Supplementary Materials:
Comparative transcriptomics across the prokaryotic tree of life

Ofir Cohen [1,2], Shany Doron [1], Omri Wurtzel [1,3], Daniel Dar [1], Sarit Edelheit [1], Iris Karunker, [1] Eran Mick [1,4], and Rotem Sorek [1,§]

1. Department of Molecular Genetics, Weizmann Institute of Science, Rehovot, 76100, Israel
2. Broad Institute of Harvard and MIT, Cambridge, MA, USA
3. Whitehead Institute for Biomedical Research, Cambridge, MA, USA
4. Department of Systems Biology, Harvard Medical School, Boston, Massachusetts, USA

[§] Corresponding author: rotem.sorek@weizmann.ac.il

Supplementary Tables

Table S1 – Transcriptomes in the study

A quantitative summary of raw transcriptome sequencing data for each organism and each condition. Values represent the number of uniquely mapped reads (in millions).

Organism	Accession	Phylum	Growth conditions	RNA-seq M reads	5' TAP M reads	5' untreated M reads	Reference
Bacillus subtilis	*NC_000964*	Firmicutes	Terrific Broth, mid-log phase	2.26	0.77	0.35	This study
Desulfovibrio vulgaris	*NC_002937*	Proteobacteria	DSMZ medium	3.70	1.17	4.03	(1)
Escherichia coli	*NC_000913*	Proteobacteria	mid-log phase	3.83	0.37	0.17	This study
Listeria monocytogenes	*NC_003210*	Firmicutes	merged	6.11	1.78	3.75	(2)
			log phase, 30°C	NA*	0.15	0.93	(2)
			log phase, 37°C	NA*	0.41	0.35	(2)
			hypoxia	NA*	0.44	0.57	(2)
			delta *prfA*	NA*	0.20	0.31	(2)
			delta *sigB*	NA*	0.41	0.73	(2)
			stationary phase, 37°C	NA*	0.16	0.87	(2)
Bdellovibrio bacteriovorus	*NC_005363*	Proteobacteria	merged	9.94	5.95	6.60	(3)
			HEPES buffer, attack phase	8.80	5.95	6.60	(3)
			HEPES buffer, growth phase	1.14	0.01	0.00	(3)
Catenulispora acidiphila	*NC_013131*	Actinobacteria	DSMZ medium	3.03	2.10	0.42	(1)
Clostridium acetobutylicum	*NC_003030*	Firmicutes	merged	7.67	0.25	0.36	This study
			pH 4.5	4.10	0.19	0.32	This study
			pH 5.7	3.57	0.06	0.04	This study
Gluconobacter oxydans	*NC_006677*	Proteobacteria	merged	3.02	6.10	1.97	This study
			glycerol	1.41	1.57	0.51	This study
			mannitol	1.61	4.53	1.45	This study

Kangiella koreensis	*NC_013166*	Proteobacteria	DSMZ medium	2.81	0.37	1.31	(1)
Lactobacillus brevis	*NC_008497*	Firmicutes	ATCC medium	10.65	0.95	0.62	(1)
Lactococcus lactis	*NC_002662*	Firmicutes	M17 medium	4.82	0.62	0.35	(1)
Pseudomonas aeruginosa	*NC_008463*	Proteobacteria	merged	5.00	1.53	0.95	(4)
			28°C, LB, early stationary phase	2.13	1.30	0.82	(4)
			37°C, LB, early stationary phase	2.87	0.23	0.13	(4)
Synechococcus WH7803	*NC_009481*	Cyanobacteria	Artificial seawater medium	5.53	3.11	0.65	(5)
Synechococcus WH8102	*NC_005070*	Cyanobacteria	Artificial seawater medium	7.74	4.58	2.50	(5)
Spirochaeta aurantia	*Saur_Contig1177*	Spirochaetes	DSMZ medium	5.71	12.16	2.27	(1)
Sulfolobus acidocaldarius	*NC_007181*	Crenarchaeotes	yeast extract, stationary phase	2.97	0.98	2.73	This study
Sulfolobus solfataricus	*NC_002754*	Crenarchaeotes	merged	14.90	4.89	4.13	(6)
			cellobiose	7.67	2.26	2.91	(6)
			glucose	4.66	2.26	2.91	(6)
			minimal	3.01	2.26	2.91	(6)
Thermus thermophilus	*NC_005835*	Deinococcus-Thermus	mid-log phase	1.51	2.26	2.91	This study

* In *Listeria*, RNA-seq (coverage) exists only for the "merged" growth condition. In the different growth conditions in *Listeria*, inference of TSS is based on growth-condition specific 5' reads but a single "merged" coverage.

