CHARACTERIZING HUMAN TRANSFER RNAS BY HYDRO-TRNASEQ AND PAR-CLIP

A Thesis Presented to the Faculty of

The Rockefeller Uniersity

in Partial Fullfillment of the Requirements for

the degree of Doctor of Philosophy

by

Tasos Gogakos

June 2017

Abstract

CHARACTERIZING HUMAN TRANSFER RNAS BY HYDRO-TRNASEQ AND PAR-CLIP

Tasos Gogakos, Ph.D.

The Rockefeller University 2017

The participation of transfer RNAs (tRNAs) in test2 (tEsT) fundamental aspects of biology and disease necessitates an accurate, experimentally confirmed annotation of tRNA genes, and curation of precursor and mature tRNA sequences. This has been challenging, mainly because RNA secondary structure and nucleotide modifications, together with tRNA gene multiplicity, complicate sequencing and sequencing read mapping efforts. To address these issues, we developed hydro-tRNAseq, a method based on partial alkaline RNA hydrolysis that generates fragments amenable for sequencing. To identify transcribed tRNA genes, we further complemented this approach with Photoactivatable Crosslinking and Immunoprecipitation (PAR-CLIP) of SSB/La, a conserved protein involved in pre-tRNA processing. Our results show that approximately half of all predicted tRNA genes are transcribed in human cells. We also report predominant nucleotide modification sites, their order of introduction, and identify tRNA leader, trailer and intron sequences. By using complementary sequencing-based methodologies we present a human tRNA atlas, and determine expression levels of mature and processing intermediates of tRNAs in human cells.

Acknowledgments

First, I would like to thank my

contents

Table of Contents

List of Figures

List of Tables

Glossary

List of Abbreviations

ncRNA noncoding RNA.

tEsT test2.

tRNA transfer RNA.

Chapter 1

Introduction

1.1 Overview

Transfer RNAs (tRNAs) are essential factors for the expression of genetic information, serving as the adaptor molecules that decode the genetic code during protein synthesis cite Crick tie club letter, and are among the earliest studied noncoding RNA (ncRNA) non-coding RNA molecules [1, 2]. Despite their highly conserved participation in the translational machinery, there is growing evidence that they play roles in other cellular processes, including non-coding RNA-mediated gene silencing and responses to cellular stress. The biological importance of tRNAs and their associated proteins is underscored by the pathologic conditions that are related to aberrations in their expression and function or The biological significance of tRNAs and their protein interactions is underscored by the number of human diseases caused by mutations in tRNAs and tRBPs 8,11-15 from TRP.

Yet, in recent years tRNAs received new attention in the context of codon-resolved translational control [3–8], and due to the involvement of their metabolic byproducts in regulation and cross-talk with processing and effector functions of

other classes of non-coding RNAs (ncRNAs) [9–11]. Nevertheless, the lack of reliable methods for tRNA quantification has hampered such analyses, and necessitated the use of predicted tRNA gene copy number as a surrogate index of expression [7, 12, 13]. This hinged on the assumption that predicted tRNA gene loci are all expressed constitutively and equally, even though there has been experimental evidence against it [14]. Similarly, experimental tRNA gene annotation in the past had to focus on RNA polymerase III (POLR3) ChIP-seq [15] [16] [17] or hybridization-based approaches [18] [19]. The former, however, were impeded by their restricted genomic resolution and the assumption that POLR3 binding always leads to productive tRNA expression followed by complete processing, while the latter fell short of providing absolute counts and did not address the discovery of new transcripts and genes, assuming also normal hybridization rules for modified nucleosides.

An improvement in tRNA quantification has arisen from recent efforts that employed modification-reverting enzymes prior to sequencing, in order to minimize stalling of reverse transcriptase at modified sites [20] [21]. However, an extensive annotation of human genes and transcripts was foregone because the focus was either on mature tRNAs only [21] or on tRNA fragments not inclusive of full-length precursor tRNA (pre-tRNA) transcripts [20]. Thus, to-date an experimentally validated list of curated mature and pre-tRNA sequences and annotating tRNA genes in human is still missing.

We have combined complementary high-throughput techniques for obtaining the sequence composition and abundance of tRNAs in human embryonic kidney cells (HEK293). We developed hydro-tRNAseq, a modified small RNA sequencing protocol based on partial alkaline hydrolysis of input RNA, in order to identify and quantify tRNAs, and provided evidence for the validity of this approach when deter-

mining the accumulation of disease-associated tRNA intron fragments caused by mutations in the tRNA splicing machinery [22]. Here we extend this approach by applying it to tRNA-enriched size fragments with the aim to annotate and curate all tRNAs. Since tRNA processing, such as precursor trimming and intron removal, is a fast process[23], we also aimed to enrich specifically for pre-tRNAs in order to identify and annotate the corresponding unique tRNA gene template. Thus, we performed PAR-CLIP on SSB, a conserved and ubiquitous protein involved in 3 tRNA processing [24] [25] [26].

1.1.1 tRNA biogenesis

tRNA genes are transcribed by RNA polymerase III (POLR3) that uses promoters internal to the DNA sequence of the tRNA gene (tDNA). The primary transcript is a precursor tRNAs (pre-tRNA) with a 5' triphosphate. In humans, a minority of tRNA transcripts (see section XXX) harbor introns. A dedicated tRNA splicing complex composed fo core and accessory proteins carries out tRNA splicing cite references. Pre-tRNAs comprise the mature tRNA sequence, and 5' leader and 3' trailer extensions, which are trimmed in a coordinated manner by endonucleases and other processing factors. The ribonucleoprotein (RNP) complex RNase P removes the 5' leaders, leaving a 5' monophospahte, and ELAC2, the human homolog of tRNase Z trims the 3' trailer, leaving a 3' hydroxyl (OH). Next, the universally conserved 3' terminal CCA tail is added by TRNT1, the tRNA nucleotidyll transferase 1 (TRNT1), and acts as the acceptor of the amino acid. tRNAs are further modified by chemical nucleotide modifications (see section XXX), exported from the nucleus to the cytoplasm where they can undergo further modifcations, are aminoacylated with their cognate amino acid by aminoacyl tRNA synthetases, and are finally presented to the ribosome by translation factors to participate in

protein synthesis Fig. 1.1. cite 16,17 from TRP

Although these processes allow for multiple levels of regulation, variation in tRNA expression across tissues or between normal and pathologic conditions has not been studied extensively, mainly for two reasons. First, until recently there was the assumption that their essentiality obviated a need for any specialized transcriptional or post-transcriptional control. Second, the lack of an extensively curated and experimentally validated tRNA profile prevented quantitative and systematic studies. Nevertheless, it is now clear that the expression of tRNAs can be dynamic and can indeed exhibit tissue specificity1 18 from TRP. Importantly, abnormal tRNA expression levels have been correlated and causally associated with pathologic conditions, such as cancer 14 from TRP.

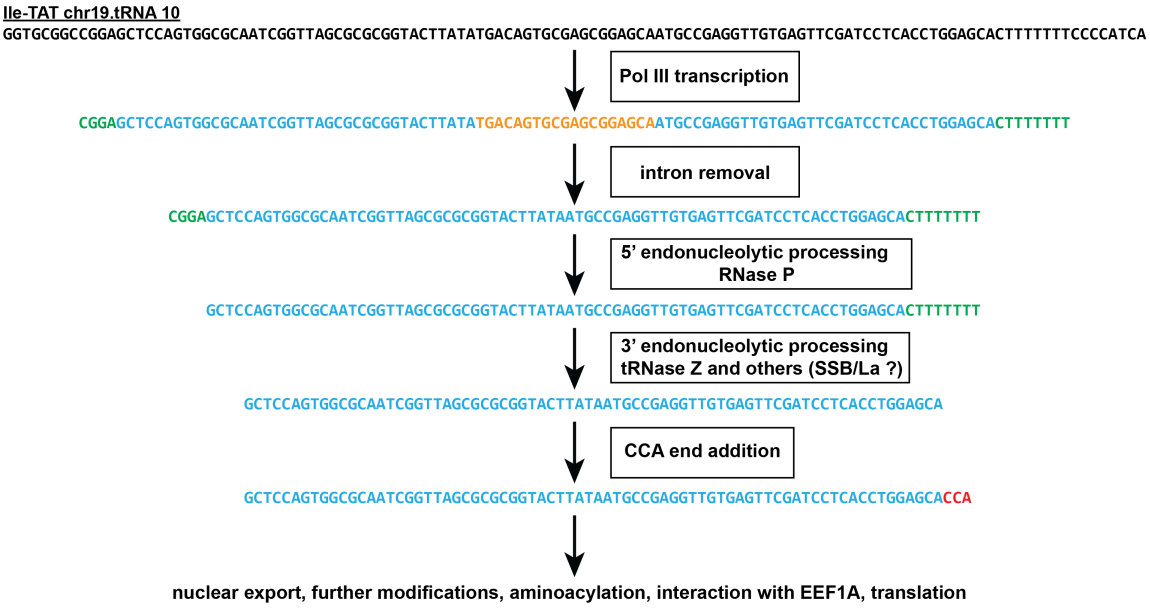

Figure 1.1: Overview of tRNA biogenesis and processing. thththt

1.1.2 tRNA sequencing

Evidently the biogenesis pathway of tRNAs is quite a complex one. Already some of the possible problems associated with tRNA annotation become apparent. Issues that complicate tRNA sequencing and analysis pertain to both experimental and bioinformatic problems: A) Experimental:

i) stable 2o and 3o structures. The highly structured tRNA impede ligation steps employed in traditional protocols of small RNA sequencing.

ii) extensive post-transcriptional processing. The extensive chemical modification of nucleosides causes stalls, blocks or errors during reverse-transcription RT steps

iii) aminoacylation of the 3' end of tRNAs. The 3' aminoacyl-tRNA bind prevents ligation of adapters at the 3' end of tRNAs

