UNIVERSITY OF CALIFORNIA

SANTA CRUZ

Analyzing Regulation of tRNAs, tRNA Fragments, and mRNAs in Whole Genomes

A dissertation submitted in partial satisfaction
of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

BIOMOLECULAR ENGINEERING & BIOINFORMATICS

by

Andrew David Holmes

September 2018

The Dissertation of Andrew D. Holmes
is approved:

____

Professor Todd M. J. Lowe, Chair

____

Professor Chad Saltikov

____

Professor Jeremy Sanford

____

Lori Kletzer

Vice Provost and Dean of Graduate Studies

Analyzing Regulation of tRNAs, tRNA Fragments, and mRNAs in Whole Genomes

Contents

List of Figures

List of Tables

Abstract

Analyzing Regulation of tRNAs, tRNA Fragments, and mRNAs in Whole Genomes

by

Andrew David Holmes

Rather than focus on individual pathways or genes, whole-genome analysis of regulation allows for the discovery of the biological processes that drive cells decisions and fate in single-celled and multicellular organisms. I report on three different forms of whole-genome analysis of gene regulation. The first is analysis of tRNA fragments using RNA-sequencing. I develop a computational method to analyze changes in RNA sequencing data from tRNAs and tRNA fragments and use this method to analyze changes in RNA-sequencing results from the use of demethylase treatment, PNK end treatment, and RNA sequencing kit to determine how RNA-sequencing results of tRNA can change from experiment to experiment. I also analyze tRNA sequencing on a chromatin level using previously published genome-wide chromatin data. Here, I categorize tRNAs based on their expression and compare these tRNAs to examine what the determinants of expression are. Using this expression data, I also build a method to compare tRNAs across related species and find how tRNA conservation varies with expression. Finally, I examine cis-regulatory structural mRNA elements in archaea. I find that the known types of cis-regulatory element in bacteria are conserved, and then I use a combination of RNA sequencing data and methods for finding conserved RNA structure to predict these elements in three archaeal species. I build a list of candidate structural elements including new ribosomal autoregulatory elements in each species.

Acknowledgements

I'd like to thank David Bernick, Julie Murphy, and Andrew Smith for developing and performing RNA-seq in archaea along with growing archaeal species. I'd also like to thank David Bernick in particularly for starting the work of performing small-RNA sequencing in archaea. I would like to thank Aaron Cozen for a great deal of very useful feedback on the computational tools that were developed into TRAX, as well as building the epigenomic classifications. I'd like to thank Eva Robinson for doing the RNA-seq work for tRNA sequencing as well as discussing the implications of RNA-sequencing preparations. I would like to thank David Bernick for helping to teach me about archaea and computational biology when I was just starting as an undergraduate researcher and I'd like to thank Andrew Uzilov for helping me when I was starting out and teaching me how to perform computational biology. I'd like to thank Jonathan Howard for help in editing manuscripts. I'd like to thank Lauren Lui, Brian Lin, and Bryan Thornlow for general discussion. Todd Lowe, in addition to general advising, suggested small RNA reads could be signs of mRNA secondary structure. He also gave me my start in computational biology an undergraduate working in his lab.

Introduction

Gene regulation can take many forms, at the transcription step, at the translation step, or just through the degradation of the RNA. Each gene can pass through numerous steps at which it's expression can be regulated, and the gene shut down. Each step can be stopped in multiple ways to shut down expression of that gene. This is the case for both non-coding and coding RNAs, although it occurs through different methods. Understanding these steps is crucial for understanding how the cell functions and attempting to alter the activity of the cell either through changing its substrate or through manipulation of the genome. More than understanding it for individual genes though, building methods for analyzing entire genomes allows for the understanding of the patterns that guide the entire cell, which is useful for understanding phenomena like aging, cancer, or growth of single-celled organisms.

This thesis develops methods for analyzing tRNA fragmentation, tRNA expression through chromatin state, and mRNA cis-regulation through mRNA structure. Analyzing RNA fragmentation is performed with a pipeline I developed (TRAX) and then used on a set of data both already published and sequenced in-lab for examining the consequences of differences in library preparation. Expression through chromatin state was performed by using published data from the epigenomic road and applying it to tRNAs while comparing it to Pol III data, tRNA sequence, DNA methylation, tRNA duplication, and conservation in related species to find the pattern in tRNA expression. Analyzing cis-regulation in mRNA structure was done by combining a method for finding conserved RNA structure in multiple alignments with small-RNA sequencing reads that appear to act as a signal of cis-regulatory elements

Small RNA sequencing for the archaeal work was performed by David

Bernick, Julie Murphy and Andrew Smith. Small RNA sequencing of tRNAs that was not from previously published work was performed by Eva Robinson or Erin Quartley, and TRAX was developed with input from Aaron Cozen who also did the analysis of ARM-seq effects on m1a modification in human and yeast cells.[1] Epigenomic group categorization was done Aaron Cozen who assisted in building the method for determining tRNA states and Todd Lowe who developed the rules for tRNA epigroup categorization.

Background

Transfer RNAs

Transfer RNAs, or tRNAs, are a form of non-coding RNA that are a part of the process of translation. tRNAs function in the translation process as the adapter molecule that converts the three-base codons into their specific amino acid. tRNA molecules do this translation by having a specific three-base anticodon that base-pairs with the mRNA codon to allow decoding. This anticodon-codon pairing in the ribosome is what allows a new amino acid to be added to the peptide chain.[2]

Codon-anticodon pairing in tRNAs has a specific pattern that allows it to decode mRNA codons. The structure of the ribosome requires that the first two bases of the codon use Watson-Crick base pairing rules, but the third base in the codon is allowed to form G-U base pairs that allow a single tRNA to decode multiple codons. This position is known as the wobble position due to the structure of these non-Watson-Crick base pairs. This third position also is sometimes modified to an Inosine base that allows that tRNA to decode adenine, cytosine, and uracil bases. This is partly responsible for the degeneracy of the genetic code and is why some anticodons are often missing in the genome.[3]

Amino acids are attached to tRNAs by proteins that are known as tRNA synthetases. These molecules read the tRNA sequence and structure, especially the anticodon, and attach the specific amino acid that the tRNA decodes. Generally, each amino acid has a specific tRNA synthetase that can match all tRNAs that decode it.[4]

Mature transfer RNAs have a specific RNA cloverleaf secondary structure that is composed of 4 arms. The first arm, the acceptor arm, is the only arm not containing a loop and is where the tRNA synthetase attaches the amino acid. The anticodon arm is the arm responsible for decoding tRNAs and contains the anticodon sequence in a loop. The other two arms are known as the D arm and the T arm due to the RNA modifications that are commonly found in them. Some tRNAs also possess another arm, known as the variable loop.[5]

Genomic tRNAs in eukaryotes

Eukaryotes and prokaryotes have different configurations of tRNA genes in their genomes. In prokaryotic species, there will often be the minimum number of tRNAs required to decode all anticodons, while others can have two or sometimes three copies of a tRNA. In eukaryotes, tRNAs exist in many exact and inexact copies.[6] One purpose for this tRNA duplication is to increase the transcription of tRNAs, this is required for tRNAs because unlike protein-coding genes there is no step after transcription that can be used to amplify output.[7] Another reason for duplication of tRNAs is to increase allow for more fine-grained control over translation of different codons, this can be used in conjunction with mRNA codon frequency to favor or disfavor certain genes at certain cell stages.[8] Mitochondrial genomes possess their own tRNAs, and their tRNAs are configured similarly to prokaryotes with one copy of each tRNA. Mitochondrial tRNAs also can be aberrant, sometimes missing arms entirely. [9]

Transfer RNAs in the genome also can possess introns. These introns are spliced out during the maturation process using specialized tRNA intron processing pathway that is unrelated to mRNA splicing. Most tRNA genes in the genome do not possess introns, they are present in a small proportion of human tRNAs. Introns are present even in prokaryotes, indicating that this is an ancient property of tRNAs.[10]

Duplication of tRNAs can occur for functional reasons, but tRNAs are also known to be highly transposed genes. In eukaryotes, this can happen in the form of SINE's[11] and segmental duplications[12] that include tRNA genes. This duplication results in many tRNA copies in eukaryotic genomes that can be either functional or non-functional. tRNAs are also thought along with some forms of ribosomal RNA genes to be subject to concerted evolution[13], where genes that exist in multiple copies in the genome are homogenized.[14]

tRNAs are known to be concentrated in the genome, often existing in clusters of genes in close proximity and on specific chromosomes.[15] In the human genome, tRNAs are most abundant in two locations, one location on chromosome one and one location on chromosome six. The chromosome six cluster is located near the major histocompatibility complex.[12] In mouse, there is a single tRNA cluster on chromosome thirteen.

The annotated set of tRNAs genes in most sequenced organisms are those predicted using the program tRNAscan-SE.[16] This program searches genomes for regions that match the sequence and structure of known tRNAs using a covariance model[17] to find tRNA genes in a sequenced genome. The quality of the match to known sequence is expressed in the form of a covariance model score that ranks how close tRNAs are to the canonical tRNA structure.

Transcription of tRNAs

tRNAs transcription is performed by a gene known as RNA polymerase III.[18] The main signal for tRNA transcription is the presence of the A and B boxes, small segments of RNA located in the D and T arms of the RNA that allow for the tRNA to be transcribed. Transcription of tRNAs transcripts is done with a specific terminator sequence consisting of a run of T bases, known as poly-T terminator.[19]

The mechanisms of tRNA regulation are not well understood. One primary regulator of tRNA genes in the Maf1 protein. This protein is not known to target specific DNA sequences it instead functions as a universal regulator of tRNA transcription that binds RNA polymerase III directly.[20] It is known however that merely sequence does not determine tRNA transcription, that multiple identical tRNAs can have different transcription patterns.[7]

Post-transcriptional processing of tRNAs

tRNAs when they are initially transcribed in eukaryotes contain extra sequence on the 5-prime and 3-prime end as well as missing the CCA tail and base modifications present in the tRNA.[21] The 5-prime bases are removed by RNAse P, an RNA/protein complex that uses RNA as its catalytic element. The 3-prime end is removed by RNAse Z. Splicing occurs with some tRNAs after this step, a specific "Bulge-helix-Bulge" structure in the anticodon loop is removed by an endonuclease and a ligase. After this, the 3-prime end has the sequence "CCA" added to it as the site of aminoacylation. After this, the tRNA base modification and aminoacylation can occur.[2]

There are many types of tRNA modifications that exist across all domains of life, including A-to-I editing, base methylation, pseudouridylation, and 2' O methylation.[21] Some of these modifications occur in all tRNAs, but many exist in only some tRNAs.[22] A-to-I editing, for instance, is found in the first base of the anticodon, and is partially responsible for the effect of the "wobble" third base position in tRNA decoding.[23] A-to-I editing is also found in the 3-prime end of the anticodon loop, where the inosine base is methylated. While the determinants for some modifications are understood, the determinants of many modification enzymes is not known.[24]

Base methylation can positively affect the stability of tRNA secondary structure.[25] Very few of these modifications are critical for tRNA function, but lack of modifications can target tRNAs for degradation. There are many types of base methylations, targeting all 4 standard RNA bases in addition to inosine. Specific bases are generally methylated at specific positions in the tRNA structure.[23] Changes to tRNA modifications are associated with several disease, including neurologic conditions, diabetes, and cancer.[26]

In addition to methylation, base editing occurs in some tRNAs. In addition to the conversion of adenosine to inosine, cytosine is converted to uridine, and uridine is converted to pseudouridine. While inosine editing can affect base-pairing in the anticodon, the function of many base modifications is not clear.[25]

These modifications also can can hinder any attempt to reverse transcribe these RNAs. Of the modifications present in tRNAs, 1-methyladenosine, 1-methylguanosine, N2,2-methylguanosine, and 3-methylcytosine are known to cause stops in reverse transcription.[27] pseudouridine is known to causes pauses in reverse transcription while wyobutosine can cause it to terminate.[27] The inosine base does not cause halting in reverse transcription, but is read as a guanosine instead of the original adenosine base.[27] The halting effect that these bases have on reverse transcription has been used to to find these modifications in the past using primer extension, and it also affects any attempt to sequence and map tRNAs.[27,28]

Base modification is also known to be used in other types of RNA, although this has not been studied extensively. 3-methylcytosine is the most commonly found mRNA modification, but 6-methylcytosine and possibly 1-methyladenosine has also been found in human mRNA.[29]

tRNA expression

Recent work on tRNAs have shown the effect that tRNA levels have on cancer and cell state.[30] Small RNA sequencing has been done on both tRNA and tRNA fragments to determine expression of tRNAs and levels of tRNA fragments.[1,31] However, the duplication of tRNAs in the genome makes RNA sequencing insufficient to determine the expression of RNA, as reads cannot be mapped to a unique tRNA locus.[32]

Chapter 1: Analyzing RNA-seq data in tRNAs using TRAX (TRna Analysis of eXpression), a new computational pipeline

Abstract

Small RNA Sequencing of tRNAs has been used to measure expression of tRNAs and tRNA fragments, but features specific to tRNA such as RNA modification, duplication, structure and chemistry different ribonucleases make analyzing tRNA sequencing data difficult. Here, I develop a computation method for analysis of tRNA sequencing data to solve some the problems in sequencing analysis and use it to show the effects of different sequencing kits, demethylation, end treatment, and linking protocols to show how tRNA sequencing results can differ between different experiments.

Background

tRNA fragmentation

While the function of tRNAs in protein translation is well known, recent work studied the creation and use of tRNA fragments. tRNA fragments were originally thought to be the nonfunctional result of RNA breakdown. More recent work as shown that that these tRNA fragments are functional products created as a product of specific pathways that have functional roles in the cell.[33]

One form of tRNA fragments is tRNA that has been well-studied are those generated by angiogenin. The angiogenin protein cleaves tRNA right before the anticodon into two tRNA halves.[34] Another common form of tRNA fragment are the dicer-derived fragments. Dicer cleaves tRNAs at the D or T stems, leaving shorter tRNA fragments.[35] Pre-tRNAs are also a source for some tRNA fragments that are generated as part of the maturation process. RNAse P cleavage of pre-tRNAs generates a product that in some human tRNAs appears to continue to remain in the cell.[36]

Functionally, tRNA fragments have been shown to modulate global gene expression through Argonaute-dependent silencing mechanisms, reduction of protein translation efficiency, and stimulation of target RNA degradation. tsRNAs also been implicated in a number of stress-induced regulatory mechanisms associated with viral infections and disease. Three-prime tRNA fragments and dicer-derived fragments both have been reported to function as microRNAs. However, there are several features of tRNAs and tRNA fragments that make them difficult to sequence and analyze using standard small-RNA sequencing protocols.[4] What regulates tRNA fragmentation is not fully known, but tRNA fragmentation is known to be associated with aging,[37] and cancer.[38] The growing interest in the importance of tRNA fragments necessitates protocols uniquely adapted to these atypical RNAs.

Small-RNA sequencing

Read mapping is a critical part of any small RNA-seq analysis. Transfer RNAs and their fragments have several attributes that make their mapping difficult. One of these factors is that tRNAs exist in many copies in the cell, both as perfect and imperfect copies spread throughout the genome.[5] It is often impossible to uniquely map a tRNA fragment to a specific tRNA locus, and any analysis must either accept some ambiguity in tRNA fragment mapping or remove large numbers of reads.

tRNAs also have difficulty with their post-transcription processing. All mature tRNA genes contain a CCA tail that is added post-transcriptionally to the mature tRNA.[39] Introns are also present in some tRNAs that are not present in the mature sequence from which fragments are generated.[10] These changes create differences between the mature tRNA sequence and fragments of the mature sequence and the sequence of the genome.

In addition to analysis, several issues exist with most protocols for generating tRNA sequencing libraries. One of these issues is the effect of RNA modification during cDNA synthesis. The reverse transcriptases used to create cDNA for RNA-seq library preparation have difficulty transcribing these modifications.[7] Often, these modifications cause the reverse transcriptase to "fall off" of the RNA, leading to truncated cDNAs that are not amplified during library PCR, or misincorporation of bases during read-through.[8–11] While reverse transcription termination can generate partial reads in some RNA-seq protocols, small-RNA sequencing protocols that require a 5′ linker termination such as Illumina TruSeq and New England Biolabs NEBNext small RNA sequencing kits results in a fragment that does not complete the full RNA sequence being unreadable.[40]

Another issue with sequencing tRNAs and tRNA fragments results from the linker ligation step. Most sequencing protocols require a RNA molecule containing a 3'- OH and a 5' phosphate for appropriate linker ligation.[40] While mature tRNAs do have this chemistry[13], tRNA fragments may not due to endonuclease fragmentation, resulting in subpopulations of tRNA fragments that are unlinkable. There is also the closely related issue of end accessibility that can affect RNA ligation. In mature tRNAs, the 5' end is involved in base-paired acceptor stem secondary structure, which can negatively influence linker ligation steps.[14] As a consequence of difficulties in linker ligation, and reverse transcriptase's failure to read through RNA modifications small-RNA sequencing rarely sequences full mature tRNAs.[12]

Recently published sequencing protocols have taken into account many of the aforementioned difficulties of tRNA and tsRNAs sequencing. For example, ARM-seq and DM-tRNA-seq use pre-treatments of RNA with E. coli AlkB to remove commonly methylated tRNA residues, facilitating read-through of reverse transcriptase. New sequencing library preparation protocols utilize thermostable group II intron reverse transcriptase (TGIRT) which does not require ligation and instead uses template switching on the five-prime end during adaptor attachment.[15]

Results

TRAX analysis pipeline

TRAX as a pipeline designed to analyze RNA-sequencing data. This pipeline takes FASTQ files as input, along with a sample description that provides information on experiment replication, and outputs a set of results in the form of tables and plots.

Instead of mapping to just the genome, TRAX uses a gene-centric approach that maps reads to the mature sequences of tRNAs. These mature sequences lack introns and include the CCA tail present in mature tRNAs. However, due to the duplication of tRNA sequences, TRAX maps to unique tRNA sequences found in the genome rather than to sequences of tRNA loci. To allow this, TRAX creates a tRNA database for each organism that uses the genome sequence of the organism and a set of tRNA predictions. Reads are mapped to both the mature tRNA sequence and the genome sequence to ensure that reads mapped to tRNAs match tRNA sequence better than another region in the genome.

Unlike other small-RNA analysis pipelines, TRAX by default allows for sequence mismatches in read mapping. While this can result in mismapping, the presence of RNA modifications mean that the actual mature sequence of the genome may not be known. Additionally, this allows for the sequencing of reads that have modification-induced mismatches from RNA modification.

Even when duplicate tRNAs are not included separately in the sequencing database, many tRNAs are almost duplicates and therefore mapping can be ambiguous. If multiple tRNAs can all map equally well to the read, then all mappings are reported. Mappings that are worse than the best mapping are not reported. If a read maps equally well to a tRNA sequence and a region of the genome, then only the reads that map to the tRNA sequence are reported. The number of tRNAs, isodecoders, and isoacceptors that each tRNA can map to is recorded as part of the mapping analysis.

TRAX can separate tRNA reads into specific mutually exclusive fragment types. These fragment types include five-prime fragments, three-prime fragments, whole tRNAs, and trailer fragments derived from RNAse P cleavage of pre-tRNAs. These correspond to published fragment types that have been characterized in previous papers.[36] Counts for tRNA fragment types are combined with counts from non-coding RNAs to normalize reads counts between experiments using DESeq2[41], and read counts of all fragment types and non-coding RNA genes supplied are output. Separation of tRNAs into fragment types can be disabled, this is recommended with dm-tRNAseq or other protocols for sequencing full-length mature tRNAs which otherwise may result in mischaracterization of reverse transcriptase dropoff as tRNA fragmentation.

This system allows for small-RNA sequencing reads to be easily mapped to the genome. This method is not designed with any specificity of input beyond small-RNA sequencing, TRAX can be used to analyze micro-RNA experiments for tRNA data to find information that may have been missed in the initial analysis.

Visualization of tRNAs and tRNA fragments

Due to the complexity of tRNA fragmentation, read counts cannot provide the full picture of tRNA fragmentation. tRNA fragments can exist in sizes between 20 and 60 base pairs and be grouped using many methods. Rather than define a grouping, TRAX provides tools for visualizing tRNA fragment distribution. TRAX also includes visualizations and tools for performing quality analysis of experiments and getting statistics on entire sets of small RNA experiments.

To aid in visualizing the tRNA fragment distribution, TRAX includes methods for visualizing base-wise coverage of tRNA fragments and pre-tRNAs using the R ggplot2 package. These “coverage plots,” similar to those are used on the UCSC genome browser[42], are made by computing base-wise coverage for the set of aligned tRNAs and by using the tRNA alignments to combine these coverages for all tRNAs. To help the user visualize any multiple mapping, read coverage is colored by ambiguity in read mapping and shows the level of specificity, including whether the read transcript specific, anticodon specific, and acceptor specific coverages. These coverage plots can optionally be combined with data from the modomics[22] database, allowing for visualization of tRNA modifications known to cause stops in reverse transcription.[27]

In addition to these tools, TRAX generates several more visualizations for use with tRNAs and tRNA fragments. One of these is a chart showing read share among different gene types, which can show when an experiment or protocol causes a qualitative change in read profile. Another is a scatter plot showing normalized read counts for comparing selected sample pairs for the four different fragments typed. Finally, there is a summary plot that shows summed read coverage for all tRNAs in a sample that allows for a quick comparison of different experimental conditions or protocols.

While TRAX was designed with its own analysis and visualization tools, it is also designed to work with other RNA-Seq analysis tools that the user prefers. The results of the mapping are stored in the BAM format[43] and contain mappings to the genome and mature tRNA sequences. The table of read and fragment counts can also be used with any other tool for analyzing or visualizing read counts.

Using TRAX to find m1a modification in yeast and human tRNAs

RNA modification poses a problem for any sequencing of small RNA. These modifications prevent reverse transcriptase and result in tRNA sequencing experiments that could be missing a great number of tRNA fragments. To solve this, we used a demethylase known as AlkB to treat RNA before reverse transcription. This enzyme is known to demethylate RNA, particularly the methyl-1-adenosine (m1a) modification that is known to occur in tRNAs.[22,44] This protocol is known as ARM-seq.