Table S2 – Features used for TSS inference

Genomic and transcriptomic features used for the inference of TSSs based on a random forest classifier

Feature name	Feature details	Feature Type
Distance	Distance of site from annotated gene	Genomic
TAP	Number of TAP treated 5’ reads	5' RNA-seq
Ratio	Ratio of TAP treated vs. untreated 5' reads	5' differential RNA-seq
Diff	Number of TAP treated +1 divided by untreated +1	5' differential RNA-seq
Δ Coverage short	Expression increase within 20 bp (log)	RNA-seq
Δ Coverage long	Expression increase within 80 bp (log)	RNA-seq
Coverage ratio	Ratio of average coverage 80 bp downstream and 40 bp upstream (log)	RNA-seq
Coverage downstream	Average coverage 80 bp downstream	RNA-seq
Alternative sites in vicinity (Diff)	Number of adjacent sites with higher "Diff" value in 100 bp vicinity	5' differential RNA-seq
Alternative sites in vicinity (TAP)	Number of adjacent sites with higher "TAP" value in 100 bp vicinity	5' differential RNA-seq
Alternative sites in vicinity	Number of adjacent sites with both higher "Diff" and "TAP" value in 100 bp vicinity	5' differential RNA-seq
Processing value (Ratio)	Ratio of TAP treated vs. the sum of all untreated 5' reads up to 100 bp upstream	5' differential RNA-seq
Processing value (Diff)	TAP treated minus the sum of all untreated 5' reads up to 100 bp upstream	5' differential RNA-seq
Coverage compatibility with gene	Ratio between coverage at TSS and min coverage till gene's end (40 bp slide window)	RNA-seq + Genomic
Is overlapping other gene	True, if putative upstream TSS is located within adjacent gene	Genomic

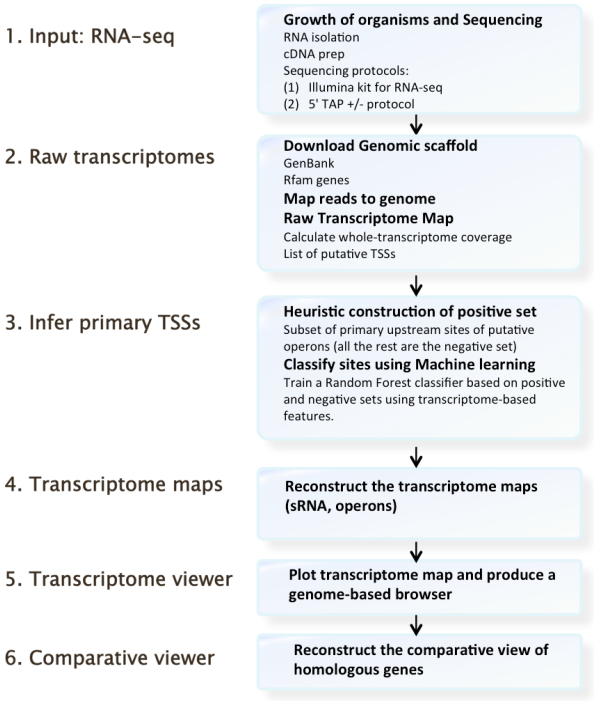
Table S3 – gene families with recurring long 5'UTRs

Gene families (COGs) with recurring propensity for recurring long 5'UTRs across organisms. Gene families with 5'UTRs of at least 100 bp found within least 5 different genera are presented.

