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

A Thesis Presented to the Faculty of

The Rockefeller Uniersity

in Partial Fullfillment of the Requirements for

the degree of Doctor of Philosophy

by

Tasos Gogakos

June 2017

Abstract

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

Tasos Gogakos, Ph.D.

The Rockefeller University 2017

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

Στους γονείς και τον αδερφό μου

Acknowledgments

First, I would like to thank my

contents

Table of Contents

List of Figures

List of Tables

Glossary

List of Abbreviations

ncRNA noncoding RNA.

tEsT test2.

tRNA transfer RNA.

Chapter 1

Introduction

1.1 Overview

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

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

other classes of non-coding RNAs (ncRNAs) [9–11]. Nevertheless, the lack of reliable methods for tRNA quantification has hampered such analyses, and necessitated the use of predicted tRNA gene copy number as a surrogate index of expression [7, 12, 13]. This hinged on the assumption that predicted tRNA gene loci are all expressed constitutively and equally, even though there has been experimental evidence against it [**Gingold:2014iz**]. Similarly, experimental tRNA gene annotation in the past had to focus on RNA polymerase III (POLR3) ChIP-seq [**Kutter:2011ff**] [**Moqtaderi:2010hc**] [**Oler:2010fb**] or hybridization-based approaches [**Dittmar:2004fb**] [**Goodarzi:2016gd**]. 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 [**Cozen:2015ds**] [**Zheng:2015dw**]. However, an extensive annotation of human genes and transcripts was foregone because the focus was either on mature tRNAs only [**Zheng:2015dw**] or on tRNA fragments not inclusive of full-length precursor tRNA (pre-tRNA) transcripts [**Cozen:2015ds**]. Thus, to-date an experimentally validated list of curated mature and pre-tRNA sequences and annotating tRNA genes in human is still missing.

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

tify and quantify tRNAs, and provided evidence for the validity of this approach when determining the accumulation of disease-associated tRNA intron fragments caused by mutations in the tRNA splicing machinery [**Karaca:2014em**]. Here we extend this approach by applying it to tRNA-enriched size fragments with the aim to annotate and curate all tRNAs. Since tRNA processing, such as precursor trimming and intron removal, is a fast process[**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:2009cx**] [**Bayfield:2010cs**] [**Stefano:1984wp**].

1.1.1 tRNA biogenesis

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

from the nucleus to the cytoplasm where they can undergo further modifcations, are aminoacylated with their cognate amino acid by aminoacyl tRNA synthetases, and are finally presented to the ribosome by translation factors to participate in protein synthesis Fig. 1.1. cite 16,17 from TRP