We first used this to examine tRNA fragments and modification in human and S. cerevisiae. Comparing ARM-Seq to control sequencing reveals that arm-seq more than doubles the number of sequencing reads from modified RNAs. The pipeline showed that these RNAs are rarely mature tRNAs and are most often small RNA fragments derived from the three-prime fragment that contains the m1A modification at position 58. (Figure 1)

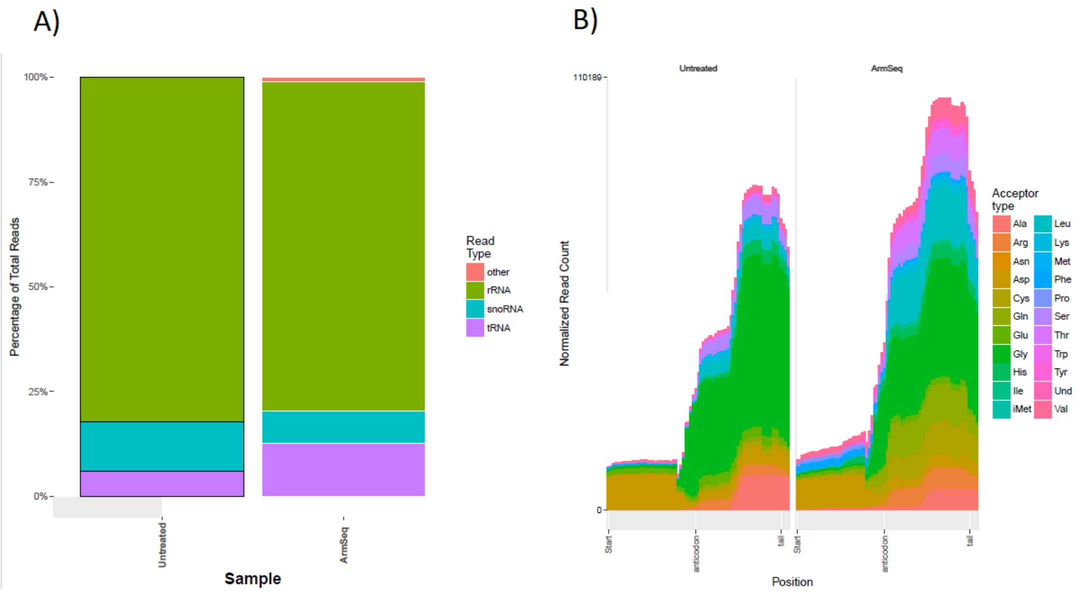


Figure 1: Results of ARM-seq on yeast small RNA sequencing. A) Plot of distribution of reads among gene types showing ARMseq increases the percentage of reads that are from tRNA based by more than two times. B) plot of read coverage for all tRNA gragment types showing ARMseq increases the number of three-prime tRNA reads.

TRAX correctly noted the differences in read abundance for 24 out of 26 modified yeast tRNAs, with the two remaining showing changes in read abundance that were not noted as significant. This includes finding two modified tRNAs that were not previously known, but were confirmed with primer extension. TRAX also correctly identified 18 out of 19 yeast tRNAs that did not contain modified tRNAs.

These results show that TRAX is able to perform tRNA sequencing analysis that can reveal changes in tRNA fragment abundance. These changes can then be corroborated directly showing the accuracy of the TRAX pipeline. This also shows how TRAX can be used to analyze entire RNA experiments quickly to find differences between experimental conditions.

Using TRAX to find modifications in pre-tRNAs

TRAX can separate mature tRNA transcripts from transcripts that have not been through processing with RNAse P and RNAse Z. Previous work had shown that pre-tRNAs can contain m1A modifications in Xenopus laevis, unlike most modifications that occur after RNAse P and RNAse Z cleavage.45,46 We used ARM-seq to determine if we could find modifications in pre-tRNA sequences.

While present in the ARM-seq results, pre-tRNAs are far less abundant than tRNA fragments, but TRAX did find at least one locus in most acceptor types with pre-tRNA reads, with 33 decoder types and 86 loci. Of these tRNAs, TRAX showed a difference between AlkB and control in 38 loci. This indicates that the m1a modification may be an early modification that happens before RNAse Z and RNAse P cleavage.[2]

TRAX comparison of sequencing kits in human small-RNA

Many tRNA sequencing analysis experiments have been published, but the question of how comparable they are remains. Previous work has been done to compare small RNA sequencing kits[47], but this has not been done for tRNA. With tRNAs being an unusual gene, there are reasons to believe that different protocols will achieve different small RNA sequencing results.

To test this, we compared a test of ARMseq[1] from two different commonly used preparation kits for sequencing small-RNA, the NEBNext kit and the TruSeq kit. Both of these kits were used to prepare small RNA libraries derived from MCF7 cells prepared using AlkB treatment and a buffer control.

The TRAX plot of read share among gene types shows a substantial increase in tRNA reads as a percentage of read total with AlkB treatment, with AlkB treated TruSeq having the largest percentage of tRNA reads. (Figure 2a) Read coverage across all tRNAs samples show more 3′ fragments than 5′, but this difference is decreased with AlkB treatment in both TruSeq and NEBNext kits. (Figure 2b)

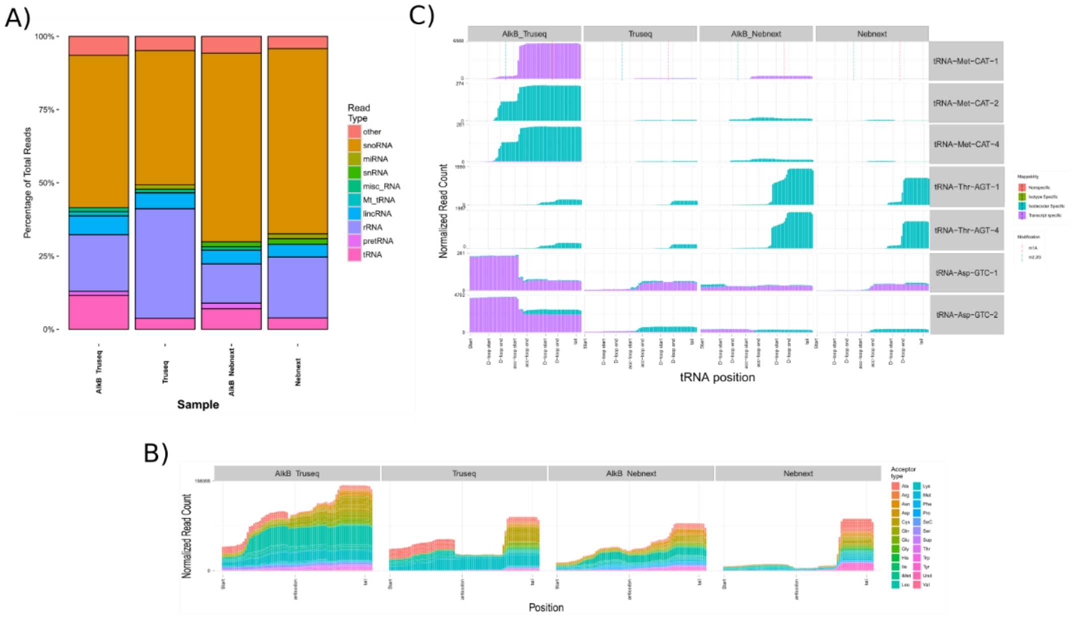


Figure 2: Comparison of TruSeq and NEBnext sequencing kits in sequencing of tRNA fragments. A) Plot of reads by gene type in Truseq and AlkB with and without AlkB treatment. AlkB increases the percentage of tRNA reads significantly for both kits, while maximum tRNA read percentage is achieved with Truseq and AlkB. B) Profile of tRNA fragment reads for all tRNAs colored b amino acid showing differences between protocols. AlkB allows for the sequencing of more RNAs, especially reads in Nebnext. C) Read coverage plots of selected tRNAs colored by mappability showing differences between the protocols. Some methionine and asparagine 5-prime fragments are only visible here with both Truseq and AlkB, and this occurs in multiple tRNA sets. Three-prime fragments of threonine are only visible in Nebnext and not all fragments require AlkB.

Both kits show an increase in both 5′ and 3′ fragments for many tRNA genes. Some of the strongest effects occur for 3′ fragments in Histidine tRNAs, and strong 5′

effects in proline tRNA. (Figure 2c) The use of ARM-seq on methionine tRNAs reveals a 3′ fragment that is enriched when both TruSeq and Armseq are used, and a similar 5′ fragment for Asparagine tRNAs. For threonine tRNAs, the effect of ARMseq is much smaller than the choice of library preparation, and the number of 3′ fragment reads is enriched substantially with NEBNext sequencing (Figure 2b)

All fragment types show a difference between the two library preparation methods. Valine and glutamine tRNAs show more 5′ fragments using the TruSeq kit, while methionine and glycine tRNAs show more 3′ fragments. Conversely, NEBNext sequences more threonine, tyrosine, and glutamic acid 3′ fragments and more cysteine 5′ fragments.

TRAX shows effects of PNK treatment in yeast small-RNA sequencing

We also wanted to use TRAX to measure the effect of PNK treatment on AlkB treated samples. PNK, or T4 polynucleotide kinase, is an enzyme that changes the end phosphorylation of RNAs, changing it to a 5′ monophosphate and a 3′ OH. This configuration is required for linker ligation in small-RNA sequencing protocols. While many mature small RNAs have this configuration, ribonucleases such as angiogenin and other processes such as RNA hydrolysis can leave alternate end configurations that should not be linkable.[34] PNK repairs these alternate transcript ends into suitable targets for the ligases used in small RNA sequencing kits. The effect of PNK treatment was studied in yeast samples sequenced using the ARMseq[1] protocol.

The TRAX pipeline shows that the primary effect of PNK treatment on tRNAs is on the levels of 5′ fragment (Figure 3). 13 5′ fragments show a greater than four-fold increase with PNK treatment, including those from Alanine AGC and Glutamine TTC tRNAs. Some tRNAs, such as the Phe tRNA, show a change in the fragment types under PNK treatment while Valine 5′ fragments show no measurable response to PNK treatment.

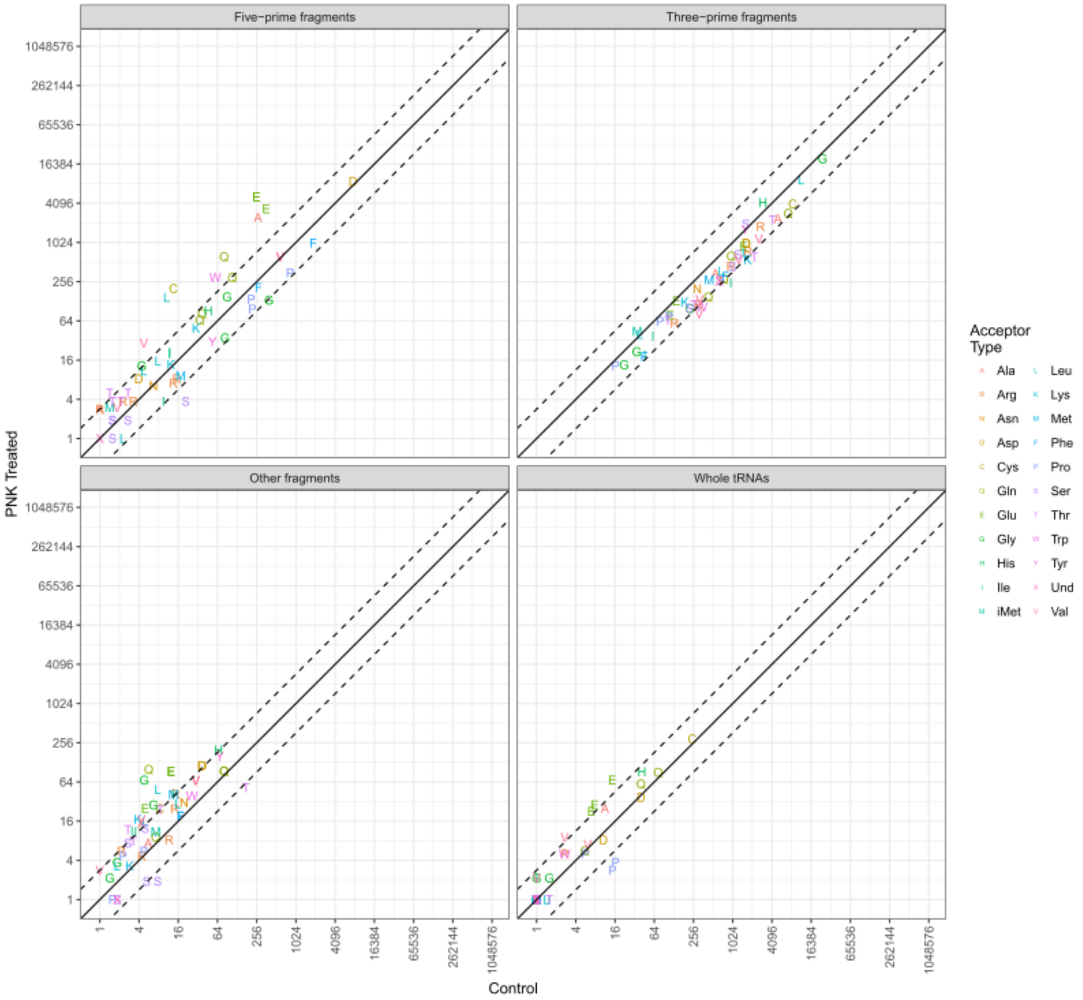


Figure 3: Read counts for different tRNA fragment types both with and without PNK treatment showing difference in five-prime tRNA fragments. Most five-prime fragments and all three-prime fragments show no difference with and without treatment, but some types such as glutamic acid and Alanine show differences with PNK treatment

While only some tRNA 5′ fragments respond to PNK treatment, this represents a set of tRNA fragments that thus far has remained largely missing from tRNA fragment analysis. While the reason for these tRNAs responding to PNK treatment is not clear, this suggests that PNK treatment is indicated for maximal sequencing of tRNA fragments. The effect of PNK treatment on human cell RNA remains uncertain.

Using TRAX to compare exosome and extracellular sequencing of tRNAs

Some recent work has been done on sequencing small RNA in exosomes and other types of extracellular RNA. [20,21] Many of these studies are focusing on microRNAs and ignore the tRNA population entirely. However, this means that a great deal of tRNA fragment data could exist unanalyzed in published datasets.

To show this pipeline can be used to identify tRNA fragments in extracellular material, we use TRAX to examine tRNAs from two published studies, one comparing cellular RNA to exosome RNA and one sequencing RNA blood serum.[48,49] Both extracellular samples have a greater percentage of tRNA than the sample from MCF7 cells (Figure 4). tRNA sequencing in exosomes and serum also have a distinct pattern of coverage, with unexpectedly high counts of 5′ fragments (Figure 4b). While these studies sequence different samples, both show a pattern of high numbers of valine, glycine, and glutamic acid 5′ reads with fewer reads of other tRNAs.

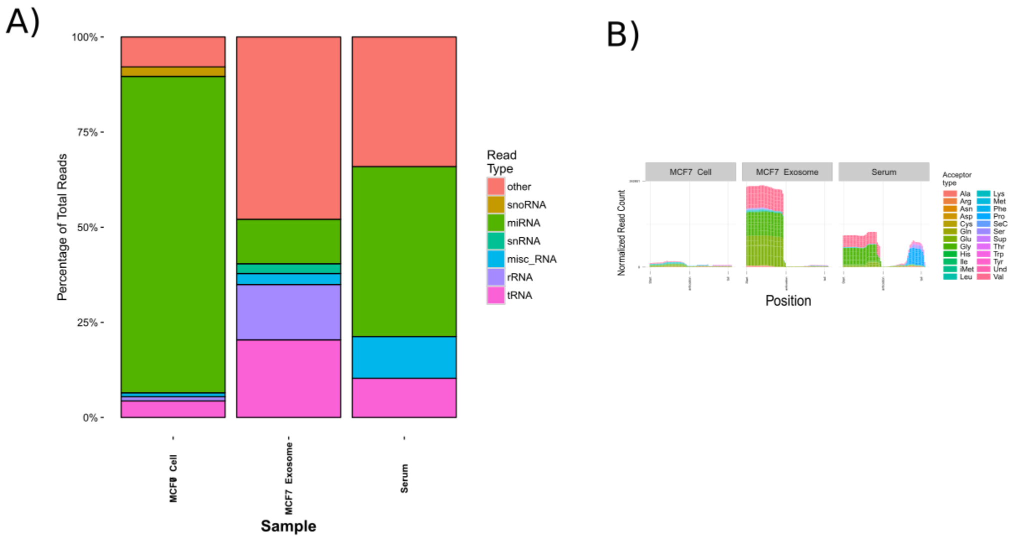


Figure 4: Comparisons of RNA-seq data in extracellular RNA vs cell RNA. A) Plot of reads by gene type showing both the matching extracellular RNA sample and one from a separate experiment show a larger percentage of tRNA than cellular MCF7 RNA. B) Read profiles of all tRNA fragments colored by amino acid showing similarties in extracellular samples which both have a large number of five-prime tRNA fragments of a few tRNA isotypes.

End independent sequencing

While RNA-Seq protocols based on micro-RNA protocols are capable of sequencing tRNA fragments, few library preparation methods are capable of sequencing whole mature tRNAs. What these methods have in common is the lack of a 5′ RNA linker ligation before reverse transcription. 5′ independent sequencing ensures that RNA with inaccessible 5′ ends or hard stops in reverse transcription can still be sequenced. One recent example of this type of sequencing is the TGIRT[50] protocol, which uses no RNA ligation but instead a strand-jumping reverse transcriptase that can read across gaps in the backbone to create cDNA.

To show the results of TRAX on sequencing generated using the TGIRT protocol, we used a dm-tRNA-seq dataset to analyze human tRNAs.[31] This study had four samples, composing of two conditions. One condition was the presence or absence of the demethylation step to remove RNA modifications. The second was a step to select specifically for tRNAs by requiring that the final base in the sequence be an A, matching the final base in a CCA tail. All four combinations with and without these steps were sequenced using dm-tRNAseq.[31]

With this 5′ independent protocol, greater than 75% of reads are tRNA-derived, which is consistent with tRNAs being the most common RNA transcript by molecule (Figure 5a). The coverage plots also show both full tRNAs and modification induced reverse transcription halts at the sites of several known modifications, including methyl-1-adenosine and methyl-2,2-guanosine. (Figure 4b) While the modifications make these reads incomplete, they still represent complete tRNAs and can be used to measure tRNA levels.

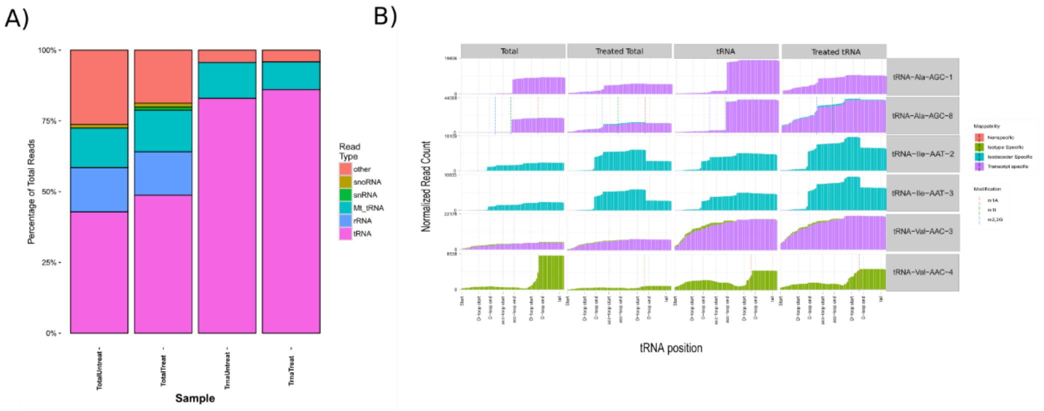


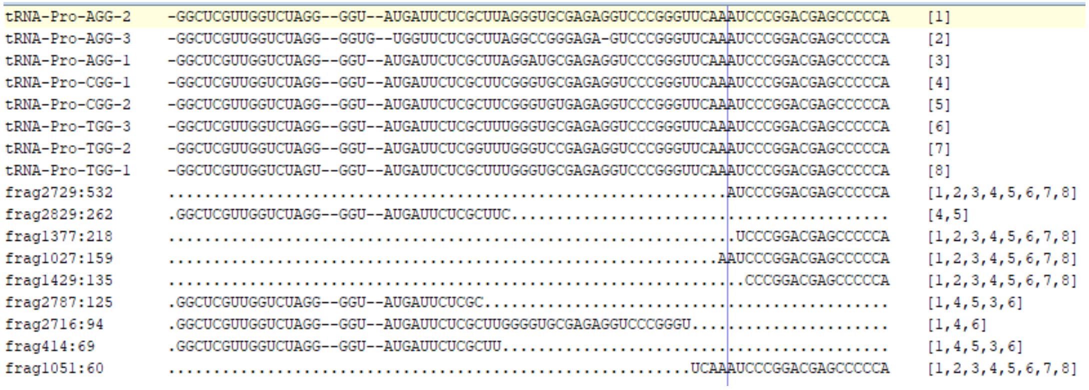
Figure 5: Analysis results for published DM-tRNA sequencing of whole tRNAs. A) Plot of reads by gene type reveals a much higher percentage of total reads are derived from tRNAs than in tRNA fragment sequencing. C) This shows

TRAX has the ability to show differences in read abundance, positions of possible modification-induced reverse transcriptase termination, and show differences between different RNA-seq preparations. Here, it shows the power of the demethylase treatment and the relative lack of power in the tRNA selection.

Using TRAX to analyze individual tRNA fragment sequences.

These visualizations can examine trends in tRNA fragmentation but they are unsuited to examining individual tRNA fragment sequences. Extracting the sequence or recovering the most abundant individual tRNA fragment is not possible using these charts for examining coverage. To solve this, I developed a method for analyzing tRNA fragments by sequence.

