Supplementary Information for "Unusual Transcripts in Procaryotic Transcriptome Sequencing Data"

Gero Doose, Maria Alexisa, Rebecca Kirsch, Sven Findeiß,
David Langenberger, Rainer Machné, Steve Hoffmann,
Mario Mörl and Peter F. Stadler

May 6, 2013

Table S1: Number of annotation items for each species. From NCBI the annotation of CDS, tRNAs and rRNAs genes are used. The "other genes" class contains as gene annotated protein fragments, frame shift regions, ncRNAs and pseudo-genes and has not been used for annotation overlaps. To get a comprehensive set of regulatory non-coding regions the annotation of ncRNAs, regulatory elements and autocatalytically spliced introns for each species are taken from the Rfam database. Numbers in brackets indicate the amount of splited CDSs and tRNAs if annotated in a certain species.

Species	**Accession**	**NCBI**				**Rfam**		
		# CDS genes	**# tRNAs genes**	**# rRNAs genes**	**# other genes**	**# ncRNAS**	**# regul. elements**	**# auto.spl. introns**
Bacteria								
B. cereus	NC_003909	5605 (6)	98	36	36	10	114	8
	NC_005707	241	0	0	1	0	0	0
E. coli	NC_000913	4378 (111)	89	22	65	101	47	0
H. pylori	NC_000915	1577 (8)	36	7	11	10	2	0
P. aeruginosa PA14	NC_008463	5892	59	13	13	81	29	0
S. enterica	NC_016856	5323	86	22	89	143	46	0
	NC_016855	101	0	0	2	0	0	0
Synechocystis	NC_000911	3179	43	6	1	11	6	0
	NC_005229	132	0	0	0	0	0	0
	NC_005230	105	0	0	0	0	0	0
	NC_005231	49	0	0	0	0	0	0
	NC_005232	110	0	0	0	0	0	0
Archaea								
N. equitans	NC_005213	540	50 (16)	3	0	0	0	0
I. hospitalis	NC_009776	1434	47	4	10	2	0	0
P. furiosus	NC_003413	2127 (10)	46	4	53	271	1	0
S. solfataricus	NC_002754	2978	46	4	6	211	1	0

Table S2: Group I intron candidates as detected by our infernal search based on splitting of the Rfam seed alignment (RF00028) 14 alignments of intron subtypes from the GISSD database (`http://www.rna.whu.edu.cn/gissd/`, [**?**]). Host gene annotations were manually collected from genome annotation files and the respective publications (**Publ.**). [a]: IStrons, fusions of an intron with an insertion element, including a IS605 transposase ORF, described in [**?**].

Coordinates			**Host gene**	**Publ.**	**# split reads**
			B. cereus		
1,109,803	1,109,454	-	ADH	[**?**][a]	0
1,462,483	1,462,845	+	nrdE	[**?**]	0
2,995,548	2,995,210	-		[**?**][a]	0
3,235,296	3,234,947	-	Bcr/CflA	[**?**][a]	0
3,563,680	3,563,421	-	recA	[**?**, **?**]	**9**
4,248,229	4,247,901	-			0
			Synechocystis		
2,791,491	2,791,707	+	fMet-tRNA	[**?**]	**39,014**
			P. aeruginosa PA14		
5,750,048	5,755,197	+			0

Table S3: Group II intron candidates identified by blasting known intron sequences (**DB ID**) from strains of the respective species deposited at the group II intron database (`http://webapps2.ucalgary.ca/~groupii/`, [**?**]). Host gene annotations were manually collected from genome annotation files and the respective publications (**Publ.**). [a]: 506 nucleotides at the 3' end of the query intron sequence were not present in the blast hit. [b]: these introns were also featured in the Rfam annotation as gII intron families RF00029 and RF02004.

DB ID	**Coordinates**			**Host gene**	**Publ.**	**# split reads**
				E. coli		
E.c.I4[a]	271,415	273,178	+			0
n.a.[b]	4,499,355	4,499,216	-			0
				B. cereus		
B.c.I1[b]	3,134,418	3,132,060	-	intergenic	[**?**]	0
B.c.I2[b]	3,446,987	3,444,106	-	unknown transcript	[**?**]	0
B.c.I3[b]	3,606,193	3,603,527	-	DnaPOLIII s.u.α	[**?**]	**21**
				B. cereus, plasmid pBc10987		
B.c.I1	77,150	74,792	-	intergenic	[**?**]	0
B.c.I2	109,197	112,078	+	hypo. prot.	[**?**]	0
B.c.I4	35,608	32,766	-	hypo. prot.	[**?**]	**1**
B.c.I5	84,166	86,938	+	prophage protein	[**?**]	**1**
				P. aeruginosa PA14		
P.ae.I2[b]	643,335	640,489	+			0

Table S4: tRNA in Archaea with circularized introns.

