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 regulated 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 atlas, and determine expression levels of mature and processing intermediates of tRNAs in human cells.

The technical advances provide by hydro-tRNAseq are applied towards the molecular diagnosis of a genetic neurodevelopmental syndrome, caused by mutations in the tRNA processing factor, CLP1. Since then, it has also been widely

used on multiple other fronts, some which are outlined in the appendix of this thesis.

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

First, I would like to thank my

Table of Contents

List of Figures

List of Tables

Glossary

List of Abbreviations

ChIP-seq	chromatin immunoprecipitation sequencing.
ncRNA	noncoding RNA.
OH	hydroxyl.
POLR3	human RNA polymerase III.
pre-tRNA	precursor tRNA.
RNP	ribonucleoprotein.
RT	reverse transcriptase.
tRNA	transfer RNA.

List of Terms

tRNA isoacceptor	tRNA molecules that decode synonymous codons.

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 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. First, I developed hydro-tRNAseq, a modified small RNA sequencing protocol based on partial alkaline hydrolysis of input RNA that succeeded in identifying and quantifying tRNAs.

I used the results of this approach to annotate and curate all mature and pre-tRNAs. Since tRNA processing, such as precursor trimming and intron removal, is a fast process[**Foretek:2016ea**], we also aimed to enrich specifically for pre-

tRNAs in order to identify and annotate the corresponding unique tRNA gene template. Thus, we performed PAR-CLIP on SSB, a conserved and ubiquitous protein involved in 3' tRNA processing [**Bayfield:2010cs**, 46, 47]

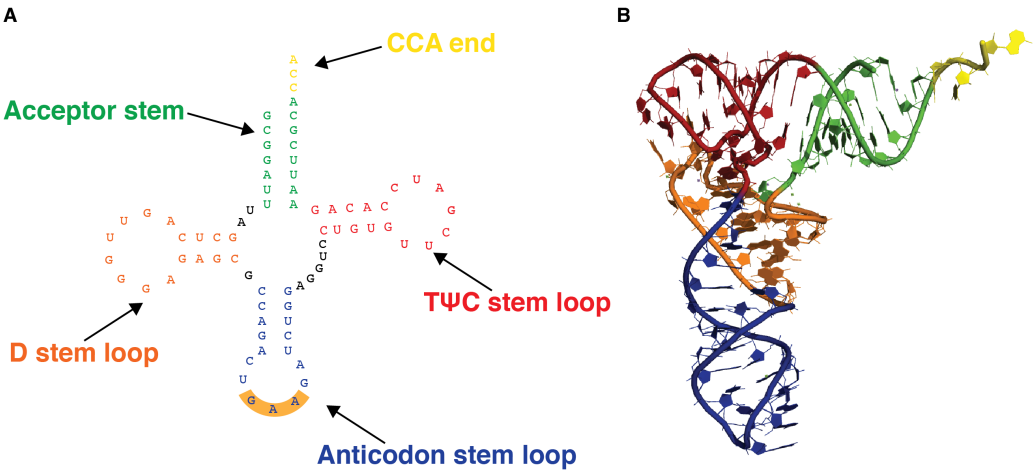

Figure 1.1: tRNA structure. (A) tRNA transcripts, such as the phenylalanine tRNA shown here, adopt the typical "cloverleaf" secondary, which in turns adopts an L-shaped tertiary structure as shown in (B). The structurally conserved stems and stemloops are indicated in A, color-coded, and their coordinates are reflected in the 3-dimensional structure in B (PDB 1EHZ).

1.2 tRNA and disease

talk about javier's papers, schimmel and dreyfuss review

1.3 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' triphosphate. In humans, a minority of tRNA transcripts harbor introns (see section refintrons). 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 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 (2.12), 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. 1.2**)[34–36].

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].

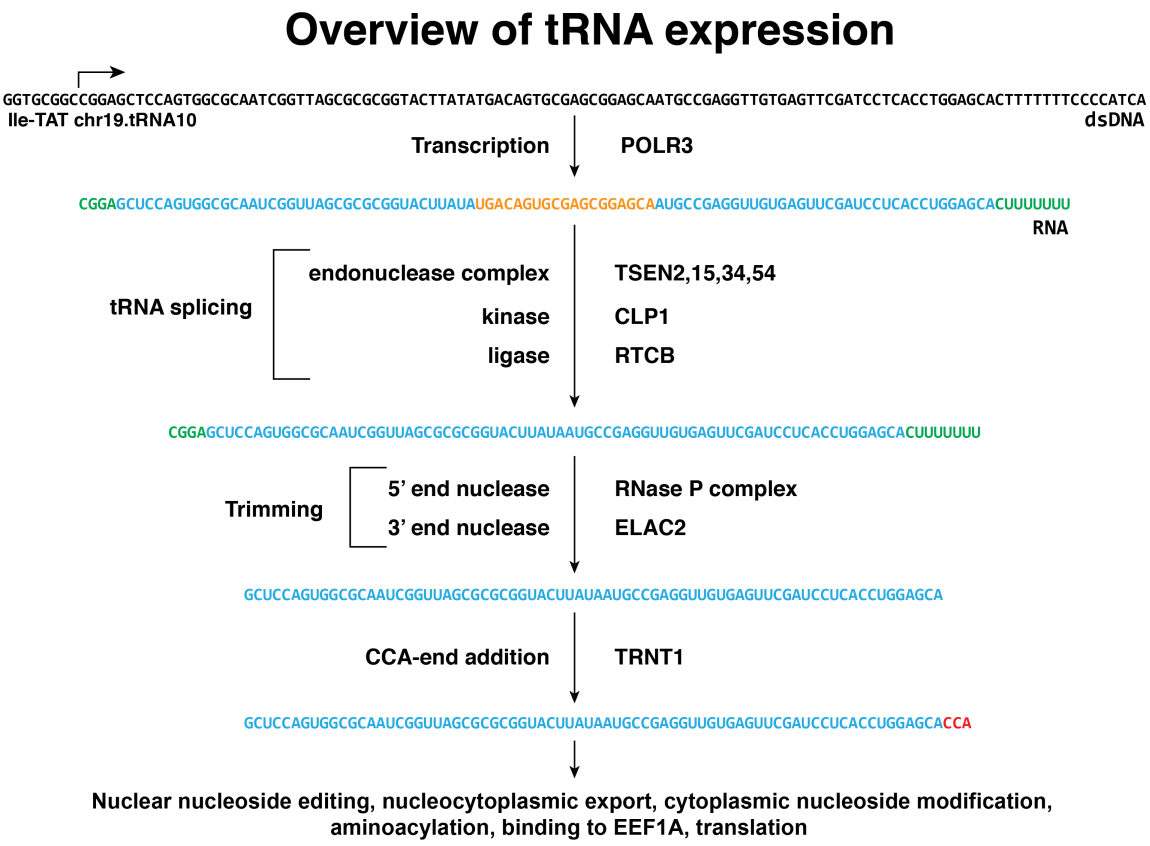

Figure 1.2: Overview of tRNA biogenesis and processing. tRNAs are transcribed by POLR3. If present, tRNA introns are removed by the tRNA splicing complex, and mature halves are ligated by the tRNA ligase (RTCB). Pre-tRNA leaders are trimmed by the RNase P complex, and 3' trailer by ELAC2. The 3' terminal 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.

1.4 tRNA sequencing

This 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-

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 1.2, 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.

1.5 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 that use 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 modifications21, 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 [**Hanada:2013bk**, **Weitzer:2014bi**, **Karaca:2014em**].

1.6 Small RNA sequencing protocol

In order to obtain RNA-seq reads, I decided to first apply the well established protocol for sequencing small RNAs, established in the Tuschl lab [44] (**Fig. 1.3**). The experimental procedure takes advantage of the 5' monophosphate (p) and 3' OH groups present in small RNAs, such as micro RNAs (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 on Illumina instruments using sequencing primer sites present in the adapter sequences. The different sequences of the 3' and 5' adapters preserves 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.

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:

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

2. mature tRNAs and miRNAs both have a 5' monophosphate and 3' OH, which

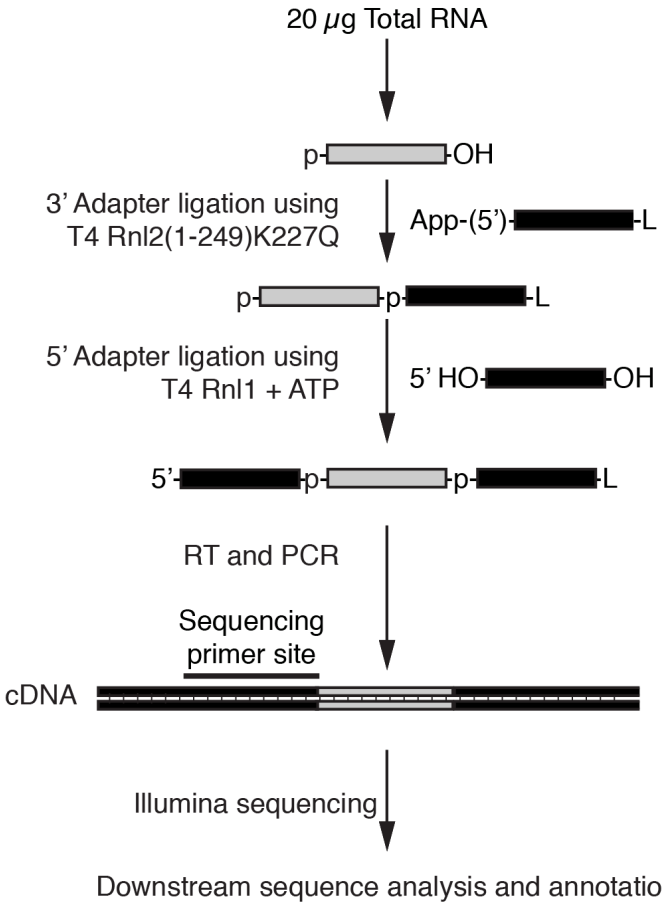

Figure 1.3: Small RNA sequencing protocol. Schematic overview of the conventional small RNA sequencing protocol, as it has been described previously [44]

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 protocol.

RNA type	% Total reads	Mean length (nt)
rRNA	35.8%	60.5
no match	24.1%	76.2
no annotation	17.8%	64.2
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 RNAseq 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 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 (e.g. 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 is

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

where n = number of modified nucleosides affecting RT. Given that full length tRNAs are longer than the hydrolysis-derived fragments, and modifications are usually concentrated in the loops of the tRNA (see (**Fig. 1.1** and **Fig. 2.22**), 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 our deepest dataset (**Table 2.1**). We named this procedure hydro-tRNAseq (**Fig. 2.1A**).

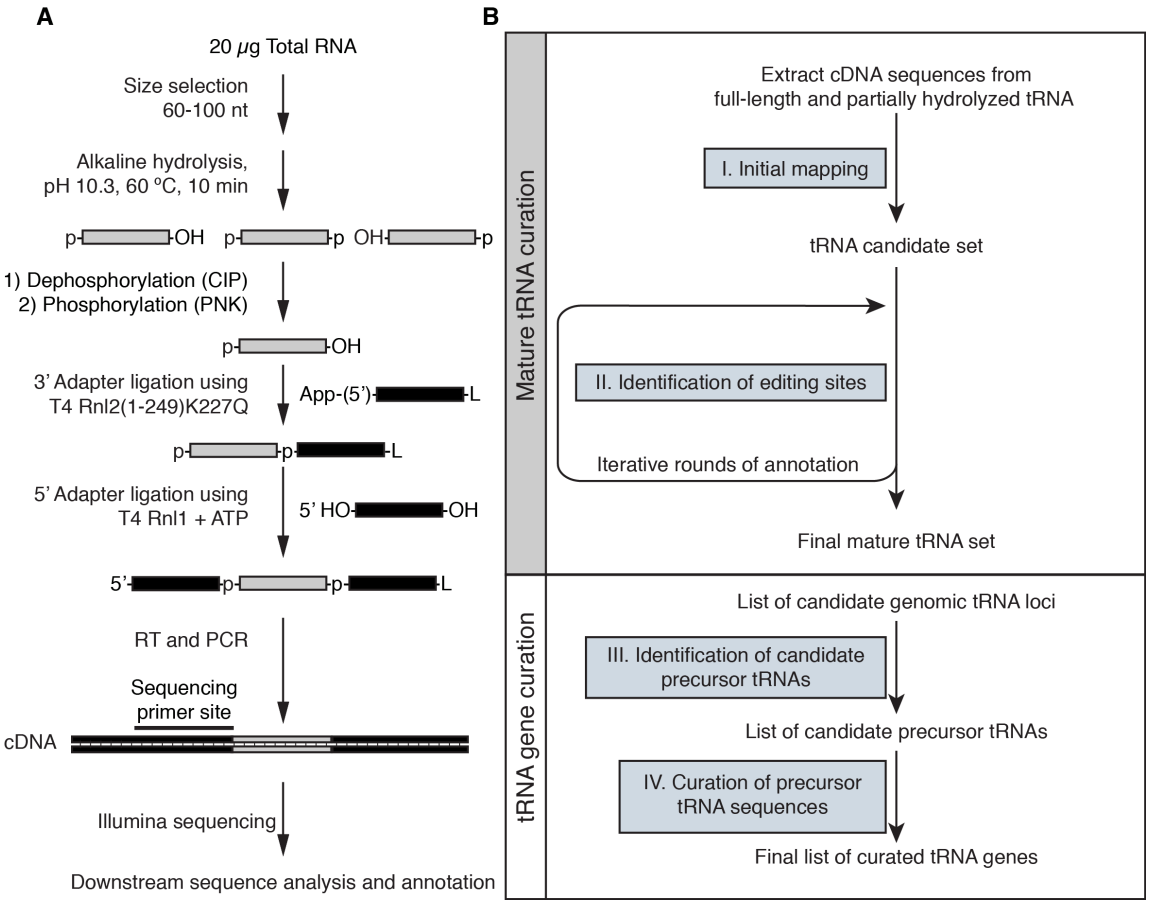

Figure 2.1: Hydro-tRNAseq experimental and bioinformatic pipeline for tRNA annotation and reference transcript curation by hydro-tRNAseq.(A) tRNAs and pre-tRNAs were size-selected from HEK293 total RNA and subjected to limited alkaline hydrolysis, followed by dephosphorylation, rephosphorylation and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

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%
tRNA prec	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, we had to build a bioinfroamtic 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 (`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 [**Juhling:2009ip**, 40], 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 1-2 than 0. 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 (**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 (Fig. S2), suggesting that our analysis was exhaustive, at least for tRNAs in HEK293 cells.

