Supplementary Data

riboSeed: leveraging prokaryotic genomic architecture to assemble across ribosomal regions

Nicholas R. Waters,[1,2] Florence Abram,[1] Fiona Brennan,[1,3] Ashleigh Holmes,[4] and Leighton Pritchard[2*]

[1] *Department of Microbiology, School of Natural Sciences, National University of Ireland, Galway, Ireland*
[2] *Information and Computational Sciences, James Hutton Institute, Invergowrie, Dundee DD2 5DA, Scotland*
[3] *Soil and Environmental Microbiology, Environmental Research Centre, Johnstown Castle, Wexford, Ireland*
[4] *Cell and Molecular Sciences, James Hutton Institute, Invergowrie, Dundee DD2 5DA, Scotland*

*To whom correspondence should be addressed: leighton.pritchard@hutton.ac.uk

Compiled: 2017/10/25 20:39:00

Extended Methods

Reference Selection Recommendations

Using a close reference sequence maximizes chances of a successful assembly. We have outlined two methods that to select an appropriate reference for a given isolate: a robust method, and a quick method.

Method 1: Kraken

Kraken is a kmer-based phylogeny tool that can be used to idenify the strains present in a metagenomic dataset; the installation and usage instructions can be found here: https://ccb.jhu.edu/software/kraken/. Download and install Kraken, along with the MiniKraken database from their website. Run Kraken on the isolate's reads, and generate the Kraken report.

The MiniKraken database was built from all the complete genomes from RefSeq, allowing the user identify which strain in the database has the closest match to the sequenced isolate.

Method 2: reads2type and cgFind

reads2type is also a kmer-based phylogeny tool, but it relies on a lightweight, prebuilt database of 55-mers that allows the analysis to be performed in the web browser, and it does not require the user to upload the whole read files, allowing it to perform well when either speed or network access is limitted. It works by taking one read at a time from your file, generating 55-mers, and comparing to a prebuild databse. If there is not enough taxanomic information to indentify the isolate off of that read alone, additional reads will be processed until a single taxonomy is achieved. This method works best on trimmed reads. Instructions and the webserver can be found at https://cge.cbs.dtu.dk/services/Reads2Type/

Now, given the genus and species from reads2type, users can make use of cgFind, our web tool developed to provide easy access to downloadable genomes based on the complete prokaryotic genomes found in NCBI. The tool can be found at https://nickp60.github.io/cgfind.

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

The following 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: 0.0, 0.0025, 0.0050, 0.0075, 0.0100, 0.0150, 0.0200, 0.0250, 0.0500, 0.0750, 0.1000, 0.1250, 0.1500, 0.1750, 0.2000, 0.2250, 0.2500, 0.2750, 0.3000. 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 pseudo random number generation may differ between operating systems, comparable but not identical results can be expected.

Performance on Archaeal Data

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* (Illumina HiSeq 2000, 100bp paired-end reads) and *Methanobacterium formicicum st. JCM10132* (DRR017790, Ion Torrent PGM, 89bp single-end reads). *Methanobacterium formicicum st. BRM9* 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 with seqtk (https://github.com/lh3 *Methanosarcina barkeri str. Wiesmoor* (CP009526.1) 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.

Key Parameters

`--ref_as_contig`

The assembly that results from including riboSeed's "long reads" is sensitive to the manner in which they are incorporated into the *de novo* assembly. Here, for our analyses, we used the SPAdes assembler, as it has built-in ways to include contigs (using the "–trusted-contigs" or "–untrusted-contigs") in FASTA format. Other assemblers could be used, but most require long reads to have a quality score associate with them, preventing direct use of riboSeeds long reads.

As mentioned in the Methods section, riboSeed uses the reference rDNA region in the initial subassembly; in subsequent subassemblies, the longest contig of the previous subsassembly is used. The manner in which these regions are can be one of four options to `--ref_as_contig`: `trusted`, `untrusted`, `inferr`, or `ignore`. Additionally, if the user is worried that the reference rDNA will too heavily influence the initial subassembly, the can enable the `--initial_consensus` flag to use a mapping consensus assembly instead of the de Bruijn graph based assembly from SPAdes.

The default manner in which rDNA regions (either from the reference or from the previous iteration's subassembly) behaviour is to infer (`--ref_as_contig infer`): if the percent of reads mapping to the (whole) reference sequence is over 80%, than the rDNA region will be included as a trusted contig. If below 80%, the reads will be treated as untrusted.

