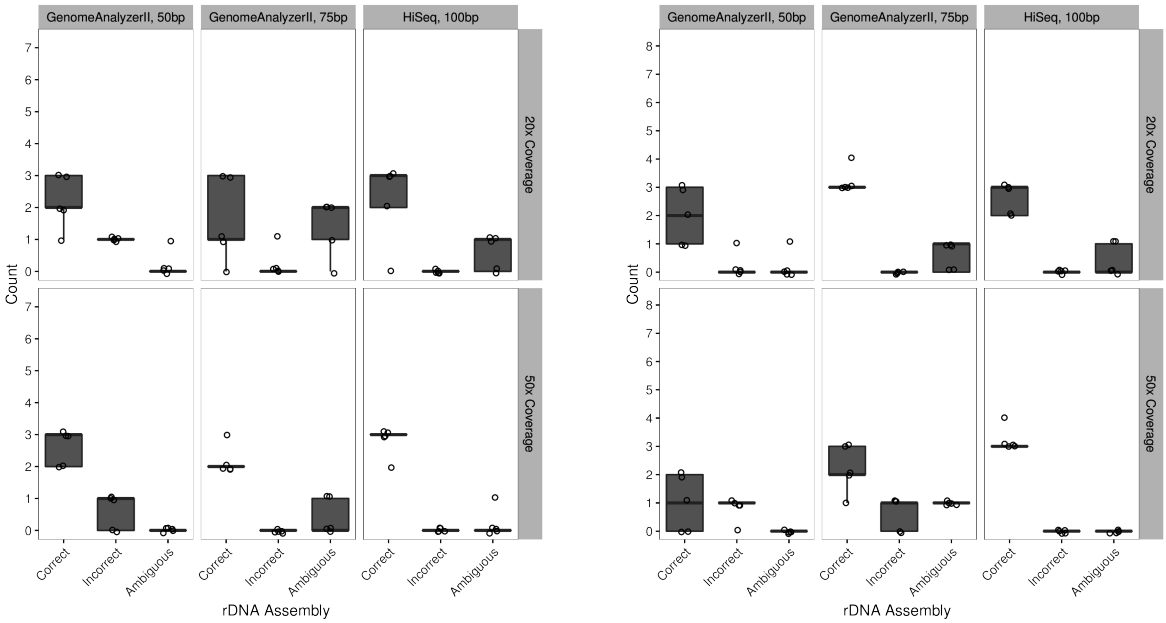
Figure 5 indicates that the more similar the reference sequence is to the genome being assembled, the greater the likelihood of correctly assembling through rDNA regions. When mutating only the flanking regions (Figure 5), which more closely resembles the relative sustitution frequencies of the rDNA regions, the procedure correctly assembles rDNAs with tolerance to substitution frequencies up to approximately 30 substitutions per kbp. Using a average nucleotide identity species boundary of 95% [14], it could be concluded that riboSeed requires a reference within the same species for optimal performance, and that moderate success can be achieved even when using a more distant reference.

Simulated reads with *E. coli Sakai* and *K. pneumoniae* Genomes

To investigate the effect of short read length on riboSeed assembly, pIRS [19] was used to generate paired-end reads from the complete *E. coli MG1655* and *K. pneumoniae NTUH-K2044* genomes, simulating datasets at a range of read lengths most appropriate to the sequencing technology. In all cases, 300bp inserts with 10bp standard deviation

bioRxiv preprint doi: https://doi.org/10.1101/159798; this version posted July 14, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

and the built-in error profile were used. Coverage was simulated at 20x to emulate low coverage runs and at 50x to emulate coverage close to the optimized values determined by Miyamoto [29] and Desai [12]. *De fere novo* assembly was performed with riboSeed using *E. coli Sakai* and *K. pneumoniae HS11286* as references, respectively, and the results were scored with riboScore.py (Figure 6).



(a) *E. coli* **(b)** *K. pneumoniae*

Figure 6: Comparison of *de fere novo* assemblies of simulated reads generated by pIRS. In most cases, increasing coverage depth and read length resulted in fewer misassemblies. Assemblies were scored using riboScore.py. N=5.

At either 20x or 50x coverage, *de novo* assembly was unable to resolve any rDNAs with any of the simulated read sets. *de fere novo* assembly with riboSeed showed modest improvement to both the *E. coli* and *K. pneumoniae* assemblies. Increasing depth of coverage and read length improves rDNA assemblies.

Benchmarking against Hybrid Sequencing and Assembly

To establish whether riboSeed performs as well with short reads obtained by sequencing a complete prokaryotic chromosome as with simulated reads, we attempted to assemble short reads from a published hybrid Illumina/PacBio sequencing project. The hybrid assembly using long reads was able to resolve rDNAs directly, and provides a benchmark against which to assess riboSeed performance in terms of: (i) bridging sequence correctly across rDNAs, and (ii) assembling rDNA sequence accurately within each cluster.

Sanjar, et al. published the genome sequence of *Pseudomonas aeruginosa BAMCPA07-48* (CP015377.1) [38], assembled from two libraries: ca. 270bp fragmented genomic DNA with 100bp paired-end reads sequenced on an Illumina HiSeq 4000 (SRR3500543), and long reads from PacBio RS II. The authors obtained a closed genome sequence by hybrid assembly. We ran the riboSeed pipeline on only the HiSeq dataset in order to compare *de fere novo* assembly to the hybrid assembly and *de novo* assembly of the same reads, using the related genome *P. aeruginosa ATCC 15692* (NZ_CP017149.1) as a reference.

de fere novo assembly correctly assembled across all 4 rDNA regions, whereas *de novo* assembly failed to assemble

bioRxiv preprint doi: https://doi.org/10.1101/159798; this version posted July 14, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

any rDNA regions (Table 3).

Comparing the *BAMCPA07-48* reference to the *de fere novo* assembly, we found a total of 9 SNPs in the rDNA flanking regions (Table 4). The same regions from the *ATCC 15692* reference used in the *de fere novo* assembly showed 103 SNPs compared to the *BAMCPA07-48*. This demonstrates that this subassembly scheme successfuly recovers the correct sequence dispite a large number of differences between the reference and the sequenced isolate.

Thus, we find that the *de fere novo* assembly using short reads performs better than *de novo* assembly using short reads alone. Comparison of the *de fere novo* assembly to the hybrid assembly allows assessment of *de fere novo* accuracy, and reveals that *de fere novo* can recover the rDNA sequence correctly placed in their genomic context, and with a low error rate.