COG ID	Type	Genes description	# Rfam ID or reference	# Genera with long 5'UTRs (# reported in RFAM)
COG0052	Ribosomal Leader	30S ribosomal subunit protein S2	RF00127	11 (3)
COG0244	Ribosomal Leader	50S ribosomal subunit protein L10	RF00557	10 (3)
COG0441	T-box	threonyl-tRNA synthetase	RF00230	10 (1)
COG0290	Ribosomal Leader	protein chain initiation factor IF-3	RF00558	9 (3)
COG0102	Ribosomal Leader	50S ribosomal subunit protein L13	RF00555	9 (3)
COG0085	Lacto-rpoB	RNA polymerase, beta subunit	RF01709	9 (2)
COG1158	Pseudomon-Rho	transcription termination factor Rho	RF01720	8 (1)
COG0048	Ribosomal Leader	30S ribosomal subunit protein S12	(7)	8
COG1278	Thermometer	cold-shock csp-family	RF01766	8 (8)
COG0261	Ribosomal Leader	50S ribosomal subunit protein L21	RF00559	7 (3)
COG0513	Thermometer	DEAD/DEAH box RNA helicase	(8)	7 (1)
COG0192	S-box	S-adenosylmethionine synthetase	RF00162	7 (2)
COG0209	Riboswitch, Cobalamin	vitamin B12-dependent ribonucleotide reductase	RF00174	7 (1)
COG0051	Ribosomal Leader	30S ribosomal subunit protein S10	(9)	7 (1)
COG0422	Riboswitch, TPP	thiamine biosynthesis protein ThiC	RF00059	7 (6)
COG0776	Novel candidates	HU, DNA-binding transcriptional regulator	NA	7
COG1271	Novel candidates	Cytochrome bd-type quinol oxidase, subunit 1	NA	6
COG0550	Novel candidates	DNA topoisomerase	NA	6
COG0539	Ribosomal Leader	30S ribosomal subunit protein S1	(10)	6 (1)
COG0525	T-box	Valyl-tRNA synthetase	RF00230	6 (3)
COG0060	T-box	Isoleucyl-tRNA synthetase	RF00230	6 (2)
COG0495	T-box	Leucyl-tRNA synthetase	RF00230	6 (2)
COG1622	Novel candidates	cytochrome o ubiquinol oxidase subunit II	NA	6
COG0833	Riboswitch, Lysine	lysine transporter (permease)	RF00168	6 (2)
COG0119	Leucine operon leader (leuA)	Isopropylmalate/homocitrate/citramalate synthases	RF00512	5 (1)
COG0838	mini-ykkC	NADH:ubiquinone oxidoreductase	RF01068	5 (1)
COG0117	Riboswitch, FMN	riboflavin biosynthesis protein	RF0005	5 (2)
COG0449	glmS ribozyme	glucosamine--fructose-6-phosphate	RF00234	5 (1)
COG0404	Riboswitch, Glycine	Glycine cleavage system T protein (aminomethyltransferase)	RF00504	5 (2)
COG0752	T-box	Glycyl-tRNA synthetase, alpha subunit	RF00230	5 (1)
COG0234	Pseudomon-GroES	Co-chaperonin GroES (HSP10)	RF01721	5 (1)
COG0522	Ribosomal Leader, candidate	30S ribosomal subunit protein S4	NA	5
COG1544	Ribosomal Leader, candidate	sigma 54 modulation protein/30S ribosomal	NA	5
COG1418	Novel candidates	Predicted HD superfamily hydrolase	NA	5
COG0227	Ribosomal Leader, candidate	50S ribosomal subunit protein L28	NA	5
COG0605	Novel candidates	superoxide dismutase (Fe, Mn)	NA	5
COG0568	Novel candidates	RNA polymerase, sigma 70 (rpoD)	NA	5
COG0016	T-box	Phenylalanyl-tRNA synthetase alpha subunit	RF00230, RF01859	5 (3,1)
COG0779	rimP leader	ribosome maturation factor for 30S subunits	(11)	5 (1)
COG2252	Riboswitch, Purine	xanthine/uracil/vitamin C permease, membrane transporter	RF00167	5 (2)
COG1077	Novel candidates	rod shape-determining protein MreB	NA	5

