Benchmarked programs

meta_rna version 3 was downloaded from:
`http://weizhong-lab.ucsd.edu/meta_rna/`

The required HMMER-3.0 was downloaded from:
`http://selab.janelia.org/software/hmmer3/3.0/`

BLAST version 2.2.25 was downloaded from:
`ftp://ftp.ncbi.nih.gov/blast/executables/release/LATEST/`

riboPicker version 0.4.2 was downloaded from:
`http://ribopicker.sourceforge.net`

Commands used to identify rRNA sequences

meta_rna3:

```
$ rna_hmm3.py -k "arc,bac" -L <dir> -i <file.fasta>
```

BLAST:

```
$ blastall -p blastn -m 8 -a 1 -d <db> -i <file.fasta>
```

riboPicker:

```
$ perl ribopicker.pl -no_seq_out -keep_tmp_files -z 3 -dbs <db> -
out_dir <dir> -f <file.fasta>
```

Fine-tuning the options based on the characteristics of the input data may yield better performance. The benchmarks were performed on a machine with two Intel® Xeon® X5650 6C 12T 2.66GHz processors.

Reference data

Simulated data

NCBI bacterial genomes were downloaded from (10/16/2011):
`ftp://ftp.ncbi.nlm.nih.gov/genomes/Bacteria/all.gbk.tar.gz`

Human Microbiome Project (HMP) bacterial genomes were downloaded from:
`http://downloads.hmpdacc.org/data/reference_genomes/all_gbk_20110912.tar.gz`

The bacterial genome sequences were split into rRNA sequences and non-rRNA sequences based on their annotations (given in the GBK files). Sequences were considered as rRNA if they were annotated with the Feature "rRNA" and with a product that suggested a 5S, 16S, or 23S sequence (annotations were corrected for typos and ambiguities).

The rRNA sequences were used to generate three datasets each with 5,000 randomly selected sequences of 200 bp, 500 bp and 1,000 bp length. The non-rRNA sequences were used to generate three datasets each with 45,000 randomly selected sequences of 200 bp, 500 bp and 1,000 bp length. Error rates of exactly 2% and 5% (with 15% indels and 85% substitutions) were then simulated for each of the datasets resulting in 27 datasets total (available at the SourceForge repository of riboPicker).

Real data

To assess the performance of riboPicker compared to BLASTn and meta_rna3, we used public datasets available from the NCBI SRA. Datasets were downloaded in SRA-lite format and converted to FASTA format using the NCBI SRA Toolkit available at:

`http://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?cmd=show&f=software&m=software&s=software`

The datasets were then preprocessed using PRINSEQ (Schmieder and Edwards, 2011) to remove short reads (most likely of low quality), read duplicates and reads with more than 1% of the ambiguous base N. The following command was used:

```
$ perl prinseq-lite.pl -log –verbose -derep 12345 -min_len 200 -
ns_max_p 1 -fasta <file.fasta> -out_good <file.out>
```

Table 1: Overview of the datasets used to compare the programs.