Table 3: Assembly of Hybrid-Sequenced *P. aeruginosa BAMCPA07-48*

Sequenced Strain Name	Platform	Length	Depth	Reference Strain		*de novo*			*de fere novo*		
				Name	rDNAs	✓	–	×	✓	–	×
P. aeruginosa BAMCPA07-48	HiSeq	100	200x	ATCC 15692	4	1	3	0	4	0	0

✓ correct assembly
– skipped assembly
× incorrect assembly

Table 4: rDNA region SNPs between *de fere novo* assembly and hybrid assembly of *P. aeruginosa BAMCPA07-48* in rDNA regions, including 1kb upstream and downstream of the rDNA

rDNA region (5S,16S,23S with flanking region)	CP015377.1 Location	
398001–405418	402331	T → C
	402332	C → T
	404332	C → T
	404380	G → T
1039539–1045687	—	—
1862045–1869194	1864462	A → G
	1868402	A → C
	1868426	A → T
2809154–2816303	2811180	G → A
	2813886	T → C

Case Study: Closing the assembly of *S. aureus UAMS-1*

Staphylococcus aureus UAMS-1 is a well-characterized, USA200, methicillin-sensitive strain isolated from an osteomyelitis patient. The corresponding published genome was sequenced using Illumina MiSeq generating 300bp reads, and the assembly refined with GapFiller as part of the BugBuilder pipeline [1]. Currently, the genome assembly is represented by two scaffolds (JTJK00000000), with several repeated regions acknowledged in the annotations [39]. As the rDNA regions were not fully characterized in the annotations, we proposed that *de fere novo* assembly might resolve some of the problematic regions.

Using the same reference *S. aureus MRSA252* [18] (BX571856.1) with riboSeed as was used in the original assembly, *de fere novo* assembly correctly bridged gaps corresponding to two of the five rDNAs in the reference genome (Table 5). Furthermore, *de fere novo* assembly bridged two contigs that were syntenic with the ends of the scaffolds in the published assembly, indicating that the regions resolved by riboSeed could allow closure of the genome.

bioRxiv preprint doi: https://doi.org/10.1101/159798; this version posted July 14, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Table 5: Assembly of the *S .aureus UAMS-1*

Sequenced Strain Name	Platform	Length	Depth	Reference Strain		*de novo*			*de fere novo*		
				Name	rDNAs	✓	–	×	✓	–	×
S.aureus UAMS-1	MiSeq	300	110x	MRSA252	5	0	5	0	2	3	0

✓ correct assembly
– skipped assembly
× incorrect assembly

We modified the BugBuilder pipeline (`https://github.com/nickp60/BugBuilder`) used in the published assembly to incorporate pseudocontigs from riboSeed, resulting in a single scaffold of 7 contigs. In this case, riboSeed was able assist in bringing an existing high-quality scaffold to completion.

Benchmarking against GAGE-B Datasets

We used the Genome Assembly Gold-standard Evaluation for Bacteria (GAGE-B) datasets [26] to assess the performance of riboSeed against a set of well-characterized assemblies. These datasets represent a broad range of challenges; low GC content and tandem rDNA repeats prove challenging to the riboSeed procedure. *Mycobacterium abscessus* has only a single rDNA operon and does not suffer from the issue of rDNA repeats, so it was excluded from this analysis.

When the reference used in the GAGE-B study came from the sequenced strain (as was the case for *R. sphaeroides* and the *B. cereus*), we chose an alternate reference, as using the true reference sequence would provide an unfair advantage to riboSeed. The GAGE-B datasets include both raw and trimmed reads; in all cases, the trimmed reads were used. Results are shown in Table 6.

Compared to the *de novo* assembly, *de fere novo* assembly improved the majority of assemblies. In the case of the *S. aureus* and *R. sphaeroides* datasets, particular difficulty was encountered for all of the references tested. In the case of *B. fragilis*, the entropy plot (Figure S3g) shows that the variability on the 5' end of the operon is much lower than the other strains, likely leading to the misassemblies.

Table 6: Assembly of GAGE-B datasets

Sequenced Strain Name	Platform	Length	Depth	Reference Strain		*de novo*			*de fere novo*		
				Name	rDNAs	✓	–	×	✓	–	×
A. hydrophila SSU	HiSeq	101	250	ATCC 7966	10	0	10	0	4	6	0
B. cereus VD118	HiSeq	101	300	ATCC 10987	12	0	12	0	1	11	0
B. cereus ATCC 10987	MiSeq	250	100	NC7401	14	0	14	0	12	2	0
B. fragilis HMW 615	HiSeq	101	250	638R	6	0	5	1	0	3	3
R. sphaeroides 2.4.1	HiSeq	101	210	ATCC 17029	4	0	4	0	1	3	0
R. sphaeroides 2.4.1	MiSeq	251	100	ATCC 17029	4	1	2	1	1	2	1
S. aureus M0927	HiSeq	101	250	USA300_TCH1516	5	0	5	0	3	2	0
V. cholerae CO 0132(5)	HiSeq	100	110	El Tor str. N16961	8	0	8	0	5	3	0
V. cholerae CO 0132(5)	MiSeq	250	100	El Tor str. N16961	8	0	8	0	4	4	0
X. axonopodis pv. Manihotis UA323	HiSeq	101	250	pv. Citrumelo	2	0	1	1	2	0	0

✓ correct assembly
– skipped assembly
× incorrect assembly

bioRxiv preprint doi: https://doi.org/10.1101/159798; this version posted July 14, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Discussion

We show that the regions flanking equivilent rDNAs from related strains show a high degree of conservation in related organisms. This allows us to infer the location of rDNAs within a newly sequenced isolate, even in absence of the resolution that would be provided by long read sequencing. Comparing the regions flanking rDNAs within a single genome, we observed that when considering sufficiently large flanking regions, flanking sequences show enough variability to differentiate each instance of the rDNAs. Taken together, the cross-taxon homology allows inference of the location (i.e. the flanking regions) of rDNAs, and the variability of these flanking regions within a genome enables unique identification of reads likely belonging to each cluster.

The extent of sequence similarity between the sequenced isolate and the reference influences the resulting *de fere novo* assembly. To prevent spurious joining of contigs, if less than 80% of the reads map to the reference, the resulting pseudocontigs will be treated as "untrusted" contigs by SPAdes. Figure 5 shows that although one should use the closest complete reference available for optimal results, the subassembly method is robust against moderate discrepancies between the reference and sequenced isolate's flanking regions.

The method of constructing pseudocontigs implemented by riboSeed relies on having a relevant reference sequence, where the rDNA regions to act as "bait", fishing for reads that likely map specifically to that region. Although we show this to be an effective way to partition the appropriate reads, perhaps a more robust and supervision-free method would be use a probabilistic representation of equivalent rDNA regions for a particular taxon. By developing a database of sequence profiles (e.g. hidden Markov Models) from each of the rDNAs in a taxon, perhaps the step of choosing a single most appropriate reference could be circumvented. For datasets where the choice of reference determines riboSeed's effectiveness, a probabilistic approach may improve performance.

Several checks are implemented after the subassembly to ensure that the resulting pseudocontig is fit for inclusion in the next mapping/subassembly iteration or the final *de fere novo* assembly. If a subassembly's longest contig is greater than 3x the particular pseudocontig length or shorter than 6kb (a conservative minimum length of a 16S, 23S, and 5S operon), this is taken to be a sign of poor parameter choice so the user is warned, and by default no further seedings will occur to avoid spurious assembly. Such an outcome can be indicative of any of several factors: improper clustering of operons; insufficient or extraneous flanking sequence; sub-optimal mapping; inappropriate choice of k-mer length for subassembly; inappropriate reference; or other issues. If this occurs, we recommend testing the assembly with different k-mers, changing the flanking length, or trying alternative reference genomes. Mapping depth of the rDNA regions is also reported for each iteration; a marked decrease in mapping depth may also be indicative of problems.