Obtaining an RNA-Seq based atlas of human tRNAs is hindered by multiple obstacles. First, sequencing of tRNAs is technically arduous due to their relatively small size, and their stable structure that proves to be a challenge for enzymes used in cDNA library preparations. Second, numerous (¿100) tRNA pseudogenes are interspersed in the human genome19. Third, tRNAs undergo extensive post-transcriptional processing, which involves the removal of the 5 leader and 3 trailer sequences of the primary transcript, removal of tRNA introns, addition of the universally conserved 3 CCA end, and addition of a 5 guanosine to all histidine tRNAs17. Fourth, tRNAs are subjected to extensive chemical modifications on numerous nucleosides, which are likely to lead to mismatches upon the reverse transcription step of the RNA cloning protocols20,21. Some modifications are universally conserved and required for proper tRNA function (e.g. adenosine to inosine deamination at the wobble position of the anticodon and methylation of

adenosine in the TpsiC loop)20,22. Since alignment algorithms cannot tolerate multiple mismatches, it is likely that significant numbers of tRNA reads are excluded even if non-default mapping parameters are used. Fifth, tRNA isoacceptors (tRNA molecules that decode synonymous codons) share a large degree of sequence similarity that makes the distinction between alternative isoacceptors and editing products equivocal. Finally, eukaryotic cells harbor two distinct populations of tRNAs, nuclear and mitochondrial, whose length, structure, genomic organization, and processing differ considerably, and thus call for customized annotation procedures. Owing to all these hurdles, the normal genetic makeup and variation of the tRNA population in human cells has not been probed with RNA-Seq tools. Instead information about tRNA sequences and genes comes from bioinformatic predictions19,23. Such approaches take into account base-pair covariance, secondary structure predictions of the classical cloverleaf fold of tRNAs, and the tRNA promoter and termination architecture, and scan the human genome in order to identify sequences that are likely to obtain the typical tRNA structure. These analyses have resulted in the most comprehensive standard for whole-genome, predictive annotation of tRNAs so far, and the sequences they have predicted have been used extensively as bona fide tRNAs23.

Thus, it may come as no surprise that obtaining an accurate annotation of tRNA genes and curation of tRNA transcripts is challenging. We wanted to obtain an RNA-seq validated list of human nuclear and mitochondrial tRNA gene, and their processing intermediates This was my goal. To design a method for seequencing and a

1.1.3 Previous efforts for genome-wide tRNA annotation

To date, no direct and rigorous experimental validation of tRNA sequences has been carried out. Instead, experimental evidence for tRNA expression has been indirect, through: a) chromatin immunoprecipitation and sequencing (ChIP-Seq) studies focusing on the occupancy of genomic locations by POLR3 and/or its transcription factors and b) tRNA microarrays that use the predicted tRNA sequences as the reference for the creation of array probes42-45. These methods, though, have several limitations. ChIP-Seq, for example, uses chromatin occupancy as a proxy for productive RNA synthesis. Conversely, tRNA microarrays have limited sensitivity and specificity thresholds due to off-target hybridization that is potentiated by nucleoside modifications21, while their dynamic range is considerably narrower than RNA-Seq. Finally, neither method is appropriately equipped to determine definitively precursor tRNAs (pre-tRNAs) or their transcription start and termination sites. This is an important limitation, as pre-tRNA fragments have been associated with neurodegenerative diseases46. To address the lack of a global and unbiased analysis of the human tRNA profile, I will develop an experimental and computational methodology for the generation of a reference tRNA atlas. To overcome existing experimental challenges, I will use a customized RNA-Seq technique (HydroRNAseq). To efficiently analyze the sequencing data in silico, I will develop a systematic and iterative bioinformatics platform.

1.1.4 Small RNA sequencing protocol

First, I applied the protocol that the Tuschl lab had previously developed for sequencing small RNAs [27] (**Fig. 1.2**). The experimental procedure resulting in small RNA cDNA library preparation begins with the ligation of barcoded 3'

oligonucleotide adapters, pooling of several multiplexed samples, ligation of a 5' adapter, reverse transcription and PCR amplification, followed by high-throughput Illumina sequencing. The different sequences for the 3' and 5' adapters preserves the strandedness of the original RNA sequence, enhancing ncRNA discovery and curation.

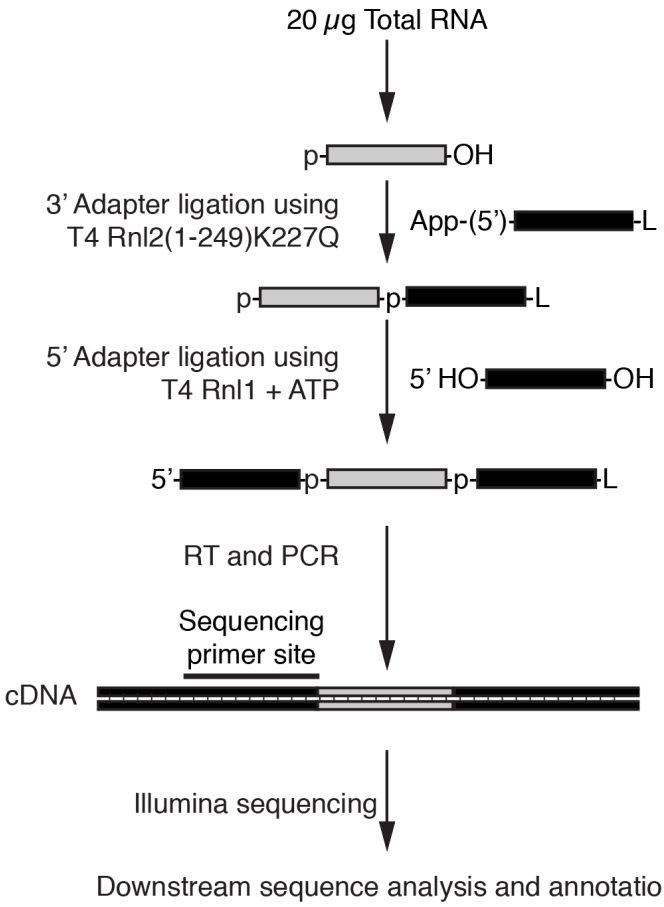

Figure 1.2: Small RNA sequencing protocol. thththtfahflahflahdflahdflahdflad-fladhflasdhfladhfladhfladhfladhfaldjsfhaldskfjaldsfhalsdfhasdf

The utility of this protocol is documented for the discovery and study of miRNAs. Indeed, the decision to employ this protocol for sequencing of tRNAs is reasonable because:

- tRNAs, which are on average 75 nucleotides nts long are closer in length

than most other ncRNAs (typically longer than 100 nts).

- mature tRNAs and miRNAs have a monophosphate at their 5' ends, which acts as the nucleophilic attacking group in the 5' ligation step.

The application of this protocol for tRNA sequencing, though, resulting in RNAseq datasets with only ~2% tRNA content, with an average length of 59 nts (**Fig. 1.1**). These suggested that tRNAs were refractory to the small RNA sequencing protocol, and necessitated the development of a novel sequencing protocol.

RNA type	% Total reads	Mean length (nt)
rRNA	35.8%	60.5
no match	24.1%	76.2
no annotation	17.8%	64.2
sn/snoRNA	15.1%	62.5
repeat	3.8%	59.1
tRNA	2.0%	59.1
miscRNA	1.3%	63.1
miRNA	0.1%	22.2

Table 1.1: RNA category from small RNA sequencing protocol

Chapter 2

Results

2.1 Hydrolysis-based tRNA sequencing

In order to overcome the problems associated with tRNA sequencing, we tried to identify the minimal number of simplest steps that could tackle the maximal number of problems. Thus, to curate and quantify human tRNAs, we isolated 60-100 nt-sized total RNA from HEK293 cells comprising both precursor and mature tRNAs, but being devoid of most other abundant RNAs and short tRNA turnover products [11]. Full-length tRNAs have thermodynamically stable secondary and tertiary structures and are heavily modified by RNA editing, all of which compromise reverse transcription (RT) and RNAseq analysis. To overcome these problems, we implemented a limited alkaline hydrolysis step, which generates shorter RNA fragments with less structure and fewer modifications per fragment, and consequently more amenable to small RNA cDNA library preparation and deep sequencing. Furthermore, basic conditions also cleave the aminoacyl-tRNA bond, freeing the 3 terminal hydroxyl group required for 3 adapter ligation during RNA cDNA library preparation. This approach increased the tRNA read

content to ¿40% in our deepest dataset Table S1. We named this procedure hydro-tRNAseq (**Fig. 2.1**). In summary, partial hydrolysis of tRNAs overcame technical limitations of suboptimal adapter ligation and RT and albeit this resulted in shorter reads, it also resulted in fewer errors at sites of modification per sequenced read, which ultimately improved the performance of the mapping algorithm.

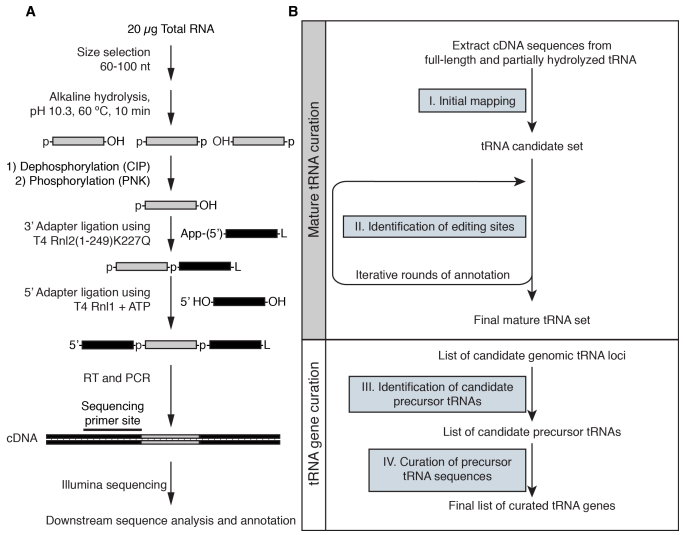

Figure 2.1: Experimental and bioinformatic pipeline for tRNA annotation and reference transcript curation by hydro-tRNAseq.(A) tRNAs and pre-tRNAs were size-selected from HEK293 total RNA and subjected to limited alkaline hydrolysis, followed by dephosphorylation, rephosphorylation and conventional small RNA sequencing as described previously (Hafner et al., 2012). (B) An iterative mapping and annotation protocol was used to first annotate and curate fully processed and nucleotide-modified mature tRNAs. Leftover reads that spanned the mature-precursor junctions were used to identify transcribed tRNA genes.