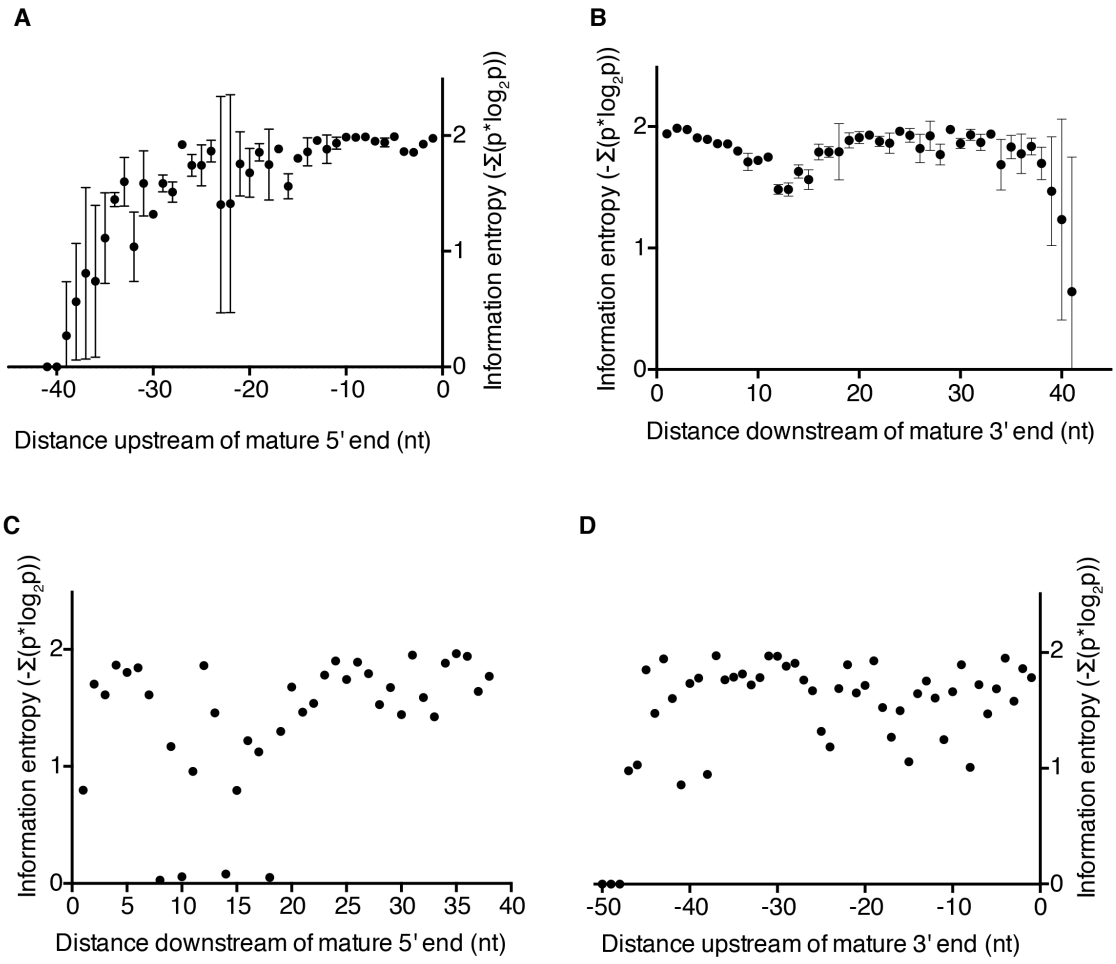

Figure 2.2: Information entropy in pre-tRNA segments and mature body (A,B) Information entropy $H = -\sum_{i=1}^{n} p(i) * log(p(i))$, (where p is the frequency of each nucleotide at a given position, i, and n the total number of transcripts) was calculated using read evidence from hydro-tRNAseq (four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

2.3 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 datasets.

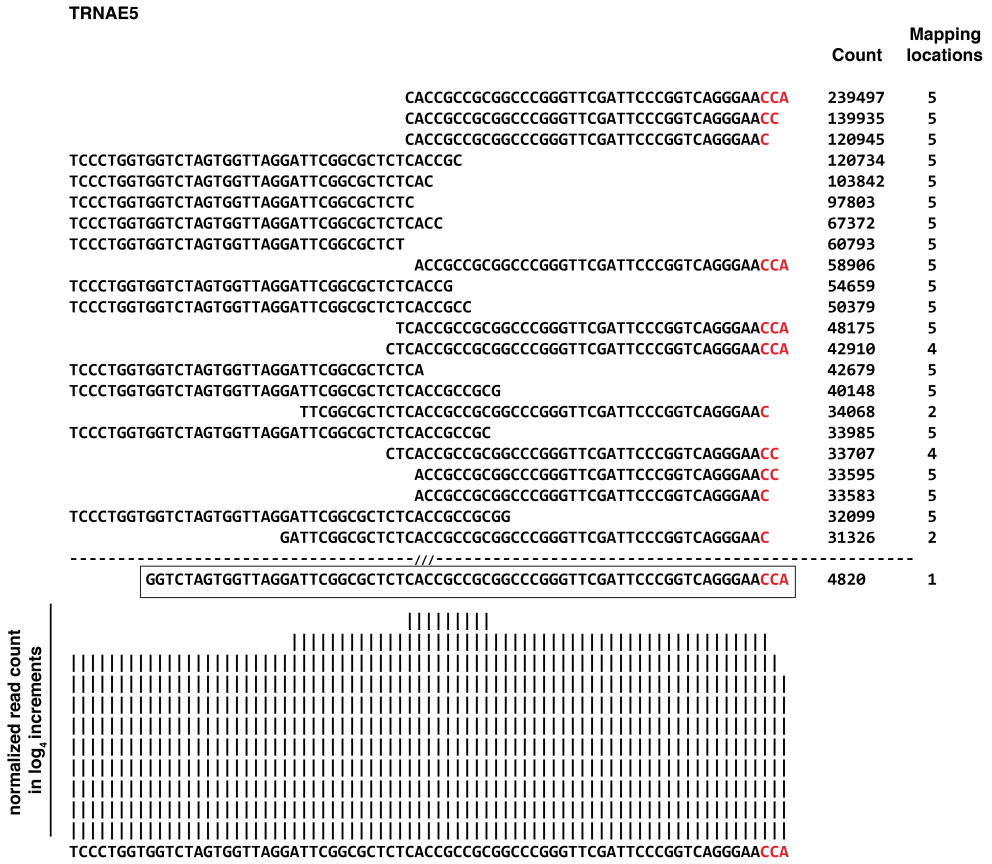

Figure 2.3: Mature tRNA alignment. An indicative alignment to mature TRNAE5 is shown including, sequences of reads, abundance, mapping locations, and $\log_4$ bin-normalized abundance.

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, 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.

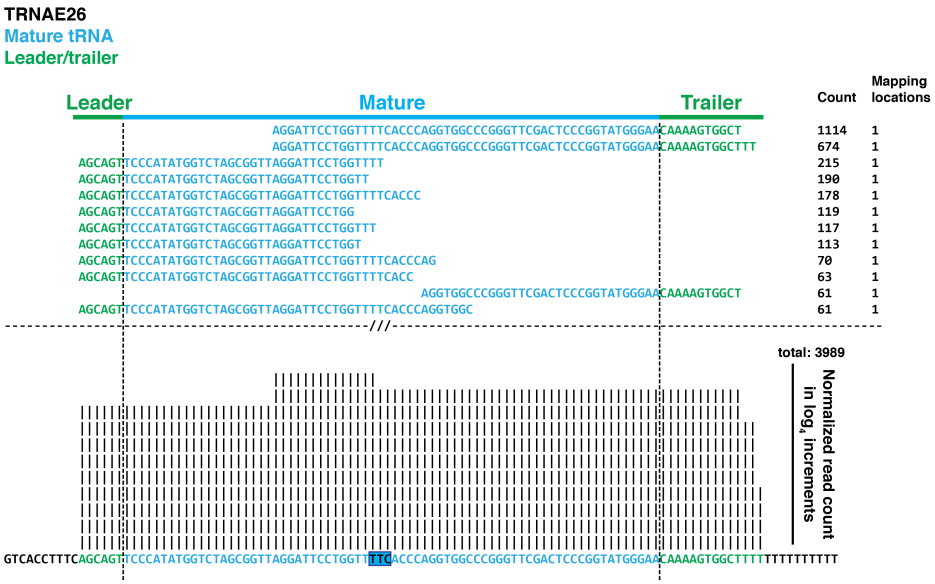

Figure 2.4: Pre-tRNA alignment. An indicative alignment to pre-tRNA TRNAE5 is shown. Mature sequence is in blue and precursor-specific sequence segments in green. The anticodon is boxed in blue for orientation purposes.

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 2.9, 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, 2.1**). This raised the possibility that we might have missed reads corresponding to lowly expressed or very rapidly processed pre-tRNAs. We did this by performing

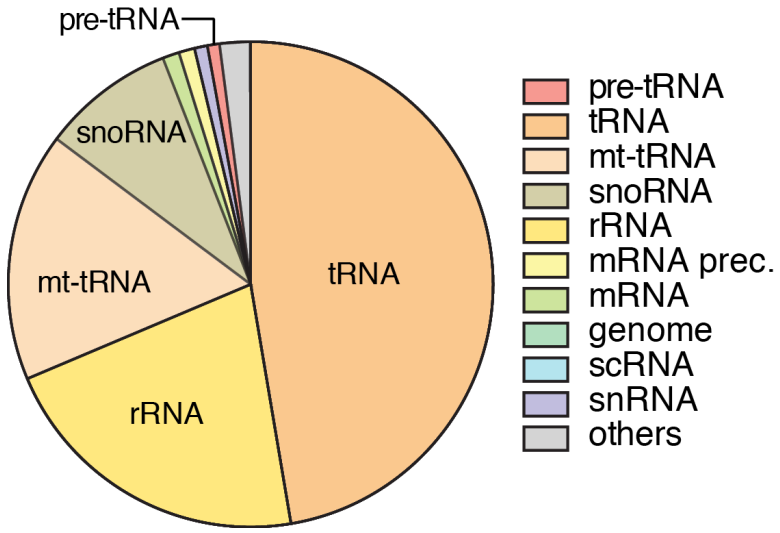

Figure 2.5: Composition of hydro-tRNAseq libraries. Total RNA composition of the 60-100 nt size fraction from hydro-tRNAseq according to RNA classes.

photoactivatable-ribonucleoside-enhanced crosslinking and immunoprecipitation

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

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 [**Konig:2011jv**]. 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**[**Hafner:2010kr**]). 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).

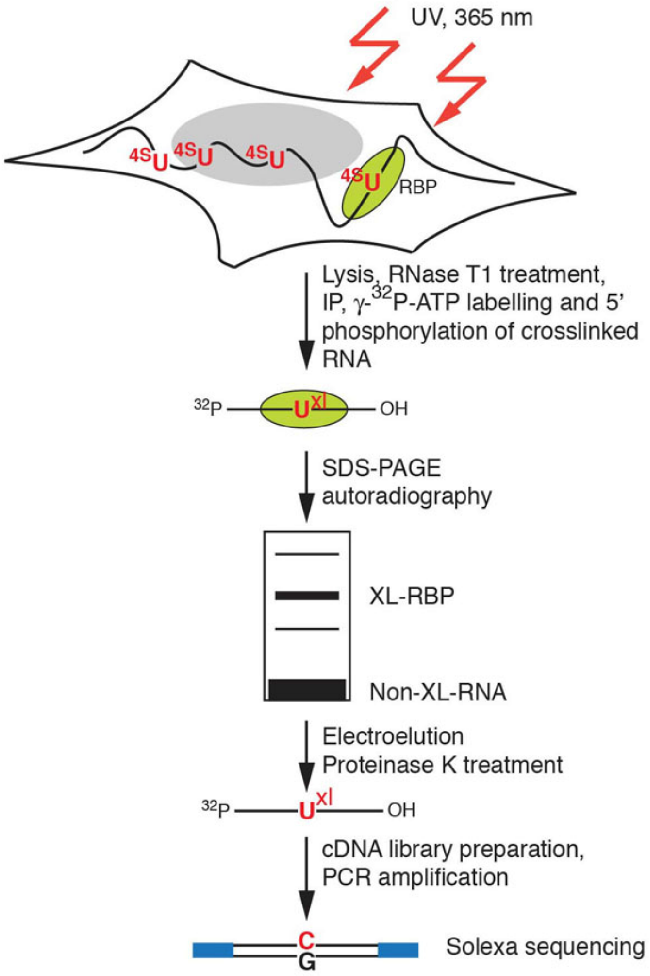

Figure 2.6: PAR-CLIP. Outline

2.6 SSB PAR-CLIP

Therefore, 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). SSB has been shown biochemically to bind to the short 3' oligoU tails [46] that acts as termination signal for POLR3 [45]. Therefore, we reasoned that SSB should bind all tRNA precursors, and that if we could isolate its targets, we would be able to reliably identify transcribed tRNA loci.

SSB exhibited a striking binding preference for pre-tRNAs and showed a dras-

tic enrichment in precursor tRNAs compared to hydro-tRNAseq (**Fig. 2.8**), which confirmed our hypothesis, as well as previous observations [47]. 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. 2.8**), the overwhelming majority of which showed the characteristic T-to-C transition, indicative of crosslinking (**Fig. 2.8**, table S3.