If a user wishes to have the subassemblies only using the reads (true *de novo* assembly), they can use the `ignore` option. We only recommend this with very close references.

Further, if the user wishes to explicity define the behaviour, `trusted` or `untrusted` can be provided to the `--ref_as_contig` argument.

`--score_min`

By default, the accepted alignment score for BWA mapping is $\frac{1}{2}$ the read length. If need, this can be increased for greater stringency when dealing with more divergent references, or decreased to include more reads, which may be adventageous when assembling a low coverage dataset.

Excluding GAGE-B HiSeq *B. cereus*

The GAGE-B paper ¡cite¿ notes that the *B. cereus* HiSeq dataset proved particularly difficult to assemble. After noticing this irregularity, we re-assembled the trimmed reads downloaded from the GAGE-B website with metaSPAdes ¡cite¿ using default parameters. Then, blastn was used to search the resulting contigs against NCBI's nt database (May, 2017) to get a list of hits according to the blobtools ¡cite¿ specifications. Blobtools was then used to plot the

hit coverage, taxonomy, and GC-content of the contigs. This revealed what appears to be a contamination. **??**A. As the GC content of the contaminating organisms did not differ from *B. cereus*, we believe that many tools that use GC-skew to detect contamination would not have detected the problem with this dataset.

To further show the contamination, we split reads into those pairs mapping to the *B. cereus* ATCC 10987 reference genome and those unmapped. BWA MEM was used to map the 12039737 reads to the reference genome; samtools was used to separate the 7500534 reads (0.62%) that mapped from the 3984200 reads (0.33%) that failed to map with default parameters*.Each of these sets of reads was then assembled, BLASTed against the nt database, and plotted with blobtools **??**B and C.

Further, MaxBin, Kraken, and MBBC also supported the hypothesis that the sample is contaminated with approximately one third of reads originating from a non-*B. cereus* strain.

*The remaining percentage are those pairs where only one read aligned to the reference..

Table S1: 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 S2: Accessions for 25 *E. coli* genomes

GCA_000005845.2_ASM584v2
GCA_000019385.1_ASM1938v1
GCA_000026245.1_ASM2624v1
GCA_000026345.1_ASM2634v1
GCA_000026545.1_ASM2654v1
GCA_000146735.1_ASM14673v1
GCA_000257275.1_ASM25727v1
GCA_000520055.1_ASM52005v1
GCA_000732965.1_ASM73296v1
GCA_001007915.1_ASM100791v1
GCA_001442495.1_ASM144249v1
GCA_001469815.1_ASM146981v1
GCA_001660565.1_ASM166056v1
GCA_001660585.1_ASM166058v1
GCA_001753565.1_ASM175356v1
GCA_001888075.1_ASM188807v1
GCA_001901025.1_ASM190102v1
GCA_001936315.1_ASM193631v1
GCA_002056065.1_ASM205606v1
GCA_002078295.1_ASM207829v1
GCA_002156825.1_ASM215682v1
GCA_002163935.1_ASM216393v1
GCA_002192275.1_ASM219227v1
GCA_002220265.1_ASM222026v1
GCA_900096795.1_Ecoli_AG100_Sample3_Doxycycline_Assembly

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

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

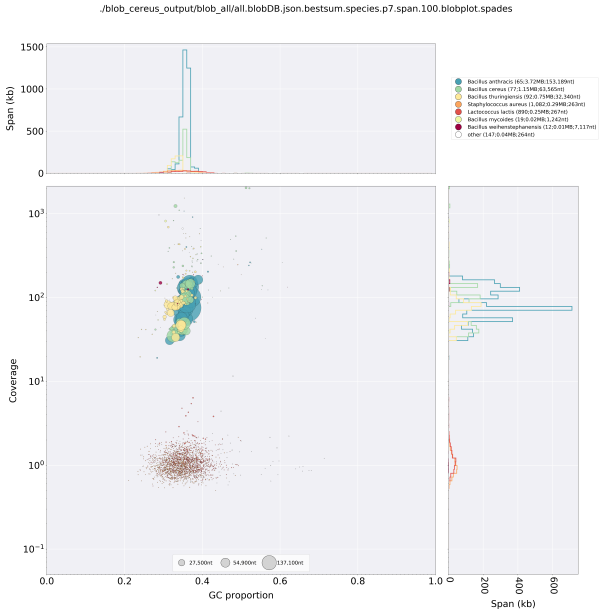
Strain Name	Accession
E. coli MG1655	NC_000913.3
A. hydrophila ATCC 7966	NC_008570.1
B. cereus ATCC 10987	AE017194.1
B. cereus NC7401	NC_016771.1
B. fragilis 638R	FQ312004.1
R. sphaeroides ATCC 17029	NC_009049.1, NC_009050.1
S. aureus TCH1516	NC_010079.1
S. aureus MRSA252	BX571856.1
V. cholerae El Tor 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

