Comparative transcriptomics across the prokaryotic tree of life

Ofir Cohen1,2, Shany Doron1, Omri Wurtzel1,3, Daniel Dar1, Sarit Edelheit1, Iris Karunker1, Eran Mick1,4 and Rotem Sorek1,*

1Department of Molecular Genetics, Weizmann Institute of Science, Rehovot 76100, Israel, 2Broad Institute of Harvard and MIT, Cambridge, MA 02142, USA, 3Whitehead Institute for Biomedical Research, Cambridge, MA 02142, USA and 4Department of Systems Biology, Harvard Medical School, Boston, MA 02115, USA

Received February 08, 2016; Revised April 20, 2016; Accepted April 28, 2016

ABSTRACT

Whole-transcriptome sequencing studies from recent years revealed an unexpected complexity in transcriptomes of bacteria and archaea, including abundant non-coding RNAs, cis-antisense transcription and regulatory untranslated regions (UTRs). Understanding the functional relevance of the plethora of non-coding RNAs in a given organism is challenging, especially since some of these RNAs were attributed to ‘transcriptional noise’. To allow the search for conserved transcriptomic elements we produced comparative transcriptome maps for multiple species across the microbial tree of life. These transcriptome maps are detailed in annotations, comparable by gene families, and BLAST-searchable by user provided sequences. Our transcriptome collection includes 18 model organisms spanning 10 phyla/subphyla of bacteria and archaea that were sequenced using standardized RNA-seq methods. The utility of the comparative approach, as implemented in our web server, is demonstrated by highlighting genes with exceptionally long 5′UTRs across species, which correspond to many known riboswitches and further suggest novel putative regulatory elements. Our study provides a standardized reference transcriptome to major clinically and environmentally important microbial phyla. The viewer is available at http://exploration.weizmann.ac.il/TCOL, setting a framework for comparative studies of the microbial non-coding genome.

INTRODUCTION

The growing availability of next-generation sequencing technologies has led to a burst of studies characterizing the transcriptomes of many prokaryotic species, often charting transcription start sites (TSSs) at single-nucleotide resolution (1–9). These studies unraveled unexpected complexity in the transcriptomes of bacteria and archaea including numerous long regulatory 5′UTRs, non-coding RNAs, alternative operon structures, internal promoters and abundant cis-antisense transcription (10–12).

Comparative transcriptome studies of various prokaryotes had shown that transcriptional maps may differ even among closely related species, challenging functional interpretations of ncRNA transcription with suspected transcriptional noise (13–17). While the functional importance of non-conserved transcription remains debated (18), cases in which evolutionary conservation is observed were shown to have the potential to highlight functionally important features in the organisms’ transcriptional dynamics (19–21). To date, comparative transcriptomic studies were limited to few organisms, usually comparing closely related species.

A major challenge in comparative transcriptomics studies is the construction of reliable and comparable transcriptome maps across multiple diverse species. The first published transcriptome maps were based on manual curation of multiple putative TSSs. With the dramatic reduction of sequencing costs, however, using manual curation for construction of transcriptome maps becomes increasingly prohibitive because it is not scalable to large number of organisms and multiple conditions. Thus, in the last couple of years several methods had been developed to automatically infer TSSs. Some of these methods are tailored analyses for differential 5′-RNA sequencing data (dRNA-seq, (11)) as in the primary transcriptome of H. pylori (1). These include methods that are primarily based on analyzing two RNA-seq libraries, one untreated and one enzymatically treated to enrich for primary transcripts (22–24). Additional methods that analyze TSSs integrate comparative signals from closely related species (25), while other methods attempt to delineate transcript boundaries (both TSS and termination) based on statistically significant local differences in RNA-seq coverage (26,27). However, methods that can infer pri-
mary TSSs based on multiple transcriptomic and genomic features, taken together, are still lacking.