Many published genome finishing tools and approaches offer improvements when applied to suitable datasets, but none (including the approach presented in this paper) is able in isolation to resolve all bacterial genome assembly issues. One constraint on the performance of riboSeed is the quality of rRNA annotations in reference strains. Although it is impossible to concretely confirm it is the case *in silico*, we (and others [28]) have found several reference genomes of the course of this study that we suspect have collapsed rDNA repeats. We recommend using a tool such as 16Stimator [34]

bioRxiv preprint doi: https://doi.org/10.1101/159798; this version posted July 14, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

or *rrn*DB [41] to estimate number of 16s (and therefore rDNAs) prior to assembly, or riboStack.py to assess mapping depths after running riboSeed.

As riboSeed relies on de Bruijn graph assembly, the results can be affected by assembler parameters. Care should be taken to find the most appropriate settings, particularly in regard to read trimming approach, range of k-mers, and error correction schemes.

One difficulty in determining the accuracy of rDNA counts in reference genome occures becuase genome sequences are often released without publishing the reads used to produced the genome. This practice is a major hindrance when attempting to perform coverage-based quality assessment, such as to infer the likelihood of collapsed rDNAs. While data transparency is expected for gene expression studies, that stance has not been universally adopted when publishing whole-genome sequencing results. To ensure the highest quality assemblies, it is imperative the researchers allow the scientific community to scrutinize the raw whole genome sequencing data with the same rigor that would be applied to any other type of high-throughput sequencing project.

Conclusions

Demonstration that rDNA flanking regions are conserved across taxa and that flanking regions of sufficient length are distinct within a genome allowed for the development of riboSeed, a *de fere novo* assembly method. riboSeed utilizes rDNA flanking regions to act as barcodes for repeated rDNAs, allowing the assembler to correctly place and orient the rDNA. *de fere novo* assembly can improve the assembly by bridging across ribosomal regions, and, in cases where rDNA repeats would otherwise result in incomplete scaffolding, can result in closure of a draft genome when used in conjunction with existing polishing tools. Although riboSeed is far from a silver bullet to provide perfect assemblies from short read technology, it shows the utility of using genomic reference data and mixed assembly approaches to overcome algorithmic obstacles. This approach to resolving rDNA repeats may allow further insight to be gained from large public repositories of short read sequencing data, such as SRA, and when used in conjunction with other genome finishing techniques, provides an avenue towards genome closure.

List of abbreviations

rDNA: DNA region coding for ribosomal RNA operon; rRNA: ribosomal RNA; SRA: Sequence Read Archive; ENA: European Nucleotide Archive; IG: intergenic, GAGE-B: Genome Assembly Gold-standard Evaluation for Bacteria

Availability of data and materials

The riboSeed pipeline and the datasets generated during the current study are available in the riboSeed GitHub repository, `https://github.com/nickp60/riboSeed`. The software is released under the MIT licence. Supplementary data can be found in the riboSeed repository under `Waters_et_al_2017`. The modified BugBuilder pipeline can be found at `https://github.com/nickp60/BugBuilder`. Reference strains used for this study can be found in Table S3.

bioRxiv preprint doi: https://doi.org/10.1101/159798; this version posted July 14, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Competing interests

The authors declare that they have no competing interests.

Funding

The work was funded through a joint studentship between The James Hutton Institute, Dundee, Scotland, and the National University of Ireland, Galway, Ireland.

Authors' contributions

NRW wrote all the bugs.

Acknowledgements

We thank Anton Korobeynikov for his helpful tips on optimizing SPAdes. Yoann Augagneur, Shaun Brinsmade, and Mohamed Sassi graciously provided access to the *S. aureus UAMS-1* genome sequencing data.

bioRxiv preprint doi: https://doi.org/10.1101/159798; this version posted July 14, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

References

[1] James Abbott. BugBuilder - An Automated Microbial Genome Assembly and Analysis Pipeline. *bioRxiv*, 2017.

[2] Can Alkan, Saba Sajjadian, and Evan E Eichler. Limitations of next-generation genome sequence assembly. *Nature Methods*, 8(1), 2011.

[3] Anton Bankevich, Sergey Nurk, Dmitry Antipov, Alexey A Gurevich, Mikhail Dvorkin, Alexander S Kulikov, Valery M Lesin, Sergey I Nikolenko, Son Pham, Andrey D Prjibelski, Alexey V Pyshkin, Alexander V Sirotkin, Nikolay Vyahhi, Glenn Tesler, Max A Alekseyev, and Pavel A Pevzner. SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *Journal of Computational Biology*, 19(5):455–477, 2012.

[4] Marten Boetzer, Walter Pirovano, DR Zerbino, E Birney, JT Simpson, K Wong, SD Jackman, JE Schein, SJ Jones, I Birol, R Li, W Fan, G Tian, H Zhu, L He, J Cai, Q Huang, Q Cai, B Li, Y Bai, Z Zhang, Y Zhang, W Wang, J Li, F Wei, H Li, M Jian, J Li, Z Zhang, R Nielsen, D Li, W Gu, Z Yang, Z Xuan, OA Ryder, FC Leung, Y Zhou, J Cao, X Sun, Y Fu, Marten Boetzer, CV Henkel, HJ Jansen, D Butler, Walter Pirovano, A Dayarian, TP Michael, AM Sengupta, IJ Tsai, TD Otto, M Berriman, B Langmead, C Trapnell, M Pop, SL Salzberg, H Li, R Durbin, H Li, B Handsaker, A Wysoker, T Fennell, J Ruan, N Homer, G Marth, G Abecasis, R Durbin, DR Kelley, MC Schatz, SL Salzberg, SL Salzberg, AM Phillippy, A Zimin, D Puiu, T Magoc, S Koren, TJ Treangen, MC Schatz, AL Delcher, M Roberts, G Marçais, M Pop, JA Yorke, I MacCallum, D Przybylski, S Gnerre, J Burton, I Shlyakhter, A Gnirke, J Malek, K McKernan, S Ranade, TP Shea, L Williams, S Young, C Nusbaum, and DB Jaffe. Toward almost closed genomes with GapFiller. *Genome Biology*, 13(6), 2012.

[5] Balázs Brankovics, Hao Zhang, Anne D. van Diepeningen, Theo A. J. van der Lee, Cees Waalwijk, G. Sybren de Hoog, C Hahn, L Bachmann, B Chevreux, RE Green, AS Malaspinas, J Krause, AW Briggs, PLF Johnson, C Uhler, IJ Tsai, TD Otto, M Berriman, D Hernandez, P François, L Farinelli, M Osterås, J Schrenzel, D Hernandez, R Tewhey, JB Veyrieras, L Farinelli, M Østerås, P François, DR Zerbino, E Birney, GSC Slater, E Birney, L Guo, L Han, L Yang, H Zeng, D Fan, Y Zhu, G Fourie, NA van der Merwe, BD Wingfield, M Bogale, B Tudzynski, MJ Wingfield, DM Hillis, and MT Dixon. GRAbB: Selective Assembly of Genomic Regions, a New Niche for Genomic Research. *PLOS Computational Biology*, 12(6):e1004753, 2016.