This method combines identical tRNA fragments and plots them as an alignment with the set of tRNAs that the sequences map to. This allows for easy searching most common fragment type among many fragments that could differ only by one or two base pairs. (Figure 6)



```
tRNA-Pro-AGG-2    -GGCUCGUUGGUCUAGG--GGU--AUGAUUCUCGCUUAGGGUGCGAGAGGUCCCGGGUUCAAAUCCCGGACGAGCCCCCA    [1]
tRNA-Pro-AGG-3    -GGCUCGUUGGUCUAGG--GGUG--UGGUUCUCGCUUAGGCCGGGAGA-GUCCCGGGUUCAAAUCCCGGACGAGCCCCCA    [2]
tRNA-Pro-AGG-1    -GGCUCGUUGGUCUAGG--GGU--AUGAUUCUCGCUUAGGAUGCGAGAGGUCCCGGGUUCAAAUCCCGGACGAGCCCCCA    [3]
tRNA-Pro-CGG-1    -GGCUCGUUGGUCUAGG--GGU--AUGAUUCUCGCUUCGGGUGCGAGAGGUCCCGGGUUCAAAUCCCGGACGAGCCCCCA    [4]
tRNA-Pro-CGG-2    -GGCUCGUUGGUCUAGG--GGU--AUGAUUCUCGCUUCGGGUGUGAGAGGUCCCGGGUUCAAAUCCCGGACGAGCCCCCA    [5]
tRNA-Pro-TGG-3    -GGCUCGUUGGUCUAGG--GGU--AUGAUUCUCGCUUUGGGUGCGAGAGGUCCCGGGUUCAAAUCCCGGACGAGCCCCCA    [6]
tRNA-Pro-TGG-2    -GGCUCGUUGGUCUAGG--GGU--AUGAUUCUCGGUUUGGGUCCGAGAGGUCCCGGGUUCAAAUCCCGGACGAGCCCCCA    [7]
tRNA-Pro-TGG-1    -GGCUCGUUGGUCUAGU--GGU--AUGAUUCUCGCUUUGGGUGCGAGAGGUCCCGGGUUCAAAUCCCGGACGAGCCCCCA    [8]
frag2729:532      ..............................................................AUCCCGGACGAGCCCCCA    [1,2,3,4,5,6,7,8]
frag2829:262      .GGCUCGUUGGUCUAGG--GGU--AUGAUUCUCGCUUC..........................................    [4,5]
frag1377:218      ...............................................................UCCCGGACGAGCCCCCA    [1,2,3,4,5,6,7,8]
frag1027:159      .............................................................AAUCCCGGACGAGCCCCCA    [1,2,3,4,5,6,7,8]
frag1429:135      ................................................................CCCGGACGAGCCCCCA    [1,2,3,4,5,6,7,8]
frag2787:125      .GGCUCGUUGGUCUAGG--GGU--AUGAUUCUCGC.............................................    [1,4,5,3,6]
frag2716:94       .GGCUCGUUGGUCUAGG--GGU--AUGAUUCUCGCUUGGGGUGCGAGAGGUCCCGGGU......................    [1,4,6]
frag414:69        .GGCUCGUUGGUCUAGG--GGU--AUGAUUCUCGCUU...........................................    [1,4,5,3,6]
frag1051:60       ..........................................................UCAAAUCCCGGACGAGCCCCCA    [1,2,3,4,5,6,7,8]
```

Figure 6: Plot of alignment of individual tRNA fragments in a set of proline tRNAs This shows only those fragments with greater than 50 reads. Each tRNA is given an individual fragment identifier “frag2729” and the number of reads with that sequence is reported afterwards. Each fragment is aligned to a set of tRNAs with the numbers in square brackets after the sequence indicating either the identifier for that tRNA or the set of tRNAs that the fragment can map to.

Methods

TRAX description

TRAX is a software pipeline written primarily in Python for the analysis of small-RNA sequencing data. As input, TRAX takes a set of tRNAs derived from the genomic tRNA database and a set of fastq files. Specialized tRNA mapping and quantification are done, and visualizations are created for the tRNAs and text versions of these files are outputted for any custom analysis the user wishes to perform.

tRNA analysis pipeline

TRAX prevents mismapping by creating a specialized tRNA database for use in mapping and analysis. This database is created by processing a set of known tRNAs derived from the genomic tRNA database[51] or from the output of tRNAscan-SE.[16] This database includes a bowtie2[52]genome index consisting of both the set of unique mature tRNAs and the genome sequence of the organism. These mature tRNA sequences are created by using the tRNA predictions to create a mature tRNA sequence by adding CCA tails, removal of introns, and addition of the histidine post-transcriptional 5′ G base. As an additional step, theses sequences are padded with 20 "N" bases to allow for extra bases off the end of the tRNA such as potential CCACCA ends. For analysis of basewise coverage, sequence alignments of both mature tRNA sequences and of genomic tRNA are created using the Infernal package[17]and covariance models from tRNAscan-SE.[16]

The read mapping of this pipeline is performed by bowtie2 in very-sensitive mode ignoring quality scores and allowing a maximum of 100 mappings. Mappings returned from this are post-processed to return all best mappings, with an exception for reads that map equally well to mature tRNAs and genome sequence, for which reads mapping to the genome are removed. Using this method, reads that contain genomic sequence flanking the tRNA loci in the genome are considered to be from pre-tRNAs, while reads that contain no flanking sequence will be considered as fragments of mature tRNA. Additionally, for reads mapping to tRNAs, the number of unique transcripts, anticodons, and acceptor types are counted and added to the mapping results as a set of custom SAM flags. Due to the presence of modifications and editing in tRNAs, TRAX does not by default use a mismatch cutoff/

TRAX separates reads that map to mature tRNAs into four fragment types. These are whole tRNAs, 5′ fragments, 3′ fragments, and a final "other" category. Fragment types are judged by the distance of the read mapping to the start and end of the tRNA. Reads where both the 5′ end and 3′ end lie within five bases of their respective ends on the mature tRNA are called as "whole tRNAs." Reads where only the 5′ end of the tRNA is close to the end of the transcript are called as "5′ fragments", and "3′ fragments" are similarly called as those close to the 3′ end of the tRNA. Fragments that do not meet any of these criteria are called "Other fragments." The 5′ and 3′ fragment types loosely correlate to the known TRF-5 and TRF-3 fragment types but the requirements are loosened to account for the diversity of fragment types seen in ARMseq and other sequencing experiments. These requirements are highly similar to that of a method that has been used in the past for fragment categorization. [53]

Reads that map to the location of tRNAs in the genome are from pre-tRNAs are similarly separated into fragment types. These reads are categorized into three types, "whole pre-tRNAs" that start before and end after the annotated tRNA gene, "partial pre-tRNAs" that overlap part of the tRNA gene, and "tRNA trailers" that start after the end of the tRNA gene, corresponding to tRF-1 fragments.[36] As a part of TRAX, counts for fragment types are combined with counts for non-tRNA genes and the full set of counts for fragments and genes is used as input to DESeq2.[41]

Reads in the fragment alignment is done identically to how read alignment is done with the exception that identical reads are combined to form a single read that records the number of copies. The BAM CIGAR string alignment is used to align the read with the tRNA alignments generated as part of the genome database.

Library preparation

ARM-seq was done using the protocols described in published work.[1] Cells were prepared using both the ARM-seq protocol and with the buffer control. tRNA from MCF7 cells were prepared using the ARMseq protocol[1] using both the TruSeq and NEBNext library preparation methods. Yeast was prepared using ARMseq and the NEBNext kit and treated with T4 polynucleotide kinase. All experiments were sequenced with Illumina Miseq.

Reads were preprocessed using cutadapt[54] to remove adapters from single-end reads and seqprep(John St. John, unpublished) to remove adapters and merge paired-end reads. In both cases, reads were rejected that were less than 15 bases long. Processed reads were mapped using bowtie2 to a sequencing database consisting of the full genome sequence of the organism, and the set of unique mature tRNA sequences from the gtRNAdb.[51]

The TRAX package consists of two major components, each of which is combined into a python script. The first of these uses data from the gtRNAdb[51]or tRNAscan-SE[16] to create a sequencing database for tRNAs in that species, including a bowtie2 database, a tRNA alignment, list of tRNA genome coordinates, and information for each unique tRNA transcript.

The second function uses a tRNA database to maps and count reads. After reads are counted, this pipeline runs DESeq2 analysis and creates plots of basewise coverage, pairwise sample comparison, and read distribution by gene type.

Discussion

Using TRAX to study tRNA fragments

By using a modified genome database containing the full sequence of tRNAs and an all-best mapping approach, TRAX can successfully map and quantify sequenced tRNAs and tRNA fragments. The combination of post-transcriptional modification and gene duplication was incompletely dealt with in previous tRNA sequencing pipelines. This method is designed to account for these shortcomings.[53]

The tRNA fragment determination of TRAX can allow for more fine-grained detection of changes than a simple counting of reads. Separating tRNA reads assists in finding effects where read counts of different fragments of the same tRNA change in opposite directions. While these fragments do not strictly correspond to the canonical tRNA fragment types, they more closely fit the tRNA fragments viewed in the experiments analyzed.

The use of tRNA alignments as a component of TRAX allows tRNA fragments to be compared with tRNAs including the tRNA structure and annotated tRNA modifications such as those in modomics.[22] This alignment uses the Sprinzel canonical tRNA positions,[5]making comparison of fragments with tRNA arms straightforward. The alignment of the base-wise coverage of all tRNAs also allows for fragments to be compared between tRNAs with different sequences for signs of consistent processing. This also allows for visualization more complex than read counting, as just measuring count may confuse a difference in level of tRNA fragment for a difference of fragment types. In the case of 5′ end independent protocols such as TGIRT, this allows for visualization of read dropoff due to modification.

The separation of reads into classifications based on specificity for visualization allows the user to mitigate the problem of tRNA similarity and multiple mapping. The read coverage visualization allows for viewing both the minimum and maximum level of possible tRNA fragments for each tRNA. With these features, it’s easy to measure and visualize differences in both fragment count and fragment type between experimental conditions.

In addition to the statistical and visualization tools provided, TRAX is intended to be compatible with other statistical and visualizations tools. TRAX stores read mapping as simple BAM files That can be read with IGV[55] and similar tools to examine individual read mappings. The fragment counts combined with counts for non-tRNA genes are similarly output as a tab-separated text file that can be used with other statistical tools for read count analysis. While built to handle tRNA fragments, TRAX can also perform read counting and statistical analysis for other small RNA types.

Choice of kit and protocol has large differences on tRNA sequencing result

Many tRNA sequencing have been done to examine the population of tRNA fragments in the cell. These results suggest that any of these studies will only sequence a portion of the population of tRNA fragments. Even with the use of demethylating enzymes, differences in sequencing kit cause noticeable differences in the results of tRNA sequencing. This does not even including comparisons of other components of tRNA sequencing protocols, such as size selection or RNA purification methods.

End chemistry is another part of tRNA sequencing that is often ignored in sequencing studies. Enzymes that cut tRNAs can leave different end chemistries depending on which end the phosphate back is left, or can even result in cyclic phosphates that must be resolved.[56] Angiogenin in particular is known to leave end chemistry that should not be sequenceable without the use of PNK or other treatments.[56] However, the ability of PNK treatment to make other breakdown products sequenceable as well does mean that there are costs to PNK treatment.

These complications of tRNA sequencing indicate that any sequencing result from any single experiment is unlikely to provide a complete picture of tRNA fragmentation within cells. Changes to tRNA fragment populations present in cells sequenced with the same protocol could be due instead to changes other than RNA population but may still provide useful data. Use of tRNA sequencing to describe the total population of tRNA fragments though is not possible based on these results.

Whole tRNA sequencing and tRNA fragment sequencing

The use of the TGIRT enzyme to develop a protocol for tRNA sequencing that allows whole tRNAs to be sequenced. However, the quantity of tRNAs relative to tRNA fragments and the ability of this protocol to sequence partially transcribed prevents the study of tRNA fragments and whole tRNAs in the same sequencing run.

Chapter 2: Analyzing epigenomics and comparative genomics of tRNAs to determine patterns of tRNA expression

Abstract

While some work has been done on regulation of tRNAs, most of the work done has been piecemeal focusing on one or a few cell types. To build a full atlas of tRNA gene expression, I used data from the ENCODE epigenomic roadmap to build a full atlas of tRNA expression in over 100 cell types. This data revealed both constitutive tRNAs and a set of tRNAs that are expressed only in cell lines and stem cells. I find that tRNA expression regulation is tied but not solely determined to tRNA sequence but is not strongly tied to expression of adjacent tRNAs. I compare the human epigenomic roadmap to a similar study performed in mouse and find that expressed tRNAs are more often conserved along with their expression regulatory program. For tRNAs that are only expressed in stem cells, their conservation and sequence quality is lower than constitutively expressed tRNAs indicating that these are degraded tRNAs rather than conserved specialized tRNAs. This suggests that tRNA regulation may be largely performed by a singly regulator that modulates pol III transcription.

Background

Epigenomics and Chromatin roadmap

The chromatin roadmap was a gathering of epigenomic data for set of 127 different cell types from tissues, cell lines, and embryonic cells done as part of the ENCODE project.[57] . This roadmap included multiple types of experiments including chromatin immunoprecipitation of histone marks. This data was analyzed with ChromHMM[58], a tool for analyzing Chip-seq data using a hidden markov model, to determine chromatin states for the entire genome.

This analysis categorized every base in the human genome into one of fifteen chromatin states based on histone methylation. These states include among others transcription initiation, enhancers, transcription elongation, and quiescent DNA. These chromatin states are annotated in fixed 50 base intervals of the genome which maximizes their resolution at that level.[57]

Orthology of tRNAs in eukaryotic genomes

Duplication of tRNAs makes determining tRNA orthology more difficult than annotating orthologs in protein-coding genes. Protein-coding gene orthology is commonly determined by finding pairs of genes that most closely resemble one another with reverse-best blast.[59] The duplication of tRNAs makes this approach fail, tRNAs often have many identical copies that make finding a single best match impossible.[6]

This duplication makes other approaches to finding tRNA orthologs necessary. The primary method that has been used for finding orthologous tRNAs is to use adjacent sequence and particularly adjacent protein-coding genes to generate syntenic blocks that can then be compared, this has been done with both Drosophila[15] and one for Eukarya.[60]

Epigenomic effects of tRNA genes

tRNAs are known to influence the epigenomics of the genome. Specifically, tRNAs are known to act as chromatin insulators that block expansion of heterochromatin from proceeding past them in human and yeast.[61,62] tRNA are also known affect transcription of adjacent protein coding genes[63], and are responsible for packaging of chromatin loops due to their tendency to locate near the nucleolus.[64]

DNA methylation and CpG islands

In many eukaryotic genomes including those of mammals, much of the genome is deactivated through DNA methylation.[65] Specifically, cytosine bases in DNA are methylated into 5-methyl-cytosine. These modifications occur in cytosine bases located before guanosine bases, known as CpG positions. This methylation is used in cell differentiation to de-activate regions of the genome that differentiated cell types no longer used.[66] The mechanisms that target region of DNA for methylation are not currently well understood, but one that is known is that not being transcribed can target a region of the genome for methylation.[67]

Vertebrate genomes often contain ~1000 base pair regions of highly concentrated CpG base pairs known as CpG islands.[68] While containing many potential methylation targets, these regions are protected from DNA methylation. These regions are very common in the promoter of most protein-coding genes, including housekeeping genes.

Results

Epigenomic data can be used to determine activation state for genomic tRNAs

While previous studies have measured tRNA expression,[69] we wanted to take a broader look at tRNA expression in multiple cell types. This can allow us to find broad patterns of tRNA regulation that may not be present in smaller studies that focus on a small number of cell types, and to look for master tRNA regulators.

To do this, we used the published data from chromatin roadmap project[57] to search the set of annotated human tRNA genes to determine the transcription state of all tRNAs in the human genome. Each tRNA was classified into one category of chromatin state in the chromatin roadmap depending on which the most active state overlapping it or the flanking region. (Figure 7)

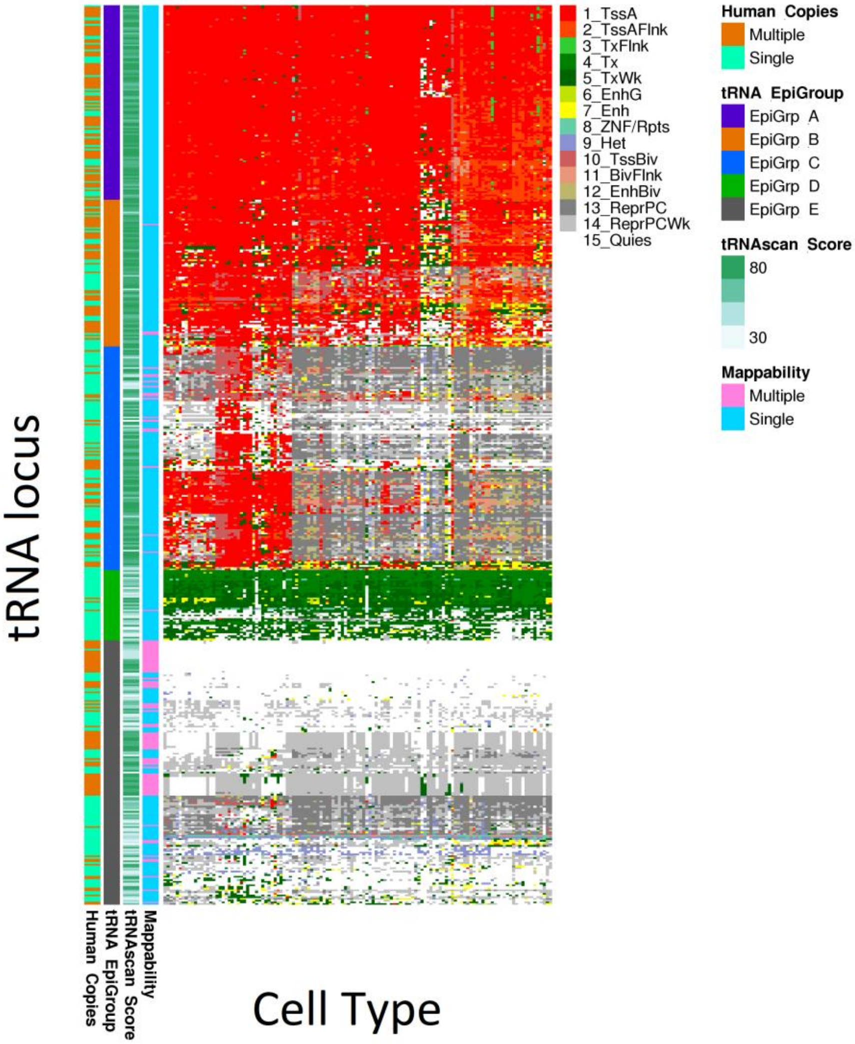


Figure 7: Heatmap of chromatin states of tRNAs in the human genome showing differences in tRNA expression. Colors represent the state of the tRNA in the chromatin roadmap. Ordered by hierarchical clusterint. Annotations include the mappability of the region, the tRNA score, the epigenomic classification of the tRNA into epigroups A through E, the number of copies of tRNAs of that sequence in the human genome, and the number of copies of that sequence in the mouse genome. tRNAs in a transcription activation state, colored here in read, are transcriptionally active. Many annotated tRNAs here are never transcribed, while some are trancribed only in a subset of cell types

We simplified the 15 chromatin states that the epigenomic roadmap classified genes into, a smaller group of states. (**Error! Reference source not found.**) Transcription start states and bivalent promoter states are believed to overlap tRNAs that are transcriptionally active. Given the 50-base window that ChromHMM classifies the genome into, tRNA genes are too short to have separate initiation and extension states. In this data, these states are associated with the presence of the H3k4me3 histone mark.[57] tRNAs that are located in the transcript of a longer gene, often the intron of a protein coding gene, are in one of the Transcription extension states. Other states are believed to correspond to inactive tRNA.[57]

These results show that there are more than just "constitutive" and "inactive" tRNAs, and specifically there seems to be a set of tRNAs that is only expressed in a subset of cells. This set of cells includes stem cell types and embryonic cell types, suggesting that this could be a subset of tRNAs that are activated in times of high tRNA demand.

Classifying tRNAs into epigenomic groups

To study these epigenomic results and prepare to search for possible regulatory methods we classified tRNAs into five epigenomic groups based on broad patterns in their chromatin regulation. These groups describe the broad patterns visible in tRNA epigenomic data based on their classification of chromatin states into transcription initiation, transcription extension, and inactive chromatin states in each of the cell types.

The first group, called "Epigroup A," is defined as RNAs that are almost always expressed. This group can be thought of as the constitutively expressed tRNAs. This set consists of 115 tRNAs, or 23% of tRNA for which epigroup can be determined.

The second group, "Epigroup B," is defined as tRNAs that are active in most cells. While not constitutively expressed, this group of tRNAs are expressed in more than half of cell types and can be thought of as the tRNAs that are generally active. This set consists of 86, tRNAs and consist of 17% of the total.

The third group, “Epigroup C,” is defined as those tRNAs that are active in only a small set of cells. The set of cells where these tRNAs are transcribed is primarily stem cells and cell lines, which is also when demand for tRNAs should be highest. This set consists of 131, or 26% of the total number of tRNAs.

The fourth group, “Epigroup D,” is defined as those tRNAs that are in the transcriptions extension state and part of a long transcript. These are tRNAs are located in the introns of protein-coding genes. These tRNAs may be either cut out of introns, transcribed independently in some cases, or just inactive pseudo-tRNAs. This set consists of 35 tRNAs, or 7% of the total.

The fifth group, “Epigroup E,” are those tRNAs that are active in none or extremely few cell types. This group consists of annotated tRNAs that are either never transcribed and tRNAs that are only transcribed in very specific circumstances. As such, this group is where tRNA pseudogenes will be classified in this group. This set consists of 141, or 28% of tRNAs.

Not all tRNAs can be places into one of these sets, as some tRNAs are in segmental duplications or other regions that make determine transcription state impossible due to the risk of multiple mapping. These tRNA are removed from future analysis steps as impossible to categorize.

These epigenomic groups can form the basis of analysis of tRNA regulation. Separating tRNAs by epigenomic group can allow us to look for common elements within groups and differences between them to reveal methods of tRNA regulation.

Polymerase III reads match chromatin states

We sought to confirm that this epigenomic roadmap data matched other data that has in the past been used to measure tRNA data. Specifically, we matched this data to RNA polymerase III CHIP-seq data[69] that was performed for liver cells. This study was taking also examining tRNA expression between cell types and species, which makes this a suitable experiment to compare chromatin roadmap data.

