מכון ויצמן למדע
WEIZMANN INSTITUTE OF SCIENCE

Thesis for the degree

עבודת גמר (תזה) לתואר

Master of Science

מוסמך למדעים

Submitted to the Scientific Council of the

Weizmann Institute of Science

מוגשת למועצה המדעית של

מכון ויצמן למדע

רחובות, ישראל

By

Asaf Blumental

מאת

אסף בלומנטל

ספירת tRNA: מדידה מדויקת של רמות tRNA בעזרת RNA-Seq

tRNA counts: Accurate measurement of tRNA levels by RNA-Seq

Advisor:

Prof. Yitzhak Pilpel

Dr. Ido Amit

מנחה:

פרופ' יצחק פלפל

ד"ר עידו עמית

February 2016

אדר א' תשע"ו

Abstract

In *S.cerevisia* (yeast) there are 275 genes that code for tRNA, those genes transcribe 56 different tRNA sequences that can be grouped into 42 different families based on their anticodon. The tRNA species are changing their expression level according to various conditions or diseases. The most common way to determine tRNA levels is by using gene copy number which gives a rough estimate, but this does not accounts for differences in expression from each gene. Other, more direct method for measuring tRNA levels is tRNA microarray, which is lacking in accuracy and sensitivity, among other things, due to cross hybridization.

RNA-seq based methods for detection were not successful in counting the tRNA levels due to secondary structure and especially the modified bases that may interfere with reverse transcription and are pervasive along the tRNA. Another unknown in large numbers of tRNA species are the location and identity of the modified bases: in yeast only for 34 out of 56 tRNA sequences, the identity and location of base modifications are known.

We developed a method for quantifying tRNA levels using RNA deep sequencing called Advanced tRNA-seq. Our method uses RNA-seq in a manner that allow us to detect also tRNA fragments that were aborted due to a modified base and uses them to characterize the tRNA pool.

In order to separate the tRNA from the total RNA for the Advanced tRNA-seq library, a SPRI beads based method was developed and used.

The method was able to detect all of the tRNA species, in a simple and a fairly quick manner. The sequencing results were further analyzed as the method allows us to investigate in high-throughput the effect the modified bases have on the nucleotide composition and the likelihood of the Reverse-Transcriptase (RT) to fall-off. The analysis of the modified bases was used to try and predict the location and identity of a few modifications in the unknown tRNAs.

Further improvements for this method should be made as it is not accurate enough, due to a bias toward several tRNA species.

Contents

1. Introduction

tRNA molecules are responsible for translating the genetic code into amino acids. For each amino acid there is usually more than one tRNA specie, those are dubbed tRNA isoacceptors, even for the same anti-codons there can be different tRNA species that have different sequence, those are dubbed tRNA isodecoders (1). The expression level of the different tRNA species can vary by organism, cell type and perturbation (1, 2).

The relative proportion of isoaccepting tRNA can influence choice for codons which are synonymous. This will determine the translation efficiency of transcripts that have different ratios of those synonymous codons, a phenomenon known as codon bias (3). Furthermore, a change in the level of certain tRNA species can have great physiological impact during development, and can even lead to cancerous transformation (4).

The most common way to estimate a tRNA level is by counting the gene copy number of varies tRNA species. The Pilpel lab, among others have shown a correlation between the gene copy number and codon usage, which seems to indicate tRNA level (5, 6). Accordingly, *B.subtilis* $tRNA^{Arg}$(ACG), that has 4 copies of the gene, will be four times more abundant than the one copy of $tRNA^{Arg}$(CCG) (3). This method assumes a static tRNA levels and a similar expression from each tRNA gene, yet the situation is actually known to be more complex both in bacteria and especially in eukaryotes (2, 7).

Another method, which increased in popularity in recent years, is the use of microarrays, a direct measurement method, unlike gene copy number, which is indirect. In a version of this method, which is specifically dedicated to tRNAs (2), one, first attaches a fluorescently labeled probe to the conserved 3' CCA sequence of the tRNA molecules, and then hybridize them to a 70-80bp DNA probes that are homologues to the tRNA sequences. In the end one can quantify the relative number of each tRNA specie by the fluorescence intensity (2). This method requires constructing or adapting a microarray to each organism, it cannot differentiate between all the tRNA species as it needs more than 10 different residues to differentiate between two tRNA species (2) and there is a very defined dynamic range of possible levels (about two orders of magnitude) (2).

RNA-Seq is an unbiased counting approach which uses the power of next generation sequencing to sequence and count millions of molecules in a few hours. RNA-Seq uses reverse transcriptase (RT) to turn the cellular RNA to cDNA, which can be then amplified by PCR with specific sequencing adapters and sequenced (8). RNA-Seq offers an unbiased approach to count accurately the global levels of cellular RNA in a high throughput manner (8). RNA-Seq is a method that has not been widely used yet to determined tRNA levels. This stems from the secondary structure that the tRNA adopts (which interferes with the RT advance), and most importantly from the presence of modified bases (which the RT can have problem interpreting) (see figure 1). There are more than 100 biochemically distinct modifications in on tRNA (9). For example in yeast's tRNA more than 25 different modifications are known (10), and on each tRNA there are 7 to 17 modified bases (11). Only part of the tRNA's have their modified bases location and identity, fully mapped. In yeast just 34 out of 275 tRNA genes and 56 unique sequences are mapped (12, 13). When the RT reaches a modification it can continue to elongate using a standard base (which can be the correct one or another one entirely) or it can stop and fall (14).

Recently, there have been several attempts to overcome those problems and utilize deep-sequencing in order to quantify tRNA levels.
One such attempt was the removal of several of the more common modification (13). By using special enzymes most of those modified bases: N1-methyladenosine (m1A), N3-methylcytosine (m3C) and N1-methylguanosine (m1G) were removed. In addition, this method used a high processivity RT in order to try and reduce the falling-off when reaching a modification (13). This method was called DM-tRNA-seq (demethylase-thermostable group II intron RT tRNA sequencing).

While this method improved the full length tRNA detected it has a few drawbacks. It fails to remove most of the different modifications and the few modifications it can remove are removed only between 70 to 80%. The correlation to gene copy number is poor. This method also needs the use of various expensive enzymes.

Another such attempt was to try and sequence tRNA even when the RT falls-off and there is only fragmented tRNA, the fragments should be sufficient to identify the tRNA

type (14). This method is using isolated tRNA to ligate adaptor to their 3' end then using the adaptor for RT, following with adding another adaptor to the 3' end of the cDNA, and finishing with PCR amplification of the tRNA(14). This method is able to detect tRNA that due to modifications failed to be fully reverse transcribed. This increases the pool of detectable tRNAs. This method is called tRNA-seq (14). This method is able to detect all tRNA types and does not require special enzymes (14).

The aim of my project was to develop method for utilizing deep-sequencing for tRNA quantification. The approach we independently developed has the same idea of the tRNA-seq method mention above (14), as it tries to determine the tRNA levels by counting even the fragmented tRNA caused by RT falling-off. The differences between our method and the published method are shown later on. The method was dubbed **Advanced tRNA-seq**.

In order to gauge the feasibility of detecting tRNA from a fragmented tRNA caused by a modification, a bioinformatics simulation was carried out. For this simulation we incorporated *S. cerevisiae* tRNA mature sequences (that contain modified bases) and *S. cerevisiae* tRNA gene sequence from the 3’ end until first modification (restrictive role, as not all modification cause RT stop) there was just 1 (0.18%) and 59 (1.24%) undistinctive pairs in mature sequences and gene sequences, respectively (see figures 2 and 3). When further investigating the tRNA gene sequences it was found that 91.3% of them share the same anti-codon (isodecoders) and 93.48% share the same amino acid (isoacceptors). This data indicate the feasibility of differentiating the *S.cerevisiae* tRNA using their 3’ end partial segment. A similar analysis was done on mouse (*M. musculus*) and human tRNA with similar results (see tables 1 and 2), suggesting a feasibility for those organisms as well.

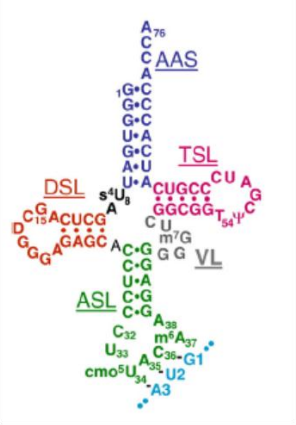

Figure 1- A carton of mature tRNA which include modified bases and secondary structure

As the simulation showed that feasibility of the idea, we independently developed a method that resembles the tRNA-seq method (14) in the general idea, dubbed Advanced tRNA-seq. However, the tRNA–seq method has a major drawbacks in that it uses HPLC in order to separate tRNA in every stage it is needed. Our protocol uses a simpler and quicker method to enrich for tRNAs which is based on magnetic beads for tRNA separation, a change which is more in tune with standard lab protocols. Other changes in this protocol allows for a quicker and simpler protocol.

As the Advanced tRNA-seq protocol detects even fragments of tRNA caused by RT falling off, possibly after reaching a modified position, it can infer the position of a modification by finding the end of the reads. We show that the likelihood of the RT to fall off at a certain position correlates to certain modifications in the known tRNAs. This makes it possible to predict several modifications in the unknown tRNAs by detecting and analyzing sharp declines in read counts over the tRNA length.

In addition, we propose, that the Advanced tRNA-seq method can be used as a tool to study the interaction between different tRNA modification and reverse transcriptase activity, be it by causing an incorporation of a different nucleotide or the falling of the RT at that position. We show this for each of the modified bases known in yeast tRNA.

2. Methods

2.1. Databases for tRNA

The tRNA genes sequences were downloaded from the Genomic tRNA Database (http://gtrnadb.ucsc.edu/) version tRNAscan-SE 2.0.for *Saccharomyces cerevisiae* S288c (yeast), *Mus musculus* (GRCm38/mm10) (mouse), Homo sapiens (GRCh37/hg19) (human) and *Escherichia coli* str. K-12 substr. MG1655. The mature tRNA with modified bases were downloaded from the Transfer RNA database (tRNAdb) from Leipzig University (http://trna.bioinf.uni-leipzig.de/) for *Saccharomyces cerevisiae* and *Mus musculus*.

2.2. Yeast strain used and RNA extraction

In all yeast experiment *S.cerevisia* BY4741 strain (MATa his3Δ1 leu2Δ0 met15Δ0 ura3Δ0) was used. Cells were grown on rich YPD media at 30°C until reaching a concentration of 10^7 cells/ml. 10ml were centrifuge and the pellet was immediately frozen in liquid nitrogen. The total RNA was extracted using MasterPure™ Yeast RNA Purification Kit (epicenter) according to the manufacturer guidelines.

For the diauxie shift experiment, overnight culture was diluted into fresh YPD media to cell concentration of $5x10^6$ cells/ml and was grown at 30°C until reaching a stationary phase. At different ti,e point (t=0, t=6h and t=9h) samples were taken and processed as described above.

2.3. Mouse cells used and RNA extraction

With the help of Yoach Rais from Yacob Hanna lab, $3.5x10^6$ mouse V6.5 ES cells were seeded on cell culture plate covered with MEF feeder cells, after two days moved to KSR/2i/LIF for 2 more days. The total RNA was extracted using NucleoSpin miRNA kit (Macherey-Nagel) according to the manufacturer guidelines.

