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 tRNAs in 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 read mapping efforts. To address these issues, I developed hydro-tRNAseq, a method based on partial alkaline RNA hydrolysis that generates fragments amenable for sequencing. To identify transcribed tRNA genes, I further complemented this approach with Photoactivatable Crosslinking and Immunoprecipitation (PAR-CLIP) of SSB/La, a conserved protein involved in pre-tRNA processing. My results show that approximately half of all predicted tRNA genes are transcribed in human cells, suggesting that the tRNA genomic space is more contracted than previously thought as a result of regulation of expression. I also report predominant nucleotide modification sites, their order of incorporation, and identify tRNA leader, trailer and intron sequences. By using complementary sequencing-based methodologies I present a human tRNA reference set, and determine expression levels of mature and processing intermediates of tRNAs in human cells.

The technical advances provided by hydro-tRNAseq are applied towards the molecular diagnosis of a genetic neurodevelopmental syndrome, caused by mutations in the tRNA processing factor CLP1.

Finally, I harness this novel experimental and computational expertise towards the identification of the endonuclease complex C3PO as a novel processing factor of human tRNAs. I carry out a transcriptome-wide analysis of C3PO targets, identify its binding sites and motifs, and provide insights into its biochemical and biological functions.

To my parents and my brother

Acknowledgments

I owe my being to my parents, but my well-being to my teacher

Alexander the Great

I would like to thank Tom, my advisor, for exposing me to a combination of intellectual rigor, scientific skill, ethical conduct, interest in society, and personal drive that I had never before thought it existed. Thank you, Tom, for making me a better (and hopefully good) scientist, and for the support on every aspect of my life; for not being hard on me when you should have, for giving me the freedom to do whatever I wanted, and for always smiling while saying "good morning."

Many thanks go out to all the members of the Tuschl lab for all their help and support: Manny Ascano for advice and supervision, Miguel Brown for my first lessons of bioinformatics, Manju Kustagi for computer help, discussions and Indian food, Kemal Akat for help with medicine and R, Pavel Morozov for help with statistics, Masashi Yamaji for insightful discussions, Aitor Garzia for experimental help and data, Klaas Max for saying "kalimera," Artem Serganov for experimental help and soccer discussions, Lodoe Lama for all the help and all the food, Jonathan Liau for eating all the food, Jenny Li for taking care of everything, and Claudia Bognanni for her friendship. And of course, Leonida for keeping me on track. I would also like to thank Jill de Jong for all the help, support and friendship.

Special thanks especially Markus Hafner (Xaipe!) for being a mentor inside and outside the lab, and for helping me with everything.

I am very grateful to Javier Martinez and Dinshaw Patel for their continuing support and for allowing me to be part of their scientific endeavors.

I would like to acknowledge all my committee members, present and past (Jim Hudspeth, Luciano Marraffini, Dinshaw Patel, Nina Papavasiliou, and Rich Maraia).

Also, the Programming for Biology Course instructors at CSHL, all my teachers at Rockefeller, and the Tri-I MD/PhD program for enabling me to fulfill my greatest ambitions so far.

I would also like to thank my Greek friends in the U.S.: Mihalis Maniatakos, Dimitris Zattas, Argyro Katsika, Palmyra Geraki, Antonis Stampoulis, Costas Arkolakis and Sotiria Palioura, and everyone back home.

Special thanks to Vasilis Kalogeridis. And to Christos Papadopoulos for having been "a collaborator and "friend."

Last, but not least, I would like to thank my brother Apostolos, for his love, for being my first teacher, and for always setting an unreachable level of excellence for me, my Giulia for her love and support, and my parents, Ilias and Evangelia, to whom I owe not only my being, but also everything else.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

4-SU	4-thiouridine.
C3PO	component 3 promoter of RISC.
ChIP-seq	chromatin immunoprecipitation sequencing.
CLP1	cleavage and polyadenylation factor I subunit 1.
EMSA	electrophoretic mobility shift assay.
HEK293	Human embryonic kidney cells 293.
La	Lupus La protein.
ncRNA	noncoding RNA.
nt	nucleotide.
OH	hydroxyl.
PAR-CLIP	photoactivatable-ribonucleoside-enhanced crosslinking and immunoprecipitation.
POLR3	human RNA polymerase III.
pre-tRNA	precursor tRNA.
RBP	RNA-binding protein.
RNA-seq	RNA sequencing.
RNP	ribonucleoprotein.
RT	reverse transcriptase.

SSB	Sjögren syndrome type B antigen.
tRBP	tRNA-binding protein.
tRNA	transfer RNA.
TSEN	tRNA splicing and endonuclease complex.
TSN	translin.
TSNAX	translin-associated protein X, also known as TRAX.

Glossary

tRNA isoacceptors	tRNA molecules that decode synonymous codons.
tRNA isodecoders	tRNA molecules that have the same anti-codon, and thus decode the same amino acid.
tRNA isotype	colection of tRNAs encoding the same amino acid.

Part I

Characterizing human tRNAs

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 [1], and are among the earliest studied noncoding RNA (ncRNA) non-coding RNA molecules [2, 3]. 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 [4–7]. Despite their highly conserved participation in the translational machinery, tRNAs have received new attention in recent years in the context of codon-resolved translational control [8–13], 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) [14–18].

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 [12, 19, 20]. 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 [21]. Similarly, experimental tRNA gene annotation in the past had to focus on human RNA polymerase III (POLR3) chromatin immunoprecipitation sequencing (ChIP-seq) [22–24] or hybridization-based approaches [25, 26]. 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 [27, 28]. However, an extensive annotation of human genes and transcripts was foregone because the focus was either on mature tRNAs only [28] or on tRNA fragments not inclusive of full-length precursor tRNA (pre-tRNA) transcripts [27]. Thus, to-date an experimentally validated list of curated mature and pre-tRNA sequences and annotating tRNA genes in human is still missing.

To address this lack of experimentally-validated tRNA reference, I combined complementary high-throughput techniques for obtaining the sequence composition and abundance of tRNAs in human cells.

1.2 tRNA biogenesis

tRNA genes are transcribed by POLR3 that uses promoters internal to the DNA sequence of the tRNA gene (tDNA), resulting in a primary transcript with a 5’

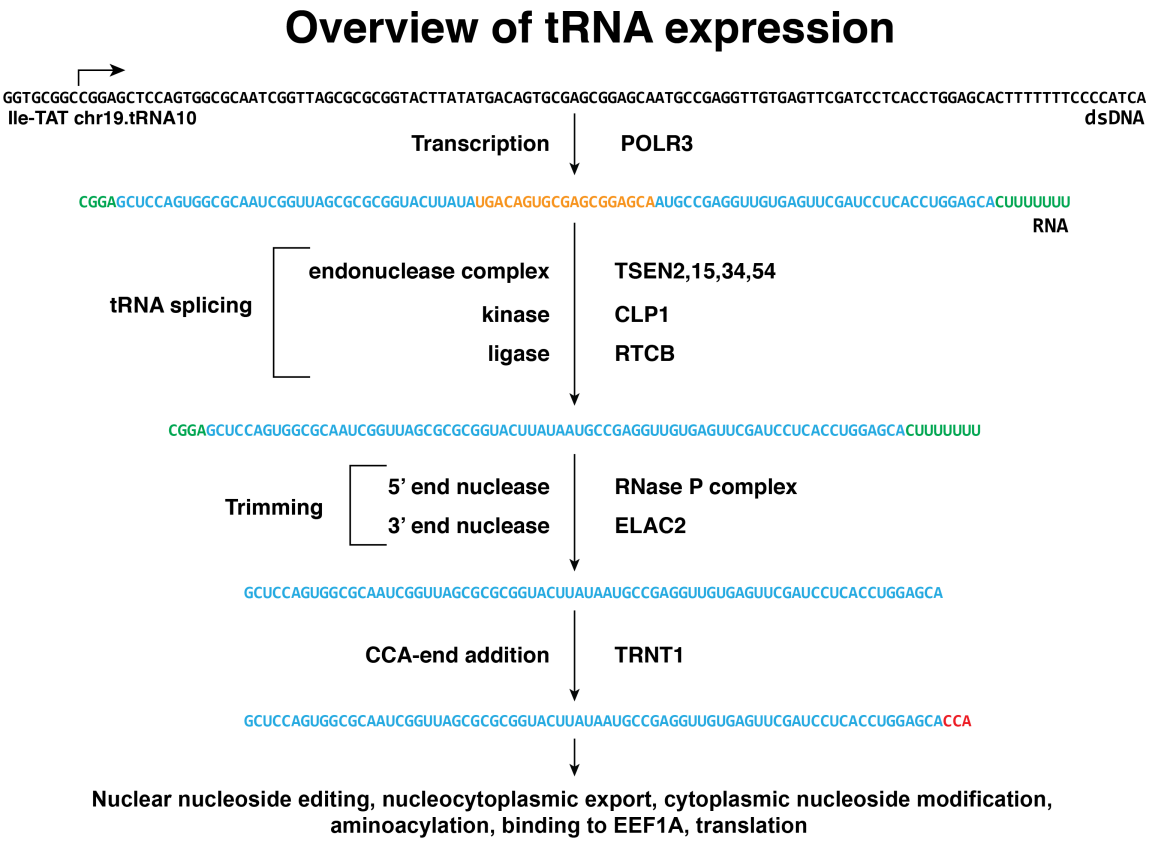


Figure 1.1: Overview of tRNA biogenesis and processing. tRNAs are transcribed by POLR3. If present, tRNA introns are removed by the tRNA splicing complex (TSEN and CLP1), and mature halves are ligated by the tRNA ligase (RTCB). Pre-tRNA leaders are trimmed by the RNase P complex, and 3' trailer by ELAC2. The 3' terminal CCA tail is added by TRNT1. tRNAs are further modified by nucleoside editing in the nucleus and the cytoplasm, are aminoacylated by cognate tRNA synthetases and are presented to the ribosome by translation factors.

triphosphate. In humans, a minority of tRNA transcripts harbor introns (see section 3.6). A dedicated tRNA splicing complex composed of core and accessory proteins carries out tRNA splicing [29–33]. 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 possibly 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 the tRNA nucleotidyl transferase 1 (TRNT1), and acts as the acceptor of the amino acid. tRNAs are further modified by chemical nucleotide modifications (3.5), 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**)[34–36]. Mature tRNAs adopt a cloverleaf secondary and L-shaped tertiary structure (**Fig. 1.2**)

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 specificity [21, 37]. Importantly, abnormal tRNA expression levels have been correlated and causally associated with pathologic conditions, such as cancer [21, 26].

1.3 tRNA sequencing

Their complex biogenesis and processing pathway adds multiple layers of difficulty to the analysis of tRNAs. Obtaining data for tRNAs is hindered by multiple obstacles:

i) sequencing of tRNAs is technically arduous due to their relatively small size, and their stable structure that impedes enzymes used in cDNA library prepa-

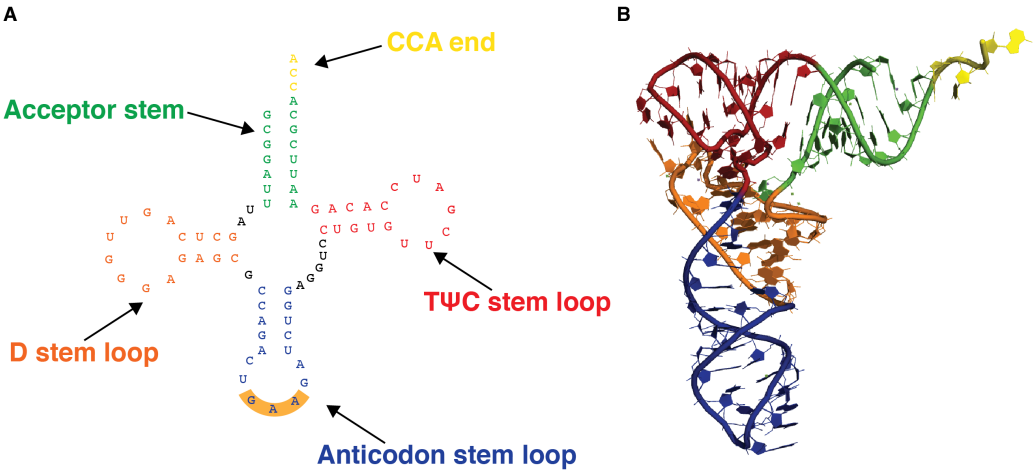


Figure 1.2: tRNA structure. (A) tRNA transcripts, such as the phenylalanine tRNA shown here, adopt the typical cloverleaf secondary structure, which in turns adopts an L-shaped tertiary structure as shown in (B). The structurally conserved stems and stemloops are indicated in (A), color-coded, and their coordinates are reflected in the 3-dimensional structure in B (PDB 1EHZ).

rations, such as RNA ligases and reverse transcriptase (RT)

ii) numerous (>100) tRNA pseudogenes are interspersed in the human genome [38, 39]

iii) all tRNAs undergo extensive post-transcriptional processing (see 1.1), while some involve extra processing steps (intron removal, addition of a 5' guanosine to all histidine tRNAs [40])

iv) tRNAs are subjected to extensive chemical modifications on numerous nucleosides, which lead to mismatches upon the reverse transcription step of the RNA cloning protocols [41, 42]. 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 TΨC loop) [41, 43]. 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.

v) tRNA isoacceptors share a large degree of sequence similarity that makes the distinction between alternative isoacceptors and editing products equivocal.

vi) 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 adequately with RNA sequencing (RNA-seq) tools. Instead, information about tRNA sequences and genes comes from bioinformatic predictions [38, 39]. Such approaches take into account base-pair covariation, 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 tRNAs.

1.4 Previous efforts for genome-wide tRNA annotation

Even though no direct and rigorous experimental validation of tRNA sequences has been carried out, there has been indirect experimental evidence for tRNA expression:

i) ChIP-seq studies, focusing on the occupancy of genomic locations by POLR3 and/or its transcription factors [22–24]

ii) tRNA microarrays, using the predicted tRNA sequences as the reference for the creation of array probes [25]

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 modifications, while their dynamic range is considerably narrower than RNA-seq. Finally, neither method is appropriately equipped to determine definitively pre-tRNAs or their transcription start and termination sites. This is an important limitation, as pre-tRNA fragments have been associated with neurodegenerative diseases [44–46].

1.5 Small RNA sequencing protocol

In order to obtain RNA-seq reads, I decided to first apply the well established protocol for sequencing small RNAs, developed in the Tuschl lab [47] (**Fig. 1.3**). The experimental procedure takes advantage of the 5' monophosphate (p) and 3' OH groups present in small RNAs, such as microRNAs (miRNAs) , in order to enrich for such RNA species over other abundant RNA molecules. The use of a truncated and mutated RNA ligase (T4 Rnl2(1-249)K227Q) that requires 5' preadenylated adapter (App-(5')-adapter) prevents on one hand the formation of secondary, circularized byproducts, and allows on the other for exclusive ligation at the 3' end of the target RNA. Rnl1 is used to ligate an adapter with a different sequence at the 5' end, by activating the 5' monophosphate of the small RNA. cDNA is obtained by RT and amplified at non-saturated levels by PCR. The derived small RNA cDNA library is submitted to high-throughput sequencing on Illumina instruments using sequencing primer sites present in the adapter sequences. The different se-

quences of the 3' and 5' adapters preserve the strandedness of the original RNA sequence, enhancing ncRNA discovery and curation. Adding short (5-nt) barcode sequences at the 3' adapter also allows for pooling of several multiplexed samples, reducing costs, processing time, and batch variability.

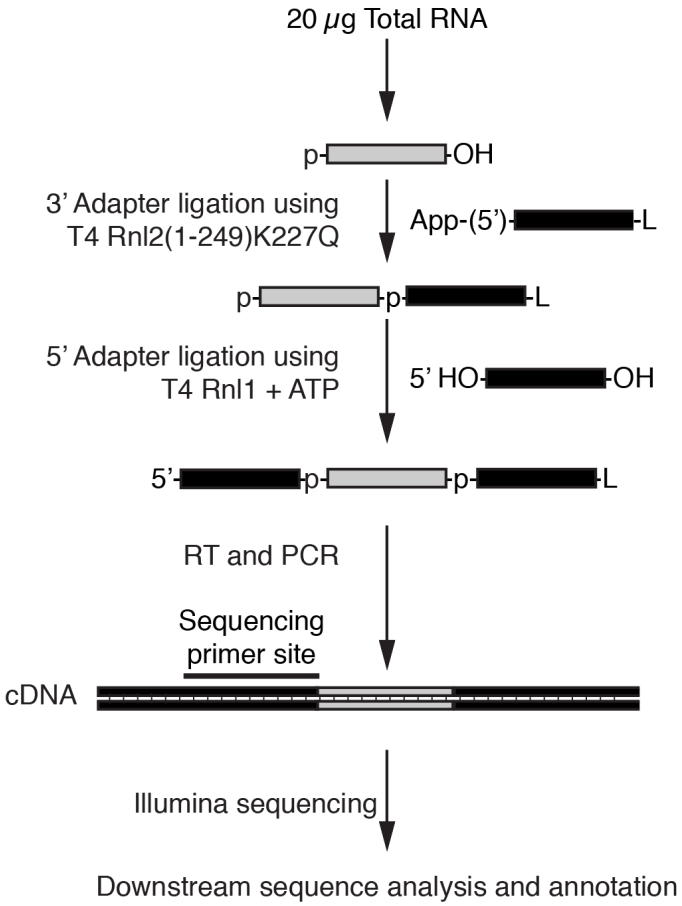


Figure 1.3: Small RNA sequencing protocol. Schematic overview of the conventional small RNA sequencing protocol, as it has been described previously [47]

Even though the utility of this protocol has been documented for the discovery and quantification of miRNAs, it was reasonable to apply towards tRNA sequencing because:

i) tRNAs, which are on average 75 nucleotides (nts) long, are closer in length than most other highly abundant ncRNAs (typically longer than 150 nts)

ii) mature tRNAs and miRNAs both have a 5' monophosphate and 3' OH, which are employed at different steps of library preparation

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

RNA type	% Total reads	Mean length (nt)
rRNA	35.8%	60.5
no match	24.1%	76.2
no annotation	17.8%	64.2
snRNA/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 types recovered by small RNA sequencing protocol. Percentage of reads mapped to indicated ncRNA type over total depth of library, and mean length of reads mapped to RNAs of each type are shown. snRNA: small nuclear RNA; snoRNA: small nucleolar RNA; repeat: repetitive DNA sequence; miscRNA: all other ncRNAs.

Chapter 2

Hydro-tRNAseq

2.1 Experimental innovation

In order to overcome the problems associated with tRNA sequencing, I tried to identify the minimal number of simplest steps that could tackle the maximal number of problems. Thus, I isolated 60-100 nt-sized total RNA from Human embryonic kidney cells 293 (HEK293) cells, comprising both pre- and mature tRNAs, but being devoid of most other abundant RNAs and short tRNA turnover products [16]. Full-length tRNAs have thermodynamically stable secondary and tertiary structures and are heavily modified by RNA editing, all of which compromise RT and RNA-seq analysis. To overcome these problems, I implemented a limited alkaline hydrolysis step. I reasoned that hydrolysis would generate shorter RNA fragments less prone to adopt stable structures, and would also reduce the number of modified nucleosides per sequenced fragment.

The value of the latter effect becomes apparent if one performs the following thought experiment. Let us assume that the probability of an RT "problem" (stall, drop or misincorporation) is the same for all modifications (p). The compound

probability of RT stalling, dropping or misincorporating a nucleoside in a given sequence is given by the product of any of these events happening at a given modified position, and is equal to:

$$P = p^n \tag{2.1}$$

where n = number of modified nucleosides affecting RT. Given that full length tRNAs are longer than the hydrolysis-derived fragments, and modifications are usually distributed over the loops of the tRNA (**Fig. 1.2** and **Fig. 3.6**), then

$$n_{full-length} \geq n_{fragment} \tag{2.2}$$

and therefore:

$$P_{full-length} \geq P_{fragment} \tag{2.3}$$

and the probability of sequencing through an RNA fragment $(1 - p)$:

$$1 - P_{full-length} \leq 1 - P_{fragment} \tag{2.4}$$

In addition to increasing read-through by RT, the reduced frequency of modified nucleosides per sequenced fragment, also improves mapping efforts by reducing the number of mismatches per sequenced read. 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. Thus, I anticipated that collectively these effects would yield RNA sequences more amenable to small RNA cDNA library preparation and deep sequencing than the refractory tRNAs. Indeed this approach increased the tRNA read content to >40% in

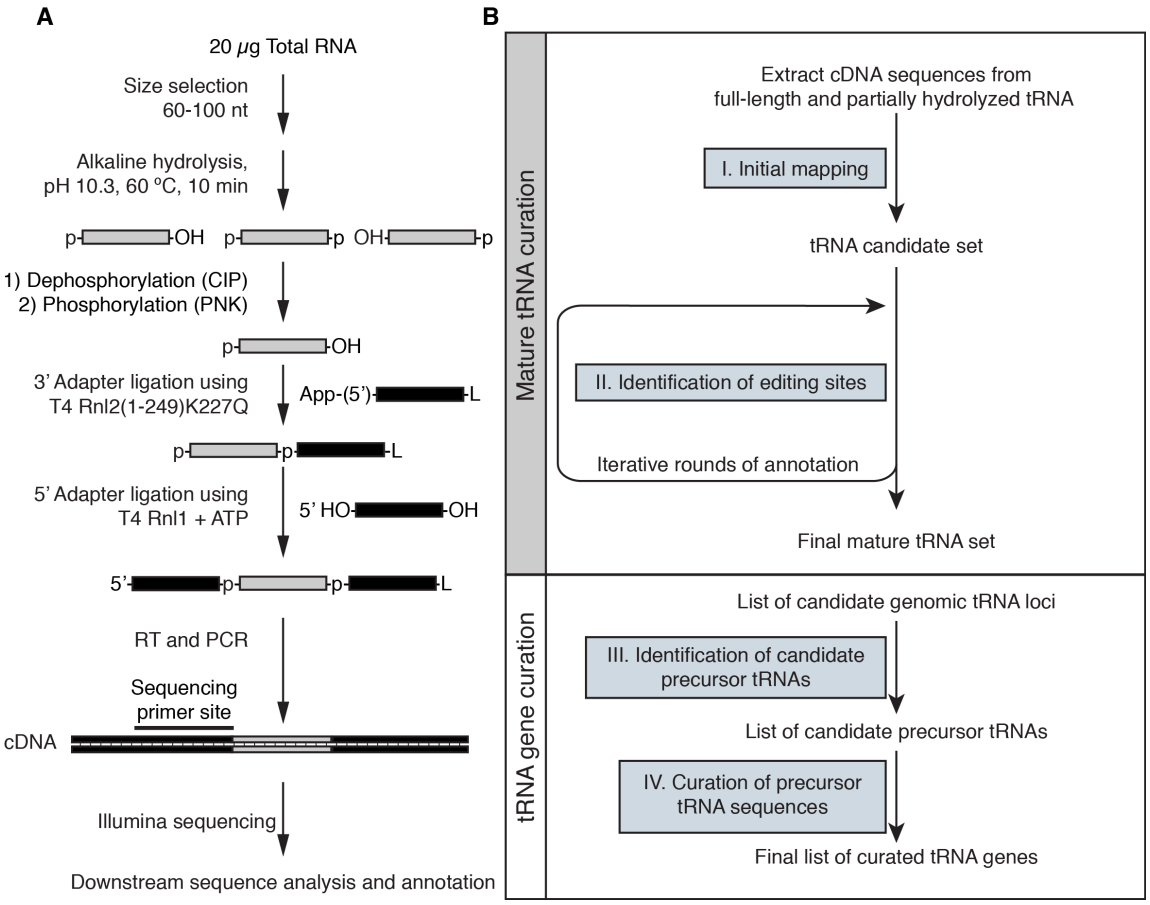


Figure 2.1: Hydro-tRNAseq experimental and bioinformatic pipeline for tRNA annotation and reference transcript curation by hydro-tRNAseq. (A) tRNAs and pre-tRNAs were size-selected from HEK293 total RNA and subjected to limited alkaline hydrolysis, followed by dephosphorylation, rephosphorylation and 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.

our deepest dataset (**Table 2.1**). We named this procedure hydro-tRNAseq (**Fig. 2.1A**).

Encouraged by the preliminary performance of the protocol, I set out to obtain a curated list of human nuclear and mitochondrial tRNAs, their genomic loci, and their processing intermediates. However, such an effort required a custom-made computational analysis pipeline.