Table S4: Software Versions

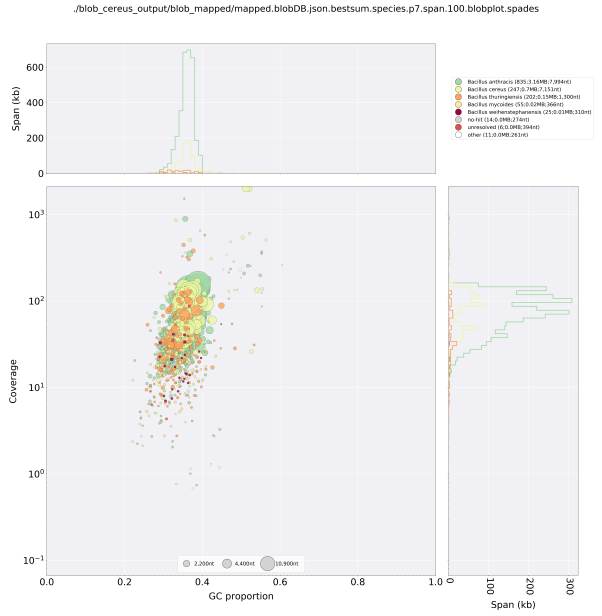
Tool	Version
Mauve	2015-02-13 build 0
BLAST+	2.2.28+
Barrnap	0.8
BWA	0.7.8-r455
samtools	1.4.1
MAFFT	v7.310
SPAdes	v3.9.0
QUAST	4.4
bedtools	2.17.0
EMBOSS	6.5.7
pIRS	2.0.2
seqtk	1.2-r94
Parsnp	v1.2

Table S5: QUAST results of *P. aeruginosa* BAMCPA07-48 assemblies comparing *de fere novo* assembly, *de novo* assembly, and reference-based assembly (where the *P. aeruginosa* ATCC 15692 reference is included in the *de novo* assembly as a trusted contig). Blue and red highlight the best and worst results, respectively. riboSeed's *de fere novo* assembly either outperforms or performs comparably to *de novo* assembly in all categories. Using the reference as a trusted contig results in longer assemblies but with a much higher rate of missmatches, indels, and missasemblies.

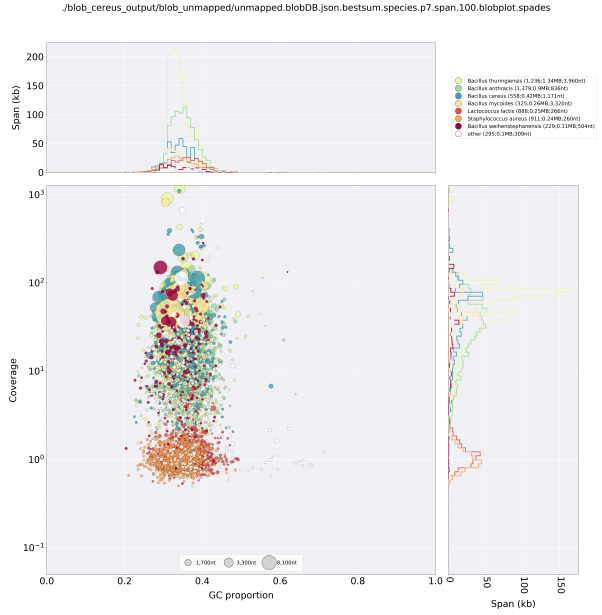
Genome statistics	***de fere novo***	***de novo***	**reference-based**
Genome fraction (%)	98.106	97.868	98
Duplication ratio	1.001	1.001	1.017
Largest alignment	630503	402463	757685
Total aligned length	6893293	6876715	6993532
NGA50	176510	176510	135376
LGA50	12	13	14
Misassemblies			
# misassemblies	2	2	9
Misassembled contigs length	212498	212498	2347560
Mismatches			
# mismatches per 100 kbp	1.89	1.69	11.66
# indels per 100 kbp	2.48	2.44	2.94
# N's per 100 kbp	0	0	0
Statistics without reference			
# contigs	154	159	388
Largest contig	630503	402463	1103106
Total length	6893293	6876715	7237564
Total length (>= 1000 bp)	6865091	6848513	7130244
Total length (>= 10000 bp)	6687664	6663031	6617370
Total length (>= 50000 bp)	6242010	6168232	5534330



