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: 2017/09/19 14:45:00

Extended Methods

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* (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.

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_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 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	0.7
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

```
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 S1: Pseudocode of riboSeed algorithm

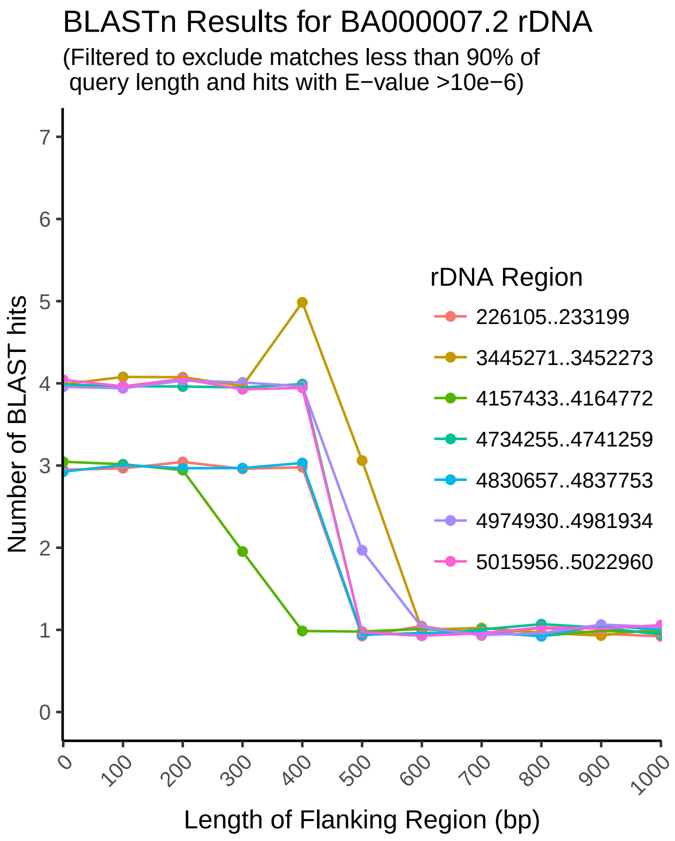


Figure S2: BLASTn was used to perform *in silico* DNA-DNA hybridization of all rDNA regions from *E. coli Sakai* with variable flanking lengths. The number of hits is a proxy for occurrences in the genome; increasing the flanking length increases the specificity. (Points are jittered to aide visibility for overlapping values.)