[6] Carlo P. J. M. Brouwer, Thuy Duong Vu, Miaomiao Zhou, Gianluigi Cardinali, Mick M. Welling, Nathalie van de Wiele, and Vincent Robert. Current Opportunities and Challenges of Next Generation Sequencing (NGS) of DNA; Determining Health and Disease. *British Biotechnology Journal*, 13(4), 2016.

[7] Rebecca J. Case, Yan Boucher, Ingela Dahllöf, Carola Holmström, W. Ford Doolittle, Staffan Kjelleberg, I. Dahllof, C. Holmstrom, W. Ford Doolittle, and Staffan Kjelleberg. Use of 16S rRNA and rpoB Genes as Molecular Markers for Microbial Ecology Studies. *Applied and Environmental Microbiology*, 73(1):278–288, 2007.

[8] Jill E Clarridge III. Impact of 16S rRNA gene sequence analysis for identification of bacteria on clinical microbiology and infectious diseases. *Clinical Microbiology Reviews*, 17(4):840–62, table of contents, 2004.

[9] Tom Coenye and Peter Vandamme. Intragenomic heterogeneity between multiple 16S ribosomal RNA operons in sequenced bacterial genomes. *FEMS Microbiology Letters*, 228:45–49, 2003.

[10] Aaron Darling, Andrew Tritt, Jonathan A. Eisen, and Marc T. Facciotti. Mauve Assembly Metrics. *Bioinformatics Advance Access*, 2011.

[11] Aaron C.E. Darling, Bob Mau, Frederick R. Blattner, and Nicole T. Perna. Mauve: Multiple Alignment of Conserved Genomic Sequence With Rearrangements. *Genome Research*, 14(7):1394–1403, 2004.

[12] Aarti Desai, Veer Singh Marwah, Akshay Yadav, Vineet Jha, Kishor Dhaygude, Ujwala Bangar, Vivek Kulkarni, and Abhay Jere. Identification of optimum sequencing depth especially for de novo genome assembly of small genomes using next generation sequencing data. *PloS ONE*, 8(4), 2013.

[13] Marco Galardini, Emanuele G Biondi, Marco Bazzicalupo, and Alessio Mengoni. CONTIGuator: a bacterial genomes finishing tool for structural insights on draft genomes. *Source Code for Biology and Medicine*, 6(11), 2011.

[14] Johan Goris, Konstantinos T Konstantinidis, Joel A Klappenbach, Tom Coenye, Peter Vandamme, James M Tiedje, and Correspondence Johan Goris. DNADNA hybridization values and their relationship to whole-genome sequence similarities. *International Journal of Systematic and Evolutionary Microbiology*, 57:81–91, 2007.

bioRxiv preprint doi: https://doi.org/10.1101/159798; this version posted July 14, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

[15] Dieval Guizelini, Roberto T. Raittz, Leonardo M. Cruz, Emanuel M. Souza, Maria B. R. Steffens, and Fabio O. Pedrosa. GFinisher: a new strategy to refine and finish bacterial genome assemblies. *Nature Scientific Reports*, 6, 2016.

[16] Alexey Gurevich, Vladislav Saveliev, Nikolay Vyahhi, and Glenn Tesler. QUAST: quality assessment tool for genome assemblies. *Bioinformatics*, 29(8):1072–1075, apr 2013.

[17] Tetsuya Hayashi, Kozo Makino, Makoto Ohnishi, Ken Kurokawa, Kazuo Ishii, Katsushi Yokoyama, Chang-Gyun Han, Eiichi Ohtsubo, Keisuke Nakayama, Takahiro Murata, Masashi Tanaka, Toru Tobe, Tetsuya Iida, Hideto Takami, Takeshi Honda, Chihiro Sasakawa, Naotake Ogasawara, Teruo Yasunaga, Satoru Kuhara, Tadayoshi Shiba, Masahira Hattori, Hideo Shinagawa, and Kozo Makino. Complete Genome Sequence of Enterohemorrhagic Eschelichia coli O157:H7 and Genomic Comparison with a Laboratory Strain K-12. *DNA Research*, 8:11–22, 2001.

[18] Matthew T. G. Holden, Edward J. Feil, Jodi A. Lindsay, Sharon J. Peacock, Nicholas P. J. Day, Mark C. Enright, Tim J. Foster, Catrin E. Moore, Laurence Hurst, Rebecca Atkin, Andrew Barron, Nathalie Bason, Stephen D. Bentley, Carol Chillingworth, Tracey Chillingworth, Carol Churcher, Louise Clark, Craig Corton, Ann Cronin, Jon Doggett, Linda Dowd, Theresa Feltwell, Zahra Hance, Barbara Harris, Heidi Hauser, Simon Holroyd, Kay Jagels, Keith D. James, Nicola Lennard, Alexandra Line, Rebecca Mayes, Sharon Moule, Karen Mungall, Douglas Ormond, Michael A. Quail, Ester Rabbinowitsch, Kim Rutherford, Mandy Sanders, Sarah Sharp, Mark Simmonds, Kim Stevens, Sally Whitehead, Bart G. Barrell, Brian G. Spratt, and Julian Parkhill. Complete genomes of two clinical Staphylococcus aureus strains: evidence for the rapid evolution of virulence and drug resistance. *Proceedings of the National Academy of Sciences of the United States of America*, 101(26):9786–91, jun 2004.

[19] Xuesong Hu, Jianying Yuan, Yujian Shi, Jianliang Lu, Binghang Liu, Zhenyu Li, Yanxiang Chen, Desheng Mu, Hao Zhang, Nan Li, Zhen Yue, Fan Bai, Heng Li, and Wei Fan. pIRS: Profile-based Illumina pair-end reads simulator. *Bioinformatics*, 28(11):1533–1535, jun 2012.

[20] Kazutaka Katoh, Kazuharu Misawa, Keiichi Kuma, and Takashi Miyata. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research*, 30(14):3059–3066, jul 2002.

[21] Yuichi Kodama, Martin Shumway, and Rasko Leinonen. The sequence read archive: explosive growth of sequencing data. *Nucleic Acids Research*, 40(Database issue):D54–56, jan 2012.

[22] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, and 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–9, aug 2009.

[23] Pinglei Liu, Peng Li, Xiaofei Jiang, Dexi Bi, Yingzhou Xie, Cui Tai, Zixin Deng, Kumar Rajakumar, and Hong-Yu Ou. Complete genome sequence of Klebsiella pneumoniae subsp. pneumoniae HS11286, a multidrug-resistant strain isolated from human sputum. *Journal of Bacteriology*, 194(7):1841–1842, apr 2012.

[24] Oksana Lukjancenko, Trudy M. Wassenaar, and David W. Ussery. Comparison of 61 Sequenced Escherichia coli Genomes. *Microbial Ecology*, 60, 2010.