2.4. Uncharging tRNA

To remove the amino acid from the tRNA at least 1ug of total RNA in a 24µl volume was used, added 3µl of 10mM EDTA and 3µl of Tris-HCl pH 9 and incubated for 30 min in 37°C. Then added 3µl 3M NaOAc pH 5.2. The RNA was precipitated by ethanol

precipitation: 1µl of 15 µg/µl GlycoBlue Coprecipitant (Ambion), 239µl of DNase free water and 900µl of 100% ethanol were added to the sample and incubated overnight at -80°C; next, this was spun at 13,000 RPM for 30min at 4°C, the supernatant was discarded and the RNA pellet washed with 80% ethanol, followed by similar condition of centrifugion for 1min to remove residual ethanol; the pellet was then air-dried for 10min and finally dissolved in 12µl DNase free water.

2.5. SPRI separation of tRNA from total RNA

This protocol allows the separation of medium size RNA including tRNA from total RNA. To a 20µl of DNase free water containing at least 1µg of total RNA, 36µl (1.8X) of Agencourt AMPure XP beads or Solid Phase Reversible Immobilization (SPRI) beads were added. After separating on magnet the supernatant containing the RNA without large RNA was transferred to a new tube. For mouse samples 0.8X of SPRI beads and 2.6X 100% isopropanol were added. For yeast samples 0.8X of SPRI beads and 3X 100% isopropanol were added. The beads were washed twice with 100µl of 85% ethanol air dried and eluted with 20µl of DNase free water. This process was repeated twice to better improve the separation.

2.6. Advanced tRNA-seq protocol

2.6.1. Alkaline Phosphatase

In order to improve the ligation of the RNA adaptor to the tRNA by ensuring the 3' end is dephosphorylated, FastAP (Thermo scientific, EF0654) was used. For around 20ng of the tRNA. The tube was incubated in 37°C for 12min.

2.6.2. Repair ends

In order to repair ends and add hydroxyl group to the 5' end, T4 PNK (NEB, M0201S) was used for a total volume of 100µl, this was incubated in 37°C for 30min.

2.6.3. Cleaning RNA after repair ends

After the PNK step the RNA was cleaned with a SPRI beads, 0.8X of beads were used with a 2.6X 100% isopropanol for mouse samples and 3X isopropanol for yeast samples. The beads were washed twice with 100µl of 85% ethanol and eluted with 10µl of DNase free water.

2.6.4. First ligation

This step adds an RNA adaptor to the 3' end of the tRNA. To a 7µl sample containing at least 10ng of cleaned, repaired and uncharged tRNA add 1µl of 20µM RNA adapter (RA31) and 1.5µl of 100% DMSO. The samples are heated for 65°C for 2min and then in cold ice. Then 11.5µl of ligation mix was added: 1.3µl T4 RNA ligase 1,Hi conc,36U (NEB, M0204S), 0.2µl 100mM ATP, 7.5µl PEG 8000 (50%), 0.3µl 100% DMSO, 2µl 10X NEB ligase buffer and 0.2µl DNase free water. This was incubated at Room-Temperature for 1hr.

The RNA adaptor (RA31) sequence: rArGrArUrCrGrGrArArGrArGrCrGrUrCrGrUrG, a 5' phosphate and 3' C3 spacer were added to the adaptor by the manufacturer (IDT).

2.6.5. Silane linker cleanup

To clean the RNA from the reaction and adaptors, Dynabeads® MyOne™ SILANE (SILANE beads) was used. For each sample a 12µl SILANE beads was taken and 12µl RLT buffer was added. The beads were separated on a magnet and the supernatant removed. The beads were re-suspended in 61µl RLT. This was added to the sample with 73µl of 100% ethanol. The attached beads were separated on a magnet and washed twice with 70% ethanol and eluted with 12µl of DNase free water.

2.6.6. Reverse Transcription

For the sample 1µl of 10µM RT31 short RT primer was added. The samples were heated to 65°C for 2min and then placed on ice. Then, an 8µl mix was added. The mix contained: 2.4µl DNase free water, 2µl of 10X RT buffer, 2µl of 100mM DTT, 0.8µl of 25mM dNTP mix and 0.8µl of AffintyScript RT enzyme. This was incubated in 54°C for 45min then in 4°C for 1min.
The primers were removed by adding 3µl of ExoSap-it (affymetrix) into each sample and incubating at 37°C for 15min.
The RT primer (RT31) is: ACACGACGCTCTTCCGA.

2.6.7. RNA degradation after RT

The RNA was degraded by adding 1µl of 0.5M EDTA and 2.5µl of 1M NaOH to the sample. This was incubated at 70°C for 12min and neutralized with 2.5µl of 1M HCl.

2.6.8. Silane clean up

To clean the DNA from the reaction, Dynabeads® MyOne™ SILANE (SILANE beads) was used. For each sample a 10µl SILANE beads was taken and 10µl RLT buffer was added. The beads were separated on a magnet and the supernatant removed. The beads were re-suspended in 90µl RLT buffer. This was added to the sample with 107µl of 100% ethanol. The attached beads were separated on a magnet and washed twice with 75% ethanol and eluted with 6.5µl of DNase free water.

2.6.9. Second ligation

This step adds a DNA adaptor with a 6mer barcode to the cDNA originated from full or partial tRNA. To 5.5µl sample, 0.6µl of 100µM DNA adaptor (DA32-DA36) and 0.8µl of 100% DMSO were added. This was heated to 75°C for 2min and then put on ice. A 13.7µl mix was added. The mix contained: 1µl DNase free water, 2µl 10X NEB buffer, 0.2µl of 100mM ATP, 9µl of PEG 8000 (50%) and 1.5µl of T4 RNA Ligase 1, HC, 45U (NEB). This was incubated at Room-Temperature overnight.

The barcoded DNA adaptors were of this format NNNNNNAGATCGGAAGAGCACA and 5' phosphate and 3' C3 spacer were added by the manufacturer (IDT). The barcodes were: ATGCAT, CGTCAT, CTTGGA, TAAGGC and CCGGTA, named DA32-DA36 respectively.

2.6.10. Silane linker clean up

To clean the DNA from the reaction and the adaptors, Dynabeads® MyOne™ SILANE (Silane beads) was used. For each sample a 5µl Silane beads was taken and 5µl RLT buffer (Qiagen, 79216) was added. The beads were separated on a magnet and the supernatant removed. The beads were re-suspended in 61µl RLT buffer. This was added

to the sample with 57µl of 100% ethanol. The attached beads were separated on a magnet and washed twice with 75% ethanol and eluted with 30µl of DNase free water.

2.6.11. PCR enrichment

The cDNA samples were amplified by PCR by adding 1µl of 25µM of forward (DPS1) and reverse (DPSB6) primers, 8µl of DNase free water and 25µl of 2× NEB Q5 HOTSTART MIX into 15µl of sample.
The PCR program ran as follows:

Temperature	Time	Number of repeats
98°C	30 Sec	
98°C	15 sec	
68°C	30 sec	6 cycles
72°C	30 sec	
98°C	15 sec	8
72°C	1 min	cycles
72°C	2min	

Primers sequences were: DPS1-
AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT, DPSB6-
CAAGCAGAAGACGGCATACGAGATGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT

2.6.12. SPRI Library cleanup

In order to clean the library 75µl of Agencourt AMPure XP (SPRI) beads were used. After binding to the beads, a magnet was used to separate the beads, the supernatant was removed and the beads were twice washed with 70% ethanol, air dried for 5min and eluted with 25µl of DNase free water.

2.7. Checking DNA and RNA concentration and length

DNA and RNA concentration was checked using Nanodrop or Qubit Fluorometer (Invitrogen).

DNA or RNA samples were analyzed using electrophoresis in the Agilent 2200 TapeStation system with High Sensitivity D1000 ScreenTape or High Sensitivity RNA ScreenTape, respectively.

2.8. Real time (RT) PCR

Samples were checked for tRNA in different stages of the library preparation with Real Time PCR machine. The samples were diluted 1:50 and for each well 4.75µl of diluted sample, 0.125µl of 10µM of forward and reverse primers and 5µl of Syber Green reagent were mixed. The RT-PCR program was the standard one for Syber Green.

2.9. First ligation calibration

The sample used was *E.coli* Lys tRNA from Sigma-Aldrich (R6018).
For this procedure part of the Advanced tRNA-seq method was used (2.6.4.-2.6.8.), for some samples the RA1 adaptor was added at 2.6.4 while for others it was replaced with water, for some samples other parts of the Advanced tRNA-seq protocol were added (2.6.1.-2.6.3.). The RT primer was changed for different samples. The samples were then checked in RT-PCR with combinations of primers against *E.coli* Lys-TTT.

The reverse primers were: RT31 (Adaptor primer)- ACACGACGCTCTTCCGA, RTlys31 (primer1)- ATTCGAACCTGCGACCAAT, RTlys32 (primer2)- GATTCGAACCTGCGACCAAT and RTlys33 (primer3)- ATTCGAACCTGCGACCAA.

The forward primers (F) were: qPlys51 (used in ligated sample and non-ligate sample1)- GGTCGTTAGCTCAGTTGGTA, qPlys52 (used in non-ligated sample2)- GCTCAGTTGGTAGAGCAGTT and qPlys53 (used in non-ligated sample3)- CTCAGTTGGTAGAGCAGTTG.

2.10. Sequencing

For sequencing an Illumina NextSeq was used for 5 million reads 60bp read from read1 primer and 15bp read from read2 primer.

2.11. Sequence analysis

The sequence data was trimmed using Homertools, and aligned using Bowtie2 with very-sensitive and local preference. It was aligned to specially created reference of unique tRNA sequences, those contained tRNA with introns and without, and also only introns. Another reference included for each mature tRNA (without introns) another copy with the added CCA.
The un-aligned reads were aligned to the genome using HISAT with the very-sensitive preference. The aligned genes were counted using HOMER (Hypergeometric Optimization of Motif EnRichment) (http://homer.salk.edu/homer/) with the analyzeRepeats.pl tool.

3. Results

3.1. In-silico simulation

As mentioned above, RNA-seq difficulty in tRNA is the modified bases which cause the RT to fall-off. We wanted to overcome this by detecting even the partial sequences after the RT fell. In order to determine if tRNA type can be ascertain from a partial sequence caused by a likely scenario of the reverse transcriptase falling off when it reaches a modified base, an in silico simulation of this was created.

We decided to focus on the most stringent scenario and therefore assumed that the reverse transcriptase will fall off on the first modified base, although we knew this is not necessarily the case for all tRNAs, as some modified bases can be read-through. Since sequencing can theoretically be done from either 5’ or 3’ end of the tRNA molecule (although reverse transcriptase is only transcribing from the 3' end), we also wanted to determine if there is a preferable end of the tRNA that will have a more informative sequence.

The known mature tRNA sequences were used to check for the most likely position of the first modification in either side of the tRNA, in order to establish the likely tRNA fragment length that will be created if the RT will fall at the first modification.

For example the histograms of length of the known tRNA mature sequences from each end until the first modified base in yeast is shown in figure 2.