(**Fig. 2.7**

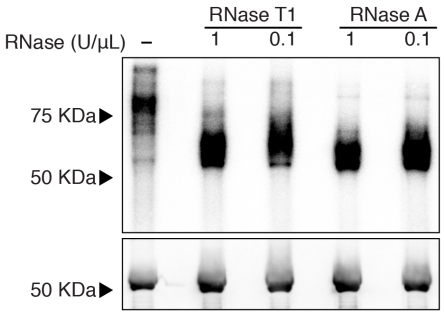

Figure 2.7: SSB crosslinking to RNA. Phosphorimage of SSB-crosslinked to radiolabeled RNA. PAR-CLIP was performed using RNase A or RNase T1, at two different concentrations to account for possible biases of RNase treatment conditions. Libraries from PAR-CLIP using 1 U/μL of RNase A and RNase T1 were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

The vast majority of crosslinking sites in pre-tRNAs were concentrated, as expected, in the oligoU tract of the 3' trailer sequence (**Fig. 2.9**A,B). We also found

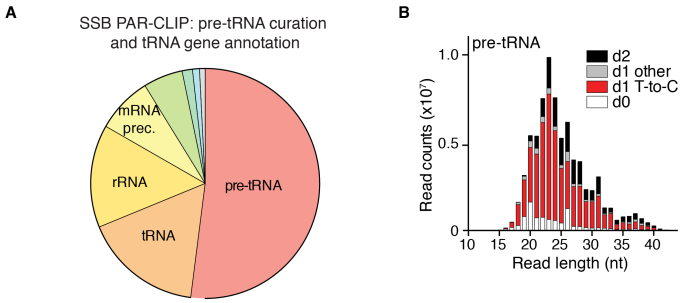

Figure 2.8: figure2cd. placeholder

that SSB crosslinked to the 5' segment of the mature tRNA body at conserved sites in the D-stemloop (**Fig. 2.9**B), which is a novel finding, hinted at by a report proposing that the affinity of SSB for a full-length pre-tRNA cannot be explained solely by its binding to the 3' oligoU tract [47]. The other major target of SSB was 5S ribosomal RNA (rRNA), which is the only POLR3-transcribed rRNA, and as such also terminates with an oligoU stretch to which SSB crosslinked (Fig. S3).

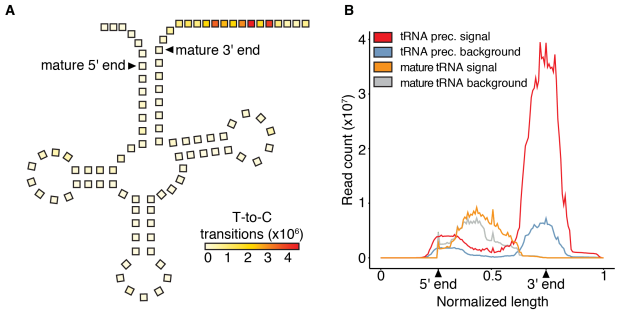

Figure 2.9: figure2ef. placeholder

2.7 tRNA gene annotation

We combined hydro-tRNAseq and SSB PAR-CLIP to identify actively transcribed tRNA genes (genomic locations that give rise to a supported pre-tRNAs). We confidently identified 288 tRNA genes as the intersection of 4 replicates of hydro-tRNAseq (**Fig. 2.10A**), and 349 tRNA genes as the intersection of two SSB PAR-

CLIP experiments. Of note, SSB PAR-CLIP confirmed the expression of an additional 7 tRNA genes that were not supported in hydro-tRNAseq replicate (e.g. **Fig. 2.10B**), further showcasing the complementarity of the two approaches. We observed a strong correlation of pre-tRNA abundances between SSB PAR-CLIP and hydro-tRNAseq (Pearson R = 0.72; **Fig. 2.10D**), providing confidence that SSB PAR-CLIP quantitatively detected pre-tRNAs, without introducing biases (e.g. artificially enriching for lowly expressed pre-tRNAs). Instead, we observed no strong correlation between precursor and mature tRNA read counts in either of the two techniques (R < 0.2; Fig. S4). The correlation of identified isoacceptor counts between SSB PAR-CLIP and hydro-tRNAseq was virtually perfect (Pearson R = 0.99; **Fig. 2.15C**), ruling out the introduction of a pronounced systematic bias from our hydrolysis-based protocol. Some anticodons seemed to be served by multiple isodecoders (e.g. 19 isodecoders for $tRNA^{Ser}_{GCA}$), while others only from one (e.g. $tRNA^{Ser}_{ACT}$;(**Fig. 2.15B**, **Fig. 2.15**).Selenocysteine was the only amino acid that, in our data, was decoded by only one tRNA gene.

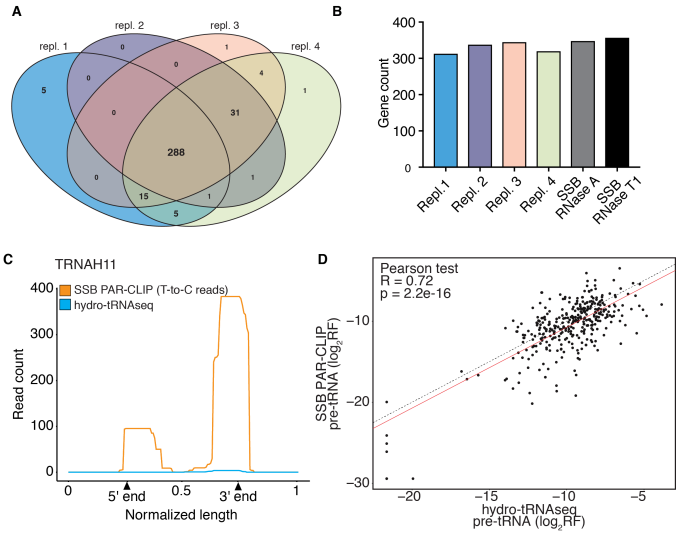

Figure 2.10: figure3. placeholder

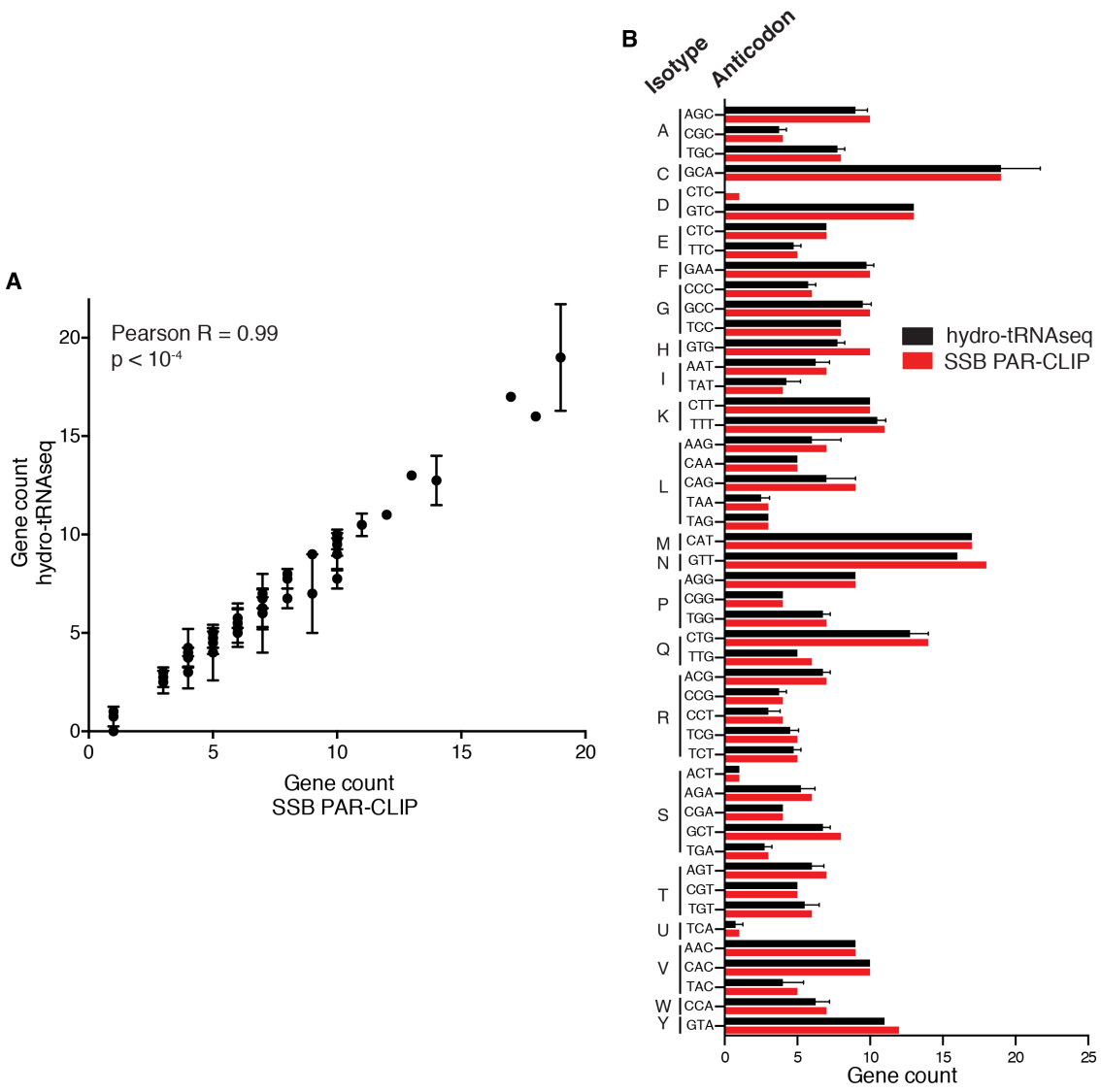

Figure 2.11: figure4. placeholder

2.8 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 we tried to address. We size-selected at a higher size window to avoid contamination

by tRNA-derived fragments (as compared to [27]). Also, we 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 us to differentiate RT stops from fragment ends. Additionally, we did not bias our sequencing protocol towards mature tRNAs, but instead we captured more precursors by both RNAseq and more importantly PAR-CLIP methods, allowing us to perform a deeper precursor 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. 2.12**).

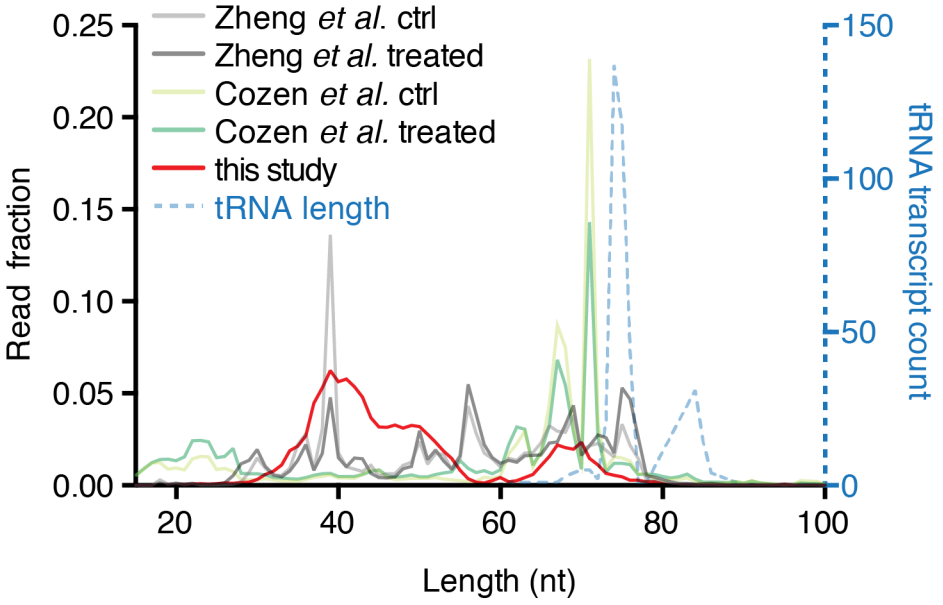

Figure 2.12: supp6. placeholder

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. 2.13**). Also, SSB PAR-CLIP was more sensitive in identifying tRNA genes, detecting 349

genes as compared to 159 and 212 in the other methods.Also, SSB PAR-CLIP was more sensitive in identifying tRNA genes, detecting 349 genes as compared to 159 and 212 in the other methods.

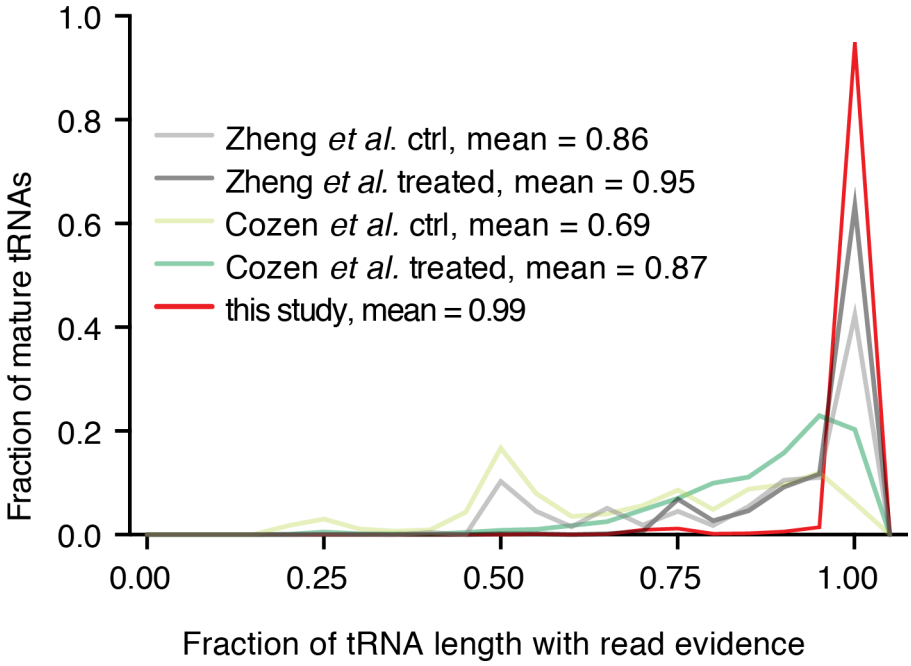

Figure 2.13: supp7. placeholder

2.9 Applications and biological insights

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

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

This was assumed in the absence of tRNA sequencing data and because in yeast it seems that all tRNA genes are epxressed and seem to contribute equally

to the mature tRNA pool. So, this result undesrcore the need for caution when reporting tRNA abunandce measurements and estimates.