SRA ID	Type	Env.	#Raw Sequences	Mean Length	#Sequences filtered	Reference
SRR099254	MT	Marine	179,643	241.09	90,472 (50.36%)	(Ottesen et al., 2011)
SRR099253	MT	Marine	160,364	241.89	79,383 (49.50%)	(Ottesen et al., 2011)
SRR171683	MT	Marine	118,595	226.80	61,467 (51.83%)	(Ottesen et al., 2011)
SRR171682	MT	Marine	203,574	224.52	113,081 (55.55%)	(Ottesen et al., 2011)
SRR060835	MT	Human	55,017	243.45	22,154 (40.27%)	(Gosalbes et al., 2011)
SRR060836	MT	Human	46,375	297.81	26,204 (56.50%)	(Gosalbes et al., 2011)
SRR060837	MT	Human	32,993	216.66	11,773 (35.68%)	(Gosalbes et al., 2011)
SRR060838	MT	Human	18,548	188.43	4,016 (21.65%)	(Gosalbes et al., 2011)
SRR060839	MT	Human	16,529	223.75	5,515 (33.37%)	(Gosalbes et al., 2011)
SRR060840	MT	Human	19,656	209.35	6,426 (32.69%)	(Gosalbes et al., 2011)
SRR060841	MT	Human	65,658	274.25	32,649 (49.73%)	(Gosalbes et al., 2011)
SRR060842	MT	Human	61,138	308.15	34,345 (56.18%)	(Gosalbes et al., 2011)
SRR060843	MT	Human	30,836	312.28	18,174 (58.94%)	(Gosalbes et al., 2011)
SRR060844	MT	Human	26,583	290.87	14,880 (55.98%)	(Gosalbes et al., 2011)
SRR059252	MT	Marine	221,751	264.58	134,637 (60.72%)	(McCarren et al., 2010)
SRR059257	MT	Marine	230,376	263.24	136,884 (59.42%)	(McCarren et al., 2010)
SRR059258	MT	Marine	251,690	261.49	151,589 (60.23%)	(McCarren et al., 2010)
ERR016008	16S	Mouse	38,114	413.92	11,885 (31.18%)	-

Reference database

The sequences for the reference database were retrieved from the latest version (as of 10/16/2011) of SILVA (Pruesse et al., 2007), RDP (Cole et al., 2009), Greengenes (DeSantis et al., 2006), Rfam (Gardner et al., 2011), NCBI (Sayers et al., 2011), and HMP DACC (The NIH HMP Working Group et al., 2009). The sequences were combined into a single file and duplicates removed with PRINSEQ using the following command:

```
$ perl prinseq-lite.pl -log -verbose -derep 12345 -fasta <file.fasta> -
out_good rrnadb -out_bad null
```

The database was formatted for BLAST and riboPicker using the following commands:

```
$ formatdb -i rrnadb.fasta -p F -o T -n rrnadb -s T
$ bwa64 index -p rrnadb rrnadb.fasta
```

Results

The outputs of the programs were parsed to identify rRNA-like sequences using a coverage threshold of 50% and an identity threshold of 75% for BLASTn and riboPicker, and an e-value threshold of 10^{-5} for meta_rna3, as there is no information about the sequence similarity available.

Table 2: Results of the programs showing the number of sequences correctly identified as rRNA and non-rRNA sequence in the 27 simulated datasets.

Read length	Error rate	BLASTn		riboPicker		meta_rna3	
		rRNA	non-rRNA	rRNA	non-rRNA	rRNA	non-rRNA
200	0%	5,000	45,000	5,000	45,000	4,981	44,996
200	0%	5,000	45,000	5,000	45,000	4,975	44,999
200	0%	5,000	45,000	5,000	45,000	4,976	44,999
200	2%	5,000	45,000	4,995	45,000	4,979	44,996
200	2%	4,999	45,000	4,991	45,000	4,975	44,999
200	2%	5,000	45,000	4,993	45,000	4,974	44,999
200	5%	5,000	45,000	4,983	45,000	4,977	44,995
200	5%	5,000	45,000	4,989	45,000	4,973	44,998
200	5%	5,000	45,000	4,988	45,000	4,974	45,000
350	0%	5,000	45,000	5,000	45,000	4,982	44,999
350	0%	5,000	45,000	5,000	45,000	4,985	44,996
350	0%	5,000	45,000	5,000	45,000	4,982	44,997
350	2%	5,000	45,000	4,991	45,000	4,982	44,999
350	2%	5,000	45,000	4,992	45,000	4,985	44,998
350	2%	5,000	45,000	4,992	45,000	4,982	44,999
350	5%	5,000	45,000	4,992	45,000	4,981	44,999
350	5%	5,000	45,000	4,991	45,000	4,985	44,999
350	5%	5,000	45,000	4,985	45,000	4,982	45,000
500	0%	5,000	45,000	5,000	45,000	4,985	44,996
500	0%	5,000	45,000	5,000	45,000	4,988	44,998
500	0%	5,000	45,000	5,000	45,000	4,981	44,999
500	2%	4,999	45,000	4,990	45,000	4,985	44,997
500	2%	5,000	45,000	4,997	45,000	4,988	44,998
500	2%	5,000	45,000	4,993	45,000	4,981	44,999
500	5%	5,000	45,000	4,986	45,000	4,985	44,998
500	5%	4,999	45,000	4,987	45,000	4,988	44,999
500	5%	5,000	45,000	4,993	45,000	4,980	45,000

Table 3: Prediction sensitivity and accuracy (in percentage) of the three programs for the 27 simulated datasets.