2.2 Hierarchical sequence read mapping

In parallel with the tRNA annotation procedure, we had to build a bioinfroamtic pipeline for processing the obtained sequence information. We developed an iterative, hierarchical approach for mapping and annotating our sequence reads.

We mapped the reads to reference tRNA genes (hg19, http://gtrnadb.ucsc.edu/) using an iterative and hierarchical protocol (Fig. 1B). We started by mapping only to mature tRNAs, which included the 3 CCA aminoacyl acceptor terminus introduced posttranscriptionally by tRNA nucleotidyl transferase, and the G-1 nucleotide added posttranscriptionally to histidine tRNAs (Gu, 2003; Juhling et al., 2009), but excluded tRNA introns. Starting with two most abundant tRNA transcripts per isotype (tRNAs encoding the same amino acid) as indicated after the first mapping round, except for selenocysteine, where only one mature tRNA sequence could be identified, we performed iterative rounds of mapping and manual reference transcript selection, focusing in every step on transcripts that collected more reads with an error distance of 1-2 than 0. If these reads with mismatches could be assigned to other tRNA isoacceptors (tRNA accepting the same amino acid), these were included in our candidate reference set. Otherwise, we reasoned that the mismatches were the results of nucleotide-modification-induced errors of RT. In those cases, we accounted for the modified nucleoside signatures by introducing a new, edited reference transcript in our set. For tRNAs that exhibited multiple positions with high modification rates (¿10% compared to reference), we compiled reference sequences with all possible combinations of modified signatures at all detectably modified positions, aiming to account for the maximum possible number of mapped sequence reads. We ended the curation cycles when there was no observed modified position that exhibited a mismatch frequency greater than or equal to 10% compared to the reference. By performing this iterative process of curation, we obtained an experimentally validated reference set of mature tRNAs accounting for modified-nucleotide-induced sequence variation upon reverse transcription (Table S2).

Now, I can automatically and fast obtain relative read counts for all classes of

ncRNAs in one experiment. Applying this pipeline to the tRNA-seq experimental resutls shows that indeed the majority of reads map to tRNAs. The depth of the library is extensive and thus allows for a reliable annotation of tRNA genes and transcripts.

2.3 Iterative manual tRNA curation

In order to identify possible tRNA gene loci, we mapped the curated tRNA sequences back to the genome, allowing for gaps to accommodate tRNA introns, as well as up to 7 mismatches to accommodate terminal and internal RNA editing events. By appending 40 nt 5 and 3 of the location of genomic mapping, we obtained a candidate pre-tRNA gene set. We mapped non-annotated residual reads to these candidates to identify 5 leader- and 3 trailer-comprising pre-tRNA reads, which also distinguished actively transcribed tRNA genes from silent ones or pseudogenes. Leader- and trailer-comprising tRNA genes show higher sequence variation, as evidence by higher information entropy values, across the leader and trailer nucleotides than internal sequences within the mature tRNA suggesting that even short precursor sequences with read coverage are sufficient for the annotation of non-redundant tRNA genes (Fig. S1). At the end of our analysis we accounted for 93% of the 114,367,140 reads in our deepest library (Table S1). Given the depth of sequencing, we are confident that we accounted for the vast majority of precursor and mature tRNAs. Indeed, a posteriori we looked for genomic regions that collected at least 50 overlapping reads throughout their whole length, fell within the 60- to 100-nt size window, and adopted a cloverleaf structure, in an effort to detect any tRNAs that might have been overlooked by our approach or in prior literature. The only sequences that we identified were U1

snRNA (pseudo)genes (Fig. S2), suggesting that our analysis was exhaustive, at least for tRNAs in HEK293 cells.

2.4 Protocol and pipeline outputs

2.4.1 Mature tRNA alignment

Representative alignment of reads to a mature tRNA reference., where the reference is shown at the bottom, the reads that map to it, and the depth of read abundance, as well as the reference locations at which its read maps.

Due to the intentional fragmentation of input RNA, we observe that the majority of reads were shorter than the full length tRNA, but there were enough long reads (like the one shown here in red) to bridge together separate segments of the obtained sequence .

Figure X. Alignmnent of tRNA-seq reads from HEK293 to mature tRNA transcript TRNAE5. Shorter reads are shown in black; a longer read, bridging the two halves of the tRNAs is shown in red. The frequency of each read (count) and the number of locations that it maps within the tRNA reference with no mismatches are indicated. Vertical lines represent the relative frequency of binned, normalized read count in log4 increments.

(**Fig. 2.2**)

2.4.2 Pre-tRNA alignment

The reads that after a first pass were mapped to a mature tRNA were set aside, and the leftover reads were then mapped to tRNA genomic locations that were extended 40 nts up- and downstream from all mature tRNA boundaries. The reads

Figure 2.2: Mature tRNA alignmentl. ththhtfahflahflahdflahdflahdfladfladhflasd-hfladhfladhfladhfladhfaldjsfhaldskfjaldsfhalsdfhasdf

that mapped to such tRNA precursors (that I wil refer to as pre-tRNAs) were used to identify actively transcribed tRNA loci in our cell system of HEK293 cells (**Fig. 2.3**)

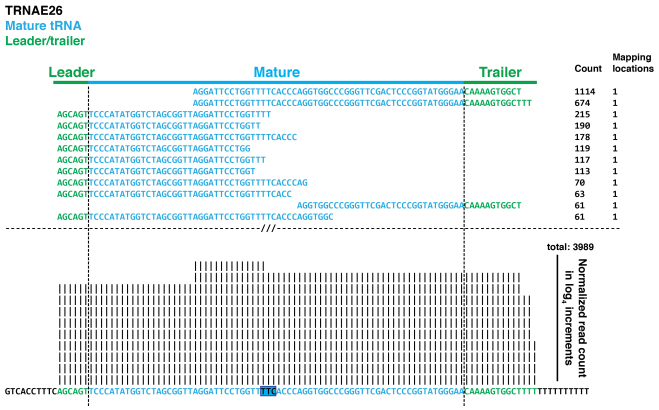

Figure 2.3: Pre-tRNA alignmentl. Placeholder

2.5 Justification for using precursor reads

Entropy (**Fig. 2.4**)

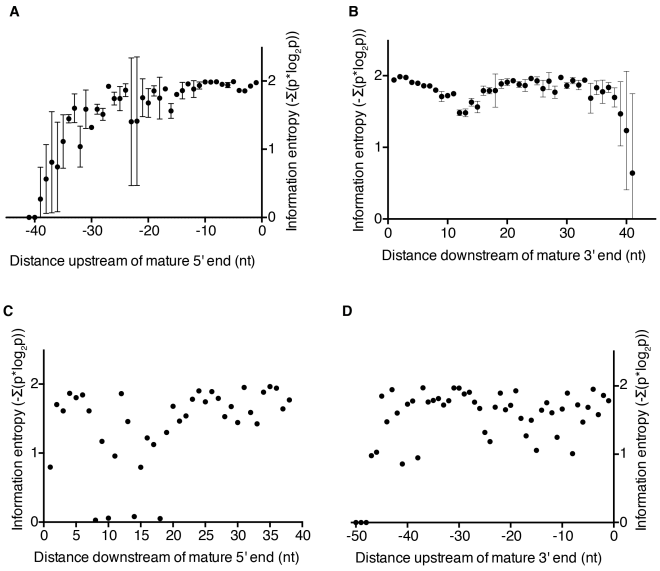

Figure 2.4: Information entropy in pre-tRNA segments and mature body (A,B) Information entropy $H = -\sum_{i=1}^{n} p(i) * log(p(i))$, (where p is the frequency of each nucleotide at a given position, i, and n the total number of transcripts) was calculated using read evidence from hydro-tRNAseq (four replicates) for the 5 leader and 3 trailers of all pre-tRNAs with positions centered at the 5 and 3 ends of mature tRNAs. (C,D) Same as before, but using the reference sequence of mature tRNAs.

2.6 Composition of hydro-tRNAseq libraries

The majority of our reads obtained from 60-100 nt size-fractionated total RNA were assigned to mature tRNAs. The improvement we observed in recovering tRNA reads was considerable, as 2/3 of our reads mapped to either mature or pre-tRNAs or mitochondrial tRNAs (**Fig. 2.5**).

2.7 Need for pre-tRNA enrichment

Even though the majority of our reads obtained from 60-100 nt size-fractionated total RNA were assigned to mature tRNAs, only 1% of the reads comprised sequences overlapping with pre-tRNA leader or trailer sequences (**Fig. 2.5**, Table

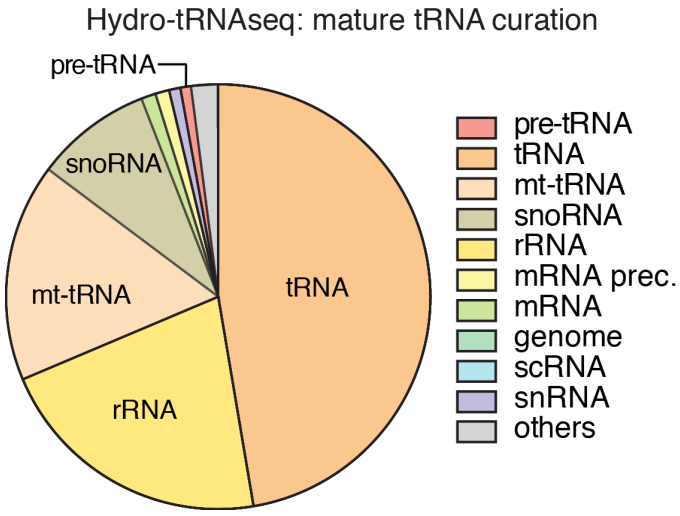

Figure 2.5: Composition of hydro-tRNAseq libraries. Total RNA composition of the 60-100 nt size fraction from hydro-tRNAseq according to RNA classes.