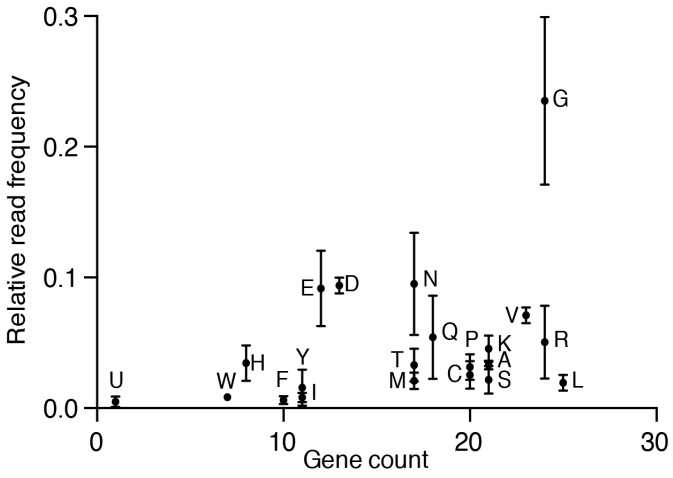

Figure 2.14: figure4A. placeholder

Although tRNA isotypes with higher relative abundances generally tend to have higher tRNA gene numbers, we did not observe a clear linear correlation between read frequency and gene count (R = 0.12; **Fig. 2.14**)), like it has been reported before [12]. We then focused on the number of tRNA isoacceptors per amino acid, and isodecoders (tRNAs with the same anticodon sequence) per anticodon. We noticed a wide range of pre-tRNA counts per isoacceptor (**Fig. 2.15B**), with our data providing read evidence for 47 out of 62 coding codons (61 canonical and 1 selenocysteine TAG).

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

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

On this graph the data are broken down by aminoacid, which you can see

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

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

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

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

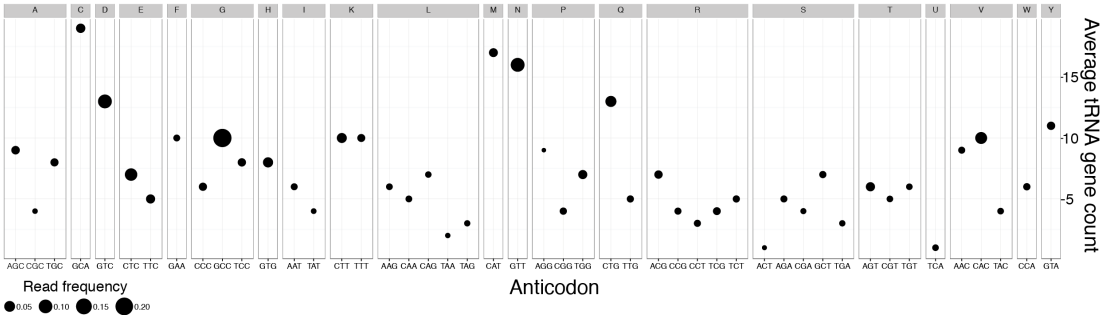

Figure 2.15: figure4D. placeholder

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

No good correlation (pearson coefficients < 0.2) as identified by our 2 separate techniques. (**Fig. 2.17**)

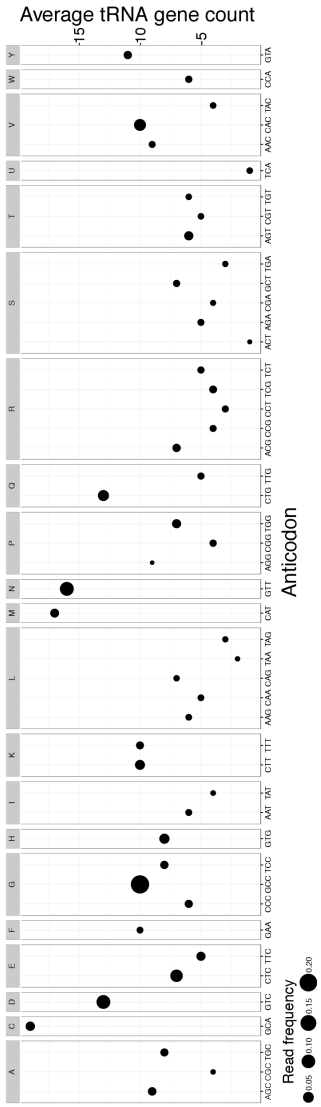

Figure 2.16: figure4Drot. placeholder

2.11 tRNA transcription initiation and termination

Besides tRNA gene annotation and quantification, our approach yielded insights about pre-tRNA 3' trailer sequences. Based on hydro-tRNAseq, we determined the median 5' leader and 3' trailer lengths to be 6 and 10 nt, respectively, with the trailer lengths showing a broader distribution (**Fig. 2.19A,B**). Interestingly, SSB PAR-CLIP revealed a subset of much longer trailers (**Fig. 2.19C**), suggesting that

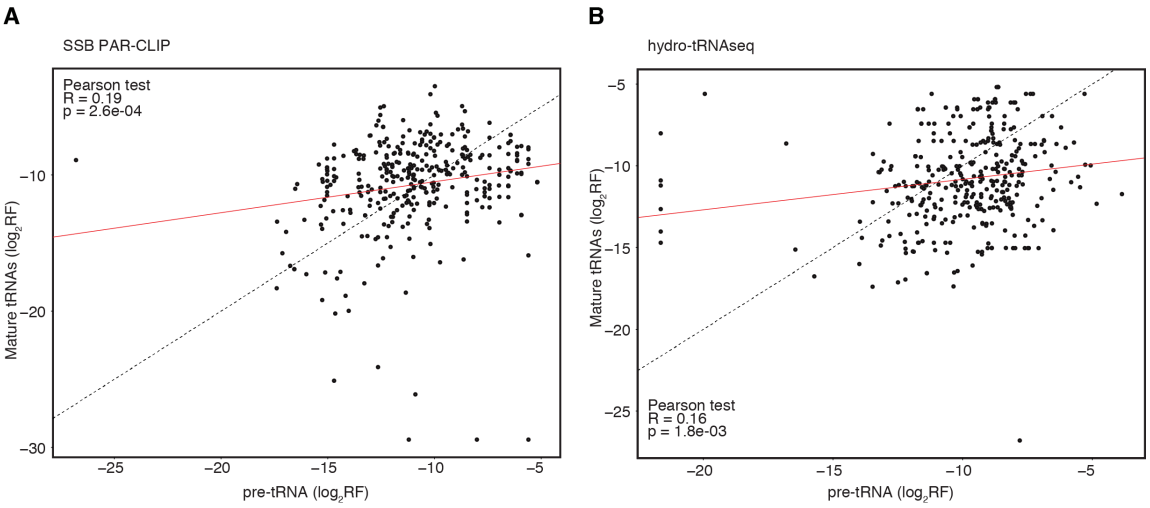

Figure 2.17: supp4. placeholder

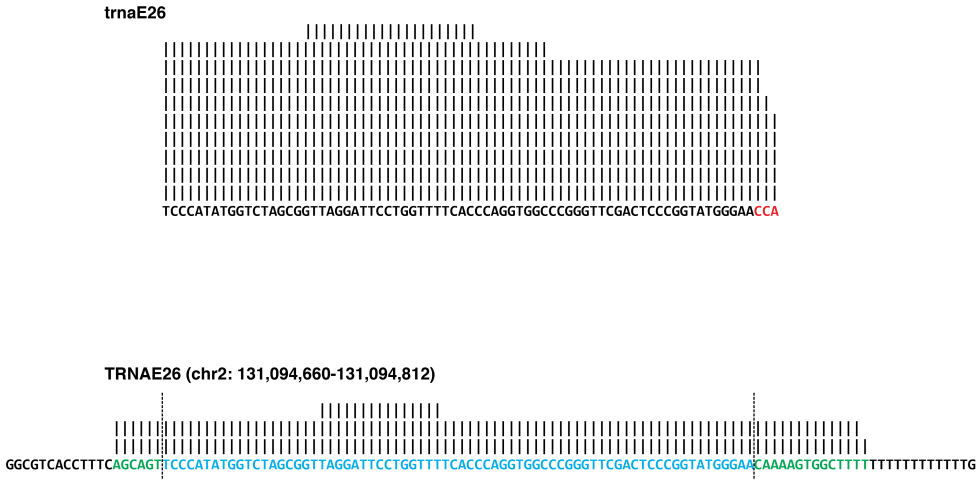

Figure 2.18: clp1 bar. placeholder

SSB PAR-CLIP captured the very initial steps of precursor tRNA processing, and accordingly that hydro-tRNAseq captures pre-tRNAs partially trimmed, either by ELAC2 (tRNase Z) or some other nuclease [34].

We next focused on the POLR3 oligoU termination signals. Various reports in the past have focused on the oligoU requirements for transcription termination in different species [48, 49]. SSB protected consistently a 3' 4 to 5 nt oligoU stretch, which was also confirmed by hydro-tRNAseq (**Fig. 2.20**). This is in agreement with previous *in vitro* results [46, 47, 50]. We also addressed the proposed re-

quirement for a stem-loop immediately upstream of the oligoU termination signal [49]. Secondary structure predictions for the trailer sequences with documented sequence evidence in hydro-tRNAseq and SSB PAR-CLIP did not detect predicted stable stem-loop structures for approximately half of all pre-tRNAs (**Fig. 2.21**). 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 [48].

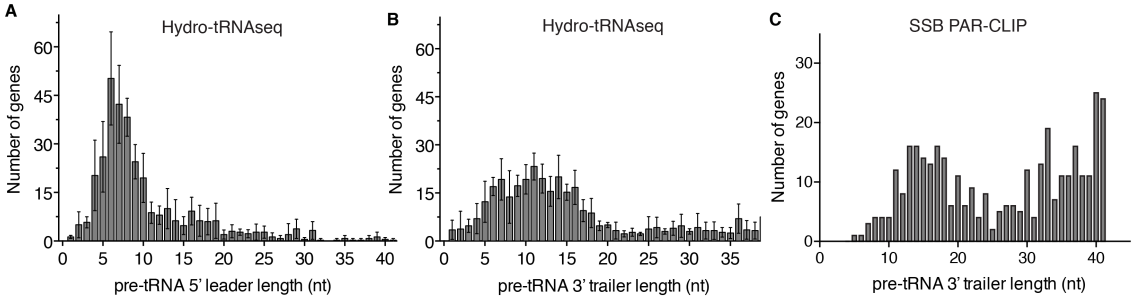

Figure 2.19: figure6. placeholder

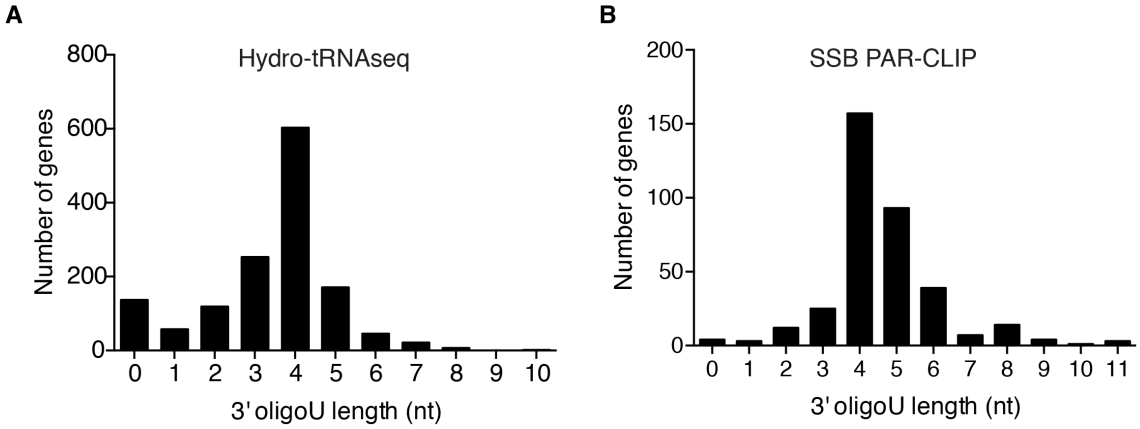

Figure 2.20: figure6de. placeholder

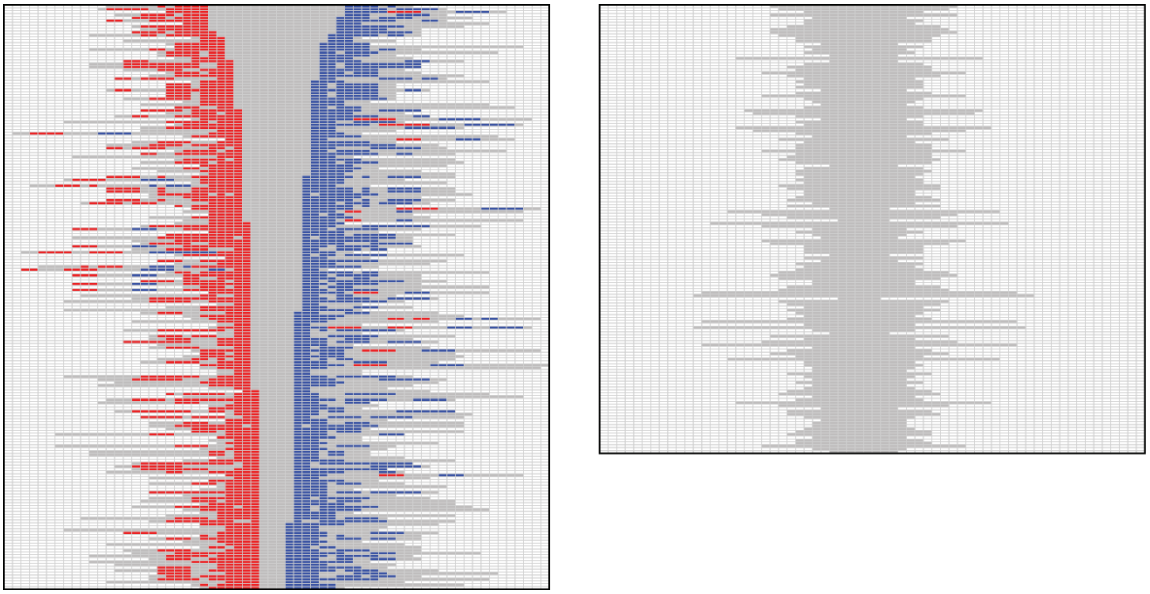

Figure 2.21: supp5. placeholder

2.12 Ribonucleotide modifications

RT across modified nucleotide-containing RNA leads to errors in cognate deoxynucleotide incorporation, revealed by mismatches in sequence reads upon mapping to reference genomic sequence. Read coverage across regions with a high degree of modifications may result in incomplete or largely uneven coverage. Therefore, we included in our mature tRNA reference the combination of all frequent mismatch signatures in all heavily modified positions. We reported the most frequently modified positions per tRNA gene (Table S5), and computed the frequencies of every nucleoside change per position across all tRNA genes (**Fig. 2.22**).