After these reads were re-mapped and reads mapping to tRNAs regions counted, we find that 89% of tRNAs where both the polymerase III CHIP reads and epigenomic roadmap chromatin state match, either both indicating an active tRNA or both indicating an inactive tRNA. Even stronger is the tendency for tRNAs with polymerase III reads to be in one of the active epigenomic groups, only two tRNAs with RNA polymerase III reads are not in epigroup A, B, or C. (Table 1)

	tRNA loci with RNA Polymerase III binding	tRNA loci without RNA Polymerase III binding
tRNAs with active chromatin	124	18
tRNAs with inactive chromatin	39	333

Table 1: Matrix of tRNA counts that are in active or inactive chromatin states and tRNAs that have pol 3 transcription. 457 tRNAs match polymerase III states while 57 tRNAs do not match in chromatin state.

While these RNA polymerase reads do not perfectly recapitulate the result, this agreement between these two forms indicate that chromatin roadmap data is a good proxy for tRNA expression. That the mismatches between the two data sets are often among those tRNAs that are expressed in some cell types suggests that this issue is primarily an issue of different sensitivity of the two methods. The few remaining differences could be result of differences between the two sets liver cells sequenced.

tRNA activation is largely dependent on tRNA sequence

One possible source of differences in tRNA expression is differences in tRNA sequences. Expression of tRNAs is known to be based on the presence of A and B boxes, which is one possible source of tRNA expression. But rather than just focus on the A and B boxes, we wanted to measure the entire tRNA gene to ensure that no other signals within the tRNA were being missed, such as a possible transcription shutdown due to malformed product.

The scores of the tRNAscan-SE covariance model were used as to measure the sequence and structure of the tRNA. This program scores the sequence and secondary structure of the tRNA sequence compared to the general model of tRNAs. High-scoring tRNAs are those that closely match the canonical sequence and structure of tRNAs, while low-scoring tRNAs are those that do not match.

We find that epigenomic category does correspond to the quality of the tRNA sequence as measured by tRNAscan-SE. (Figure 8) We find that constitutively active tRNAs are those tRNAs with a score of greater than 55, always in the case of epigroup A and with a few exceptions in epigroups B and C. However, we find that some of the epigroup E inactive tRNAs have tRNA scores of greater than 55, indicating that there is some other mechanism used to shut down tRNA.

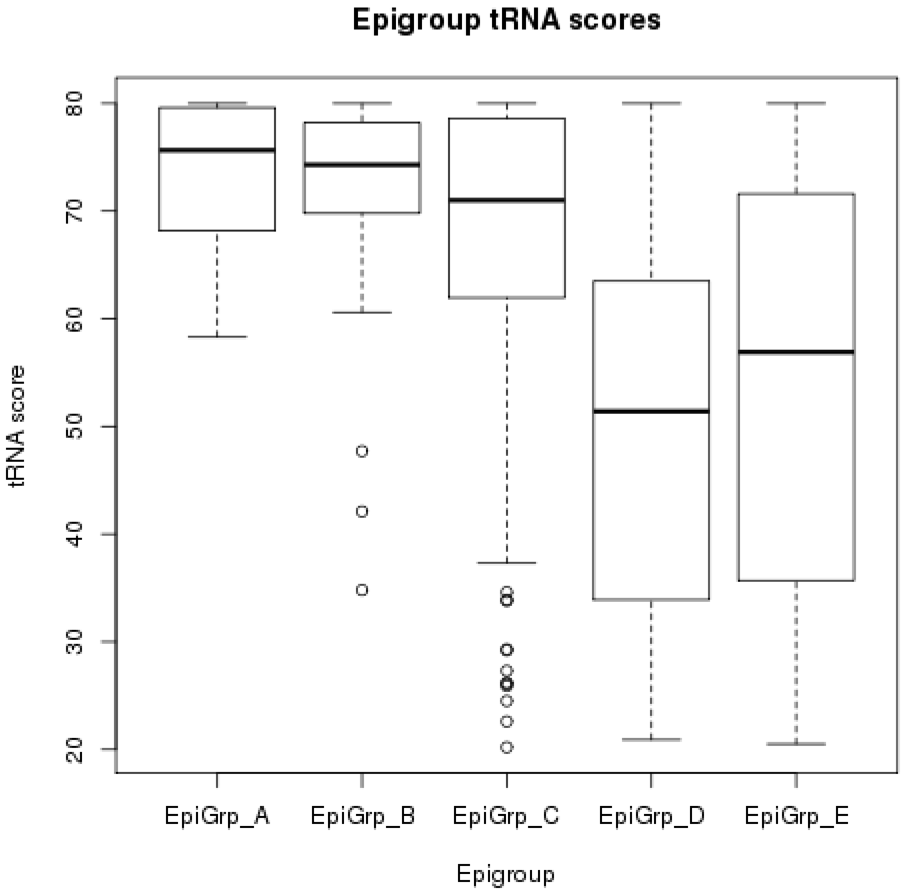


Figure 8: tRNAscan-SE score of tRNAs of in epigenomic groups. . tRNAs that are in the highly transcribed epigroups A and B contain few or no high-scoring tRNAs. tRNAs in the inactive epigroup E contain both high and low-scoring tRNAs, indicating that sequence alone is not sufficient for a tRNA to be transcribed.

This indicates that tRNA score, while playing a role in determining transcription of tRNA, cannot be the sole determinant of tRNA transcription. The presence of low-scoring tRNAs in epigroups B and C and especially the presence of high-scoring tRNAs in epigroup E indicate that there is some other mechanism that is helping to control tRNA expression.

All isotypes and isodecoders have constitutively expressed tRNAs

The epigenomic categories presented here raise questions about tRNA redundancy. The presence of multiple isodecoders per tRNA could provide some redundancy if those tRNA are active, but that redundancy could be illusory if those duplicate tRNAs are not active. We sought to determine how many tRNAs are active in each decoder group and whether certain decoder groups could be inactive in normal conditions or especially augmented in stem cells.

The isodecoder with the most epigroup A loci is the initiator methionine with six isodecoders, while those with slightly fewer epigroup A loci are the asparagine GTC, alanine AGC, and tyrosine GTA decoders. Only redundant decoder types have zero epigroup A tRNA loci. The number of epigroup C stem cell tRNAs can vary greatly, with some decoder types having none while others, such as initiator methionine and lysine TTT having eight epigroup C tRNAs. (**Error! Reference source not found.**) These numbers could be misleading though, as tRNA genes can be present in deleted or duplicated regions that a single genome assembly cannot represent entirely.

Expressed tRNA exist in multiple exact copies.

tRNAs often exist in multiple exact copies in the genome, for which it is not totally clear the source or purpose of the duplication.[60] We sought to determine whether these tRNAs were inactive copies or if they were signs of active transcription by measuring activation of single-copy RNAs or multi-copy tRNAs. This should indicate whether unique tRNAs are specialized genes to perform specific tasks, or tRNAs that are degrading.

We found that active tRNAs are more likely be duplicated, and less active tRNAs tend to exist in single copies. This is the case both for totally inactive tRNAs, and for tRNAs that are in the less active epigroup C, which tend to not be as duplicated as epigroup A tRNAs. (Table 2)

Epigenomic group	Multi copy tRNAs	Single copy tRNAs	Percent of multi copy tRNAs
A	75	40	65.2%
B	48	38	55.8%
C	43	88	32.8%
D	3	32	8.6%
E	54	87	38.3%

Table 2: Counts of multicopy vs singlecopy tRNAs in different epigenomic groups shows that more active tRNAs exist in more copies than single copy.

That epigroup A tRNAs are duplicated while epigroup C and inactive tRNAs are not suggest both that unique sequence is a signal of degradation and that epigroup C tRNAs could be degraded tRNAs rather than specialized genes that are preserved.

Clusters of tRNA are only partially co-regulated

Some proposed methods for regulating tRNAs, such as chromatin loop formation[70], point to location in the genome as a part of tRNA transcription. To determine if this is the case, we examined sets of tRNAs that are adjacent in the genome. Nearly ¾ of human annotated tRNAs are in one of these multi-tRNA clusters. Coregulation of these clusters would indicate that tRNAs are regulated by their region in the genome, while differences in regulation indicate that tRNAs are regulated individually. This would also answer the question of whether clustering is required for tRNA expression or prohibits it.

These results show that singleton tRNAs are approximately equally likely to be expressed and silent. Most active tRNA do exist in multigene clusters though, and very few tRNA clusters are entirely silent. Most multi-gene tRNA clusters are very short, consisting of two or three tRNAs, while a few large clusters with greater than

30 tRNAs exist on chromosome six. (Figure 9)

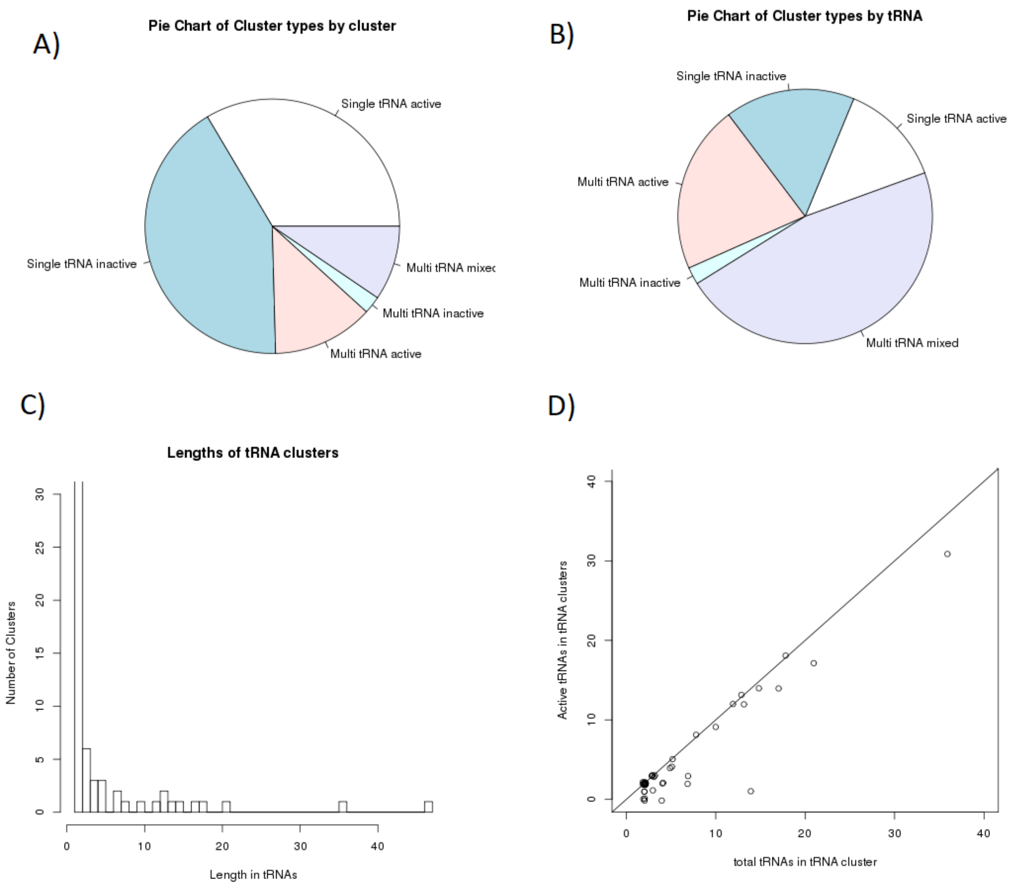


Figure 9: A) Pie chart showing distribution of tRNA clusters. Most tRNA clusters consist of just a single tRNA. B) Pie chart showing distribution of tRNAs within tRNA clusters. Most tRNAs exist in multi-tRNA clusters. C) Histogram of tRNA cluster length showing many tRNAs exist in very short clusters of two tRNAs while a few longer clusters exist. D) Plot of tRNA cluster length vs active tRNAs in cluster showing that larger clusters generally consist of active tRNAs with a few inactive tRNAs mixed in. Smaller clusters

Multigene tRNA clusters can contain both active and inactive tRNAs, including inactive tRNAs with high quality scores that would require some mechanism to suppress. The largest tRNA clusters can contain constitutively expressed tRNAs, inactive tRNAs, and tRNAs that are only activated in stem cells. (Figure 10)

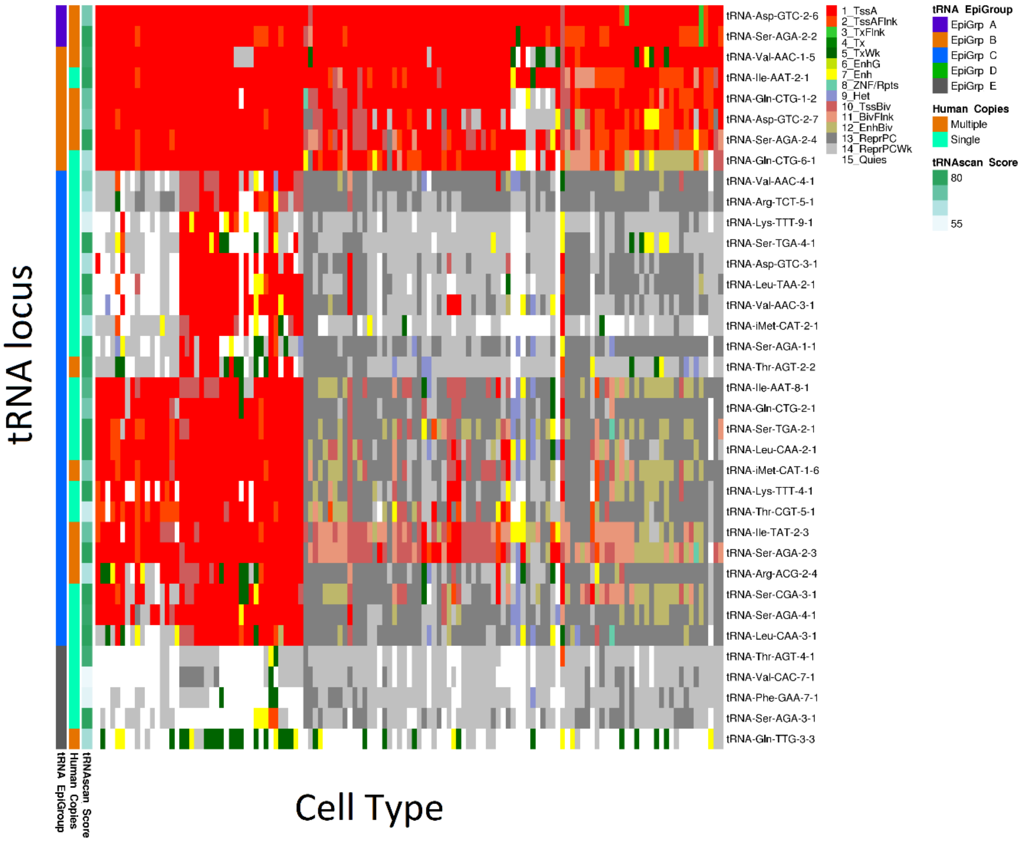


Figure 10: Chromatin state of all tRNAs that exist in the tRNA cluster at chr6:28442328-28956860 of hg19 show heterogeneity of expression within tRNA clusters. This cluster includes both constitutively expressed tRNAs and tRNAs that do not appear to be expressed in any cell types, as well as different types of partially active cell types. tRNAs are sorted by epigenomic group

Of these clusters, 70 contain at least one transcribed tRNA, and 22 of these tRNA clusters include both tRNAs that are transcribed in some tissues and tRNAs that are never transcribed. When filtered for tRNAs with a score of greater than 60 bits, then 9 of the 65 clusters include both transcribed and non-transcribed tRNAs. This suggests that tRNA deactivation is not solely due to large-scale chromatin patterns.

Analyzing clusters by epigenomics groups does show that these clusters often contain mixtures of group A, B, and C tRNAs. Almost half of all tRNAs are in the 32 tRNA clusters that contain multiple categories of transcribed tRNA.

This indicates that genomic regions is not enough to determine tRNA transcription and therefore that adjacent protein-coding genes or chromatin loops cannot explain the epigenomic differences present between tRNAs. Some other mechanism here must be playing a role in chromatin regulation of tRNA expression.

DNA methylation is a signal of tRNA suppression

+

While tRNA quality and A and B boxes may be enough to explain some of the tRNA expression pattern, there are a number of high-scoring tRNAs that are never expressed. To determine why, we looked at the bisulfite data in the epigenomic roadmap to determine if DNA methylation is a signal for deactivation of the genome. (Figure 11)

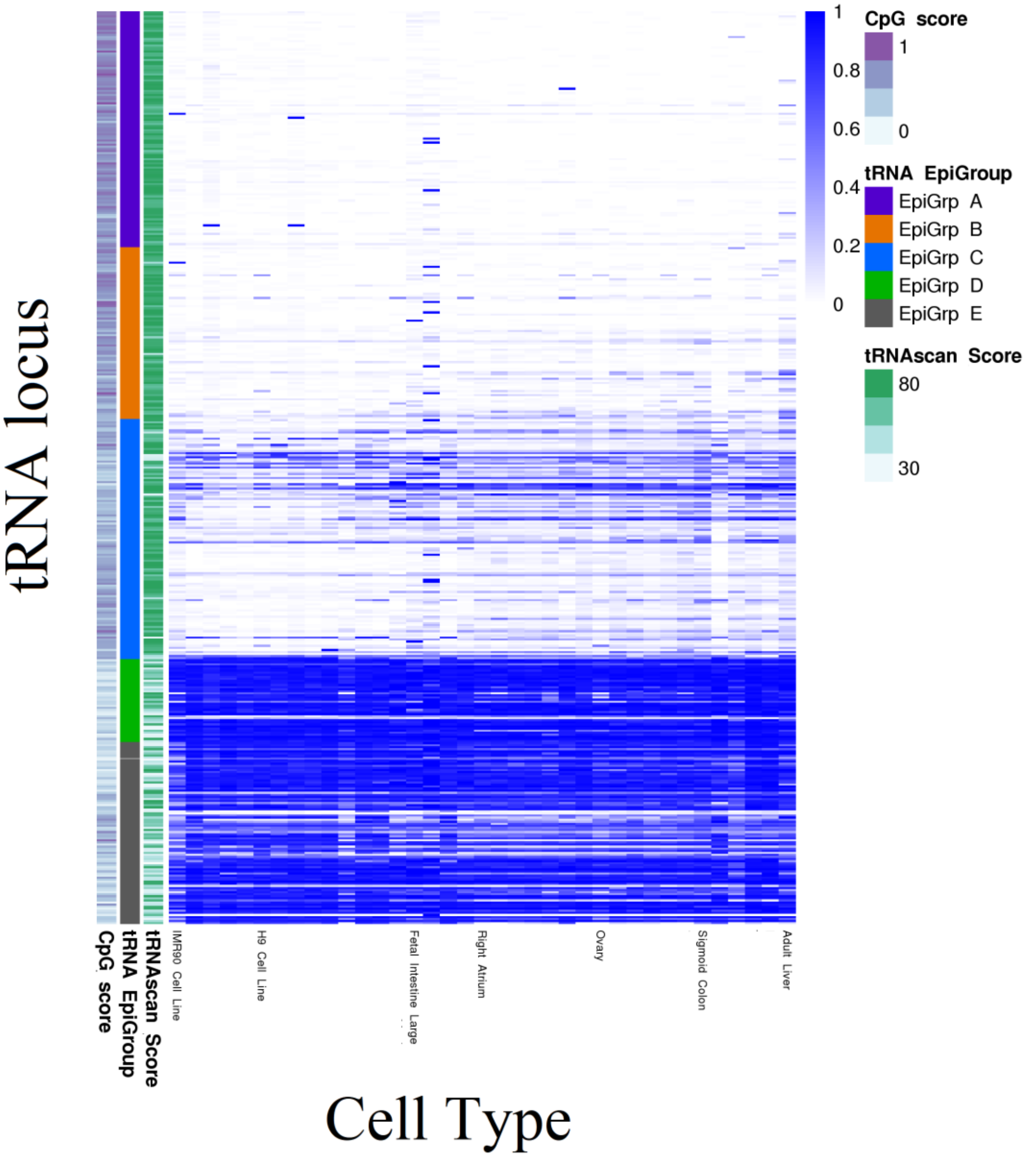


Figure 11: Heatmap of DNA average cytosine methylation results of cells in the chromatin roadmap determined using bisulfite sequencing showing methylation of inactive tRNA loci. Scale is average of percent methylated of all C bases in the tRNA and 100 flanking bases. Active tRNAs in groups A and B tend to be unmethylated, while inactive tRNAs are highly methylated. tRNAs in group C are lightly methylated in some cell types.

We found that tRNAs that are either inactive or intronic are strongly methylated, while other tRNAs are weakly methylated or unmethylated. tRNAs in epigroup B show slightly more methylation than epigroup A, and epigroup C shows more still. Even in cell types for which a tRNA is deactivated though, the level of methylation does not approach that found in tRNAs that are never transcribed.