Type	D0 counts	D1 counts	D2 counts	Total	% total starting	% over mapped reads	% D0/ total D0	D0/total (per type)	D1/total	D2/total
tRNA	41,880,208	9,444,737	2,814,103	54,139,048	44%	47%	47%	77%	17%	5%
rRNA	18,825,243	4,317,680	1,222,652	24,365,575	20%	21%	21%	77%	18%	5%
mt tRNA	15,254,805	2,976,430	736,595	18,967,830	15%	17%	17%	80%	16%	4%
snoRNA	8,126,590	1,635,120	388,833	10,150,543	8%	9%	9%	80%	16%	4%
mRNA	951,538	231,248	89,536	1,272,322	1%	1%	1%	75%	18%	7%
mRNA gene	828,458	229,342	150,541	1,208,341	1%	1%	1%	69%	19%	12%
snRNA	773,703	169,800	43,861	987,364	1%	1%	1%	78%	17%	4%
pre-tRNA	529,077	247,304	146,557	922,938	1%	1%	1%	57%	27%	16%
genome	327,950	114,402	153,951	596,303	0%	1%	0%	55%	19%	26%
mt rRNA	363,913	86,271	24,813	474,997	0%	0%	0%	77%	18%	5%
scRNA	339,135	91,056	27,098	457,289	0%	0%	0%	74%	20%	6%
marker	113,303	52,725	32,143	198,171	0%	0%	0%	57%	27%	16%
rRNA prec	106,373	49,811	12,589	168,773	0%	0%	0%	63%	30%	7%
bacterial	31,724	38,555	52,354	122,633	0%	0%	0%	26%	31%	43%
mt mRNA	47,466	12,816	4,887	65,169	0%	0%	0%	73%	20%	7%
lincRNA gene	35,934	10,686	9,516	56,136	0%	0%	0%	64%	19%	17%
mt genome	17,990	14,749	20,379	53,118	0%	0%	0%	34%	28%	38%
miRNA	5,425	7,443	38,348	51,216	0%	0%	0%	11%	15%	75%
lincRNA	26,713	6,474	1,460	34,647	0%	0%	0%	77%	19%	4%
snoRNA prec	22,776	6,544	2,774	32,094	0%	0%	0%	71%	20%	9%
plasmid	9,461	3,193	1,896	14,550	0%	0%	0%	65%	22%	13%
scaRNA	10,065	2,184	526	12,775	0%	0%	0%	79%	17%	4%
snRNA prec	7,207	2,076	1,934	11,217	0%	0%	0%	64%	19%	17%
piRNA	1,227	866	1,283	3,376	0%	0%	0%	36%	26%	38%
scRNA prec	173	43	132	348	0%	0%	0%	50%	12%	38%
adapter	0	0	160	160	0%	0%	0%	0%	0%	100%
doubtful miRNA	101	42	14	157	0%	0%	0%	64%	27%	9%
mirtron	24	5	1	30	0%	0%	0%	80%	17%	3%
scaRNA prec	10	8	1	19	0%	0%	0%	53%	42%	5%
std cali	0	0	1	1	0%	0%	0%	0%	0%	100%
total	88,636,592	19,751,610	5,978,938	114,367,140	93%	100%	100%	n/a	n/a	n/a

Table 2.1: Hydro-tRNAseq reads per RNA type. Reads assigned to each RNA type following hierarchical annotation are shown. Reads with no (d0), one (d1) or two (d2) mismatches compared to reference are shown. Hydro-tRNAseq enriches for nuclear (44%) and mitochondiral (15%) tRNAs.

2.2 Bioinformatics analysis pipeline

2.2.1 Hierarchical sequence read mapping

In parallel with the tRNA annotation procedure, I had to build a bioinformatics pipeline for processing the obtained sequence information. To account for the

multiple maturation steps between pre- and mature tRNAs, in collaboration with bioinformatician Miguel Brown, we developed an iterative, hierarchical approach for mapping and annotating our sequence reads.

First we mapped reads to reference tRNA genes for the human genome release hg19 (`http://gtrnadb.ucsc.edu/`) using an iterative and hierarchical protocol (**Fig. 2.1B**). We started by mapping only to mature tRNAs, which included the 3' CCA aminoacyl acceptor terminus, and the G_{-1} nucleotide added posttranscriptionally to histidine tRNAs [40, 48], but excluded tRNA introns. Starting with two most abundant tRNA transcripts per tRNA isotype, 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 or 2 than reads without errors. If these reads with mismatches could be assigned to other tRNA isoacceptors, 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.

2.2.2 tRNA gene annotation

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 nts upstream and downstream 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. These reads 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. 2.2**). At the end of our analysis we accounted for 93% of the 114,367,140 reads in our deepest library (**Table 2.1**).

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, suggesting that our analysis was exhaustive, at least for tRNAs in HEK293 cells.

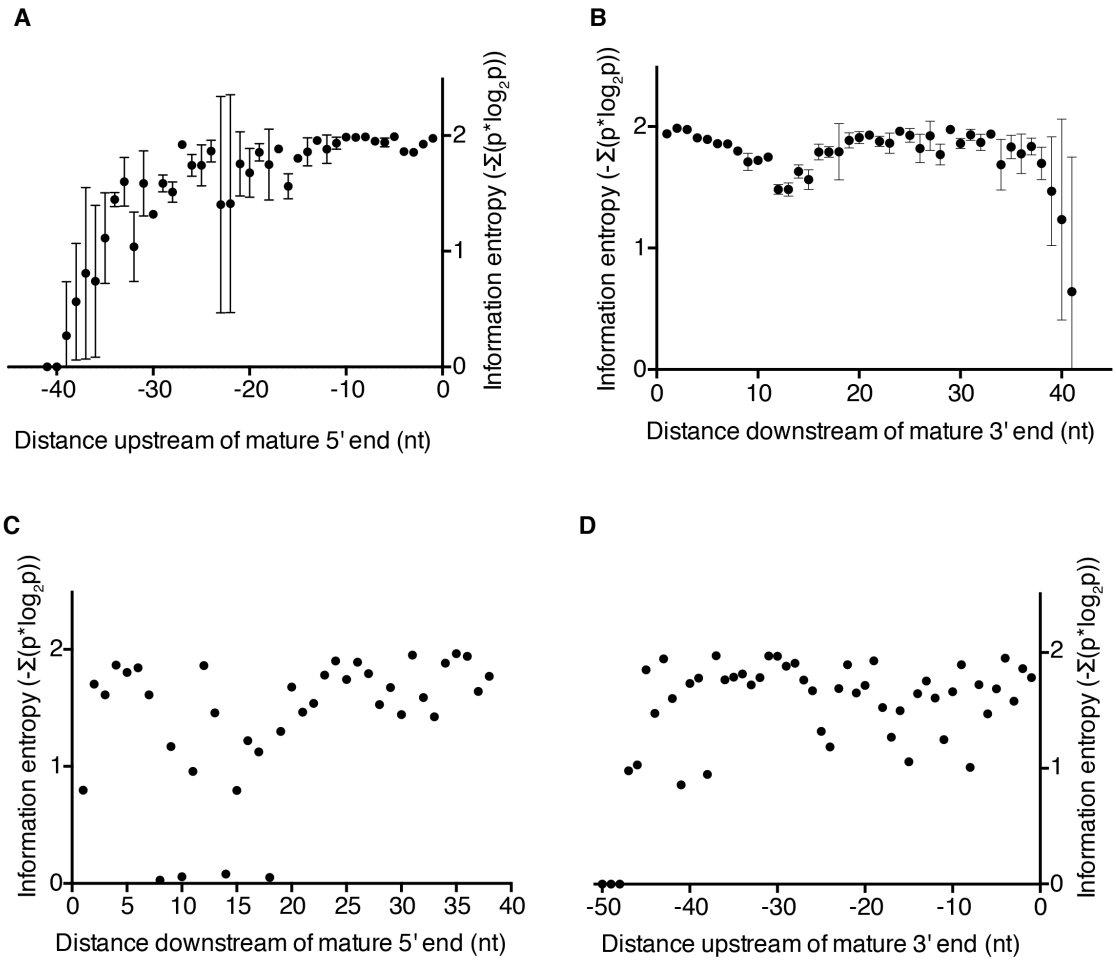


Figure 2.2: Information entropy in pre-tRNA segments and mature body. (A,B) Information entropy $H = -\sum_{i=1}^{n} p(i) * log(p(i))$, (where p is the frequency of each nucleotide at a given position, i, and n the total number of transcripts) was calculated using from hydro-tRNAseq for the 5' leader and 3' trailers of all pre-tRNAs with positions centered at the 5' and 3' ends of mature tRNAs. Pre-tRNA fragments are shorter than 15 nt; hence the drop in signal. (C,D) Same as before, but using the reference sequence of mature tRNAs.

2.3 Pipeline outputs

2.3.1 Mature tRNA alignment

Our pipeline provides individual alignments for every reference transcript included in our curated database. Each alignment is presented in a separate text file (.txt)

that can be surveyed without the requirement of any special analysis or display software, as is the case for conventional mRNA-seq packages.

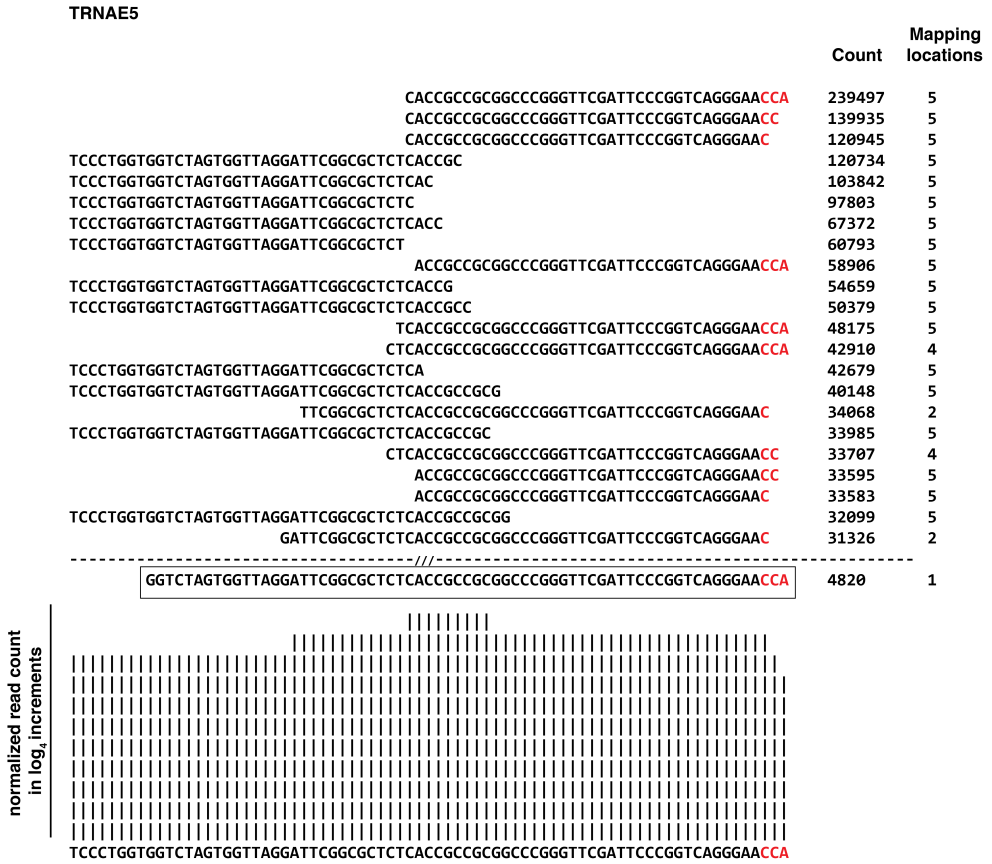


Figure 2.3: Mature tRNA alignment. An indicative alignment to mature TRNAE5 is shown, including read sequences and counts, mapping locations, and $\log_4$ bin-normalized abundance.

Figure 2.3 shows a representative alignment of reads to a mature glutamate-tRNA reference (TRNAE5). The name of the transcript is shown at the top, and the reference transcript sequence at the bottom of the file. Pile-ups of of reads that were mapped to the reference are sorted in descending order of abundance, shown in the 'count' column. Due to the intentional fragmentation of input RNA, we observe that the majority of reads were shorter than the full length tRNA. Thus, longer reads (like the one boxed at the lower part of the alignment) are used to

bridge together separate segments. The total mapping locations for multimapping reads are also indicated. Vertical lines represent the relative frequency of binned, normalized read count in $\log_4$ increments. The 3' CCA aminoacyl acceptor terminus is shown in red indicating the hydro-tRNAseq is successful in freeing the terminus from charged amino acids.

2.3.2 Pre-tRNA alignment

As part of the hierarchical mapping protocol, reads that are not accounted for during the generation of mature tRNA alignments are mapped to our curated pre-tRNA reference sequences, which span 40 nt up- and downstream from all mature tRNA boundaries. **Figure 2.4** shows a representative alignment of reads to a glutamate pre-tRNA reference (TRNAE26), similar to **Fig. 2.3**. The mature sequence is shown in blue, while the leader and trailer sequences in green.

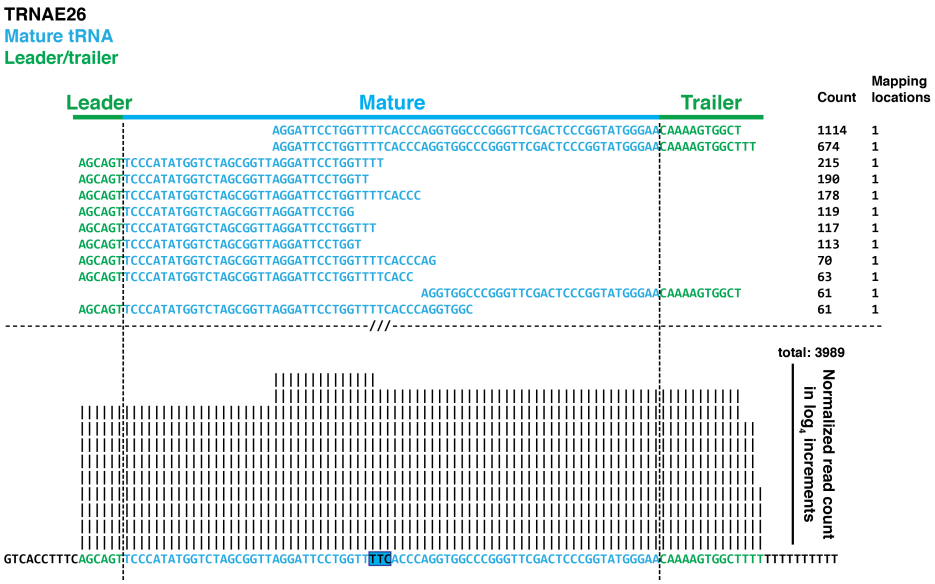


Figure 2.4: Pre-tRNA alignment. An indicative alignment to pre-tRNA TRNAE5 is shown. Mature sequence is in blue and precursor-specific sequence segments in green. The anticodon is boxed in blue for orientation purposes.

The simplicity of the output files of our pipeline renders them easily exploitable

for downstream bioinformatics analysis, necessitating only intermediate programming skills. At the same time the alignment displays can be used directly in figure making. As a matter of fact, simple, and easily customizable Perl and Python scripts were used for all the analysis presented in section 3, showcasing the ease of primary sequence data access and analysis conferred by our approach.

2.4 Need for pre-tRNA enrichment

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 (nuclear or mitochondrial) or pre-tRNAs. Nevertheless, only 1% of total reads comprised sequences overlapping with pre-tRNA leader or trailer sequences (**Fig. 2.5, Table 2.1**). This raised the possibility that we might have missed reads corresponding to lowly expressed or very rapidly processed pre-tRNAs. We controlled for this by

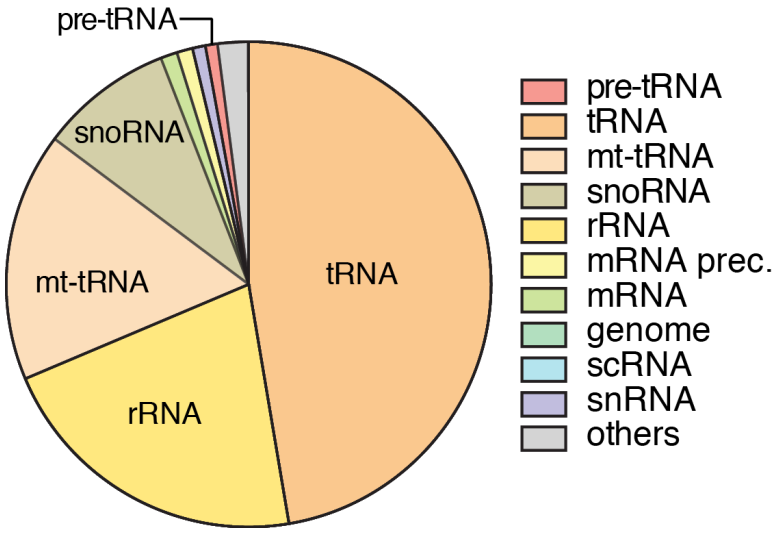


Figure 2.5: Composition of hydro-tRNAseq libraries. Total RNA composition of the 60-100 nt size fraction from hydro-tRNAseq according to RNA classes.

performing photoactivatable-ribonucleoside-enhanced crosslinking and immuno-

precipitation (PAR-CLIP) and sequencing, a technique developed in our lab to identify RNA targets of RNA-binding proteins (RBPs) with high specificity [49].

2.5 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 [50]. Our lab developed 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. 2.6**). To enable efficient RNA-RBP crosslinking using long wavelength UV light, 4-thiouridine (4-SU) 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 tRNA-binding proteins (tRBPs).

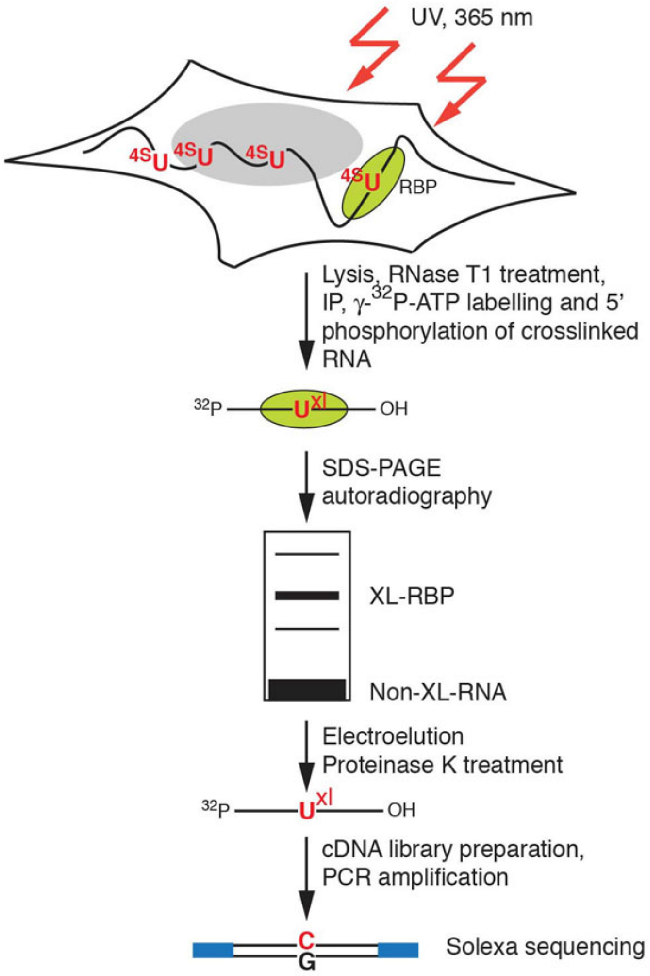


Figure 2.6: PAR-CLIP. Experimental outline of PAR-CLIP methodology, as described previously [49].

2.6 SSB PAR-CLIP

In collaboration with Aitor Garzia, we decided to complement our sequencing efforts with PAR-CLIP-sequencing of the protein Sjögren syndrome type B antigen (SSB), which is also known as Lupus La protein (La) (**Fig. 2.7**). SSB has been shown biochemically to bind to the short 3' oligoU tails [51] that acts as termination signal for POLR3 [52]. Therefore, we reasoned that SSB should bind all tRNA precursors, and that if we could isolate its targets, we would be able to reliably

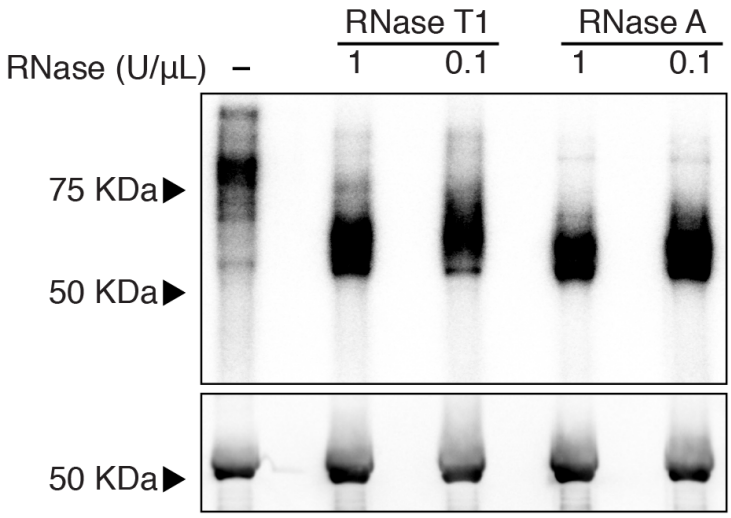


Figure 2.7: SSB crosslinking to RNA. Phosphorimage of SSB-crosslinked to radiolabeled RNA. PAR-CLIP was performed using RNase A or RNase T1, at two different concentrations to account for possible biases of RNase treatment conditions. Libraries from PAR-CLIP using 1 U/μL of RNase A and RNase T1 were prepared and submitted for sequencing. Western blot against HA, shown in the bottom, confirmed the immunoprecipitation of SSB.

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 [53]. 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 latter's 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. 2.8A**), the overwhelming majority of which showed the characteristic T-to-C transition, indicative of crosslinking (**Fig. 2.8B, Table 2.2**).

The vast majority of crosslinking sites in pre-tRNAs were concentrated, as ex-

Type	D0 counts	D1 counts	D2 counts	Total	Percent of total
pre-tRNA	8,249,237	45,310,051	18,719,097	72,278,385	46%
tRNA	8,687,821	16,903,928	7,278,511	32,870,260	21%
rRNA	3,120,687	15,480,966	2,745,431	21,347,084	14%
mRNA gene	1,446,452	4,200,866	10,423,676	16,070,994	10%
mRNA	396,184	2,788,499	7,499,514	10,684,197	7%
scRNA	243,364	1,543,873	305,333	2,092,570	1%
lincRNA gene	81,146	224,548	335,874	641,568	0%
adapter	282,626	90,077	8,183	380,886	0%
snRNA	37,462	131,211	31,302	199,975	0%
snoRNA	11,203	54,179	16,072	81,454	0%
miRNA	3,228	17,382	20,827	41,437	0%
scRNA prec	5,979	23,546	7,876	37,401	0%
piRNA	8,573	1,671	9,102	19,346	0%
plasmid	5,344	4,722	6,301	16,367	0%
mt rRNA	11,006	2,866	979	14,851	0%
mt tRNA	3,356	5,942	2,617	11,915	0%
snoRNA prec	703	5,030	2,362	8,095	0%
rRNA prec	2,076	3,549	1,548	7,173	0%
mt mRNA	1,017	1,783	3,207	6,007	0%
scaRNA	198	2,654	1,708	4,560	0%
marker	1,739	208	27	1,974	0%
lincRNA	698	515	714	1,927	0%
long cali	3	151	662	816	0%
doubtful miRNA	176	140	203	519	0%
snRNA prec	13	66	37	116	0%
scaRNA prec	6	37	7	50	0%
mirtron	2	17	6	25	0%
circRNA	0	0	14	14	0%
std cali	0	1	10	11	0%
Ome cali	0	2	5	7	0%
alt cali	0	0	3	3	0%
doubtful snoRNA	1	0	1	2	0%

Table 2.2: SSB PAR-CLIP reads per RNA type. Reads assigned to each RNA type following hierarchical annotation are shown. Reads with no (d0), one (d1) or two (d2) mismatches compared to reference are shown. Data from replicate using RNase T1 are shown. SSB PAR-CLIP enriches for pre-tRNAs (46%).

pected, in the oligoU tract of the 3' trailer sequence (**Fig. 2.9**). We also found that SSB crosslinked to the 5' segment of the mature tRNA body at conserved sites in the D-stemloop (**Fig. 2.9B**), 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 [53]. The other major target of SSB was

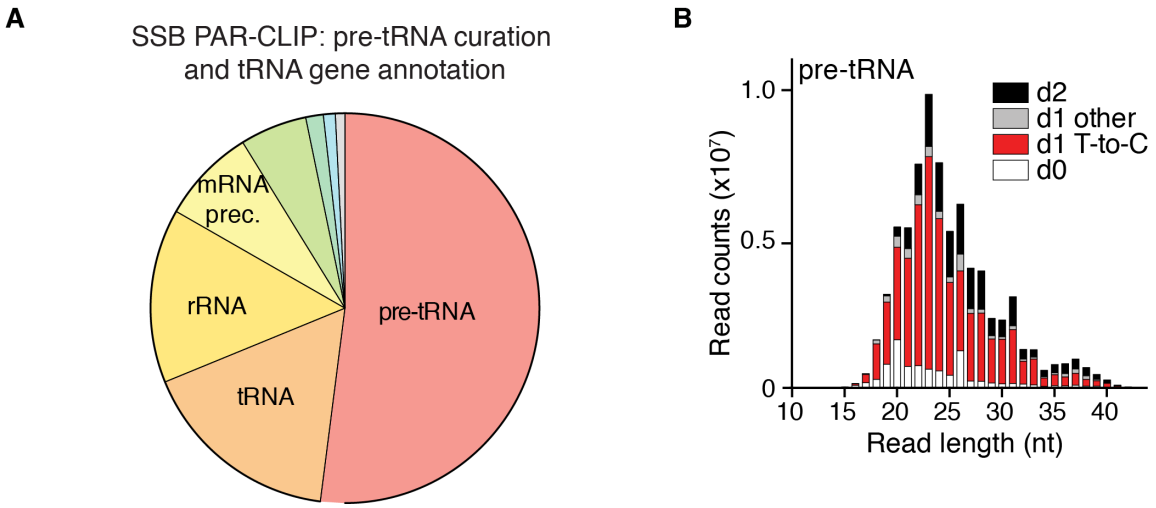


Figure 2.8: SSB binds pre-tRNAs. (A) Assignment of reads from SSB PAR-CLIP (RNase T1, 1 U/μL) to RNA classes. (B) SSB PAR-CLIP reads mapping to tRNA precursors with 0, 1 or 2 mismatches (d0, d1, d2); reads with T-to-C mismatches are separated (red) from the rest of the reads with one mismatch (gray). Read length and number of reads are represented on the x- and y-axis, respectively.