The majority of editing events were A-to-G transitions at the first position of the anticodon and at the position 3' to the anticodon (usually position 37). Both positions are known to be heavily modified, the former being deaminated to inosine, and the latter further modified (e.g. 1-methylinosine) [51]. In our data the majority

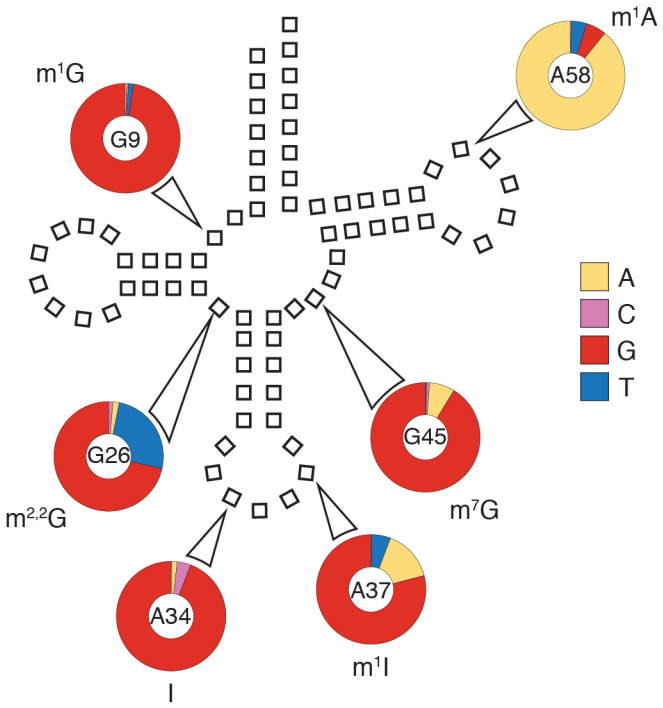

Figure 2.22: paper7a. placeholder

of the reads that mapped to the anticodon of the modified tRNAs contained the mismatches. To a lesser extent we could also detect 1-methyladenosine in the pseudouridine loop (returned as A-to-T or A-to-G), and various guanosine modifications at positions 9, 26, and 45, which most likely correspond to 1-methyl-, N2,N2-dimethyl-, and 7-methyl-guanosine, respectively [51].

The temporal resolution of tRNA modifications by RNAseq has begun to be addressed recently [52], however at a single modification level (inosine 34), and by using libraries relative poor in tRNA reads (<1% of total reads). We were appropriately poised to address this issue since our very deep sequencing set, in combination with our hierarchical annotation pipeline, offered the advantage of dissecting multiple modifications simultaneously. We focused on the inosine modifications, since they represented the majority of modified nucleosides. By inspecting read alignments with error distance 1 to the reference pre-tRNA, we noticed A-to-G

transition mismatches at position 34 in reads that retained the leader and trailer sequences of the precursor tRNA (**Fig. 2.23, top**). This confirmed that A34 deamination takes place at the precursor level, and therefore is a nuclear modification, as it has been previously reported [52]. Next, we noticed that 1-methyl-inosine at position 37 also appears at the precursor stage. Of note the A37 modification became apparent prior to A34, as the majority of the error distance 1 reads contained a mismatch at A37. Reads with two mismatches contained both modifications (**Fig. 2.23, bottom**)

In many instances we observe specific and highly abundant mismatchsignatures in pre-tRNAs, which can possibly allow us to determine and monitor by RNA-seq the editing hierarchy of tRNAs and the order of function of their modifying enzymes.

We observe editing in the D1 alignments. Those reads do not map at error distance 1, since we are following a hierarchical mapping, neither to they map back to the genome. This suggests that we can observe nucleotide editing at the precursor stage of tRNAs.

What we can say is that we observe tRNA editing leading to mismatches in RNA-seq specifc postitions edited many times more abundant than unedited This slide brings up another point. . .

We are far from experts on tRNA editing and modifications, but our protocols do allow for a careful annotation and overview of specific modifications. Those that lead to characteristic signatures of mismatches upon RNAseq In many instances we observe specific and highly abundant mismatchsignatures in pre-tRNAs, which can possibly allow us to determine and monitor by RNA-seq the editing hierarchy of tRNAs and the order of function of their modifying enzymes.

We observe editing in the D1 alignments. Those reads do not map at error

distance 1, since we are following a hierarchical mapping, neither to they map back to the genome. This suggests that we can observe nucleotide editing at the precursor stage of tRNAs.

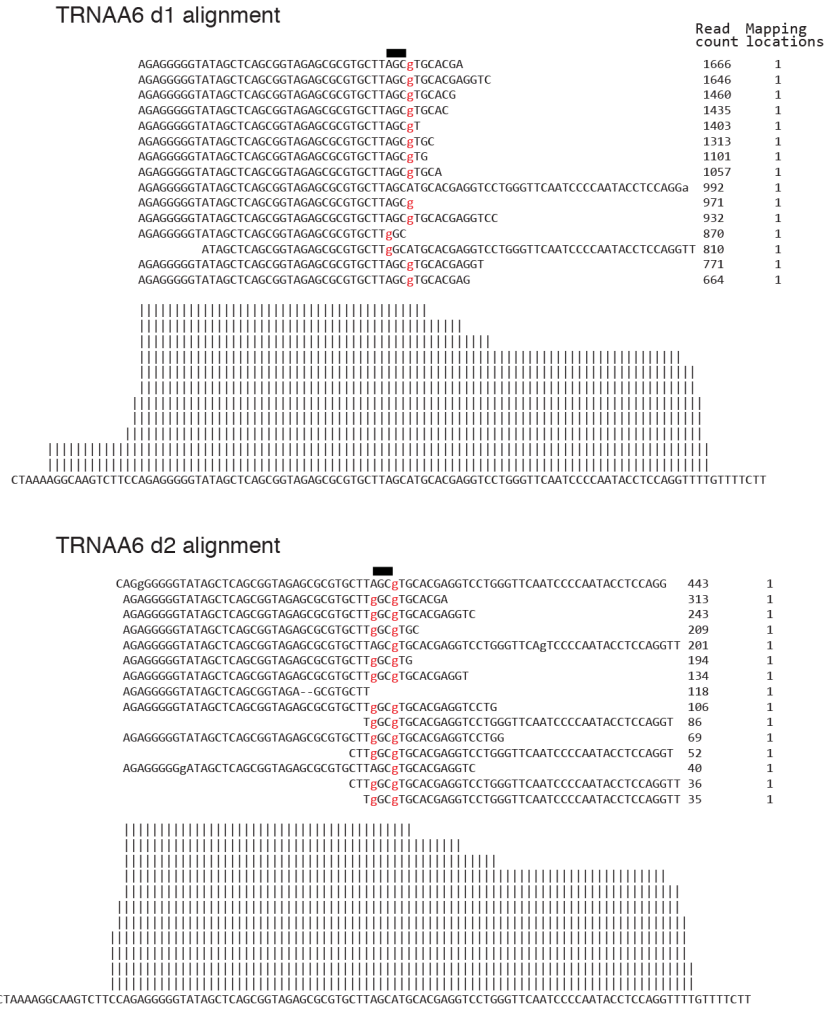

Figure 2.23: paper7b. placeholder

2.13 Annotation of intron-containing tRNA genes

Intron-containing tRNAs represent a particularly interesting set of tRNA genes, as mutations in their evolutionarily conserved, yet distinct, processing machinery have emerged recently as causes of severe neurodevelopmental syndromes,

such as pontocerebellar hypoplasia [53]. 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. We confirmed 26 out of 32 predicted intron-containing tRNAs by hydro-tRNAseq (**Fig. 2.24A**. Excluding any unknown biologically redundant mechanism, this suggests that the integrity of the tRNA splicing complex is essential for survival. To further confirm our observations, we coupled hydro-tRNAseq results with previously published PAR-CLIP data on the human tRNA ligase, RTCB [54]. Despite the shallow read depth of the dataset, we identified a crosslinked read peak at the anticodon loop of all intron-containing tRNAs annotated by our approaches. (**Fig. 2.24B,C**. One tRNA isotype and three isoacceptors are fully dependent on tRNA splicing, which suggests that despite functional redundancies that are often observed in tRNA processing steps, tRNA splicing is an esential [rocess for viability, at least in the cell system that we studied

2.14 CLP1

This came at an opportune moment because of the studies of CLP1

- Recessive mutations in children with neurodevelopmental disorders were localized to tRNA splicing factor CLP1
- Exome sequencing identified R140H recessive mutation
- R140H reduced interaction between CLP1 and the TSEN splicing complex

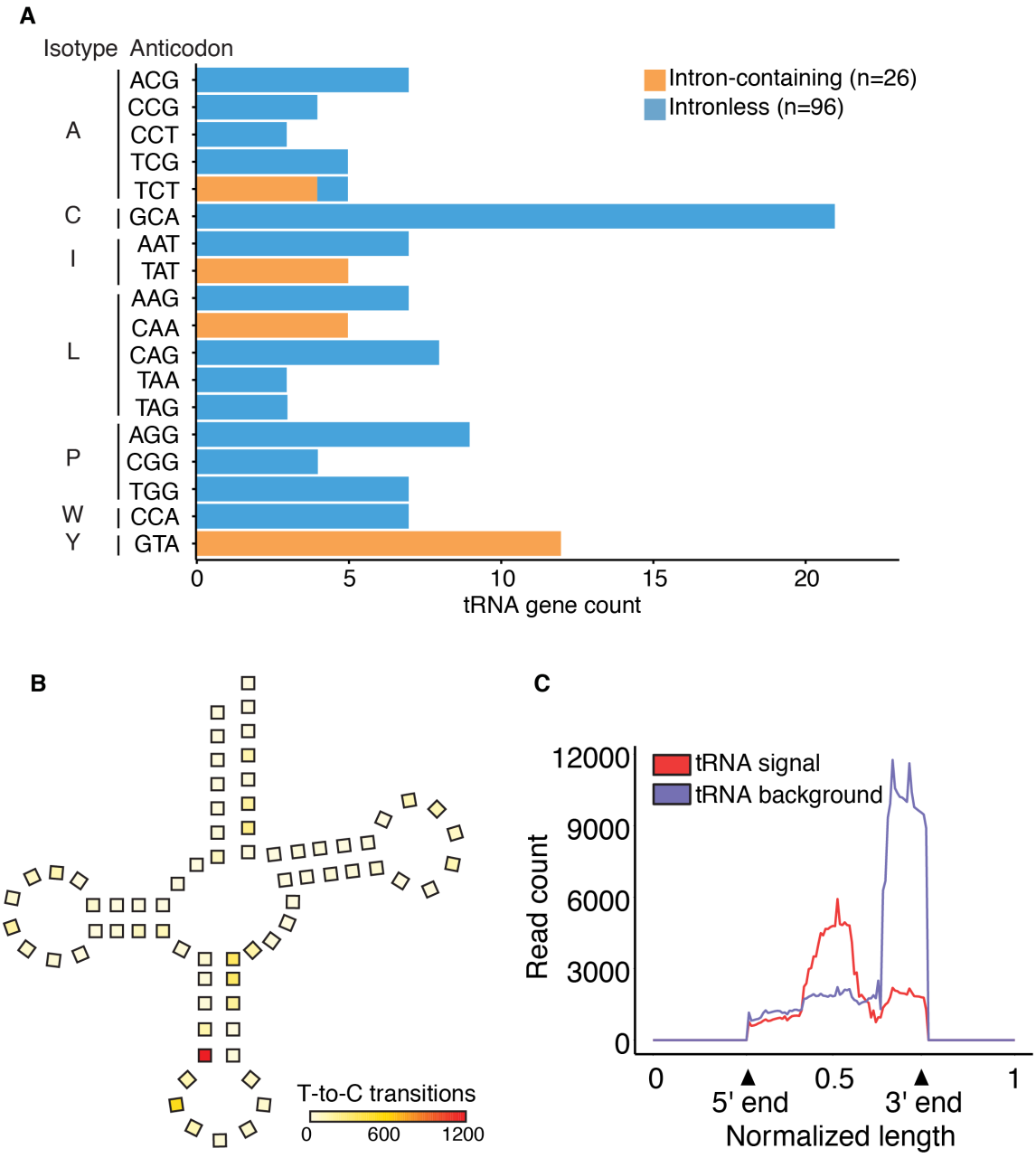

Figure 2.24: figure5. placeholder

2.15 CLP1 figures

2.15.1 Plausible pathomechanisms of CLP1 mutations

i) Reduced tRNA splicing in sensitive tissues (e.g. neurons)