Supplementary Results and Figures

Figure S1 – The transcriptome maps reconstruction pipeline



Accuracy of TSS inference

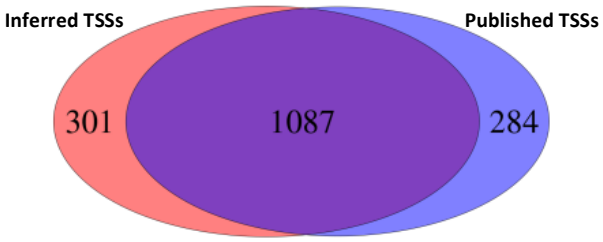
Our compendium of organisms includes three organisms with published transcriptomes maps including manually curated single-base resolution TSSs (*Listeria monocytogenes* (2)*, Pseudomonas aeruginosa* (4) and *Sulfolobus solfataricus* (6)). These published sets were used as benchmarks to evaluate the accuracy of our automatic approach (tolerance of 1 bp in position was allowed). Assuming that the published TSSs represent the true set of TSSs, we considered a site to be "True Positive" (TP) if the position, strand, and association of TSS to genes were the same as in the published set, "False Positive" (FP) if a TSS was inferred but is missing from the published set, and "False Negative" (FN) if a TSS appears in the published set but was not inferred by the automatic approach (Figure S2). We find strong agreement between the automatically inferred TSSs and the published sets. In *Listeria monocytogenes*, out of 1,371 published gTSSs we infer 1,087 (sensitivity of 79.3%) while a total of 1,388 gTSSs were inferred (precision of 78.3%). The agreement with the published sets of *Pseudomonas* and *Sulfolobus* was lower with sensitivity of 74.5% and 64.1% while maintaining precision of 70.9% and 61.6%, respectively. In all three benchmarks the False Positive Rate at the cited sensitivity was extremely low, rejecting tens of thousands of processed sites, with FPR of 0.0000035%, 0.000054%, and 0.000026% for *Listeria, Pseudomonas* and *Sulfolobus,* respectively.

Importantly, further manual inspection of the sites in which disagreement with the published set was found, suggests that only a subset of the cases deemed "false" are the result of inaccurate inference by the automatic algorithm. Manual inspection of 50 sites that are inferred by our method but missing from the published set ("FP") revealed that most of them are justified TSSs ("TP", 52%, 68%, and 96%, for *Listeria*, *Pseudomonas*, and *Sulfolobus,* respectively), while only a small fraction were wrongly inferred (16%, 8%, and none, for *Listeria*, *Pseudomonas*, and *Sulfolobus,* respectively), with the rest deemed inconclusive. Similarly, out of 50 sites that appear in the published set and were not inferred by the computational pipeline ("FN"), many are indeed not justified primary TSS ("TN", 80%, 34%, and 50%, for *Listeria*, *Pseudomonas*, and *Sulfolobus*), while only a few sites (2%, 14%, and 14%, for *Listeria*, *Pseudomonas*, and *Sulfolobus*) were clearly missing from inference, with the rest inconclusive.

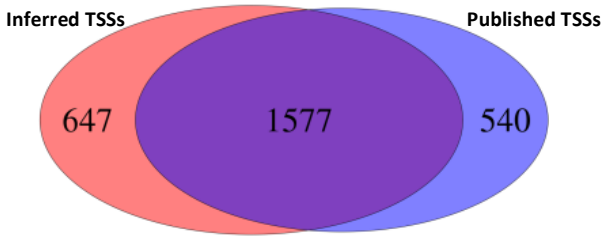
Figure S2 – Accuracy of TSS inference.

Accuracy of the automated inference of TSSs is estimated by comparison to the curated/manually annotated gTSSs of the published transcriptomes. The TP (agreement with published, center), FP (inferred but missing from published, left area) and FN (appear in published but not inferred, right area) are shown in purple, red and blue, respectively. (A) *Listeria*: sensitivity of 79.3% with precision of 78.3% (B) *Pseudomonas*: sensitivity of 74.5% with precision of 70.9% (C) *Sulfolobus*: sensitivity of 64.1% with precision of 61.6%.