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. 2.10**).

2.7 tRNA gene annotation

I combined hydro-tRNAseq and SSB PAR-CLIP to identify actively transcribed tRNA genes (genomic locations that give rise to a supported pre-tRNAs). I could confidently identify 288 tRNA genes as the intersection of 4 replicates of hydro-tRNAseq (**Fig. 2.11A**), and 349 tRNA genes as the intersection of two SSB PAR-CLIP experiments (**Fig. 2.11B**). Of note, SSB PAR-CLIP confirmed the expression of an additional 7 tRNA genes that were not supported in hydro-tRNAseq replicate (**Fig. 2.11C**), further showcasing the complementarity of the two approaches. There was a strong correlation of pre-tRNA abundances between SSB PAR-CLIP and hydro-tRNAseq (Pearson R = 0.72; **Fig. 2.11D**), providing confidence that

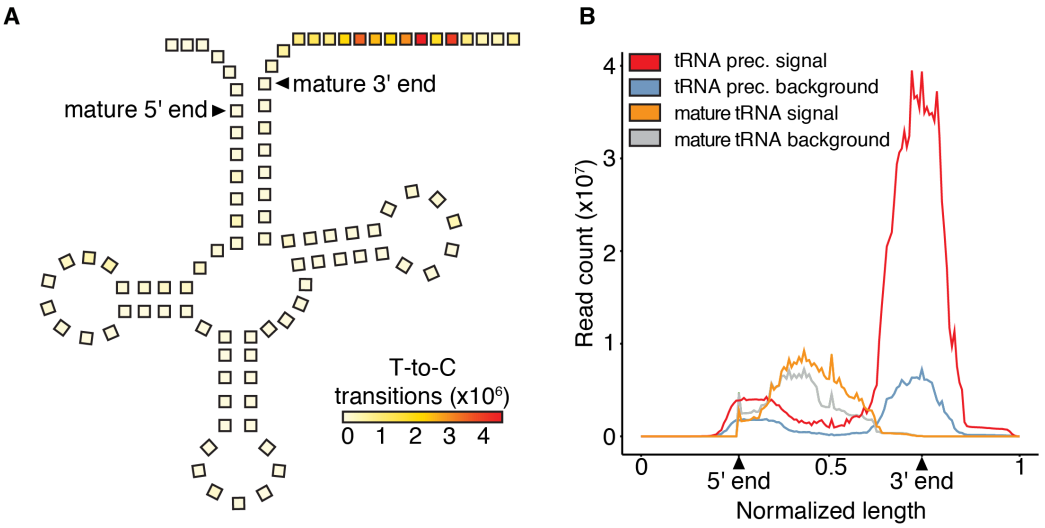


Figure 2.9: SSB binds the 3' oligoU stretch of pre-tRNAs. (A) Positional preference of SSB crosslinking by metagene analysis of all crosslinking events to precursor tRNAs. The incidence of T-to-C transitions is indicated by color intensity. (B) Positional preference of SSB crosslinking by metagene analysis of reads mapped hierarchically; first to mature and then to precursor tRNAs. For each class, PAR-CLIP signal (reads containing T-to-C) and background (reads with no mismatches) are shown. The normalized boundaries of pre-tRNAs (labeled as 0,1) and mature tRNAs (labeled as 5' end, 3' end), and read count are shown on the x- and y-axis, respectively.

SSB PAR-CLIP quantitatively detected pre-tRNAs, without introducing biases (e.g. artificially enriching for lowly expressed pre-tRNAs). The correlation of identified isoacceptor counts between SSB PAR-CLIP and hydro-tRNAseq was virtually perfect (Pearson R = 0.99; **Fig. 2.12A**), ruling out the introduction of a pronounced systematic bias from our hydrolysis-based protocol. Some anticodons seemed to be served by multiple tRNA isodecoders (e.g. 19 isodecoders for $tRNA^{Ser}_{GCA}$), while others only from one (e.g. $tRNA^{Ser}_{ACT}$;**Fig. 2.12B**).Selenocysteine was the only amino acid that, in our data, was decoded by only one tRNA gene.

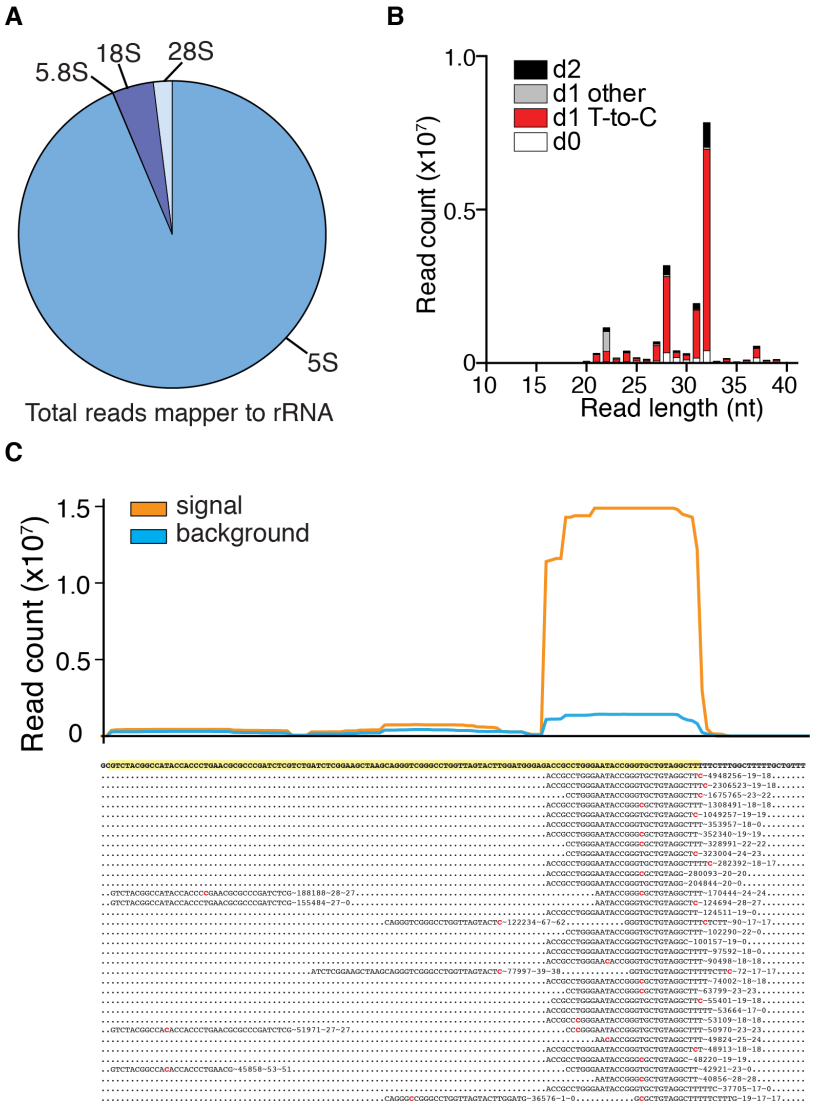


Figure 2.10: SSB binds 5S rRNA. (A) Assignment of reads from SSB PAR-CLIP to rRNAs. (B) Abundance of reads mapped to 5S rRNA with 0, 1 or 2 mismatches (d0, d1, d2) as a function of read length; reads with T-to-C mismatches are separated (red) from the rest of the reads with one mismatch (gray). Read length and number of reads are represented on the x- and y-axis, respectively. (C) Read alignments corresponding to 5S rRNA. Crosslinked positions are shown in red. The read count is shown next to each read sequence, followed by the total mapping positions, and the mapping positions that contain a T-to-C transition. Crosslinked and non-crosslinked read coverage is graphically represented.

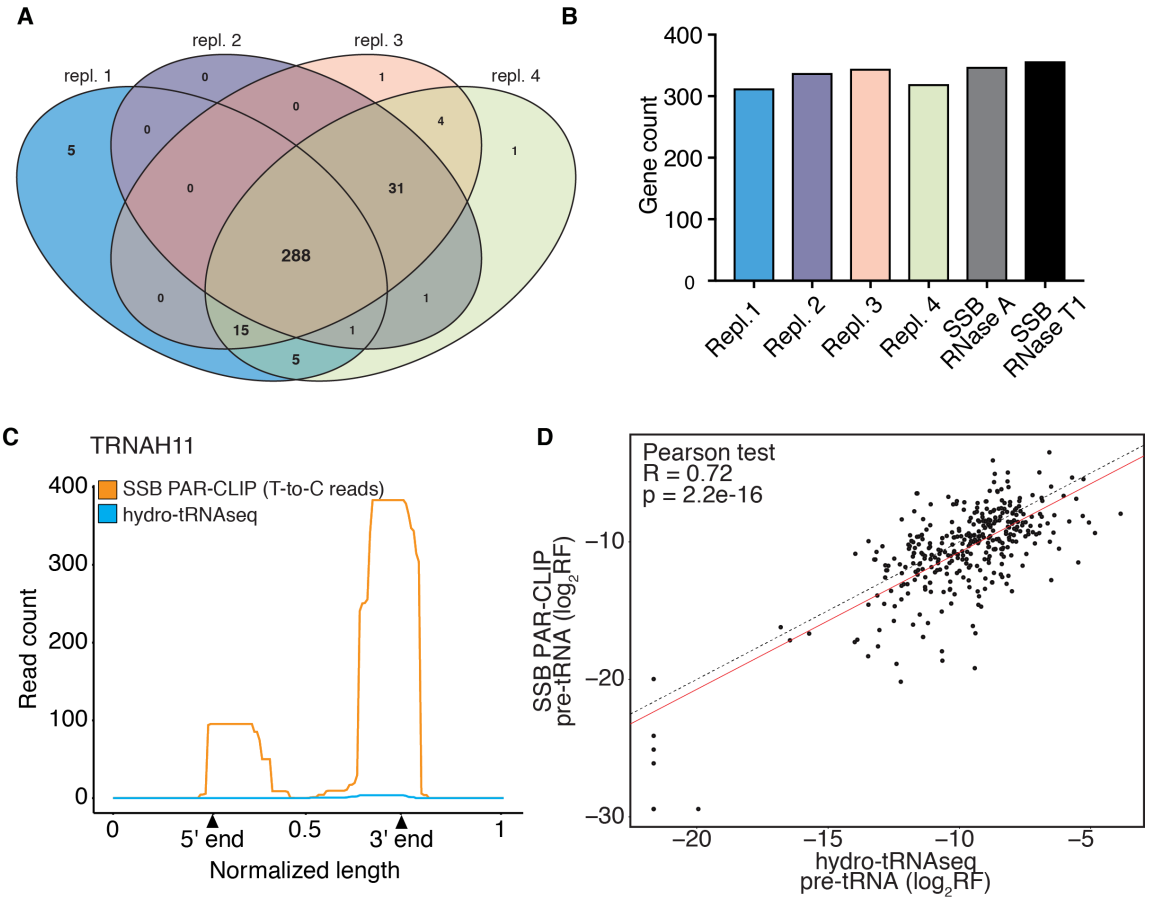


Figure 2.11: tRNA gene annotation. (A) Venn diagram of expressed tRNA genes detected by hydro-tRNAseq in HEK293 cells. Genes with read evidence in both the 5' leader and 3' trailer of the pre-tRNA were counted. (B) Bar chart showing the number of pre-tRNAs detected in each replicate of hydro-tRNAseq and SSB PAR-CLIP. (C) Example of one out of seven tRNA genes that were detected by SSB PAR-CLIP, but were not detected in any of four hydro-tRNAseq replicates. (D) Correlation of relative read frequencies (log2-transformed) of precursor tRNAs between hydro-tRNAseq (x-axis) and SSB PAR-CLIP (y-axis). Correlation was calculated using the Pearson test. Linear fit is shown in red. The y=x line (dotted) is shown for comparison.

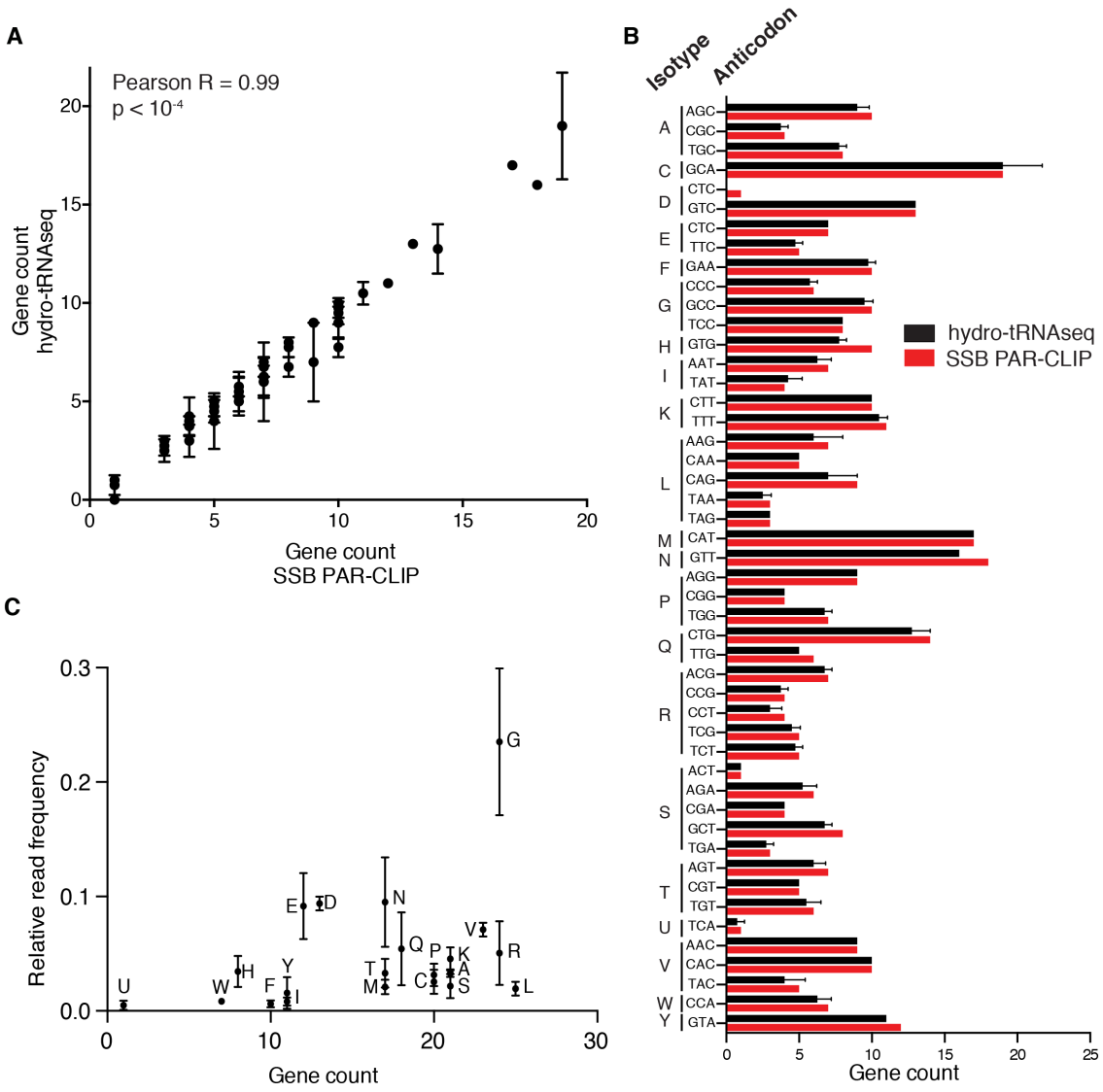


Figure 2.12: Number and relative abundance of tRNA genes per isotype and isoacceptor. (A) Correlation of gene counts for each anticodon detected by hydro-tRNAseq (y-axis) and SSB PAR-CLIP (x-axis). Each dot represents an anticodon shown in (B). Correlation was calculated using the Pearson test. (B) Number of tRNA genes for each anticodon and isotype. Hydro-tRNAseq data are shown in black (mean of 4 replicates; error bars represent standard deviation), SSB PAR-CLIP (RNase T1 treated) in red. (C) Correlation of relative read frequency (y-axis) of mature tRNAs with number of genes (x-axis) per tRNA isotype. Hydro-tRNAseq data; four replicates; mean values shown; error bars represent standard deviation.

Chapter 3

Applications and biological insights

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

Having established the rigor and validity of our experimental and computational method, I set out to address long-standing questions related to tRNA biology. First, I asked whether there is a linear relationship between number of tRNA genes (which will from now on be used interchangeably with pre-tRNAs) and the collective abundance of tRNAs for a given isotype. Due to the absence of reliable tRNAseq datasets, it was assumed that tRNA abundance increased monotonically with increasing tRNA gene count [12, 19, 20].

This proposition hinged on the assumption that tRNAs undergo insignificant transcriptional or post-transcriptional regulation gene regulation, and thus all tRNA genes are transcribed and contribute equally to the mature tRNA abundance. It dismissed, thus, any possible mechanisms of affecting steady state tRNA levels, such as modulation of transcription, processing and degradation.

Indeed, recent data, albeit scant and based on array-type experiments, sug-

gested that the tRNA repertoire can be regulated in a dynamic fashion, and might not even be associated with tumor biology [21]. My anlaysis lent credence to the concept of tRNA regulation. I showed that although tRNA isotypes with higher relative abundances generally tend to have higher tRNA gene numbers, there is no linear correlation between read frequency and gene count (R = 0.12; **Fig. 2.12C**).

3.2 tRNA gene abundance does not correlate with tRNA gene count on the isoacceptor level

The same non-monotonic relationship seems to be true also on the level of tRNA isoacceptors. The relevant data are presented in **Fig. 3.1**. The complexity of the collective data presentation merits a detailed explanation (please note that a reader-friendlier version of the figure is shown in Appendix A). In detail the figure shows:

- tRNA isotypes as headers, using single-letter amino acid codes
- anticodons (representing tRNA isodecoders) on the x-axis
- tRNA gene count (average of four hydro-tRNAseq replicates) on the y-axis
- relative mature tRNA read frequency per anticodon, normalized over all tRNAs (proportional to the area of each disc)

Even though, for example, cysteine-tRNA^{GCA} is the tRNA with the highest gene count, glycine-tRNA^{GCC} is the tRNA with the highest abundance. Also, proline tRNAs are a telling counterexample, whereby the AGG isodecoders are encoded by the largest number of genes, but represent the least abundant set. At the same time, it becomes evident that the tRNA repertoire in HEK293 cells can decode 47

out of the 62 coding codons (61 canonical and one for selenocysteine,TGA) by Watson-Crick basepairing, being thus, dependent on wobble basepairing for the remaining set.

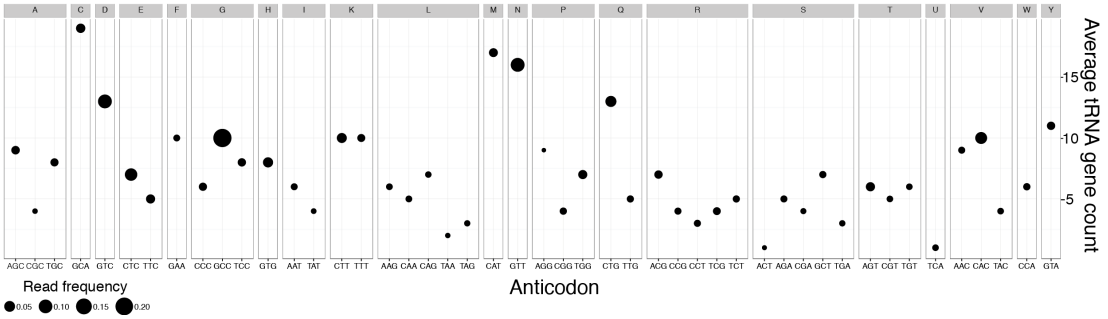


Figure 3.1: Average gene count and relative frequency for each anticodon. The gene count (mean of 4 replicates) is shown on the y-axis, anticodons on the x-axis; isotypes are indicated on the top. The area of each black disc is proportional to the fraction of reads mapping to all mature tRNAs for a given anticodon, normalized over all mature tRNAs for all anticodons.

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

The next question that arose, was whether there is a correlation between individual pre-tRNAs and their mature transcripts. If that were the case, then one could conclude that tRNA maturation is a uniform process across all tRNAs that occurs without significant regulation on individual tRNAs.

Neither of the two techniques I employed showed any strong correlation between precursor and mature tRNA counts ($R < 0.2$; **3.2A,B**). Our techniques can reproducibly quantify pre-tRNAs (**Fig. 2.11D**), and detect accumulation of pre-tRNA processing intermediates (see section 3.7). Therefore, the observed lack of correlation is likely due to marked differences in the kinetics and dynamics of

processing across tRNAs, and/or their pre-tRNA fragments' degradation, rather than due to experimental limitations in capturing processing intermediates.

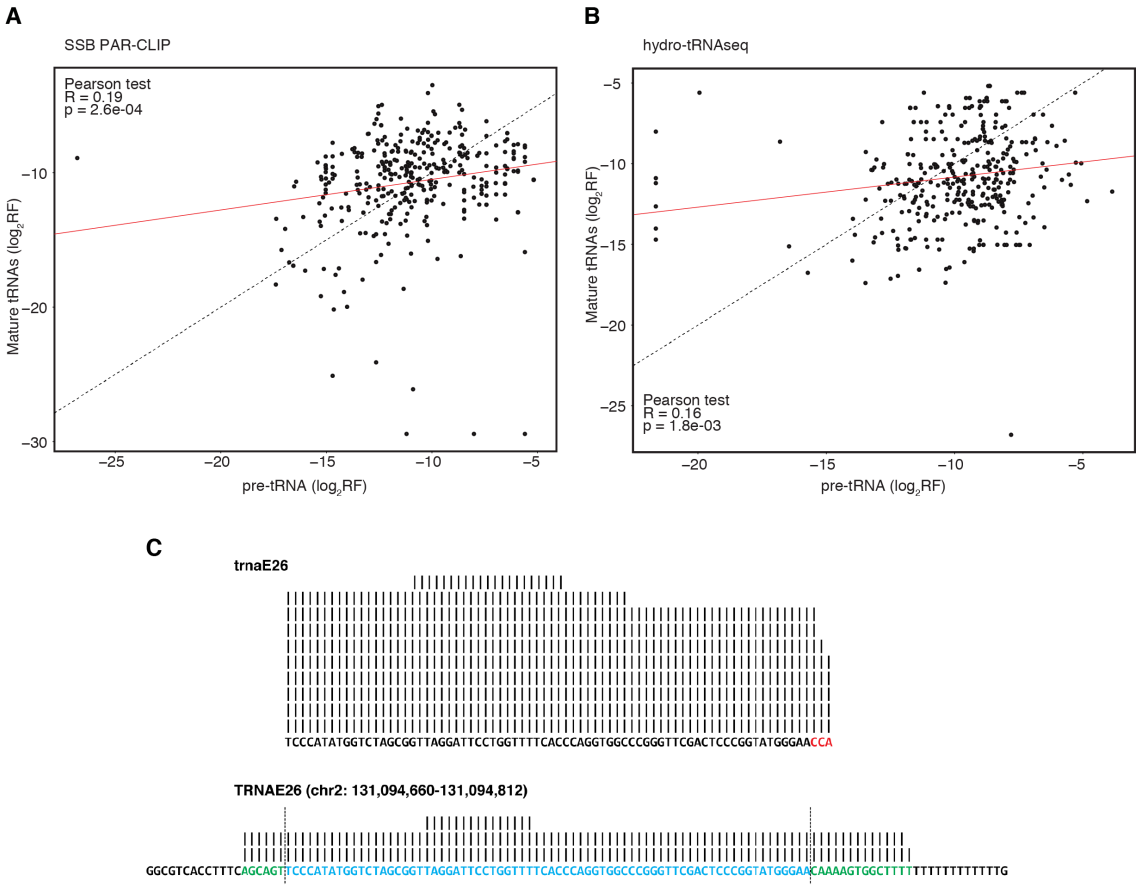


Figure 3.2: Correlation between pre-tRNA and mature tRNA read frequencies. Correlation of relative read frequencies (log2-transformed) between pre-tRNA (x-axis) and mature tRNAs (y-axis) for SSB PAR-CLIP (A) and hydro-tRNAseq (B). Correlation was calculated using the Pearson test. Linear fit is shown in red. The y=x line (dotted) is shown for comparison. (C) Representative alignments of a mature tRNA (top) and its corresponding precursor (bottom). On the pre-tRNA reference, the mature sequence is labeled in blue, and the leader and trailer sequences in green, with their borders demarcated by vertical dotted lines. The post-transcriptionally added CCA tail is shown in red on the mature reference sequence. Vertical bars represent binned, $\log_4$-transformed abundance.

With respect to individual tRNAs, however, mature tRNA read counts were always higher than the respective pre-tRNA counts (**Fig. 3.2C**). This suggests

that tRNA processing is an efficient process across the tRNA spectrum, leading to prompt removal of intermediates and preventing accumulation of fragments that could possibly be toxic to the cells [44].