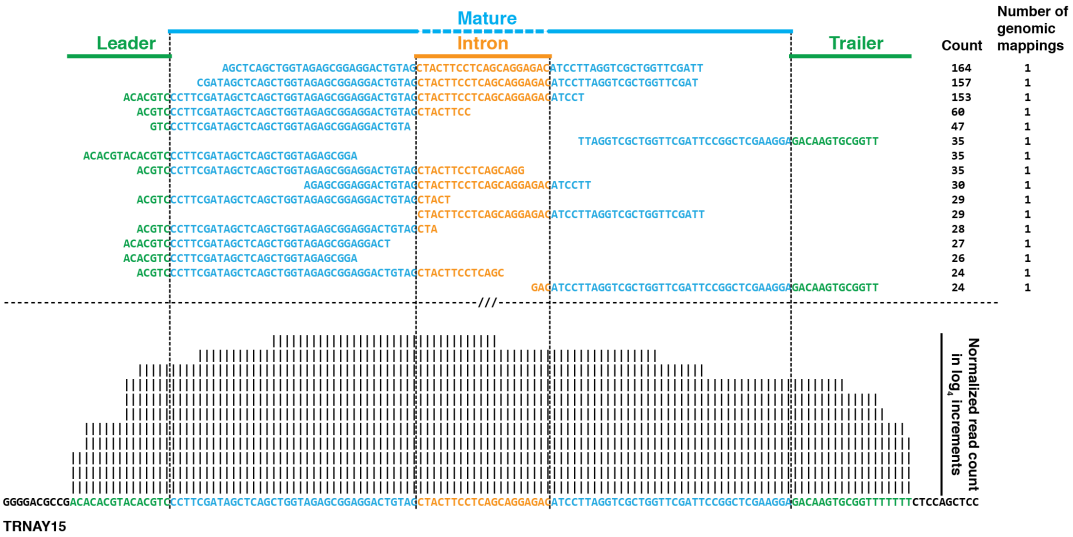

Figure 2.25: intron-containing tRNA. placeholder

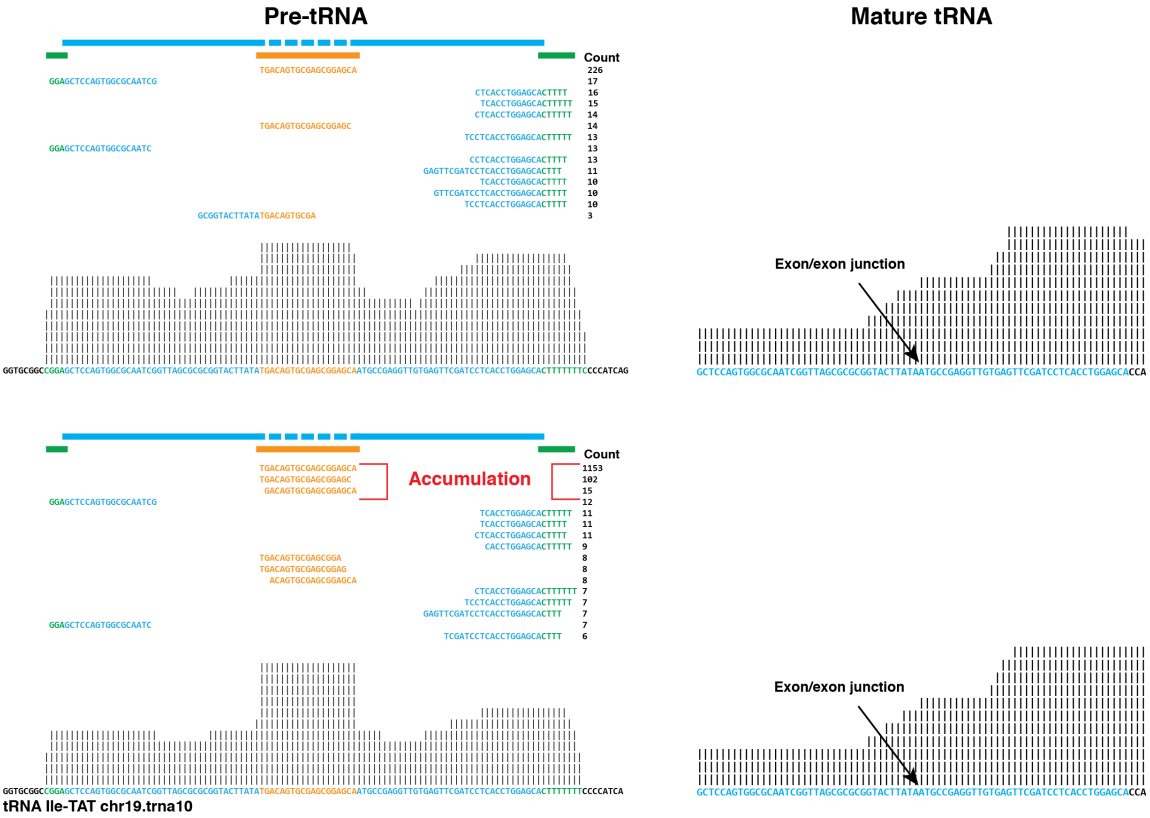

Figure 2.26: clp1 alignments. placeholder

ii) Innate immunity activation by tRNA intron accumulation

iii) tRNA-unrelated effects (e.g. CLP1 participates in mRNA polyadenylation)

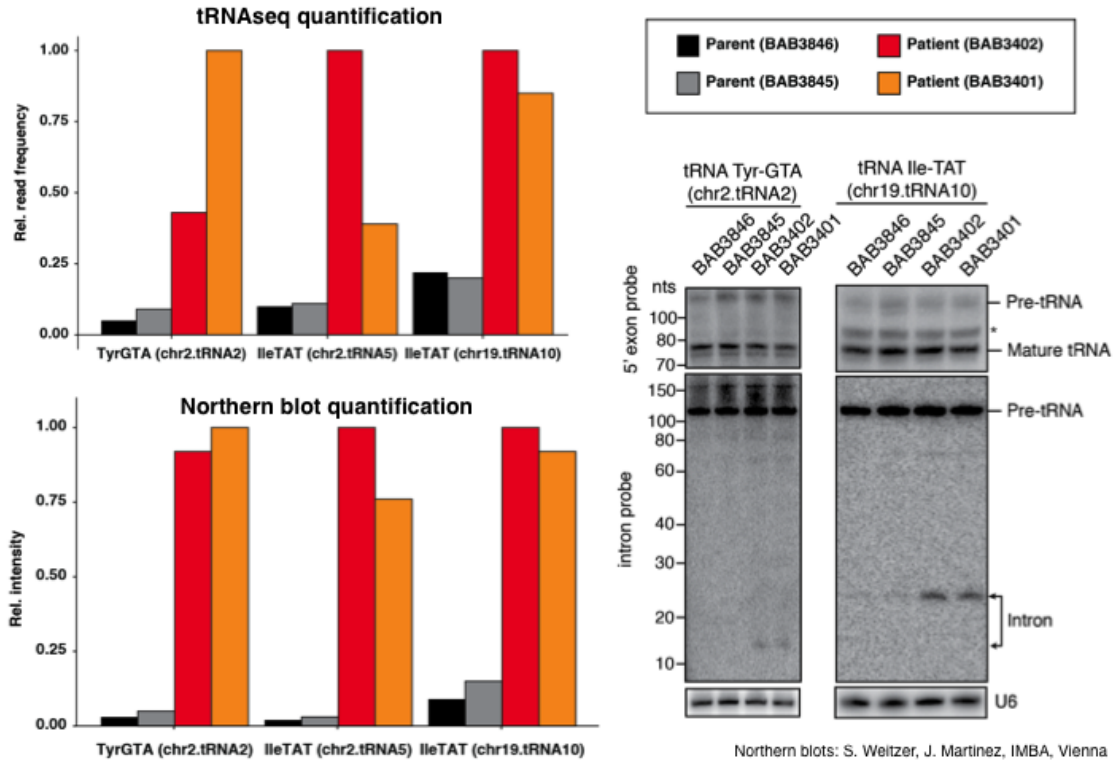

Figure 2.27: clp1 bar. placeholder

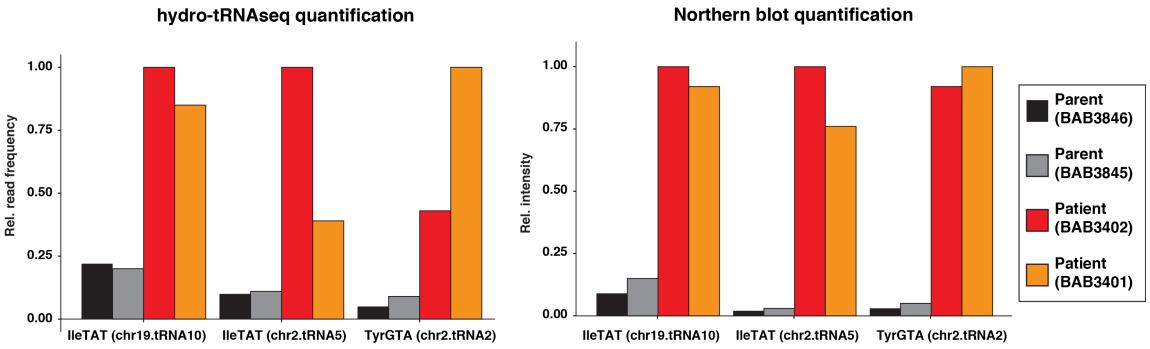

Figure 2.28: clp1 bar. placeholder

2.15.2 hydro-tRNAseq on CLP1

tRNA intron accumulation in patient with CLP1 recessive mutation compared to heterozygous parent