(A) Listeria monocytogenes



(B) Pseudomonas aeruginosa



(C) Sulfolobus solfataricus

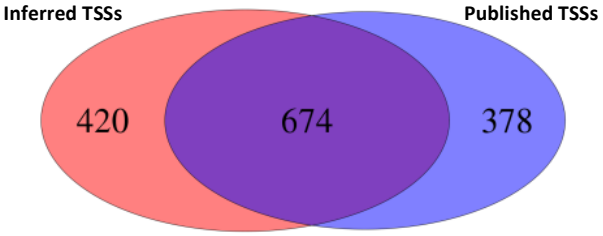
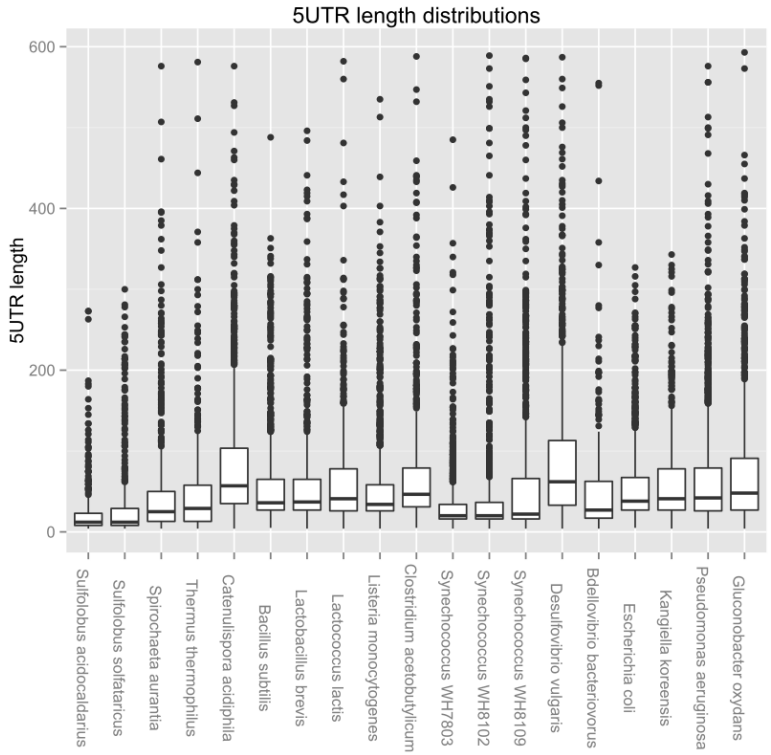


Figure S3 – distribution of 5'UTR length across the tree of life

(A) Comparable collection of 5'UTRs across the analyzed organisms (B) Percent of long 5'UTRs (at least 100 bp) per organism, across the tree of life, out of total genes.

(A)



(B)

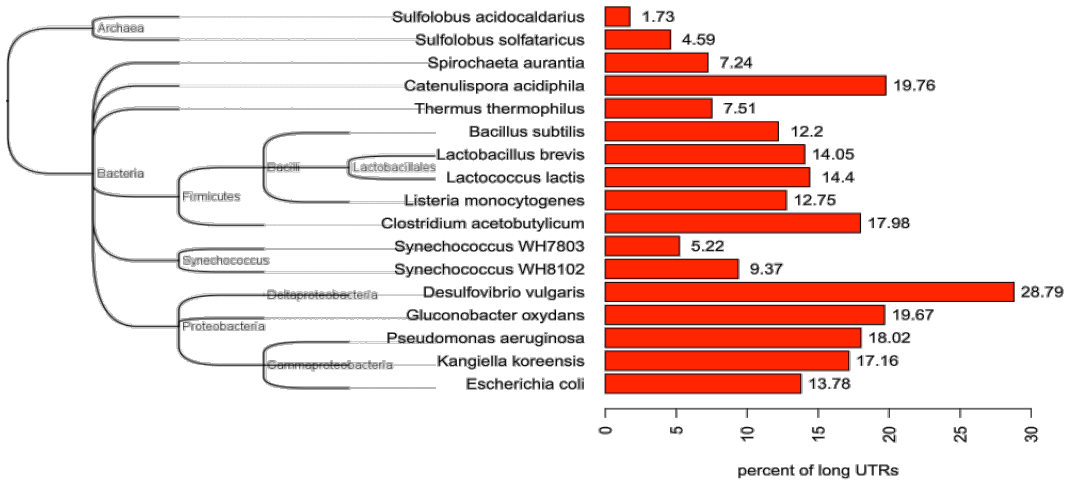
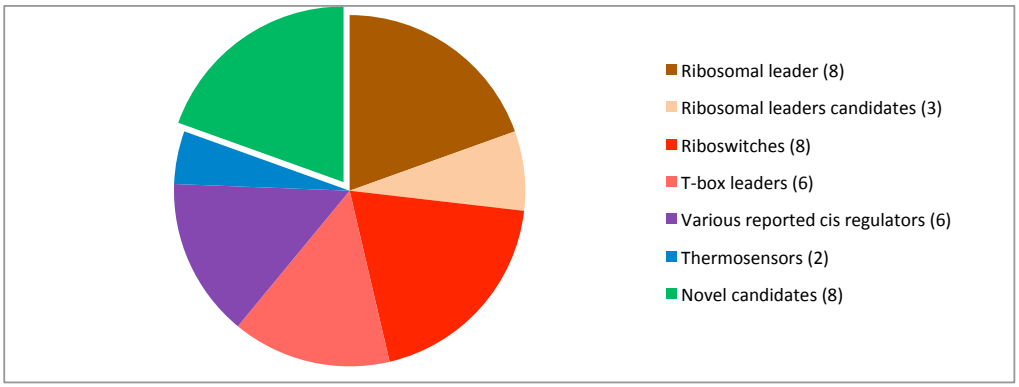