S1). This raised the possibility that we might have missed reads corresponding to lowly expressed or very rapidly processed pre-tRNAs.

We did this by performing PAR-CLIP (photoactivatable-ribonucleoside-enhanced crosslinking and immunoprecipitation), a technique developed in our lab to identify RNA targets of RNA binding proteins at nucleotide level resolution and with high specificity.

2.8 PAR-CLIP methodology for the study or RNA-RBP interactions

A series of techniques have been developed for the study of RNA-RBP interactions on a genomic scale 27 from TRP. Our lab developed Photoactivatable-Ribonucleoside-Enhanced Crosslinking and Immunoprecipitation (PAR-CLIP), coupled with deep sequencing, which is a cell-based approach that allows the determination of RBP binding sites on RNA targets at nucleotide-level resolution (**Fig. 2.7**28 from TRP. To enable efficient RNA-RBP crosslinking using long wavelength

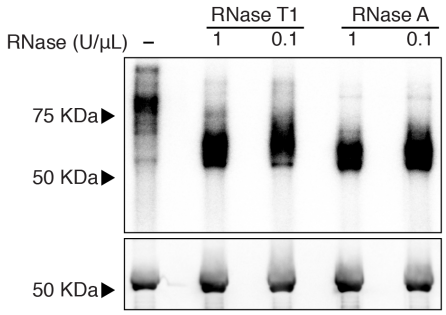

Figure 2.6: SSB crosslinking to RNA. Phosphorimage of SSB-crosslinked to radiolabeled RNA. PAR-CLIP was performed using RNase A or RNase T1, at two different concentrations to account for possible biases of RNase treatment conditions. Libraries from PAR-CLIP using 1 U/μL of RNase A and RNase T1 were prepared and submitted for sequencing. Western blot against HA, shown in the bottom, confirmed the immunoprecipitation of SSB.

UV, 4-thiouridine (4SU) is added to culture medium, taken up by cells and incorporated into nascent transcripts. The crosslinked ribonucleoprotein complex is submitted to partial RNase digestion, immunopurification and size-fractionated. Crosslinked RNA is recovered, converted into small RNA cDNA libraries, and sequenced. Importantly, crosslinking introduces a structural change in the thiouridine base, which allows pinpointing the position of crosslinking by scoring for characteristic T-to-C transitions in the sequenced cDNA. In addition, the abundant background derived from non-crosslinked fragments of co-purifying cellular RNAs do not contain these T-to-C transitions and can be filtered out. Thus, PAR-CLIP has a very low rate of false positive target identification, since the nucleotide transition signature reliably marks true crosslinking sites. PAR-CLIP has so far been applied successfully to the study of mRNA- and miRNA-binding proteins, but not

tRBPs.

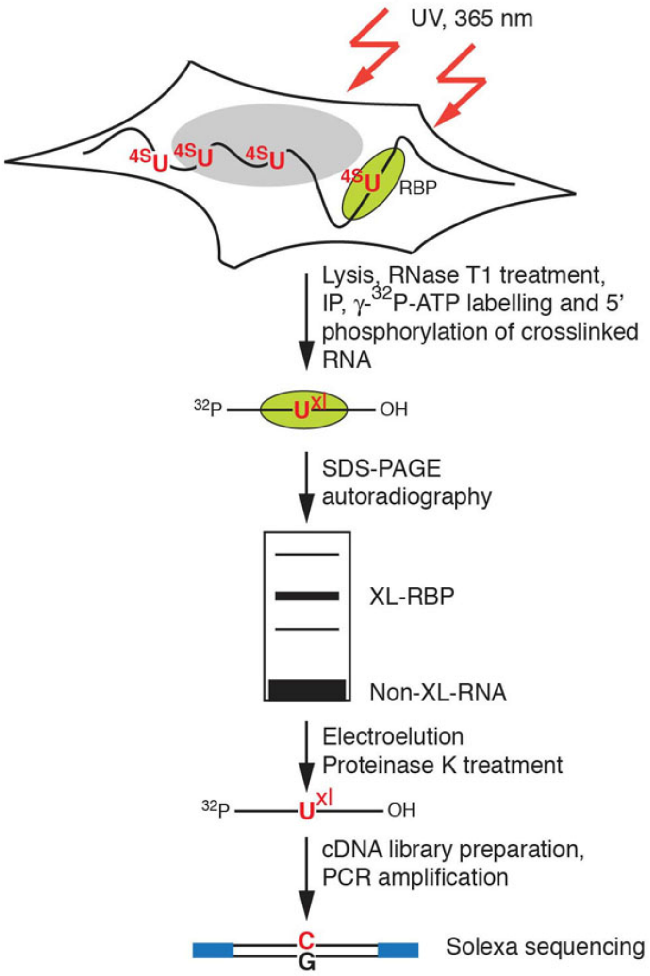

Figure 2.7: PAR-CLIP. Outline

2.9 SSB PAR-CLIP

Therefore, we decided to complement our efforts with PAR-CLIP-sequencing of SSB (**Fig. 2.6** All tRNA genes are transcribed by POLR3, which terminates upon decoding an oligo-uridine (oligoU) region [28]. SSB binds to the short pre-tRNA 3 oligoU tail [26] prior to removal of the entire 3 trailer sequence. Therefore, we reasoned that SSB should bind all tRNA precursors, and that if we could isolate

its targets, we would be able to reliably identify transcribed tRNA loci.

SSB exhibited a striking binding preference for pre-tRNAs and showed a drastic enrichment in precursor tRNAs compared to hydro-tRNAseq (**Fig. 2.8**), which confirmed our hypothesis, as well as previous observations [24]. We performed PAR-CLIP using two different nucleases to control for sequence biases at the nuclease digestion step. RNase T1 resulted in longer precursor tRNA trailer sequences than RNase A, due the latters preference for cleaving 3 to pyrimidines, which are highly abundant in the 3 trailer sequences. Overall, 46% of all PAR-CLIP reads mapped to pre-tRNAs (**Fig. 2.8**), the overwhelming majority of which showed the characteristic T-to-C transition, indicative of crosslinking (**Fig. 2.8**, table S3.

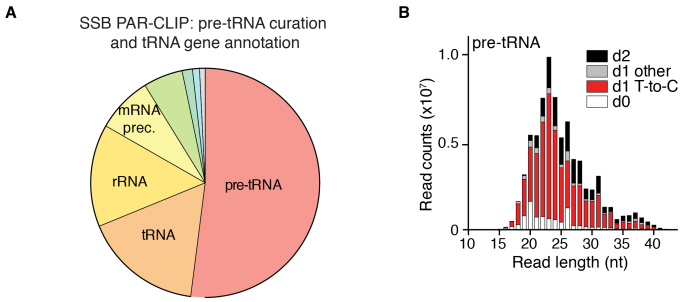

Figure 2.8: figure2cd. placeholder

The vast majority of crosslinking sites in pre-tRNAs were concentrated, as expected, in the oligoU tract of the 3 trailer sequence (**Fig. 2.9**A,B). We also found that SSB crosslinked to the 5 segment of the mature tRNA body at conserved sites in the D-stemloop (**Fig. 2.9**B), which is a novel finding, hinted at by a report proposing that the affinity of SSB for a full-length pre-tRNA cannot be explained solely by its binding to the 3 oligoU tract [24]. The other major target of SSB was 5S ribosomal RNA (rRNA), which is the only POLR3-transcribed rRNA, and as such also terminates with an oligoU stretch to which SSB crosslinked (Fig. S3).

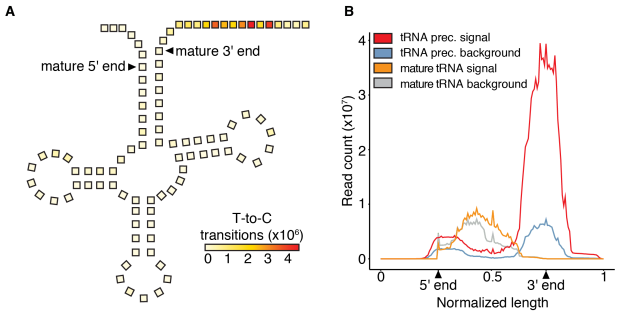

Figure 2.9: figure2ef. placeholder

2.10 tRNA gene annotation

We combined hydro-tRNAseq and SSB PAR-CLIP to identify actively transcribed tRNA genes (genomic locations that give rise to a supported pre-tRNAs). We confidently identified 288 tRNA genes as the intersection of 4 replicates of hydro-tRNAseq (**Fig. 2.10A**), and 349 tRNA genes as the intersection of two SSB PAR-CLIP experiments. Of note, SSB PAR-CLIP confirmed the expression of an additional 7 tRNA genes that were not supported in hydro-tRNAseq replicate (e.g. **Fig. 2.10B**), further showcasing the complementarity of the two approaches. We observed a strong correlation of pre-tRNA abundances between SSB PAR-CLIP and hydro-tRNAseq (Pearson R = 0.72; **Fig. 2.10D**), providing confidence that SSB PAR-CLIP quantitatively detected pre-tRNAs, without introducing biases (e.g. artificially enriching for lowly expressed pre-tRNAs). Instead, we observed no strong correlation between precursor and mature tRNA read counts in either of the two techniques (R ¡ 0.2; Fig. S4). The correlation of identified isoacceptor counts between SSB PAR-CLIP and hydro-tRNAseq was virtually perfect (Pearson R = 0.99; **Fig. 2.13C**), ruling out the introduction of a pronounced systematic bias from our hydrolysis-based protocol. Some anticodons seemed to be served by multiple isodecoders (e.g. 19 isodecoders for $tRNA^{Ser}_{GCA}$), while others only from one (e.g.

tRNA$^{Ser}_{ACT}$;(**Fig. 2.13B**, **Fig. 2.13**).Selenocysteine was the only amino acid that, in our data, was decoded by only one tRNA gene.