3.4 tRNA transcription initiation and termination

Besides tRNA gene annotation and quantification, I realized that my approach could yield insights into POLR3 transcription initiation and termination, by paying attention to the length and characteristics of pre-tRNA leader and trailer sequences. To the best of my knowledge, such an analysis had not been previously performed on a transcriptome-wide level in human cells, even though elegant work had yielded valubale insights either *in vitro* or *in silico*, predominantly in yeast [52, 54–58].

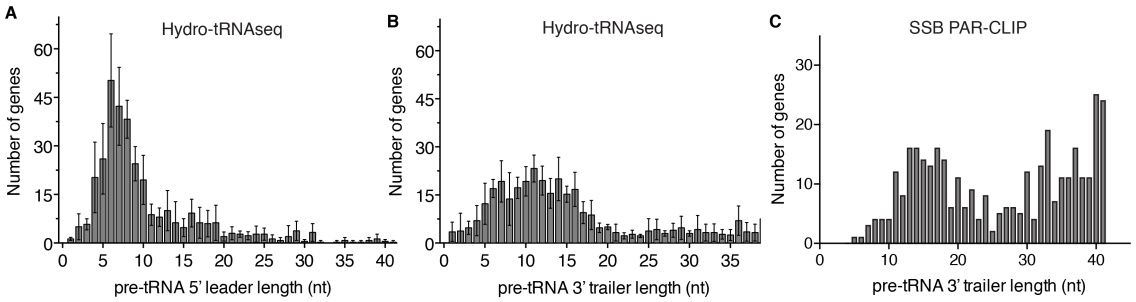


Figure 3.3: Boundaries of tRNA transcription initiation and termination. Histogram of the length distribution of precursor tRNA 5' leaders (A) and 3' trailers (B) detected by hydro-tRNAseq, and trailers detected by SSB PAR-CLIP (C). Mean ± standard deviation of four hydro-tRNAseq replicates shown in (A) and (B). Data from SSB PAR-CLIP replicate with RNase T1 shown in (C).

Based on hydro-tRNAseq, I determined the median 5' leader and 3' trailer lengths to be 6 and 10 nt, respectively, with the trailer lengths showing a broader distribution, cengered around 12 nt (**Fig. 3.3A,B**). Interestingly, SSB PAR-CLIP revealed a subset of much longer trailers (**Fig. 3.3C**), suggesting that it 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 [34].

I 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 [55, 57]. SSB protected consistently a 4 to 5 nt oligoU stretch, which was also confirmed by hydro-tRNAseq (**Fig. 3.4**). This is in agreement with previous *in vitro* results [51, 53, 59]. I also addressed the proposed requirement for a stem-loop immediately upstream of the oligoU termination signal [55]. 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. 3.5**). 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 [57].

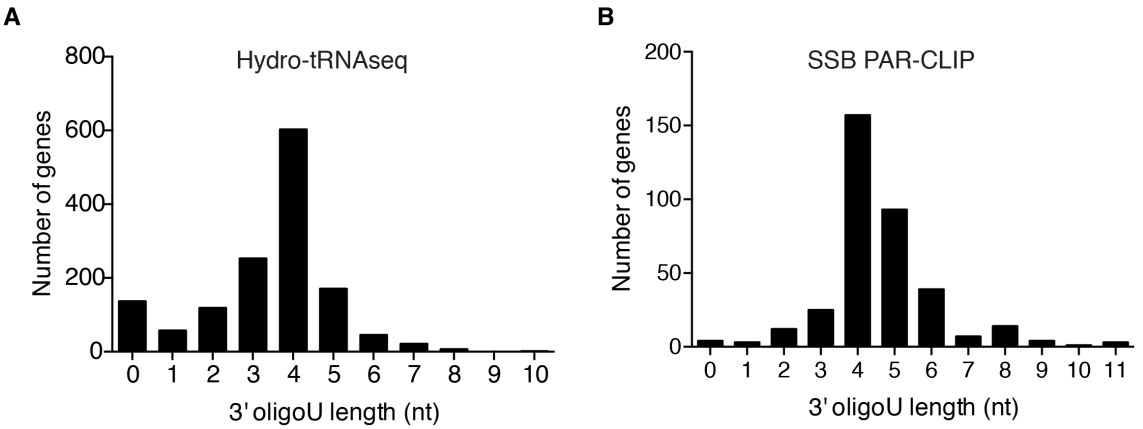


Figure 3.4: POLR3 oligoU transcription termination signals. Histogram of the length distribution for the longest oligoU tract per pre-tRNA 3' trailer detected by hydro-tRNAseq (A) (aggregate gene total of all replicates shown on the y-axis), and by SSB PAR-CLIP (B).

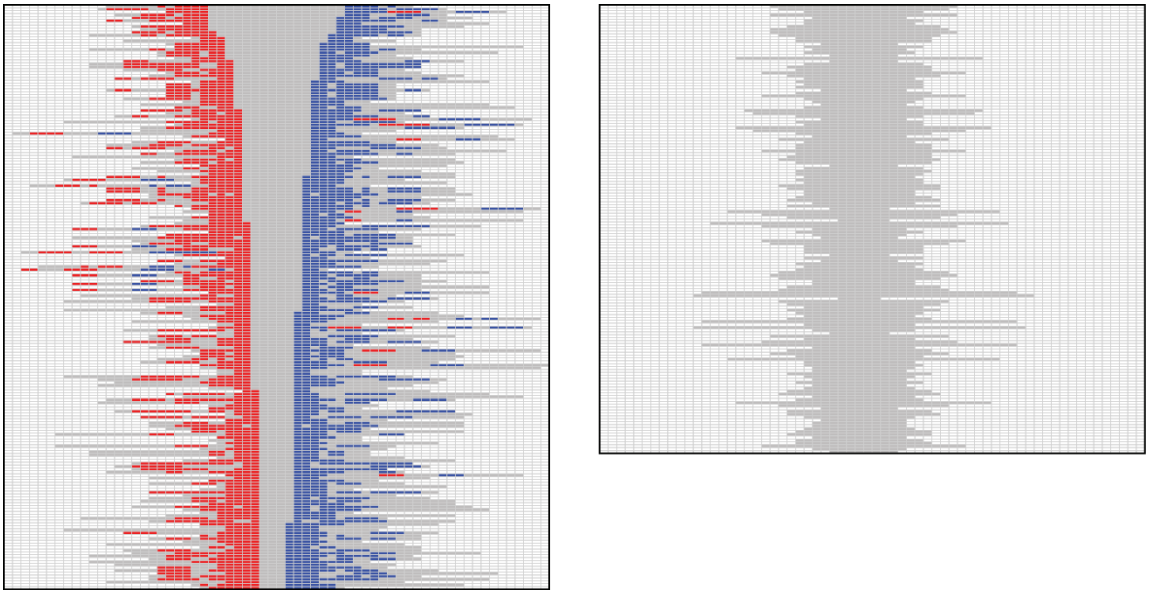


Figure 3.5: Predicted structures of precursor tRNA 3' trailers with read evidence in SSB PAR-CLIP. Every colored box indicates one nucleotide. Red and blue reflect the 5' and 3' parts of a duplex, respectively, whereas gray indicates non-basepaired nucleotides. About 1/2 of all tRNA trailers lack any predicted structure (right panel). The predicted stemloops do not have a uniform stem or loop length. Folding was performed using the Vienna folding package [60].

3.5 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, the compilation of my tRNA reference set, included the combination of all frequent mismatch signatures in all heavily modified positions. I reported the most frequently modified positions per tRNA gene (**Table 3.1**), and computed the frequencies of every nucleoside change per position across all tRNA genes (the most common ones shown in **Fig. 3.6**).

tRNA	Position	Nucleoside	Read count
trnaA10	36	G	25,957
trnaA10	36	T	1,198
trnaA10	36	A	2,206
trnaA10	36	C	51
trnaA11	36	A	100
trnaA11	36	T	61
trnaA11	36	G	95
trnaA13	6	C	2,507
trnaA13	6	T	175,171
trnaA13	6	A	1,178
trnaA13	6	G	6,873
trnaA13	10	G	185,402
trnaA13	10	T	39
trnaA13	10	A	829
trnaA13	10	C	136
trnaA13	16	C	1,134
trnaA13	16	T	189,165
trnaA13	16	G	1,038
trnaA13	33	C	112
trnaA13	33	A	386
trnaA13	33	T	110
trnaA13	33	G	205,885
trnaA13	36	C	559
trnaA13	36	T	12,584
trnaA13	36	A	4,747
trnaA13	36	G	173,085
trnaA13	39	A	30
trnaA13	39	T	3,993
trnaA13	39	C	172,312
trnaA13	45	T	319
trnaA13	45	C	1,202
trnaA13	45	G	65,025
trnaA13	56	G	126
trnaA13	56	T	7
trnaA13	56	A	36,136
trnaA13	56	C	40
trnaA13	58	A	13
trnaA13	58	T	36,111
trnaA13	58	C	117
trnaA13	58	G	6
trnaA13	66	T	62
trnaA13	66	A	32,876
trnaA13	66	C	684
trnaA2	34	G	33,023
trnaA2	34	A	206
trnaA2	34	T	11
trnaA2	34	C	9
trnaA2	37	A	1,135
trnaA2	37	T	9,316
trnaA2	37	C	250
trnaA2	37	G	23,009
trnaA23	5	G	12,936
trnaA23	5	C	593
trnaA23	5	T	250
trnaA23	5	A	277,684
trnaA23	33	C	206,326
trnaA23	33	T	215,944
trnaA23	33	A	126
trnaA23	36	A	9,456
trnaA23	36	T	18,956
trnaA23	36	G	415,552
trnaA23	46	G	992
trnaA23	46	A	366
trnaA23	46	T	102,084
trnaA23	46	C	145,919

Continued on next column

Continued from previous column

tRNA	Position	Nucleoside	Read count
trnaA23	67	C	2,990
trnaA23	67	T	145,047
trnaA23	67	A	16
trnaA28	33	A	98
trnaA28	33	T	50
trnaA28	33	C	1,180
trnaA28	33	G	38,349
trnaA28	36	G	26,060
trnaA28	36	C	54
trnaA28	36	A	922
trnaA28	36	T	1,098
trnaA28	45	T	478
trnaA28	45	A	19
trnaA28	45	C	21
trnaA28	45	G	7,579
trnaA3	34	C	25
trnaA3	34	A	205
trnaA3	34	T	26
trnaA3	34	G	101,158
trnaA3	37	G	79,503
trnaA3	37	C	673
trnaA3	37	T	28,670
trnaA3	37	A	2,886
trnaA32	36	C	155
trnaA32	36	A	107,454
trnaA32	36	T	254
trnaA32	36	G	7,740
trnaA32	45	G	74,872
trnaA32	45	C	3,688
trnaA32	45	A	75
trnaA32	45	T	1,610
trnaA34	33	G	33,690
trnaA34	33	C	15
trnaA34	33	A	219
trnaA34	33	T	20
trnaA34	36	G	27,770
trnaA34	36	T	2,594
trnaA34	36	A	1,468
trnaA34	36	C	128
trnaA6	33	G	2,435
trnaA6	33	C	7
trnaA6	33	T	6
trnaA6	33	A	5,790
trnaA6	36	G	3,318
trnaA6	36	C	31
trnaA6	36	A	4,344
trnaA6	36	T	380
trnaA7	33	A	1,183
trnaA7	33	T	1,892
trnaA7	33	C	1,117
trnaA7	33	G	27,108
trnaA7	36	G	20,964
trnaA7	36	T	12,246
trnaA7	36	A	3,841
trnaA7	36	C	594
trnaC10	6	G	15,648
trnaC10	6	C	101,951
trnaC10	6	T	2,423
trnaC10	6	A	5,244
trnaC10	16	G	26,551
trnaC10	16	C	2,047
trnaC10	16	A	1,356
trnaC10	16	T	117,667
trnaC11	16	C	10
trnaC11	16	T	783
trnaC11	16	G	4,230

Continued on next column

Continued from previous column

tRNA	Position	Nucleoside	Read count
trnaC11	44	A	82
trnaC11	44	T	10
trnaC11	44	G	803
trnaC14	6	G	4,417
trnaC14	6	C	2,002
trnaC14	6	A	762
trnaC14	6	T	92,027
trnaC14	57	A	47,747
trnaC14	57	T	5,006
trnaC14	57	C	391
trnaC14	57	G	6,331
trnaC21	42	G	33,541
trnaC21	42	C	73
trnaC21	42	A	7,594
trnaC21	42	T	181
trnaC24	6	G	5,664
trnaC24	6	T	95,339
trnaC24	6	A	1,284
trnaC24	6	C	16,078
trnaC24	57	T	9,354
trnaC24	57	A	139,890
trnaC24	57	C	339
trnaC24	57	G	9,511
trnaC7	57	G	1,048
trnaC7	57	T	931
trnaC7	57	A	25,533
trnaC7	57	C	20
trnaD10	9	G	813
trnaD10	9	T	304
trnaD10	9	A	118,147
trnaD10	9	C	89
trnaD10	53	C	19,831
trnaD10	53	A	1,716
trnaD10	53	T	472,250
trnaD10	53	G	45,707
trnaD10	56	G	367,491
trnaD10	56	C	1,093
trnaD10	56	T	846
trnaD10	56	A	144,309
trnaD10	57	G	2,914
trnaD10	57	T	16,380
trnaD10	57	A	492,772
trnaD10	57	C	791
trnaD9	24	C	468
trnaD9	24	A	4,669
trnaD9	24	T	1,114
trnaD9	24	G	2,101
trnaD9	31	C	91,720
trnaD9	31	T	2,596
trnaD9	31	A	953
trnaD9	31	G	2,238
trnaD9	56	A	114,699
trnaD9	56	T	609
trnaD9	56	C	858
trnaD9	56	G	1,051,388
trnaD9	57	G	9,918
trnaD9	57	A	1,095,714
trnaD9	57	T	59,596
trnaD9	57	C	2,145
trnaE5	9	G	913,314
trnaE5	9	A	1,503
trnaE5	9	T	1,502
trnaE5	9	C	1,330
trnaE5	19	G	668
trnaE5	19	A	198
trnaE5	19	T	1,171,172

Continued on next column

Continued from previous column

tRNA	Position	Nucleoside	Read count
trnaE5	19	C	3,141
trnaE5	24	G	18,086
trnaE5	24	A	1,368,739
trnaE5	24	T	2,320
trnaE5	24	C	2,252
trnaE5	27	G	4,750
trnaE5	27	T	27,560
trnaE5	27	A	3,180
trnaE5	27	C	1,490,638
trnaE5	34	C	1,948,790
trnaE5	34	A	1,036
trnaE5	34	T	45,989
trnaE5	34	G	1,246
trnaE5	53	G	46,498
trnaE5	53	A	3,717
trnaE5	53	T	1,971,842
trnaE5	53	C	16,977
trnaE5	57	G	27,513
trnaE5	57	C	12,646
trnaE5	57	A	1,784,527
trnaE5	57	T	163,808
trnaE7	6	G	442
trnaE7	6	T	224,221
trnaE7	6	A	428
trnaE7	6	C	622,013
trnaE7	9	G	925,734
trnaE7	9	A	1,682
trnaE7	9	T	3,498
trnaE7	9	C	1,570
trnaE7	34	G	7,239
trnaE7	34	C	63,597
trnaE7	34	T	1,701,031
trnaE7	34	A	5,025
trnaE7	44	A	3,111
trnaE7	44	T	209,800
trnaE7	44	C	1,445,839
trnaE7	44	G	8,730
trnaE7	66	C	36
trnaE7	66	T	1,122
trnaE7	66	A	154,605
trnaE7	66	G	1,018,843
trnaF5	6	A	1,703
trnaF5	6	C	36
trnaF5	6	G	1,587
trnaF5	47	G	783
trnaF5	47	T	17,158
trnaF5	47	C	2,431
trnaF5	57	A	3,199
trnaF5	57	G	16,069
trnaG15	5	G	983,413
trnaG15	5	T	477,119
trnaG15	5	A	2,009
trnaG15	5	C	5,058
trnaG15	45	G	1,455,904
trnaG15	45	T	3,715
trnaG15	45	A	131,627
trnaG15	45	C	10,529
trnaG15	56	C	2,347
trnaG15	56	T	137,083
trnaG15	56	A	802,531
trnaG15	56	G	23,277
trnaG15	66	G	13
trnaG15	66	A	556,240
trnaG15	66	C	353,038
trnaH11	5	A	687
trnaH11	5	T	1,239

Continued on next column

Continued from previous column

tRNA	Position	Nucleoside	Read count
trnaH11	5	C	243
trnaH11	5	G	9,794
trnaH11	33	A	729
trnaH11	33	T	897,345
trnaH11	33	C	5,452
trnaH11	33	G	8,216
trnaH11	58	G	20,143
trnaH11	58	C	260
trnaH11	58	T	27,065
trnaH11	58	A	870,683
trnaI5	27	G	2,177
trnaI5	27	T	1,830
trnaI5	35	G	32,621
trnaI5	35	A	107
trnaI5	60	A	3,085
trnaI5	60	T	36,960
trnaI5	69	G	36,009
trnaI5	69	T	2,962
trnaI9	27	T	2,220
trnaI9	27	A	90
trnaI9	27	C	79
trnaI9	27	G	18,350
trnaI9	35	G	54,591
trnaI9	35	T	7
trnaI9	35	A	189
trnaI9	59	G	1,438
trnaI9	59	C	229
trnaI9	59	T	3,282
trnaI9	59	A	43,060
trnaI9	60	G	9
trnaI9	60	C	27
trnaI9	60	A	3,213
trnaI9	60	T	44,664
trnaI9	69	T	5,705
trnaI9	69	C	92
trnaI9	69	G	38,915
trnaK1	54	G	12,329
trnaK1	54	T	529,905
trnaK1	54	A	2,946
trnaK1	54	C	7,353
trnaK1	58	C	1,742
trnaK1	58	T	22,380
trnaK1	58	A	435,401
trnaK1	58	G	81,542
trnaL1	27	G	52,413
trnaL1	27	T	4,682
trnaL1	27	A	2,091
trnaL1	27	C	717
trnaL1	35	C	21
trnaL1	35	T	2,285
trnaL1	35	A	546
trnaL1	35	G	68,859
trnaL1	67	G	1,793
trnaL1	67	C	221
trnaL1	67	A	55,644
trnaL1	67	T	2,456
trnaL15	10	G	12,159
trnaL15	10	T	13
trnaL15	10	A	12
trnaL15	10	C	863
trnaL15	12	G	5
trnaL15	12	T	647
trnaL15	12	A	2
trnaL15	12	C	12,503
trnaL15	27	C	80
trnaL15	27	A	316

Continued on next column

Continued from previous column

tRNA	Position	Nucleoside	Read count
trnaL15	27	T	720
trnaL15	27	G	12,945
trnaL4	27	G	21,789
trnaL4	27	T	812
trnaL4	35	G	21,913
trnaL4	35	A	202
trnaL4	67	T	206
trnaL4	67	A	3,302
trnaL4	67	G	176
trnaL5	51	G	4,824
trnaL5	51	C	24,212
trnaL5	51	T	19,749
trnaL5	51	A	2,548
trnaM1	9	G	85,934
trnaM1	9	C	6,935
trnaM1	9	A	5,666
trnaM1	9	T	23,834
trnaM1	57	A	312,240
trnaM1	57	T	18,020
trnaM1	57	C	1,202
trnaM1	57	G	12,797
trnaM13	20	G	779
trnaM13	20	T	20,233
trnaM13	20	A	393
trnaM13	20	C	2,235
trnaM13	26	A	802
trnaM13	26	T	3,204
trnaM13	26	C	113
trnaM13	26	G	34,706
trnaM13	58	G	6,039
trnaM13	58	A	372,150
trnaM13	58	T	7,682
trnaM13	58	C	894
trnaM13	59	G	141,279
trnaM13	59	C	1,112
trnaM13	59	T	243,802
trnaM13	59	A	273
trnaN1	27	C	5,696
trnaN1	27	A	5,664
trnaN1	27	T	133,510
trnaN1	27	G	250,204
trnaN1	59	T	79,049
trnaN1	59	A	827,307
trnaN1	59	C	4,741
trnaN1	59	G	127,829
trnaN14	27	G	102,061
trnaN14	27	A	4,021
trnaN14	27	T	21,768
trnaN14	27	C	1,392
trnaN14	59	G	7,434
trnaN14	59	C	706
trnaN14	59	A	151,684
trnaN14	59	T	9,968
trnaP10	19	C	4,621
trnaP10	19	A	2,631
trnaP10	19	T	411,265
trnaP10	19	G	45,430
trnaP10	33	C	183,783
trnaP10	33	T	89,056
trnaP10	33	A	1,332
trnaP10	33	G	461,368
trnaP10	39	G	3,517
trnaP10	39	T	81,988
trnaP10	39	A	972
trnaP10	39	C	592,210
trnaQ1	56	C	1,480

Continued on next column

Continued from previous column

tRNA	Position	Nucleoside	Read count
trnaQ1	56	A	1,768,755
trnaQ1	56	T	617
trnaQ1	56	G	8,734
trnaQ1	57	G	29,151
trnaQ1	57	T	4,802
trnaQ1	57	A	1,692,736
trnaQ1	57	C	2,268
trnaQ1	58	G	23,220
trnaQ1	58	T	1,535
trnaQ1	58	A	1,685,855
trnaQ1	58	C	7,182
trnaQ2	32	A	147
trnaQ2	32	T	391
trnaQ2	32	C	220,622
trnaQ2	32	G	38,543
trnaR10	19	G	880
trnaR10	19	C	25
trnaR10	19	T	77
trnaR10	19	A	66,484
trnaR10	34	T	42
trnaR10	34	A	4,136
trnaR10	34	C	51
trnaR10	34	G	115,152
trnaR10	58	G	4,908
trnaR10	58	A	68,405
trnaR10	58	T	2,439
trnaR10	58	C	227
trnaR2	19	G	6,415
trnaR2	19	T	133
trnaR2	19	A	122,886
trnaR2	19	C	97
trnaR2	34	G	238,519
trnaR2	34	C	203
trnaR2	34	A	6,830
trnaR2	34	T	160
trnaR2	58	G	11,883
trnaR2	58	A	152,011
trnaR2	58	T	4,470
trnaR2	58	C	653
trnaR22	58	C	1,653
trnaR22	58	T	28,134
trnaR22	58	A	325,452
trnaR22	58	G	53,856
trnaR25	32	C	164,623
trnaR25	32	T	25,406
trnaR25	32	A	171
trnaR25	32	G	161
trnaR25	34	G	78
trnaR25	34	C	41,819
trnaR25	34	T	187,213
trnaR25	34	A	115
trnaR25	58	C	1,903
trnaR25	58	A	242,429
trnaR25	58	T	29,879
trnaR25	58	G	51,224
trnaR4	37	C	2,958
trnaR4	37	T	10,499
trnaR4	37	A	2,141
trnaR4	37	G	384,444
trnaR4	58	G	54,001
trnaR4	58	A	392,684
trnaR4	58	T	30,994
trnaR4	58	C	1,602
trnaR7	58	A	150,051
trnaR7	58	T	5,083
trnaR7	58	C	537

Continued on next column

Continued from previous column

tRNA	Position	Nucleoside	Read count
trnaR7	58	G	3,895
trnaS16	52	C	6,033
trnaS16	52	A	2,966
trnaS16	52	G	5,330
trnaT14	35	G	87,515
trnaT14	35	C	513
trnaT14	35	T	326
trnaT14	35	A	159
trnaT14	59	G	3,635
trnaT14	59	T	10,441
trnaT14	59	A	108,722
trnaT14	59	C	746
trnaT2	26	C	209
trnaT2	26	A	16,122
trnaT2	26	T	20
trnaT2	26	G	61
trnaT2	58	G	39
trnaT2	58	T	833
trnaT2	58	A	5,715
trnaT2	58	C	18
trnaT21	35	A	19
trnaT21	35	T	490
trnaT21	35	C	48
trnaT21	35	G	9,748
trnaT21	59	A	10,638
trnaT21	59	T	240
trnaT21	59	C	59
trnaT21	59	G	1,403
trnaT4	35	C	538
trnaT4	35	T	2,319
trnaT4	35	A	472
trnaT4	35	G	210,546
trnaT4	59	G	18,477
trnaT4	59	C	1,524
trnaT4	59	T	21,160
trnaT4	59	A	325,542
trnaT4	70	G	3,512
trnaT4	70	C	4,114
trnaT4	70	A	1,371
trnaT4	70	T	345,239
trnaU1	28	C	39
trnaU1	28	A	15
trnaU1	28	T	3,002
trnaU1	28	G	97
trnaU1	64	A	61
trnaU1	64	T	7,513
trnaU1	64	C	112
trnaU1	64	G	385
trnaU1	71	T	706
trnaU1	71	A	6,359
trnaU1	71	C	10
trnaU1	71	G	67
trnaV12	34	G	10,863
trnaV12	34	C	11,001
trnaV12	34	A	122
trnaV12	34	T	221
trnaV12	67	G	59,792
trnaV12	67	A	1,838
trnaV12	67	T	344
trnaV12	67	C	376
trnaV2	19	G	4,257
trnaV2	19	C	10,750
trnaV2	19	A	1,121
trnaV2	19	T	889,930
trnaV2	20	G	202
trnaV2	20	C	361,327

Continued on next column

Continued from previous column

tRNA	Position	Nucleoside	Read count
trnaV2	20	T	546,333
trnaV2	20	A	678
trnaV2	26	G	873,859
trnaV2	26	A	5,271
trnaV2	26	T	46,502
trnaV2	26	C	7,245
trnaV2	34	C	445,948
trnaV2	34	A	1,387
trnaV2	34	T	2,668
trnaV2	34	G	948,953
trnaV2	58	G	67,711
trnaV2	58	A	946,402
trnaV2	58	T	101,713
trnaV2	58	C	5,446
trnaV27	26	G	416,151
trnaV27	26	C	3,461
trnaV27	26	T	18,759
trnaV27	26	A	1,627
trnaV27	58	G	13,799
trnaV27	58	C	1,270
trnaV27	58	T	21,819
trnaV27	58	A	168,700
trnaV8	34	A	471
trnaV8	34	T	525
trnaV8	34	C	47,407
trnaV8	34	G	84,016
trnaW2	70	C	708
trnaW2	70	A	165
trnaW2	70	T	12,759
trnaW2	70	G	1,040
trnaW5	70	G	2,127
trnaW5	70	C	1,492
trnaW5	70	T	71,434
trnaW5	70	A	320
trnaW7	57	G	2,443
trnaW7	57	C	49
trnaW7	57	A	65,298
trnaW7	57	T	331
trnaW7	70	C	1,174
trnaW7	70	T	58,462
trnaW7	70	A	291
trnaW7	70	G	2,288
trnaY12	26	G	1,265
trnaY12	26	T	131
trnaY2	26	G	1,003
trnaY2	26	T	85
trnaY2	26	C	11
trnaY2	37	G	10,310
trnaY2	37	C	365
trnaY2	37	A	37
trnaY2	37	T	461
trnaY4	26	G	8,449
trnaY4	26	A	102
trnaY4	26	T	380
trnaY4	26	C	41
trnaY4	37	T	540
trnaY4	37	A	8
trnaY4	37	C	1,197
trnaY4	37	G	22,102
trnaY4	58	G	3,462
trnaY4	58	A	26,513
trnaY4	58	T	14,753
trnaY4	58	C	190
trnaY4	59	C	98
trnaY4	59	A	9,871
trnaY4	59	T	34,855

Continued on next column

Continued from previous column

tRNA	Position	Nucleoside	Read count
trnaY4	59	G	81
trnaY8	26	G	1,450
trnaY8	26	C	32
trnaY8	26	A	45
trnaY8	26	T	122
trnaY8	51	G	5
trnaY8	51	A	14
trnaY8	51	T	17,882
trnaY8	51	C	2,101
trnaY8	58	G	610
trnaY8	58	A	17,992
trnaY8	58	T	1,254

Table 3.1: Positions resulting to modifications-induced mismatches. tRNA transcripts, the positions where modifications result in mismatches, and the read count corresponding to every observed nucleoside per position are shown.

