Robust identification of noncoding RNA from transcriptomes requires phylogenetically-informed sampling

Stinus Lindgreen[1,2,co], Sinan Uğur Umu[2,co], Alicia Sook-Wei Lai[2], Hisham Eldai[2], Wenting Liu[2], Stephanie McGimpsey[2], Nicole Wheeler[2], Patrick J. Biggs[3,4], Nick R. Thomson[5], Lars Barquist[5,6], Anthony M. Poole[2,4,*], Paul P. Gardner[2,*]

1 Section for Computational and RNA Biology, Department of Biology, University of Copenhagen, Ole Maaloes Vej, 2200 Copenhagen, Denmark
2 Biomolecular Interaction Centre, School of Biological Sciences, University of Canterbury, Christchurch, New Zealand
3 Institute of Veterinary, Animal & Biomedical Sciences, Massey University, Palmerston North, New Zealand
4 Allan Wilson Centre for Molecular Ecology & Evolution, New Zealand
5 Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridgeshire CB10 1SA, UK
6 Institute for Molecular Infection Biology, University of Wuerzburg, D-97080 Wuerzburg, Germany
co: These authors contributed equally.
∗ Correspondence to A.M.P. or P.P.G.
E-mails: anthony.poole@canterbury.ac.nz; paul.gardner@canterbury.ac.nz

Abstract

Noncoding RNAs are increasingly recognized as integral to a wide range of biological processes, including core processes such as translation, gene regulation, host-pathogen interactions and environmental sensing. While bacterial and archaeal genomics is now a mature field, our capacity to identify noncoding RNA elements in genomes is hampered by the difficulty of de novo identification. The emergence of new technologies for characterizing transcriptome outputs, notably RNA-seq, are improving noncoding RNA identification and expression quantification. However, a major challenge is to robustly distinguish functional outputs from transcriptional noise. To establish whether annotation of existing transcriptome data has effectively captured all functional outputs, we analysed 400 publicly available RNA-seq datasets spanning 40 different Archaea and Bacteria. Using comparative tools, we identify close to a thousand highly-expressed candidate noncoding RNAs. However, our analyses reveal that capacity to identify noncoding RNA outputs is strongly dependent on phylogenetic sampling. Surprisingly, and in stark contrast to protein-coding genes, the phylogenetic window for effective use of comparative methods is perversely narrow: aggregating public datasets only produced one phylogenetic cluster where these tools could be used to robustly separate unannotated noncoding RNAs from a null hypothesis of transcriptional noise. Our results show that for the full potential of transcriptomics data to be realized, a change in experimental design is paramount: effective transcriptomics requires phylogeny-aware sampling.

Introduction

Genome sequencing has transformed microbiology, offering unprecedented insight into the physiology, biochemistry, and genetics of Bacteria and Archaea [1–4]. Equally, careful examination of transcriptional outputs has revealed that bacterial and archaeal transcriptomes are remarkably complex [5], containing a diverse array of noncoding RNA (ncRNA) families, including sRNAs, sno-like sRNAs in Archaea, crRNAs and riboswitches [6–10]. However, our view of the microbial RNA world still derives from a narrow sampling [11] of microbial diversity. Additional bias comes from the fact that many microbes are not readily culturable [12]. The development of metagenomics and initiatives such as the Genomic Encyclopedia of Bacteria and Archaea (GEBA) project have sought to redress these biases, generating

genomes spanning undersampled regions of the bacterial and archaeal phylogeny [1], and sequencing uncultured or unculturable species through metagenomics [2, 13–16].

A further source of bias in our genome-informed view of microorganisms derives from a protein-centric approach to genome annotation. The majority of genome sequences deposited in public databases carry limited annotation of noncoding RNAs and cis-regulatory elements, yet it is rapidly becoming clear that RNA is essential to our understanding of molecular functioning in microbes [17].

The paucity of annotations is understandable, as RNA gene annotation is non-trivial [18,19]. However, the increasing roles for RNA uncovered through experimental and bioinformatic study make illuminating this dark matter all the more urgent. Among the remarkable discoveries made are, riboswitch-mediated regulation [20], transcriptional termination by RNA elements [21–23], identification of novel natural catalytic RNAs [24], CRISPR-mediated acquired immunity [25], temperature-dependent gene regulation [26], and sno-like RNAs in Archaea [27,28]. The Rfam database [21,29] provides a valuable platform for collating and characterising these and other families of noncoding RNA. However, a recent comparative analysis [30] revealed that fewer than 7% of RNA families within Bacteria and less than 19% in Archaea show a broad phylogenetic distribution (that is, presence in at least 50% of sequenced phyla). Crucially, that analysis revealed that underlying genome sequencing biases were a major contributor to this pattern, and that the wider genomic sampling provided by the GEBA dataset did help improve identification of broadly-conserved RNA families [30]. Tools such as RNA-seq [31] and transposon insertion sequencing [32–34] promise to complement comparative genomics tools for RNA family discovery, and it may be possible to use a mix of data types in the identification of RNA elements. However, to date, no systematic analysis of available data has been undertaken, suggesting ncRNAs may be hidden in the deluge of published data.