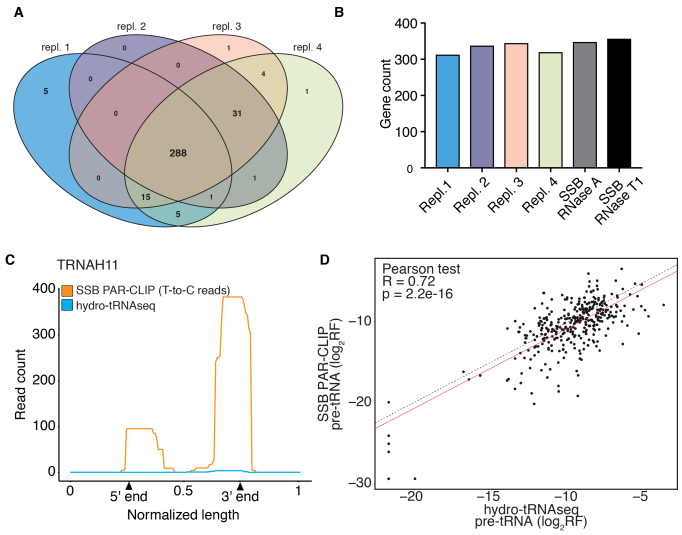

Figure 2.10: figure3. placeholder

2.11 Applications and biological insights

tRNA gene abundance does not correlate with tRNA gene count on the isotype level

There is no monotonic relationship between number of tRNA genes per amino acid and the abundance of each class/family of tRNAs. This lies in contrast with prior publications that had assumed that the number of tRNA predicted tRNA genes can be used as a proxy of tRNA expression.

This was assumed in the absence of tRNA sequencing data and because in yeast it seems that all tRNA genes are epxressed and seem to contribute equally to the mature tRNA pool. So, this result undesrcore the need for caution when reporting tRNA abunandce measurements and estimates.

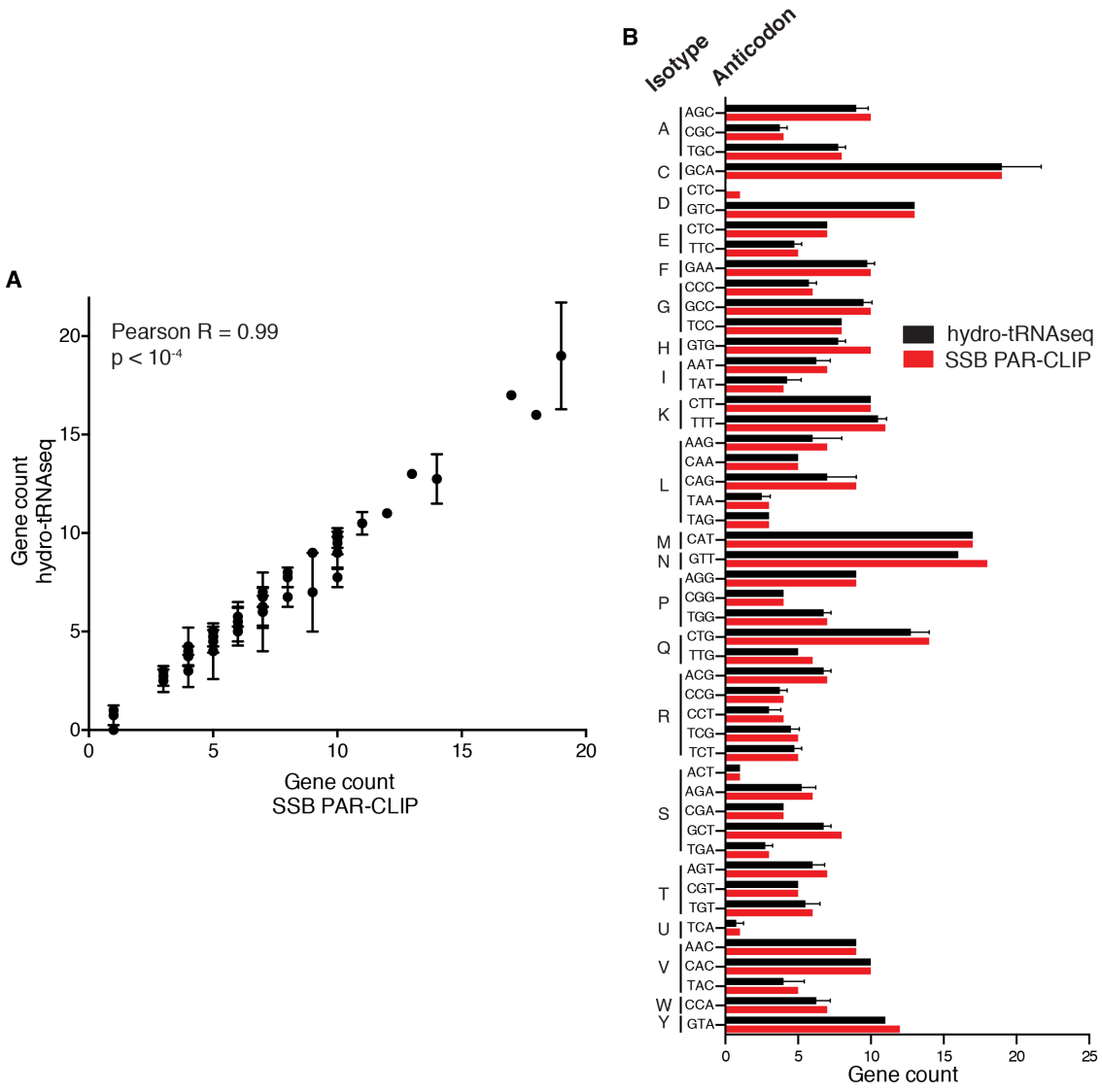

Figure 2.11: figure4. placeholder

Although tRNA isotypes with higher relative abundances generally tend to have higher tRNA gene numbers, we did not observe a clear linear correlation between read frequency and gene count (R = 0.12; **Fig. 2.12**)), like it has been reported before [7]. We then focused on the number of tRNA isoacceptors per amino acid, and isodecoders (tRNAs with the same anticodon sequence) per anticodon. We noticed a wide range of pre-tRNA counts per isoacceptor (**Fig. 2.13B**), with our data providing read evidence for 47 out of 62 coding codons (61 canonical and 1

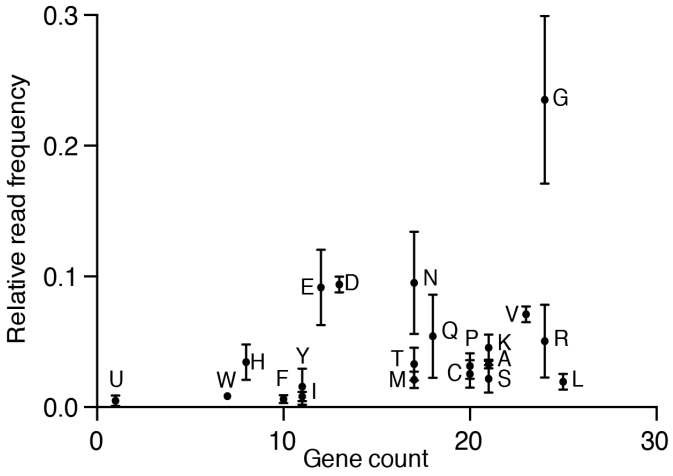

Figure 2.12: figure4A. placeholder

selenocysteine TAG).

tRNA gene abundance does not correlate with tRNA gene count on the isoacceptor (same amino acid, different anticodon) level

The same non-monotonic relationship seems to be true also on the level of isoacceptors, that is tRNAs with different anticodons that decode the same amino acid.

On this graph the data are broken down by aminoacid, which you can see as headers at the top, and then by anticodon which you can see on the bottom. The y-axis reperesnts tRNA gene count, and the size of every disc the relative abundance of all mature tRNAs with a given anticodon.

Thus, even though, for example, Cys GCA is the tRNA with the highest gene count, Gllycine GCC, is the tRNA group with the highest abundance.

Also, if you take a look at Proline, the group with highest gene count is the one with the lowest total abundance which is completely opposite by what was assumed beofre.

But how about correlation between idnividual pre- and mature tRNA levels

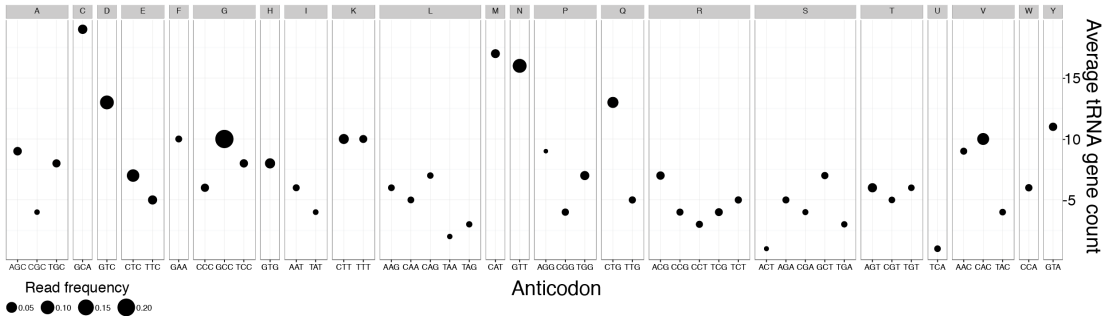

Figure 2.13: figure4D. placeholder

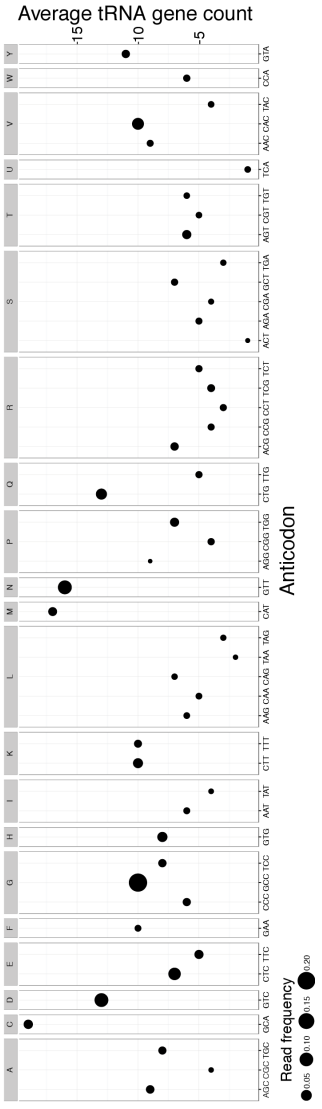

Figure 2.14: figure4Drot. placeholder

2.12 Mature tRNA abundance does not correlate with pre-tRNA abundance

No good correlation (pearson coefficients ¡ 0.2) as identified by our 2 separate techniques. (**Fig. 2.15**)