Figure S4 – Gene families with recurring long UTRs.

Functional characterization of gene families (COGs) with high propensity for recurring long 5'UTRs multiple species. Approximately three quarters of the gene families with recurring long 5'UTRs (30 out of 41) recapitulate previously reported 5'UTR regulatory elements partitioned into ribosomal leaders, riboswitches and thermosensor elements. Additional three gene families are candidates for ribosomal leaders and eight gene families are novel candidates for cis-acting 5'UTR regulators (COG0776, COG1271, COG0550, COG1622, COG1418, COG0605, COG0568, COG1077).



Extended Methods

RNA-seq protocols

RNA-seq protocols were employed as described in the individual references presented in Table S1. In cases marked as "this study", RNA-seq protocols were performed using standard Illumina strand insensitive library preparation kits as described in (1).

Determination and assessment of reliability of TSSs

Training set selection

For each organism, a "positive TSS training set" was selected as follows. We automatically selected dominant 5' sites (gTSS) upstream of genes that are the first in their operon. We defined genes as starting new operons if the distance to the nearest upstream gene on the same strand is greater than 40bp, as used previously (2, 4). Two criteria were used to define the "top" (best) TSS for each gene (gTSS): (1) maximal number of TAP-treated reads uniquely mapped to the site (2) minimal level of processing at the site or upstream to it, as defined by TAP(+) reads mapped to the site minus the TAP(-) reads in the range of 100bp upstream (and including the site itself). If, for a specific gene, the two criteria resulted with two or more sites, both sites were included in the training set only if the distance between them was greater than 50bp and the number of TAP(+) reads in both sites was above the 75 percentile of all gTSS sites in this transcriptome. Otherwise, the site with higher number of TAP(+) reads was preferred unless the alternative site was better supported by the local coverage increase ("Coverage ratio" and "Δ Coverage short", Table S2).

Filtering of training set TSSs

To omit sites that have very low probability of representing primary TSS, sites that did not satisfy one of the following filters were excluded. We used two types of filters: **(A) simple unary conditions -** omitting sites that include (1) less than two TAP(+) reads mapped to the site, and (2) an average RNA-seq coverage of less than one read per bp downstream of the site (average calculated over 80bp window). **(B) composite conditions** – omitting sites based on rules that take into account the transcriptome-computed distributions for all observed 5' sites that are present upstream of genes (up-5' sites). The condition to omit sites from the positive training set include: (1) the ratio of TAP(+)/TAP(-) was below the median of up-5' sites, with lack of either (i) RNA-seq coverage increase starting at site was below the median for up-5' sites or (ii) number of TAP(+) reads was below the 90^{th} percentile among up-5' sites; (2) TAP(+) reads was below 90^{th} percentile among up-5' sites, and coverage increase value was below the 90^{th} percentile among up-5' sites; and (3) low values in coverage increase (below median among up-5' sites) while no compensatory signal by high number of TAP(+) reads (below 90^{th} percentile among up-5' sites).