In this study, we present the TCOL web-server (Transcriptomes Compared across the tree Of Life) that includes several important novel components. First, TCOL encompasses the largest compendium of comparable transcriptomes to date, spanning representatives of the major phyla of the prokaryotic tree of life. Second, all transcriptomes were sequenced using standardized approach, combining RNA-seq and 5’-end sequencing, generating comparable data. Third, we reconstructed the transcriptome maps using a computational pipeline that utilized machine learning to accurately infer primary TSSs based on multiple genomic and transcriptomic features. Fourth, TCOL includes a comparative transcriptome browser, allowing for simultaneous viewing of transcriptome maps for homologous genes across divergent organisms. Finally, the TCOL transcriptomes are BLAST-searchable with query sequences provided by users, broadening the utility of this web-server to the microbiology community.

MATERIALS AND METHODS

Culture conditions

Most data were derived from previously published studies, with the exception of the following organisms for which data were obtained within this study: *E. coli* was grown in LB medium overnight, diluted 1:200 into fresh LB and incubated at 37°C until OD<sub>600nm</sub> of ~0.3 (early-mid log). *B. subtilis* was grown in TB, similarly diluted into fresh TB until reaching mid log OD<sub>600nm</sub> of ~0.6. *G. oxydans* 621H was grown in pH 6 in constant temperature of 30°C in two media: (i) mannitol fermentation (50 mM glucose) (ii) glycerol fermentation (50 mM glycerol) until reaching OD<sub>600nm</sub> of 1.12 and OD<sub>600nm</sub> of 0.9, for mannitol and glycerol, respectively. *C. acetobutylicum* was cultured in phosphate limited conditions at (i) pH 4.5 (ii) pH 5.7 until OD<sub>600nm</sub> of 3.7. *T. thermophilus* was cultured in TB medium at 70°C until OD<sub>600nm</sub> of 3.0. *S. acidocaldarius* was cultured in ‘Brock medium’ (28) with 0.1% tryptone added, until early stationary phase. For all other organisms culture conditions details were published (Supplementary Table S1).

RNA-seq protocols and reads mapping

Total RNA-seq (strand insensitive) and 5’ end transcriptome (strand sensitive) libraries were produced as described previously (19,29) (Supplementary Materials). List of transcriptomes in this study is depicted in Table 1, and sequencing depth and growth conditions of all organisms are detailed in Supplementary Table S1. Briefly, the 5’ differential RNA-seq protocol included the construction of 2 libraries: A treated library in which RNA is incubated with Tobacco Acid Pyrophosphatase (TAP, Epicentre), which we term TAP(+) and an untreated library termed TAP(−). The relative number of TAP(+) and TAP(−) reads that are mapped to the same site is indicative of the likelihood of that site to represent primary TSS. Higher TAP(+)/TAP(−) ratios typically represent primary TSSs while lower values represent processing sites (Supplementary Materials).

Determination and assessment of reliability of TSSs

A machine learning approach was utilized to infer TSSs from all 5’ mapped sites (the set of putative TSSs). The inference algorithm consists of two main steps: (i) Heuristic determination of the training set and (ii) supervised Random Forest learning (pipeline steps in Supplementary Figure S1). The learning is based on a set of 15 features which capture relevant TSS characteristics including genomic, whole-transcriptome RNA-seq coverage patterns, and 5’ differential RNA-seq (TAP(+)/(−)), see list in Supplementary Table S2). A Random Forest machine learning procedure was used to infer and quantify reliability of the TSS determination. Finally, the classification of TSS types (e.g. gTSS as gene-related TSS or nTSS as non-coding TSS) was performed automatically based on genomic and transcriptomic features (Supplementary Extended Methods).

Inference of operons, sRNAs and putative ORFs

The general approach to infer operons (transcriptional units, TUs), sRNAs and putative ORFs was as previously described (19,29). Briefly, operon beginnings are defined by location of primary gTSSs while operon ends are defined by the genomic coordinates of the last gene which is likely to be expressed within the same TU. Inclusion of downstream genes in the same TU is dependent on both genomic and transcriptomic features (Supplementary Extended Methods). sRNAs inference is based on the presence of a TSS and expression in intergenic regions (Supplementary Extended Methods). Putative ORFs are inferred based on bioinformatic detection of putative start and stop codons in intergenic regions. Putative ORFs of at least 30 amino acids were subjected to BLASTX search against the non-redundant protein database.