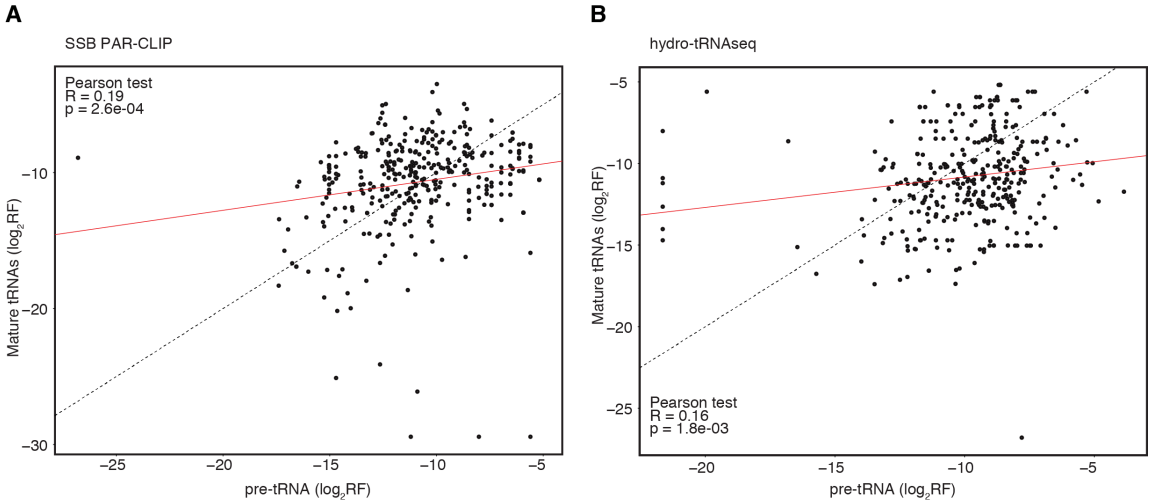

Figure 2.15: supp4. placeholder

2.13 tRNA transcription initiation and termination

Besides tRNA gene annotation and quantification, our approach yielded insights about pre-tRNA 3 trailer sequences. Based on hydro-tRNAseq, we determined the median 5 leader and 3 trailer lengths to be 6 and 10 nt, respectively, with the trailer lengths showing a broader distribution (**Fig. 2.16A,B**). Interestingly, SSB PAR-CLIP revealed a subset of much longer trailers (**Fig. 2.16C**), suggesting that SSB PAR-CLIP captured the very initial steps of precursor tRNA processing, and accordingly that hydro-tRNAseq captures pre-tRNAs partially trimmed, either by ELAC2 (tRNase Z) or some other nuclease [29].

We next focused on the POLR3 oligoU termination signals. Various reports in the past have focused on the oligoU requirements for transcription termination in different species [**Nielsen:2013be**, 30]. SSB protected consistently a 3 4 to 5 nt oligoU stretch, which was also confirmed by hydro-tRNAseq (**Fig. 2.17**). This is in agreement with previous *in vitro* results [**Teplova:2006dv**, 24, 26]. We also addressed the proposed requirement for a stem-loop immediately upstream of the oligoU termination signal [**Nielsen:2013be**]. Secondary structure predictions for the trailer sequences with documented sequence evidence in hydro-tRNAseq and SSB PAR-CLIP did not detect predicted stable stem-loop structures for approximately half of all pre-tRNAs (**Fig. 2.18**). This argued against a formal requirement for a stem-loop in the termination process of POLR3, at least on tRNA genes, in accordance with previous biochemical evidence [30].

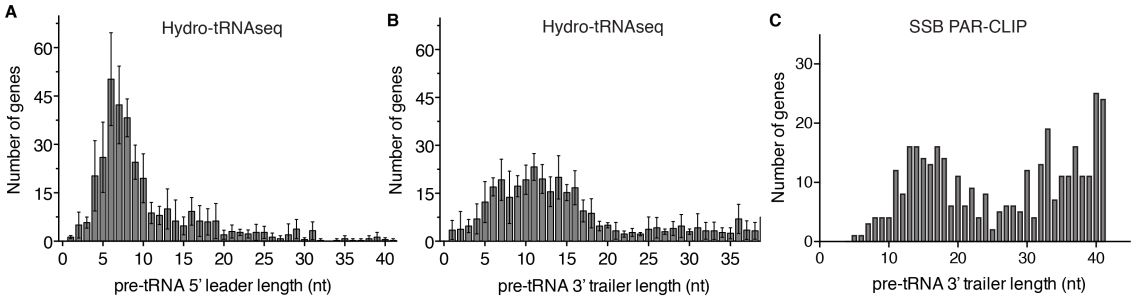

Figure 2.16: figure6. placeholder

2.14 Ribonucleotide modifications

RT across modified nucleotide-containing RNA leads to errors in cognate deoxynucleotide incorporation, revealed by mismatches in sequence reads upon mapping to reference genomic sequence. Read coverage across regions with a high degree of modifications may result in incomplete or largely uneven coverage. Therefore, we included in our mature tRNA reference the combination of all

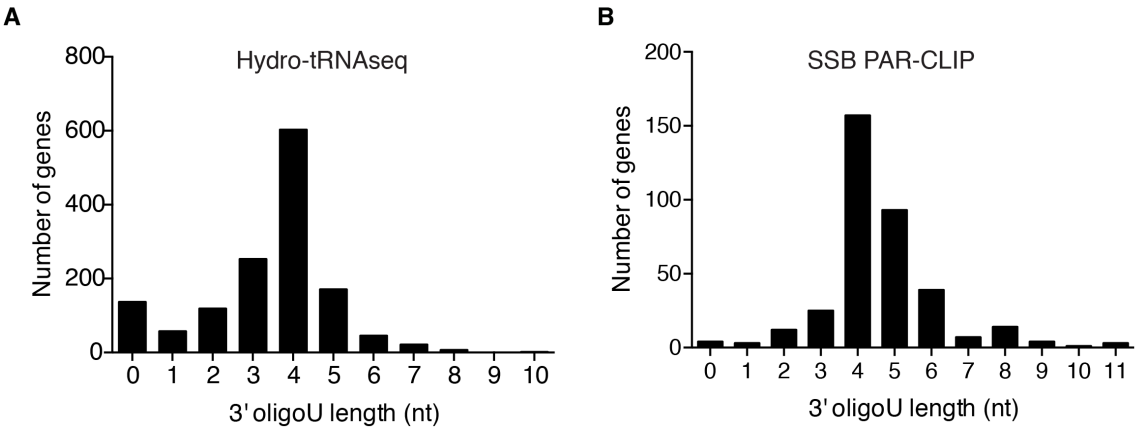

Figure 2.17: figure6de. placeholder

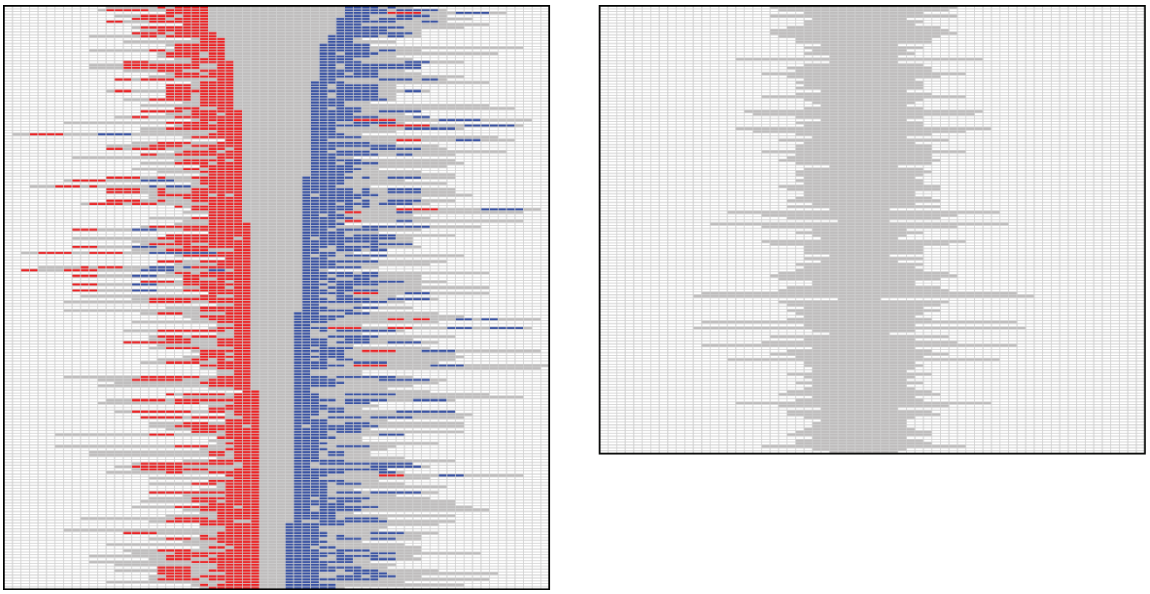

Figure 2.18: supp5. placeholder

frequent mismatch signatures in all heavily modified positions. We reported the most frequently modified positions per tRNA gene (Table S5), and computed the frequencies of every nucleoside change per position across all tRNA genes (**Fig. ??**).