These automated selections were found to fit well with the benchmark of previously published sets of manually curated gTSS, quantified as accuracy in classification of curated gTSSs, we found TP=949, FP=211, and FN=422 for *Listeria monocytogenes* (2), TP=1,393, FP=481, and FN=724 for *Pseudomonas aeruginosa* (4), and TP=544, FP=288, and FN=508 for *Sulfolobus solfataricus* (6).

.

Random Forest learning procedure

The positive training set of TSSs for a typical transcriptome consisted of several hundred sites. We referred to the remaining sites (defined as sites supported by at least 2 TAP(+)) as the negative training set (consisting of the rest of the putative 5' sites, typically few tens of thousands). Therefore, the sizes of training sets produced by the automated selection procedure are imbalanced. Thus, to allow efficient learning we used "Random Forest" classifier (12, 13) with the negative training set sampled to maintain a 1:10 ratio of positive vs. negative sites. The learning accuracy and reproducibility was improved by averaging the score for each site over ten independent sampling of the negative training set, with each forest consisting of 1,000 trees. The cutoff score for sites was set for sensitivity of 95% of the positive training set. The cutoff for "high reliability score" was defined by the score set for sensitivity of 90% of the positive training set (these sites appear in bold color in the transcriptome maps). We note that the cutoff was set according to sensitivity rather than false positive rate due to the large, highly variable number of negatives in the trainning set.

Association of TSSs to genes

For many genes the learning process infers multiple alternative gTSSs in close proximity. Aiming to reconstruct useable transcriptome maps, we present only the top scoring sites in a 50bp region. A major challenge is to associate TSSs in cases where the TSS is located in an unusual distance from the start of the gene, because it could belong to an exceptionally long 5'UTRs (gTSSs), but also to a ncRNA (nTSSs). Initial association was performed prior to the learning with initial simplified rules reflecting that long 5'UTRs are rare in Archaea (6) but are more common in bacteria (14) . Sites were defined as gTSS if the distance to the closest gene was smaller than 300bp and 30bp in bacteria and archaea, respectively. If the distance was in the range of 300 to 600bp in bacteria and 30 to 300bp in archaea, the association remained ambiguous allowing for either gTSS or nTSS. The final association of these ambiguous sites was performed by taking into account the RNA-seq coverage. Distant sites were only associated with genes if there was a minimal level of continuity in the expression level between the TSS and the gene with maximal coverage drop along a 10 bp sliding window from the TSS to the ORF start was below the median of this value measured for all gTSS.

Inference of operons, and sRNAs

Operons

Operons (Transcriptional Units, TUs) prediction was performed similarly to the previously used approach (2, 4) with the main difference of allowing overlapping TUs. That is, each gTSS defines the beginning of a new TU, however in this study the existence of gTSS does not necessarily define the end of previous TU. An inferred TU was terminated if the downstream gene was located on the opposite strand, or the intergenic distance to the downstream gene was more than 200bp. In addition, an inferred TU was terminated based on coverage - if: (1) The expression level of downstream gene is below average coverage of one read covering each position, (2) there was a substantially different RNA-seq coverage level in the downstream gene,

with the threshold determined from calculating the distribution of coverage ratios between the first and the second half of single genes: If the ratio between the expression levels of two adjacent genes is above the 90 percentile of ratios within genes, the adjacent genes are not included in the same TU. (3) Termination of TU is also derived from substantial drop in coverage in the intergenic region, with the threshold determined from the distribution of the observed coverage drops within genes: If the coverage drop observed in the intergenic region is above the 75 percentile of coverage drops within genes, the TU is considered as terminated.

Small non-coding RNAs (sRNAs)

The prediction of small RNAs (sRNAs) was performed similarly to (2, 4), with a modified algorithm that allows sRNAs to be inferred within long 5'UTRs. The 5' of a putative sRNA was inferred from a TSS within intergenic region, which is not associated with a downstream gene, and only if sufficiently high coverage was observed (>= 20 percentile of the coverage in genome-wide intergenic regions). The 3' termination site of the sRNA was defined by 3-fold drop in expression as compared to the average coverage of that putative sRNA. An association of the sRNA transcript to the downstream gene was defined based on the maximal coverage drop between the TSS and the downstream gene start position (along a 10 bp sliding window). If the maximal drop from this TSS to the gene was above the median maximal drop measured for all TSSs and genes in this transcriptome, the sRNA transcript was defined as terminating at its inferred 3' end. Otherwise, the sRNA transcript was defined as part of a long 5'UTR associated with the downstream gene.