Species	**tRNA**	**intron**	**read support**
Sulfolobus solfataricus	Trp-CCA	72,767-72,831	374
Pyrococcus furiosus	Leu-TAA	1,443,658-1,443,696	1
Pyrococcus furiosus	Trp-CCA	937,405-937,475	22
Nanoarchaeum equitans	Gly-TCC	54,080-54,150	1
Nanoarchaeum equitans	Leu-TAA	278,944-279,011	1
Nanoarchaeum equitans	Met-CAT	327,385-327,472	1
Nanoarchaeum equitans	Met-CAT	327,400-327,464	251
Ignicoccus hospitalis	Leu-GAG	1,219,487-1,219,534	1
Ignicoccus hospitalis	Met-CAT	868,198-868,268	1
Ignicoccus hospitalis	Asn-GTT	881,933-882,008	3
Ignicoccus hospitalis	Tyr-GTA	169,603-169,696	1
Ignicoccus hospitalis	Tyr-GTA	169,603-169,672	2
Ignicoccus hospitalis	Trp-CCA	1,229,063-1,229,121	1
Ignicoccus hospitalis	Trp-CCA	1,229,027-1,229,107	1
Ignicoccus hospitalis	Trp-CCA	1,229,037-1,229,086	3
Ignicoccus hospitalis	Trp-CCA	1,229,037-1,229,097	3
Ignicoccus hospitalis	Trp-CCA	1,229,048-1,229,107	7
Ignicoccus hospitalis	Trp-CCA	1,229,037-1,229,107	421

Table S5: Overview

species	experiment	ref. genome	# input reads	# mappable reads	% mappable reads
		Eubacteria			
Bacillus cereus	ERR031734	NC_003909	15,498,220	15,264,233	98.49%
Escherichia coli	SRR441585	NC_000913	52,515,346	44,429,568	84.60%
Salmonella enterica	SRR066805	NC_016856	31,924,568	27,752,771	86.93%
Pseudomonas PA14	SRR363800-SRR363807	NC_008463	78,141,620	65,573,260	83.92%
Helicobacter pylori 26695	SRA010186	NC_000915	82,847,902	40,152,294	48.47%
Synechocystis PCC6803	PCC6803 2012	NC_000911	31,985,927	15,080,656	47.15%
		Archaea			
Nanoarchaeum equitans / *Ignicoccus hospitalis*	SRR514574-SRR514578	NC_005213 NC_009776	17,253,447	11,173,688 5,302,517	94.95%
Pyrococcus furiosus	SRR032433, SRR358786-SRR358787	NC_003413	16,449,461	8,691,213	52.84%
Sulfolobus solfataricus	SRR030761-SRR030768	NC_002754	17,356,356	11,965,214	68.94%

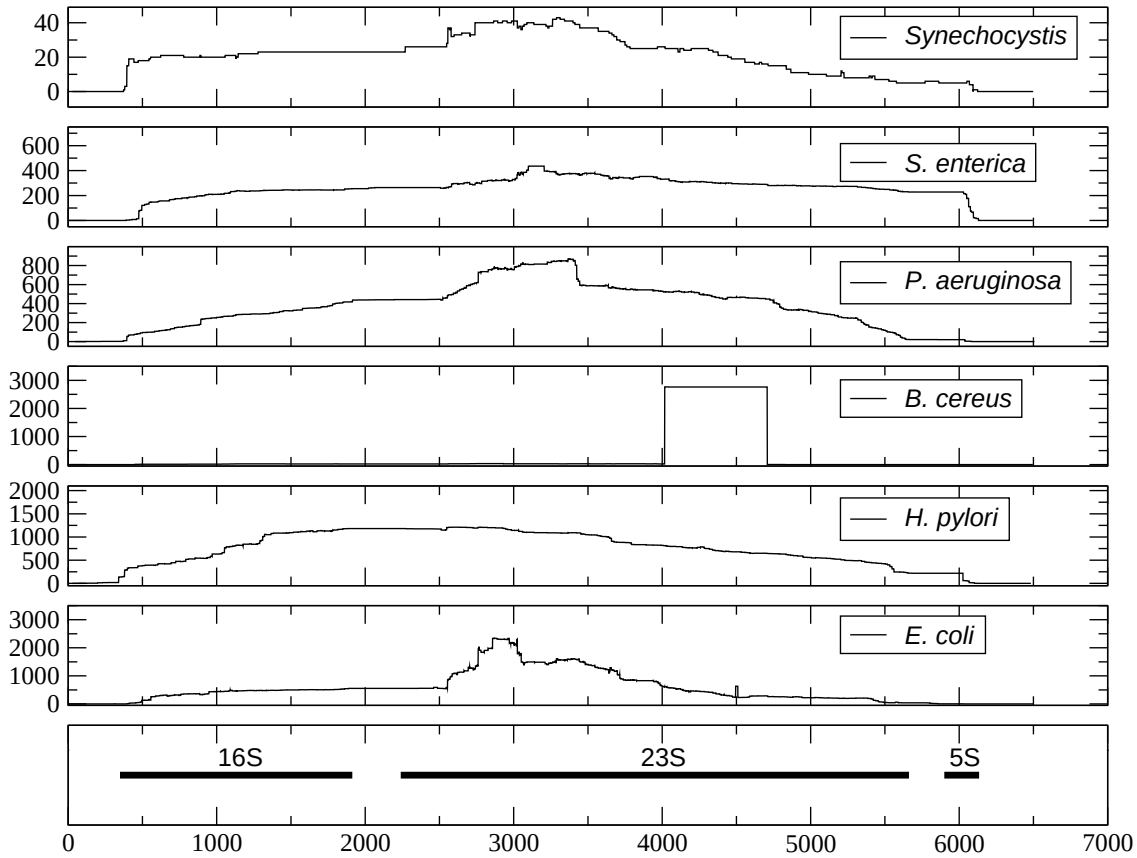
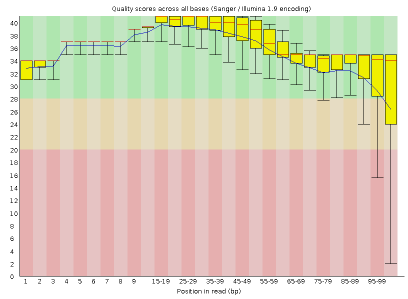
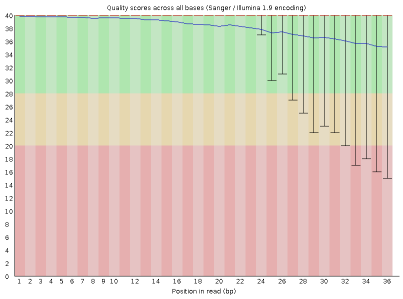