[25] Ruibang Luo, Binghang Liu, Yinlong Xie, Zhenyu Li, Weihua Huang, Jianying Yuan, Guangzhu He, Yanxiang Chen, Qi Pan, Yong Yunjie Liu, Jingbo Tang, Gengxiong Wu, Hao Zhang, Yujian Shi, Yong Yunjie Liu, Chang Yu, Bo Wang, Yao Lu, Changlei Han, David W Cheung, Siu-Ming Yiu, Shaoliang Peng, Zhu Xiaoqian, Guangming Liu, Xiangke Liao, Yingrui Li, Huanming Yang, Jun Jian Wang, Tak-Wah Lam, and Jun Jian Wang. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *GigaScience*, 1(1):18, 2012.

[26] Tanja Magoc, Stephan Pabinger, Stefan Canzar, Xinyue Liu, Qi Su, Daniela Puiu, Luke J. Tallon, and Steven L. Salzberg. GAGE-B: an evaluation of genome assemblers for bacterial organisms. *Bioinformatics*, 29(14):1718–1725, 2013.

[27] Diego CB Mariano, Felipe L Pereira, Preetam Ghosh, Debmalya Barh, Henrique CP Figueiredo, Artur Silva, Rommel TJ Ramos, and Vasco AC Azevedo. MapRepeat: an approach for effective assembly of repetitive regions in prokaryotic genomes. *Bioinformatician*, 11(6):276–279, 2015.

[28] Diego César Batista Mariano, Thiago De Jesus Sousa, Felipe Luiz Pereira, Flávia Aburjaile, Debmalya Barh, Flávia Rocha, Anne Cybelle Pinto, Syed Shah Hassan, Tessália Diniz, Luerce Saraiva, Fernanda Alves Dorella, Alex Fiorini De Carvalho, Carlos Augusto Gomes Leal, Henrique César, Pereira Figueiredo, Artur Silva, Rommel

bioRxiv preprint doi: https://doi.org/10.1101/159798; this version posted July 14, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Thiago, Jucá Ramos, Vasco Ariston, and Carvalho Azevedo. Whole-genome optical mapping reveals a misassembly between two rRNA operons of Corynebacterium pseudotuberculosis strain 1002. *BMC Genomics*, 17, 2016.

[29] Mari Miyamoto, Daisuke Motooka, Kazuyoshi Gotoh, Takamasa Imai, Kazutoshi Yoshitake, Naohisa Goto, Tetsuya Iida, Teruo Yasunaga, Toshihiro Horii, Kazuharu Arakawa, Masahiro Kasahara, and Shota Nakamura. Performance comparison of second-and third-generation sequencers using a bacterial genome with two chromosomes. *BMC Genomics*, 15, 2014.

[30] Claudia Moreno, Jaime Romero, and Romilio T. Espejo. Polymorphism in repeated 16S rRNA genes is a common property of type strains and environmental isolates of the genus Vibrio. *Microbiology*, 148:1233–1239, 2002.

[31] Francesca Nadalin, Francesco Vezzi, and Alberto Policriti. GapFiller: a de novo assembly approach to fill the gap within paired reads. *BMC Bioinformatics*, 13:12–14, 2012.

[32] Niranjan Nagarajan, Christopher Cook, MariaPia Di Bonaventura, Hong Ge, Allen Richards, Kimberly A Bishop-Lilly, Robert Desalle, Timothy D Read, and Mihai Pop. Finishing genomes with limited resources: lessons from an ensemble of microbial genomes. *BMC Genomics*, 11(1):242, 2010.

[33] Makoto Ohnishi, Takahiro Muratal, Keisuke Nakayama, Satoru Kuhara, Masahiro Hattori, Ken Kurokawa, Teruo Yasunaga, K Atsushi Yokoyamas, Kozo Makinos, Hideo Shinagawa, and Tetsuya Hayashi. Comparative Analysis of the Whole Set of rRNA Operons Between an Enterohemorrhagic Escherichia coli 0157:H7 Sakai Strain and an Escherichia coli K-12 Strain MG1655. *Systematic and Applied Microbiology*, 23:315–324, 2000.

[34] Matthew Perisin, Madlen Vetter, Jack A Gilbert, and Joy Bergelson. 16Stimator: statistical estimation of ribosomal gene copy numbers from draft genome assemblies. *The ISME Journal*, 10(4):1020–1024, apr 2016.

[35] Vitor C Piro, Helisson Faoro, Vinicius A Weiss, Maria Br Steffens, Fabio O Pedrosa, Emanuel M Souza, and Roberto T Raittz. FGAP: an automated gap closing tool. *BMC Research Notes*, 7, 2014.

[36] Aaron R. Quinlan and Ira M. Hall. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–84210, 2010.

[37] Peter Rice, Ian Longden, and Alan Bleasby. EMBOSS: the European Molecular Biology Open Software Suite. *Trends in Genetics*, 16(6):276–7, jun 2000.

[38] Fatemeh Sanjar, S L Rajasekhar Karna, Tsute Chen, Ping Chen, Johnathan J Abercrombie, and Kai P Leung. Whole-Genome Sequence of Multidrug-Resistant Pseudomonas aeruginosa Strain BAMCPA07-48, Isolated from a Combat Injury Wound. *Genome Announcements*, 4(4), 2016.

[39] Mohamed Sassi, Deepak Sharma, Shaun R Brinsmade, Brice Felden, and Yoann Augagneur. Genome Sequence of the Clinical Isolate Staphylococcus aureus subsp. aureus Strain UAMS-1. *Genome Announcements*, 3(1), 2015.

[40] Armin O. Schmitt and Hanspeter Herzel. Estimating the Entropy of DNA Sequences Introduction: Order and Disorder of Sequences. *Journal of Theoretical Biology*, 1888:369–377, 1997.

[41] Steven F. Stoddard, Byron J. Smith, Robert Hein, Benjamin R. K. Roller, and Thomas M. Schmidt. rrnDB: improved tools for interpreting rRNA gene abundance in bacteria and archaea and a new foundation for future development. *Nucleic Acids Research*, 43(D1), 2014.

[42] Todd J Treangen, Brian D Ondov, Sergey Koren, and Adam M Phillippy. The Harvest suite for rapid core-genome alignment and visualization of thousands of intraspecific microbial genomes. *Genome Biology*, 15(524), 2014.

[43] Todd J Treangen and Steven L Salzberg. Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nature Reviews Genetics*, 13(1):36–46, 2011.

[44] Isheng J Tsai, Thomas D Otto, and Matthew Berriman. Improving draft assemblies by iterative mapping and assembly of short reads to eliminate gaps. *Genome Biology*, 11, 2010.

[45] Sagar M Utturkar, Dawn M Klingeman, Miriam L Land, Christopher W Schadt, Mitchel J Doktycz, Dale A Pelletier, and Steven D Brown. Evaluation and validation of de novo and hybrid assembly techniques to derive high-quality genome sequences. *Bioinformatics*, 30(19):2709–2716, 2014.

[46] Benjamin P Vandervalk, Chen Yang, Zhuyi Xue, Karthika Raghavan, Justin Chu, Hamid Mohamadi, Shaun D Jackman, Readman Chiu, René L Warren, and Inanç Birol. Konnector v2.0: pseudo-long reads from paired-end sequencing data. *BMC Medical Genomics*, 8:2–5, 2015.

bioRxiv preprint doi: https://doi.org/10.1101/159798; this version posted July 14, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

[47] Tomáš Větrovský and Petr Baldrian. The Variability of the 16S rRNA Gene in Bacterial Genomes and Its Consequences for Bacterial Community Analyses. *PLoS ONE*, 8(2), 2013.