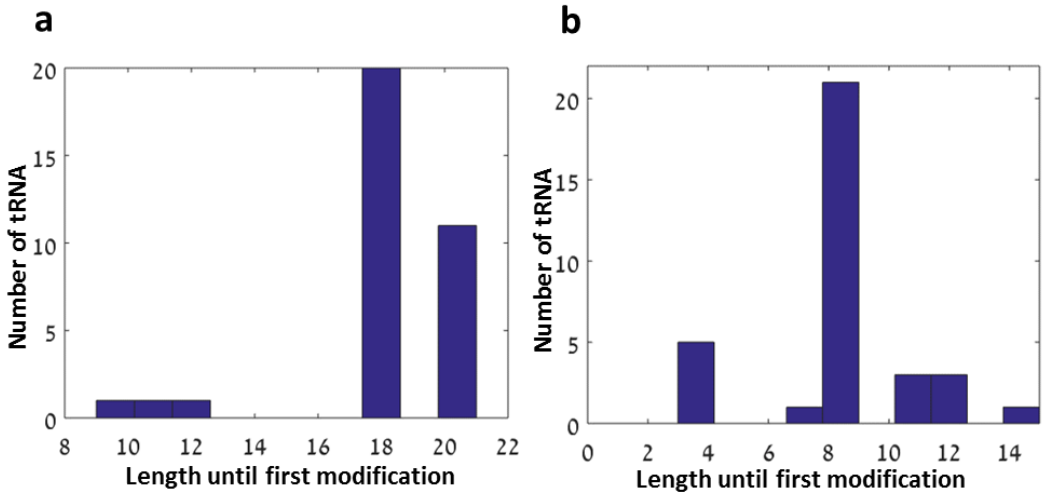

Figure 2- Histogram of length until first modification in both ends of known yeast tRNA. **(a)** shows the length from the 3' end and **(b)** shows the length from the 5' end.

As can be seen by the histogram, we found that the most common modified base appear after 8nt from the 5' end and after 18nt after the 3' end of the mature sequences. This result was not only true for yeast tRNA but also for all other species checked (mouse, human and *E.coli*).

We next set up to check whether such short sequences (i.e. 8nt from the 5’ end and 18 from the 3’ end) are sufficient to determine the identity of the tRNA from which they were generated. For that we took the 5' end 8nt sequence was taken and from the 3' end 15nt sequence was taken (as 3 bases from the 3' end of mature tRNA are the post-transcription added CCA sequence).

Those partial sequences were compared against each other in order to evaluate the ability to differentiate between the different tRNA types. The number of differences among the pairs were counted and shown in figure 3 (for example in yeast):

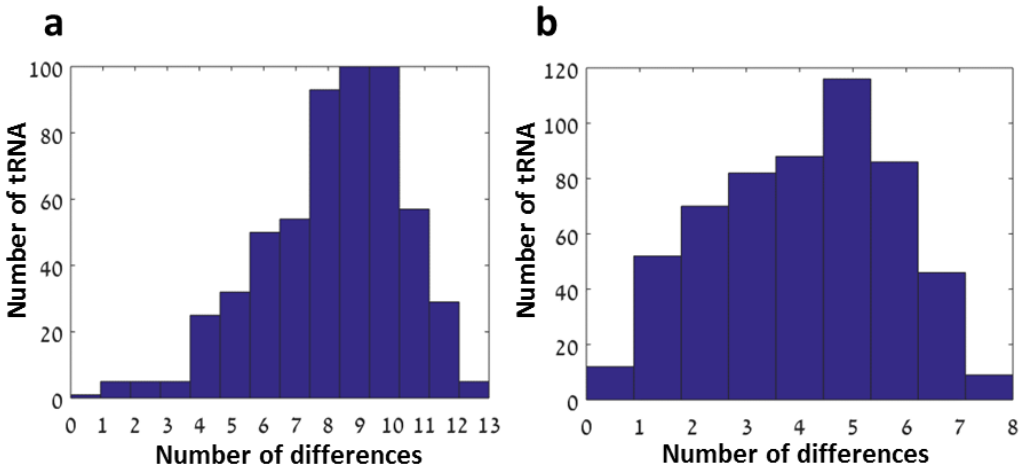

Figure 3- Histogram of differences from one end until estimated first modification in unique yeast tRNA genes. **(a)** shows the number differences in 15nt from the 3' end and **(b)** shows it in 8nt from the 5' end.

Figure 3 shows that among the different tRNA sequences most of the partial sequences show difference in the sequence, also it can be seen that for most sequences the difference is more than one (in 3' end the most common difference number is 9 differences and in 5' end it is 5 differences).

The percentage of identical pairs (0 differences), were examined in table 1.

table 1-Percentage of identical pairs

Organism	Human	Mouse	Yeast	E.coli
	%			
5' end	1.26	1.68	1.66	0.65
3' end	0.37	0.82	1.24	0.33
Total pairs	105,111	31,125	4,753	1225

The table show the percentage of identical pairs among tRNA genes from different organisms, when comparing 15nt from the 3' end and 8nt from the 5' end.

Table 1 shows that the percentage of identical tRNA sequence pairs is low (less than 2%) among the number of species checked. It is also shows that the percent of identical pairs

is lower in the 3' end. Not unexpected as the 3' end partial sequence is longer and thus has more possible combinations.

The identical sequences that were shown in table 1 were explored further and separated into those that share the same anti-codon (isodecoders) or the same amino-acid (isoacceptors). Those pairs allow us to still differentiate the tRNA at least based on anti-codon or amino-acid and by doing so gives us valuable information (table 2).

Table 2-Percentage of same anti-codon or amino acid among identical pairs

Organism	Human genes		Mouse genes		Yeast genes		E. coli genes	
	Same anti-codon (%)	Same amino acid (%)	Same anti-codon (%)	Same amino acid (%)	Same anti-codon (%)	Same amino acid (%)	Same anti-codon (%)	Same amino acid (%)
5' end	56.68	68.84	65.46	91.41	68.18	83.33	50	87.5
3' end	66.75	84.16	70.82	96.11	91.3	93.48	100	100

The percentage of identical pairs in regard to the length until first modification, that share anti-codon of amino acid in different organisms.

Table 3 summarize table 1 and 2 as it deduces the number (and not percentage) of pairs among identical pairs of partial sequence that don't share the same anti-codon or amino-acid. This table allows us to recognize the theoretical limits of the method under the stringent assumption that the reverse transcriptase falls off at the first modification. The number of pairs that don't share amino-acids are those that theoretically can't be distinguished with this method.

Table 3-Number of different anti-codon or amino acid among identical pairs

Organism	Human genes		Mouse genes		Yeast genes		E. coli genes	
	different anti-codon	different amino acid	different anti-codon	different amino acid	different anti-codon	different amino acid	different anti-codon	different amino acid
5' end	574	413	181	45	25	13	4	1
3' end	129	62	74	10	5	4	0	0

This table determine the number of pairs that can't be separated that don't share the same anti-codon or amino acid. The pairs are from tRNA genes ends until estimated first modification in different organisms.

Those results gave us confidence that by looking at partial sequence, preferably from the 3' end, we can distinguish between most of the tRNA types.

3.2. Separation of tRNAs from total RNA

Encouraged by the bioinformatics analysis above we set out to develop the tRNA deep-seq protocol (Advanced tRNA-seq). The first step is enriching for tRNA molecules out of all cellular RNA. In order to separate tRNA from total RNA we chose to use Solid Phase Reversible Immobilization (SPRI) beads. Those polystyrene beads are paramagnetic and coated with carboxyl groups. When adding different concentrations of salts and alcohols to a solution containing polynucleotide it allows differential precipitation and adherence to the beads of the polynucleotides based on their length.

There were no protocols for the use of those beads with tRNA separation so, it was needed to be calibrated.

We wanted to find the right conditions that allows to retain RNA in the area of 72nt (tRNA length) with high concentration and that allows as little as possible the inclusion of other RNA lengths. As SPRI beads allows the separation of polynucleotides based on different concentration of salts and alcohols we checked various combinations of those. For the salts we used different volume of SPRI beads (as the beads included in a buffer while only small amount of beads is actually needed for the separation), and an additional

SPRI buffer. As for the alcohol we used 100% isopropanol in different volumes. This was first calibrated on a yeast isolated tRNA and after this used on yeast total RNA samples (Figure 4a,b).

In order to separate the tRNA, two steps are required, as the beads can only attach to RNA from a certain size and above. The first step is without isopropanol and is used to remove the large (above 200nt) RNA. The second step is with isopropanol and is used to attach the medium size RNA (from 50 to 200nt), and separate it from the small RNA. This step was calibrated to get more tRNA size RNA from the medium size RNA than the other types of medium size RNA (figure 4b). It was found that a second round of those two steps get an even better separation (figure 4c).

This calibration was then fine tuned to the V6.5 mouse ES cells (figure 4d).

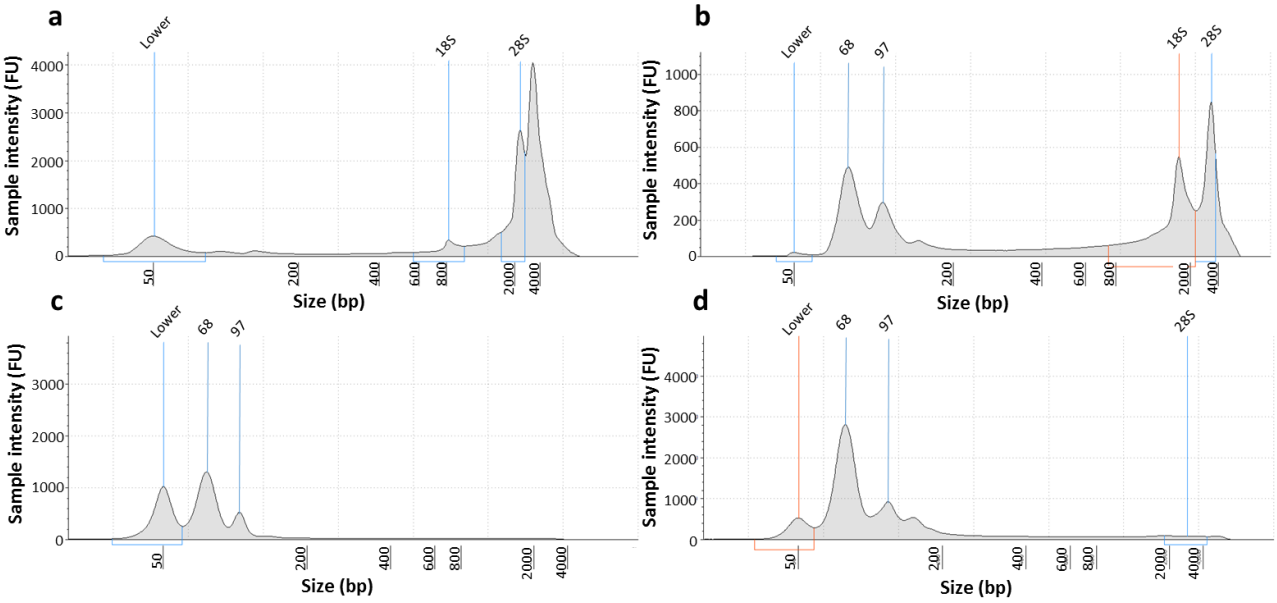

Figure 4- Length Histogram of RNA using tapestation: **(a)** yeast total RNA without SPRI separation, **(b)** Yeast RNA after one round of 0.8X SPRI beads and 3X isopropanol**, (C)** Yeast RNA after two rounds of 0.8X SPRI beads and 3X isopropanol, and **(d)** V6.5 Mouse RNA after two rounds of 0.8X SPRI beads and 2.6X isopropanol

3.3. Generating Advanced tRNA-seq libraries

In order to sequence the partial tRNA sequences, the scheme shown in figure 5 was proposed (based on unpublished protocol from M. Guttman group):