We have therefore assessed the value of RNA-seq data for identification of unannotated non-coding and cis-regulatory RNA elements in bacterial and archaeal genomes. We show that numerous, hitherto uncharacterised, expressed RNA families are lurking in publicly available RNA-seq datasets. We find that poor sequence conservation for RNA families limits the capacity to identify evolutionarily conserved, expressed ncRNAs from existing genomic and transcriptomic data. Our results suggest that maximising phylogenetic distance, a sampling strategy effective for identification of novel protein families [1,2], is not the most effective strategy for ncRNA identification. Instead, our results show that, for RNA element identification, sequencing clusters of related microbes will generate the greatest benefit.

Results

Non-coding RNA elements dominate bacterial and archaeal transcriptional profiles

To assess the relative contribution of noncoding RNAs and protein-coding genes to transcriptional output, we collected all publicly-available bacterial and archaeal RNA-seq datasets (available as of **August 2013**), spanning 37 species/strains and 400 datasets. For all datasets, we supplemented publicly available genome annotations with screening for additional loci against Pfam and Rfam databases, followed by manual identification of expressed unannotated regions. This latter annotation yielded 922 expressed RUFs (RNAs of unknown function).

We next examined the relative abundance of transcripts within each RNA-seq dataset, yielding an expression rank for individual transcripts. This analysis reveals that most transcriptomes are dominated by highly expressed non-coding RNA outputs (Figure 1) (P-value $<< 0.0001$, Chi-square test of observed vs. expected ratios and Fisher's Exact test on the counts). In addition to well-characterised RNAs (rRNA, tRNA, tmRNA, RNase P RNA, SRP RNA, 6S and sno-like sRNAs), and known cis-regulatory elements (riboswitches, leaders and thermosensors - Table S2), the top 50 most abundant transcriptional outputs (Figure 1) across the 32 Bacteria and 5 Archaea in our dataset included a total of 308 RUFs.

Comparative analyses reveal that highly expressed transcripts are often poorly conserved

To assess whether highly expressed RUFs possess features commonly associated with function, we employed three criteria: 1) evolutionary conservation, 2) conservation of secondary structure, 3) evidence of expression in more than one RNA-seq dataset. For this analysis, we compared and ranked transcriptional outputs across species (see Methods for details). Based on the relative rank across RNA-seq datasets and the maximum phylogenetic distance observed across all genomes, each transcript was classified as high, medium or low expression, and high, medium or low conservation. This yielded a set of highly expressed transcripts consisting of 162 Rfam families, 568 RUFs and 1429 Pfam families. As expected [35–37], conserved, highly expressed outputs are dominated by protein-coding transcripts (Figure 2B&C). In contrast, transcripts that are highly expressed but poorly conserved are primarily RUFs (Figure 2A). Of the 568 RUFs identified, only 25 are supported by all three criteria (conservation, secondary structure and expression) (Figure 2D). A further 138 RUFs are supported by two criteria (Figure 2D). Consequently, on these criteria, the vast majority of RUFs appear indistinguishable from transcriptional noise. However, as these RUFs are among the mostly highly expressed transcripts in public RNA-seq data, we next considered whether our criteria were sufficiently discriminatory to identify functional RNAs. It is well established that not all functional RNAs exhibit conserved secondary structure – antisense base pairing with a target is common, and does not require intramolecular folding [38]. This indicates that criterion 2 will apply to some, but not all functional RNA elements. Criteria 1 and 3 both derive from comparative analysis: criterion 1 requires an expressed RUF to be conserved in some other genome, while criterion 2 requires an expressed RUF to be expressed in another of the datasets in our study. We therefore sought to examine how effective our comparative analyses are given that the available data represent a small sample (transcriptomes from 37 species) and given that biases in genome sampling across bacterial and archaeal diversity impact comparative analysis of RNAs [30].

Comparative analysis reveals a 'Goldilocks Zone' for ncRNA identification

Effective comparative analysis requires appropriate phylogenetic distances between species under investigation (Eddy 2005). The transcriptomes available for our analyses are limited in number and phylogenetic diversity, and efforts to redress phylogenetic biases in genome sequencing have focused on maximising phylogenetic diversity across the tree of life [1, 2]. Moreover, underlying biases in genome sampling have been shown to affect the assessment of ncRNA conservation [30]. We therefore considered how species selection impacts comparative analysis as a tool for distinguishing signals of conservation in ncRNAs from transcriptional noise.

To assess the effect of species selection on our capacity to identify RNA families using comparative analysis, we first generated F84 phylogenetic distances between 2562 bacterial species and 154 archaeal species using SSU rRNA sequences from each species (Methods). Next, for each Rfam RNA family and Pfam protein family, we identified the maximum phylogenetic distance between any two species that encode a given family. We then calculated the fraction of conserved RNA and protein families for a given phylogenetic distance.

This reveals a dramatic difference in evolutionary conservation of Rfam and Pfam families (Figure 3). While 80% of protein families are still conserved at the broad evolutionary distances that separate Bacteria and Archaea, the phylogenetic distance at which 80% of RNA families are conserved lies somewhere between the taxonomic levels of genus and family (Figure 3). The explanation for this rapid decay of RNA family conservation across long evolutionary time-scales is likely to be a combination of the limited abilities of existing bioinformatic tools to correctly align RNA sequences [39] and genuine rapid turnover of non-coding RNAs during evolution [30].