[48] Dapeng Wang, Jiayue Xu, and Jun Yu. KGCAK: a K-mer based database for genome-wide phylogeny and complexity evaluation. *Biology Direct*, 10(53), 2015.

[49] William G. Weisburg, Susan M. Barns, Dale A. Pelletier, and David J. Lane. 16S ribosomal DNA amplification for phylogenetic study. *Journal of Bacteriology*, 173(2):697–703, 1991.

[50] N. Whiteford, Niall Haslam, Gerald Weber, Adam Prügel-Bennett, Jonathan W. Essex, Peter L. Roach, Mark Bradley, and Cameron Neylon. An analysis of the feasibility of short read sequencing. *Nucleic Acids Research*, 33(19), 2005.

[51] Carl R Woese, Otto Kandlert, and Mark L Wheelis. Towards a natural system of organisms: Proposal for the domains Archaea, Bacteria, and Eucarya. *Proceedings of the National Academy of Sciences of the United States of America*, 87:4576–4579, 1990.

[52] Xing Yang, Daniel Medvin, Giri Narasimhan, Deborah Yoder-Himes, and Stephen Lory. CloG: a pipeline for closing gaps in a draft assembly using short reads. *2011 IEEE 1st International Conference on Computational Advances in Bio and Medical Sciences*, pages 202–207, 2011.

bioRxiv preprint doi: https://doi.org/10.1101/159798; this version posted July 14, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Supplementary Data

Making the artificial test genome

The artificial genome used for testing was constructed using the `makeToyGenome.sh` script included in the GitHub repository under the `scripts` directory. Briefly, the 7 rDNA regions from the *E. coli Sakai* genome were extracted with 5kb flanking sequence upstream and downstream; these sequences were then concatenated end to end to form a single, ~100kb sequence containing the 7 rDNAs as well as their flanking context.

Effect of reference sequence identity on riboSeed performance: Methods

A range of substitutions were introduced into a artificial genome using the `runDegenerate.sh` script (included in the GitHub repository under the `scripts` directory), which facilitates the following procedure. An artificial test genome is constructed (see above), and reads simulated using pIRS (100bp, 300bp inserts, stdev 10, 30-fold coverage, built-in error profile). Then, for each of a range of substitution frequencies, substitutions are introduced into the simulated genome, either just in the flanking regions or throughout. riboSeed is run on the reads using the mutated genome as the reference, and the results are evaluated with riboScore. This script was run 100 times, using a different random seed each time. As the algorithm used by python's pseudo random number generator may differ between operating systems, comparable but not identical results can be expected.

Archaeal Datasets

We assessed the effectiveness of riboSeed with assembling archaeal genomes. Most (~55%) archaeal genomes have only a single rDNA, and none has been observed to have more than four. As riboSeed requires a sequencing dataset and a reference genome, applicability was limited; of the 104 entries in *rrn*DB with multiple rDNAs, only 7 had multiple entries at the species level. Among those, only 2 had publicly available short read data. We used riboSeed to re-assemble *Methanosarcina barkeri Fusaro DSMZ804* (Ion Torrent PGM, 89bp single-end reads) and *Methanobacterium formicicum st. BRM9* (Illumina HiSeq 2000, 100bp paired-end reads). *Methanobacterium formicicum st. JCM10132* (DRR017790) and *Methanosarcina barkeri Fusaro DSMZ804* (SRR2064286) were the only ones that were suitable for riboSeed, meaning that there was publicly available short read data and that there is a related genome at the species level which is complete.

M. formicicum st. JCM10132 was sequenced on an Ion Torrent PGM, generating 106.5Mbp of single-end data. *M formicicum BRM9* (CP006933.1) was used as a reference. The resulting *de fere novo* assembly resulted in assembly of 1 of 2 rDNA gaps. This represents the first application of riboSeed to Ion Torrent data.

Methanosarcina barkeri Fusaro DSMZ804 was sequenced using an Illumina HiSeq2000 with 101bp paired-end reads, with an average fragment length of 400bp. We downsampled to use 5% of the 19.4Gbp dataset. *Methanosarcina barkeri str. Wiesmoor* was used as a reference. The resulting riboSeed assembly showed correct assembly of 3 of 3 rDNAs, while *de novo* assemble failed to resolve any.

Taken together, we show that given appropriate datasets, archaeal datasets can be processed in the same manner used for bacteria.

~

bioRxiv preprint doi: https://doi.org/10.1101/159798; this version posted July 14, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Table S1: Accessions for 25 *E. coli* genomes

GCA_000021125.1_ASM2112v1
GCA_000023665.1_ASM2366v1
GCA_000026545.1_ASM2654v1
GCA_000262125.1_ASM26212v1
GCA_000273425.1_Esch_coli_MG12655_V1
GCA_000299255.1_ASM29925v1
GCA_000714595.1_ASM71459v1
GCA_000967155.1_HUSEC2011CHR1
GCA_000974405.1_ASM97440v1
GCA_000974465.1_ASM97446v1
GCA_000974575.1_ASM97457v1
GCA_001020945.2_ASM102094v2
GCA_001566675.1_ASM156667v1
GCA_002012245.1_ASM201224v1
GCA_001750845.1_ASM175084v1
GCA_001886755.1_ASM188675v1
GCA_001901145.1_ASM190114v1
GCA_002012145.1_ASM201214v1
GCA_900096815.1_Ecoli_AG100_Sample2_M9_Assembly
GCA_002116715.1_ASM211671v1
GCA_002118095.1_ASM211809v1
GCA_002125925.1_ASM212592v1
GCA_001612475.1_ASM161247v1
GCA_001651965.1_ASM165196v1
GCA_001721125.1_ASM172112v1

All available at `ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/`

Table S2: Hits resulting from searching the SRA database for various sequencing technologies as of January, 2017

Search term	Hits	Percentage
illumina	2242225	94.27
pacbio	21131	0.89
ion	30560	1.28
roche	42445	1.78
oxford	12301	0.52
solid	29791	1.25
Total	2378453	100

Table S3: Strain names and accessions for reference genomes used in this study

Strain Name	Accession
E. coli MG1655	NC_000913.3
A. hydrophila ATCC 7966	NC_008570.1
B. cereus ATCC 10987	AE017194.1
B. cereus NC7401	NC_016771.1
B. fragilis 638R	FQ312004.1
R. sphaeroides ATCC 17029	NC_009049.1, NC_009050.1
S. aureus TCH1516	NC_010079.1
S. aureus MRSA252	BX571856.1
V. cholerae El Tor str. N16961	NC_002505.1, NC_002506.1
X. axonopodis pv. Citrumelo	CP002914.1
P. aeruginosa BAMCPA07-48	CP015377.1
P. aeruginosa ATCC 15692	NZ_CP017149.1

bioRxiv preprint doi: https://doi.org/10.1101/159798; this version posted July 14, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Table S4: Software Versions