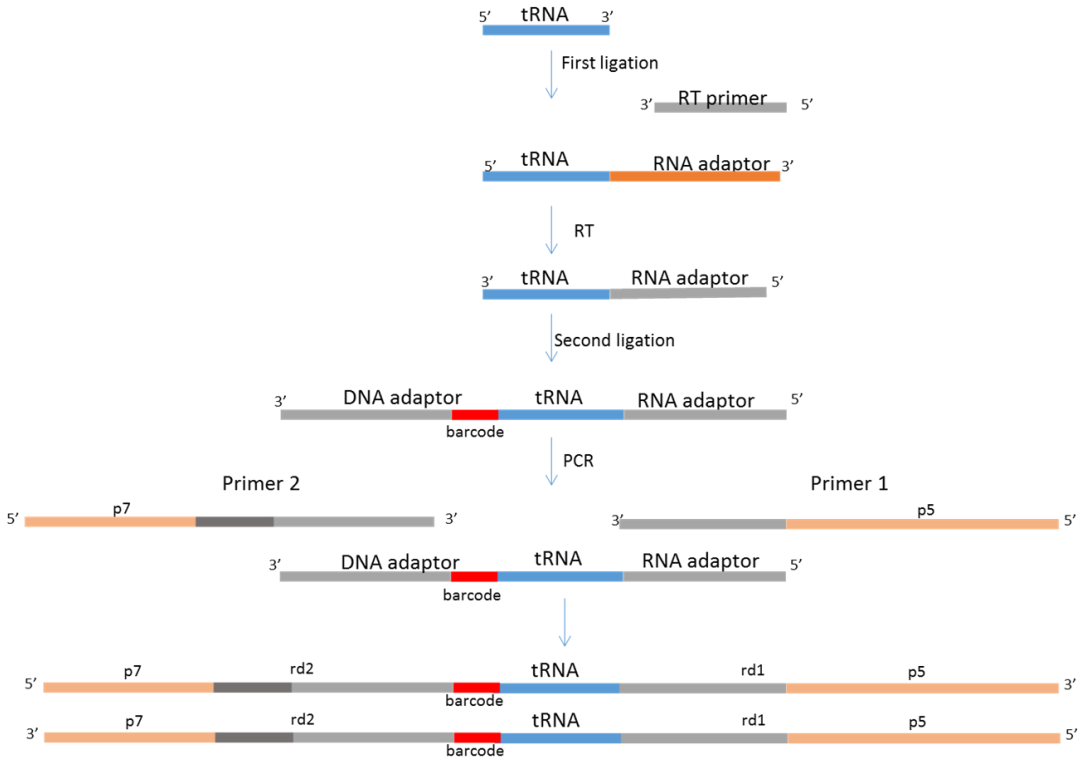

Figure 5–Scheme of Advanced tRNA-seq library preparation.
The p5 and p7 are sequences needed for the illumina primers. The 6 mer barcode is in red.

In this scheme (Figure 5), tRNA molecules separated from total RNA are used. First an RNA adaptor is ligated to the 3' end of the molecule. Then a reverse transcription step is performed using a primer for the RNA adaptor previously attached. It is assumed the reverse transcriptase will fall if it reaches certain modified bases. After this another adaptor is ligated, this time a DNA adaptor is ligated to the new 3' cDNA (that corresponds to the 5' end of the original tRNA). At the end those two adaptors are used to enrich the tRNA by PCR using primers that contains specific barcodes and illumine sequencing adaptors and thus preparing it for sequencing.

3.4. First ligation calibration

The first ligation process is a very important step in the Advanced tRNA-seq protocol, which is because only RNAs that have an adaptor attached can be reverse transcribed and eventually sequenced. We wanted to make sure that the method is indeed efficient and a representative sample is being ligated and reverse transcribed. We also wanted to check if

additional steps will improve the ligation efficiency. Our main concern was that the ligation may be hindered by the amino acids attached to the tRNA at the 3' end.

The method that was proposed to check this step has its pros and cons.

The method was to quantify the tRNA with adaptor that underwent reverse transcription and compare it to all the tRNA in the sample that underwent reverse transcription without adaptor ligation (figure 6). This will allow us to quantify the tRNAs with adaptor to the total tRNA. This will be problematic if used in a heterogeneous sample, with many types of tRNA, as the reverse transcription step requires a specific primer. This problem was reduced by using a tRNA isolated for a specific amino acid. Another difficulty is the difference in primer sequence for the reverse transcription. This was addressed by using several primers and using similar Tm for all of them (2.9. in methods).

The experiment was done on Lys tRNA from E.coli as all those tRNA types share a very similar sequence, and so it also reduces the problem of reverse transcribing only a small portion of the tRNA.

The experiment was done to samples that didn't have any additional treatment, and on samples that had. The treatment included amino acid removal, alkaline phosphatase usage and repair ends.

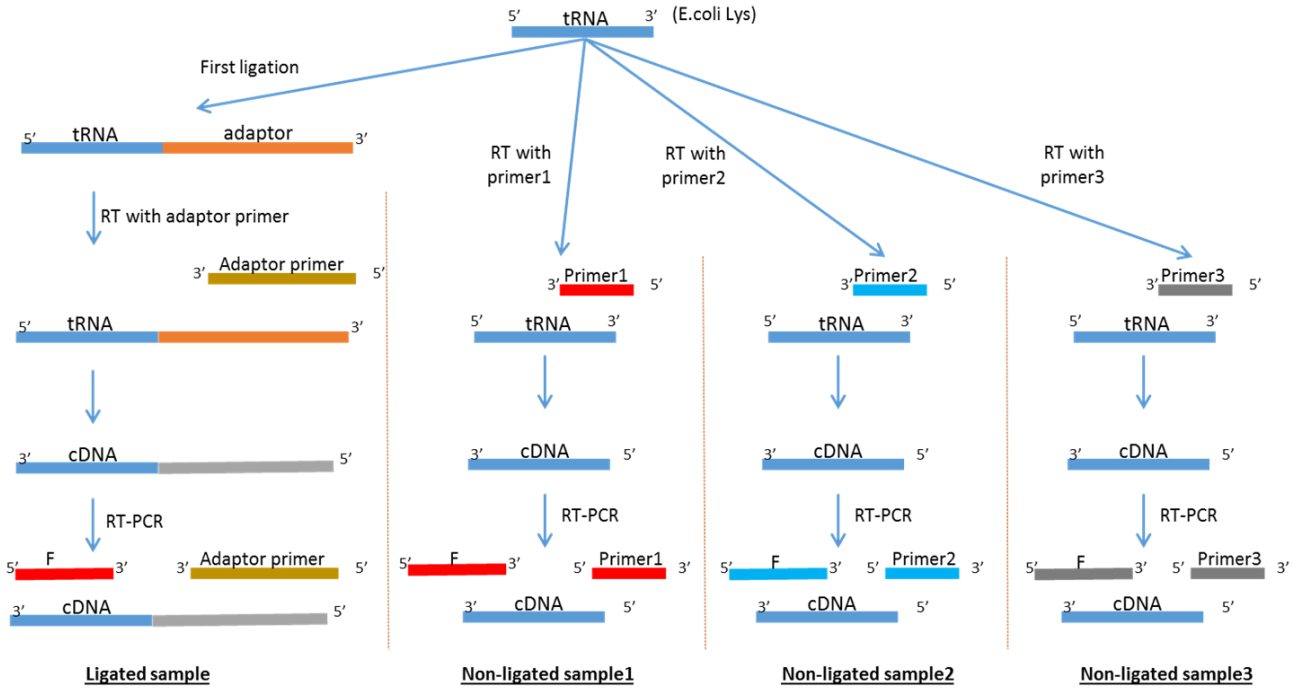

Figure 6- Scheme for the checking first ligation efficiency method.
The left panel show the scheme after the first ligation while the other panels show scheme without the ligation. For each panel there are different primers for different places on the tRNA or the adaptor. The bottom show samples names.

This was checked in RT-PCR in order to estimate the tRNA in the different samples. The results show that when comparing the additional treatment to the no additional treatment, the number of tRNA detected in the samples that were ligated (1/2) increases by 2^9, while the number of tRNA in the non-ligated samples (3/4, 5/6 and 7/8) don't change much. This means that using alkaline phosphatase and repair ends treatment improved the 3' ligation.

We also, can see that before the treatment, the ligated sample tRNA are lower than the non-ligated samples, but after the treatment the number of tRNAs in the ligated samples are much higher the non-ligated samples. This indicates that we didn't succeed in nullifying the PCR bias toward the adaptor primer and so when there are more ligated tRNA (after treatment) the difference between those groups increases exponentially. A use of UMIs (Unique Molecular Identifiers) may be able to remove this bias in future experiments.

Another thing that we could measure in the RT-PCR was the difference in the abundance of a single tRNA from the ligated sample and from the non-ligated sample. If the efficiency of ligation is high there shouldn't be a difference between those samples as most of the tRNA got an adaptor and then reverse transcribed. We got that the ligated samples have a quarter of the specific tRNA than the non-ligated samples. This indicated the ligation efficiency is 25%. This result increases the suspicion that the reason for the 2^9 ratio between the ligated to non-ligated samples is partly due to PCR bias as there we used different primers.

3.5. Advanced tRNA-seq analysis

The Advanced tRNA-seq protocol was used several times in different conditions and samples: V6.5 mouse ES cells, yeast cells in different treatments and order of the protocol and in different stages of diauxic shift. It is important to notice that in the length histogram of the final libraries there were 2 or 3 major peaks, indicating that not only tRNA size RNA got sequenced.

The sequencing was evaluated according to the percentage of alignment and the correlation between fractions of different anti-codons in the reads to the gene copy number (only in yeast). The best result was when we used repair ends treatment and alkaline phosphatase and removed the amino-acids (removing AA before and after SPRI size selection didn't have any considerable effect). In this case the percentage of alignment was between 25-35% and the correlation was between 0.27 to 0.5. We used mouse V6.5 samples, 4741 yeast grown on rich medium and 4741 yeast before, during and after diauxic shift. The analysis was mostly focused on the yeast grown on rich medium, dubbed here “yeast1”.

The reads that were not aligned to the tRNA genes were then aligned to the yeast genome at more than 70% alignment (the remaining 30%, are those that didn't align to either tRNA genes or the yeast genome). Among the genes that they were aligned more than half were mapped to medium size genes, 80% of them were snRNA, which correspond to the SPRI beads size selection range.

After the reads were aligned to the tRNA reference, they were separated into each anti-codon (isoacceptor). We were able to detect 42 out of 42 yeast isoacceptors and 69 out of 69 different tRNA sequences in the yeast genome (including non-mature ones). We then compared the number of reads detected for each isoacceptor to the gene copy number that codes for this isoacceptor family (Figure 7). This analysis yield relatively mild correlation of 0.4.