It is not clear if methylation is what blocks tRNA transcription of high-scoring tRNAs, but transcription of tRNAs has been shown to be blocked by DNA methylation in Xenopus[71]. While it is not known what prevents methylation of transcribed tRNAs, active tRNAs tend to be located in genomic regions that are enriched for CpG, suggesting that CpG islands may play a role in activating tRNA loci. (Figure 12)

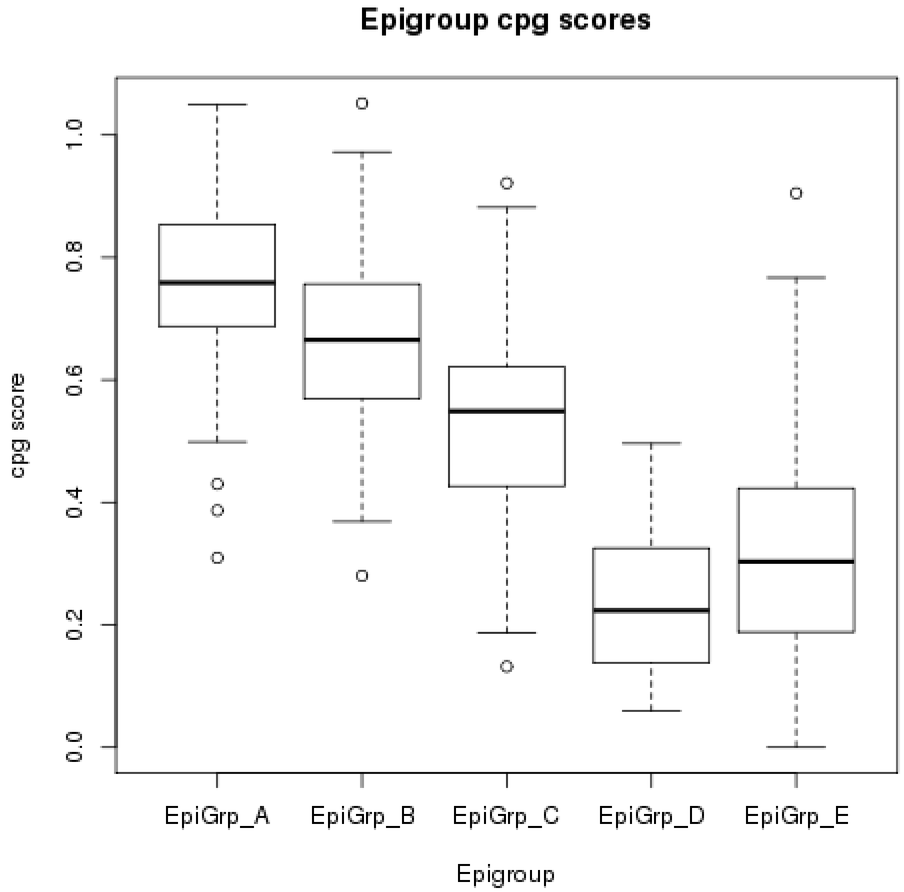


Figure 12 Box plot of CpG scores from regions flanking tRNA genes showing active tRNAs have higher CpG enrichment. tRNAs in more active epigroups tend to have higher CpG scores suggesting the presence of CpG Islands enriched around actively transcribed tRNAs.

Expressed tRNAs show conservation across mammals

tRNAs are highly mobile and redundant, and many tRNAs are not conserved in even closely related species. We sought to determine the degree to which tRNA conservation was related to tRNA activation to determine whether active tRNAs we conserved or the constant duplication and re-arrangement of tRNAs makes conservation of individual tRNAs irrelevant.

In mouse, only 39% of human tRNAs have a conserved ortholog, compared with 72.8% of protein-coding genes. Epigroup A tRNAs are more likely to be conserved, with 51% of active tRNAs having an ortholog in mouse. For the more closely related orangutan this trend is weaker but still present, with 59.27% of active tRNAs have a conserved tRNA ortholog compared to 56.77% total.

This indicates that conservation is a feature of active and especially constitutive tRNAs and suggests that despite the duplication of tRNAs. This may be partially due to the conservation of location of tRNA position in the genome, or conservation of position may just be caused by lack of tRNA elimination or need to retain a small set of flanking sequence.

tRNAs in mouse show a similar pattern of chromatin activation

While conservation of active genes provides some information, we wanted to study conservation of tRNA expression itself. To do this, we used a mouse epigenomic roadmap that was constructed using histone mark pulldowns and chromhmm.[72]

The mouse epigenomic roadmap has fewer cell types but similar profiles of tRNA expression. To compare these results, using an analogous classification system to that used for human. Using this system, the mouse epigenome has 180 inactive tRNAs, 176 constitutively active tRNAs, 41 differentially active, and 55 active in embryonic stem cells. Mouse epigenomics shows a similar pattern tRNA activation, with active tRNAs having high tRNA scores and conservation. (Figure 13)

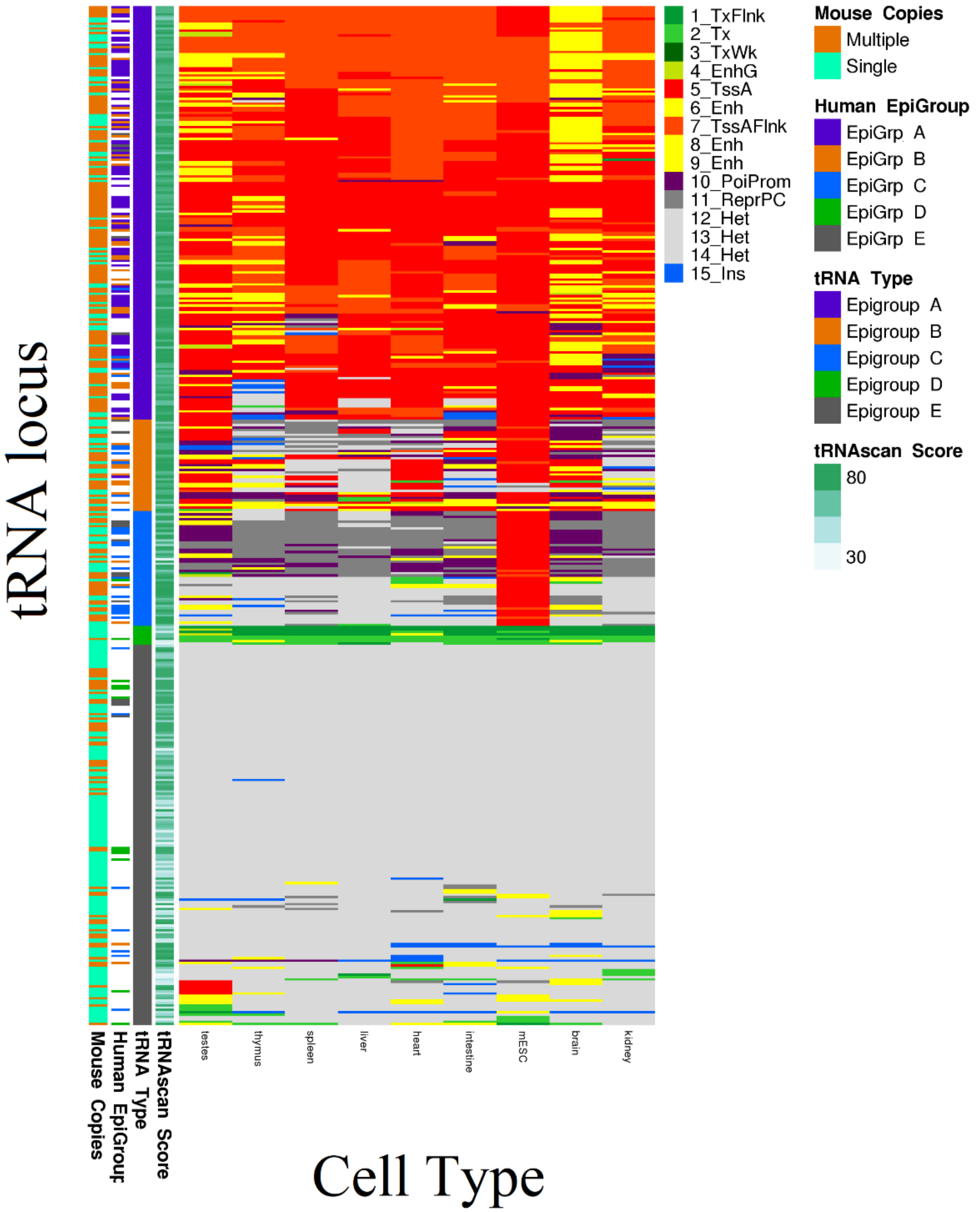
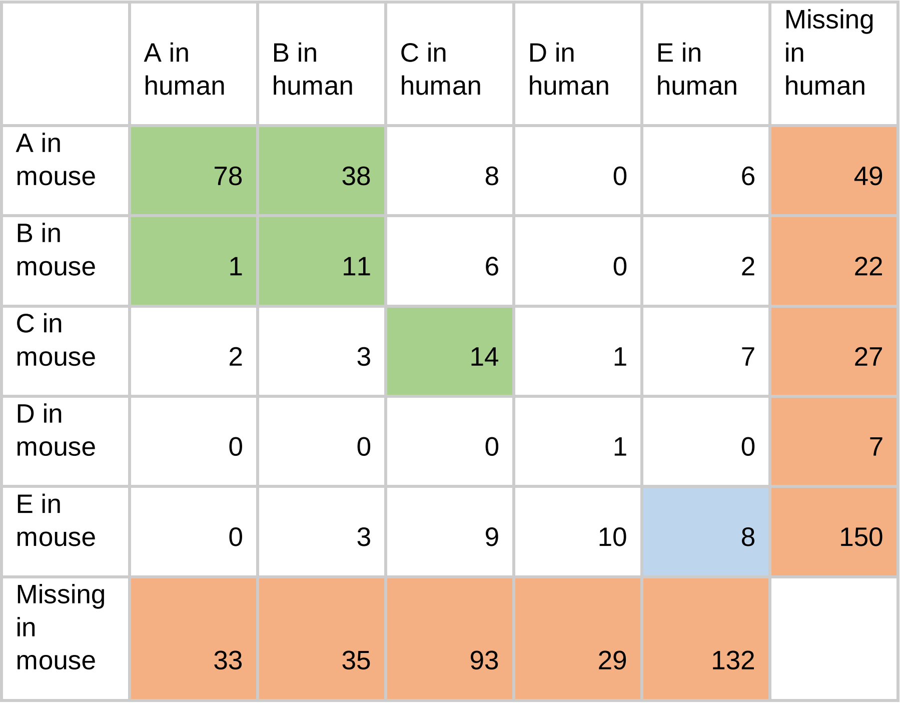


Figure 13: Heatmap of mouse tRNA chromatin state showing similar patterns to those in human chromatin. "Human Epigroup" shows the epigenomic category of the ortholog of that tRNA in the human genome. Mouse tRNAs also tend share chromatin status to their human orthologs for highly expressed tRNAs, while tRNA that is less transcribed or not transcribed tend to be not conserved

In addition, the chromatin regulation of individual mouse tRNAs is similar to their human counterparts. 90% of active mouse tRNAs that have an ortholog in human are also active in mouse, while 61% of the inactive tRNAs that are conserved in mouse are also inactive in human. (Table 3)



	A in human	B in human	C in human	D in human	E in human	Missing in human
A in mouse	78	38	8	0	6	49
B in mouse	1	11	6	0	2	22
C in mouse	2	3	14	1	7	27
D in mouse	0	0	0	1	0	7
E in mouse	0	3	9	10	8	150
Missing in mouse	33	35	93	29	132	

Table 3: Epigroups of human tRNAs compared to epigroups of their mouse ortholog. tRNAs that are active in human, in epigroup A or B, tend to be in epigroup A or B in the mouse genome. Inactive tRNAs in epigroup E tend to not be conserved at all in the orthologous genome.

The conservation of expression both in general and in individual tRNAs suggest that despite duplication and therefore redundancy of tRNAs that could make this form of conservation unnecessary, tRNAs are conserved just like other forms of gene and expression information from one genome will be informative to tRNAs in another genome if that tRNA is conserved. That epigroup C tRNAs are the least well conserved raises questions as to their role in the cell and how they are created.

Methods

Determining epigenomic states of tRNAs in cell types

Chromatin roadmap data was taken for hg19 for all cell types in the 127-sample set. From this, states were assigned to the tRNAs taken from the genomic tRNA database[51] based on the most active state within the window 250 bases upstream and 100 bases downstream of the tRNA start. This was used to generate a chromatin state for all 622 genes in all 127 samples.

tRNAs that were deemed to be unmappable using CHIP-seq were specially marked. This was done by mapping the tRNA and 100 bases of flanking sequence to the genome and mapping to the genome using bowtie2. Sequences that mapped with a phred score of greater than -10 were marked as multiply mapped tRNAs and removed from epigroup analysis.

The fifteen chromatin states were simplified into five categories, those of transcription start, bivalent states, enhancers, transcription extension, and inactive states. For heatmaps, these were hierarchically clustered using the pheatmap R package.

Categorizing tRNAs into epigenomic groups

tRNAs were classified into epigenomic groups according to a set of rules. Samples were separated according to whether they came from primary tissues or other cell types. tRNAs that were active in more than 90% of all samples and more than 85% of primary tissues were classified into group A. tRNAs active in 27% to 90% of all samples and 20% to 90% of primary tissues. Group C tRNAs were those that were active in 3 to 43% of all samples and 0 to 20% of all tissues. Group D tRNAs are those that are active in less than 1% of samples and are in the transcription extension state in 36% to 99% of samples. Group E tRNAs are those that do not fit in any of the above categories, those that are active in less than 3% of total tissues and in a transcription extension state in less than 30% of tRNAs

Classification of mouse epigenomic states

Epigenomic data for mouse was taken from a previous study of chromatin states on mouse[73]. Data for mm9 was transcluded to mm10 using liftover.[74] Chromatin states were then assigned to the mouse tRNAs using the same method as used for human with 250 base upstream and 100 bases downstream.

The epigenomic groups for mouse were assigned according to the following set of rules to achieve a set of states as similar as possible conceptually to those used for human, with the exception that enhancer epigenomic categories were considered to be "active" transcribed types. tRNA loci active in more than 5 samples are in group A, loci active in more than one are group B, tRNAs in neither of those categories but active in the embryonic stem cell sample were group C, tRNAs in states 1, 2, and 3 in more than 5 samples are group D, and tRNAs in none of those categories are group E.

Verification with RNA polymerase III chip-seq

Pol 3 results were taken from a previous paper[69] that performed pol III CHIP-seq in liver of human and mouse. Reads from this paper were mapped using bowtie2 using default parameters, and read counts were taken from a 100-base window around the tRNA. These were compared to epigenomic sample E066 in the human chromatin roadmap. tRNAs with more than 100 reads in that region were considered to be Pol III positive, while tRNAs with less than 100 reads were considered to be Pol III negative.

Predicting orthologous tRNAs

To predict orthologous tRNAs, a method was used based on nearby adjacent orthologous protein-coding genes. Sets of orthologous genes between a pair of species were predicted using reverse-best blast. Intervals between human genes with orthologs were then matched to the other species to find syntenic regions with the orthologous genes on the same chromosome and the same orientation. In addition to these intergene intervals, the genes themselves were used as intervals in cases where they were overlapped by tRNAs.

These syntenic blocks were then searched for orthologous tRNAs. To predict them, these tRNAs were locally aligned using smith-waterman using the nucleotide identity of the tRNAs to create a scoring function of ten minus the total number of mismatches of the tRNAs aligned with EMBOSS stretcher.

Analysis of bisulfite DNA methylation data and CpG states

Bisulfite data was taken from the epigenomic roadmap project.[57] A methylation level for tRNA genes was calculated by averaging the tRNA methylation level for the window of 100 bases around the tRNA gene.

CpG scores were calculated using the formula Number of CpG * N / (Number of C * Number of G) for calculating CpG enrichment.[68] This was calculated for the 400 base window around the tRNA gene to generate a CpG score.

Discussion

Epigenomic classification of tRNAs

The classification of tRNAs performed here allows for more understanding of how many tRNAs are transcribed in human cells. Previous work done in single or a handful of cell types[69] has suggested that around 224 tRNAs are active, compared with the 334 human tRNAs that are transcribed in this study. Despite differing number of annotated tRNAs, the mouse genome has similar numbers of active tRNAs, with 275 active tRNAs.

This confirms previous work[7] showing that tRNA sequence cannot be the only driver of tRNA expression profile, as identical tRNAs can be in different epigenomic states. However, it does show that tRNA sequence is a major driver of tRNA expression and “tRNAs” without sequences matching the canonical tRNA sequence cannot be transcribed.

This study allows for a comprehensive approach that allows for the determination of the large trends in tRNA regulation. What we find is that there is a broad trend of tRNA regulation that. The major trend we find here is the discovery of a large group of tRNAs that are only active in cell lines. While this work does not indicate how this subset of tRNAs are suppressed, this is consistent with published results on the effect of the MAF1 gene, which has been shown to both be deactivated in cell lines and to target a subset of tRNA genes.[75]

Previous work looking at a small set of tissue types using Pol III showed that said that while tissues could differ, the total number of tRNAs per isotype remained largely the same.[69] This work suggests that differences between tissue showing both that these differences between tissues were relatively minor and that the number of active tRNAs can change significantly in stem cells.

Regulation of Genomic tRNA clusters

The power of the epigenomic categories suggests that other methods that have been shown to affect tRNA expression, such as large-scale chromatin structure[70] cannot be the major drivers of tRNA expression. tRNA expression is not highly individualized is not wholly dependent on related genes, instead, adjacent tRNAs can be differentially regulated, both in the sense of including annotated tRNAs that are never transcribed and those can be transcribed, and in the sense of containing transcribed tRNAs with different expression profiles.

Methylation of tRNA genes and CpG islands

Methylation of tRNA genes has been shown to affect their transcription in Xenopus,[71] making DNA methylation a possible explanation for the deactivation of high-quality tRNAs. However, methylation can be caused by lack of expression, and therefore methylation of tRNAs may be just a result of lack of expression.[66] If this is the case, then this may merely mean that tRNAs are not exempt from this methylation of tRNA.

The heightened CpG frequency near tRNAs suggests a mechanism by which this can happen. The decrease in CpG that steps down as tRNA become less and less active and the presence of methylation in the tRNA increases. This suggests that methylation could be a mechanism for control of tRNA expression.

Conservation and activation of tRNAs

The duplication of tRNA genes within the genome poses two possibilities for their regulation and role. These tRNAs could be superfluous copies, tRNAs that were copied in the genome by transposons but inactive. If these was the case, then the cell might have many specialized tRNAs that have unique attributes. These tRNAs could instead be a method for gene amplification, a way to ensure that tRNAs are able to be transcribed at high rates that can allow mRNAs that contain codons matching those tRNAs to be translated at a different rate. The results of this study, showing that duplicated tRNAs are more likely to be active, suggests that multicopy tRNAs are amplified, they are more likely to be transcribed and tend to be high-scoring.

Activation of tRNAs correlates with conservation of tRNAs across many dimensions. Active tRNAs tend to have sequence and structure closer to the canonical tRNA. Active tRNAs tend to exist in multiple copies, and active tRNAs tend to be conserved in multiple species. Less active tRNAs tend to be less copied, less conserved, and lower scoring.

This suggests that rather a specialization process where tRNAs are selected to be conditionally active with specific signaling markers, tRNAs instead move into that category by simply decaying, either decaying in-place as may be the case for a redundant tRNA or being copied and then slowly mutating. This may be what allows them to escape the concerted evolution[13] process that synchronizes tRNAs.

If there is no conserved set of partially active tRNAs, then tRNAs may just be slowly entering this category from the set of constitutively active tRNAs across evolutionary time and then either being deleted or decaying into unrecognizability. Why tRNAs do not simply have a conserved set of tRNAs for this purpose is not clear, it may be because a conserved set runs the risk of overwriting the constitutive tRNAs through concerted evolution.

Causes of tRNA activation

The coarse nature of tRNA regulation, with many tRNAs always on, many never on, and a specific subset of cells having certain tRNAs expressed, suggest that tRNAs expression is broadly controlled by a single master regulator that adjusts the overall level of tRNA expression in a cell. Some tRNAs have strong enough promoters or CpG islands that this regulator cannot turn them off, such as those in epigroup A, and some are sufficiently weak that no tRNA activation level is capable of turning them on, such as those in epigroup E. tRNAs in epigroups B and C are variable and can respond to specific tRNA signals.

This is consistent with previous work on tRNA master regulators, including MAF1, that act as a global regulator of Pol 3 transcription. This work suggests that Maf1 and similar processes maybe the major regulator of tRNA expression. Other work has shown that adjacent protein-coding genes have an effect on tRNA expression, but the differences between regulation of tRNAs within clusters suggest that this is a minor effect on transcript.

Chapter 3: Cis-regulatory elements in the Archaea

Abstract

Cis-regulatory alements have been shown to be responsible for autoregulation of ribosomal proteins and tRNA synthetases in E. coli. However, these elements have not been well studied in the archaeal domain and very few are known to exist. To find these elements, I developed a method to searching genomes for these elements by aligning orthologous groups of protein-coding and finding signs of conserved structure. These results were combined with small-RNA data which has shown overlap with known cis-regulatory elements in archaea to screen for a set of candidate cis regulatory elements in the archaea in three species. Here I find that ribosomal proteins and transmembrane proteins are highly represented and could be subjects of cis-regulatory regulation

Background

Prokaryotic cis-regulation of transcription and translation

Post-transcriptional regulation of mRNA exists in many forms across all domains of life. Ribosomal proteins are one common target of post-transcriptional modification, as an excess of ribosomal protein is toxic to cells.[76] In eukaryotes, many ribosomal protein genes have autoregulatory structural elements in their RNA that bind the translated protein and an intron from being spliced out, resulting in non-viable transcripts that are then degraded by the nonsense-mediated decay process.[77] The ribosomal proteins S13, S26, and, S16 are known to be regulated using this method in eukaryotes.[78,79]

Bacterial ribosomal proteins are post-transcriptionally autoregulated by the binding of the ribosomal protein to the mRNA, but instead of blocking splicing, bacterial autoregulation generally functions by blocking translation initiation.[78,79] Ribosomal autoregulatory are well-studied in Escherichia coli and related bacteria and several ribosomal protein genes are known to be regulated in this way including

L20[80,81], S15[82], S8[83] and L10[81] ribosomal protein genes. In this form of autoregulation, the rRNA target titrates the ribosomal protein away from its rRNA.[83]

The L1 autoregulatory element is a well-studied example of this type of element.[84] This element exists upstream of the L11 ribosomal protein which is co-transcribed with the L1 ribosomal protein. (Figure 1) In prokaryotic poly-cistronic transcripts, disrupting translation initiation of one open reading frame is often sufficient to disrupt translation of the entire transcript. This disruption is referred to as translational coupling,[84] the mechanism for which is not fully understood. One suggested mechanism is that the non-initial genes in the polycistron are only translated due to the scanning of the 30s ribosomal subunit after the open reading frame ends. It has also been suggested that without ribosomes translating the transcript, a mRNA secondary structure forms that can cause early transcription termination.[84]

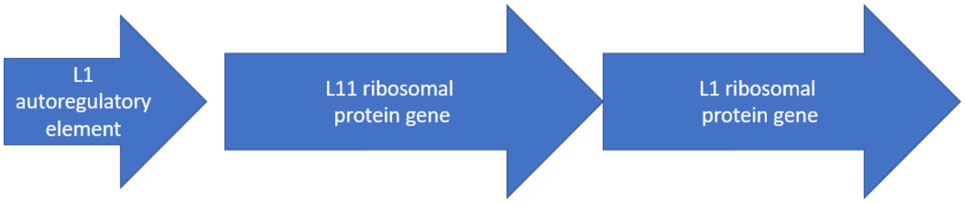

L1 autoregulatory element
L11 ribosomal protein gene
L1 ribosomal protein gene

Figure 14: Structure of the L1 autoregulatory element in E. coli and gamma-proteobacteria. The L1 autoregulatory element exists upstream of the L11 ribosomal protein which upstream of the L1 ribosomal protein in a polycistron. The L1 element can regulate both genes through translational coupling.