The majority of editing events were A-to-G transitions at the first position of the anticodon (position 34) and at the position 3' to the anticodon (position 37). Both positions are known to be heavily modified, the former being deaminated to inosine, and the latter further modified (e.g. 1-methylinosine) [61]. In my data the majority of the reads that mapped to the anticodon of the modified tRNAs contained mismatches. To a lesser extent I 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 [61].

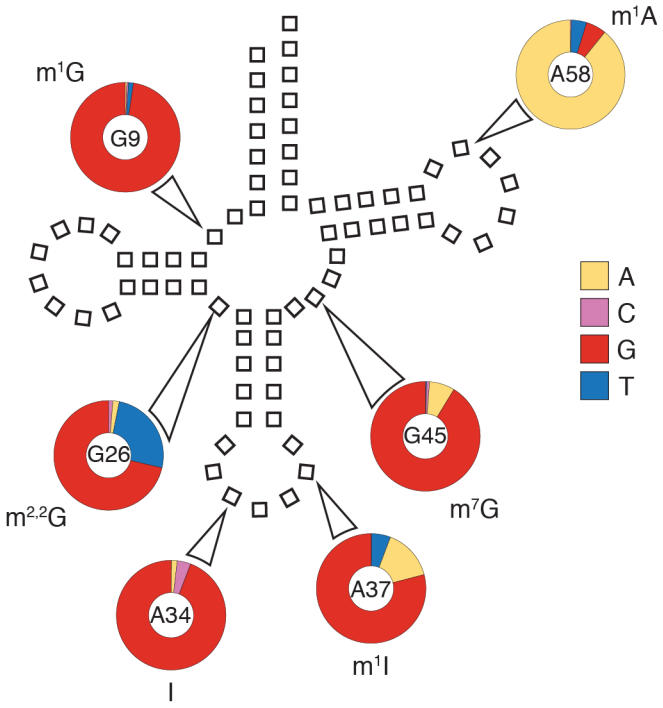


Figure 3.6: tRNA modifications detected by hydro-tRNAseq. The positions that resulted in the most common mismatches over all tRNAs, along with the reference nucleotide and the most likely modification are indicated in the center of each ring. The relative frequency of each returned nucleoside is proportional to the corresponding color-coded area of the ring (m^1G: 1-methylguanosine; $m^{2,2}G$: N2,N2-dimethylguanosine; I: inosine; m^1I: 1-methylinosine; m^7G: 7-methylguanosine; m^1A: 1-methyladenosine).

The temporal resolution of tRNA modifications by RNAseq has begun to be ad-

dressed recently [62], however at a single modification level (inosine 34), and by using libraries relative poor in tRNA reads (<1% of total reads). I was appropriately poised to address this issue since my very deep sequencing set, in combination with my hierarchical annotation pipeline, offered the advantage of dissecting multiple modifications simultaneously.

I focused on inosine editing events, since they represented the majority of modified nucleosides. By inspecting read alignments with error distance 1 to the reference pre-tRNA, I noticed A-to-G transition mismatches at position 34 in reads that retained the leader and trailer sequences of the precursor tRNA (**Fig. 3.7, top**). This confirmed that A34 deamination takes place at the precursor level, and therefore is a nuclear modification, as it has been previously reported [62].

Next, I 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. 3.7, bottom**).

All in all, far from being a comprehensive analysis of tRNA editing and modifications, the work presented here outlines how a careful examination and analysis of deep hydro-tRNAseq datasets can yield valuable insights in the and type and temporal occurrence of such events. Such analysis is paramount to avoiding misassignment of sequencing reads when dealing with highly similar reference sequences.

3.6 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

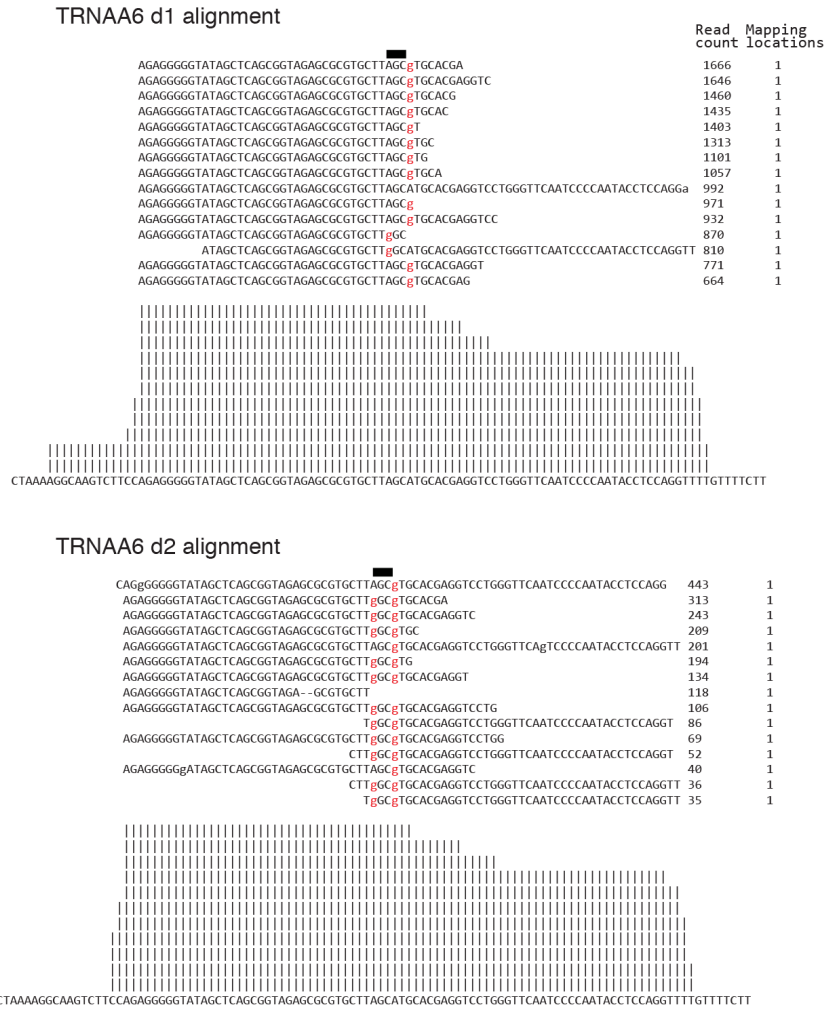


Figure 3.7: Temporal resolution of tRNA modifications. Read alignments for precursor TRNAA6 with one (top) and two mismatches (bottom). The position of the anticodon is marked by the black rectangle at the top of the alignments. Mismatches of the heavily modified adenosines in the anticodon loop are indicated by red lowercase letters. Read counts and mapping locations for each read are shown on the right side. Vertical bars represent binned, log4-transformed, and normalized read counts for each alignment.

have emerged recently as causes of severe neurodevelopmental syndromes, such as pontocerebellar hypoplasia [4, 45, 63]. Therefore, there is documented need for a comprehensive annotation of human intron-containing tRNAs, which should be revisited as markers or disease-causing candidates in phenotypically similar conditions.

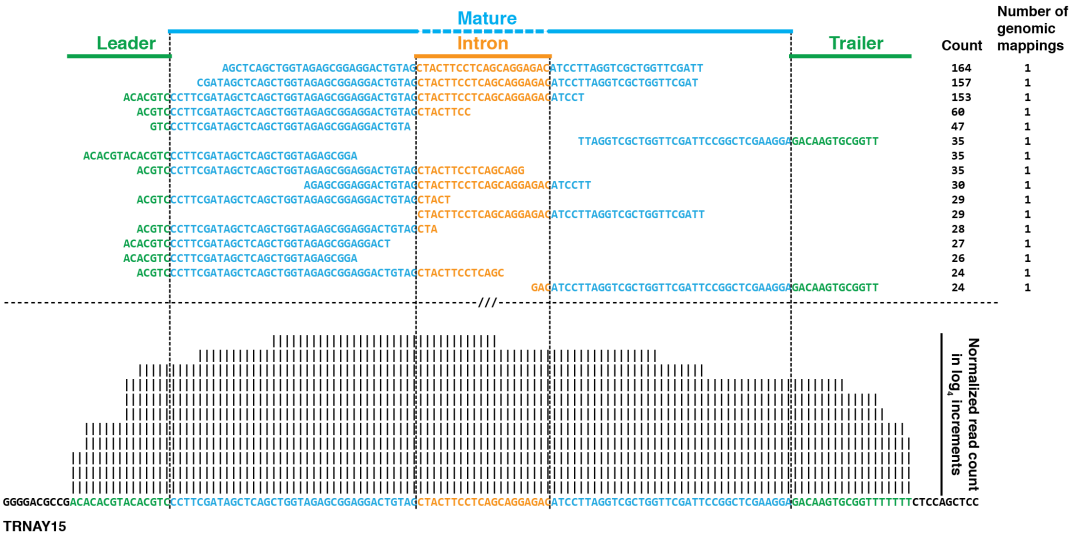


Figure 3.8: Intron-containing pre-tRNA alignment. An indicative alignment to pre-tRNA TRNAY15 is shown. Mature sequence is in blue, leaders and trailers in green, and intron sequences in orange. Precursor/mature tRNA borders are demarcated by vertical, dotted lines.

I set out to document all intron-containg tRNAs. Read coverage across both exon/intron boundaries in the pre-tRNA alignments was a requirement for the validation of intron-containing tRNAs (e.g. **Fig. 3.8**).I confirmed 26 out of 32 predicted intron-containing tRNAs by hydro-tRNAseq (**Fig. 3.9A**). Intriguingly, I could not confirm expression of any predicted intronless tyrosine pre-tRNA gene. The same was true for isoleucine-TAT and leucine-CAA isoacceptors.Excluding any unknown biologically redundant mechanism, this suggests that the integrity of the tRNA splicing complex is essential for survival, at least in the studied cellular system.

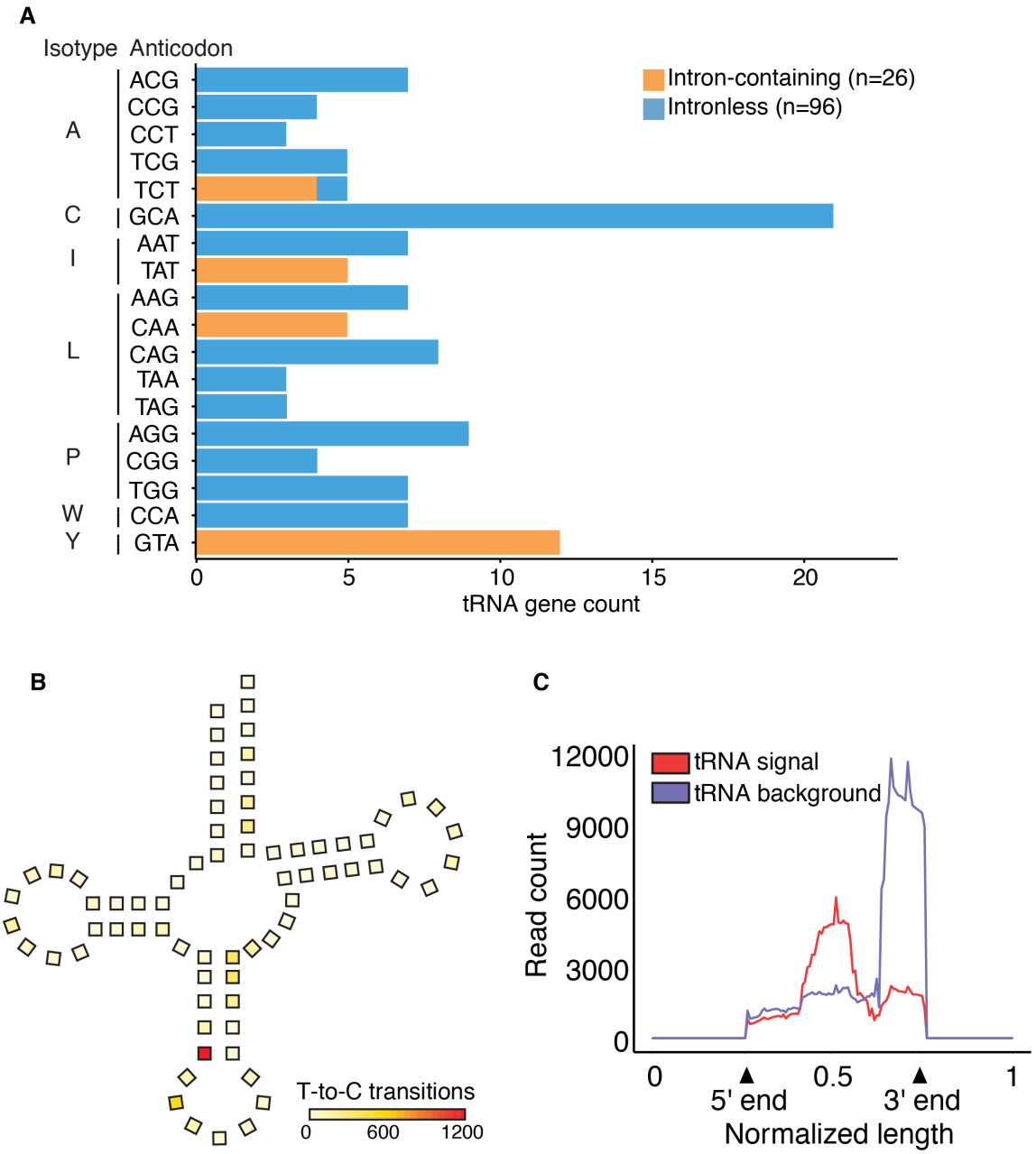


Figure 3.9: Annotation of intron-containing tRNA genes. (A) Number of validated intron-containing (orange) and intronless (blue) tRNA genes for all tRNA isotypes with possible intron-containing tRNAs. (B), (C) Analysis of PAR-CLIP for the tRNA ligase, RTCB (previously published in [64]). Positional preference of crosslinking by metagene analysis of all crosslinking events (B) and all crosslinked reads (C) to mature tRNAs. The incidence of T-to-C transitions is indicated by color intensity in (B). PAR-CLIP signal (reads containing T-to-C), background (reads with no mismatches), normalized boundaries of the pre-tRNAs (labeled as 0, 1) and mature tRNAs (labeled as 5' end, 3' end) are shown in (C).

To further confirm my observations, I coupled hydro-tRNAseq results with previously published PAR-CLIP data on the human tRNA ligase, RTCB [64]. Despite the shallow read depth of the dataset, I identified a crosslinking peak (RBP binding site) at the anticodon loop of all intron-containing tRNAs annotated by my approach (**Fig. 3.9B,C**.

3.7 Hydro-tRNAseq application in human disease

These observations came at an opportune moment because of novel contributions of tRNA processing ongoing efforts to characterize the human disease involvement of cleavage and polyadenylation factor I subunit 1 (CLP1), a 5' RNA kinase involved in tRNA splicing [31]. Before my studies, it was shown that mouse models with CLP1 that lacked its catalytic activity develop severe neuromuscular defects accentuated by loss of motor neurons and muscular paralysis [44].

Human exome capture and sequencing from individuals with ponotcerebellar hypoplasia identified a homozygous missense mutation in CLP1 (chr11:g.57,427,367 G > A [hg19], p.R140H) in five unrelated, but consanguineous families, resulting in an autosomal recessive pattern of inheritance (**Fig. 3.10**). Biochemical studies showed that this autosomal mutation reduces the interaction affinity of CLP1 for the splicing complex [46].

However, it was not clear whether this had any molecular effect on tRNA processing. I applied hydro-tRNAseq on RNA isolated from parental and patient fibroblasts. Initially, there was no significant difference in the steady-state levels of either mature (for example **Fig. 3.11**, right part) or pre-tRNAs as reported by our bioinformatics pipeline.

Interestingly, however, manual inspection of all pre-tRNA alignments revealed

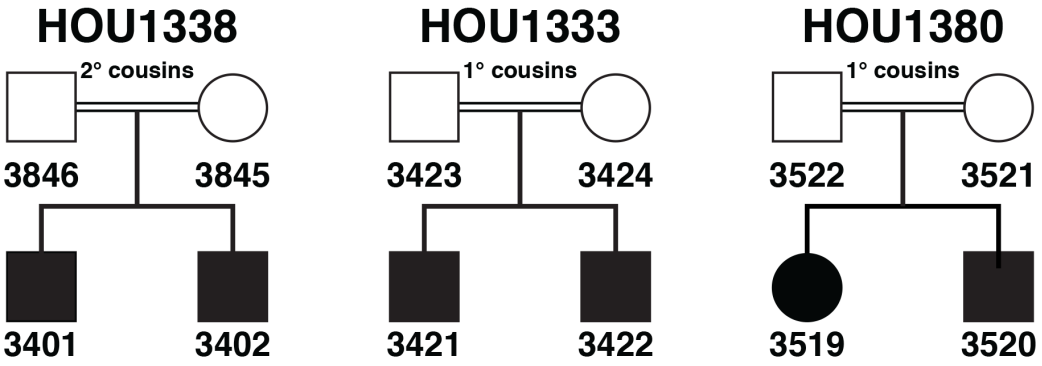


Figure 3.10: Pedigrees of families with CLP1-induced syndromes. Pedigrees of three nuclear families with patients showing microcephaly, pontocerebellar hypoplasia, dysmorphic facial features and severe neuromuscular abnormalities. CLP1 p.R140H follows autosomal recessive inheritance. Double lines indicate consanguineous marriages; squares and circles represent male and female subjects, respectively. Black-filled shapes indicate patients.

a reproducible accumulation of intronic sequences for specific pre-tRNAs (**Fig. 3.11**, left part) in patient samples. The same pattern was observed in two more pre-tRNAs, all of which showed >4-fold accumlation of intronic sequences (**Fig. 3.12**).

These results were confirmed independently by our collaborators (Stefan Weitzer, Javier Martinez, IMBA Vienna) who carried out northern blot analysis on the same RNA samples. Consistent with my results, they failed to detect significant differences in steady-state mature or pre-tRNA levels between parents and patients (**Fig. 3.13**, top-most pannels), but detected an accumulation of the salient tRNA introns of my analysis (**Fig. 3.13**, middle panels). Densitometric analysis of the northern blots also confirmed the trend and range of intron accumulation across the studied samples (**Fig. 3.12**), proving that the two techniques are in good agreement.

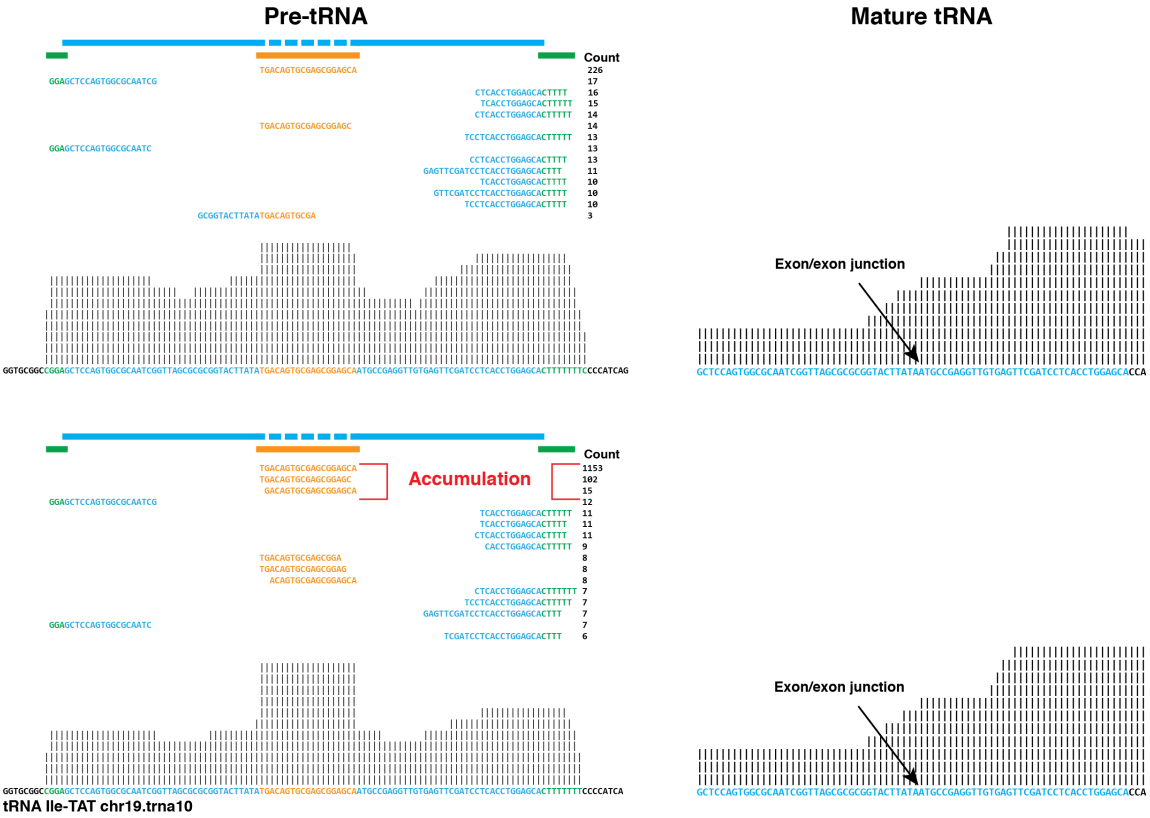


Figure 3.11: Hydro-tRNAseq analysis of CLP1 patient fibroblasts. Example of an isoleucine pre-tRNA (left part) and mature tRNA (right part) alignment for a matched parent (top) - patient (bottom) pair. Accumulating intronic reads are bracketed. The exon/exon junction is indicated in the mature tRNA alignments.

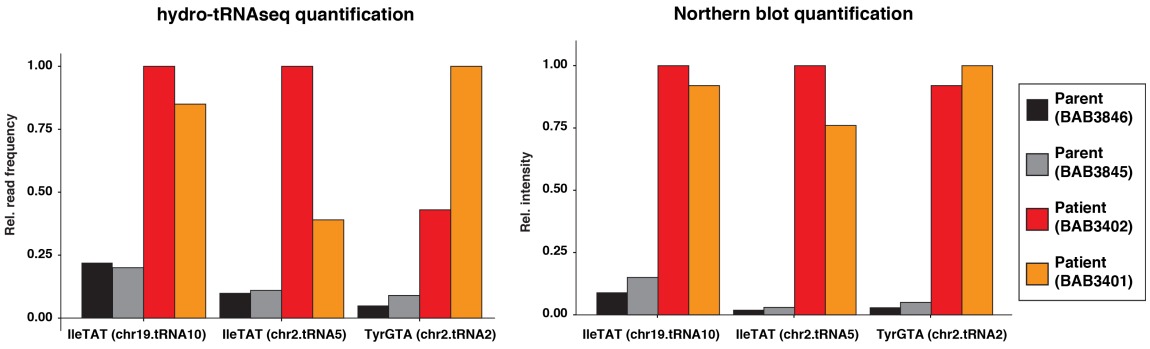


Figure 3.12: CLP1 mutation leads to intronic read accumulation. Quantification of intronic reads for three different pre-tRNAs for parents (black and grey bars) and patients (red and orange bars). Hydro-tRNAseq data shown on the left, and northern blot data on the right.