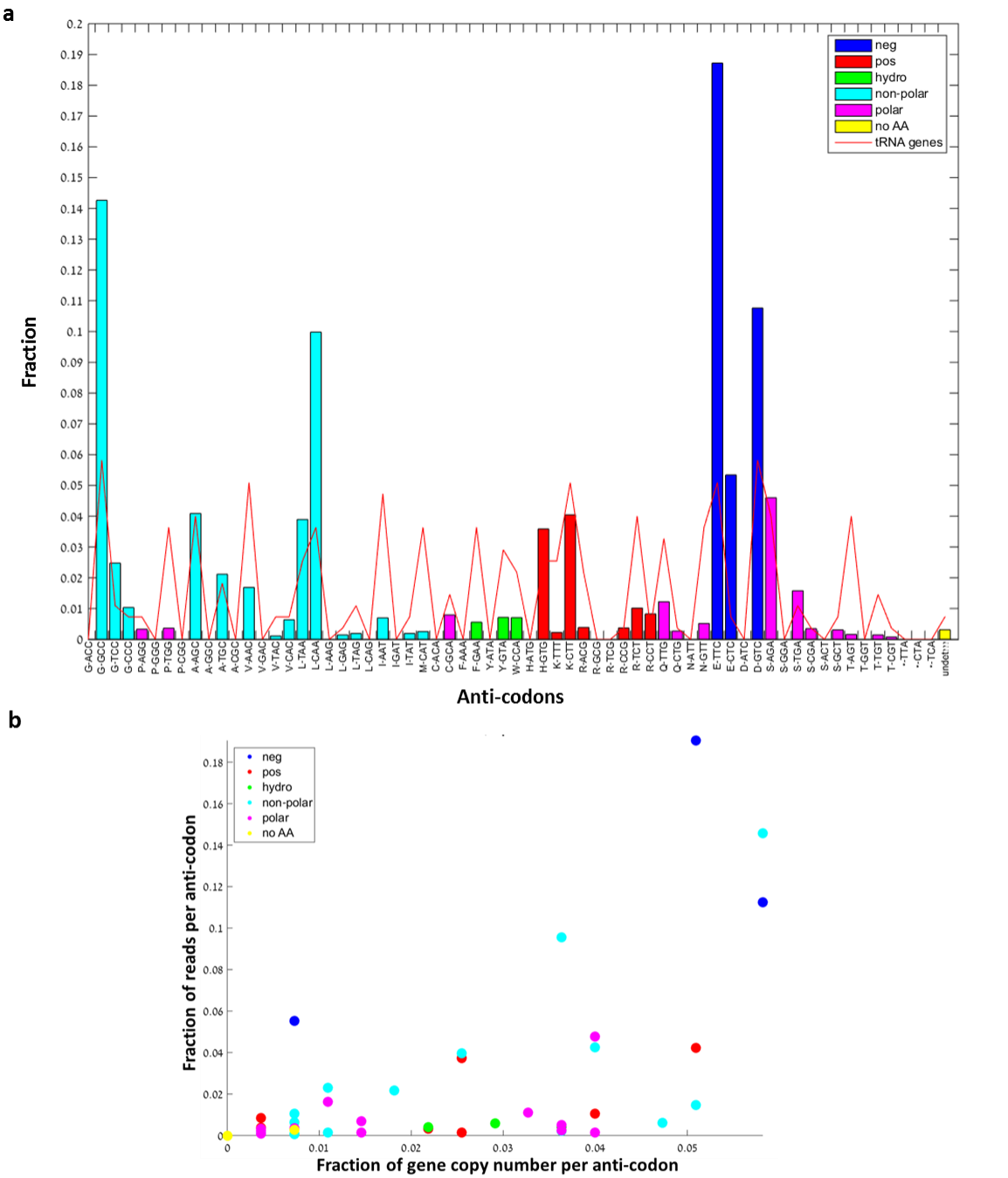

Figure 7- Fraction of anti-codons separated tRNA and fraction of gene copy number of tRNA in mapped reads of yeast1 sample.
(a) The bars indicate the fraction of anti-codon from the total mapped reads. **(b)** Scatter of fraction of anti-codons separated tRNA against fraction of gene copy number of tRNA.
The colors of the anti-codons are according to their amino-acid group. The red plot shows the fraction of the specific genes among the tRNA genes. This graph includes also undefined anti-codons and stop codons tRNA genes. The Pearson R^2 correlation of the reads separated to anti-codons and the gene copy number is around 0.4.

Another approach to corroborate the Advanced tRNA-seq results is to compare it to today's gold standard, which is microarray for tRNA. For this purpose, we sequenced yeast tRNA samples that the lab previously analyzed using tRNA custom microarrays (6). These samples were taken at different time points during the growth of *S.cerevisae* on rich media and cover cells before and after the switch from fermentation (t=0 [t0] and t=6h [t6]) to respiration (t=9h [t9]). In order to compare them we had to look at the ratio between samples t6 and t9 to sample t0, as was done in the microarray.

In figure 8 we see scatter plots of the normalized ratios of t6 and t9 to t0 for each isoacceptor family in both the microarray and Advanced tRNA-seq. In figure 8 appears also, the correlation between the two methods. The correlation between the methods is good (~0.6-0.7) indicating a similar trend in the tRNA detection, although we do see some tRNA isoacceptors that are relatively high in the Advanced tRNA-seq while other are relatively high in the microarray (figure 8).

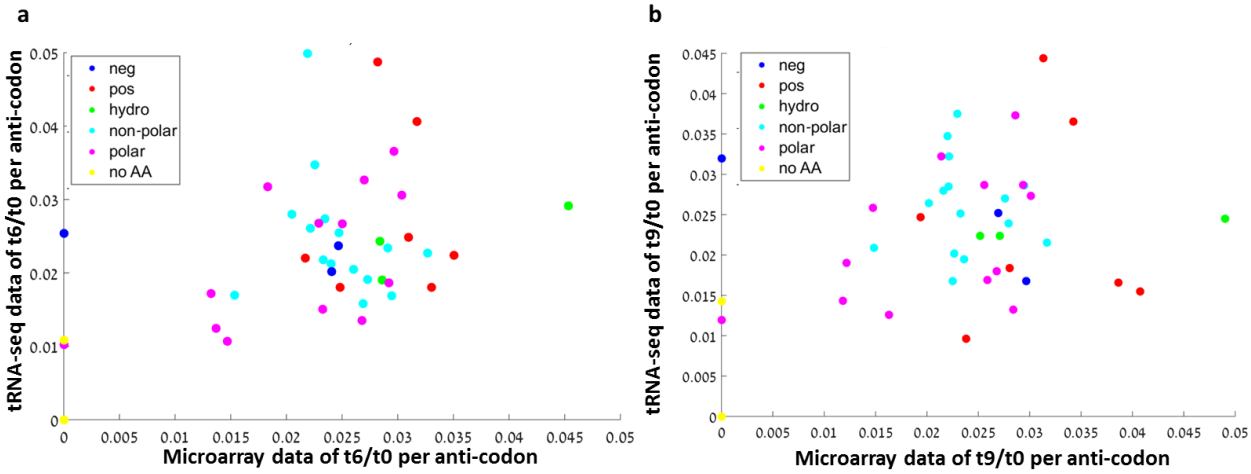

Figure 8- Comparison between microarray and Advanced tRNA-seq in t6/t0 and t9/t0 yeast samples.
This figure looks at the normalized to total of the read count ratio between time point t6 or t9 to time point t0. The scatter plots shows a ratio separated to anti-codon in microarray against Advanced tRNA-seq. The colors of the anti-codons are according to their amino-acid group. **(a)** shows the ratio of t6/t0 while **(b)** shows the ratio of t9/t0. The R^2 of t6/t0 is 0.658 and of t9/t0 is 0.593.

3.6. Analysis of tRNA modifications

The modified bases in the tRNA can possibly affect the RT by either causing it to insert a different nucleotide or by stopping the reverse-transcription and the RT falling off. The way the RT is responding to every type of modification is not yet clear and a large portion of tRNA are not characterized by the location and type of modifications. The

Advanced tRNA-seq method lets us delve into the modification effect on the RT, and can help us characterize the modifications.

The reads were separated according to their length before and after the alignment (figure 9). This gives an estimate of the fraction of molecules that have a modified base that cause the reverse transcriptase to fall off prematurely. We can see that before the alignment the largest fraction of the reads are at full length and there is falling off at length of 18nt and leading to 65nt. The picture completely changes after alignment as we see that most of the reads are shorter than full length, the majority are at lengths 71, 38, 47 and 59. The falling-off are probably caused by certain modification that are prevalent at those positions. As only around 30% of reads were aligned, it seems possible that a huge faction of the un-aligned reads were not tRNA, as other medium size RNA don't usually have modification that can cause the RT to fall. In order to check this we did a follow-up alignment to the genome of those reads that didn't align to the tRNA. This assumption seems correct, as it was found that more than half of them aligned to genes that are shorter than 200bp (and thus included in the SPRI beads separation), among those more than 80% were snRNA genes. The histogram of length of the reads that were aligned to the genome showed that more than 98% of them were full length, this is also an indication that those reads originate from non-tRNA RNA as they lack modified bases and thus are at full length. The reads that were aligned to the genome had 70% alignment rate. The reads that weren’t aligned to either the tRNA reference or the Yeast genome are consisting of more than 95% full length reads.

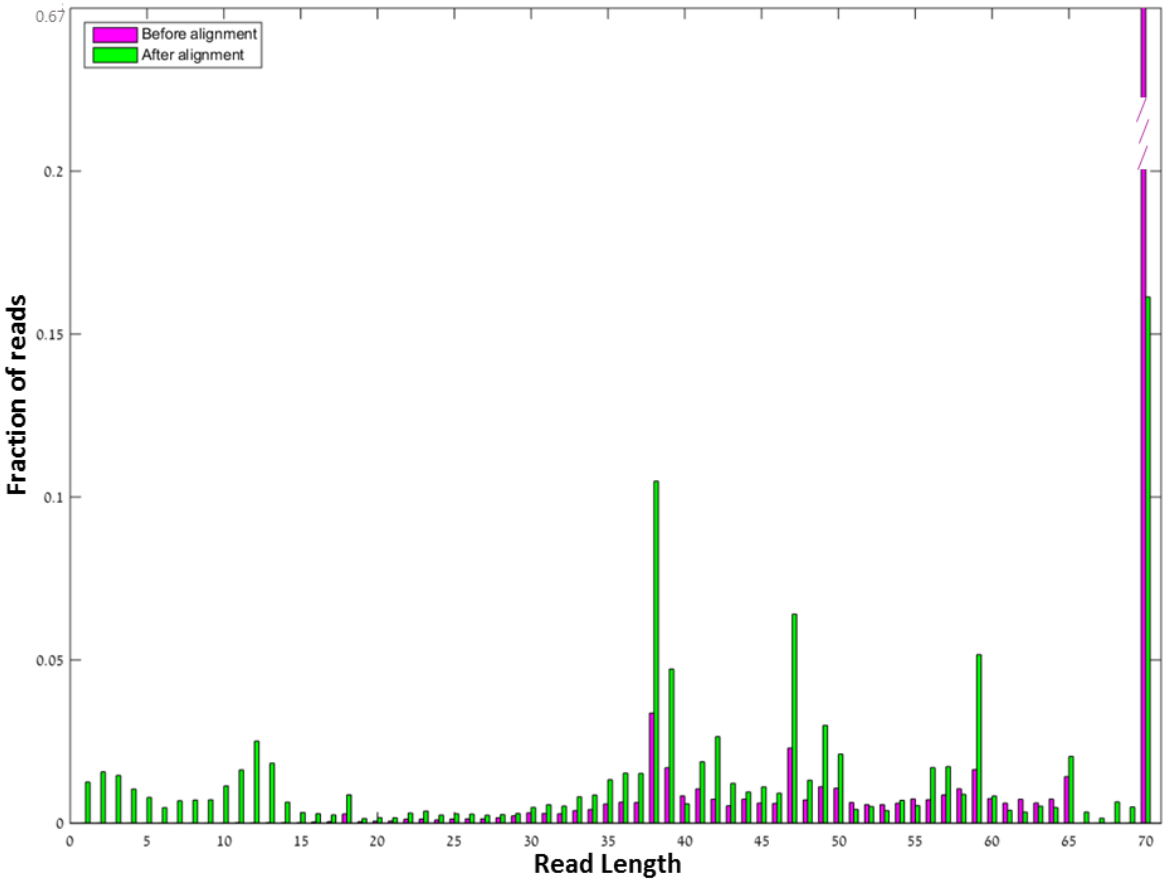

Figure 9-Length histogram of read size before and after alignment in yeast1. The before alignment (purple) is reads that were trimmed for the adaptors, and the after alignment (green) is the size of the aligned part in reads that were mapped to tRNAs.