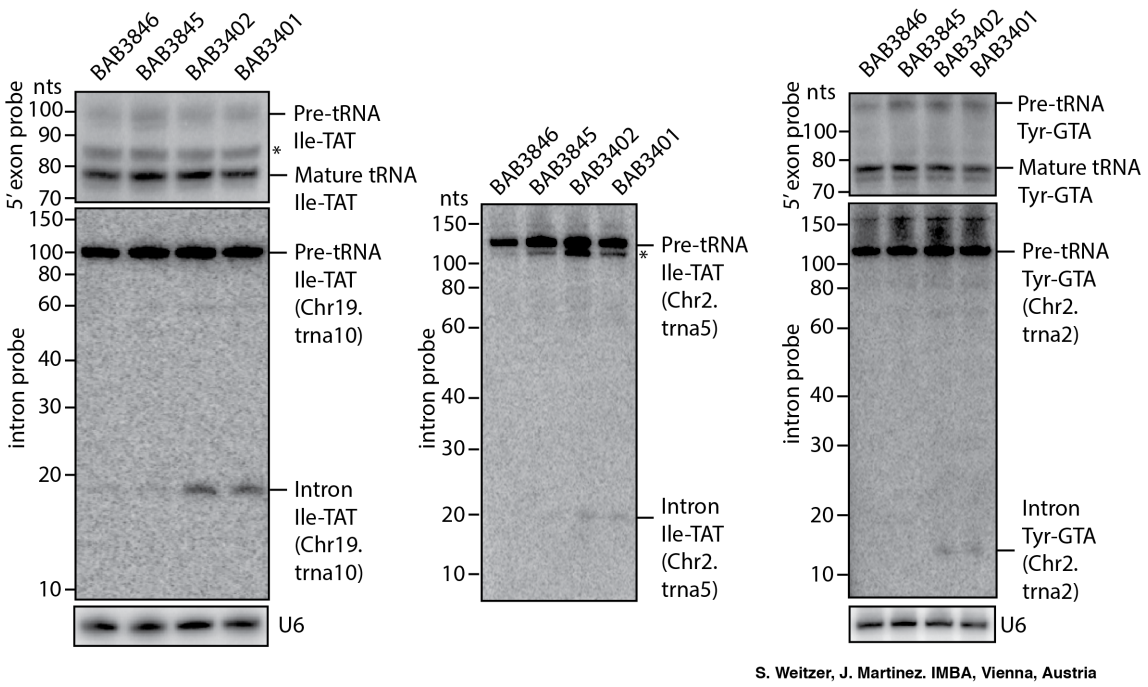

Figure 2.29: NB. placeholder

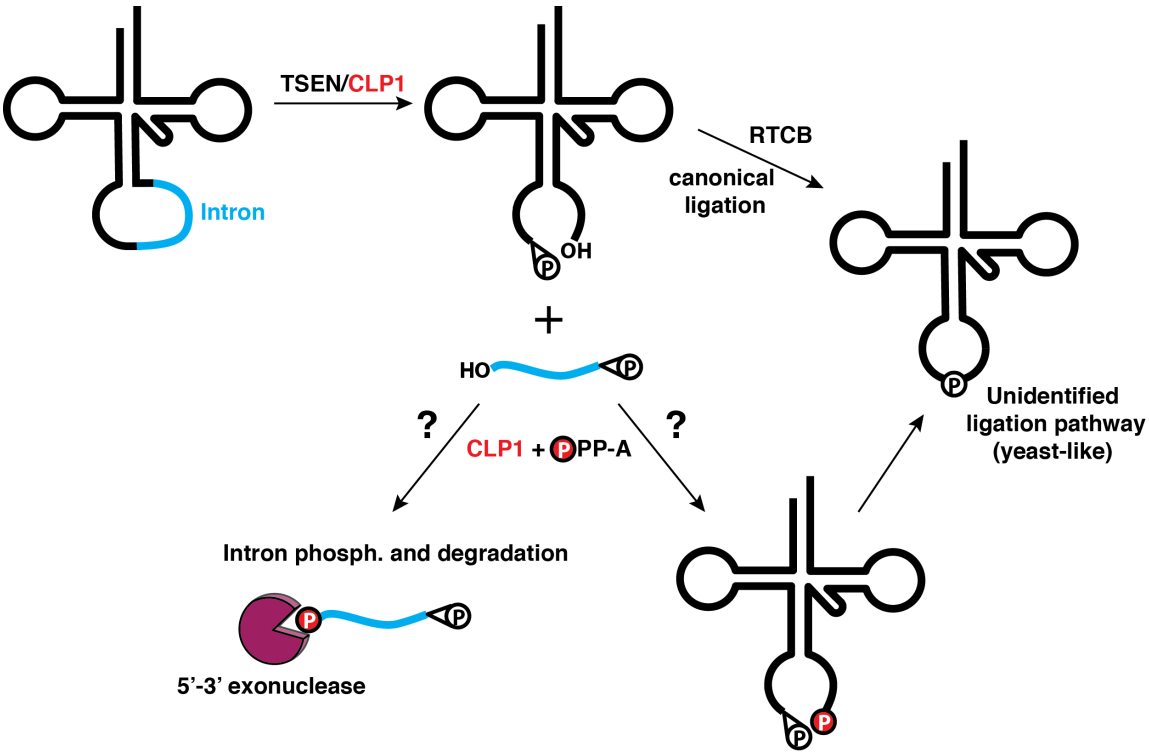

Figure 2.30: splicing. placeholder

2.16 tRNA enzyme screen

With respect to more functional aspects of tRNA biology, I focused on the results from knockdowns from a series of tRNA modifying and processing enzymes. Surprisingly, even though I targeted key components of the tRNA biogenesis and processing pathway, I could not observe major defects in tRNA processing or abundance patterns with the exception of the tRNA ligase (RTCB) that led to a very modest accumulation of 5' end tRNA halves. This suggested either that the residual enzymatic activity after knockdown is sufficient for proper tRNA processing or that there is unknown redundancy in several steps of tRNA biogenesis.

This publication steered interest in the connection of tRNA and disease, a link that was previously not well examined. To further showcase the utility and usefulness of my method in this field I decided to use it in a proof of principle example. I have carried out an siRNA-mediated gene silencing screen targeting a series of tRNA processing and modifying enzymes (i.e. the CCA terminal transferase, the tRNA methyltransferase TRMT10A, the tRNA splicing ligase HSPC117, a member of the RNase P complex, and also TRANSLIN, a putatively novel tRNA processing enzyme – see aim 2). In the cases where a reliable antibody was available, the knockdowns were validated by western blot analysis. In all other cases, knockdowns are currently being validated by mRNAseq. RNA from the validated knockdowns will be subjected to tRNAseq, with the expectation to identify tRNA processing defects that prior to our method would have been overlooked. In addition, I have performed tRNAseq from spinal cord-derived fibroblasts of knockout mice for the tRNA ligase. There I was already able to show tRNA processing is affected characterized by an accumulation of unspliced tRNA halves.

2.17 Discussion

We 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, we 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. 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, we took advantage of the pre-tRNA binding properties of SSB protein, which coordinates posttranscriptional processing and maturation of tRNAs [45], 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 U-rich 3' ends in all organisms examined so far [50, 55], our approach can be adapted towards tRNA annotation in other species.

Our data suggest that, at least in our 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 pseu-

dogenes. 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]. Our analysis shows that selenocysteine seems to be the only monogenic tRNA species, suggesting an increased sensitivity to mutations due to the lack of functional redundancy.

Our 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 [56]. We also 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 [57, 58] as well as structural data documenting the capacity of SSB to accommodate 4 Us in its binding site [46, 50]. 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 pol III *in vitro* [56].

We 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 [46, 47]. 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.

Our 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, we were able to carry out a careful overview and tRNA modifications that result in characteristic mismatch signatures. We 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 we 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, we were 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 we used PAR-CLIP of SSB for the annotation of pre-tRNAs and tRNA genes, we 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. Collectively, our results give a census of the tRNA transcriptome in human cells. We provide a method and computational flowchart for the analysis of tRNAs. We expect that these results will inform studies of tRNA-related human disease.

2.18 Summary

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, they can be used to probe long-standing questions of tRNA biogenesis, processing and function

iv) Hydro-tRNAseq can be applied for studies of human genetic diseases

Chapter 3

Materials and methods

3.1 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 MgCl2, 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 [44], and sequenced on an Illumina HiSeq 2500 instrument. Adapters were trimmed using 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.

3.2 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 [59]. Expression of FH-SSB was induced by addition of 1 μg /mL doxycycline for 24 h. 4SU PAR-CLIP was performed as described previously, using either RNase T1 or RNase A [60]. 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.

3.3 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, all available upon request. Graphs were created in R and Prism (Graphpad).

3.4 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.

Chapter 4

C3PO

4.1 Introduction

Finally, recent studies provide growing evidence for the participation of a novel endoribonuclease complex in the lifecycle of tRNAs. C3PO (component 3 promoter of RISC) is a multimeric complex of the RNA-binding protein TRANSLIN (TSN) and the nuclease TRANSLIN-ASSOCIATED PROTEIN X (TSNAX), originally shown to localize to DNA break points36,37. Recently, however, it was shown to promote the activity of the RNA-induced silencing complex in drosophila. Simultaneously, it was reported that TSNAX possesses RNA endonucleolytic activity both in vitro and in vivo, and the crystal structure of the C3PO apo-complex from various organisms was reported38-40. Its biological importance is underscored by the severe neurodegenerative phenotypes observed in mice lacking components of the complex41. Interestingly, it has also been observed that C3PO mutations in the fungus N.crassa and in mouse fibroblasts result in accumulation of tRNA fragments (tRFs), elevated levels of mature tRNAs and protein translation, and increased resistance to cell death-inducing agents10. This unexpected finding

suggests that C3PO might be a novel tRNA processing enzyme. Despite these studies, though, the biological targets and the details of the C3PO's biochemical activity 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.

Recent work from our group and others suggests that C3PO, a multimeric complex of TRANSLIN (TSN) and the nuclease TRANSLIN-ASSOCIATED PROTEIN X (TSNAX), apart from its previously known roles in DNA-damage response and enhancement of RISC activity, possesses also a tRNA processing activity. In agreement with these observations, mutations in the components of C3PO phenocopy mutations in other tRNA processing enzymes by resulting in neurological and behavioral phenotypes54. Also, lack of C3PO activity is associated with accumulation of tRFs10. Thus, it is possible that C3PO might be a regulator of gene expression that could bridge tRNA processing with PTGR. In this aim, I plan to apply the annotated list of tRNAs from aim 1 and the experimental framework for the study of tRNA-tRBP interactions from aim 2 to the study of C3PO, a regulator of gene expression that could possibly provide a new bridge between tRNA processing and PTGR.

C3PO notes. TRANSLIN (TSN) was initially described as a protein that would bind DNA breakpoint junctions at conserved sequences. Initial studies identified that TSN was specifically localized in the nucleus of lymphoid cell lines, suggesting that it might play a role in DNA repair, replication or recombination (Aoki et al. 1994, 1995).

TSN contains five potential protein kinase C phosphorylation sites, three tyrosine kinase phosphorylation sites (Aoki et al. 1995). TSN was detected in the cytosolic fraction of nonlymphoid and lymphoid cells, and in the nuclear ex-

tract of lymphoid cells. It was also detected with a slightly faster mobility in non-hematopoetic cells (e.g. HeLa). In conclusion, the nuclear relocalization of TSN is “selectively controlledin lymphoid lineage cells”.

In this study, TSN was identified as a ssDNA binding protein that binds conserved sequences at the breakpoint junctions in lymphoid cells that have undergone DNA translocations. A proposed role for TSN binding of these DNA loci was “DNA-unwinding, which makes these regions more susceptible to nuclease cleavage.”

“Since DNase I or S1 nuclease hypersensitive sites were observed near these breakpoints, and potential Z-DNA elements have been shown to be important in homologous recombination, it is conceivable that the above DNA structures influence chromatin structure and render a localized region of DNA more accessible to recombinase.”

“These results raise the intriguing possibility that a mechanism of active nuclear transport for TSN exists and that TSN might be associated with lymphoid specific processes, such as Ig/TCR rearrangement.”

Regarding the cytosolic function of TSN in non-lymphoid cells “one possibility is that, analogous to RecA protein, TSN may be required for repair of DNA damage caused by radiation or chemicals.”

The high expression pattern of TSN in testes, a tissue that is know to undergo extensive DNA rearrangements provides further evidence that the protein may be involved in DNA damage control.

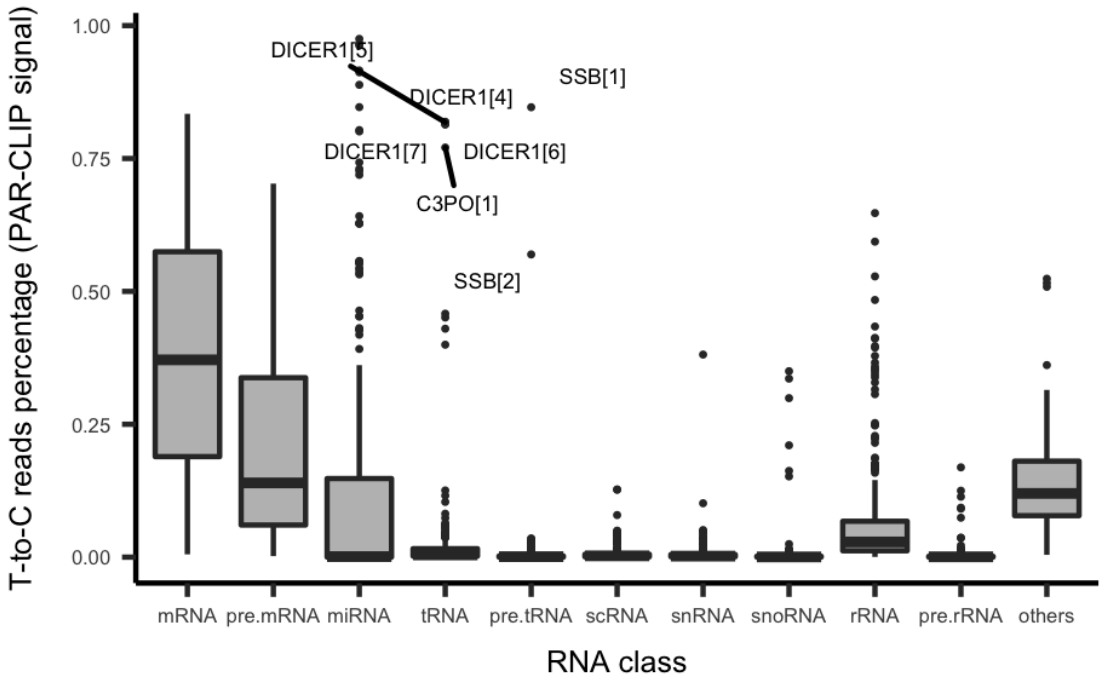

Figure 4.1: parclips. placeholder

4.2 From manny

C3PO is an RNA/DNA-binding protein complex of TRANSLIN (TSN) and TRANSLIN-ASSOCIATED FACTOR X (TSNAX), which was originally reported to co-localize with breakpoint junctions of chromosomal translocations. Tsn-/- mice were created with the hypothesis that its absence would affect genome stability. Interestingly, mice lacking Tsn exhibited learning and memory defects indicating that it is necessary for brain development or neurological maintenance. C3PO was recently reported to enhance RNAi activity in Drosophila and humans, wherein TSNAX possessed a novel and highly conserved ribonucleolytic activity. In collaboration with Dinshaw Patel's group (MSKCC), I biochemically characterized its nuclease activity and we solved the structure of Drosophila C3PO. Concurrently, the Liu laboratory (UTSW) crystallized human C3PO. Despite these recent efforts,

neither the endogenous targets nor the mechanism of C3PO assembly and nuclease activity are well understood. Given the positions of the nuclease catalytic sites, it is not clear if C3PO dynamically assembles onto its substrates, or if RNA molecules are fed inside the structure.

I have biochemical data showing that human and Drosophila C3PO possess a length-dependent ribonuclease activity (Fig. 4). C3PO nuclease activity is much reduced for RNA substrates >80 nt. The physiological importance of the length-dependent activity is not known, though one clue may come from a recent publication on a newly reported function of C3PO regarding maturation of tRNAs, which are typically between 70 and 90 nt in length. Independently, Andrew Fire and colleagues recently described small RNAs derived from tRNAs as a potential class of non-coding RNA regulators of RNA-induced silencing complex (RISC) activity. Given its purported roles in RNAi and tRNA maturation, it is tempting to hypothesize that C3PO fulfills aspects of both independently described mechanisms. Consistent with this working hypothesis, I find high crosslink evidence within the D-arm of tRNAs from a TSN PAR-CLIP (Fig. 5). These results indicate that TSN associates with specific tRNA hairpin loops containing an invariant sequence, a sequence I also find in the mRNAs that TSN binds. I plan to determine and characterize the endogenous RNA substrates of TSN and C3PO. Since TSN can form multimeric complexes without TSNAX it is plausible that specific RNAs are binding targets of only TSN, as opposed to the C3PO complex. As the relationship between TSN-only vs. C3PO complexes is not well characterized, I have created stable cell lines that express TSN, TSNAX, or the entire C3PO complex in an effort to determine whether there are target overlaps between the two types of TSN complexes. I will biochemically validate the RNA-protein interactions identified, prioritize these targets by determining TSN or C3PO enrichment with

RNAs and develop assays that, depending on the RNA category, functionally define the role of TSN and/or TSNAX. I will initially focus on assays that investigate RNA stability, RISC-dependence, and tRNA involvement, since these areas reflect the current understanding in the field. In future studies, I plan to characterize C3PO-containing mRNP and tRNP complexes in order to better understand its role among other PTGR components and processes.