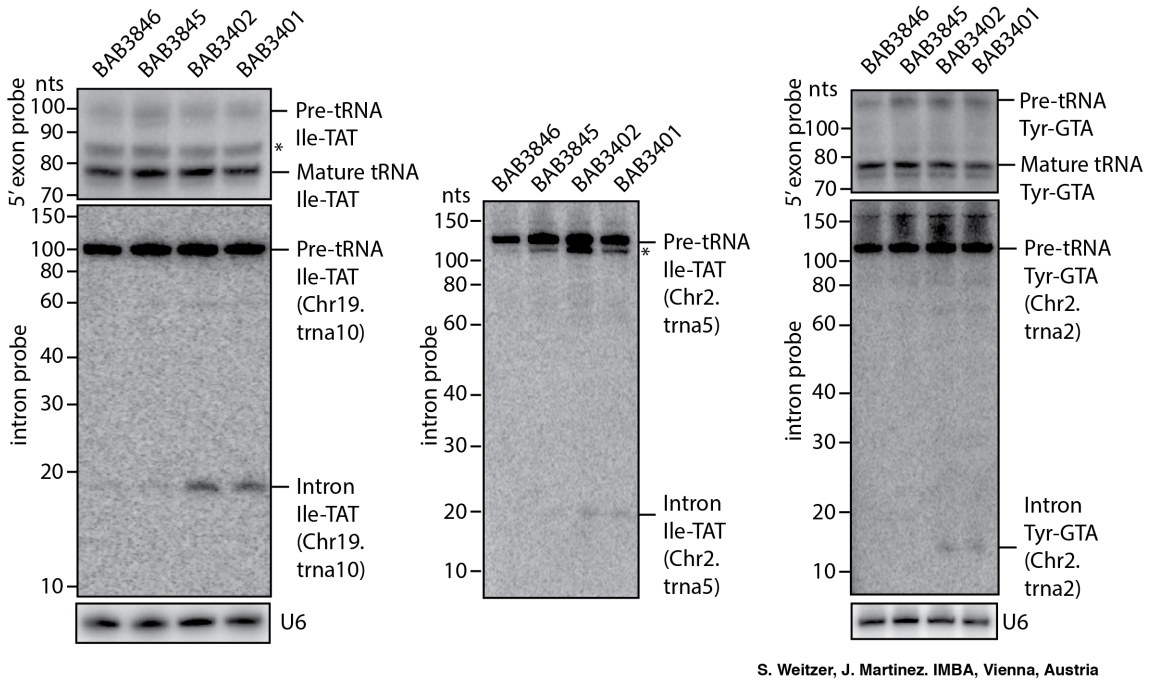


Figure 3.13: Northern blot analyses of RNA from parental and patient fibroblasts. A probe complementary to the 5' exon of isoleucine-TAT and tyrosine-GTA tRNAs was used to detect mature and pre-tRNA species (top panels). Probes specifically directed against intron sequences were used to detect pre-tRNAs and tRNA introns of two isoleucine-TAT and one tyrosine-GTA tRNA (middle panels). U6 snRNA served as loading control (bottom panels). Asterisks denote truncated pre-tRNA species.

3.7.1 Plausible pathomechanisms of CLP1 mutations

Thus, northern analysis and deep sequencing data indicate that the CLP1 R140H mutation in fibroblasts influences processing of pre-tRNAs, resulting in the accumulation of tRNA introns, whereas pre- and mature tRNA levels remain largely unaffected. Nevertheless, at this point the biological mechanisms responsible for the pathogenicity of CLP1 mutations are not clear.

During intron removal, cleavage by tRNA splicing and endonuclease complex (TSEN) leaves a 3' cyclic phosphate on the 5' tRNA exon and a 5' OH on the 3' exon (**Fig. 3.14**). These moieties are substrates of the tRNA ligase (RTCB), which

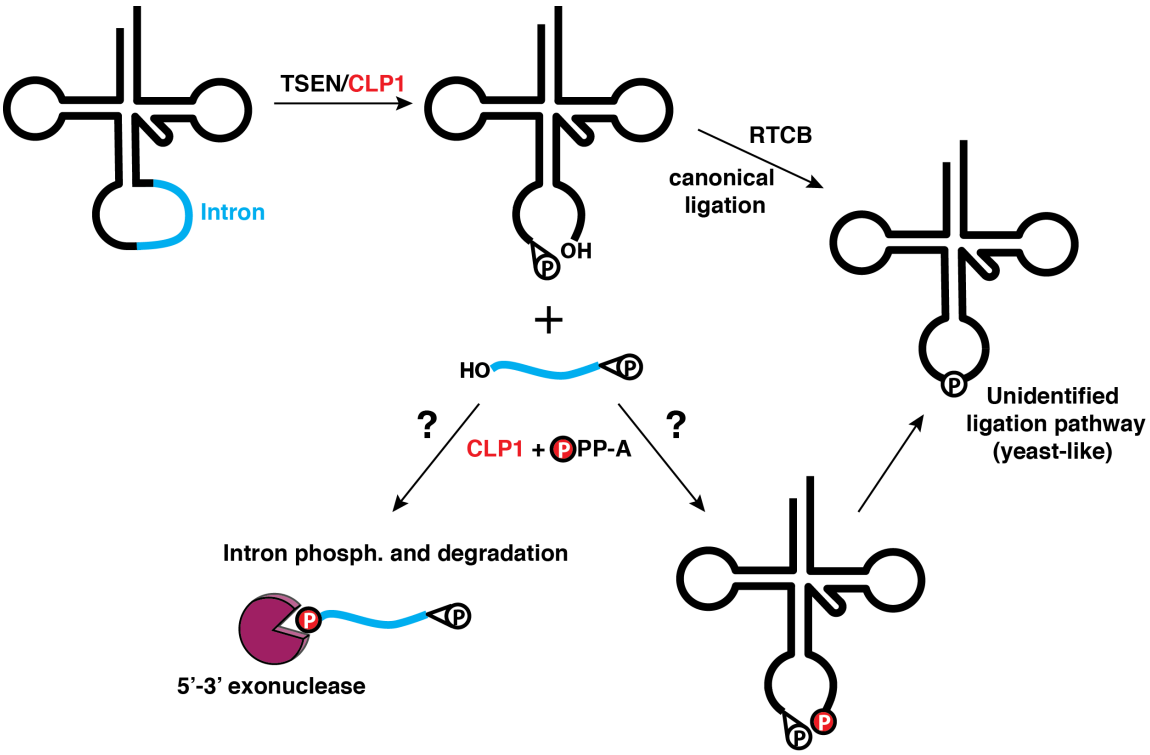


Figure 3.14: Model of CLP1 involvement in tRNA splicing. CLP1 associates with the TSEN complex, which endonucleolytically removes tRNA introns (labeled in blue). The cleavage results in a 3' cyclic phosphate and 5' OH on the 5' and 3' exons, respectively, which are ligated by RTCB (canonical ligation). CLP1 can also phosphorylate the 3' exon which could then be ligated to the 5' half by an unknown ligase. CLP1 could also phosphorylate the removed intron facilitating its removal by an exonuclease.

carries out the ligation of the two tRNA halves. CLP1 is capable of phosphorylating the 5' end of the 3' exon, possibly allowing the ligation of the two halves by a yet unknown ligase [31], which would be reminiscent of the tRNA splicing pathway in yeast [45].

Alternatively, CLP1 could be phosphorylating the excised intron, which harbors a 5' OH immediately after cleavage. The "healing" of the 5' end of the linear intron is a requirement for its degradation, at least in yeast, by the 5'-3' exonuclease Xrn1 [65]. Altogether, taking into account both the documented and hypothesized functions of CLP1, one could come up with the following testable hypotheses re-

garding the plausible pathomechanisms of CLP1 mutations:

- tRNA splicing and mature tRNA levels are reduced in metabolically in sensitive tissues (e.g. neurons), while they are unaffected in fibroblasts.
- tRNA intron accumulation leads to innate immunity activation.
- tRNA splicing proteins have pleiotropic, and possibly tRNA-unrelated effects (e.g. participation in mRNA polyadenylation [30, 45])

Chapter 4

Comparison with other methods

Recently tRNA sequencing methods have been developed that employ dealkylating enzymes and/or highly thermostable reverse transcriptase to overcome respectively the hurdles of modifications and stable structures that impede tRNA sequencing [27, 28]. However, they both have specific limitations that I tried to address. I size-selected at a higher size window to avoid contamination by tRNA-derived fragments (as compared to [27]). Also, I used two sequential adapter ligation methods to make sure that only full-length fragments were sequenced and the sequencing reads were not results of blocks during RT (as compared to [28]), which allowed me to differentiate RT stops from fragment ends. Additionally, I did not bias our sequencing protocol towards mature tRNAs, but instead captured more precursors by both hydro-tRNAseq and more importantly PAR-CLIP methods, which enabled a deeper pre-tRNA curation. Importantly, despite the reportedly high processivity conferred by dealkylating methyl modifications, in the previous studies only a small fraction of reads mapped at a given transcript were full-length reads (<1% of all reads), with a marginal increase compared to untreated controls (**Fig. 4.1**).

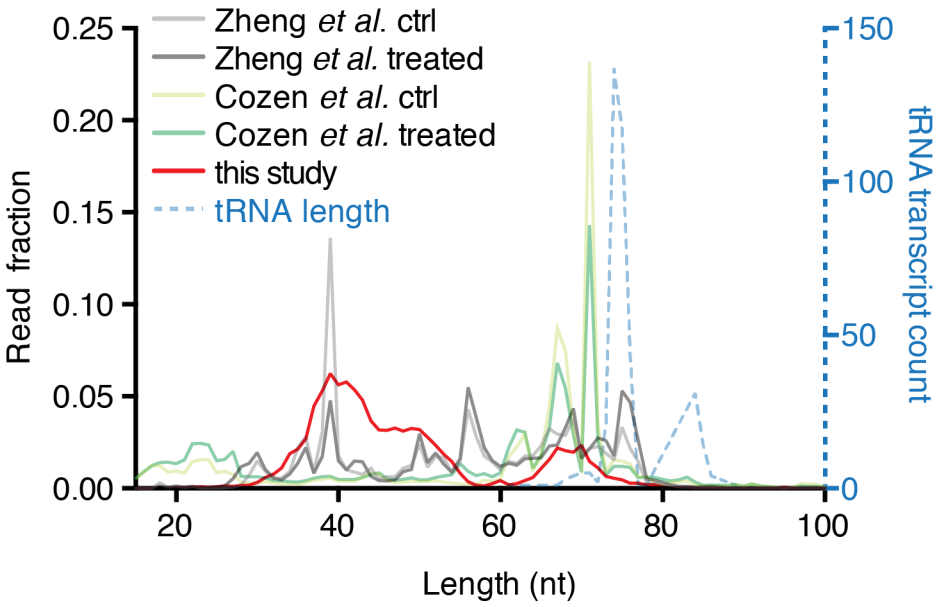


Figure 4.1: Read length distribution for hydro-tRNAseq and dealkylating sequencing methods. Histogram representing the fraction of normalized mature tRNA transcript length with ungapped and overlapping read evidence in hydro-tRNAseq and tRNA sequencing methods employing dealkylation steps (control and subjected to treatment). The mean fraction is indicated next to each method.

In contrast, hydro-tRNAseq yielded a higher cumulative fraction of mature tRNAs with read evidence across their whole length with a mean read coverage of 0.99 of the full length (compared to 0.95 and 0.87 in previous studies; **Fig. 4.2**). Also, SSB PAR-CLIP was more sensitive in identifying tRNA genes, detecting 349 genes as compared to $\sim$ 159 and $\sim$ 212 in the other methods.

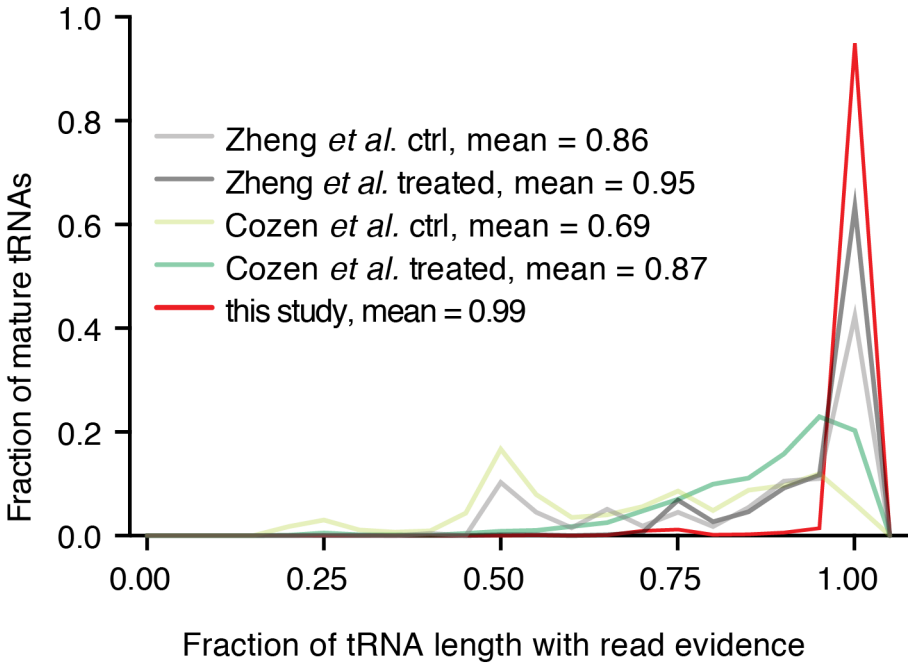


Figure 4.2: Mature tRNA read coverage by hydro-tRNAseq and dealkylating sequencing methods. The fraction of reads with a given length is indicated for hydro-tRNAseq, as well as untreated and treated samples from dealkylating methods. The distribution of tRNA lengths is shown in dotted blue lines on the right y-axis for comparison.

Chapter 5

Discussion

I have combined two complementary transcriptome-wide approaches to provide experimentally validated annotation for mature and pre-tRNA transcripts and their respective genes, in addition to furnishing an accurate quantification of tRNA abundance in human cells.

First, I developed hydro-tRNAseq, a fragmentation-based protocol for overcoming hurdles of tRNA sequencing, and obtained deep sequencing sets that enabled the annotation of tRNA genes and derivation of mature tRNA reference sequences for accurately assigning sequence reads to the otherwise edited and nucleotide-modified original tRNA pool. Alkaline hydrolysis of the tRNA-containing pool relieved thermodynamically stable structural constraints that impair ligation steps in the cDNA library preparations, reduced the number of modified nucleosides per sequenced fragment, resulting in high read-through in the RT step, and release of the 3' hydroxyl group of the otherwise aminoacylated tRNA 3' end.

Then, I took advantage of the pre-tRNA binding properties of SSB protein, which coordinates posttranscriptional processing and maturation of tRNAs [52], to enrich for tRNA precursors and allow for a comprehensive curation of pre-tRNA

transcripts and annotation of tRNA genes. Of note, since SSB interacts with pre-tRNAs and other small nuclear RNA's U-rich 3' ends in all organisms examined so far [59, 66], this approach can be adapted towards tRNA annotation in other species.

My data suggest that, at least in my experimental system, the tRNA gene space is considerably more contracted than it has been previously predicted by bioinformatics, evidenced by the fact that almost half of the predicted tRNA loci were transcriptionally silent, presumably representing retrointegrated tRNA pseudogenes or epigenetically silenced regions. It would be interesting to examine whether such an observation holds for various cell types and at different stages of development or disease, in order to confirm the differential expression and regulation of tRNA gene expression that has been reported before [21, 26].

Furthermore, my approach allowed the elucidation of relevant issues regarding POLR3 transcription such as the length of pre-tRNA leaders and trailers, the length of oligoU required for recognition by SSB, both of which have shown species specificity [58]. I detected that 4 sequential Us act as the transcription termination signal for POLR3, confirming similar predictions based on genomic sequences that suggested a requirement for 4 Us for efficient termination in vertebrates [56, 67], as well as structural data documenting the capacity of SSB to accommodate 4 Us in its binding site [51, 59]. At the same time, the length distribution of the oligoU tract identified in our experiments reflects the heterogeneity of termination signal lengths that has been noted as an intrinsic property of polymerase III *in vitro* [58].

I could also confirm a second binding site of SSB in the 5' half of the mature tRNA sequence, in support of previous observations proposing the presence of additional pre-tRNA binding sites besides the 3' tail [51, 53]. It has been previously

noted that the binding mode of SSB to tRNA is more complex than the recognition alone of the 3' tails, and that one of the RNA recognition motifs present in SSB, RRM1, could bind elsewhere in the tRNA. This stems from two observations:

i) SSB has a higher affinity for precursor over mature tRNAs,

ii) structural data show that RRM1 is unoccupied when SSB is bound to UUU-3'-OH substrate.

My data seem to validate this observation, and could shed light into new modes of SSB-mediated processing of pre-tRNAs into either mature tRNAs or other kinds of ncRNAs [14].

Moreover, I were able to carry out a careful overview and tRNA modifications that result in characteristic mismatch signatures. I introduced all possible combinations of all "mutated" nucleotides at the most prominently modified positions in every tRNA in order to collect as many reads that could be having RT misincorporations at the modified positions. This created a large number of similar tRNA sequences, and therefore I allowed for extensive multimapping, but split the read counts in order to avoid artificial read count inflation. By making use of our hierarchical annotation pipeline, I was able to dissect the temporal order of inosine modifications at the tRNA anticodon, confirming that A34 deamination occurs in the nucleus prior to the nucleolytic processing of the pre-tRNA, and establishing that the same holds true for A37 modifications, which in fact precede A34 deamination.

Accounting for modification signatures was also important for the reason that CLIP-seq, and especially PAR-CLIP, depends on apparent mismatches (in the case of PAR-CLIP, T-to-C conversions) for the identification of RBP binding sites on target RNAs. Since I used PAR-CLIP of SSB for the annotation of pre-tRNAs

and tRNA genes, I first examined uridine modifications that result in T-to-C conversions. Only a small minority of modification signatures were T-to-C transitions, suggesting that it is highly unlikely that our PAR-CLIP data were artificially inflated.

Then, I applied my protocols towards the elucidation of mechanisms of human disease, in a multi-collaborative effort to characterize a novel, severe neurodevelopmental syndrome caused by mutations in the tRNA kinase CLP1 [46]. By applying the hydro-tRNAseq protocol and analysis pipeline, I identified the accumulation of intronic sequences in the patient samples, a finding which was confirmed independently by northern blot analysis. This constituted a demonstrative proof-of-principle that tailored-RNAseq protocols can be of considerable value in the context of diagnostics.

5.1 Summary

In summary I have shown that:

i) hydro-tRNAseq is a facile and efficient method for sequencing tRNAs

ii) PAR-CLIP of SSB/La informs the annotation tRNA genes and curation of pre-tRNAs

iii) combined together, the two techniques can help answer long-standing questions of tRNA biogenesis, processing and function

iv) hydro-tRNAseq and the associated analysis can be applied for studies of human genetic diseases

Chapter 6

Materials and methods

6.1 Hydro-tRNAseq

Total RNA from HEK293 (Flp-In T-REx, Invitrogen) was isolated using TRIzol (Invitrogen). For each sample 20 μg of total RNA were resolved on 12% urea-polyacrylamide gels and recovered within a size window of 60-100 nt. The eluted fraction was subjected to limited alkaline hydrolysis in a 15 μL buffer of 10 mM Na_2CO_3 and 10 mM $NaHCO_3$ at pH 9.7 either at 65 ℃ for 10 min (replicate 1) or 1 h (replicates 2-4).

The partially hydrolyzed RNA was dephosphorylated with 10 U of calf intestinal phosphatase (NEB) in a 50 μL reaction of 100 mM NaCl, 50 mM Tris-HCl, pH 7.9 at 25 ℃, 10 mM $MgCl_2$, 1 mM DTT, 3 mM Na_2CO_3 and 3 mM $NaHCO_3$, at 37 ℃ for 1 h. The resulting RNA was re-phosphorylated with 10 U of T4 polynucleotide kinase (NEB) in a 20 μL reaction of 70 mM Tris-HCl, pH 7.6, 10 mM $MgCl_2$, 5 mM DTT and 1 mM ATP, at 37 ℃ for 1 h. Fragments of 19-35 nt were converted into barcoded small RNA cDNA libraries, as previously described [47], and sequenced on an Illumina HiSeq 2500 instrument. Adapters were trimmed us-

ing cutadapt http://journal.embnet.org/index.php/embnetjournal/article/view/200/458. Sequencing read alignments were performed using the Burrows-Wheeler aligner against an in-house curated and annotated list of mature and precursor tRNAs containing predicted tRNA sequences for human genome version hg19
http://gtrnadb.ucsc.edu. Sequencing reads were first mapped against mature tRNAs. Remaining reads were mapped against genomic tRNA sequences that included 5' leader and 3' trailer sequences, as well as tRNA introns.

6.2 SSB PAR-CLIP

Flp-In T-REx HEK293 cells (Invitrogen) were grown in high glucose DMEM supplemented with 10% (v/v) FBS, 1 mM sodium pyruvate, 100 U/mL penicillin, 100 μg/mL streptomycin, 100 μg/mL zeocin and 15 μg/mL blasticidin. Cell lines stably expressing FLAG/HA- (FH-) SSB were generated as described previously [68]. Expression of FH-SSB was induced by addition of 1 μg /mL doxycycline for 24 h. 4-SU PAR-CLIP was performed as described previously, using either RNase T1 or RNase A [69]. PAR-CLIP cDNA libraries were sequenced on an Illumina HiSeq 2500 instrument. Adapter extracted reads were aligned against an in-house curated and annotated list of mature and precursor mRNAs and ncRNAs. Bioinformatic analysis was performed using a analysis pipeline based on a curated and annotated reference RNA collection, which we organized into categories, such as rRNA, tRNA, snoRNA, mRNAs, etc. This pipeline is available at https://rnaworld.rockefeller.edu/PARCLIP_suite/. T-to-C conversion frequency, indicative of binding, was calculated for each annotated category of RNA.

6.3 Bioinformatic analysis

Reads were mapped to our transcriptomic database with error distance 0 (d0), 1 (d1) or 2 (d2), allowing mismatches, insertions and deletions. Assignment of reads with more than one mapping locations that belong to different RNA classes followed a hierarchical procedure reflective of the cellular abundance of each RNA class. Mature RNA sequences (e.g. fully processed tRNAs) received priority compared to precursors (e.g. pre-tRNAs), thus minimizing multimapping events. A tRNA gene was considered to be expressed when there were reads spanning the precursor/mature junctions, including exon/intron junctions for intron-containing tRNAs. For abundance reports, multimapping reads were split equally over the number of their mapping locations, and all reads mapping to edited and non-edited versions of the same tRNA transcript were summed for quantification of a given tRNA. Naming of tRNAs followed HUGO guidelines, with edited variants of reference tRNAs exhibiting the edited position and the identity of induced mismatch in their naming. The same analysis pipeline was applied to mitochondrial tRNAs, with the exception that no tRNA precursors were annotated due to continuous transcription of the mitochondrial genome. Remaining reads that did not map to any annotated transcript were mapped to the human genome. The mapped read annotation process was based on a hierarchical procedure that assigned priority to reads mapping in their entirety to mature sequences, followed by reads that spanned the precursor-mature junctions. Bioinformatic analysis was performed by custom Perl and Python scripts. Graphs were created in R and Prism (Graphpad).

6.4 Accession codes

The RNAseq and PAR-CLIP sequence data have been deposited in the National Center for Biotechnology Information (NCBI) Sequence Read Archive under accession numbers GSE95683.

6.5 Code availability

All scripts used for the analysis (written in Perl, Python and R) are available upon request, and will also be deposited on `https://github.com/tgogakos` upon publication of the presented work.

Part II

C3PO

Chapter 7

Introduction

Even though tRNA biology has been studied extensively in the past, there are still fundamental parts in the lifecycle of tRNAs that remain unexplored in humans. For example, the 3' end processing of human tRNAs is not completely understood [52]. On one hand, this can be attributed to divergent evolutionary trajectories of tRNA processing pathways within Eukarya, resulting, for example in great differences in tRNA splicing and nucleocytoplasmic dynamics (this has already been alluded to here in section 3.7.1), but has been surveyed in much greater detail and eloquence elsewhere [34, 35, 70]). On the other hand, tRNA processing and expression is characterized by multiple levels of evolutionary and functional redundancy (e.g. multiple translation factors, editing enzymes, multicopy genes). One could speculate, and explain the emergence of such redundant mechanisms in light of the essentiality of tRNA biogenesis for life.

Given the incomplete knowledge of human tRNA processing proteins, I was interested in studying potentially novel such factors. Thus, I turned my attention to component 3 promoter of RISC (C3PO), which is a conserved, multisubunit complex of the RBP translin (TSN) and translin-associated protein X, also known as

TRAX (TSNAX) [71]. This complex has been associated with various biological functions, ranging from double stranded DNA (dsDNA) damage response, RNA interference (RNAi) enhancement, and has also been involved neurologic development [71–75].

Both members of the complex contain a previously undescribed RNA recognition motif, and TSNAX possesses RNA endonucleolytic activity both *in vitro* and *in vivo*, with a preference for single stranded RNA (ssRNA) [76, 77]. The crystal structure of the C3PO complex, with or without substrates, from various organisms has also been reported, and has determined a tight octameric assembly, consisting of six TSN and two TSNAX monomers [76–78]. Its biological importance is underscored by the neurodegenerative phenotypes observed in mice lacking components of the complex [74].