Tool	Version
Mauve	2015-02-13 build 0
BLAST+	2.2.28+
Barrnap	0.7
BWA	0.7.12-r1039
samtools	1.3.1
MAFFT	v7.215
SPAdes	v3.9.0
QUAST	4.1
bedtools	2.17.0
EMBOSS	6.6.0
pIRS	2.0.2

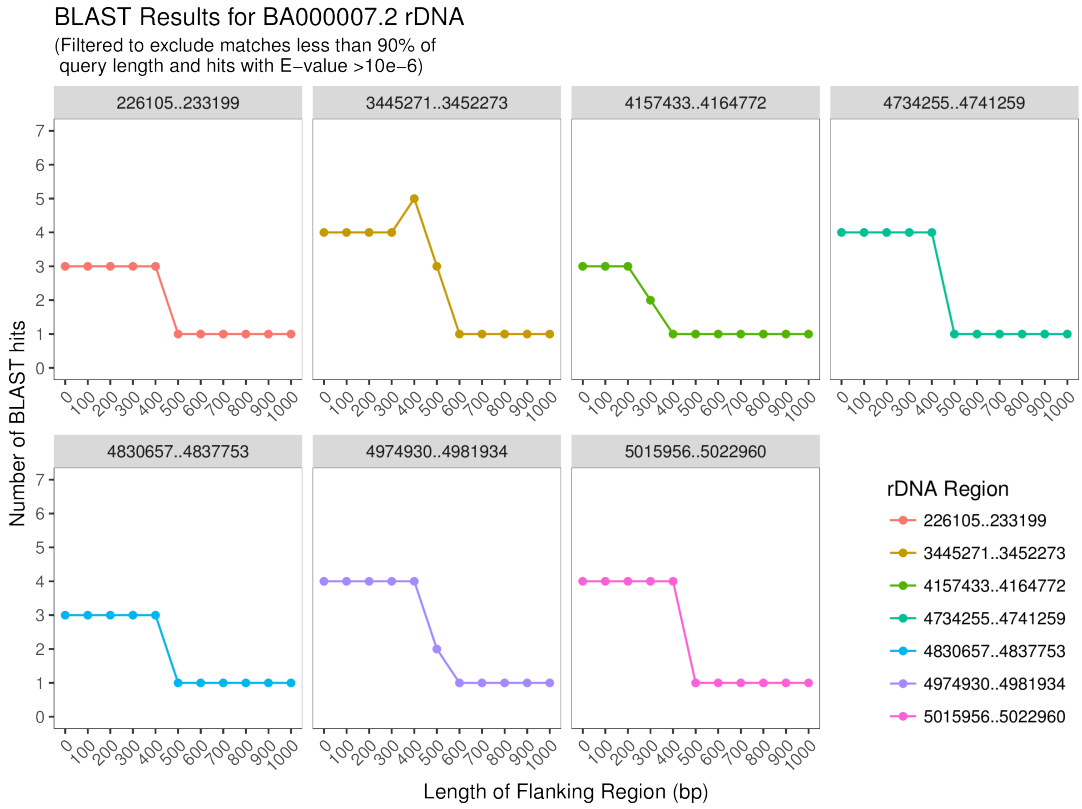


Figure S1: BLASTn was used to perform *in silico* DNA-DNA hybridization of all rDNA regions from *E. coli Sakai* with variable flanking lengths. The number of hits is a proxy for occurrences in the genome; increasing the flanking length increases the specificity.

bioRxiv preprint doi: https://doi.org/10.1101/159798; this version posted July 14, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

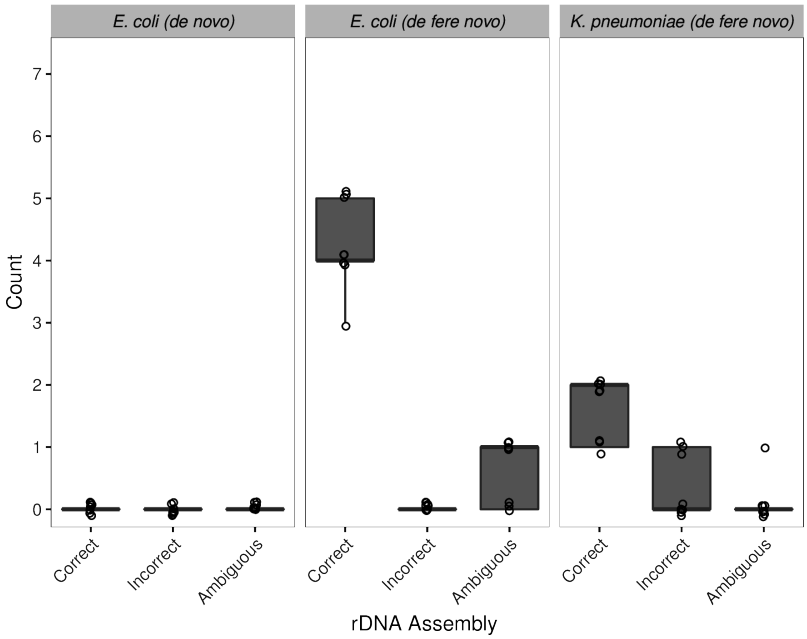
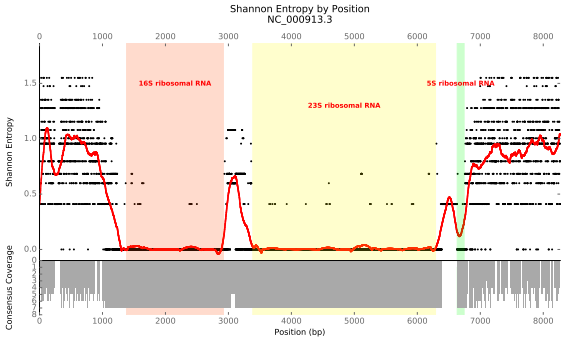
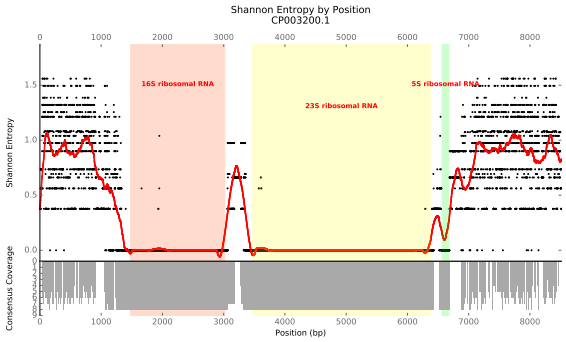


Figure S2: Assembly of artificial genome. *De fere novo* results in closure of 3-5 rDNAs with the correct reference; only 1-2 rDNAs are correctly assembled using *K. pneumoniae*. No rDNAs are assembled with *de novo* assembly. Scored with riboScore.py. N=8.