As each isoacceptor tRNA can include different tRNAs that differ in sequence and potentially modification, the length distribution of the reads was further separated according to the different tRNA species, in order to find the "behavior" of the RT in different molecules (figure 10). It was found to have very large difference between the different tRNAs. Some tRNAs have mostly full length reads, dubbed here “class I”, this can be explained by either not having any modification that can lead to RT falling off, or that only full length tRNA are being aligned (this option is less likely as was seen in the in-silico simulation). Other tRNAs have mostly a short length reads (full length is less than 10%), dubbed here “class II”, this is not saying that the length of the short reads are uniform. This can be explained by the presence of modification that the RT cannot read-through. The last type of tRNAs, dubbed here “class III”, are those that have a mixture of

reads in various lengths, including full length (full length between 10-50%). This can be explained by having modifications that cause the RT to fall off in only some of the cases. We found that from the mature and non-mature tRNAs 56% of the tRNAs belong to "class I", 29% to "class II" and 15% to "class III" (Figure 10d). While we found that from only mature tRNAs 4% belong to "class I", 80% to "class II" and 16% to "class III" (Figure 10d). It seems that the non-mature tRNA contribute mostly to "Class I", as expected, as they should not be modified.

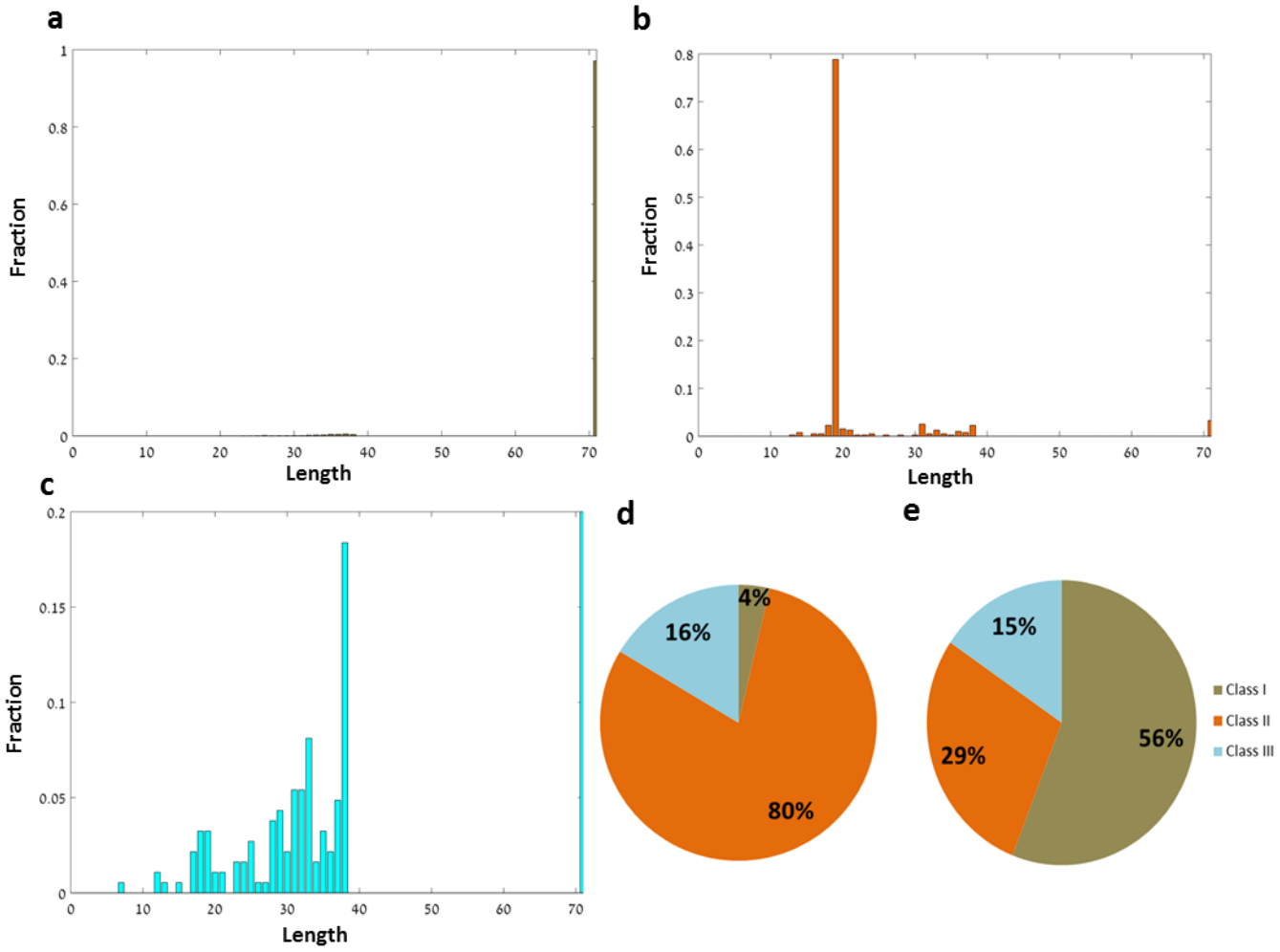

Figure 10- examples of the 3 classes of tRNA seq read length distributions. **(a)** Class I- yeast1 Glu-TTC, **(b)** Class II- yeast1 Pro-TGG , **(c)** Class III- yeast1 Thr-CGT. **(d)** Pie chart of the percentage of mature tRNAs from each class, from a total of 55 tRNAs species. **(e)** Pie chart of the percentage of mature and non-mature tRNAs from each class, from a total of 86 tRNAs species.

The tRNAs were also analyzed by their read coverage along the tRNA, or in other words, the number of reads that were mapped to each position along the tRNA (figure 11). This gives us a general understanding of the possible falling-off events and their positions. The tRNAs were divided according to their coverage profile to 3 types: Type 1 (figure 11a) which has high coverage (more than 50% of the maximum) along the entire length; Type 2 (figure 11b) which are the majority in the mature tRNAs, and have low coverage at the beginning of the tRNA (5' end), this can be caused by a modified base that interfere with the RT; and Type 3 (figure 11c) which are the majority in the mature and non-mature tRNAs and a small minority in only the mature ones, and have low coverage at both ends of the tRNA and high coverage in the middle (the middle is defined by being not the end or beginning). The Types classifications are not to be confused with the Class classifications that appear at figure 10, as the Classes are separating tRNAs according to the percentage of full length reads while the Types are separating the tRNA according to the position of reads along the entire molecule, that being said, Class I are necessarily Type 1 although not the other way around as long reads can align not from the 3' end and could count as Class II or III. Figure 11 allows comparison between only mature tRNA (those that end with CCA without introns) and the mature and non-mature tRNA sequences which include also tRNAs with introns. The non-mature tRNA include mostly type 3 and also the remaining 14 that can be called type 4 and have low coverage on the 3' end but high coverage in the beginning (5' end). It seems that the tRNAs that contains introns are abundant in type 3 and the highest part of the coverage located in the intron area (Figure 12a). This indicate that either there is some alignment issues with the reads that are not introns (those reads align to the mature sequences even when they originate from the non-mature tRNA) or that there is some tRNA fragments that are the result of intron containing tRNAs cleavage. From figure 12 we can also observe that there don't seems to be a relation between the mature and non-mature forms in regard to the coverage, so the low coverage in the non-mature are not the result of alignment to the mature form. Not all intron containing tRNA have the same profile as Ile-TAT, some are Type 1 or Type 2, and the intron area is in some cases highest while in others there is a low coverage there while others are the same height.

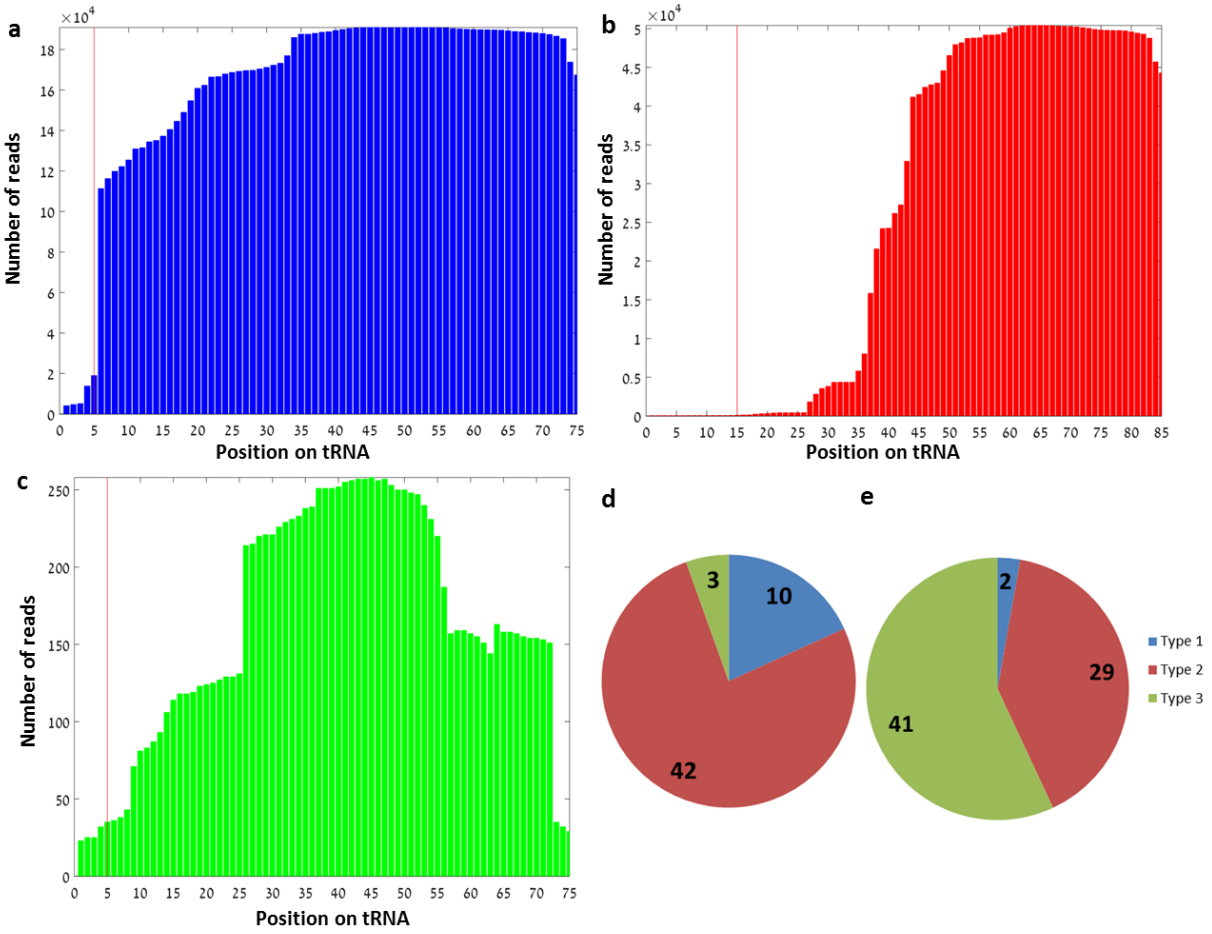

Figure 11-Different coverage profiles of tRNAs.
Those example show types of coverage defined by different levels of coverage profile in the beginning, middle and end of the tRNA. Low coverage was defined as having less reads than half of the maximum read. The red line indicate the end of the sequencing. **(a) Type 1**-Glu-TTC-1-1-most of the reads reach the beginning and thus are full length, **(b) Type 2**-Ser-AGA-1-1-most of the reads don't cover the beginning probably because of falling-off, **(c) Type 3**-iMet-CAT-1-1- most of the reads are in the middle with less coverage at the end, can be caused by fragmented tRNA or alignment problems. **(d)** Pie chart of the number of mature tRNAs from each type, from a total of 55 tRNAs species. **(e)** Pie chart of the number of mature and non-mature tRNAs from each type, from a total of 86 tRNAs species (the remaining 14 comprise another type that have more coverage at the beginning than the end).