(a) All reads



(b) Reads aligning to *B. cereus* ATCC 10987



(c) Reads failing to align to *B. cereus* ATCC 10987

Figure S1: Assessing contamination in the GAGE-B HiSeq *B. cereus* dataset with blobtools. Reads were assem-

```
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 S2: Pseudocode of riboSeed algorithm

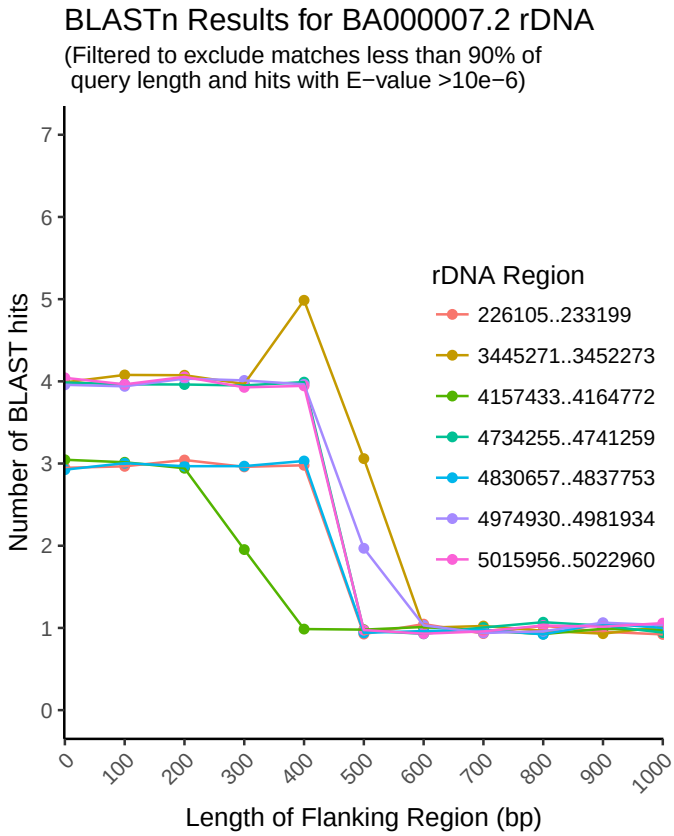


Figure S3: BLASTn was used to perform *in silico* DNA-DNA hybridization of all rDNA regions from *E. coli Sakai* with variable flanking lengths. The number of hits is a proxy for occurrences in the genome; increasing the flanking length increases the specificity. (Points are jittered to aide visibility for overlapping values.)