These results in turn indicate that appropriate evolutionary distances for optimising comparative analysis differ greatly for protein- and RNA-coding genes. Figure 3 confirms the utility of the GEBA

sampling strategy [1, 2] for protein-coding gene identification, since maximising phylogenetic diversity yields strong identification of conserved protein-coding genes. In contrast, at the largest phylogenetic distances, around 60% of RNA families are invisible to comparative analysis. These results define a 'Goldilocks Zone' (an evolutionary distance neither too close nor too distant) for RNA family comparative analysis.

To assess how phylogenetic sampling impacts comparative analysis of transcriptome data, we mapped the pairwise distances between species covered by the RNA-seq datasets in this study. Of the 506 possible pairs (excluding Bacteria vs Archaea), only 11 are in the Goldilocks Zone for RNA (phylogenetic distance between 0.0118 and 0.0542) covering 9 species/strains. While five pairs of datasets are 'too hot' (i.e. too close phylogenetically), the remaining 490 comparisons are 'too cold' for effective comparative RNA analysis (Figure 3). The datasets in the Goldilocks Zone span three distinct clades covering five Enterobacteria, two Pseudomanada, and two Xanthomonada (Figure 3). Subsequent analyses focused on the Enterobacterial clade, as this was the only clade containing transcriptomes from multiple species.

We next calculated the percentage of conserved RUFs for all Enterobacterial species pairs. On average, 83% of RUFs are conserved across the Goldilocks Zone. The two E. coli strains are extremely similar, and share 99% of RUFs, suggesting that these species are too similar for us to robustly separate expression of bona fide RNAs and noise. While these outputs could be genuine RNAs, these strains are in the 'too hot' region: if everything is conserved, comparative power is lost. In contrast, only 12% of RUFs are conserved between species pairs in the 'too cold' region (spanning clades; Figure 3). Indeed, most of the currently available datasets fall outside the Goldilocks Zone for RNA discovery (Figure 3): of the 197 RUFs found in the datasets that cover this region, only 19 show evidence of expression in another transcriptome. This suggests that the low number of RUFs from Figure 2D showing both conservation and expression in other AREBA species is primarily a consequence of limited sampling. We also extracted a RUF in the Enterobacterial clade that had homologs in the other four species. In all cases, the homologs showed evidence of expression when examining the RNA-seq data, and the five homologs show evidence of a conserved secondary structure (Figure 3).

In summary, the Goldilocks Zone for RNA is surprisingly narrow, and suggests that optimal species selection for RNA comparative analyses should comprise strains of the same species, members of the same genus, and closely related taxonomic families (Figure 3). Thus, the Goldilocks Zone for RNA is not encompassed by the sampling regimes currently being employed for protein family discovery.

Discussion

Our analysis of over 400 publicly-available bacterial and archaeal RNA-seq datasets reveals that evidence for large numbers of RNAs of unknown function is present in public data. We find evidence for close to 1000 unannotated noncoding transcriptional outputs, but, given that RNA-seq experiments provide a snapshot of gene expression under specific experimental conditions, this number is likely to be far lower than the complete set of transcriptional outputs. Thus, the dataset we assembled for this project, which includes data generated by a number of labs and derives from various species and strains grown under a range of experimental conditions, is expected to represent a broad, though partial, census of total expression outputs across the species represented. Equally striking is the fact that, for the 922 RUFs identified in our study, over half (568) are among the most abundant transcripts. These results suggest that ncRNA may play an even greater role in the molecular workings of Bacteria and Archaea than hitherto realised.

This use of transcriptome data clearly improves our capacity to identify noncoding outputs: applying three criteria (sequence conservation, conservation of secondary structure, and expression in multiple species) we have identified 163 high-confidence expressed RUFs from public data (Figure 2). An additional 405 RUFs are highly expressed across the transcriptomes we have examined, yet these do not show clear signs of sequence or structural conservation in other sequenced genomes. Given their high expression

level, these seem unlikely to be transcriptional noise. Some may represent technical artefacts, but many could be bona fide lineage-specific ncRNAs with potentially novel functions.

Our results indicate that the greatest gain in analytical power for ncRNA discovery will come from phylogenetically-informed experimental design. Indeed, we find that this is critical to successful element identification, since the 'Goldilocks Zone' for optimal comparative analysis of RNA elements is surprisingly narrow. Hence, existing efforts to maximise phylogenetic coverage of genome space [1, 2] need to be complemented with fine-scale sampling of the tips (Figure 4). Indeed, analysing the few transcriptomes that span the Goldilocks Zone reveals a remarkable enrichment of transcripts showing evidence of structure, conservation and expression in other species.

Given that isolation, cultivation and study of individual bacterial and archaeal strains can be extremely challenging (Stewart 2012), successful phylogeny-informed comparative RNA-seq will be a demanding endeavour, requiring complex sets of expertise spanning advanced culturing and isolation techniques, functional genomics capability and RNA bioinformatics. This places such a project beyond the reach of most individual labs. We therefore propose that comprehensive resolution of the comparative RNA-seq problem can best be resolved via a community-driven initiative: in recognition of the success of the GEBA project, we have dubbed this An RNA Encyclopedia of Bacteria and Archaea (AREBA). The appropriateness of this acronym will be especially clear to Japanophones, as, in Japanese, the phrase 'areba' (あれば) translates to 'if there'.