Read length	Error rate	Sensitivity			Accuracy		
		BLASTn	riboPicker	meta_rna3	BLASTn	riboPicker	meta_rna3
200	0%	100.00	100.00	99.62	100.00	100.00	99.95
200	0%	100.00	100.00	99.50	100.00	100.00	99.95
200	0%	100.00	100.00	99.52	100.00	100.00	99.95
200	2%	100.00	99.90	99.58	100.00	99.99	99.95
200	2%	99.98	99.82	99.50	100.00	99.98	99.95
200	2%	100.00	99.86	99.48	100.00	99.99	99.95
200	5%	100.00	99.66	99.54	100.00	99.97	99.94
200	5%	100.00	99.78	99.46	100.00	99.98	99.94
200	5%	100.00	99.76	99.48	100.00	99.98	99.95
350	0%	100.00	100.00	99.64	100.00	100.00	99.96
350	0%	100.00	100.00	99.70	100.00	100.00	99.96
350	0%	100.00	100.00	99.64	100.00	100.00	99.96
350	2%	100.00	99.82	99.64	100.00	99.98	99.96
350	2%	100.00	99.84	99.70	100.00	99.98	99.97
350	2%	100.00	99.84	99.64	100.00	99.98	99.96
350	5%	100.00	99.84	99.62	100.00	99.98	99.96
350	5%	100.00	99.82	99.70	100.00	99.98	99.97
350	5%	100.00	99.70	99.64	100.00	99.97	99.96
500	0%	100.00	100.00	99.70	100.00	100.00	99.96
500	0%	100.00	100.00	99.76	100.00	100.00	99.97
500	0%	100.00	100.00	99.62	100.00	100.00	99.96
500	2%	99.98	99.80	99.70	100.00	99.98	99.96
500	2%	100.00	99.94	99.76	100.00	99.99	99.97
500	2%	100.00	99.86	99.62	100.00	99.99	99.96
500	5%	100.00	99.72	99.70	100.00	99.97	99.97
500	5%	99.98	99.74	99.76	100.00	99.97	99.97
500	5%	100.00	99.86	99.60	100.00	99.99	99.96

Table 4: Results of the programs BLASTn (B), riboPicker (R), and meta_rna3 (M) showing the number of sequences identified as rRNA-like in the publicly available real datasets. All numbers are exclusive. Column "Any" shows the number of identified rRNA-like sequences across all three programs.

SRA ID	B	R	M	B+R	B+M	R+M	B+R+M	Any
SRR059252	241	10	1,606	1,054	851	29	120,212	124,003
SRR059257	281	9	1,299	961	847	28	121,236	124,661
SRR059258	54	1	1,173	206	594	16	136,003	138,047
SRR060835	0	0	558	6	125	3	21,440	22,132
SRR060836	0	0	496	15	43	14	25,624	26,192
SRR060837	1	0	361	22	179	2	11,194	11,759
SRR060838	0	0	124	2	45	2	3,835	4,008
SRR060839	0	0	109	4	28	0	5,371	5,512
SRR060840	0	0	192	17	82	4	6,128	6,423
SRR060841	1	0	792	39	354	5	31,432	32,623
SRR060842	1	0	835	43	261	15	33,156	34,311
SRR060843	0	0	902	5	721	11	16,519	18,158
SRR060844	0	0	587	4	257	8	14,010	14,866
SRR099253	615	52	1,046	3,550	496	20	53,420	59,199
SRR099254	842	61	1,478	4,825	655	18	57,444	65,323
SRR171682	741	7	1,832	3,255	478	21	36,489	42,823
SRR171683	307	3	886	1,723	284	15	18,446	21,664
ERR016008	0	0	1	0	1	0	11,883	11,885