The genomic browser and comparative analysis of gene families

The genomic browser is based on GenBank annotation for each organism in this study. The genomic browser annotation is augmented by sRNA and *cis*-regulators if found in Rfam repositories (30) or found with covariance models (31) homology search against the Rfam database (E-value < 10<sup>−5</sup>). The comparative approach is focused on protein families. We constructed the set of homologous proteins by first producing the network of all homologies with blastp E-value < 10<sup>−3</sup> and delineation of best bi-directional hits (BBH) between homologs. For readability we will refer to sets of BBH homologs genes as orthologs throughout the manuscript. For each locus across all organisms in our study all BBH loci are listed and a comparative viewer with aligned transcriptomes for the entire set is available. Further comparative analysis of gene families is facilitated by depicting, for each locus, the entire list of homologs (orthologs and paralogs) found in our compendium with the same COG (Clusters of Orthologous Groups) affiliation (32,33). The COG affiliation is shown separately for three phylogenetic levels: ‘TOL’, ‘Domain’ and ‘Phylum’ for orthologous groups across the entire Tree Of Life, Domain-specific, or Phylum-specific, respectively.
**RESULTS**

General approach and the TCOL transcriptome browser

To facilitate comparison of transcriptomes of diverged organisms we use standardized sequencing protocols and a fully automated computational pipeline for the reconstruction of transcriptome maps, including accurate inference of TSSs, and transcription-dependent prediction of sRNAs and operons. We produced a user-friendly web server that visualizes the transcriptome maps, available at [http://exploration.weizmann.ac.il/TCOL](http://exploration.weizmann.ac.il/TCOL). The transcriptome browser provides a multi-track view of homologous genes across the tree of life, providing a unique comparison with these benchmarks, both the sensitivity and precision were high, with sensitivity (True Positive, TP) of 79.3%, 74.5% and 64.1%, and precision (TP/TP+FP) of 78.3%, 70.9% and 64.1% for *Listeria, Pseudomonas* and *Sulfolobus*, respectively (Supplementary Materials Figure S2). Further inspection of a subset of TSSs in which the automatic annotation and the benchmark, manually curated data, disagreed showed that in the vast majority of cases the automatic inference is more likely to report the correct TSS (Supplementary Materials).

Reconstruction of transcriptome maps

A quantitative summary of all transcriptome maps in this study is presented in Table 1. Primary TSSs were inferred using a machine learning approach, integrating relevant genomic and transcriptomic information in an unbiased manner using multiple features (Supplementary Table S2). In order to estimate the accuracy of our automated inference of TSSs, we compared these to a set of manually curated TSSs in three organisms for which published TSS maps are available - *Listeria monocytogenes* (19), *Pseudomonas aeruginosa* (29) and *Sulfolobus solfataricus* (4). In comparison with these benchmarks, both the sensitivity and precision were high, with sensitivity (True Positive, TP) of 79.3%, 74.5% and 64.1%, and precision (TP/TP+FP) of 78.3%, 70.9% and 64.1% for *Listeria, Pseudomonas* and *Sulfolobus*, respectively (Supplementary Materials).

Using the TCOL transcriptome viewer

Comprehensive documentation of the functions available in TCOL is detailed in the ‘help’ page of the web server. Figure 2 illustrates some of the main features of TCOL. Users can start by choosing a representative organism (Figure 2A). The next step is choosing a desired sample (e.g. growth condition), and pressing ‘Show’ to open the transcriptome viewer (Figure 2C). A user can navigate the browser by genome position, gene locus name or annotated Rfam genes. Genomic and transcriptomic information appears as ‘tooltips’ upon hovering with the cursor, including TSS information (e.g. hover over plotted TSS arrow to...
Figure 1. Representative transcriptomes presented on the tree of life. The compendium of transcriptomes in this study is denoted as colored lineages across the prokaryotic tree of life consisting of 18 representative species across different phyla including Gamma, Alpha and Delta Proteobacteria, Cyanobacteria, Firmicutes, Deinococcus-Thermus, Actinobacteria, Spirochaetes and the Archaeal Crenarchaeotes phylum. Tree was taken from (44,45), reconstructed with (45).

reveal position, number of supporting reads and reliability score of TSS), coverage information (number of RNA-seq reads per position), information on genes and regulatory elements and operon inferences. The user may click on a gene to open a ‘Gene page’ (Figure 2D) containing specific gene and homologs information, including links to the homologs transcriptional maps, and a link to the comparative viewer that visualizes all homologs together (Figure 2E; Figure 3). Furthermore, in each ‘Gene page’ the ‘COG-based homology’ section contains links to the transcriptomes of all the genes in our data set with the same COG (Clusters of Orthologous Groups) (32,33).