Materials and Methods

Preprocessing and mapping

All available bacterial and archaeal genomes were downloaded from the European Nucleotide Archive (ENA) (2,562 and 154 genomes, respectively) [40]. RNA-seq datasets published as of March 2013 were collected, spanning 37 species/strains, 44 experiments and 413 lanes of sequencing data (Table S1). Most of these datasets are generated on the Illumina platform [41], with a few lanes from the SOLiD platform [42] and 454 platform [43]. Where possible, FastQ files were downloaded, scanned for residual adapter sequences using AdapterRemoval [44], and mapped to the reference genome using Bowtie2 [45] for Illumina and 454 data and BFAST [46] for SOLiD data.

Producing consistent genome annotations

All genomes were re-annotated for both RNA genes and protein coding genes. Non-coding RNA genes were annotated using cmsearch [47] to identify homologs of RNA families from the Rfam database [21,29]. Protein coding genes were annotated using three approaches: First, annotations were parsed from the ENA files. Secondly, Glimmer was run on all genomes to predict open reading frames [48]. Thirdly, all genomes were translated into all possible amino acid sequences of length 15 or more and scanned for homologs of entries in the Pfam database of protein families [49,50].

Identification of novel RNAs

From the mapped RNA-seq data, potential novel RNA genes (designated RNAs of Unknown Function, or RUFs) were picked manually by locating regions in the genomes that showed high levels of expression without overlapping annotated protein coding or RNA genes. Only RUFs of lengths 50 to 400 nucleotides were included, yielding a total of 844 RUFs in Bacteria and 78 RUFs in Archaea.

Homology search and structure prediction

Homologs of the identified RUFs were found in all the downloaded genomes using nhmmer [51] in an iterative fashion: First, the RUF sequence alone is used in the scan; then, all hits with $E-value < 0.001$ are included and a HMM is built, and this is iterated 5 times. The alignments from the RUF homology search were analyzed further by investigating the potential for secondary structure formation using RNAz [52] and alifoldz [53]; the protein coding potential using RNAcode [54], and overlaps between potential RUF homologs in other species and all the annotations in the respective genomes as described above.

Comparative expression and conservation analysis

For each species, the available RNA-seq datasets were pooled and a list was created of transcripts showing expression in that species in at least one experiment (defined as a transcript having a median depth of at least 10 reads in any experiment). A RUF homolog was defined as being expressed if the median read depth of the homologous region was at least 10X. Transcripts were ranked for each species based on median expression (i.e. the most highly expressed transcript will have rank 1), which makes relative comparison across species and datasets possible. The final set comprises 452 different Rfam families, 922 different RUFs, and 7249 different Pfam domains.

For comparative analysis, if a gene was found to be expressed in more than one species, the minimum rank was used (i.e. showing the relatively most abundant expression of the gene). This ensures that transcripts that are always low abundant will remain low abundant, whereas genes that are highly abundant in at least one of the sampled time points and conditions will be treated as such. The ranking is used as a measure of expression.

Conservation is based on SSU rRNA alignments of all Bacteria and Archaea, respectively. For each genome, the best hit to the Rfam model of SSU rRNA was extracted (RF00177 for Bacteria and RF01959 for Archaea). The sequences were aligned to the model using cmalign [47]. Finally, a distance matrix was calculated using dnadist [55] with the F84 model [56, 57] which both allows for different transition/transversion rates and for different nucleotide frequencies. The pairwise species distances produced in this manner estimate the total branch lengths between any pair of species. For any gene found in two or more species, the maximum pairwise distance is used as the conservation score. Upper and lower quartiles of the distributions are used to define sets of high, medium and low expression and conservation, respectively. (Expression, upper quartile: 204. Expression, lower quartile: 1660. Conservation, upper quartile: 0.478. Conservation, lower quartile: 0.267).

Quality control of RNA-seq datasets

We ranked datasets based on the following quality control metrics (values reported in Table S3).

Strand correlation: We calculated correlation between the reads on the two strands. If the dataset is unstranded, we expect a correlation close to 1.

Expression of core genes: We defined a set of 40 core protein-coding genes based on [**?**] and 16 noncoding RNA genes (the union of tRNA, RNaseP, tmRNA, SRP, 6S and rRNA RNA families) [21,29]. If the median read depth greater than 10X, we defined the gene as expressed. For each dataset, we report the fraction of the 40 core genes that are expressed.

Coverage: We calculated the coverage as the fraction of the genome covered by at least 10 mapped reads.

Fraction mapped reads: For each dataset, we ascertained the fraction of mapped reads.

Concordance: To measure how well a given RNA-seq dataset corresponds to the annotated genes in a genome, we developed a concordance metric. For this, we define true positives (TP) to be the number of annotated positions that are expressed; false positives (FP) to be the number of unannotated positions that are expressed; true negatives (TN) to be the number of unannotated positions that are

not expressed; and false negatives (FN) to be the number of annotated positions that are not expressed. Note, not all annotated genes are expected to be expressed, and not all unannotated positions are false. Therefore, we calculate the positive predictive value (PPV):

$$PPV = \frac{TP}{TP + FP}$$

This measures the fraction of expressed positions that are annotated. We also calculate the fraction of the genome that is annotated:

$$ANN = \frac{TP + FN}{TP + FP + TN + FN}$$

To make the PPV more robust, our final concordance metric normalizes PPV by ANN.