Another common form of autoregulatory elements in prokaryotic mRNA are those for the tRNA synthetase genes.[85–87] These elements mimic their tRNA target and allow for regulation in response to an excess of uncharged tRNA, which can happen in response to amino acid starvation.[88] Similarly to auto-regulatory elements in protein-genes, binding of the tRNA synthetase to the mRNA represses the gene. This allows expression to be induced by starvation of amino acids and enhances the efficiency of scavenging amino acids when those amino acids are in short supply. Expression can be reduced when uncharged tRNAs are abundant, which prevents expression of a gene whose function may not be performed. Autoregulation of tRNA synthetases has been shown to occur with threonyl[85], phenylalanine[86], and alanine[87] synthetases.

There are many methods that prokaryotes can use to post-transcriptionally regulate their genes. One common method is to alter the RNA structure to a transcription termination stem, which prevents expression of that gene.[89] Other regulators function by attracting a ribonuclease that degrades the mRNA.[90] Some elements do not down-regulate and instead increase expression rather than decreasing it.[90] These elements can instead allow access to the ribosome binding site or prevent early transcription termination. Most of this published work in post-transcription regulation in prokaryotes has been done in bacteria, and it is currently unknown to what extent these elements exist in archaea.

Riboswitches are another method of transcriptional regulation that works independently of proteins. Riboswitches consist of RNA structures that are capable of binding small molecule ligands without assistance, which causes a change in RNA structure that can block translation, halt transcription, or cleave RNA similarly to how protein cis-acting elements function.[91]

mRNA regulatory structures in the archaea

There are few known mRNA regulatory elements currently known in archaea. One of these elements was found in the transcript of the L1 ribosomal protein in Methanococcus vanellei.[92] This element is found in the 5' UTR of the gene, and functions by preventing translation initiation of that gene. While this autoregulatory element was only experimentally confirmed to exist in that species, this element was predicted to exist in many of the euryarchaea based on comparisons of RNA sequence and structure and is translationally coupled with the other genes in its transcript.[92]

Another mRNA autoregulatory element was found at the 5' end of the thymidylate synthase gene ThyX in Pyrococcus furiosus. This element also is thought to disrupt translation initiation. This gene is known to be auto-regulated in bacteria, but the autoregulatory motif was not shown to exist in other archaea.[93]

The final autoregulatory element known to exist in the archaea is the selenocysteine insertion sequence, commonly abbreviated as the SECIS element. The SECIS element is a conserved structure necessary for the decoding of the selenocysteine codon during translation.[94] Normally, this codon is interpreted as a stop codon, but when the SECIS element is present than this element can be decoded by a special selenocysteine tRNA. This element is found in the 3' UTR of the mRNA of archaea and eukarya.[95,96] In bacteria, this element instead exists in the coding region immediately after the suppressed stop codon.[96]

Finding conserved structure is a commonly used method for finding novel non-coding RNAs, and there are some programs that will perform this task. RNAz, evofold, and qRNA are three published methods for finding conserved secondary structure in homologous sequences.[97] All of these programs take as input a multiple sequence alignment, and score the probability of a conserved structure in that alignment. These programs use either stochastic context-free grammars or compare minimum free energy structures, or MFE structures[98] to compute their scores and any predicted common structure.

RNAz computes MFE RNA structure in a multiple alignment of RNA sequence to find conserved secondary structure. Sequence conservation, number of shared base-pairs, and sequence composition are used as input to a support vector machine used to calculate the probability of conserved secondary structure in the alignment. This was originally used to perform a screen of orthologous genes in the human, mouse, rat, fugu, and zebrafish genomes and was able to identify all known existing elements as well as detecting a number of possible novel elements.[99]

Small RNA sequencing

Small RNAs require their own form of RNA sequencing different from the form used for mRNA analysis.[100] In small RNA sequencing, the RNA is not fragmented prior to sequencing. In commonly used library preparation protocols such as Illumina Truseq or New England Biolabs NEBNext, the RNA is not fragmented prior to reverse transcription and instead has five-prime and three-prime adapter ligated prior to reverse transcription. When this is combined with size selection, it is possible to identify and quantify RNAs of a specific size range in the cell.[101]

For these small RNA-sequencing protocols to function, several conditions must be met. First, the RNA must be linkable by both adapters or the remaining steps in the preparation will not function, this can be prevented by secondary structure or some RNA modifications.[101] Secondly, the reverse transcriptase must be able to read the RNA, which can be prevented by some RNA modifications.[27]

Methods

Finding known RNA structure in new species.

Known secondary structures were found using INFERNAL covariance models. An INFERNAL covariance model for the L1p archaeal element was created using information from both the original paper[92], a published co-crystal structure[102], and this was used to create a covariance model. Covariance models for the SECIS element were created based on the published structure of the archaeal SECIS element. For the L5 and L1 in bacteria covariance models were created based on structures from the original published works[103] and combined with hits in related species including Salmonella enteridis, Shigella flexneri, Photorhabdus luminescens, Aeromonas hydrophila, and Shewanella ANA-3.

Models for the autoregulatory elements for threonine, leucine, and tryptophan synthetases were downloaded from RFAM as models RF00506, RF00512, and RF00513. Cutoffs were taken from the trusted cutoff created for those models.

Gene orthology and alignments

Orthologous clusters are generated from a set of ortholog pairs between all species. These pairs are generate using blastp with default parameters on the protein sequence of these genes, requiring 30% identity and requiring that at least half the gene align. Ortholog pairs are turned into multigene ortholog clusters by combining sets in which all genes are orthologs of each other.

Ortholog groups were generated for three sets of species. The set for searching in Pyrococcus furiosus includes Pyrococcus horokshii, Pyrococcus abyssi, Thermococcus gammotolerans, Thermococcus onnuensis, Thermococcus sibi, Thermococcus 4557, Pyrococcus yayanossi, and Thermococcus kodakerensis. The set for searching Methanocaldococcus janaschii includes Methanosphaera stadtmanae, Methanobrevibacter ruminantium, Methanothermobacter marburgensis, Methanobacterium thermoautotrophicum, Methanococcus maripaludis S2, Methanococcus maripaludis C5, Methanocaldococcus FS406, Methanococcus maripaludis C5, Methanocaldococcus vulcanius, Methanocaldococcus infernus, Methanocaldococcus fervens, Methanopyrus kandleri, and Methanobrevibacter smithii. The set for searching Pyrobaculum aerophilum includes Pyrococcus calidifontis, Pyrobaculum islandicum ,Pyreobaculum neutrophilus, Pyrobaculum oguneiense, Pyrobaculum 1860, Vulcanasaeta distribute, Thermoproteus tenax, and Caldivirga maquilensesis.

The nucleotide sequences of these gene clusters including the open reading frame and 100 bases of flanking sequence were aligned using the MAFFT multiple aligner. These alignments were then turned into MAF format files with a custom python script. These MAF files were then used as input to RNAz

Genome-wide screens for novel RNA structure

Gene alignments are turned into tiled windows that are 100 bases long with 10 base pair steps. These alignments are used as inputs into RNAz with default parameters for the entire genome. Regions with greater than .95 probability and greater than three species in the alignment are considered to be potential sites of conserved RNAs. As a null set, the 100 base tiled windows that are used as input to RNAz are shuffled into random alignments using multiperm[104] with default parameters. These shuffled alignments are then used as the input to RNAz to compute false positives.

Lists of potential targets were created by searching the alignments overlapping Methanocaldococcus janaschii, Pyrococccus furiosus, and Pyrobaculum aerophilum for regions with overlapping RNAz hits with probabilities greater than .95. This set was intersected with RNA elements with >20 reads to create the final candidate set.

Small RNA sequencing and analysis

Small RNA from organisms were prepared from cells grown in the lab by David Bernick using an protocol developed by him for sequencing small RNA sequencing generated and mapped using previously described protocols.[105] This sequencing was done using an in-house sequencing protocol sequencing reads less than tRNA size as identified on a gel.

Sequencing reads were mapped using blat allowing no introns, a minimum score of 40 and a tile size of 15 and filtered for all-best reads. Small RNA transcripts were generated by finding regions with a read coverage of greater than 20.

Results

Messenger RNA autoregulatory elements are broadly conserved and detectable

While many autoregulatory elements have been found in bacteria, there is little known about the spread of these elements in related species. Before performing screens for these elements, it is necessary to know how these known elements are conserved in bacteria.

To determine the spread and detectability of prokaryotic ribosomal autoregulatory sequences, I used infernal to build models based on the published sequences and structure of the L1 ribosomal protein autoregulatory elements found in E. coli K12 upstream of the L11 ribosomal protein gene and expanded it with sequences from five additional gamma-proteobacteria genomes. These models were then used to search 27 genomes in the gamma-proteobacteria.

The L1 gene is present upstream of the L11 protein in 24 out 27 of the searched genomes. While the score rises above the .05 significance threshold for some of these species, the location of this element in front of the L11 suggests that this is merely a less well-matching form of the autoregulatory structure. Strong hits to this element with a bit-score of greater than 25 bits and an e-value of less than .068 are all matches to the correct region in the genome. (Table 4)

In these matching species, the element is located 92 or 93 bases upstream of the start codon, making the end of the element overlap the start of the gene. The presence of the element 92-93 nucleotides upstream of genes that are not the L11 ribosomal protein could suggest that an element like this is engaged in alternate regulatory pathways in some species.

species	strand	E-value	Locus name	gene description	distance
Escherichia coli K12	+	1.30E-14	b3983	50S ribosomal subunit protein L11	92
Shigella flexneri 2a	+	1.40E-14	SF4056	50S ribosomal protein L11	92
*Salmonella enterica serovar Enteritidis P125109	+	2.00E-13	SEN3933	50S ribosomal protein L11	92
Photorhabdus luminescens	+	1.40E-09	plu0435	50S ribosomal protein L11	92
Haemophilus influenzae Rd KW20	-	1.10E-08	HI0517	50S ribosomal protein L11	93
Yersinia pestis CO92	-	3.30E-08	YPO3751	50S ribosomal protein L11	93
Actinobacillus pleuropneumoniae L20	+	4.10E-08	APL_1718	50S ribosomal protein L11	92
Aeromonas hydrophila ATCC 7966	-	2.50E-06	AHA_4032	50S ribosomal protein L11	93
Shewanella ANA-3	+	2.50E-05	Shewana3_0188	50S ribosomal protein L11	92
Methylococcus capsulatus Bath	+	3.10E-05	MCA1062	ribosomal protein L11	92
Sodalis glossinidius morsitans	+	0.00012	SG0130	50S ribosomal protein L11	92
Vibrio vulnificus YJ016	-	0.00036	VV3163	50S ribosomal protein L11	93
Acinetobacter sp ADP1	+	0.002	ACIAD0302	50S ribosomal protein L11	92
Dichelobacter nodosus VCS1703A	-	0.0011	DNO_1286	50S ribosomal protein L11	93
Xylella fastidiosa	-	0.0025	XF2637	50S ribosomal protein L11	93

Chromohalobacter salexigens DSM 3043	+	0.0043	Csal_0410	50S ribosomal protein L11P	92
Pseudoalteromonas haloplanktis TAC125	+	0.0062	PSHAa0218	50S ribosomal subunit protein L11	92
Hahella chejuensis	-	0.034	HCH_06228	ribosomal protein L11	93
Coxiella burnetii	+	0.03	CBU_0226	ribosomal protein L11	92
Saccharophagus degradans 2-40	+	0.095	Sde_0920	50S ribosomal protein L11P	92
Francisella tularensis tularensis	+	0.068	FTT0140	50S ribosomal protein L11	92
Thiomicrospira crunogena XCL-2	+	0.52	Tcr_0346	chaperonin Cpn60/TCP-1	92
Colwellia psychrerythraea 34H	+	1.5	CPS_4845	branched-chain amino acid aminotransferase	92
Alcanivorax borkumensis SK2	+	1	ABO_0374	50S ribosomal protein L11	92
Alkalilimnicola ehrlichei MLHE-1	+	1.5	Mlg_0447	50S ribosomal protein L11P	92
Idiomarina loihiensis L2TR	+	2.1	IL0341	50S ribosomal protein L11	92
Nitrosococcus oceani ATCC 19707	-	4.9	Noc_2373	hypothetical protein	-189

Table 4: Table of gamma-proteobacterial L1p model hits to genomes showing wide spread of this element within the gamma-proteobacteria, sorted by e-value. The model hits in the same location in 26/27 of those species, with 23/27 having a score of less than .05. Distance is the number of bases from the start of the open reading frame to the start of the regulatory element, with negative numbers indicating that the element overlaps the gene.

The L5 element model was also constructed using published sequence and structure that was expanded to five additional genomes in the gamma-proteobacteria. This element is less common than the L1 autoregulatory element, present upstream of the L5 polycistron in 14/27 of the gamma-proteobacteria genomes searched.

To determine if this method works for autoregulatory elements that are not ribosomal, I searched for autoregulatory elements of tRNA synthetases. Using models from RFAM[106], the presence of the threonine is present in 11 species, the presence of the tryptophan autoregulatory elements is present in 11 species, and the presence of the leucine element in 10 species. While these elements are less conserved than ribosomal autoregulatory elements, their conservation is still detectable.

This evidence suggests that these mRNA protein-binding structural elements are conserved widely, if not universally, and that these elements can be detected in related species using the structure and sequence of known elements. This also suggests that these elements may be detectable using methods for finding conserved RNA structure.

New method for detecting these RNA Conserved structure from groups of related species

To further develop this as a method for finding new RNA structural elements, I developed a method for searching mRNAs for conserved secondary structure that can be compared with the small-RNA sequencing method. This method finds orthologous gene clusters in sets of related species and then aligns their mRNA sequences using a multiple aligner. This creates a set of multiple sequence alignments that can be used as an input to RNAz. These regions can then be compared with these known sites of bacterial secondary structure to determine if this is a method suitable for finding cis-regulatory elements.

This method was used with Escherichia coli and a set of related genomes to predict possible sites of conserved secondary structure. This predicts conserved structure overlapping the predicted L1 element with probabilities greater than .99, suggesting that this can be used to detect these cis-regulatory RNA elements.

In addition to predicting the known structure, RNAz also predicts additional conserved structure upstream of this element that does not overlap the canonical autoregulatory element. which could be additional conserved sequence that either binds the protein or additional structure. (Figure 15)

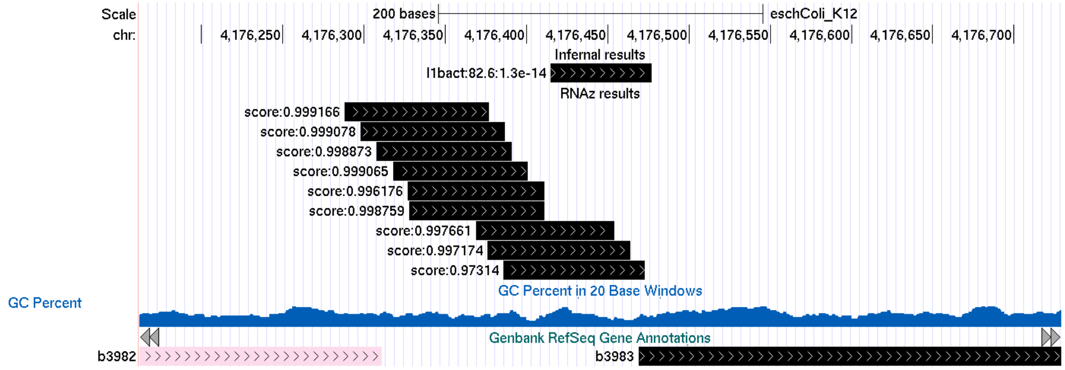


Figure 15: Location of the L1 autoregulatory region of the E coli K12 genome upstream of the L11 ribosomal protein gene b3983. The RNAz tool for predicting conserved secondary predicts structure overlapping this region as well as several hundred bases upstream.

The L1p autoregulatory element is found across the euryarchaeal

While predictions of conservation were made when the L1p ribosomal protein was discovered,[92] new sequenced species and new methods of finding conserved structure can be used to augment these predictions to find the spread of this element in the archaea. This can provide evidence on how these elements will be conserved in archaea.

To detect the l1p element in the archaea, I built a model for the L1p ribosomal autoregulatory element based on the original paper a published X-ray structure[102] (Figure 14) This model included 13 euryarchaeal species, 4 halophiles, 5 thermophiles, and 4 from other euryarchaeal groups. Searching with this model reveals that this structure exists at the five-prime end of 18 euryarchaeal genomes spread over all major euryarchaeal groups including methanogens, halophiles, and thermophiles, while it is missing in 6 of the searched euryarchaeal genomes.   This element is not present in any genome outside the euryarchaea. (Table 5)

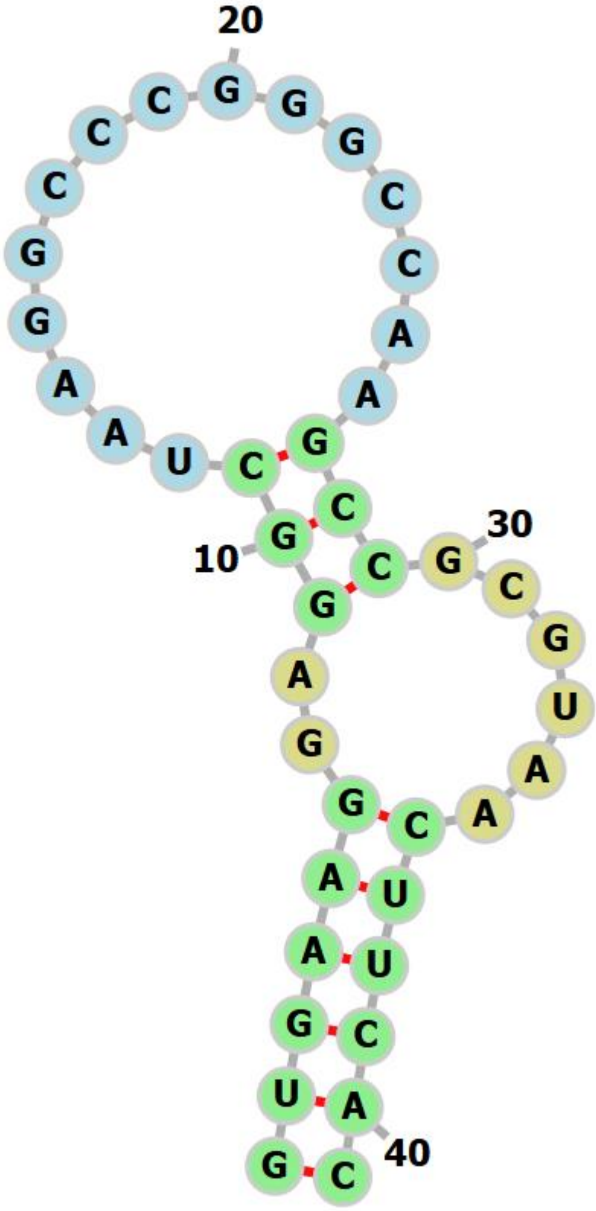


Figure 16: Structure of L1p autoregulatory element in Pyrococcus furiosus generated using cmalign showing secondary structure of rhe L1p element.

species	Phylum	strand	E-value	locus name	gene description	distance
Pyrococcus furiosus	Euryarchaea	+	0.00014	PF1992	50S ribosomal protein L1P	26
Pyrococcus yayanosii	Euryarchaea	+	0.00013	PYCH_18190	50S ribosomal protein L1P	26
Thermococcus onnurineus	Euryarchaea	+	0.00014	TON_0180	50S ribosomal protein L1P	26
Thermococcus barophilus MP	Euryarchaea	+	0.00015	TERMP_00194	50S ribosomal protein L10Ae (L1p)	26
Thermococcus kodakarensis	Euryarchaea	-	0.0002	TK1417	50S ribosomal protein L1P	27
Methanococcus maripaludis S2	Euryarchaea	-	0.00017	MMP0260	50S ribosomal protein L1P	33
Archaeoglobus veneficus SNP6	Euryarchaea	+	0.00061	Arcve_0942	50S ribosomal protein L1	109
Methanocaldococcus fervens	Euryarchaea	+	0.00081	Mefer_0703	ribosomal protein L1	32
Ferroglobus placidus	Euryarchaea	+	0.0015	Ferp_0266	ribosomal protein L1	109
Methanothermobacter thermautotrophicus	Euryarchaea	+	0.003	MTH1680	50S ribosomal protein L1P	-1
Methanobacterium sp. SWAN-1	Euryarchaea	-	0.0049	MSWAN_0393	50S ribosomal protein L1	36
Halogeometricum borinquense DSM 11551	Euryarchaea	-	0.018	Hbor_12250	LSU ribosomal protein l1p	110
Haloarcula hispanica ATCC 33960	Euryarchaea	+	0.037	HAH_1998	50S ribosomal protein L1P	109
Methanosphaera stadtmanae	Euryarchaea	-	0.018	Msp_1265	50S ribosomal protein L1P	36
Methanococcus aeolicus	Euryarchaea	-	0.019	Maeo_0189	50S ribosomal protein L1P	33
Halorhabdus utahensis DSM 12940	Euryarchaea	+	0.049	Huta_0254	ribosomal protein L1	109
Methanocorpusculum labreanum Z	Euryarchaea	-	0.043	Mlab_1581	50S ribosomal protein L1P	114
Acidilobus saccharovorans 345-15	Crenarchaea	-	0.036	-	-	-
Pyrococcus abyssi	Euryarchaea	-	0.052	PAB1208	triosephosphate isomerase	-176

Pyrobaculum calidifontis	Crenarchaea	-	0.1	Pcal_1117	DEAD_2 domain-containing protein	-731
Aeropyrum pernix	Crenarchaea	+	0.11	-	-	-
Halomicrobium mukohataei DSM 12286	Euryarchaea	-	0.25	Hmuk_2186	ribosomal protein L1	110
Thermofilum pendens	Crenarchaea	+	0.17	Tpen_0853	tRNA-modifying enzyme	-382
Sulfolobus tokodaii	Crenarchaea	-	0.29	ST2411	hypothetical protein	-87
Halorubrum lacusprofundi ATCC 49239	Euryarchaea	-	0.57	Hlac_2535	ribosomal protein L1	109
Thermoproteus neutrophilus V24Sta	Crenarchaea	+	0.33	-	-	-
Halalkalicoccus jeotgali B3	Euryarchaea	+	0.73	HacjB3_08560	50S ribosomal protein L1P	109
Ignicoccus hospitalis	Crenarchaea	+	0.26	Igni_1390	molybdenum cofactor synthesis domain-containing protein	-661
Methanopyrus kandleri	Euryarchaea	+	0.85	MK0566	flap endonuclease-1	-899
Candidatus Korarchaeum cryptofilum OPF8	Korarchaeota	+	1.3	Kcr_0601	glycine dehydrogenase subunit 1	-541
Cenarchaeum symbiosum A	Thaumarchaeota	-	1.8	CENSYa_0522	hypothetical protein	222
Methanosarcina mazei	Euryarchaea	+	4	MM_2515	sensory transduction histidine kinase	-1987
Haloquadratum walsbyi	Euryarchaea	-	3.5	HQ3564A	hypothetical protein	250
Methanosarcina acetivorans	Euryarchaea	-	7.3	MA1137	hydrogenase expression/formation protein	-374
Vulcanisaeta distributa DSM 14429	Crenarchaea	+	3.7	Vdis_0828	beta-lactamase domain-containing protein	-242

Caldivirga maquilingensis	Crenarchaea	-	3.8	Cmaq_1974	RdgB/HAM1 family non-canonical purine NTP pyrophosphatase	43
Desulfurococcus kamchatkensis	Crenarchaea	+	5.2	DKAM_0196	gamma-glutamyltranspeptidase	-1142
Thermoplasma acidophilum	Euryarchaea	+	9.7	Ta0477	hypothetical protein	-606
Picrophilus torridus	Euryarchaea	-	-	-	-	-
Sulfolobus solfataricus	Crenarchaea	-	-	-	-	-
Nanoarchaeum equitans	Nanoarchaeota	-	-	-	-	-
Nitrosopumilus maritimus SCM1	Thaumarchaeota	-	-	-	-	-

Table 5: Hits of the archaeal L1p Infernal model to archaeal genomes showing hits in most euryarchaeal genomes. Model was generated using alignment of 12 species, searched on a set of 42 archaeal species. E-values are reported for searches on each genome individually. Distance is the number of bases upstream of the start position of the gene, negative positions are hits that overlap the RNA.