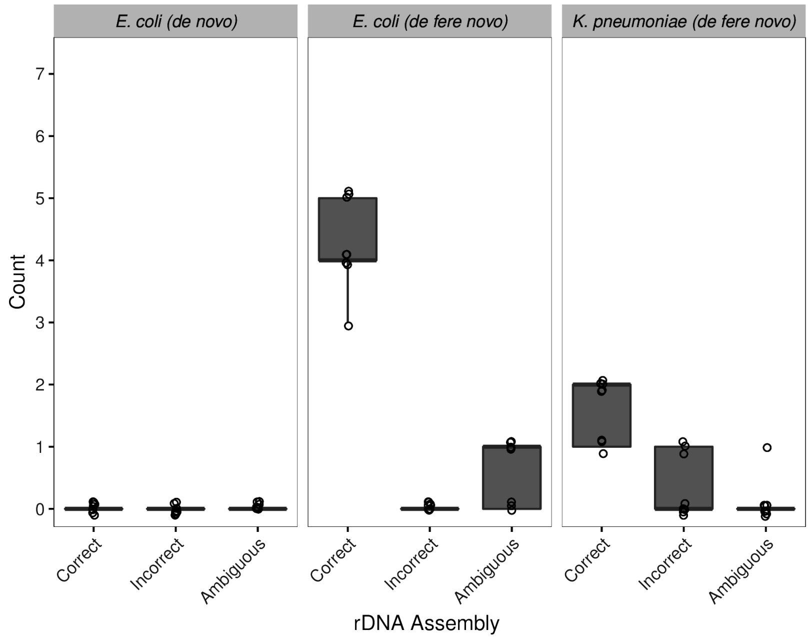
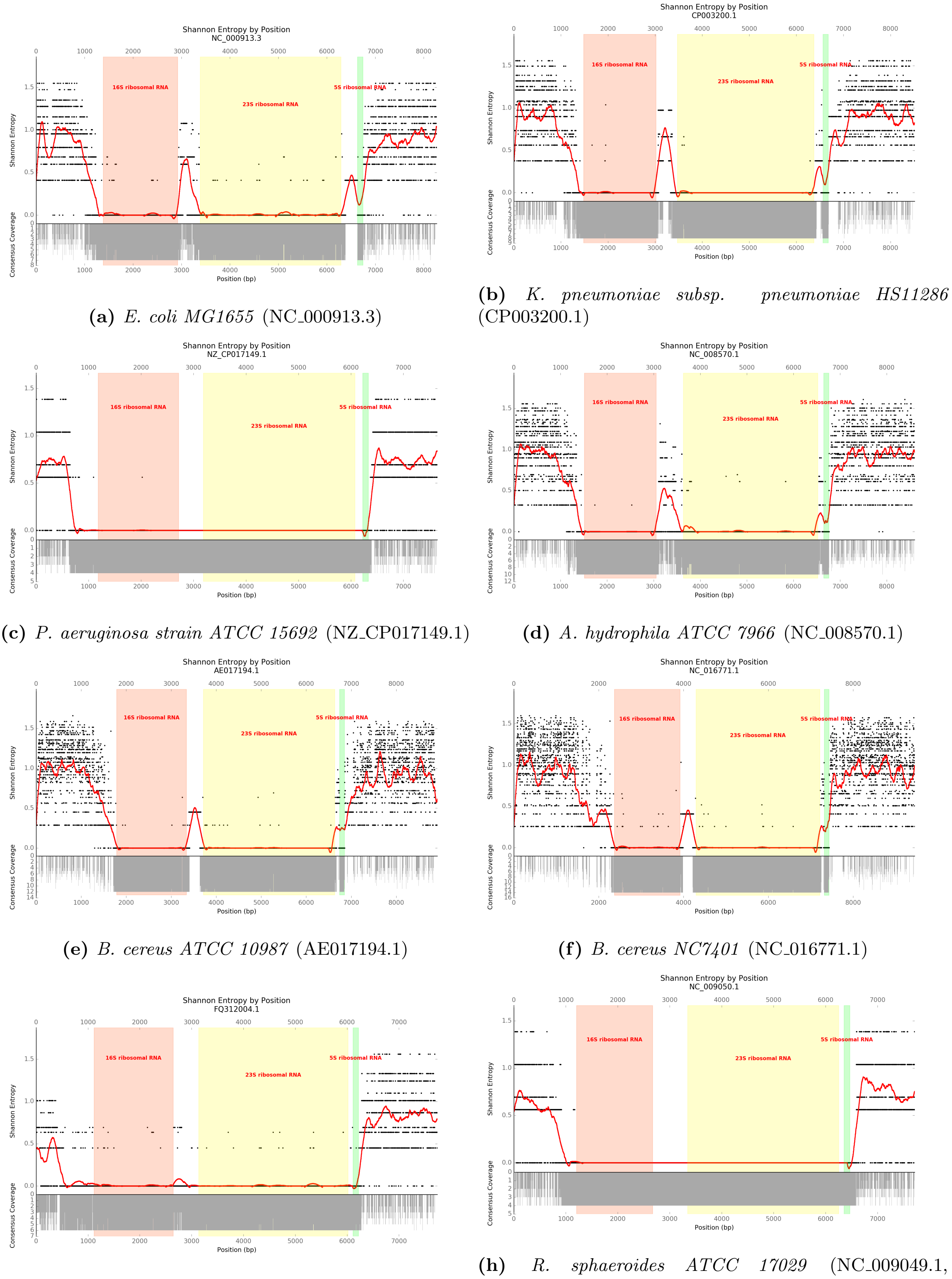


Figure S3: Assembly of artificial genome. *De fere novo* results in closure of 3-5 rDNAs with the correct reference; only 1-2 rDNAs are correctly assembled using *K. pneumoniae*. No rDNAs are assembled with *de novo* assembly. Scored with riboScore.py. N=8.



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

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

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

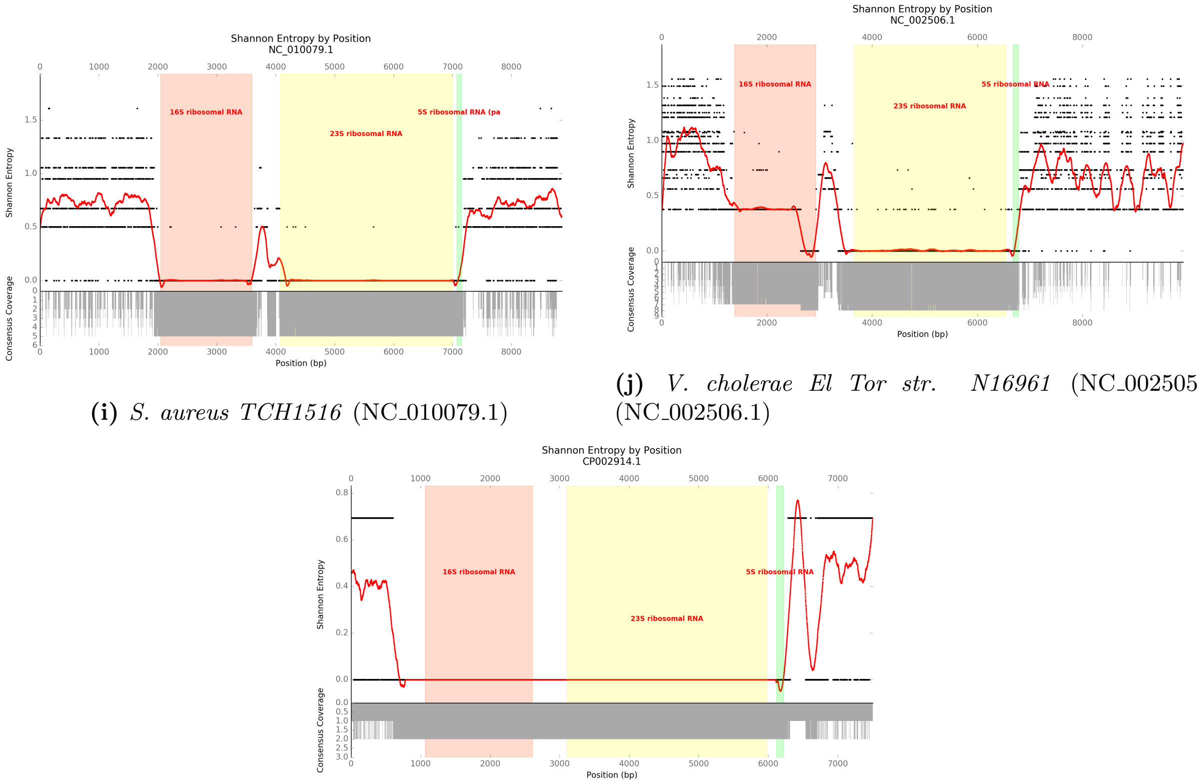
(d) *A. hydrophila ATCC 7966* (NC_008570.1)

(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 S4: 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.