Alternatively, browsing can start by BLAST searching a query sequence against the database of the TCOL compendium (Figure 2B). Each of the returning hits includes links to both the comparative transcriptome map as well as the maps for individual species. There are three different BLAST options: blastp, blastn and blastx.

Users interested in browsing all TSS or RNA-seq data globally can use the ‘Supplementary files’ tab that contains downloadable detailed information on the transcriptome annotation, including summary statistics (‘TSS statistics’), details of the TSS for each gene (‘TSS’), summary of total RNA-seq data statistics (‘RNA-seq summary’), number of reads mapped to each gene (‘Expression per gene’) and a summary file with additional transcriptome features including RFAM genes, operons and identified sRNAs (‘Additional transcriptome annotations’).

Comparative transcriptomics highlights cis-acting RNA regulators

One of the main applications of the comparative transcriptomics approach is to highlight cases in which recurring transcriptomic features are found across homologs, suggestive of regulatory importance and evolutionary conservation. We demonstrate this by focusing on orthologous genes showing long 5’ UTRs across the tree of life.

Long regulatory 5’ UTRs have the capacity to integrate environmental and inner-cellular cues into regulation of
Figure 2. The Transcriptomes Compared across the Tree of Life (TCOL) online web-server. Interactive browser with detailed transcriptional maps of 18 representative species across the microbial tree of life. Transcriptome maps can be accessed either by (A) choosing an organism or (B) by using BLAST with a user-provided query sequence. (C) Transcriptome maps can be navigated with multiple functions and are information-rich. Some of the information appears as tooltips when hovering above the features with the cursor. (D) ‘Gene page’ (when clicking on a gene in the browser) containing information about the gene including its homologs. (E) The comparative transcriptome viewer shows aligned transcriptomes of homologous genes, centered on the queried gene. This view is accessible from the ‘Gene page’ and from the BLAST results page.
Figure 3. Comparative transcriptome views of homologous genes with long 5’UTRs. (A) Threonyl-tRNA synthetase (*thrS*) has a known T-box (RF00230) in *Listeria*, *Clostridium* and *Bacillus*. (B) The transcription termination factor *rho* gene (COG1158) in which regulatory structure was previously inferred in *Pseudomonas* (*Pseudomon-Rho*). (C) The cytochrome d ubiquinol oxidase subunit I (COG1271) gene. (D) The cytochrome o ubiquinol oxidase subunit II (COG1622) gene. Brown horizontal arrows denote RFAM-identified *cis*-regulators.
gene expression (34,35). Previous comparative genomics studies significantly enhanced the capacity to detect putative functional 5′ regulatory elements (36,37). However, it is likely that many additional regulatory 5′ UTRs elements remain to be found, including regulatory elements with conservation that is insufficient for reliable detection from genomic data alone (38,39).

Searching for gene families with long conserved 5′ UTR in multiple organisms across the tree of life, we found 41 gene families (COGs) with exceptional propensity for recurring long 5′ UTRs (Supplementary Table S3). Approximately three quarters of the gene families within this list recapitulated previously reported 5′ UTR regulatory elements (Supplementary Figure S4). These include 14 known riboswitches (including T-box, S-box, TPP, Cobalamin, FMN, glmS ribozyme and Glycine, Purine and Lysine riboswitches); 8 ribosomal leaders; 2 thermostosensor RNA elements; and 6 cis-regulatory elements of various categories that had been reported including the rimP leader, Pseudomon-GroES, Pseudomon-Rho, LactobtpoB, mini-ykkC and leucine operon leader. We further found 8 novel putative candidates, inferred to harbor RNA-regulatory elements of unknown function in their 5′ UTRs, and additional 3 ribosomal leader candidates (Supplementary Table S3; Figure 3).