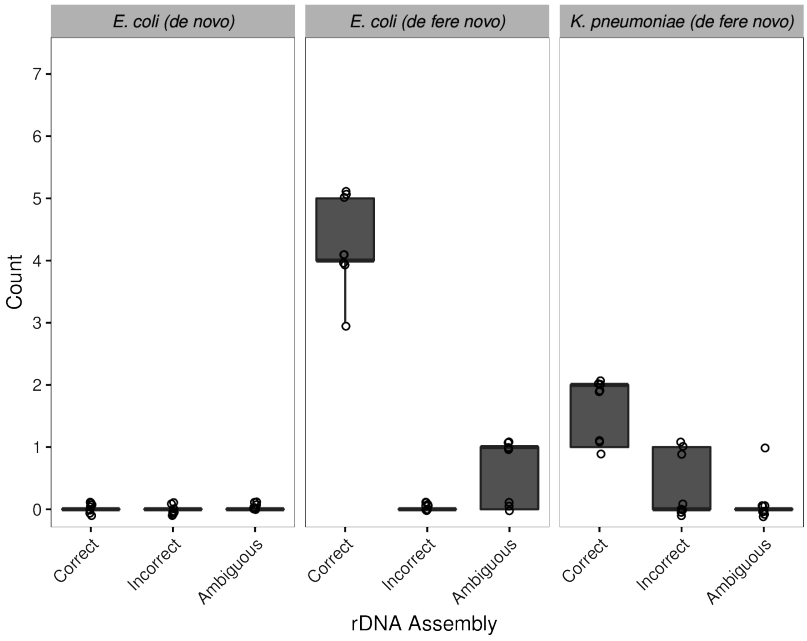
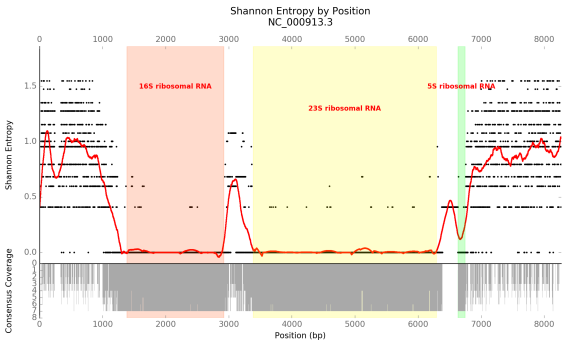
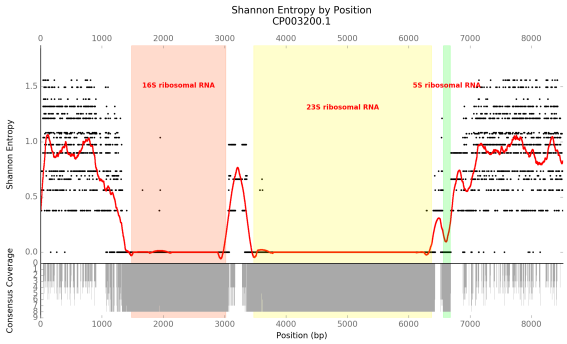


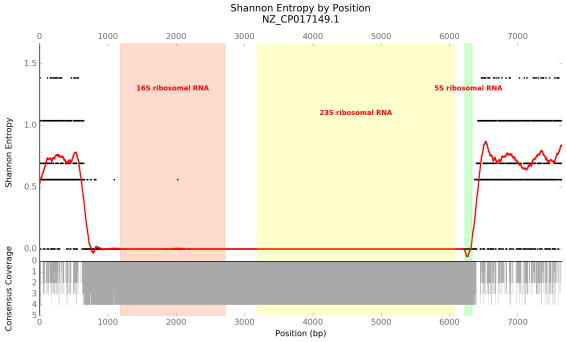
Figure S4: Assembly of artificial genome. *De fere novo* results in closure of 3-5 rDNAs with the correct reference; only 1-2 rDNAs are correctly assembled using *K. pneumoniae*. No rDNAs are assembled with *de novo* assembly. Scored with riboScore.py. N=8.



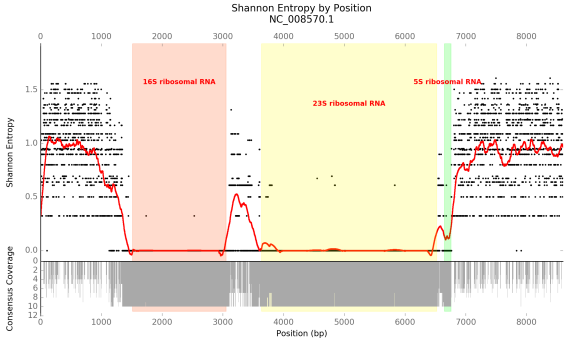
(a) *E. coli MG1655* (NC_000913.3)



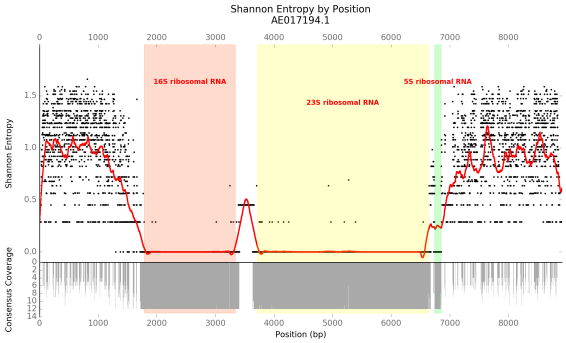
(b) *K. pneumoniae subsp. pneumoniae HS11286* (CP003200.1)