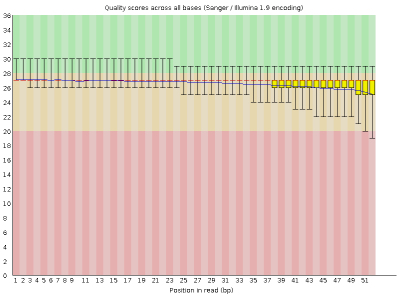
Figure S1: Density of "spliced" reads along ribosomal rRNA operons. Coordinates refer to a multiple sequence alignment of reference operons of the six bacterial species. With the exception of *Synechocystis* rRNAs are the dominating regions from which unusual RNA transcripts arise, see main text for details. Despite the fact of the huge amount of these transcripts no systematic "splicing" events, neither within species nor across species, are found.



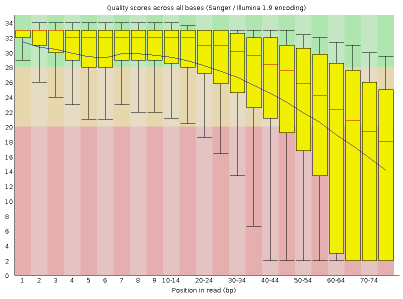
Synechocystis



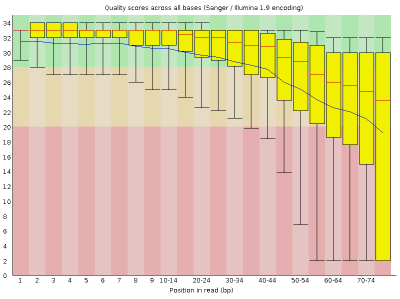
Sulfolobus sulfataricus



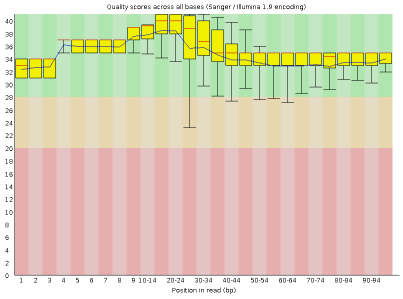
Bacillus cereus



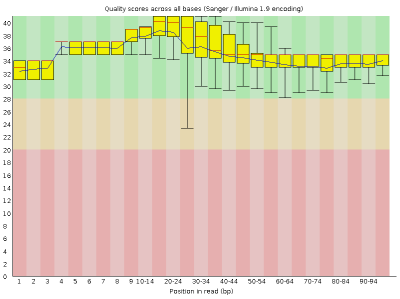
Escherichia coli



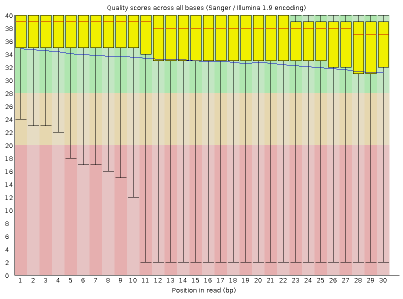
Helicobacter pylori



Ignicoccus hospitalis



Nanoarchaeum equitans



Pseudomonas PA14

Figure S2: Quality box plots for mapped reads for all used RNA-seq libraries. Adapter clipped and quality trimmed reads were mapped to the corresponding reference genomes and quality box plots were generated.