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

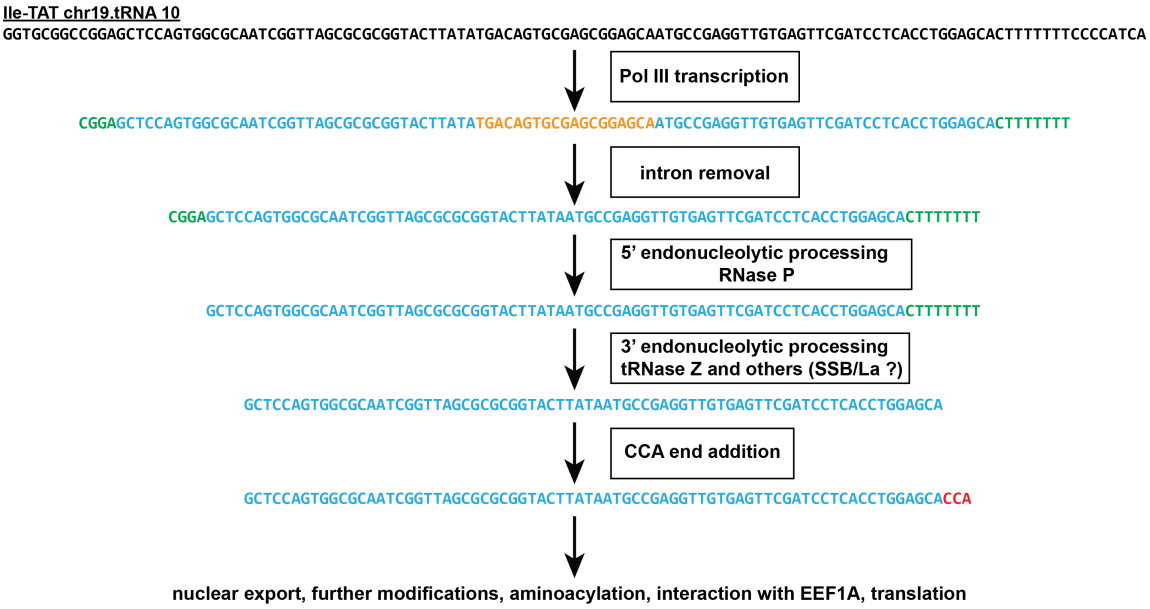

Figure 1.1: Overview of tRNA biogenesis and processing. thththt

1.1.2 tRNA sequencing

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

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

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

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

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

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

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

1.1.3 Previous efforts for genome-wide tRNA annotation

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

1.1.4 Small RNA sequencing protocol

First, I applied the protocol that the Tuschl lab had previously developed for sequencing small RNAs [**Hafner:2012eaa**] (**Fig. 1.2**). The experimental procedure resulting in small RNA cDNA library preparation begins with the ligation of bar-

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

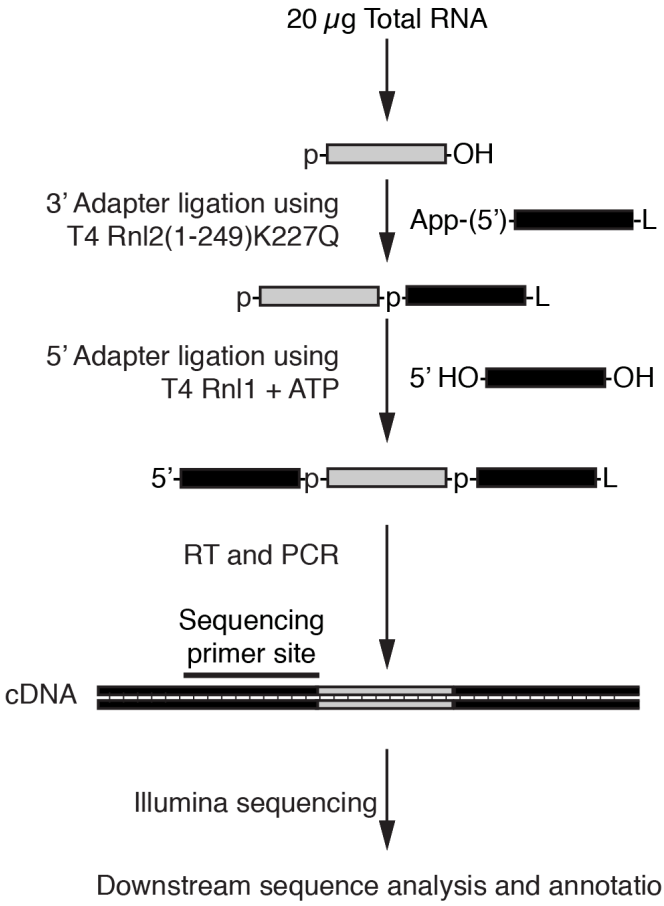

Figure 1.2: Small RNA sequencing protocol. thththtfahflahflahdflahdflahdfladfladhflasdhfladhfladhfladhfladhfaldjsfhaldskfjaldsfhalsdfhasdf

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

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

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

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

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

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

Table 1.1: RNA category from small RNA sequencing protocol

Chapter 2

Results

2.1 Hydrolysis-based tRNA sequencing (hydro-tRNAseq)

In order to overcome the problems associated with tRNA sequencing, we tried to identify the minimal number of simplest steps that could tackle the maximal number of problems. Thus, to curate

References

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

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

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

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

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

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

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

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

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

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

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

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

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

14. Arimbasseri, A. G. & Maraia, R. J. RNA Polymerase III Advances: Structural and tRNA Functional Views. *Trends in Biochemical Sciences,* 1–14 (Apr. 2016).