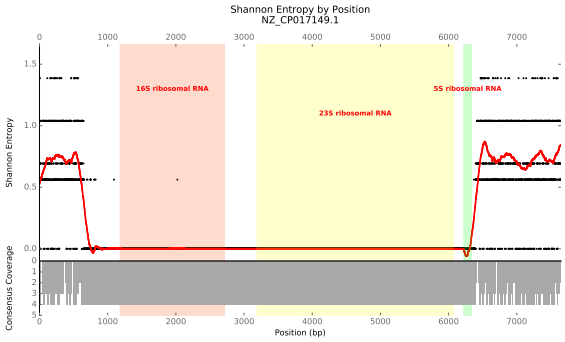
bioRxiv preprint doi: https://doi.org/10.1101/159798; this version posted July 14, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.



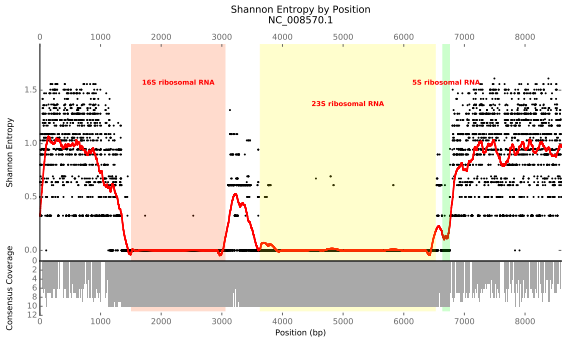
(a) *E. coli MG1655* (NC_000913.3)



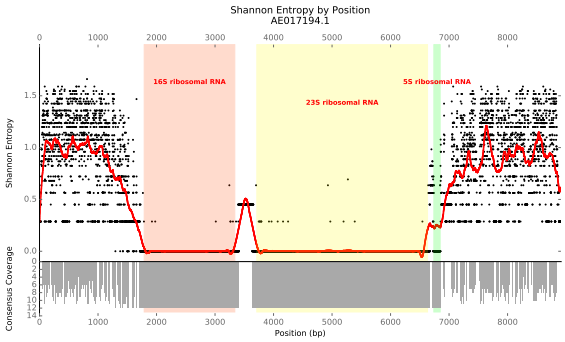
(b) *K. pneumoniae subsp. pneumoniae HS11286* (CP003200.1)



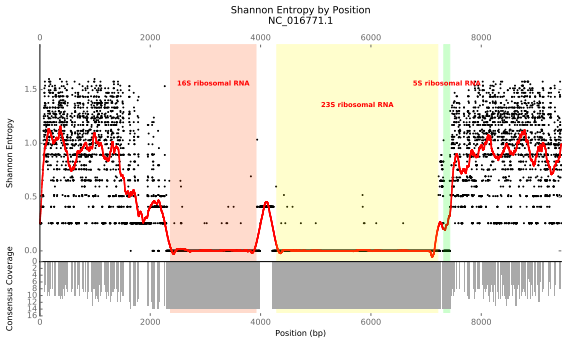
(c) *P. aeruginosa strain ATCC 15692* (NZ_CP017149.1)



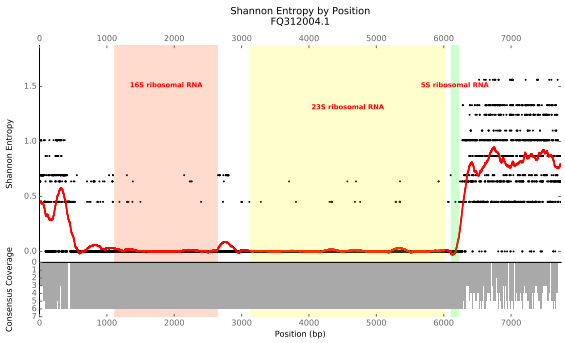
(d) *A. hydrophila ATCC 7966* (NC_008570.1)



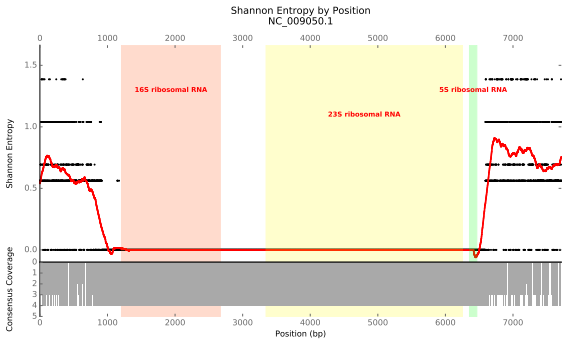
(e) *B. cereus ATCC 10987* (AE017194.1)



(f) *B. cereus NC7401* (NC_016771.1)

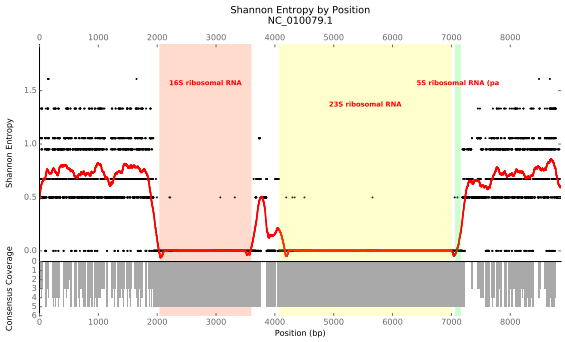


(g) *B. fragilis 638R* (FQ312004.1)

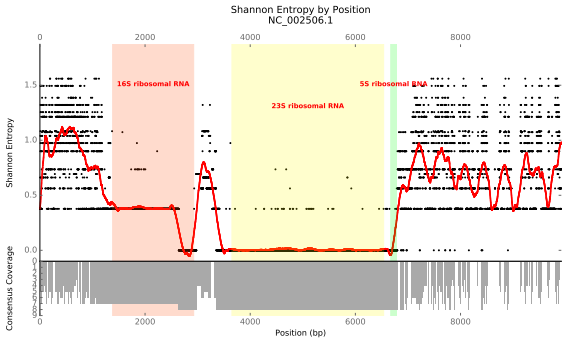


(h) *R. sphaeroides ATCC 17029* (NC_009049.1, NC_009050.1)

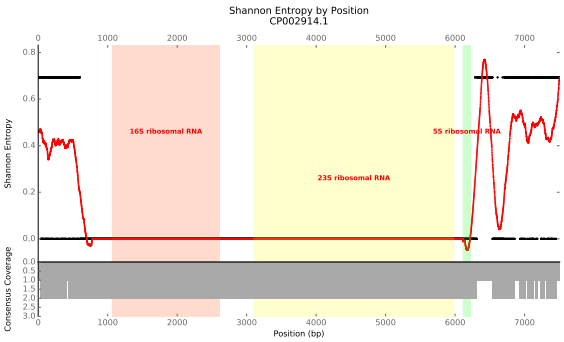
bioRxiv preprint doi: https://doi.org/10.1101/159798; this version posted July 14, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.



(i) *S. aureus TCH1516* (NC_010079.1)



(j) *V. cholerae El Tor str. N16961* (NC_002505.1) (NC_002506.1)



(k) *X. axonopodis pv. Citrumelo* (CP002914.1)

Figure S3: riboScan.py,riboSelect.py, and riboSnag.py were run on all the genomes used as references for *de fere novo* assemblies. Consensus alignment depth (grey bars) and Shannon entropy (black points, smoothed entropy as red line) for aligned rDNA regions.