Acknowledgments

References

1. Wu D, Hugenholtz P, Mavromatis K, Pukall R, Dalin E, et al. (2009) A phylogeny-driven genomic encyclopaedia of bacteria and archaea. Nature 462: 1056-60.

2. Rinke C, Schwientek P, Sczyrba A, Ivanova NN, Anderson IJ, et al. (2013) Insights into the phylogeny and coding potential of microbial dark matter. Nature 499: 431-7.

3. Loman NJ, Constantinidou C, Chan JZ, Halachev M, Sergeant M, et al. (2012) High-throughput bacterial genome sequencing: an embarrassment of choice, a world of opportunity. Nat Rev Microbiol 10: 599-606.

4. Chun J, Rainey FA (2014) Integrating genomics into the taxonomy and systematics of the bacteria and archaea. Int J Syst Evol Microbiol 64: 316-24.

5. Sorek R, Serrano L (2011) Bacterial genomes: from regulatory complexity to engineering. Curr Opin Microbiol 14: 577-8.

6. Storz G, Vogel J, Wassarman KM (2011) Regulation by small rnas in bacteria: expanding frontiers. Mol Cell 43: 880-91.

7. Dennis PP, Omer A, Lowe T (2001) A guided tour: small rna function in archaea. Mol Microbiol 40: 509-19.

8. Horvath P, Barrangou R (2010) Crispr/cas, the immune system of bacteria and archaea. Science 327: 167-70.

9. Breaker RR (2012) Riboswitches and the rna world. Cold Spring Harb Perspect Biol 4.

10. Cech TR, Steitz JA (2014) The noncoding rna revolution-trashing old rules to forge new ones. Cell 157: 77-94.

11. Pagani I, Liolios K, Jansson J, Chen IM, Smirnova T, et al. (2012) The genomes online database (gold) v.4: status of genomic and metagenomic projects and their associated metadata. Nucleic Acids Res 40: D571-9.

12. Stewart EJ (2012) Growing unculturable bacteria. J Bacteriol 194: 4151-60.

13. Elkins JG, Podar M, Graham DE, Makarova KS, Wolf Y, et al. (2008) A korarchaeal genome reveals insights into the evolution of the archaea. Proc Natl Acad Sci U S A 105: 8102-7.

14. Könneke M, Bernhard AE, de la Torre JR, Walker CB, Waterbury JB, et al. (2005) Isolation of an autotrophic ammonia-oxidizing marine archaeon. Nature 437: 543-6.

15. Tyson GW, Chapman J, Hugenholtz P, Allen EE, Ram RJ, et al. (2004) Community structure and metabolism through reconstruction of microbial genomes from the environment. Nature 428: 37-43.

16. Woyke T, Teeling H, Ivanova NN, Huntemann M, Richter M, et al. (2006) Symbiosis insights through metagenomic analysis of a microbial consortium. Nature 443: 950-5.

17. Mandin P, Toledo-Arana A, Fouquier d'Hérouel A, Repoila F (2013) Rna-mediated control of bacterial gene expression: Role of regulatory non-coding rnas. Encyclopedia of Molecular Cell Biology and Molecular Medicine .

18. Freyhult EK, Bollback JP, Gardner PP (2007) Exploring genomic dark matter: a critical assessment of the performance of homology search methods on noncoding RNA. Genome Res 17: 117-125.

19. Nawrocki EP, Kolbe DL, Eddy SR (2009) Infernal 1.0: inference of RNA alignments. Bioinformatics 25: 1335-7.

20. Barrick JE, Breaker RR (2007) The distributions, mechanisms, and structures of metabolite-binding riboswitches. Genome Biol 8: R239.

21. Gardner PP, Daub J, Tate J, Moore BL, Osuch IH, et al. (2011) Rfam: Wikipedia, clans and the "decimal" release. Nucleic Acids Res 39: D141-5.

22. Chen YJ, Liu P, Nielsen AA, Brophy JA, Clancy K, et al. (2013) Characterization of 582 natural and synthetic terminators and quantification of their design constraints. Nat Methods 10: 659-64.

23. Cambray G, Guimaraes JC, Mutalik VK, Lam C, Mai QA, et al. (2013) Measurement and modeling of intrinsic transcription terminators. Nucleic Acids Res 41: 5139-48.

24. Roth A, Weinberg Z, Chen AG, Kim PB, Ames TD, et al. (2014) A widespread self-cleaving ribozyme class is revealed by bioinformatics. Nat Chem Biol 10: 56-60.

25. Barrangou R, Fremaux C, Deveau H, Richards M, Boyaval P, et al. (2007) Crispr provides acquired resistance against viruses in prokaryotes. Science 315: 1709-12.

26. Loh E, Kugelberg E, Tracy A, Zhang Q, Gollan B, et al. (2013) Temperature triggers immune evasion by neisseria meningitidis. Nature 502: 237-40.

27. Omer AD, Lowe TM, Russell AG, Ebhardt H, Eddy SR, et al. (2000) Homologs of small nucleolar rnas in archaea. Science 288: 517-22.

28. Gardner PP, Bateman A, Poole AM (2010) Snopatrol: how many snorna genes are there? J Biol 9: 4.

29. Burge SW, Daub J, Eberhardt R, Tate J, Barquist L, et al. (2013) Rfam 11.0: 10 years of rna families. Nucleic Acids Res 41: D226-32.