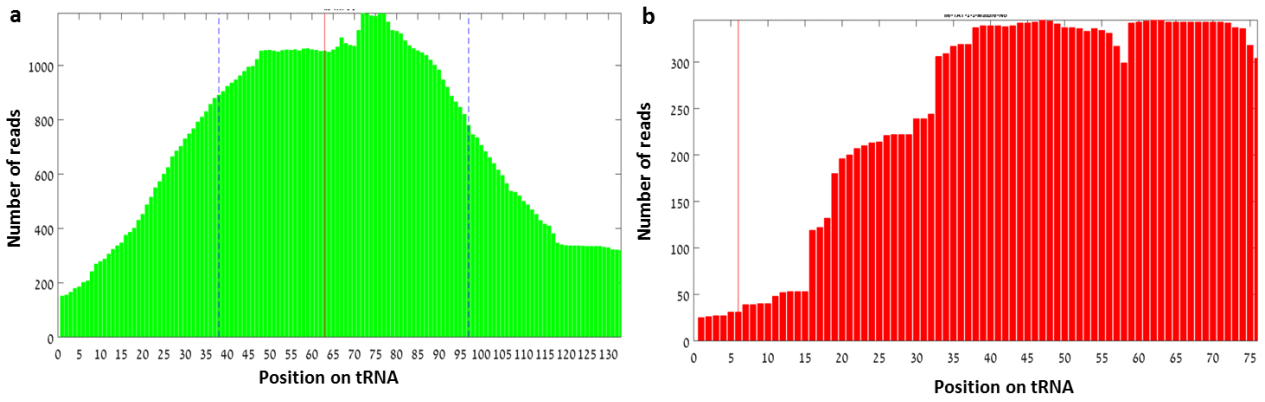

Figure 12-Coverge of Ile-TAT-1-1 in non mature and mature tRNA.
(a) The coverage along the non-mature form of Ile-TAT-1-1- a Type 3. The blue dotted vertical lines indicate the start and end of the intron. **(b)** The coverage along the mature form of Ile-TAT-1-1- a Type 2. The vertical red line indicate the position of the end of sequencing if the read would have started from the 3′ end of the molecule.

The reads were also separated according to their end position on the tRNA, this was done according to each tRNA sequence.

For example for the tRNA of Thr-AGT-1-1 to which there is a known mature tRNA sequence with the modification we can see that peaks position is in the same positions of the modification (figure 13).

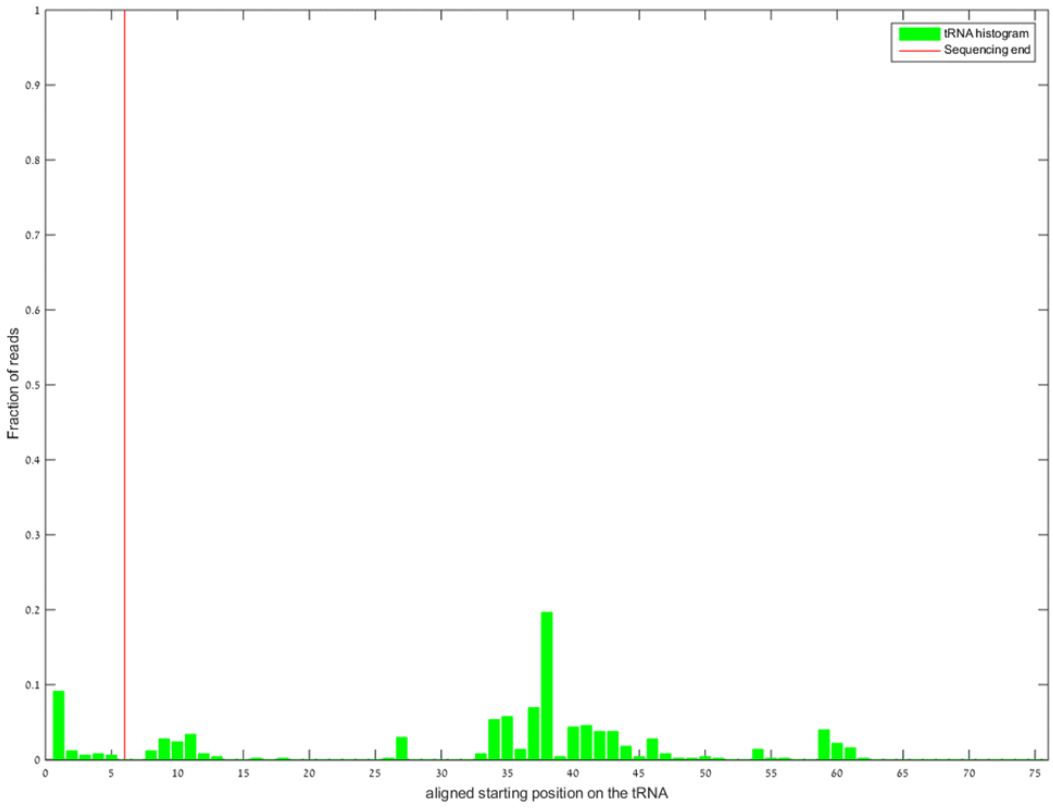

Figure 13- Positions of falling off in tRNA Thr-AGT-1-1 in yeast1.
The fraction of falling off at position is indicated by the green bars, the red line indicate the length of the sequencing assuming the read started at the 3' end of the tRNA.

In order to determine the likelihood of the RT to fall off at a certain position, the end position was normalized to the number of reads at that position, so for each position along the tRNA we get the fraction of reads that fall-off at that position from the remaining reads. In addition, in order to be able to connect the modified bases positions to the fall-off of the RT, as was assumed to be the likely case, we looked at the tRNA with the known modifications types and locations (known tRNAs) and found their likely gene by blast and were able to map the modifications to some of the reference tRNAs (29 out of the 34 known tRNA were matched). An example of this figure for tRNA Phe-GAA-1-1 with the modified bases positions is shown in figure 14.

In those types of figures (figure 14), it can be seen that some falling off seems to occur at, before or after a modification, but some modification don't seems to cause a falling off,

while some falling off peaks don't seem to be related to a modification. Also, the fraction of falling-off is different for different modifications. In Phe-GAA-1-1 shown in this graph (figure 14) it can be seen that the large falling off peak at position 38 can be the result of the modified base P (Pseudouridine) that located at position 39 or from Y (wybutosine) from position 37. If we remember that the RT is reverse transcribing from the 3' end of the molecule, we can assume that Y is the cause of the falling off as the RT falls before it reaches it. Also, it seems that P don't cause falling-off in other places it appear at.

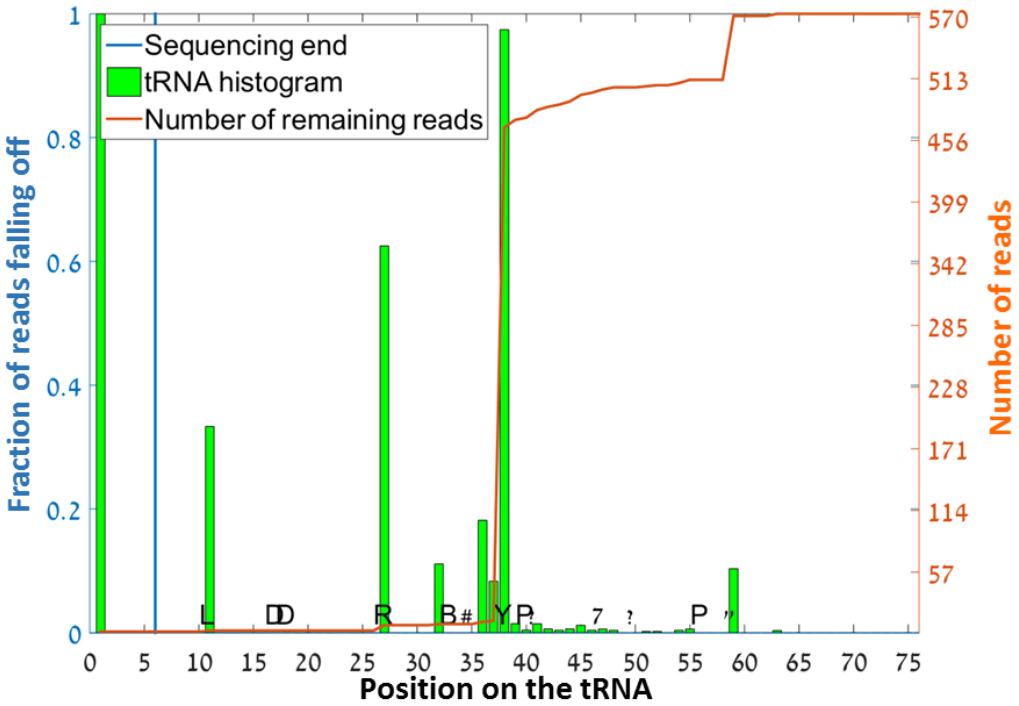

Figure 14- Chance of RT to fall off at different position across tRNA Phe-GAA-1-1.
The probability of the RT to fall off along the tRNA is indicated by the bars which correspond to the left y-axis, the blue line indicate the length of the sequencing assuming the read started at the 3' end of the tRNA, the orange plot shows the number of remaining reads and the modification type appears at the bottom of the y-axis.

Another possible effect of base modification can be the insertion of incorrect base during the reverse transcription process. Our data set provide us with unique opportunity to examine the extent of this phenomenon in a comprehensive manner. Therefore, in order

to find the sequencing miss-incorporations of nucleotides because of a modified position, the known modified position in the tRNA's were examined to calculate the fraction of each nucleotide in the position. The results were compared to upstream position in the tRNA without modification (figure 15).