References

1. He,S., Wurtzel,O., Singh,K., Froula,J.L., Yilmaz,S., Tringe,S.G., Wang,Z., Chen,F., Lindquist,E.A., Sorek,R., *et al.* (2010) Validation of two ribosomal RNA removal methods for microbial metatranscriptomics. *Nat Methods*, **7**, 807–812.
2. Wurtzel,O., Sesto,N., Mellin,J.R., Karunker,I., Edelheit,S., Bécavin,C., Archambaud,C., Cossart,P., Sorek,R. and Becavin,C. (2012) Comparative transcriptomics of pathogenic and non-pathogenic Listeria species. *Mol. Syst. Biol.*, **8**, 583.
3. Karunker,I., Rotem,O., Dori-Bachash,M., Jurkevitch,E. and Sorek,R. (2013) A global transcriptional switch between the attack and growth forms of Bdellovibrio bacteriovorus. *PLoS One*, **8**, e61850.
4. Wurtzel,O., Yoder-Himes,D.R., Han,K., Dandekar,A.A., Edelheit,S., Greenberg,E.P., Sorek,R. and Lory,S. (2012) The Single-Nucleotide Resolution Transcriptome of Pseudomonas aeruginosa Grown in Body Temperature. *PLoS Pathog.*, **8**, e1002945.
5. Doron,S., Fedida,A., Hernández-Prieto,M.A., Sabehi,G., Karunker,I., Stazic,D., Feingersch,R., Steglich,C., Futschik,M., Lindell,D., *et al.* (2015) Transcriptome dynamics of a broad host-range cyanophage and its hosts. *ISME J.*, 10.1038/ismej.2015.210.
6. Wurtzel,O., Sapra,R., Chen,F., Zhu,Y.W., Simmons,B.A. and Sorek,R. (2010) A single-base resolution map of an archaeal transcriptome. *Genome Res*, **20**, 133–141.
7. Naville,M. and Gautheret,D. (2010) Premature terminator analysis sheds light on a hidden world of bacterial transcriptional attenuation. *Genome Biol.*, **11**, R97.
8. Hunger,K., Beckering,C.L., Wiegeshoff,F., Graumann,P.L. and Marahiel,M.A. (2006) Cold-induced putative DEAD box RNA helicases CshA and CshB are essential for cold adaptation and interact with cold shock protein B in Bacillus subtilis. *J. Bacteriol.*, **188**, 240–8.
9. Zengel,J.M. and Lindahl,L. (1992) Ribosomal protein L4 and transcription factor NusA have separable roles in mediating terminating of transcription within the leader of the S10 operon of Escherichia coli. *Genes Dev.*, **6**, 2655–2662.
10. Tchufistova,L.S., Komarova,A. V and Boni,I. V (2003) A key role for the mRNA leader structure in translational control of ribosomal protein S1 synthesis in gamma-proteobacteria. *Nucleic Acids Res.*, **31**, 6996–7002.
11. Nord,S., Bylund,G.O., Lövgren,J.M. and Wikström,P.M. (2009) The RimP protein is important for maturation of the 30S ribosomal subunit. *J. Mol. Biol.*, **386**, 742–53.
12. Breiman,L. (2001) Random Forests. *Mach. Learn.*, **45**, 5–32.
13. Khoshgoftaar,T.M., Golawala,M. and Hulse,J. Van (2007) An Empirical Study of Learning from Imbalanced Data Using Random Forest. In *19th IEEE International Conference on Tools with Artificial Intelligence(ICTAI 2007)*. IEEE, Vol. 2, pp. 310–317.
14. Sorek,R. and Cossart,P. (2010) Prokaryotic transcriptomics: a new view on regulation, physiology and pathogenicity. *Nat. Rev. Genet.*, **11**, 9–16.