Interestingly, it was recently observed thatlloss of C3PO in the fungus N. crassa and in mouse fibroblasts results in accumulation of tRNA fragments (tRFs), elevated levels of mature tRNAs, enhanced protein translation, and increased resistance to cell death-inducing agents [79]. This unexpected finding suggested that C3PO might be a novel tRNA processing enzyme, at least in the two studied species. Despite these studies, though, the biological targets and the details of the biochemical activity of C3PO remain elusive. Moreover, it is not yet clear whether C3PO is important for the biogenesis of tRNAs, the generation of competent stress or non-stress related tRNA fragments or if it is simply involved in tRNA turnover.

Chapter 8

C3PO PAR-CLIP

I carried out PAR-CLIP in order to identify the transcriptome-wide targets of C3PO. I used inducible cell lines expressing a FLAG-HA-TSN or FLAG-HA-TSN-(2A)-TSNAX*-HA construct, under the control of a doxycycline-inducible promoter. TSNAX* had two mutations in the endonuclease active site, because induction of a catalytically active TSNAX proved lethal in our system (data not shown). Both ORFs were cloned in tandem, with an intervening self-cleaving 2A intein peptide, in order to control their concomitant expression and their relative stoichiometry as tightly as possible.

PAR-CLIP on C3PO confirmed its role as a bona fide tRNA binding protein, since tRNAs collected the largest fraction of sequenced reads (**Fig. 8.1A**), with the characteristic T-to-C signature (**Fig. 8.1B**). Of note, my PAR-CLIP results seem to bring into question previous reports that implicated C3PO in enhancing RNAi activity by promoting miRNA-mediated gene silencing, since only 0.2% of the total reads mapped to miRNAs. However, a large percentage of binding sites (referred to as clusters) was in 5' UTRs ($\sim$22% of total), a result that came as a surprise, as C3PO has not been reported to interact with 5' UTRs before.

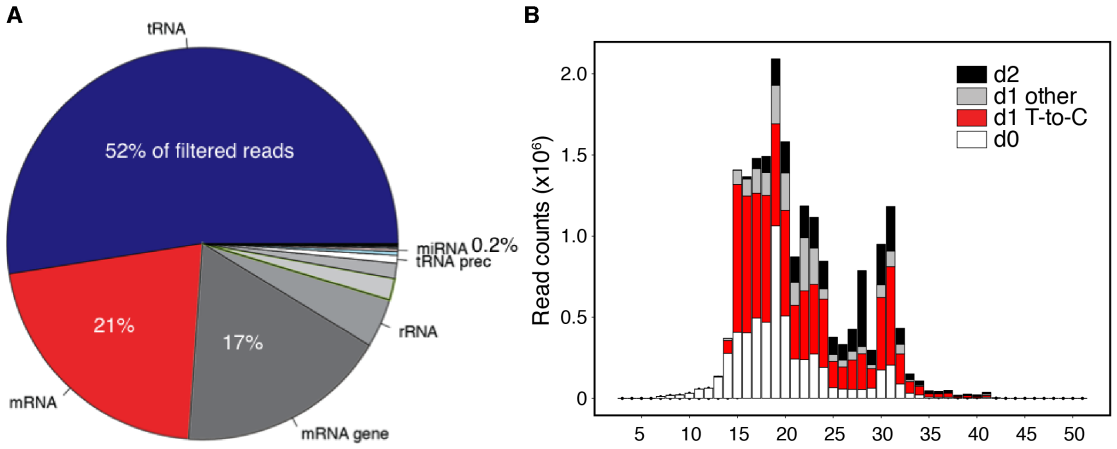


Figure 8.1: C3PO crosslinks to tRNAs. (A) Pie chart, showing distribution of C3PO PAR-CLIP reads. (B) C3PO PAR-CLIP reads mapping to mature tRNAa with 0, 1 or 2 mismatches (d0, d1, d2); reads with T-to-C mismatches are separated (red) from the rest of the reads with one mismatch (gray). Read length and number of reads are represented on the x- and y-axis, respectively.

Metagene analysis of the tRNA clusters and crosslinking sites showed that C3PO interacts with the dihydro-uridine stemloop (D stemloop) (**Fig. 8.2**), and it shows preference for Us located adjacent to the conserved di-G sequence in the loop (nt 16-19 in **Fig. 8.3**).

Since the detected binding site on tRNAs is conserved both in sequence and in structure, it was not clear whether the di-G motif was necessary for binding or whether C3PO was just recognizing the stemloop structure. Motif enrichment analysis showed that the same or a highly similar motif always containing the di-G sequence was present in non-tRNA targets of C3PO (**Fig. 8.4**), which would suggest the necessary contribution of the sequence motif to recognition and binding by C3PO.

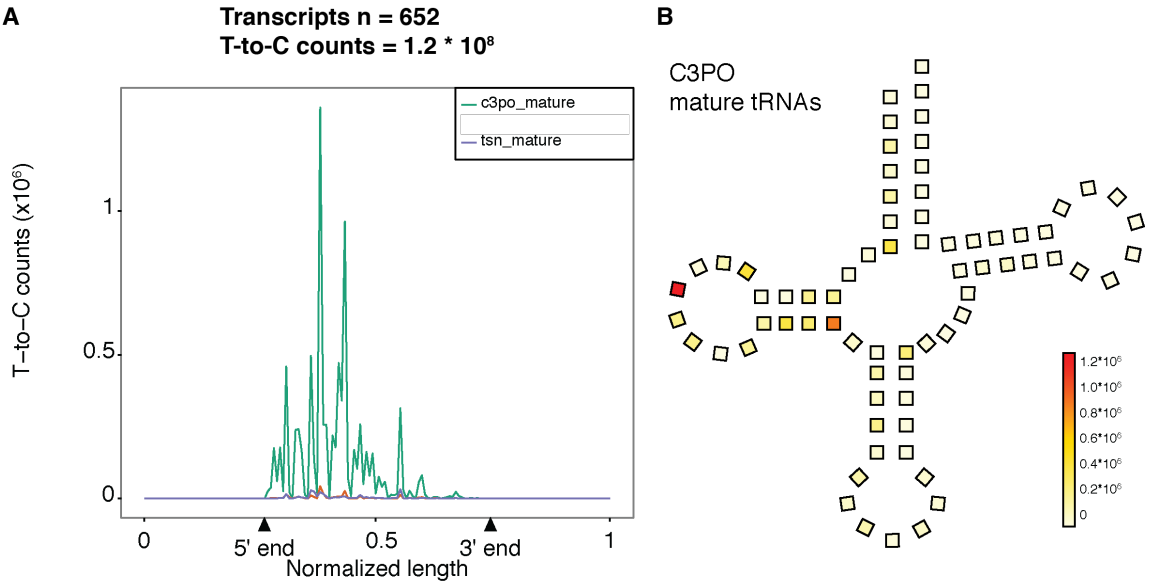


Figure 8.2: Metagene analysis of C3PO crosslinking to mature tRNAs. (A) Positional preference of C3PO crosslinking by metagene analysis of reads mapped hierarchically; first to mature and then to precursor tRNAs. PAR-CLIP signal (reads containing T-to-C) from two replicates is shown (C3PO and TSN). The normalized boundaries of pre-tRNAs (labeled as 0,1) and mature tRNAs (labeled as 5' end, 3' end), and T-to-C counts are shown on the x- and y-axis, respectively. (B) Positional preference of C3PO crosslinking by metagene analysis of all crosslinking events to mature tRNAs. The incidence of T-to-C transitions is indicated by color intensity.

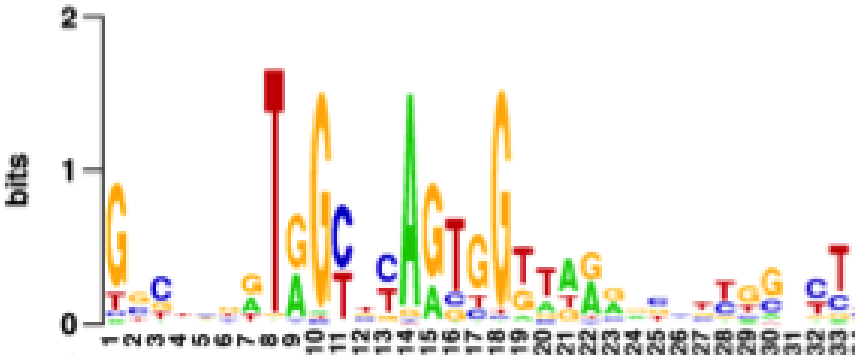


Figure 8.3: Weblogo of 5' tRNA sequences. Information entropy across the 5' half of all tRNAs shows high nucleotide conservation at various positions, and especially at the C3PO binding site(nt 16-19. Weblogo was created using `http://weblogo.berkeley.edu/logo.cgi`. See also **Fig. 2.2**.

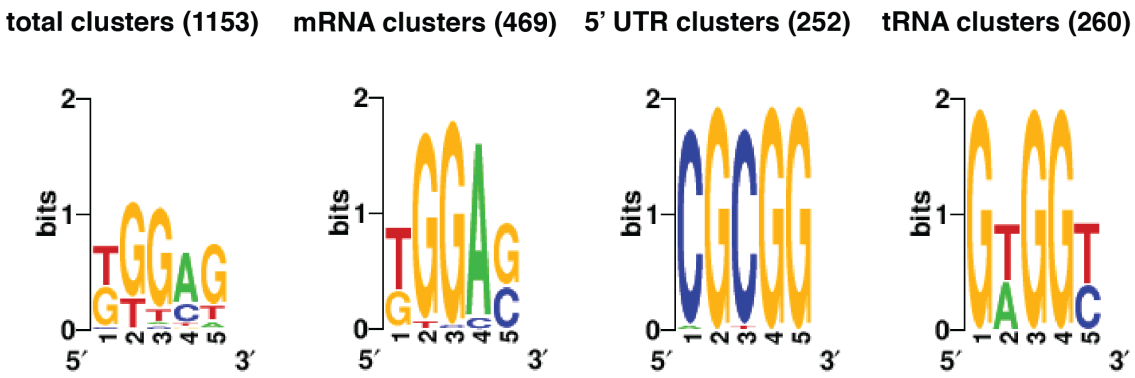


Figure 8.4: Motif analysis of C3PO RNA targets. Predicted RNA recognition motifs shown for total clusters, as well as the most common types of clusters grouped by RNA category. 5' UTR clusters are shown separately from total mRNA clusters, due to their high enrichment. Number of corresponding clusters is shown in parenthesis. Motifs were predicted using MEME (`http://meme-suite.org/tools/meme`).

Chapter 9

Biochemical characterization

9.1 C3PO possesses a length- and structure-dependent endonucleolytic activity

It was previously observed in our lab that the endonucleolytic activity of C3PO is much reduced for RNA substrates longer than 80 nt (unpublished data). This cutoff is very close to the average length of a mature human tRNA (75 nt). The length cut-off of C3PO's activity along with the presence of tRNAs in its PAR-CLIP target list prompted the in vitro study of C3PO's nuclease activity on tRNAs.

I performed cleavage assays on 5' or 3' labeled, in vitro transcribed histidine ($tRNA^{His}$, 76 nt) and serine-tRNA ($tRNA^{Ser}$, 85 nt) (**Fig. 9.1**). These substrates were chosen because of their confirmed presence in the PAR-CLIP dataset, and because they lie inside and outside the length-dependent window of optimum nuclease activity, respectively. My data suggested that $tRNA^{His}$ is a favored substrate, consistent with the length preference of C3PO. More importantly, I noticed that C3PO elicits a structure-dependent cleavage pattern with respect to tRNAs. C3PO activity does not lead to processive degradation of the tRNA substrates

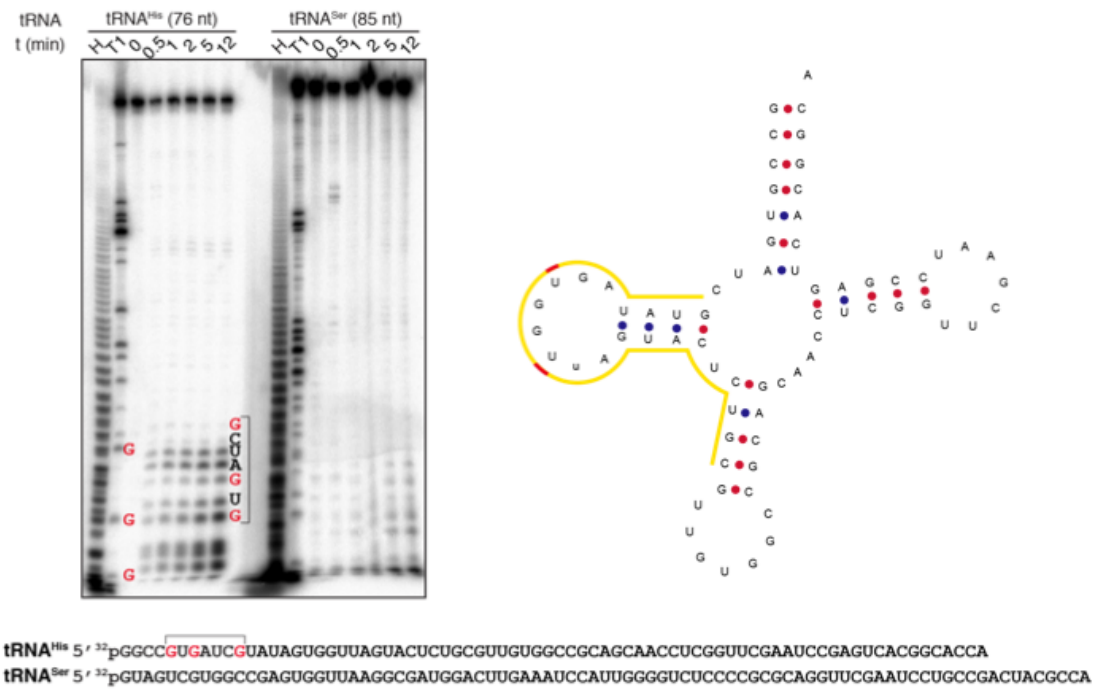


Figure 9.1: C3PO *in vitro* cleavage assay. Recombinant C3PO and 5' *in vitro* transcribed and radiolabeled tRNAs, identified as PAR-CLIP targets were used in cleavage assays (left panel). RNase T1 and hydrolysis were used as reference to map cleavage sites. The sequences used are shown in the bottom, and the major cleavage sites for $tRNA^{His}$ are labeled on the figure. The secondary structure of the same tRNA is shown on the right, with C3PO binding site labeled in yellow , and crosslinked residues in red.

(see hydrolysis ladder [H] in **Fig. 9.1**), but rather results in products of specific length ($\leq$ 9 nts). This suggests that C3PO cleaves preferentially at the shoulders of the acceptor stem of tRNAs.

9.1.1 EMSAs

In order understand the requirements for interactions of C3PO with its substrates, I performed electrophoretic mobility shift assays (EMSAs) with either TSN or cat-

alytically inactive C3PO, and a wide range of substrates, including G-homo-oligomers, single stranded GGU repeats and short RNA stemloops containing C3PO's putative binding motif (**Fig. 9.2**).

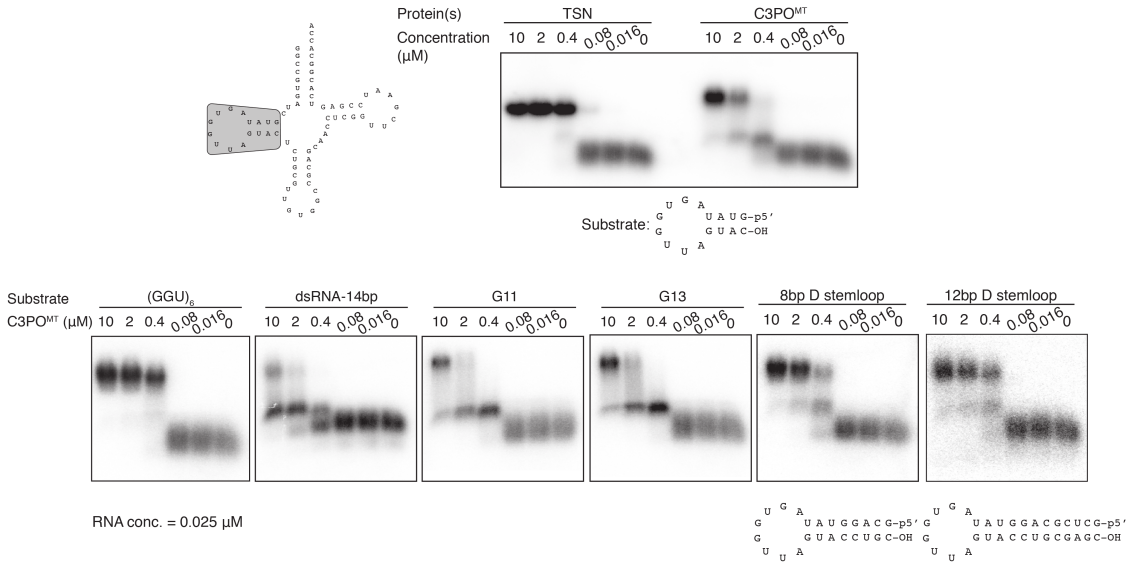


Figure 9.2: EMSA analysis of TSN and C3PO. Recombinant TSN or C3PO (with mutated catalytic site) was incubated with chemically synthesized, 5' radiolabeled oligonucleotides. Stemloop substrates are shown below the respective panels. G11, G13: oligo-G substrates; $(GGU)_6$: 18-nt-long oligo of GGU tandem repeats.

The results showed the following:

- both TSN and C3PO bind to PAR-CLIP identified stemloops
- C3PO shows a stepwise binding pattern, indicated by the formation of an intermediate product
- C3PO prefers GGU repeats over G-homo-oligomers, which suggests direct contribution of U residues to binding
- C3PO binds stemloops better than linear or dsRNA sequences
- extending the length of the stem in a stemloop increases binding affinity slightly.

Using this information, the Patel group (MSKCC) obtained well-diffracting crystals of TSN, crystallized in presence of ether single stranded RNA sequence 5'-(UG)3U(UG) or the tRNA dihydrouridine stem loop sequence (5'-GUAUAGUGGUUAGUAC). The structures were solved at 2.2 Å and 2.74 Å resolutions, respectively, and revealed minor differences in the arrangement of octameric TSN, but no bound RNA substrate in the expanded hollow interior of the closed-barrel structures, thus failing to improve on the previously reported human C3PO structure [80].

9.1.2 Immunoprecipitations

In order to further probe the biochemical function of C3PO, I tried to identify interacting proteins. Immunoprecipitation of either TSN or TSNAX, using anti-FLAG or anti-TSNAX antibodies resulted only in the co-immunoprecipitation of the other member of the complex, as indicated by the colloidal blue stain and WBs in **Fig. 9.3**. This was also observed in even mild salt conditions (150 mM NaCl) and in the presence of the the cell-permeable protein corsslinking agent DSP. Together, these results suggest that C3PO does not form stable interactions with other protein factors, at least in our cellular system.

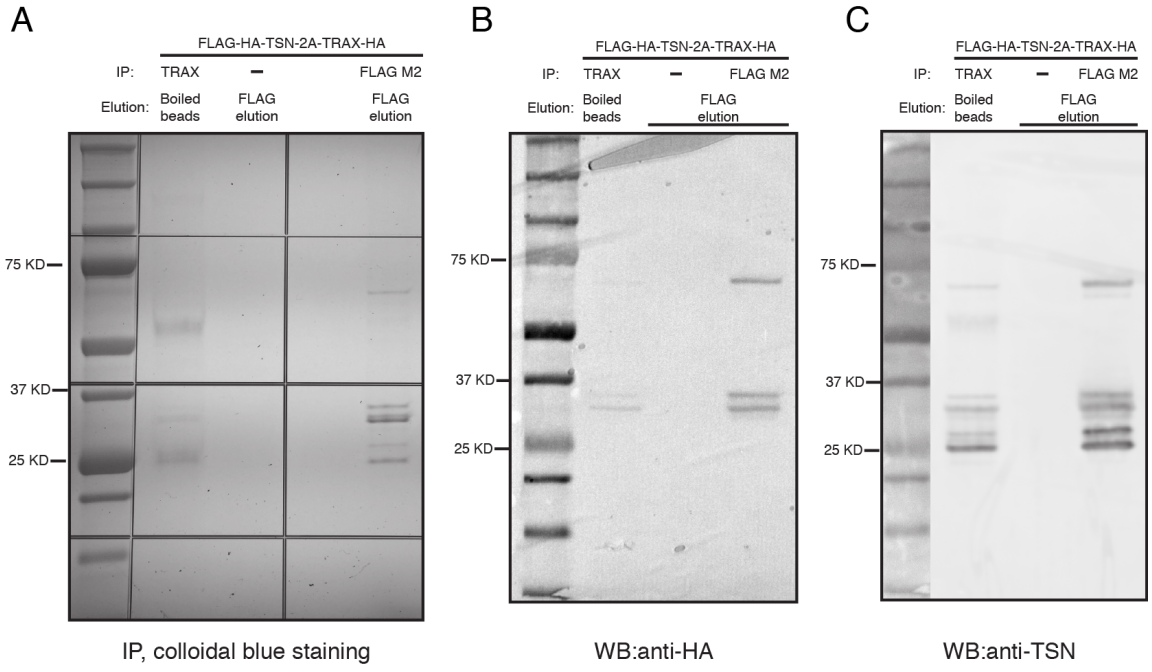


Figure 9.3: C3PO co-immunoprecipitation. TSN or TSNAX were immunoprecipitated from HEK293 lysate after induction. (A) Colloidal blue staining of eluate. (B) WB against HA tag, present on both TSN and TRAX. (C) WB against TSN on the same membrane as (B). TSN runs at $\sim$25 KDa and TSNAX at $\sim$30 KDa. The band around 75 KDa most likely corresponds to uncleaved 2A-containing construct.

Chapter 10

Biological characterization

10.1 Loss-of-function studies

To study the biological importance of C3PO in HEK293 cells, first I performed siRNA-mediated transient knockdown of its members, followed by high-throughput sequencing. Surprisingly, neither mRNA-seq or hydro-tRNAseq showed any reproducible or statistically significant effect on the levels of tRNA or mRNA targets of C3PO (data not shown). The same was true for smallRNA-sequencing, which captures miRNAs and tRNA fragments. Therefore, transient depletion of C3PO had no effect on the abundance of its RNA targets.

In light of these results, I generated TSN and TSNAX knockout cell lines by CRISPR-Cas9-mediated genome editing. Since knockdowns did not have a global pronounced effect on RNA levels, I reasoned that the depletion achieved by the siRNA methodology ($\sim$70% reduction in protein level) was not impactful enough to result in an observable effect on global tRNA levels. Even though hydro-tRNAseq analysis still pending, preliminary mRNA-seq analysis of multiple knockout clones once again failed to show major and consistent changes on mRNA abundance

(**Fig. 10.1**).

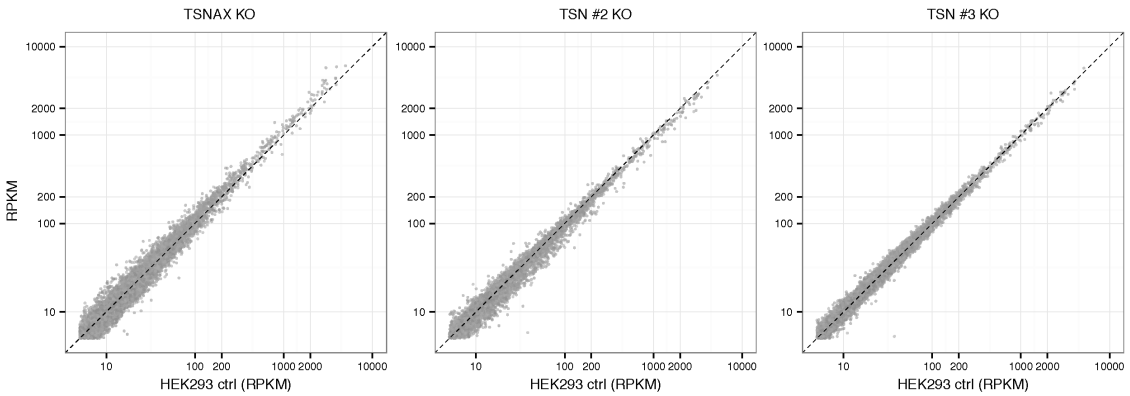


Figure 10.1: mRNA-seq analysis of TSN and TSNAX knockout clones. Scatterplot of mRNA expression values (in RPKMs). Parental HEK293 cell line always corresponds to the x-axis. The knockout cell line and clone, as indicated on the titles, corresponds to the y-axis.

10.2 Translational effects

Even though C3PO levels did not seem to have a detectable effect on RNA levels, I reasoned that the complex could perhaps affect directly and specifically the translational efficiency of its 5' UTR targets. In fact, the binding properties of C3PO towards specific 5' UTRs of some mRNAs reflect those of known translational factors, such as eIF3 [81], inasmuch as they include short (<50 nt) and single binding sites (**Fig. 10.2A,B**), with GC-content significantly higher than randomly sampled, size-matched sequences from the 5' UTR background context (**Fig. 10.2C**), and many predicted structural elements (**Fig. 10.2D**).