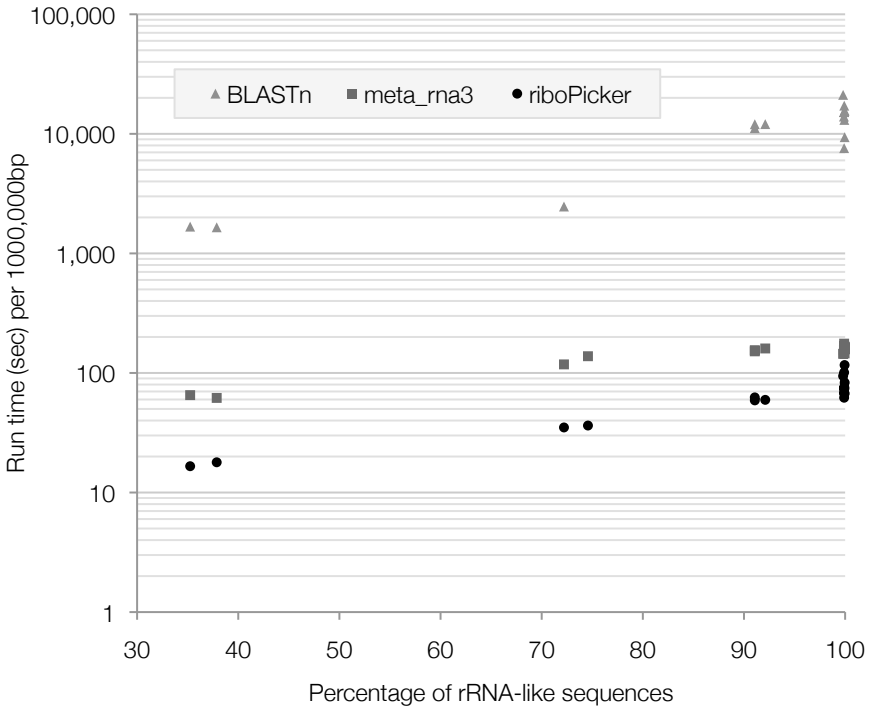

Figure 1: Graph showing the required time to process 10^6 bp of metatranscriptomic sequence data against the percentage of rRNA-like sequences identified in each dataset. The data shows that riboPicker processes the real datasets more than twice as fast as meta_rna3 and more than 100 times faster than BLASTn.

References

Cole,J.R. et al. (2009) The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. *Nucleic Acids Res.*, **37**, D141-145.

DeSantis,T.Z. et al. (2006) Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl. Environ. Microbiol.*, **72**, 5069-5072.

Gardner,P.P. et al. (2011) Rfam: Wikipedia, clans and the "decimal" release. *Nucleic Acids Res.*, **39**, D141-145.

Gosalbes,M.J. et al. (2011) Metatranscriptomic approach to analyze the functional human gut microbiota. *PLoS ONE*, **6**, e17447.

McCarren,J. et al. (2010) Microbial community transcriptomes reveal microbes and metabolic pathways associated with dissolved organic matter turnover in the sea. *Proc. Natl. Acad. Sci. U.S.A.*, **107**, 16420-16427.

Ottesen,E.A. et al. (2011) Metatranscriptomic analysis of autonomously collected and preserved marine bacterioplankton. *The ISME Journal*.

The NIH HMP Working Group et al. (2009) The NIH Human Microbiome Project. *Genome Res.*, **19**, 2317-2323.

Pruesse,E. et al. (2007) SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Res*, **35**, 7188-7196.

Sayers,E.W. et al. (2011) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **39**, D38-51.

Schmieder,R. and Edwards,R. (2011) Quality control and preprocessing of metagenomic datasets. *Bioinformatics*, **27**, 863-864.