The majority of editing events were A-to-G transitions at the first position of the anticodon and at the position 3 to the anticodon (usually position 37). Both positions are known to be heavily modified, the former being deaminated to ino-

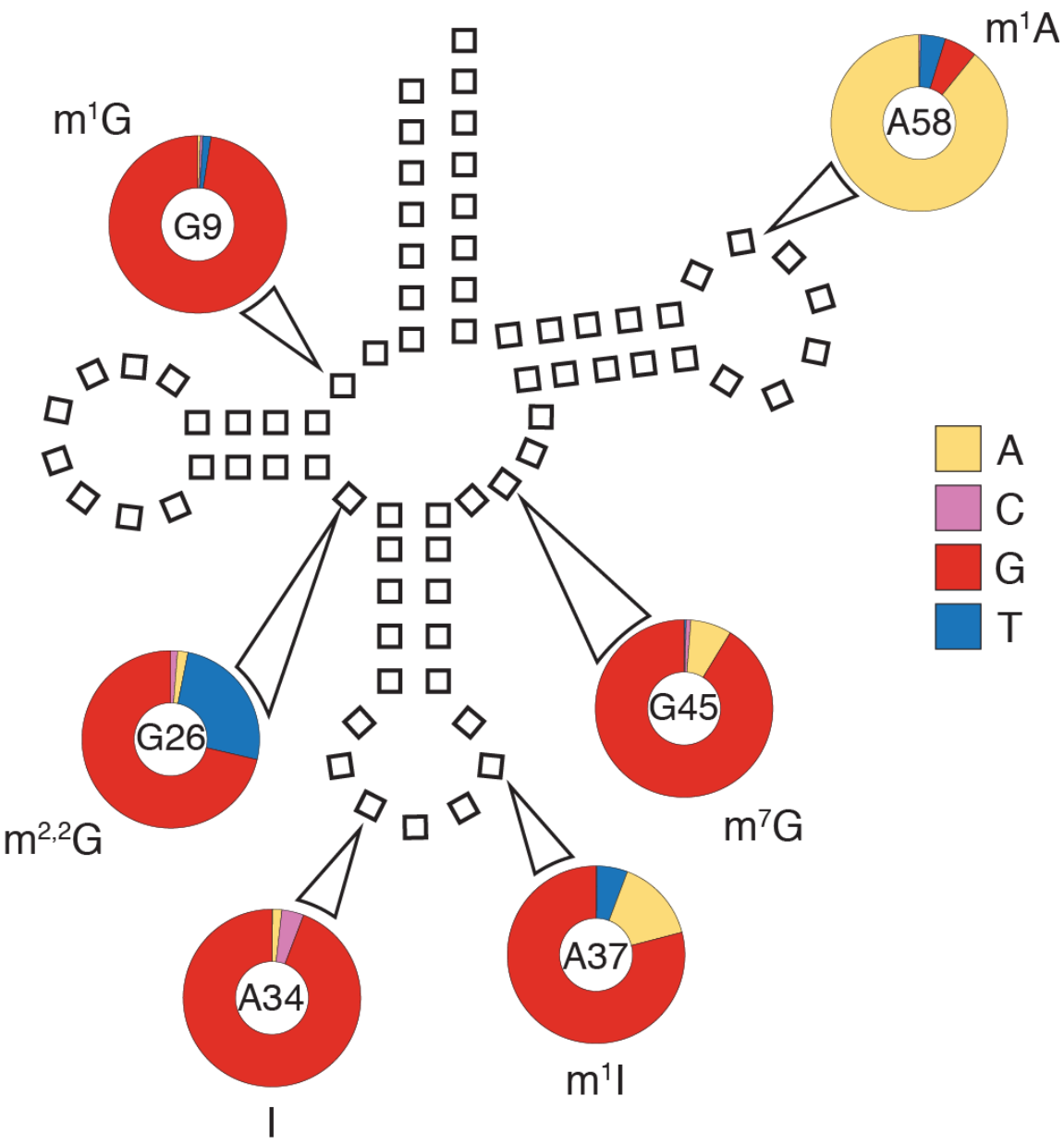

Figure 2.19: paper7a. placeholder

sine, and the latter further modified (e.g. 1-methylinosine) [**Machnicka:2013ky**]. In our data the majority of the reads that mapped to the anticodon of the modified tRNAs contained the mismatches. To a lesser extent we could also detect 1-methyladenosine in the pseudouridine loop (returned as A-to-T or A-to-G), and various guanosine modifications at positions 9, 26, and 45, which most likely

correspond to 1-methyl-, N2,N2-dimethyl-, and 7-methyl-guanosine, respectively [**Machnicka:2013ky**].

The temporal resolution of tRNA modifications by RNAseq has begun to be addressed recently [**Torres:2015ed**], however at a single modification level (inosine 34), and by using libraries relative poor in tRNA reads (¡1% of total reads). We were appropriately poised to address this issue since our very deep sequencing set, in combination with our hierarchical annotation pipeline, offered the advantage of dissecting multiple modifications simultaneously. We focused on the inosine modifications, since they represented the majority of modified nucleosides. By inspecting read alignments with error distance 1 to the reference pre-tRNA, we noticed A-to-G transition mismatches at position 34 in reads that retained the leader and trailer sequences of the precursor tRNA (**Fig. ??, top**). This confirmed that A34 deamination takes place at the precursor level, and therefore is a nuclear modification, as it has been previously reported [**Torres:2015ed**]. Next, we noticed that 1-methyl-inosine at position 37 also appears at the precursor stage. Of note the A37 modification became apparent prior to A34, as the majority of the error distance 1 reads contained a mismatch at A37. Reads with two mismatches contained both modifications (**Fig. ??, bottom**)

2.15 Annotation of intron-containing tRNA genes

Intron-containing tRNAs represent a particularly interesting set of tRNA genes, as mutations in their evolutionarily conserved, yet distinct, processing machinery have emerged recently as causes of severe neurodevelopmental syndromes, such as pontocerebellar hypoplasia [31]. Therefore, there is documented need for a comprehensive annotation of human intron-containing tRNAs, which should be

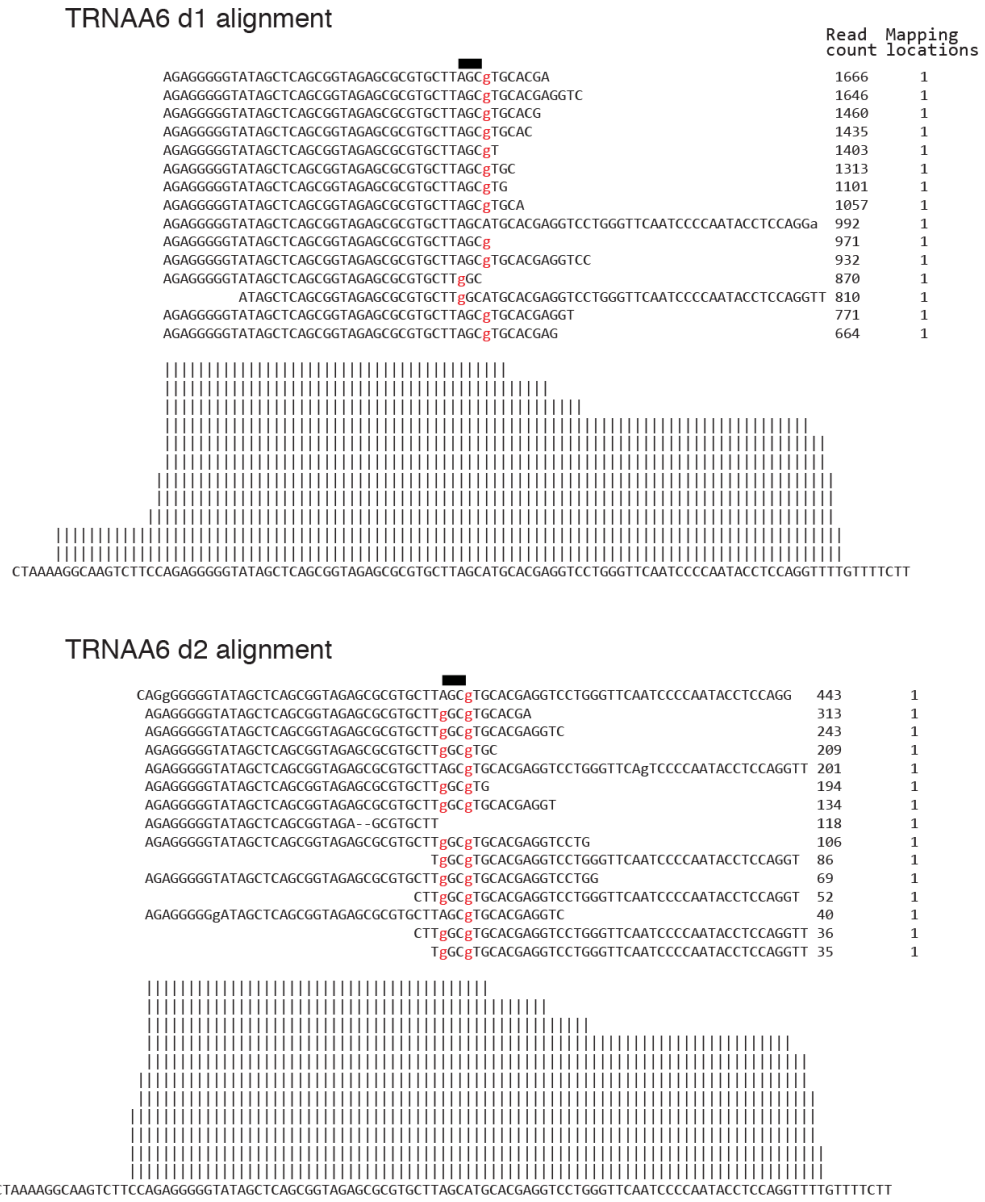

Figure 2.20: paper7b. placeholder

revisited as markers or disease-causing candidates in phenotypically similar conditions. We confirmed 26 out of 32 predicted intron-containing tRNAs by hydro-tRNAseq (**Fig. 2.20A**. Excluding any unknown biologically redundant mechanism, this suggests that the integrity of the tRNA splicing complex is essential for survival. To further confirm our observations, we coupled hydro-tRNAseq re-

sults with previously published PAR-CLIP data on the human tRNA ligase, RTCB [**Baltz:2012bh**]. Despite the shallow read depth of the dataset, we identified a crosslinked read peak at the anticodon loop of all intron-containing tRNAs annotated by our approaches (**Fig. 2.20B,C**.