30. Hoeppner MP, Gardner PP, Poole AM (2012) Comparative analysis of rna families reveals distinct repertoires for each domain of life. PLoS Comput Biol 8: e1002752.

31. Croucher NJ, Thomson NR (2010) Studying bacterial transcriptomes using rna-seq. Curr Opin Microbiol 13: 619-24.

32. van Opijnen T, Camilli A (2013) Transposon insertion sequencing: a new tool for systems-level analysis of microorganisms. Nat Rev Microbiol 11: 435-42.

33. Barquist L, Boinett CJ, Cain AK (2013) Approaches to querying bacterial genomes with transposon-insertion sequencing. RNA Biol 10: 1161-9.

34. Barquist L, Langridge GC, Turner DJ, Phan MD, Turner AK, et al. (2013) A comparison of dense transposon insertion libraries in the salmonella serovars typhi and typhimurium. Nucleic Acids Res 41: 4549-64.

35. Rocha EP, Danchin A (2003) Essentiality, not expressiveness, drives gene-strand bias in bacteria. Nat Genet 34: 377-8.

36. Pál C, Papp B, Hurst LD (2001) Highly expressed genes in yeast evolve slowly. Genetics 158: 927-31.

37. Drummond DA, Bloom JD, Adami C, Wilke CO, Arnold FH (2005) Why highly expressed proteins evolve slowly. Proc Natl Acad Sci U S A 102: 14338-43.

38. Gottesman S, Storz G (2011) Bacterial small rna regulators: versatile roles and rapidly evolving variations. Cold Spring Harb Perspect Biol 3.

39. Gardner PP, Wilm A, Washietl S (2005) A benchmark of multiple sequence alignment programs upon structural rnas. Nucleic Acids Res 33: 2433-9.

40. Cochrane G, Alako B, Amid C, Bower L, Cerdeño-Tárraga A, et al. (2013) Facing growth in the european nucleotide archive. Nucleic Acids Res 41: D30-5.

41. Shendure J, Porreca GJ, Reppas NB, Lin X, McCutcheon JP, et al. (2005) Accurate multiplex polony sequencing of an evolved bacterial genome. Science 309: 1728-32.

42. Valouev A, Ichikawa J, Tonthat T, Stuart J, Ranade S, et al. (2008) A high-resolution, nucleosome position map of c. elegans reveals a lack of universal sequence-dictated positioning. Genome Res 18: 1051-63.

43. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, et al. (2005) Genome sequencing in microfabricated high-density picolitre reactors. Nature 437: 376-80.

44. Lindgreen S (2012) Adapterremoval: easy cleaning of next-generation sequencing reads. BMC Res Notes 5: 337.

45. Langmead B, Salzberg SL (2012) Fast gapped-read alignment with bowtie 2. Nat Methods 9: 357-9.

46. Homer N, Merriman B, Nelson SF (2009) Bfast: an alignment tool for large scale genome resequencing. PLoS One 4: e7767.

47. Nawrocki EP, Eddy SR (2013) Infernal 1.1: 100-fold faster rna homology searches. Bioinformatics 29: 2933-5.

48. Delcher AL, Bratke KA, Powers EC, Salzberg SL (2007) Identifying bacterial genes and endosymbiont dna with glimmer. Bioinformatics 23: 673-9.

49. Punta M, Coggill PC, Eberhardt RY, Mistry J, Tate J, et al. (2012) The pfam protein families database. Nucleic Acids Res 40: D290-301.

50. Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, et al. (2014) Pfam: the protein families database. Nucleic Acids Res 42: D222-30.

51. Wheeler TJ, Eddy SR (2013) nhmmer: Dna homology search with profile hmms. Bioinformatics 29: 2487-9.

52. Gruber AR, Findeiss S, Washietl S, Hofacker IL, Stadler PF (2010) Rnaz 2.0: improved noncoding rna detection. Pac Symp Biocomput : 69-79.

53. Washietl S, Hofacker IL (2004) Consensus folding of aligned sequences as a new measure for the detection of functional rnas by comparative genomics. J Mol Biol 342: 19-30.

54. Washietl S, Findeiss S, Müller SA, Kalkhof S, von Bergen M, et al. (2011) Rnacode: robust discrimination of coding and noncoding regions in comparative sequence data. RNA 17: 578-94.

55. J F. Phylip (phylogeny inference package) version 3.6. Distributed by the author. Department of Genome Sciences, University of Washington, Seattle.

56. Kishino H, Hasegawa M (1989) Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from dna sequence data, and the branching order in hominoidea. J Mol Evol 29: 170-9.