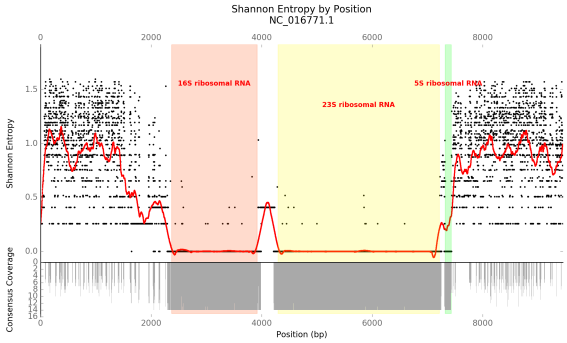
(c) *P. aeruginosa strain ATCC 15692* (NZ_CP017149.1)



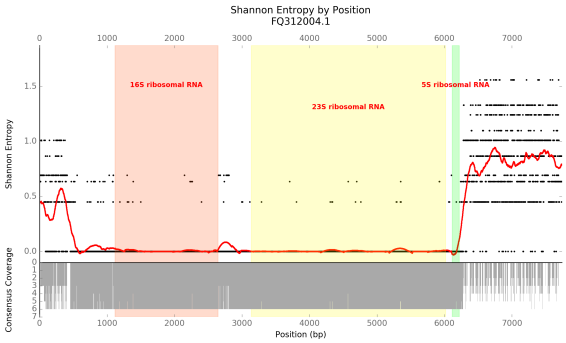
(d) *A. hydrophila ATCC 7966* (NC_008570.1)



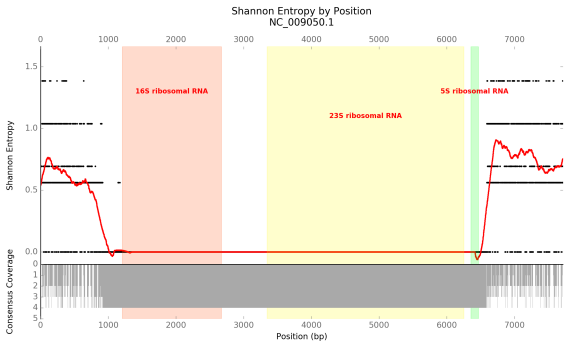
(e) *B. cereus ATCC 10987* (AE017194.1)



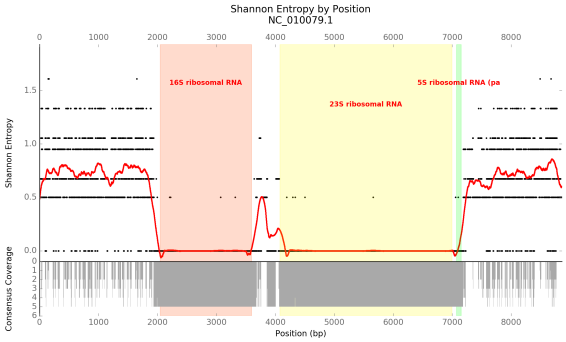
(f) *B. cereus NC7401* (NC_016771.1)



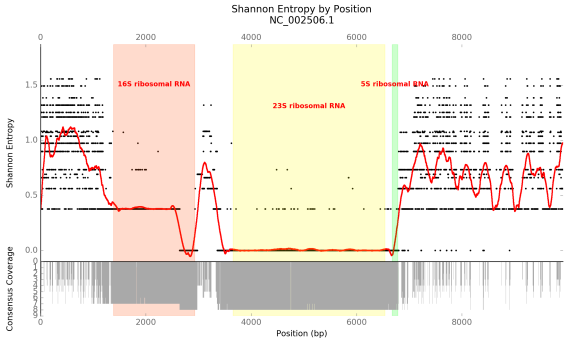
(g) *B. fragilis 638R* (FQ312004.1)



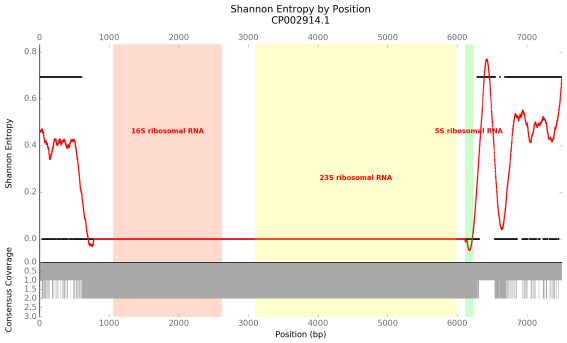
(h) *R. sphaeroides ATCC 17029* (NC_009049.1, NC_009050.1)



(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 S5: 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.