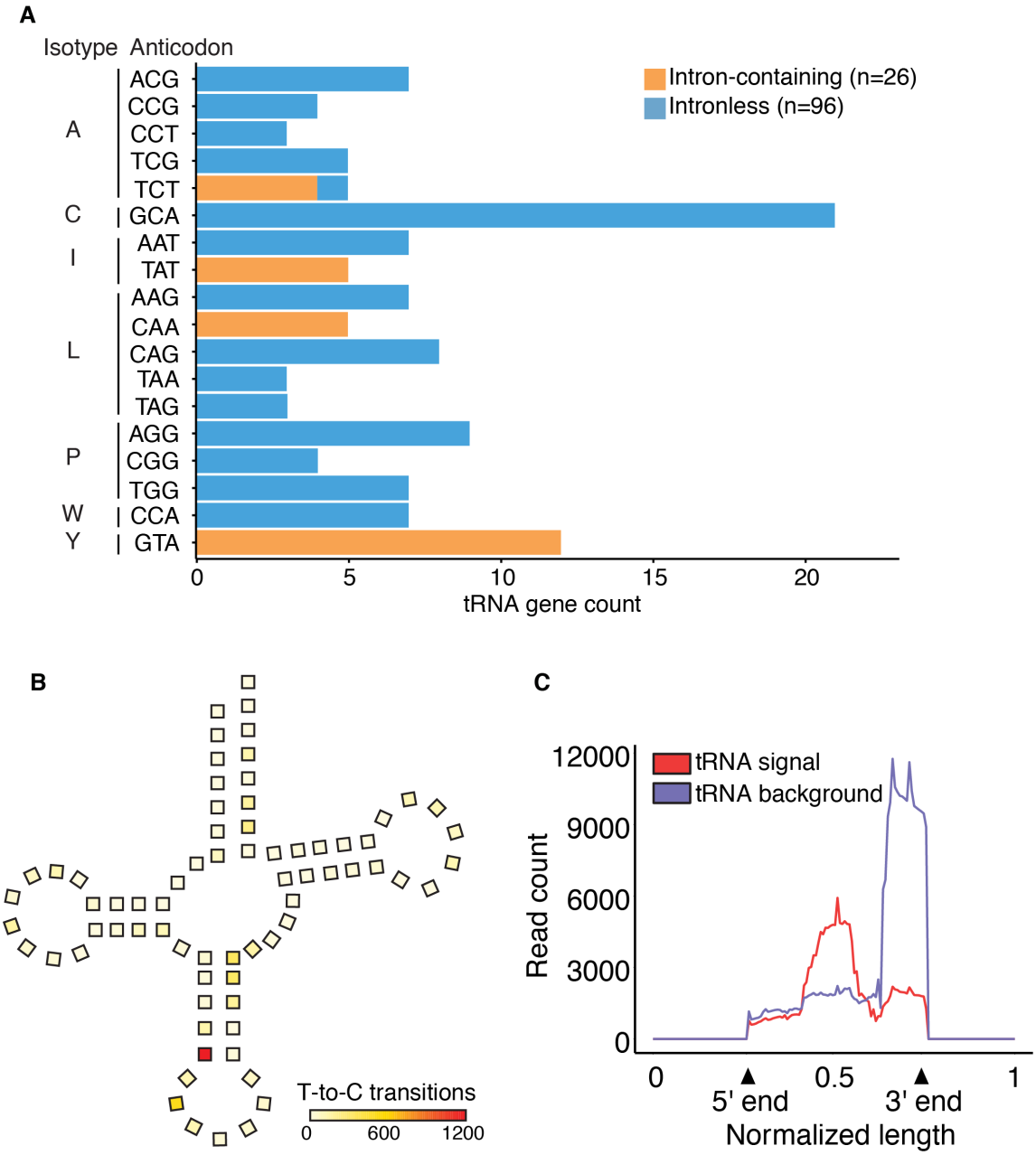

Figure 2.21: figure5. placeholder

Chapter 3

CLP1

Chapter 4

C3PO

References

1. Woese, C. *The Genetic Code. The Molecular basis for Genetic Expression* 1st ed. (Harper, 1967).

2. Soll, D. & RajBhandary, U. *tRNA: Structure, Biosynthesis and Function* 1st ed. (ASM press, 1995).

3. Dana, A. & Tuller, T. Determinants of Translation Elongation Speed and Ribosomal Profiling Biases in Mouse Embryonic Stem Cells. *PLoS Computational Biology* **8,** e1002755–11 (Nov. 2012).

4. Dana, A. & Tuller, T. Mean of the typical decoding rates: a new translation efficiency index based on the analysis of ribosome profiling data. *G3 (Bethesda, Md.)* **5,** 73–80 (Dec. 2014).

5. Mahlab, S., Tuller, T. & Linial, M. Conservation of the relative tRNA composition in healthy and cancerous tissues. *RNA* **18,** 640–652 (Mar. 2012).

6. Plotkin, J. B. & Kudla, G. Synonymous but not the same: the causes and consequences of codon bias. *Nature Reviews Genetics* **12,** 32–42 (Jan. 2011).

7. Tuller, T. *et al.* An Evolutionarily Conserved Mechanism for Controlling the Efficiency of Protein Translation. *Cell* **141,** 344–354 (Apr. 2010).

8. Weinberg, D. E. *et al.* Improved Ribosome-Footprint and mRNA Measurements Provide Insights into Dynamics and Regulation of Yeast Translation. *CellReports* **14,** 1787–1799 (Feb. 2016).

9. Hasler, D. *et al.* The Lupus Autoantigen La Prevents Mis-channeling of tRNA Fragments into the Human MicroRNA Pathway. *Molecular Cell* **63,** 110–124 (July 2016).

10. Ivanov, P., Emara, M. M., Villen, J., Gygi, S. P. & Anderson, P. Angiogenin-Induced tRNA Fragments Inhibit Translation Initiation. *Molecular Cell* **43,** 613–623 (Aug. 2011).

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

12. Iben, J. R. & Maraia, R. J. tRNA gene copy number variation in humans. *Gene* **536,** 376–384 (Feb. 2014).

13. Pechmann, S. & Frydman, J. Evolutionary conservation of codon optimality reveals hidden signatures of cotranslational folding. *Nature Publishing Group* **20,** 237–243 (Dec. 2012).

14. Gingold, H. *et al.* A Dual Programfor Translation Regulationin Cellular Proliferation and Differentiation. *Cell* **158,** 1281–1292 (Sept. 2014).

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

16. Moqtaderi, Z. *et al.* Genomic binding profiles of functionally distinct RNA polymerase III transcription complexes in human cells. *Nature Structural & Molecular Biology* **17,** 635–640 (Apr. 2010).

17. Oler, A. J. *et al.* nsmb.1801. *Nature Structural & Molecular Biology* **17,** 620–628 (Apr. 2010).

18. Dittmar, K. A., Mobley, E. M., Radek, A. J. & Pan, T. Exploring the Regulation of tRNA Distribution on the Genomic Scale. *Journal of Molecular Biology* **337,** 31–47 (Mar. 2004).

19. Goodarzi, H. *et al.* Modulated Expression of Specific tRNAs Drives Gene Expression and Cancer Progression. *Cell* **165,** 1416–1427 (June 2016).

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

21. Zheng, G. *et al.* Efficient and quantitative high-throughput tRNA sequencing. *Nature Methods,* 1–5 (July 2015).

22. Karaca, E. *et al.* Human CLP1 Mutations Alter tRNA Biogenesis, Affecting Both Peripheral and Central Nervous System Function. *Cell* **157,** 636–650 (Apr. 2014).

23. Foretek, D., Wu, J., Hopper, A. K. & Boguta, M. Control of Saccharomyces cerevisiae pre-tRNA processing by environmental conditions. *RNA,* 1–12 (Jan. 2016).

24. Bayfield, M. A. & Maraia, R. J. Precursor-product discrimination by La protein during tRNA metabolism. *Nature Structural & Molecular Biology* **16,** 430–437 (Mar. 2009).

25. Bayfield, M. A., Yang, R. & Maraia, R. J. Conserved and divergent features of the structure and function of La and La-related proteins (LARPs). *BBA - Gene Regulatory Mechanisms* **1799,** 365–378 (May 2010).

26. Stefano, J. E. Purified lupus antigen La recognizes an oligouridylate stretch common to the 3' termini of RNA polymerase III transcripts. *Cell* **36,** 145–154 (Jan. 1984).

27. Hafner, M. *et al.* Barcoded cDNA library preparation for small RNA profiling by next-generation sequencing. *Methods* **58,** 164–170 (Oct. 2012).

28. Maraia, R. J. & Lamichhane, T. N. 3 processing of eukaryotic precursor tRNAs. *Wiley Interdisciplinary Reviews: RNA* **2,** 362–375 (Nov. 2010).

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

30. Arimbasseri, A. G., Kassavetis, G. A. & Maraia, R. J. Transcription. Comment on "Mechanism of eukaryotic RNA polymerase III transcription termination". *Science* **345,** 524–524 (Aug. 2014).

31. Namavar, Y., Barth, P. G., Poll-The, B. T. & Baas, F. Classification, diagnosis and potential mechanisms in Pontocerebellar Hypoplasia. *Orphanet Journal of Rare Diseases* **6,** 50 (July 2011).