4.3 TSN binds tRNAs

Our laboratory has identified a limited set of the RNA targets of human C3PO by performing PAR-CLIP analysis of its components. Expression of TSNAX alone did not yield any crosslinked RNAs – presumably due to its nuclease activity. However, PAR-CLIP with TSN did yield crosslinked RNAs. Preliminary analysis of this dataset shows that TSN exhibits high crosslinking efficiency to tRNAs, as well as mRNAs. In particular its consensus targets include the conserved UGGU motif of the dihydrouracil (DHU) stem-loop of tRNAs. A similar motif is also observed in the most prominent mRNA targets of TSN (Figure 5).

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

4.5 Biochemical characterization of C3PO's tRNA processing activity

I plan to repeat the PAR-CLIP analysis of TSN, using the tools and expertise obtained from aims 1 and 2 in order to validate its interaction with tRNAs. At the same time, in order to gain insight into the substrates of the full complex, I plan to perform PAR-CLIP studies after co-expression of TSN with a catalytically inactive form of TSNAX, due to a single amino acid mutation at its active site40. Based on structural studies, a catalytically inactive TSNAX is expected to form a similar stoichiometric complex with TSN as its wild-type form and bind its natural RNA targets, without being able to cleave them, thereby allowing their identification by PAR-CLIP. The targets identified by PAR-CLIP will be validated, in part, by in vitro nuclease assays so as to determine the details of the catalytic mechanism of C3PO on its substrates.

4.6 Functional validation of C3PO's targets

The impact of C3PO's activity on its targets will be examined in over-expression and knockdown experiments. I plan to perform HydroRNAseq and RIP-Seq studies to investigate the impact of C3PO expression levels on the stability of its RNA targets, and to further dissect its biological function. Lack of C3PO activity has been linked with elevated levels of mature tRNAs and pre-tRNA fragments10. At

the same time, C3PO activity is suggested to promote activity of RISC leading to enhanced RNA silencing38. If the former in the true, I expect to observe an accumulation of pre-tRFs and mature tRNAs in knockdown versus control or over-expression. If the latter holds, then I expect a change in the stability of the mRNA targets of C3PO. Finally, if time permits, I intend to study the effect of C3PO's activity on the protein level, by performing western blot and mass-spectrometry studies of the proteins whose mRNAs will be validated targets.

4.7 C3PO sumary report 2014

Performing PAR-CLIP on C3PO confirmed its role as a bona fide tRNA binding protein. The evidence for this is twofold: a) tRNAs collect the largest number of sequenced reads, 31%, with the 2nd most represented class, mRNAs, collecting only 9% (table 1), b) tRNAs represent the largest percentage of PAR-CLIP clusters both within the total number of clusters, as well as within the top 100 clusters (ranked by read abundance) (figures 1, 2, table 2). Of note, my PAR-CLIP results seem to bring into question previous reports that implicated C3PO in enhancing RISC activity and promoting miRNA-mediated gene silencing, since only 4% of the total clusters map to miRNA and 3' UTRs, respectively (table 2). However, 5' UTRs represent 22% of all clusters, a result that comes as a surprise, as C3PO has not been reported to interact with 5' UTRs before. I have also determined the tRNA binding motif of C3PO (figure 3), using motif prediction algorithms (GIMSAN and MEME). Currently I am carrying out RIP-seq experiments in order to validate and rank the targets of C3PO.

In order to investigate the biochemical function of C3PO, I am trying to identify other interacting proteins. A preliminary co-immunoprecipitation analysis yielded

no clear candidates, which suggests that either C3PO functions in no close association with any other protein or that the conditions of immunoprecipitation were not appropriate. At the moment, I am repeating the experiment at different salt concentrations, as well as in the presence of a cell-permeable crosslinking agent (DSP). I hope that in this way I will be able to stabilize the possible interactions of C3PO with other proteins, which I will then identify by using the mass-spectrometry core facility. Moreover, I have already carried out in vitro endonuclease cleavage assays, using recombinant C3PO. These experiments confirmed that C3PO cleaves in vitro transcribed tRNAs in a length- and structure-dependent manner. I will carry out similar studies using tRNA targets predicted from PAR-CLIP and confirmed by RIP-seq.

Since C3PO has been previously implicated in a plethora of biological processes, I am interested in elucidating the processes in which it partakes. For this purpose, I am carrying out a gene ontology analysis of the PAR-CLIP targets. Also, I have performed siRNA knockdown experiments against the two components of the C3PO complex, TSN and TSNAX. I have validated commercially available antibodies, designed siRNAs, and have successfully knocked down C3PO in HEK293 cells by 3- to 5-fold. As the next step, I will perform two series of RNA-sequencing experiments in the context of C3PO knockdown: mRNA-seq (poly A selection) and HydroRNAseq, to determine the effect of C3PO expression on the mRNA and tRNA population, respectively.

Finally, a series of reports implicate tRNA metabolism in stress responses. Specifically, it has been suggested that upon cellular damage, tRNA endonucleolytic cleavage leads to translational arrest via the accumulation of tRNA halves. I subjected HEK293 cells to treatment with sodium arsenite (an inducer of oxidative stress), isolated RNA, and then performed small RNA sequencing that confirmed

the accumulation of stable 5' tRNA fragments. As a control, the miRNA or snRNA population remained largely unchanged. I confirmed these results by northern blots. Stress-induced endonucleolytic cleavage of tRNAs has been reported to be a function of the nuclease angiogenin. Nevertheless, angiogenin is not appreciably expressed in our cell culture system, neither in normal growth conditions nor upon oxidative stress. Therefore, it is intriguing to assume that C3PO might play a previously uncharacterized role in the metabolism of tRNAs during stress. For this purpose, I have carried out a PAR-CLIP experiment under conditions of oxidative stress and I am currently awaiting the sequencing results.

4.8 C3PO summary from annual report 2015

C3PO is a multimeric complex of the RNA binding protein TRANSLIN and the RNA endonuclease TRAX. 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 had previously identified C3PO as a tRNA binding protein, and showed that it does 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 effect of C3PO in mRNA and tRNA stability, I performed mRNA and tRNAseq upon knockdown as well as induction of C3PO. To my surprise no significant effect was observed in either class of RNAs. There results can be interpreted in two ways. First, 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. Sec-

ond, C3PO has an effect on translation rather than on mRNA or tRNA stability. I am testing the former hypothesis, by performing sequencing on RNA isolated from Translin and Trax knockout flies, and the latter by performing polysome analysis and mRNA reporter translation assays. Finally, in order to elucidate the molecular mechanism of action by C3PO, I have engaged in a collaboration with the lab of Dinshaw Patel at MSK, trying to obtain a crystal structure of C3PO with a minimal or full-length RNA target identified in my PAR-CLIP analysis.

4.9 C3PO summary from annual report 2016

Despite extensive efforts in collaboration with the Patel lab (MSK), we have not been able to obtain a crystal structure of C3PO (complex of TRANSLIN and TRAX) bound to its RNA targets, which I have previously identified by PAR-CLIP. I have performed extensive electrophoretic mobility shift assays that have specified optimal substrates (e.g. full length tRNA, single stranded GGU repeats of various lengths, short RNA stemloops containing C3PO's putative binding motif TGGW - W= A or T). Using these substrates, the Patel group obtained well-diffracted crystals of TRANSLIN crystallized in presence of ether single stranded RNA sequence 5'-(UG)3U(UG) or the tRNA dihydrouridine stem loop sequence (5'-GUAUAGUGGUUAGUAC) that belong to two different space groups. The structures were solved at 2.2 Å and 2.74 Å resolutions, respectively, and revealed minor differences in the arrangement of octameric TRANSLIN, but no bound RNA substrate in the expanded hollow interior of the closed-barrel structures. They have also obtained small crystals of the truncated wild-type C3PO in the RNA-free state, which are currently being optimized to attain diffraction quality.

In parallel, I have been characterizing the in vivo effects of C3PO, by loss-of-

function studies. Unexpectedly, I observed that siRNA knockdowns of TRANSLIN led to increased translation of mRNA targets identified by PAR-CLIP. At the same time, more than 30 ribosomal proteins were upregulated upon TRANSLIN knockdown. Therefore, there is evidence that C3PO has a role in regulating translational efficiency. I have confirmed by RNAseq that C3PO is ubiquitously expressed at moderate to high levels. Therefore, there are reasons to think that C3PO might be involved in fundamental and global regulation of translation, either by acting directly at the translational process or indirectly, by modulating tRNA levels. In fact, the binding properties of C3PO towards specific 5' UTRs of some mRNAs reflect those of known translational factors, inasmuch as they are usually single, short binding sites with GC-content significantly higher than randomly sampled, size-matched sequences from the 5' UTR background context19.

Since TRANSLIN knockdowns did not have a global pronounced effect on tRNA levels, I reasoned that the depletion achieved by the siRNA methodology (70% reduction in protein level) was not impactful enough to result in an observable effect on global tRNA levels. Therefore, I am obtaining CRIPSR knockouts20 of TRANSLIN and TRAX, which I will start characterizing after the submission of my tRNA manuscript.

Although previous attempts to co-immunoprecipitate (co-IP) interactors of C3PO had been fruitless, I have optimized the IP conditions, managing to observe stoichiometric interactors. By western blot analysis of the interactors I identified a component of the RNase P complex, which renders further validation in the involvement of C3PO in tRNA processing. Having established appropriate IP conditions, I am scaling up and preparing for proteomic analysis of the C3PO interactors. Finally, I am setting up stable isotope labeling by amino acids in cell culture (SILAC21) experiments for C3PO wild-type versus C3PO knockout cell lines in

order to confirm C3PO's involvement in the regulation of translation.

Finally, recent publications have provided evidence for abundant stable tRNA fragments with various functions, but to date no factor has been identified as responsible for the biogenesis of these fragments22. To examine whether C3PO is involved I have prepared small RNA sequencing datasets upon gain- and loss-function of C3PO, by enriching for 5'-end phosphate containing small RNAs in the appropriate size range of 19-35 nts, such as tRNA fragments and miRNAs.

Chapter 5

Things that didn't work

As I had proposed last year, I am ultimately interested in applying the expertise and data obtained form tRNA sequencing towards understanding the interdependence (if any) between tRNA and mRNA abundance on a codon-by-codon level, as well as determine whether the balance between the two affects translational rates and mRNA stability. First, I retrieved all annotated isoforms of human mRNA transcripts, and obtained computationally an expression-weighted frequency table for all codons, by multiplying the incidence of every codon per every transcript with the abundance of each transcript. After accounting for wobble effects, I examined the relationship between codon and respective anticodon abundance. Interestingly, even though for the majority of codons, an increase in their abundance is mirrored by an increase in their cognate anticodon, there are outlier codons for which there is a discrepancy, with the codon being unusually more or less abundant than its anticodon. This led to the idea that the codon-anticodon abundance imbalance might have an observable effect on the translational efficiency of mRNA harboring such codon-anticodon pairs. To investigate this possibility I focused on ribosome profiling, a technique that determines the translational efficiency of mRNAs into

proteins by quantifying the number of ribosome-protected mRNA fragments by RNAseq following isolation of actively translating ribosomes12. Recently, a very deep ribosome profiling dataset in HEK293 was published13. I started performing analysis using this dataset. At a first step, I created scripts that allow for general quality control of the data. With these I confirmed the read length distribution as well as the characteristic phasing of the 5' ends of ribosome profiling reads with respect to the first translation frame12,14. At the same time, I was able to quantify ribosome protected fragments per every single codon. This now allows me to correlate translational efficiency with the relationship between codon and anticodon abundance. Finally, I am investigating whether there is a connection between the translational efficiency of an mRNA and its stability, since it has been previously postulated that stalled ribosomes on inefficiently translated messages can trigger mRNA degradation resulting in decreased mRNA abundance15-18.

Finally, recognizing that I can obtain tRNA expression and abundance profiles in a more reliable manner than previously published ones, such as tRNA arrays, I decided to approach long-standing fundamental questions about the nature of the genetic code. Particularly, I was interested in combining mRNA abundance data, and translational efficiency data obtained by ribosome profiling, with tRNA abundance data in order to understand translational dynamics better. The hypothesis behind this work was that codon biases observed in specific contexts could be directed by the availability of tRNAs.

For this reason, I turned my attention towards ribosome profiling. Since our lab is still in the process of establishing this methodology, I focused on already published data about our cell culture system. I developed a series of scripts that allow me to perform tailored analysis of the data. Although this analysis is far from complete, there have been some important observations already. First, as

previously shown by other groups, I showed (using our own RNAseq data) that although highly expressed mRNAs are also highly occupied by ribosomes, there are indeed cases where transcripts are efficienytly transcribed but inefficiently translated and vice versa. This led to the hypothesis that specific tRNA availability may be the bottleneck in the translation of some transcripts. Using my tRNA abundance data, and after accounting for wobble effects, I computed an anticodon scoring index, which is essentially the weighted sum of the relative frequency of all anticodons required to decode a given open reading frame. Traditionally, it has been postulated that efficiently translated messages harbor common codons, reflecting codon optimization. Surprisingly, my preliminary data indicate that the most highly represented genes in ribosome profiling (top 10%) rank in the bottom 10% in anticodon score index. At the same time, these genes are highly transcribed, as evidenced by mRNAseq data. If these results are true (and not an experimental artifact, which I will validate by performing further replicates), then this would argue towards a lack of selective pressure for optimized codons in highly translated genes. One could then argue that increasing transcription rates for these genes leads to high protein expression levels, obviating the need for codon optimization at the DNA level.

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. 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).

23. 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).

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

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

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

46. 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).

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

48. 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).

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

50. 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).

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

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

53. 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).

54. 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).

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

56. 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).

57. 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).

58. 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).

59. 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).

60. 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).