This motif exists at varying distances upstream of the annotated start site, often overlapping the annotated gene start. This may be a result of misannotation, the original paper that discovered this element suggested that the annotated start codon was incorrect.[92]

No L1p autoregulatory structure is detect in any of the crenarchaea species searched, including the hyperthermophiles and acidophiles. In the crenarchaea, the L1 gene does not lead a ribosomal protein polycistron like it does in many euryarchaeal species and is instead located downstream of the L11 ribosomal protein gene. While this is the only known ribosomal autoregulatory element in archaea, this suggests that archaeal ribosomal autoregulatory elements are conserved similarly to how those in bacteria are.

Small RNA sequencing is a signature for protein/RNA interactions

In small-RNA sequencing that was done to detect small RNAs[105], it was found that there are many small RNAs that overlap protein-coding genes. These fragments, ostensibly parts of mRNA, are often compact and numerous enough to suggest that they are the result of a specific biological process. (Figure 17)

Small-RNA sequencing performed on archaeal species[105] revealed reads overlapping the predicted location of the L1p autoregulatory element. In the species for which small-RNA sequencing was performed, small RNA reads were found overlapping the predicted structure in Thermococcus kodakerensis, Pyrococcus furiosus, Haloferax volcanii, and Methanocaldococcus jannaschii.

The selenocysteine insertion element or SECIS element is a similar RNA structural element found in mRNA,[94] this structure is found at the three-prime of the gene and is responsible for suppressing a stop codon and allowing instead the insertion of a selenocysteine protein.[107] Sequencing of methanogens shows reads at the location of the SECIS element in Methanocaldococcus jannaschii. The ThyX element in Pyrococcus furiosus[93] also contains overlapping reads.

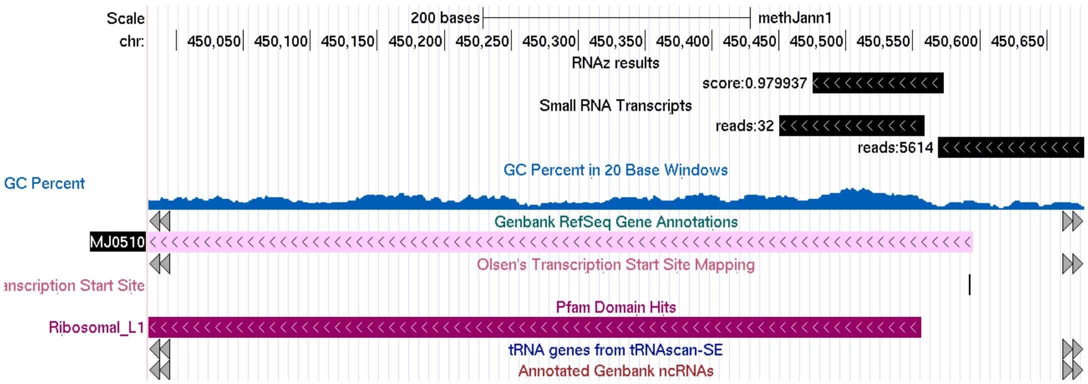


Figure 17: Start of L1p gene in Methanocaldoccus janaschii showing RNAz results. These RNAz hits are the probability from the multiple gene alignments on tiled windows across the genome, while small RNA transcripts show transcript locations

and total read counts. This shows a RNAz detected structured RNA overlapping a predicted transcript

These reads are part of population that can be called cis RNA fragments. These cis RNA fragments are present in all sequenced archaeal species, making up a comparable percentage of reads to cis-antisense RNAs. (Figure 18) These can make up between 3% and 20% of reads depending on the genome. These reads are concentrated in specific regions of the mRNA rather than being scattered.

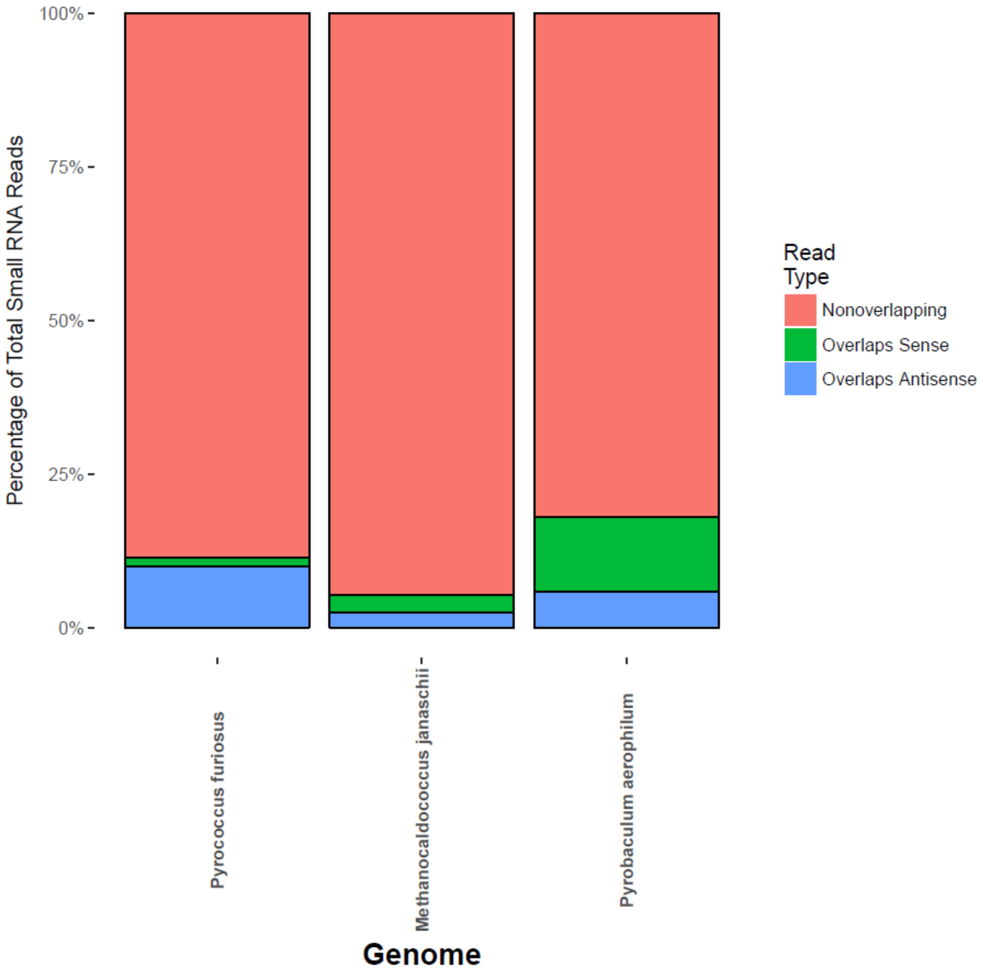


Figure 18: Read coverage of feature types for RNA sequencing runs in multiple species. Reads overlapping protein-coding genes make up 10% of the reads in Pyrobaculum aerophilum while representing smaller fractions in Methanocaldococcus jannaschii and Pyrococcus furiosus

RNAz analysis of orthologous gene sets suggests new elements in archaea

To determine the effectiveness of the RNAz pipeline developed for searching genomes for mRNA conserved sequence, I compared RNAz scores generated from alignments of genes with RNAz scores generated from gene alignments shuffled using an alignment shuffling algorithm that preserves dinucleotide frequency and alignment structure.[104] As the nucleotide frequency and alignment structure are used by RNAz, this method should overestimate the number of false positives present in RNAz results. The scores from these shuffled alignments in Pyrococcus furiosus were compared to the results from the genome to determine if RNAz determined structural elements are enriched.

This test revealed that more RNAz hits exist in gene alignments than are found in randomly shuffled alignments. For alignments with a probability score greater than .95, more than 4 times as many hits are found in shuffled alignments. This is consistent with RNAz detecting a number of real RNA structural elements using this method, although it does indicate that false positives will be an issue. (Figure 19)

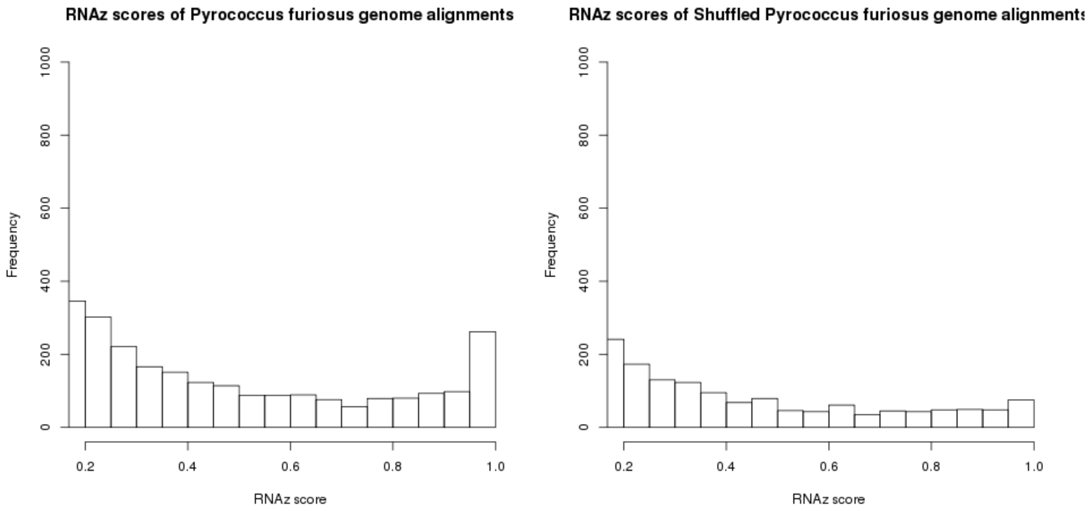


Figure 19: Plots of RNAz scores for Pyrococcus furiosus compared to scores of shuffled alignments showing high scoring tRNA. Shuffled alignments were generated using multiperm[104].

Potential new structural elements found using both conserved structure prediction and small-RNA sequencing

Using the combination of small RNA reads and conserved structure prediction then yields several new potential sites of cis-regulatory elements. Searching the species Pyrococcus furiosus, Methanocaldococcus janaschii, and Pyrobaculum aerophilum reveals potential new regulatory structures within their mRNA.

This pipeline searches the genome for regions of overlapping RNAz hits and outputs those regions that also includes overlapping small RNA reads. In addition to detecting cis-regulatory elements, this method detects RNAs including tRNAs and other non-coding RNAs that must be filtered out. For the final pipeline, regions with RNAz scores of greater than .95 and more than 20 reads were chosen as the set of potential new autoregulatory elements.

In Pyrococcus furiosus this method detects 21 potential autoregulatory regions. (Table 6) Six of these are ribosomal proteins, including the L44e, l15e, s19e, l12p, and s15p ribosomal proteins. In addition to ribosomal proteins, this detects an RNA polymerase subunit, The NOP5/NOP56 related protein which is part of the C/D box RNP.

score	reads	gene	Gene description
0.951095	96	PF0060	NOP5/NOP56 related protein
0.987863	527	PF0217	L44e ribosomal protein
0.99628	56	PF0298.1n	hypothetical gene
0.993551	13480	PF0488	S6e ribosomal protein
0.9897	148	PF0669	hypothetical protein
0.999021	39	PF0693	hypothetical gene
0.985463	38	PF0722	Peroxiredoxin
0.999335	186	PF0820	cytilidate kinase
0.98103	23	PF0825	prolyl endopeptidase
0.954695	182	PF0876	L15e ribosomal protein
0.980384	101	PF0987	hypothetical gene
0.999564	254	PF1061	hypothetical gene
0.990916	74	PF1499	S19e ribosomal protein
0.99727	149	PF1541	Ribosomal protein L37e
0.995258	132	PF1562	DNA-directed RNA polymerase subunit A"
0.966447	553	PF1586	hydrogenase expression/formation protein
0.984353	767	PF1622	n-type ATP pyrophosphatase superfamily protein
0.974811	37	PF1687	hypothetical gene
0.990304	638	PF1778	serine hydroxymethyltransferase
0.995663	200	PF1992	L1p ribosomal protein
0.998988	112	PF1994	L12p ribosomal protein
0.955607	1205	PF2056	S15p ribosomal protein

Table 6: Table of novel cis-regulatory elements in Pyrococcus furiosus.. List of genes that contain both RNAz hits with >.95 probability and overlapping small RNA reads.

In Methanocaldococcus jannaschii, there are 35 potential targets including the S15p, S13p, L18e, s10p, l14e, p0 and l39e ribosomal proteins.(Table 7) In addition to these genes, an RNA polymerase subunit, two transcription initiation factors, and flagella genes which are known to be autoregulated in some bacterial species.[108]

score	reads	gene	Gene description
0.980128	141	MJ0035	ABC transporters subunit
0.999739	327	MJ0036	S15p ribosomal protein
0.990781	76	MJ0189	S13p ribosomal protein
0.99573	361	MJ0192	DNA-directed RNA polymerase, subunit D (rpoD)
0.95943	36	MJ0193	L18e ribosomal protein
0.960664	279	MJ0215	hypothetical protein
0.98843	160	MJ0216	V type ATP synthase subunit B
0.95373	107838	MJ0226	Xanthosine triphosphate pyrophosphatase
0.95213	21	MJ0262	Transcription initiation factor IF-2
0.955026	78	MJ0275.1	hypothetical protein
0.953476	704	MJ0278	FKBP-type peptidyl-prolyl cis-trans isomerases 2
0.975805	94	MJ0282	hypothetical protein
0.985069	227	MJ0322	S10p ribosomal protein
0.999859	561	MJ0464	RNase p component 1
0.979456	316	MJ0509	acidic ribosomal protein p0
0.979937	32	MJ0510	L1p ribosomal protein
0.996948	33	MJ0590	Acyl-CoA synthetase (NDP forming)
0.996826	55	MJ0609	Amino acid transporter
0.999453	755	MJ0657	L14e ribosomal protein
0.999732	169	MJ0689	L39e ribosomal protein
0.950621	197	MJ0739	hypothetical protein
0.997176	58	MJ0782	transcription initiation factor IIB (TFIIB)
0.955149	541	MJ0843	Methyl coenzyme M reductase, subunit D

0.993601	382	MJ0893	flagellin
0.984578	847	MJ0897	Putative archaeal flagellar protein F
0.991925	34	MJ1036	Predicted divalent heavy-metal cations transporter
0.999603	175	MJ1190a	Coenzyme F420-reducing hydrogenase, delta subunit
0.979118	34	MJ1270	ABC-type branched-chain amino acid transport system, permease component
0.971099	32	MJ1330	Phosphatidylglycerophosphatase A fused to adenosylcobinamide amidohydrolase, CbiZ
0.986095	53	MJ1406	Aspartate carbamoyltransferase, regulatory subunit
0.962692	14241	MJ1569	Cobalt transport protein CbiM
0.977401	218	MJ1635	Transposase
0.991883	334	MJ1635	Transposase

Table 7: List of potential cis-regulatory elements in Methanocaldoccous janaschii. List of genes that contain both RNAz hits with >.95 probability and overlapping small RNA reads.

In Pyrobaculum aerophilum, there are 24 potential targets, including the L15e and L18 ribosomal proteins. (Table 8) Similar to Methanocaldoccus janaschii, flagellar genes are also present in these results, along with transmembrane proteins. Many of the predicted genes are annotated as hypothetical genes, this is likely due to the high number of such annotations in the genome.

Score	Reads	gene	Gene description
0.98688	70	PAE0057	Hypothetical protein
0.998411	25	PAE0098	Hypothetical protein
0.975151	37	PAE0252	diadenosine 5'5'''-P1,P4-tetraphosphate pyrophosphohydrolase (mutT/nudix family protein)
0.983217	39	PAE0923	Uncharacterized conserved protein
0.967452	80	PAE1001	Uncharacterized conserved protein
0.968377	32	PAE1002	Dipeptidyl aminopeptidase/acylaminoacyl-

			peptidase
0.994448	42	PAE1077	Uncharacterized conserved protein
0.993079	103	PAE1079	Uncharacterized conserved protein
0.978332	82	PAE1421	Hypothetical protein
0.999708	85	PAE1501	Triosephosphate isomerase
0.997155	398	PAE1833	Ribosomal protein L15E
0.999234	6817	PAE2020	Predicted membrane protein
0.988444	470	PAE2040	Transglutaminase-like enzymes, putative cysteine proteases
0.993032	482	PAE2101	Ribosomal protein L18
0.993939	27	PAE2159	Dehydrogenases (flavoproteins)
0.98852	22	PAE2387	Uncharacterized conserved protein
0.973274	31	PAE2388	Predicted permease
0.996912	42	PAE2420	cystathionine gamma-synthase
0.979383	42	PAE2511	Uncharacterized conserved protein
0.951243	22	PAE2981	acetylpolyamine aminohydrolase, putative
0.991325	23	PAE3162	HIT family protein
0.9535	94	PAE3209	permease of the drug/metabolite transporter (DMT) superfamily
0.98723	33	PAE3271	DNA-directed RNA polymerase subunit K
0.999283	137	PAE3330	RNA-binding protein involved in rRNA processing

Table 8: Table of potential cis-regulatory elements in Pyrobaculum aerophilum. List of genes that contain both RNAz hits with >.95 probability and overlapping small RNA reads.

In addition to ribosomal proteins, RNA polymerase subunits are found in all three of these species, the RNA polymerase subunit A" in Pyrococcus furiosus, the RNA polymerase, subunit D in Methanocaldococcus jannaschii, and the RNA polymerase subunit K in Pyrobaculum aerophilum all contain overlapping potential targets. While the elements for Pyrococcus furiosus and Pyrobaculum aerophilum are located on the three-prime end of the annotated gene, the element for Methanocaldococcus jannaschii is located at the five-prime end of the gene.

Ribosomal proteins also show up in these species, although most of these proteins are not shared between species. Apart from the L1 protein, the S15p protein is the only one that appears in more than one species, also showing up in

Methanocaldococcus jannaschii and Pyrococcus furious. The other type of protein commonly found here are transmembrane proteins, including transporters and permeases.

While the accuracy of these this method in finding cis-regulatory elements is not known, this method appears to work as a screen for selecting candidate structural elements that can be followed up on directly. This could allow us to determine the prevalence of these elements in archaea.

Discussion

Known cis-regulatory elements

Cis-regulatory elements were previously found in many cases in partial genome sequences that were produced before full genome sequencing became practical. The motivation for this project was an attempt to continue this project using whole-genome sequencing, modern computational techniques for finding RNA secondary structure, and RNA sequencing.

The presence of ribosomal autoregulatory elements in all three domains of life suggest these elements are crucial. While in prokaryotes autoregulatory ribosomal elements often function by blocking translation start of the first gene of the polycistron, in eukaryotes where these elements have been found they function by blocking splicing and triggering nonsense-mediated decay.[109] That these elements can be so universal and yet so diverse raises questions about how these elements are created and destroyed over evolutionary time.

While autoregulatory elements are present in all domains of life, the individual elements present in any genome can differ. In prokaryotic genomes, as ribosomal proteins are reshuffled into different operon positions the nature of the elements can change. The l1 archaeal element was only conserved in the euryarchaeal genomes, and the elements in E. coli can be conserved across all gamma-proteobacteria or only a subset of them.

Finding novel cis-acting regulatory elements

This method described in this paper allows for high-throughput searches for mRNA secondary structure. While secondary structure prediction has been used to detect cis-regulatory elements[110], adding an additional layer of data from small RNA sequencing can allow for more accurate searching. The wide range of these elements in bacteria suggests that there may be many more of these to discover in archaea. While the prevalence of extremophiles may mean that these elements are less common, even extremophilic archaea still use RNA structure such as snoRNAs, and in fact thermophilic archaea possess more snoRNA than mesophiles.[111] There is no reason why these elements should not be prevalent in archaea.