I observed that siRNA knockdowns of TSN led to increased translation of mRNA targets identified by PAR-CLIP, including the genes HNRNPA0, GADD45a, CJUN and JNK1 (**Fig. 10.3**). This did not seem to be the result of a confounding cascade effect, due to the participation of many of the aforementioned genes

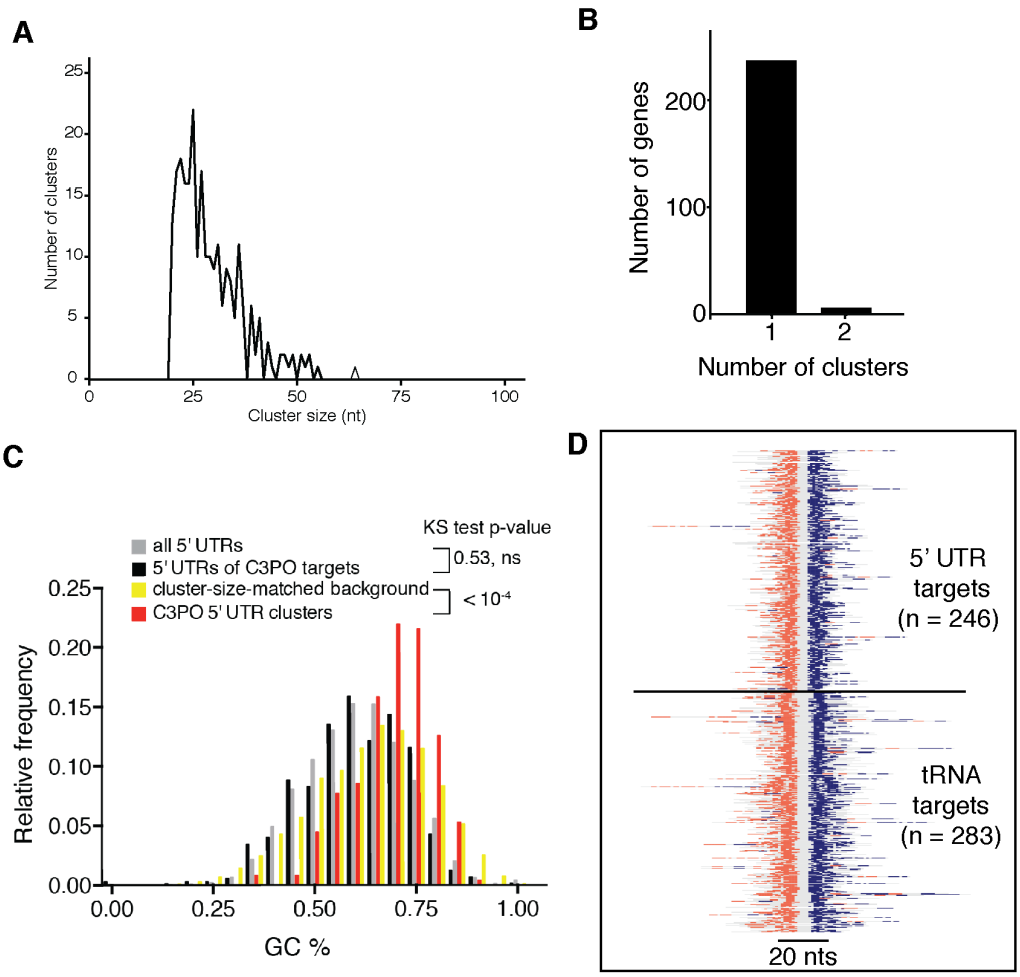


Figure 10.2: C3PO 5' UTR targets. (A) Length distribution of C3PO 5' UTR clusters. (B) Number of C3PO clusters in target genes. (C) GC content of C3PO 5' UTR clusters (red), size-matched background sampled from all 5' UTR sequences (yellow), complete 5' UTR sequences of C3PO targets (black) and all 5' UTRs (grey). Kolmogorov-Smirnov test p-values shown. (D) Predicted secondary structures of 5' UTR targets of C3PO (and tRNA targets for comparison) - analysis performed as in 3.5.

in shared signaling pathways, as another member of the pathway that was not a C3PO target (ERK1/2) remained unaffected.

Nevertheless, such results should be taken with a grain of salt, given the high inter-experimental variability across replicates of siRNA knockdowns, and also the occasionally inaccurate quantification by WB densitometry.

For this reason, I decided to examine the proteome-wide effects of C3PO loss-

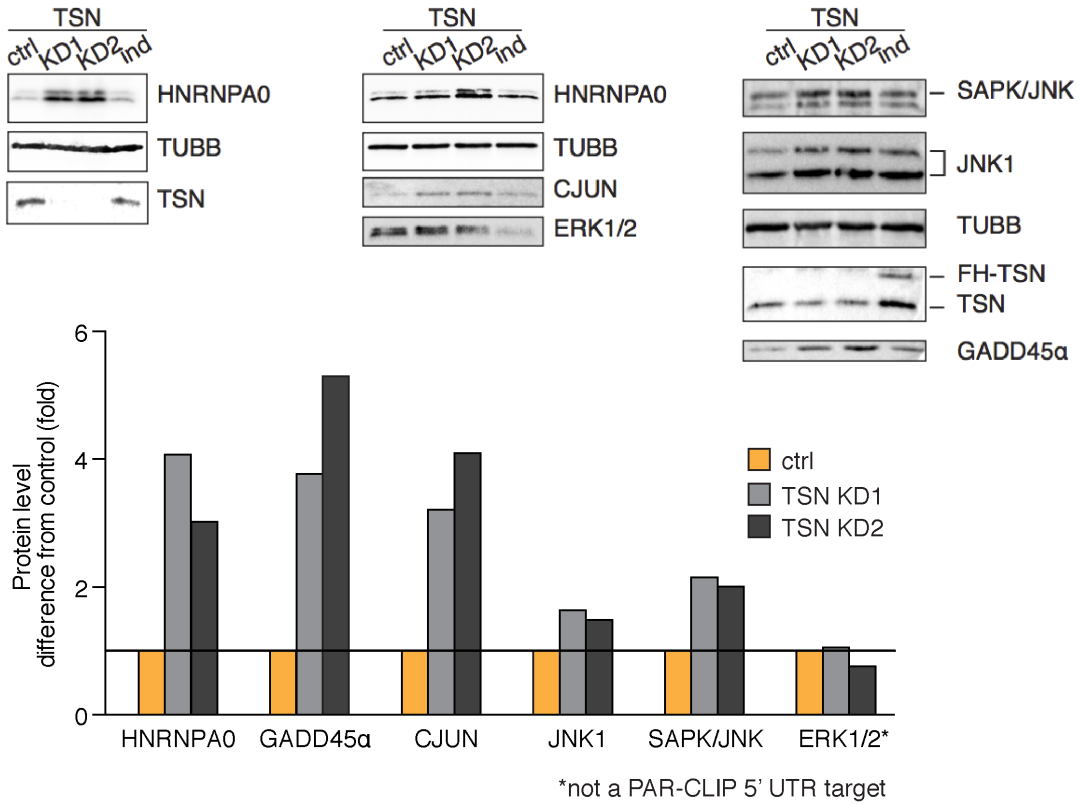


Figure 10.3: Translation effects of TSN knockdown on mRNA targets. Top: WB against 5' UTR targets of C3PO upon two knockdown repicates (KD1, KD2) and induction of expression of TSN. TUBB used as loading control and ERK1/2 as non-target control (not detected by PAR-CLIP). Bottom: densitometric analysis of presented WBs.

of-function. I performed SILAC (stable isotope labeling by amino acids in cell culture) followed by mass spectrometry, comparing C3PO knockout cell lines to parental HEK293 cells. Rather surprisingly, again, there was not a single differentially expressed protein in the overlap of three biological replicates carried out in triplicate (data not shown).

Chapter 11

Discussion

C3PO is a multimeric complex of the RNA binding protein TSN and the RNA endonuclease TSNAX. A plethora of functions have been assigned to this complex, such as a role in the repair of DNA breaks, enhancement of RNAi activity, and tRNA processing.

By performing PAR-CLIP I identified C3PO as a tRNA binding protein, and showed that it did not bind to miRNAs or 3' UTRs of mRNAs, arguing against a role in RNAi. Except for tRNAs, the other main target of C3PO was 5' UTRs, something unusual for mRNA binding proteins apart from translation initiation factors.

In order to investigate the effects of C3PO in mRNA and tRNA stability, I performed hydro-tRNAseq and small-RNAseq after depletion of C3PO by siRNA knockdowns, and mRNAseq after both knockdowns and CRISPR-Cas9 knockouts. To my surprise, no significant effect was observed transcriptome-wide.

These results can be interpreted in two ways:

i) Since C3PO is an enzymatic complex, perhaps partial loss of function conferred by siRNA knockdown is not sufficient to yield an observable phenotype, at least under the tested experimental conditions. Given that hydro-tRNAseq

analysis of the knockout clones is still pending, the possibility of an effect on tRNAs, albeit improbable, cannot be ruled out.

ii) C3PO has an effect on translation rather than on mRNA or tRNA stability. This hypothesis was reinforced initially by preliminary results, showing increased protein levels of 5' UTR targets identified, upon siRNA knockodown of TSN. SILAC results, though, came in stark contrast as no significant proteome-wide differences were observed after TSN or TSNAX knockout.

Trying to explain the lack of observable molecular phenotypes in the absence of C3PO, one could invoke evolutionary mechanisms of redundancy. One hypothesis could be that C3PO shares functions with another conserved nuclease (e.g. RNase P), requiring the loss of both proteins/RNPs for a molecular phenotype to arise. It would be interesting to create genetic models that allow combinatorial loss of function of such candidates in either a transient or permanent fashion. Such an approach would likely elucidate the interplay between C3PO and other factors and unravel their epistatic relationships.

Alternatively, loss of C3PO may not result in dramatic steady-state defects, but might sensitize cells to external toxic stimuli. In this case, one would need to challenge C3PO-deplete cells with environmental or chemical stresses in order to tip the cellular balance over and allow for the molecular effects of C3PO to be seen.

Chapter 12

Materials and methods

12.1 PAR-CLIP

C3PO PAR-CLIP was performed as described in section6.2 with the following differences: Two RNase T1 steps were used. The crosslinked RNP was first transferred to a nitrocellulose membrane, followed by on-membrane proteinase K digestion and RNA recovery from excised membrane pieces, in order to minimize background contamination of highly abundant tRNAs.

12.2 Cleavage assays

In vitro cleavage assays were performed as described previously by our lab [77]

12.3 Immunoprecipitation

Immunoprecipitations were carried out using PAR-CLIP buffer conditions for lysis and wash buffers (see section 6.2), with varying salt concentrations. DSP crosslinking was performed as described previously [33].

12.4 EMSAs, siRNA knockdowns, and SILAC

EMSAs and siRNA knockdowns were performed as described previously by our lab [82]

12.5 CRISPR-Cas9 knockouts

CRISPR-Cas9-mediated genome editing was performed as described previously [83]. Three small guide RNAs (sgRNAs) cognate to the coding region of TSN (5'-GAGGTTTGTGTTGCAGCGCT, 5'-GTTGCAGCGCTTGGTCTTCT,
and 5'- TGAAATCCTTTCTCCCGATC) or TSNAX (5'- CATTTAAAGGCCAACAT-CAC, 5'-TATAAGGATCAGGAGGGTTC, and 5'- GAGACTTGTGAAACTTAGTC). PCR products were cloned using the Zero Blunt PCR Cloning Kit (Thermo Fisher Scientific), six clones were sequenced per cell line to verify successful genome editing, and three clones were kept and used in experiments.

12.6 mRNAseq and analysis

Oligo(dT)-selected RNA was converted into cDNA for polyA mRNA-sequencing using the Illumina TruSeq RNA Sample Preparation Kit v2 according to the instructions of the manufacturer and sequenced on an Illumina HiSeq 2500 platform using 100 nt single-end sequencing. Analysis was performed using the Tuxedo suite [84] and in-house scripts Perl, Python and R scripts, available upon request.

References

1. Crick, F. H. C. *From DNA to protein On degenerate templates and the adapter hypothesis: a note for the RNA Tie Club* 1955.
2. Woese, C. *The Genetic Code. The Molecular basis for Genetic Expression* 1st ed. (Harper, 1967).
3. Soll, D. & RajBhandary, U. *tRNA: Structure, Biosynthesis and Function* 1st ed. (ASM press, 1995).
4. Cooper, T. A., Wan, L. & Dreyfuss, G. RNA and Disease. *Cell* **136,** 777–793 (Feb. 2009).
5. Park, S. G., Schimmel, P. & Kim, S. Aminoacyl tRNA synthetases and their connections to disease. *Proceedings of the National Academy of Sciences of the United States of America* **105,** 11043–11049 (Aug. 2008).
6. Griffiths, E. J. in, 249–267 (Springer Netherlands, Dordrecht, Dec. 2011).
7. McFarland, R., Elson, J. L., Taylor, R. W., Howell, N. & Turnbull, D. M. Assigning pathogenicity to mitochondrial tRNA mutations: when 'definitely maybe' is not good enough. *Trends in Genetics* **20,** 591–596 (Dec. 2004).
8. 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).

9. 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).

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

11. 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).

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

13. 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).

14. 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).

15. 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).

16. 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).

17. Haussecker, D. *et al.* Human tRNA-derived small RNAs in the global regulation of RNA silencing. *RNA* **16,** 673–695 (Mar. 2010).

18. Babiarz, J. E., Ruby, J. G., Wang, Y., Bartel, D. P. & Blelloch, R. Mouse ES cells express endogenous shRNAs, siRNAs, and other Microprocessor-independent, Dicer-dependent small RNAs. *Genes & Development* **22,** 2773–2785 (Oct. 2008).

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

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

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

22. 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).

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

24. 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).

25. 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).

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

27. 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).

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

29. Trotta, C. R. & Abelson, J. *The RNA World* Second edition (eds Gesteland, R. F., Cech, T. R. & Atkins, J. F.) 561–584 (Cold Spring Harbor Laboratory Press, 1999).

30. Paushkin, S. V., Patel, M., Furia, B. S., Peltz, S. W. & Trotta, C. R. Identification of a human endonuclease complex reveals a link between tRNA splicing and pre-mRNA 3' end formation. *Cell* **117,** 311–321 (Apr. 2004).

31. Weitzer, S. & Martinez, J. The human RNA kinase hClp1 is active on 3´ transfer RNA exons and short interfering RNAs. *Nature* **447,** 222–226 (May 2007).

32. Popow, J. *et al.* HSPC117 is the essential subunit of a human tRNA splicing ligase complex. *Science* **331,** 760–764 (Feb. 2011).

33. Popow, J., Jurkin, J., Schleiffer, A. & Martinez, J. Analysis of orthologous groups reveals archease and DDX1 as tRNA splicing factors. *Nature* **511,** 104–107 (June 2014).

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

35. Hopper, A. K., Pai, D. A. & Engelke, D. R. Cellular dynamics of tRNAs and their genes. *FEBS Letters* **584,** 310–317 (Jan. 2010).

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

37. Dittmar, K. A., Goodenbour, J. M. & Pan, T. Tissue-Specific Differences in Human Transfer RNA Expression. *PLoS Genetics* **2,** e221 (2006).

38. Lowe, T. M. & Eddy, S. R. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Research* **25,** 955–964 (Mar. 1997).

39. Chan, P. P. & Lowe, T. M. GtRNAdb: a database of transfer RNA genes detected in genomic sequence. *Nucleic Acids Research* **37,** D93–D97 (Jan. 2009).

40. Gu, W. tRNAHis maturation: An essential yeast protein catalyzes addition of a guanine nucleotide to the 5' end of tRNAHis. *Genes & Development* **17,** 2889–2901 (Dec. 2003).

41. Jackman, J. E. & Alfonzo, J. D. Transfer RNA modifications: nature's combinatorial chemistry playground. *Wiley Interdisciplinary Reviews: RNA* **4,** 35–48 (Nov. 2012).

42. Lee, J.-H., Ang, J. K. & Xiao, X. Analysis and design of RNA sequencing experiments for identifying RNA editing and other single-nucleotide variants. *RNA* **19,** 725–732 (June 2013).

43. Gustilo, E. M., Vendeix, F. A. & Agris, P. F. tRNA's modifications bring order to gene expression. *Current Opinion in Microbiology* **11,** 134–140 (Apr. 2008).

44. Hanada, T. *et al.* CLP1 links tRNA metabolism to progressive motor-neuron loss. *Nature,* 1–7 (Mar. 2013).

45. Weitzer, S., Hanada, T., Penninger, J. M. & Martinez, J. CLP1 as a novel player in linking tRNA splicing to neurodegenerative disorders. *Wiley Interdisciplinary Reviews: RNA* **6,** n/a–n/a (Aug. 2014).

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

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

48. Juhling, F. *et al.* tRNAdb 2009: compilation of tRNA sequences and tRNA genes. *Nucleic Acids Research* **37,** D159–D162 (Jan. 2009).

49. Hafner, M. *et al.* Transcriptome-wide Identification of RNA-Binding Protein and MicroRNA Target Sites by PAR-CLIP. *Cell* **141,** 129–141 (Apr. 2010).

50. König, J., Zarnack, K., Luscombe, N. M. & Ule, J. Protein-RNA interactions: new genomic technologies and perspectives. *Nature Publishing Group* **13,** 77–83 (Feb. 2011).

51. 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).

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

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

54. Arimbasseri, A. G., Rijal, K. & Maraia, R. J. Transcription termination by the eukaryotic RNA polymerase III. *BBA - Gene Regulatory Mechanisms* **1829,** 318–330 (Mar. 2013).

55. Nielsen, S., Yuzenkova, Y. & Zenkin, N. Mechanism of eukaryotic RNA polymerase III transcription termination. *Science* **340,** 1577–1580 (June 2013).

56. Arimbasseri, A. G. & Maraia, R. J. Distinguishing Core and Holoenzyme Mechanisms of Transcription Termination by RNA Polymerase III. *Molecular and Cellular Biology* **33,** 1571–1581 (Mar. 2013).

57. 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).

58. Arimbasseri, A. G. & Maraia, R. J. Mechanism of Transcription Termination by RNA Polymerase III Utilizes a Non-template Strand Sequence-Specific Signal Element. *Molecular Cell* **58,** 1124–1132 (June 2015).

59. Teplova, M. *et al.* Structural Basis for Recognition and Sequestration of UUUOH 3´ Temini of Nascent RNA Polymerase III Transcripts by La, a Rheumatic Disease Autoantigen. *Molecular Cell* **21,** 75–85 (Jan. 2006).

60. Lorenz, R. *et al.* ViennaRNA Package 2.0. *Algorithms for molecular biology : AMB* **6,** 26 (Nov. 2011).

61. Machnicka, M. A. *et al.* MODOMICS: a database of RNA modification pathways–2013 update. *Nucleic Acids Research* **41,** D262–7 (Jan. 2013).

62. Torres, A. G. *et al.* Inosine modifications in human tRNAs are incorporated at the precursor tRNA level. *Nucleic Acids Research,* 1–13 (Apr. 2015).

63. 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).

64. Baltz, A. G. *et al.* The mRNA-Bound Proteome and Its Global Occupancy Profile on Protein-Coding Transcripts. *Molecular Cell* **46,** 674–690 (June 2012).

65. Wu, J. & Hopper, A. K. Healing for destruction: tRNA intron degradation in yeast is a two-step cytoplasmic process catalyzed by tRNA ligase Rlg1 and 5'-to-3' exonuclease Xrn1. *Genes & Development* **28,** 1556–1561 (July 2014).

66. Maraia, R. J. & Bayfield, M. A. The La Protein-RNA Complex Surfaces. *Molecular Cell* **21,** 149–152 (Jan. 2006).

67. Iben, J. R. & Maraia, R. J. tRNAomics: tRNA gene copy number variation and codon use provide bioinformatic evidence of a new anticodon:codon wobble pair in a eukaryote. *RNA* **18,** 1358–1372 (July 2012).

68. Spitzer, J., Landthaler, M. & Tuschl, T. *Rapid Creation of Stable Mammalian Cell Lines for Regulated Expression of Proteins Using the Gateway® Recombination Cloning Technology and Flp-In T-REx® Lines* 1st ed. (Elsevier Inc., 2013).

69. Garzia, A., Meyer, C., Morozov, P., Sajek, M. & Tuschl, T. Optimization of PAR-CLIP for transcriptome-wide identification of binding sites of RNA-binding proteins. *Methods,* 1–17 (Oct. 2016).

70. Hopper, A. K. & Shaheen, H. H. A decade of surprises for tRNA nuclear–cytoplasmic dynamics. *Trends in Cell Biology* **18,** 98–104 (Mar. 2008).

71. Liu, Y. *et al.* C3PO, an Endoribonuclease That Promotes RNAi by Facilitating RISC Activation. *Science* **325,** 750–753 (Aug. 2009).

72. Aoki, K. *et al.* A novel gene, Translin, encodes a recombination hotspot binding protein associated with chromosomal translocations. *Nature Genetics* **10,** 167–174 (June 1995).

73. Ishida, R. *et al.* A role for the octameric ring protein, Translin, in mitotic cell division. *FEBS Letters* **525,** 105–110 (Aug. 2002).

74. Stein, J. M. Behavioral and Neurochemical Alterations in Mice Lacking the RNA-Binding Protein Translin. *Journal of Neuroscience* **26,** 2184–2196 (Feb. 2006).

75. Claussen, M., Koch, R., Jin, Z. Y. & Suter, B. Functional Characterization of Drosophila Translin and Trax. *Genetics* **174,** 1337–1347 (Sept. 2006).

76. Ye, X. *et al.* Structure of C3PO and mechanism of human RISC activation. *Nature Structural & Molecular Biology* **18,** 650–657 (May 2011).

77. Tian, Y. *et al.* Multimeric assembly and biochemical characterization of the Trax–translin endonuclease complex. *Nature Structural & Molecular Biology* **18,** 658–664 (May 2011).

78. Parizotto, E. A., Lowe, E. D. & Parker, J. S. Structural basis for duplex RNA recognition and cleavage by Archaeoglobus fulgidus C3PO. *Nature Structural & Molecular Biology,* 1–9 (Jan. 2013).

79. Li, L. *et al.* The translin–TRAX complex (C3PO) is a ribonuclease in tRNA processing. *Nature Structural & Molecular Biology* **19,** 824–830 (July 2012).

80. Ye, X. *et al.* Structure of C3PO and mechanism of human RISC activation. *Nature Publishing Group* **18,** 650–657 (May 2011).

81. Lee, A. S. Y., Kranzusch, P. J. & Cate, J. H. D. eIF3 targets cell-proliferation messenger RNAs for translational activation or repression. *Nature* **522,** 111–114 (Apr. 2015).

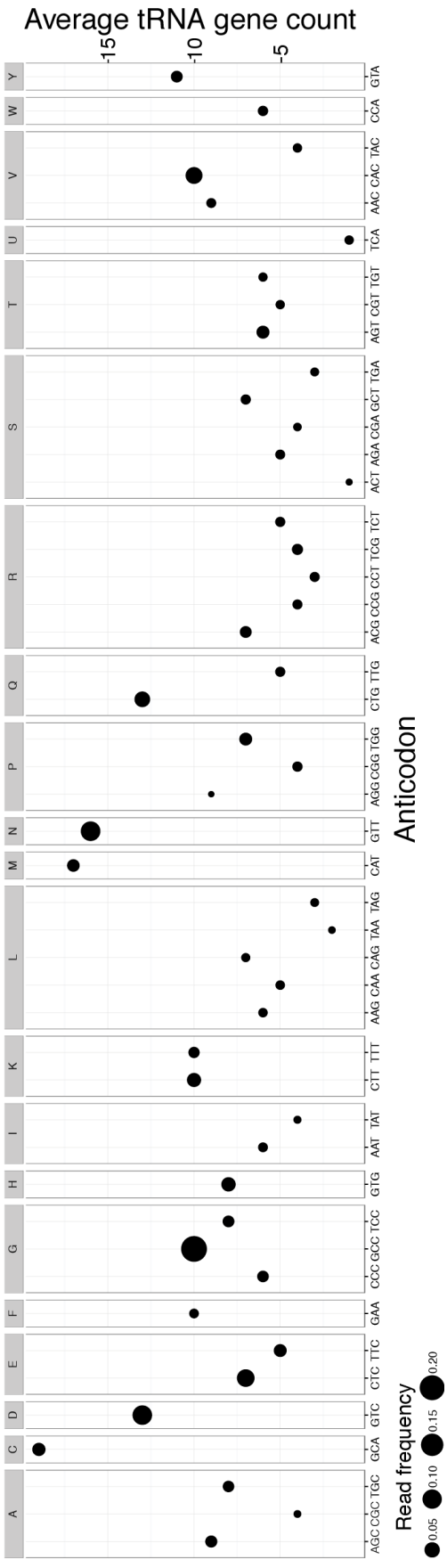
82. Hafner, M. *et al.* Identification of mRNAs bound and regulated by human LIN28 proteins and molecular requirements for RNA recognition. *RNA* **19,** 613–626 (May 2013).

83. Ran, F. A. *et al.* Genome engineering using the CRISPR-Cas9 system. *Nature Protocols* **8,** 2281–2308 (Oct. 2013).

84. Trapnell, C. *et al.* Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature Protocols* **7,** 562–578 (Mar. 2012).

Appendix A

Expanded version of figure 3.1


Average tRNA gene count
-15
-10
-5
Anticodon
A
C
D
E
F
G
H
I
K
L
M
N
P
Q
R
S
T
U
V
W
Y
AGC CGC TGC
GCA
GTC
CTC TTC
GAA
CCC GCC TCC
GTG
AAT TAT
CTT TTT
AAG CAA CAG TAA TAG
CAT
GTT
AGG CGG TGG
CTG TTG
ACG CCG CCT TCG TCT
ACT AGA CGA GCT TGA
AGT CGT TGT
TCA
AAC CAC TAC
CCA
GTA
Read frequency
0.05
0.10
0.15
0.20