57. Felsenstein J, Churchill GA (1996) A hidden markov model approach to variation among sites in rate of evolution. Mol Biol Evol 13: 93-104.
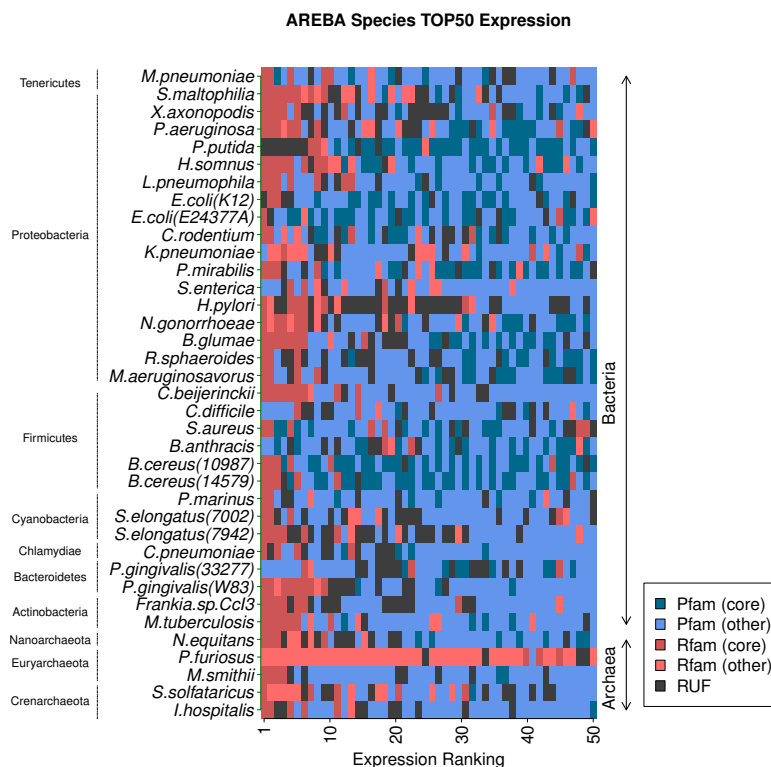
Figure Legends

Figure 1. Identification of transcribed elements across publicly-available RNA-seq data.
Non-coding RNA elements show high expression across transcriptomes. Both annotated Rfam families
(red – core Rfam families are dark red, all others are light red) and expressed RUFs (black) are among
the highest expressed outputs in transcriptomes. Protein-coding regions annotated using Pfam are also
shown (blue). For each species we generated relative rankings of expression spanning protein coding
genes, RNA genes and candidate RUF genes. Genes were extracted from mapping data by comparing
to the annotation. Going from mapped reads to expression level can be confounded by a number of
factors (e.g. sample preparation, overall sequencing depths, rRNA depletion etc.). For consistency, we
therefore ranked genes for each species and compared rankings instead of comparing the read depths
directly between species. For a given species, the annotated genes were ranked based on the median
read depth of the annotated region. RUFs were manually picked by masking out annotated genes and
selecting regions showing evidence of expression by inspecting read depth across the genome. This
yielded 844 gene candidate sequences in Bacteria (78 in Archaea). The plot contains the 50 most highly
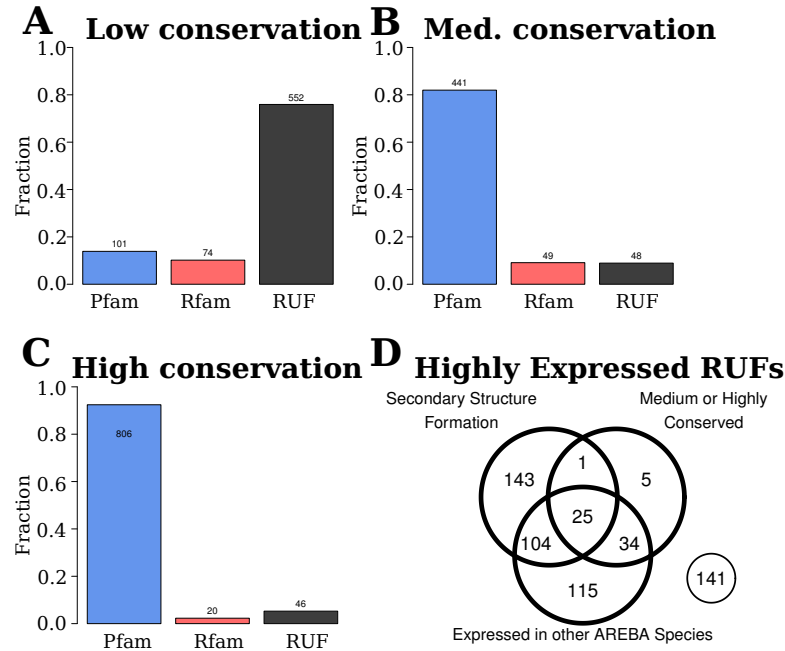expressed elements for each species.