What causes these small RNA footprints in the small RNA sequencing data is not currently known. Some of these elements are known to block translation, and it is possible that this triggers a process similar to nonsense-mediated decay that cuts non-translating mRNA. One possibility is that these elements are preserved from ordinary RNA turnover due to the fact that their protein binding targets protect them. If this is the case, then riboswitches will not be detected using small RNA reads.

Comparison with existing methods

This is not the first method to attempt high-throughput searches for RNA structure or mRNA structure, previously this has been done in an attempt to find riboswitches, such as cmfinder.[112] This method differs from riboswitch finding methods in a number of respects.

CMfinder and other methods require a specific target in the form of an RNA alignment or set of alignable sequences. This method creates the alignments itself using the annotation of protein-coding genes and blastp. This allows for easy searching of whole genomes.

The second difference is in dealing with sequences within the open reading frame of protein-coding genes. Ribosomal autoregulatory elements are often located overlapping or within the annotated ORF of a protein coding gene. While this can be a product of misannotation, this requires that searches for conserved structure can successfully handle open reading frames and not confuse the change in nucleotide and dinucleotide frequencies in an ORF from the change associated with RNA structure. The use of shuffled alignments here is used to test these methods.

Rip-seq and similar methods have been used for for high-throughput searching for RNA-protein interaction.[113] The advantage of this method over rip-seq is that it allows for a broad search that does not start with a specific protein. Unfortunately, this also means that it is not possible to tell what the target is. In the case of ribosomal proteins or RNA interacting protein, then autoregulation makes target prediction straightforward, but in other cases it may be harder to find the regulatory partner.

Bibliography

1. Cozen, A. E. *et al.* ARM-seq: AlkB-facilitated RNA methylation sequencing reveals a complex landscape of modified tRNA fragments. *Nat Meth* **12,** 879–884 (2015).

2. Phizicky, E. M. & Hopper, A. K. tRNA biology charges to the front. *Genes Dev* **24,** 1832–1860 (2010).

3. Agris, P. F., Vendeix, F. A. P. & Graham, W. D. tRNA's wobble decoding of the genome: 40 years of modification. *J. Mol. Biol.* **366,** 1–13 (2007).

4. Pang, Y. L. J., Poruri, K. & Martinis, S. A. tRNA synthetase: tRNA aminoacylation and beyond. *Wiley Interdiscip Rev RNA* **5,** 461–480 (2014).

5. Sprinzl, M. & Gauss, D. H. Compilation of tRNA sequences. *Nucl. Acids Res.* **12,** r1–r57 (1984).

6. Goodenbour, J. M. & Pan, T. Diversity of tRNA genes in eukaryotes. *Nucleic Acids Res.* **34,** 6137–6146 (2006).

7. Sagi, D. *et al.* Tissue- and Time-Specific Expression of Otherwise Identical tRNA Genes. *PLOS Genetics* **12,** e1006264 (2016).

8. Dana, A. & Tuller, T. The effect of tRNA levels on decoding times of mRNA codons. *Nucleic Acids Res* **42,** 9171–9181 (2014).

9. Salinas-Giegé, T., Giegé, R. & Gieg*é, P. tRNA biology in mitochondria. *Int J Mol Sci* **16,** 4518–4559 (2015).

10. Yoshihisa, T. Handling tRNA introns, archaeal way and eukaryotic way. *Front Genet* **5,** 213 (2014).

11. Kramerov, D. A. & Vassetzky, N. S. Origin and evolution of SINEs in eukaryotic genomes. *Heredity* **107,** 487–495 (2011).

12. Mungall, A. J. *et al.* The DNA sequence and analysis of human chromosome 6. *Nature* **425,** 805–811 (2003).

13. Liao, D. Concerted Evolution: Molecular Mechanism and Biological Implications. *The American Journal of Human Genetics* **64,** 24–30 (1999).

14. Amstutz, H., Munz, P., Heyer, W.-D., Leupold, U. & Kohli, J. Concerted evolution of tRNA genes: Intergenic conversion among three unlinked serine tRNA genes in S. pombe. *Cell* **40,** 879–886 (1985).

15. Rogers, H. H., Bergman, C. M. & Griffiths-Jones, S. The Evolution of tRNA Genes in Drosophila. *Genome Biol Evol* **2,** 467–477 (2010).

16. Lowe, T. M. & Eddy, S. R. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic acids research* **25,** 955–64 (1997).

17. Nawrocki, E. P. & Eddy, S. R. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics (Oxford, England)* **29,** 2933–5 (2013).

18. Weinmann, R. & Roeder, R. G. Role of DNA-Dependent RNA Polymerase III in the Transcription of the tRNA and 5S RNA Genes. *PNAS* **71,** 1790–1794 (1974).

19. Lin-Marq, N. & Clarkson, S. G. Efficient synthesis, termination and release of RNA polymerase III transcripts in Xenopus extracts depleted of La protein. *EMBO J.* **17,** 2033–2041 (1998).

20. Orioli, A., Praz, V., Lhôte, P. & Hernandez, N. Human MAF1 targets and represses active RNA polymerase III genes by preventing recruitment rather than inducing long-term transcriptional arrest. *Genome Res* **26,** 624–635 (2016).

21. Hopper, A. K. Transfer RNA post-transcriptional processing, turnover, and subcellular dynamics in the yeast Saccharomyces cerevisiae. *Genetics* **194,** 43–67 (2013).

22. Machnicka, M. A. *et al.* MODOMICS: a database of RNA modification pathways–2013 update. *Nucleic acids research* **41,** D262–7 (2013).

23. El Yacoubi, B., Bailly, M. & de Crécy-Lagard, V. Biosynthesis and function of posttranscriptional modifications of transfer RNAs. *Annual review of genetics* **46,** 69–95 (2012).

24. Phizicky, E. M. & Alfonzo, J. D. Do all modifications benefit all tRNAs? *FEBS letters* **584,** 265–71 (2010).

25. Paris, Z., Fleming, I. M. C. & Alfonzo, J. D. Determinants of tRNA editing and modification: avoiding conundrums, affecting function. *Seminars in cell & developmental biology* **23,** 269–74 (2012).

26. Torres, A. G., Batlle, E. & Ribas de Pouplana, L. Role of tRNA modifications in human diseases. *Trends in Molecular Medicine* **20,** 306–314 (2014).

27. Motorin, Y., Muller, S., Behm-Ansmant, I. & Branlant, C. Identification of modified residues in RNAs by reverse transcription-based methods. *Methods in enzymology* **425,** 21–53 (2007).

28. Ryvkin, P. *et al.* HAMR: high-throughput annotation of modified ribonucleotides. *RNA* **19,** 1684–1692 (2013).

29. Liu, J. & Jia, G. Methylation modifications in eukaryotic messenger RNA. *Journal of genetics and genomics = Yi chuan xue bao* **41,** 21–33 (2014).

30. Zhou, Y., Goodenbour, J. M., Godley, L. A., Wickrema, A. & Pan, T. High levels of tRNA abundance and alteration of tRNA charging by bortezomib in multiple myeloma. *Biochem. Biophys. Res. Commun.* **385,** 160–164 (2009).

31. Zheng, G. *et al.* Efficient and quantitative high-throughput tRNA sequencing. *Nat Meth* **12,** 835–837 (2015).

32. Loher, P., Telonis, A. G. & Rigoutsos, I. MINTmap: fast and exhaustive profiling of nuclear and mitochondrial tRNA fragments from short RNA-seq data. *Sci Rep* **7,** (2017).

33. Anderson, P. & Ivanov, P. tRNA fragments in human health and disease. *FEBS Lett.* **588,** 4297–4304 (2014).

34. Ivanov, P., Emara, M. M., Villen, J., Gygi, S. P. & Anderson, P. Angiogenin-induced tRNA fragments inhibit translation initiation. *Molecular cell* **43,** 613–23 (2011).

35. Cole, C. *et al.* Filtering of deep sequencing data reveals the existence of abundant Dicer-dependent small RNAs derived from tRNAs. *RNA* **15,** 2147–2160 (2009).

36. Lee, Y. S., Shibata, Y., Malhotra, A. & Dutta, A. A novel class of small RNAs: tRNA-derived RNA fragments (tRFs). *Genes Dev.* **23,** 2639–2649 (2009).

37. Dhahbi, J. M. *et al.* 5’ tRNA halves are present as abundant complexes in serum, concentrated in blood cells, and modulated by aging and calorie restriction. *BMC genomics* **14,** 298 (2013).

38. Honda, S. *et al.* Sex hormone-dependent tRNA halves enhance cell proliferation in breast and prostate cancers. *PNAS* **112,** E3816–E3825 (2015).

39. The CCA-adding enzyme: A central scrutinizer in tRNA quality control. - PubMed - NCBI. Available at: https://www.ncbi.nlm.nih.gov/pubmed/26172425. (Accessed: 23rd July 2018)

40. Dard-Dascot, C. *et al.* Systematic comparison of small RNA library preparation protocols for next-generation sequencing. *BMC Genomics* **19,** 118 (2018).

41. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome biology* **15,** 550 (2014).

42. Karolchik, D. The UCSC Genome Browser Database. *Nucleic Acids Research* **31,** 51–54 (2003).

43. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25,** 2078–2079 (2009).

44. Ougland, R. *et al.* AlkB restores the biological function of mRNA and tRNA inactivated by chemical methylation. *Molecular cell* **16,** 107–16 (2004).

45. Anderson, J. *et al.* The essential Gcd10p-Gcd14p nuclear complex is required for 1-methyladenosine modification and maturation of initiator methionyl-tRNA. *Genes Dev.* **12,** 3650–3662 (1998).

46. Nishikura, K. & De Robertis, E. M. RNA processing in microinjected Xenopus oocytes. Sequential addition of base modifications in the spliced transfer RNA. *J. Mol. Biol.* **145,** 405–420 (1981).

47. Huang, X. *et al.* Characterization of human plasma-derived exosomal RNAs by deep sequencing. *BMC Genomics* **14,** 319 (2013).

48. Hannafon, B. N. *et al.* Exosome-mediated microRNA signaling from breast cancer cells is altered by the anti-angiogenesis agent docosahexaenoic acid (DHA). *Molecular Cancer* **14,** 133 (2015).

49. Noren Hooten, N. *et al.* Age-related changes in microRNA levels in serum. *Aging (Albany NY)* **5,** 725–740 (2013).

50. Mohr, S. *et al.* Thermostable group II intron reverse transcriptase fusion proteins and their use in cDNA synthesis and next-generation RNA sequencing. *RNA* **19,** 958–970 (2013).

51. Chan, P. P. & Lowe, T. M. GtRNAdb 2.0: an expanded database of transfer RNA genes identified in complete and draft genomes. *Nucl. Acids Res.* **44,** D184–D189 (2016).

52. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nature methods* **9,** 357–9 (2012).

53. Telonis, A. G. *et al.* Dissecting tRNA-derived fragment complexities using personalized transcriptomes reveals novel fragment classes and unexpected dependencies. *Oncotarget* **6,** 24797–24822 (2015).

54. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17,** 10–12 (2011).

55. Thorvaldsdóttir, H., Robinson, J. T. & Mesirov, J. P. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief Bioinform* **14,** 178–192 (2013).

56. Fett, J. W. *et al.* Isolation and characterization of angiogenin, an angiogenic protein from human carcinoma cells. *Biochemistry* **24,** 5480–5486 (1985).

57. Roadmap Epigenomics Consortium *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518,** 317–330 (2015).

58. Ernst, J. & Kellis, M. ChromHMM: automating chromatin state discovery and characterization. *Nat Methods* **9,** 215–216 (2012).

59. Moreno-Hagelsieb, G. & Latimer, K. Choosing BLAST options for better detection of orthologs as reciprocal best hits. *Bioinformatics* **24,** 319–324 (2008).

60. Bermudez-Santana, C. *et al.* Genomic organization of eukaryotic tRNAs. *BMC Genomics* **11,** 270 (2010).

61. Raab, J. R. *et al.* Human tRNA genes function as chromatin insulators. *EMBO J.* **31,** 330–350 (2012).

62. Simms, T. A. *et al.* TFIIIC binding sites function as both heterochromatin barriers and chromatin insulators in Saccharomyces cerevisiae. *Eukaryotic Cell* **7,** 2078–2086 (2008).

63. Rivas, E. & Eddy, S. R. Noncoding RNA gene detection using comparative sequence analysis. *BMC bioinformatics* **2,** 8 (2001).

64. Thompson, M., Haeusler, R. A., Good, P. D. & Engelke, D. R. Nucleolar clustering of dispersed tRNA genes. *Science* **302,** 1399–1401 (2003).

65. Moore, L. D., Le, T. & Fan, G. DNA methylation and its basic function. *Neuropsychopharmacology* **38,** 23–38 (2013).

66. Smith, Z. D. & Meissner, A. DNA methylation: roles in mammalian development. *Nat. Rev. Genet.* **14,** 204–220 (2013).

67. Lienert, F. *et al.* Identification of genetic elements that autonomously determine DNA methylation states. *Nat. Genet.* **43,** 1091–1097 (2011).

68. Gardiner-Garden, M. & Frommer, M. CpG Islands in vertebrate genomes. *Journal of Molecular Biology* **196,** 261–282 (1987).

69. Kutter, C. *et al.* Pol III binding in six mammals shows conservation among amino acid isotypes despite divergence among tRNA genes. *Nat Genet* **43,** 948–955 (2011).

70. Van Bortle, K., Phanstiel, D. H. & Snyder, M. P. Topological organization and dynamic regulation of human tRNA genes during macrophage differentiation. *Genome Biol* **18,** (2017).

71. Talwar, S., Pocklington, M. J. & Maclean, N. The methylation pattern of tRNA genes in Xenopus laevis. *Nucleic Acids Res.* **12,** 2509–2517 (1984).

72. Shen, Y. *et al.* A map of the cis-regulatory sequences in the mouse genome. *Nature* **488,** 116–120 (2012).

73. Bogu, G. K. *et al.* Chromatin and RNA Maps Reveal Regulatory Long Noncoding RNAs in Mouse. *Mol. Cell. Biol.* **36,** 809–819 (2016).

74. Rosenbloom, K. R. *et al.* The UCSC Genome Browser database: 2015 update. *Nucleic Acids Res.* **43,** D670-681 (2015).

75. Regulation of tRNA synthesis by the general transcription factors of RNA polymerase III - TFIIIB and TFIIIC, and by the MAF1 protein - ScienceDirect. Available at: https://www.sciencedirect.com/science/article/pii/S1874939917303644?via%3Dihub. (Accessed: 11th July 2018)

76. Dennis, P. P. & Nomura, M. Stringent control of ribosomal protein gene expression in Escherichia coli. *Proceedings of the National Academy of Sciences of the United States of America* **71,** 3819–23 (1974).

77. Pérez-Ortín, J. E., Alepuz, P., Chávez, S. & Choder, M. Eukaryotic mRNA decay: methodologies, pathways, and links to other stages of gene expression. *Journal of molecular biology* **425,** 3750–75 (2013).

78. Malygin, A. A., Parakhnevitch, N. M., Ivanov, A. V., Eperon, I. C. & Karpova, G. G. Human ribosomal protein S13 regulates expression of its own gene at the splicing step by a feedback mechanism. *Nucleic acids research* **35,** 6414–23 (2007).

79. Ivanov, A. V., Malygin, A. A. & Karpova, G. G. Human ribosomal protein S26 suppresses the splicing of its pre-mRNA. *Biochimica et biophysica acta* **1727,** 134–40 (2005).

80. Baughman, G. & Nomura, M. Translational regulation of the L11 ribosomal protein operon of Escherichia coli: analysis of the mRNA target site using oligonucleotide-directed mutagenesis. *Proceedings of the National Academy of Sciences of the United States of America* **81,** 5389–93 (1984).

81. Johnsen, M., Christensen, T., Dennis, P. P. & Fiil, N. P. Autogenous control: ribosomal protein L10-L12 complex binds to the leader sequence of its mRNA. *The EMBO journal* **1,** 999–1004 (1982).

82. Philippe, C. *et al.* Ribosomal protein S15 from Escherichia coli modulates its own translation by trapping the ribosome on the mRNA initiation loading site. *Proceedings of the National Academy of Sciences of the United States of America* **90,** 4394–8 (1993).

83. Cerretti, D. P., Mattheakis, L. C., Kearney, K. R., Vu, L. & Nomura, M. Translational regulation of the spc operon in Escherichia coli. Identification and structural analysis of the target site for S8 repressor protein. *Journal of molecular biology* **204,** 309–29 (1988).

84. Berkhout, B., Kastelein, R. A. & van Duin, J. Translational interference at overlapping reading frames in prokaryotic messenger RNA. *Gene* **37,** 171–9 (1985).

85. Moine, H. *et al.* The translational regulation of threonyl-tRNA synthetase. Functional relationship between the enzyme, the cognate tRNA and the ribosome. *Biochimica et biophysica acta* **1050,** 343–50 (1990).

86. Fayat, G. *et al.* Escherichia coli phenylalanyl-tRNA synthetase operon region. Evidence for an attenuation mechanism. Identification of the gene for the ribosomal protein L20. *Journal of molecular biology* **171,** 239–61 (1983).

87. Putney, S. D., Meléndez, D. L. & Schimmel, P. R. Cloning, partial sequencing, and in vitro transcription of the gene for alanine tRNA synthetase. *The Journal of biological chemistry* **256,** 205–11 (1981).

88. Romby, P. & Springer, M. Bacterial translational control at atomic resolution. *Trends in Genetics* **19,** 155–161 (2003).

89. Henkin, T. M. tRNA-directed transcription antitermination. *Mol. Microbiol.* **13,** 381–387 (1994).

90. Glisovic, T., Bachorik, J. L., Yong, J. & Dreyfuss, G. RNA-binding proteins and post-transcriptional gene regulation. *FEBS Lett.* **582,** 1977–1986 (2008).

91. Serganov, A. & Nudler, E. A decade of riboswitches. *Cell* **152,** 17–24 (2013).

92. Mayer, C., Köhrer, C., Gröbner, P. & Piendl, W. MvaL1 autoregulates the synthesis of the three ribosomal proteins encoded on the MvaL1 operon of the archaeon Methanococcus vannielii by inhibiting its own translation before or at the formation of the first peptide bond. *Molecular microbiology* **27,** 455–68 (1998).

93. Kanai, A., Sato, A., Imoto, J. & Tomita, M. Archaeal Pyrococcus furiosus thymidylate synthase 1 is an RNA-binding protein. *The Biochemical journal* **393,** 373–9 (2006).

94. Donovan, J. & Copeland, P. R. Threading the needle: getting selenocysteine into proteins. *Antioxid. Redox Signal.* **12,** 881–892 (2010).

95. Tujebajeva, R. M. *et al.* Decoding apparatus for eukaryotic selenocysteine insertion. *EMBO Rep.* **1,** 158–163 (2000).

96. Fischer, N. *et al.* Towards understanding selenocysteine incorporation into bacterial proteins. *Biol. Chem.* **388,** 1061–1067 (2007).

97. Backofen, R. & Hess, W. Computational prediction of sRNAs and their targets in bacteria. *RNA Biol* **7,** 33–42 (2010).

98. Zuker, M. & Stiegler, P. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Res.* **9,** 133–148 (1981).

99. Gruber, A. R., Findei\s s, S., Washietl, S., Hofacker, I. L. & Stadler, P. F. RNAz 2.0: improved noncoding RNA detection. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing* 69–79 (2010).

100. McGinn, J. & Czech, B. Small RNA library construction for high-throughput sequencing. *Methods Mol. Biol.* **1093,** 195–208 (2014).

101. Hafner, M. *et al.* RNA-ligase-dependent biases in miRNA representation in deep-sequenced small RNA cDNA libraries. *RNA (New York, N.Y.)* **17,** 1697–712 (2011).

102. Nevskaya, N. *et al.* Ribosomal protein L1 recognizes the same specific structural motif in its target sites on the autoregulatory mRNA and 23S rRNA. *Nucleic acids research* **33,** 478–85 (2005).

103. Zengel, J. M. & Lindahl, L. Diverse mechanisms for regulating ribosomal protein synthesis in Escherichia coli. *Prog. Nucleic Acid Res. Mol. Biol.* **47,** 331–370 (1994).

104. Anandam, P., Torarinsson, E. & Ruzzo, W. L. Multiperm: shuffling multiple sequence alignments while approximately preserving dinucleotide frequencies. *Bioinformatics (Oxford, England)* **25,** 668–9 (2009).

105. Bernick, D. L., Dennis, P. P., Lui, L. M. & Lowe, T. M. Diversity of Antisense and Other Non-Coding RNAs in Archaea Revealed by Comparative Small RNA Sequencing in Four Pyrobaculum Species. *Frontiers in microbiology* **3,** 231 (2012).

106. Kalvari, I. *et al.* Rfam 13.0: shifting to a genome-centric resource for non-coding RNA families. *Nucleic Acids Res* **46,** D335–D342 (2018).

107. Wilting, R., Schorling, S., Persson, B. C. & Böck, A. Selenoprotein synthesis in archaea: identification of an mRNA element of Methanococcus jannaschii probably directing selenocysteine insertion. *Journal of molecular biology* **266,** 637–41 (1997).

108. Gillen, K. L. & Hughes, K. T. Transcription from two promoters and autoregulation contribute to the control of expression of the Salmonella typhimurium flagellar regulatory gene flgM. *J Bacteriol* **175,** 7006–7015 (1993).

109. Chatr-Aryamontri, A. *et al.* Nonsense-mediated and nonstop decay of ribosomal protein S19 mRNA in Diamond-Blackfan anemia. *Hum. Mutat.* **24,** 526–533 (2004).

110. Yao, Z., Weinberg, Z. & Ruzzo, W. L. CMfinder—a covariance model based RNA motif finding algorithm. *Bioinformatics* **22,** 445–452 (2006).

111. Lui, L. & Lowe, T. Small nucleolar RNAs and RNA-guided post-transcriptional modification. *Essays Biochem.* **54,** 53–77 (2013).

112. Ruzzo, W. L. & Gorodkin, J. De novo discovery of structured ncRNA motifs in genomic sequences. *Methods Mol. Biol.* **1097,** 303–318 (2014).

113. Wheeler, E. C., Van Nostrand, E. L. & Yeo, G. W. Advances and challenges in the detection of transcriptome-wide protein-RNA interactions. *Wiley Interdiscip Rev RNA* **9,** (2018).