As examples, Figure 3 presents the comparative transcriptome viewer of four gene families with recurring long 5′ UTRs. Figure 3A illustrates the 5′ UTR of the threonyl tRNA synthetase gene family (COG0441, thrS), where T-box 5′ UTR riboswitches are found in gram-positive bacteria (40). Among the 5 presented cases, the Rfam T-box leader (RF00230) is described only in Bacillus. Among the homologs of the ‘Rho, Transcription termination factor’ gene family (COG1158), we found recurring long 5′ UTRs across multiple genera including Pseudomonas, Glucobacter, Desulfovibrio and Bacillus (Figure 3B). These results suggest that the Rho transcriptional terminator is being regulated by a leader 5′ UTR element in multiple bacterial species in addition to Pseudomonas.

The gene families ‘cytochrome d ubiquinol oxidase subunit I’ (COG1271), and ‘cytochrome o ubiquinol oxidase subunit II’ (COG1622) are part of the CydABX and CyoABCD operons, respectively. These two different cytochrome oxidase complexes, responsible for cellular respiration, are known to be regulated by the oxygen state. We found, for both of these gene families, recurring long 5′ UTRs across divergent bacteria, implying the existence of cis-acting RNA regulators within their 5′ UTRs (Figure 3C and D).

CONCLUSION

The comparative transcriptomics approach was recently shown to be useful for studying closely related species – either strains within the same species (15) or species within the same genus (16,19,21). Broader evolutionary comparison across diverse species based on aggregating results from different studies (41) is limited due to methodological differences in RNA sequencing and computational analyses. Here we provide, for the first time, a comparative transcriptome viewer of representative species spanning the prokaryotic tree of life, generated by standardized methodologies in both data generation and data analysis. This constitutes an important step toward better understanding the similarities and differences in transcriptome regulation and dynamics in various organisms.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank Sarah Burge for assisting with access to Rfam data and Zichen Yao for useful discussions and help with CMfinder. We thank W. Liebl, A. Ehrenreich and C. Doering for contributing the RNA of C. acetobutylicum, G. oxydans and T. thermophilus. We also thank Sonja-Verena Albers and Michaela Wagner for contributing S. acidocaldaris cell pellets.

FUNDING

ISF [personal grant 1303/12 and I-CORE grant 1796/12]; ERC-StG program [260432]; HFSP [RGP0011/2013]; Abisch-Frenkel foundation; Pasteur-Weizmann council grant; Minerva Foundation; Leona M. and Harry B. Helmsley Charitable Trust; DIP grant from the Deutsche Forschungsgemeinschaft; AXA postdoctoral research grant [project 711545 to O.C.], Funding for open access charge: ERC-StG program [260432].

Conflict of interest statement. None declared.

REFERENCES

1. Sharma, C.M., Hoffmann, S., Darfeuille, F., Reigner, J., Findeiss, S., Sittka, A., Chabas, S., Reiche, K., Hackermüller, J., Reinhardt, R. et al. (2010) The primary transcriptome of the major human pathogen Helicobacter pylori. Nature, 464, 250–255.
2. Mitschke, J., Georg, J., Scholz, I., Sharma, C.M., Dienst, D., Bantscheff, J., Voss, B., Steglich, C., Wilde, A., Vogel, J. et al. (2011) An experimentally anchored map of transcriptional start sites in the model cyanobacterium Synechocystis sp. PCC6803. Proc. Natl. Acad. Sci. U.S.A., 108, 2124–2129.
3. Qin, Y., Cho, B.-K., Park, Y.S., Lovley, D., Palsson, B.O. and Zhengler, K. (2010) Structural and operational complexity of the Geobacter sulfurreducens genome. Genome Res., 20, 1304–1311.
4. Wurtzel, O., Supra, 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.
5. Teledo-Arana, A., Dussurget, O., Nikitas, G., Sesto, N., Guet-Revillet, H., Balestrino, D., Loh, E., Gripenland, J., Tsiensu, T., Vaitkevicius, K. et al. (2009) The Listeria transcriptional landscape from saprophytism to virulence. Nature, 459, 950–956.