Figure 2. Many ncRNAs and RUFs are highly expressed but show limited conservation across represented species. A-C: Both known Rfam families and the RUFs identified in this analysis are often highly expressed transcripts. In contrast to protein-coding transcripts (blue), where highly-expressed transcripts are well-conserved, the opposite is true of many non-coding RNA elements (Rfam, red; RUFs, black). Notably, the greatest proportion of highly expressed Rfam-annotated RNA elements show a narrow evolutionary distribution. This is also reflected in the RUFs identified in this study. **D:** Venn diagram of the 568 highly expressed RUFs. Each RUF was analysed to look for evidence of secondary structure formation, level of conservation, and evidence of expression in at least one other RNA-seq dataset. All RUFs showing expression in other AREBA species must of course be conserved in at least two species, so the figure also shows that 219 highly expressed RUFs are conserved across a limited phylogenetic distance only.
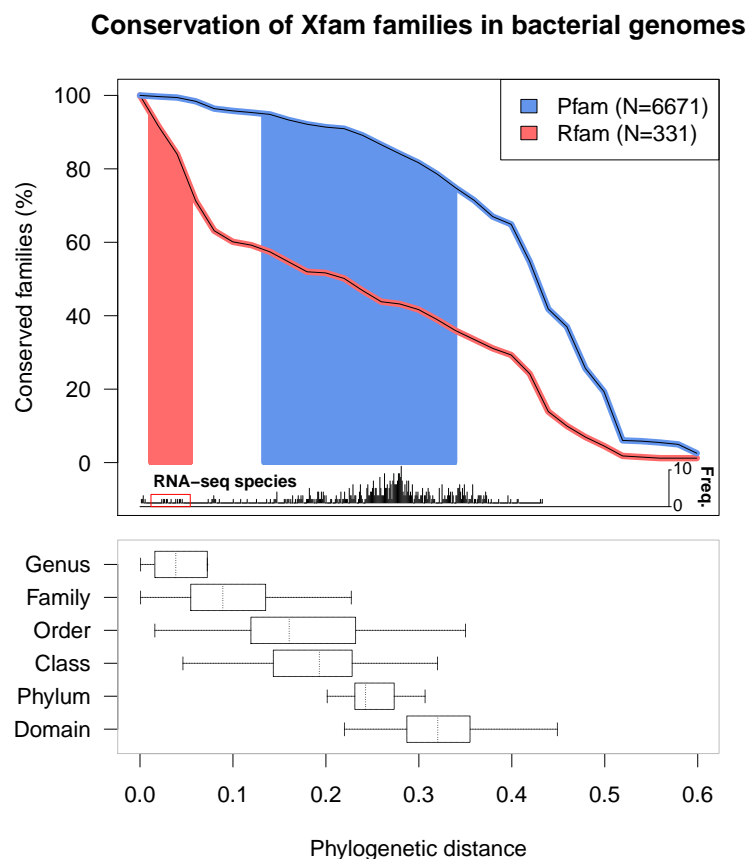
Conservation of Xfam families in bacterial genomes

Figure 3. Conservation of protein and RNA families. Top: For the protein and the RNA families we compare the levels of conservation as a function of phylogenetic distance. E.g. $\approx 60\%$ of RNA families are conserved between species from the same family, whereas $> 90\%$ of protein families are conserved within the same taxonomic range. Middle: The barplot shows the distribution of all pairwise distances between the RNA-seq datasets. Eleven pairs (boxed) are in the Goldilocks Zone (See Figure 4 for further analysis). Bottom: The ranges of phylogenetic distances for comparing species from different taxonomic groups.
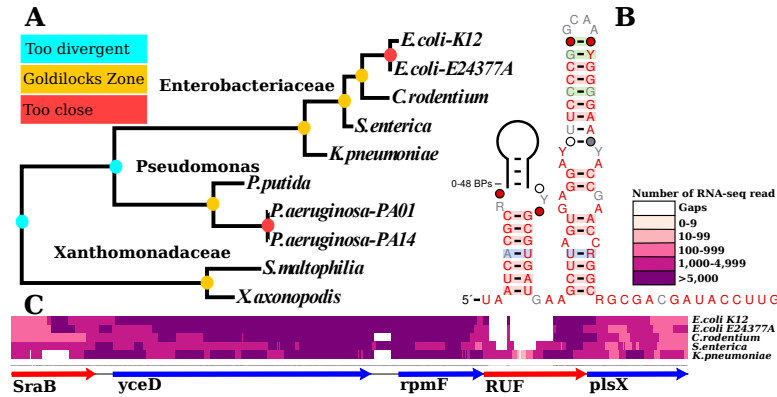
Figure 4. Comparative analysis of RNA-seq datasets in the 'Goldilocks Zone' is a powerful approach for identifying 'RNAs of Unknown Function'. (A) Ten species with available RNA-seq data and phylogenetic distances in the Goldilocks Zone (GZ) (Figure 3) have been identified. The maximum likelihood tree from a SSU rRNA alignment shows the relationships between these taxa. They fall into three clades, containing members of the families: Enterobacteriaceae and Xanthomonadaceae, and the genus: Pseudomonas. Nodes connecting taxa within the GZ are coloured gold, taxa that are too close are coloured red and those that are too divergent are coloured cyan. **(B)** The predicted secondary structure for an exemplar RNA of Unknown Function (RUF) chosen from the Enterobacteriaceae data. This RUF is highly expressed in each species (C), shows co-variation within the second hairpin (highlighted in green, structure neutral variation is highlighted in blue, highly conserved regions are highlighted in pink), it conserves a tetraloop of the GNRA or UNCG types and there have been two independent insertions of hairpins in the S. enterica and K. pneumoniae within the first hairpin. **(C)** The expression levels inferred from RNA-seq in the locus containing the RUF illustrated in B. The locus contains a ncRNA (red: SraB) and three protein coding genes (blue: yceD, rpmF and plsX) and the RUF (red). For each nucleotide in the locus the total number of reads that map there was computed, these are illustrated in the heatmap; Darker colours indicate high relative expression, lighter colours indicate low expression and white indicates a gap in the alignment of the sequences for this locus from the five Enterobacteriaceae.