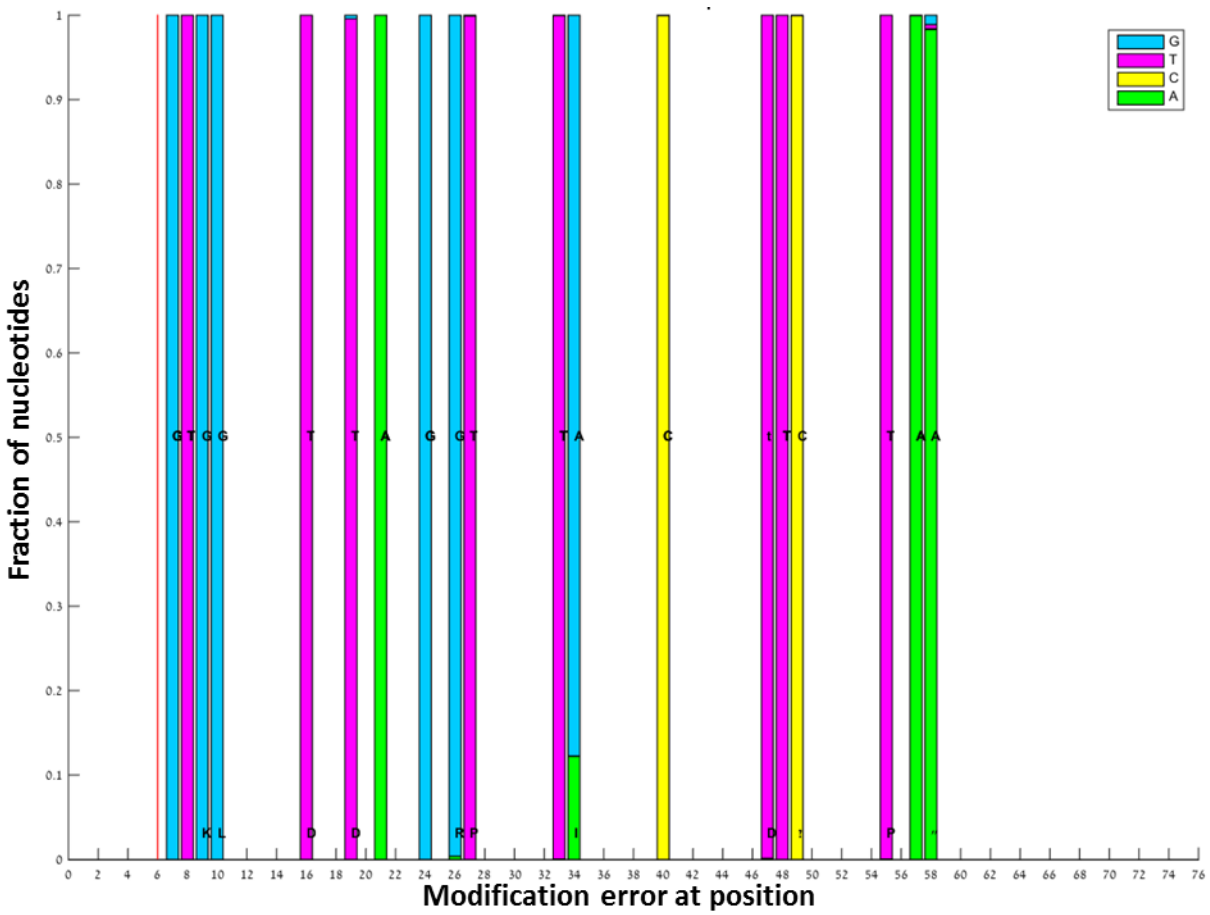

Figure 15- Fraction of nucleotides at position of modified bases in yeast tRNA Arg-ACG-1-1. For each position of modification along the tRNA it shows: the nucleotide composition among the different reads in blue (G), violet (T), yellow (C) and green (A); the modification type at the bottom of the y-axis and the original nucleotide in the middle of the y-axis.

The fraction of nucleotides was then separated according to the modification across all positions in all mature tRNAs. The modifications that has at least 5% of change from the original base are shown (figure 16). Most of the modifications don't show any change from the original base, although the modifications Pseudouridine and m3C has more than 1% of change. We can see that only at Am (2'-O-methyladenosine), Inosine and m1G the modification causes the majority (85-100%) of nucleotides to be miss-incorporated (reads

a T instead of A in Am, G instead of A in Inosine or G as a C in m1G). Also the modification m22G and m1I show a fraction of miss-incorporation (10-25%). This seems to suggest that most types of base modifications don't cause a change in nucleotide insertion, and those that do, are causing a change in various amounts.

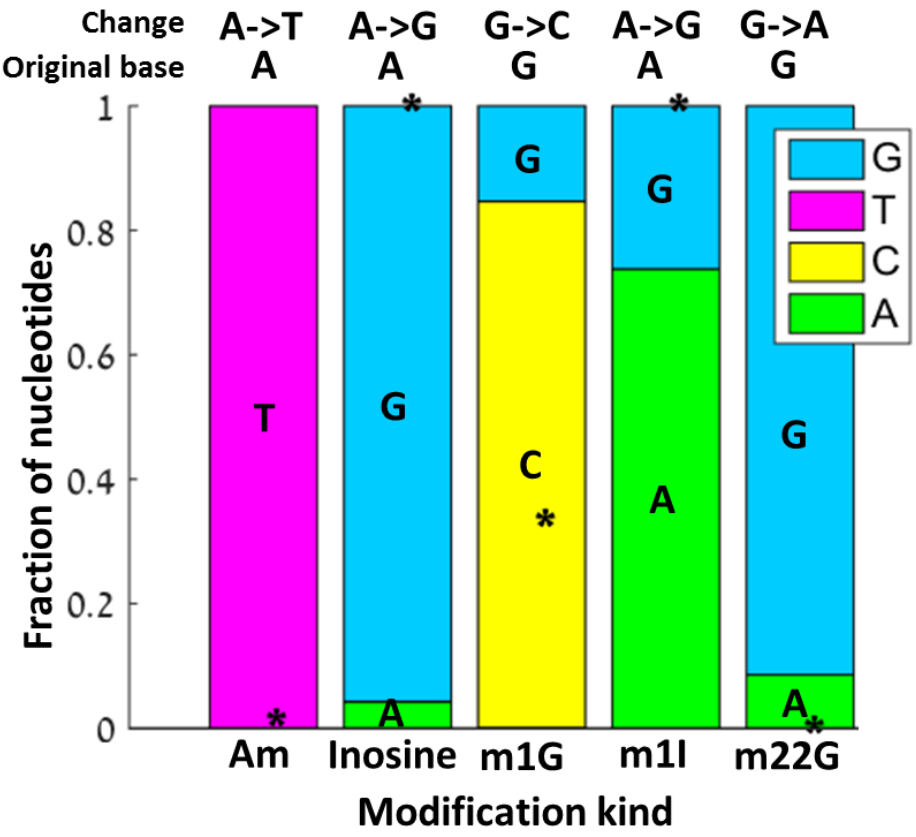

Figure 16- fraction of nucleotides per modification type across all known mature tRNA in Yeast1.
For every modification that is known in tRNA and has at least 5% variation from original base, it shows: the nucleotide composition among the different reads in blue (G), violet (T), yellow (C) and green (A); the modification type at the bottom of the y-axis and the original base above the y-axis. The Change in also located above the y-axis. The asterisks location at the y-axis indicate the fraction of falling off peaks that are explained by the modification.

Our goal was to use this data of falling off for predicting of modifications in unannotated tRNAs. We first needed to confirm that the falling off corresponds to modification, for this we checked the mature tRNAs with known modifications. When comparing the

number of modification around a falling off peak to the total amount of modification, after establishing a threshold for "noise" via trial and error of around 0.36, we got 40%. In other words, 40% of falling-off peaks above 0.36 are explained by modifications.

In order to improve that we aligned the reads again, this time separately for CCA ending tRNA and tRNA without CCA. This we hoped will separate the mature reads from the non-mature reads (that has not been added a CCA). The non-mature tRNA will not have modification, and thus will not fall off at the predicted modification position. This allowed us to improve the percentage of falling off peaks explained by modification out of all modifications to 80%, when using a threshold of 0.36.

When we took a closer look, and checked each of the modifications in the same way, we found that only a few modifications (K, R and in some cases '6') are responsible for this prediction (of 80% mentioned above) (see table 4). Some modifications (like O, Y and ')') do reach high percentage of correspondence with falling off but are appearing very few time rendering them unreliable.

Table 4-location, total number and fraction of hits for each modification type in yeast

Modification type	Modification Full name	Location on tRNA-(number at location)	Total number	Fraction of hits out of all modifications
K	M1G	9-(8),36-(1),37-(1),38-(4)	14	0.714286
D	Dihydrouridine	16-(25),17-(8),19-(13),20-(16),21-(8),22-(1),46-(4),47-(6),48-(4),49-(1)	86	0.011628
R	m22G	25-(2),26-(10),27-(5),28-(1)	18	1
I	Inosine	34-(4),35-(2)	6	0.333333
O	m1I	37-(1)	1	1
P	Pseudouridine	1-(2),13-(7),14-(1),25-(1),26-(2),27-(7),28-(5),31-(4),32-(3),33-(5),37-(3),38-(4),39-(7),40-(4),41-(1),52-(1),54-(7),55-(10),56-(4),57-(1),64-(5),66-(1),67-(1)	86	0
L	m2G	9-(1),10-(18),26-(1)	20	0.15
?	m5C	35-(1),40-(2),46-(1),47-(2),48-(6),49-(7),50-(3),57-(4),59-(1)	27	0
"	m1A	57-(5),58-(8),59-(4),60-(1),67-(1),69-(1)	20	0
1	mcm5U	33-(1)	1	0
6	t6A	36-(2),37-(4),38-(2)	8	0.375
+	i6A	36-(1),37-(2),39-(1)	4	0
7	m7G	45-(4),46-(6),47-(1)	11	0
3	mcm5s2U	34-(1)	1	0
B	Cm	4-(2),31-(1),32-(2),33-(2)	7	0.142857
.	Unknown nucleotide	34-(1)	1	0
:	Am	5-(1)	1	0
#	Gm	17-(6),18-(2),34-(2)	10	0.1
M	ac4C	12-(5)	5	0
)	cmnm5Um	34-(1)	1	1
Y	yW	37-(2)	2	1
N	?U	33-(1),34-(1)	2	0
J	Um	44-(2)	2	0
<	?C	32-(1)	1	0
'	m3C	32-(1)	1	0
&	ncm5U	35-(1)	1	0
^	Ar(p)	63-(1)	1	0

For each modification in the known tRNA this table lists: its positions and numbers (POS-(NUM)) above the threshold of 0.36; the total number of the modification; and the number of hits, a peak above the threshold, divided by the total number.

We used the most predictive and reliable modifications that also had a specific location on the tRNA, in order to try and predict their whereabouts in the unknown tRNAs (those without mapped modifications) and the known tRNA, as a control (see figure 17). The false positive percentage, the percentage of wrongly identifying a modification, was 41% for 'K', 10% for 'R' and 50% for '6'. It is important to notice that the prediction for '6' and 'K' are sometimes overlapping thus increasing this statistic. The false negative

percentage, the percentage of missing a modification, was 23% for 'K', 0% for 'R' and 50% for '6'. It needs to be mentioned that most of those were in positions that the prediction wasn't able to single this modification, and in other cases there wasn't a peak above the threshold in that position.

In the unknown tRNA a total of 2 of '6', 7 of 'K' and 9 of 'R' modifications were found.

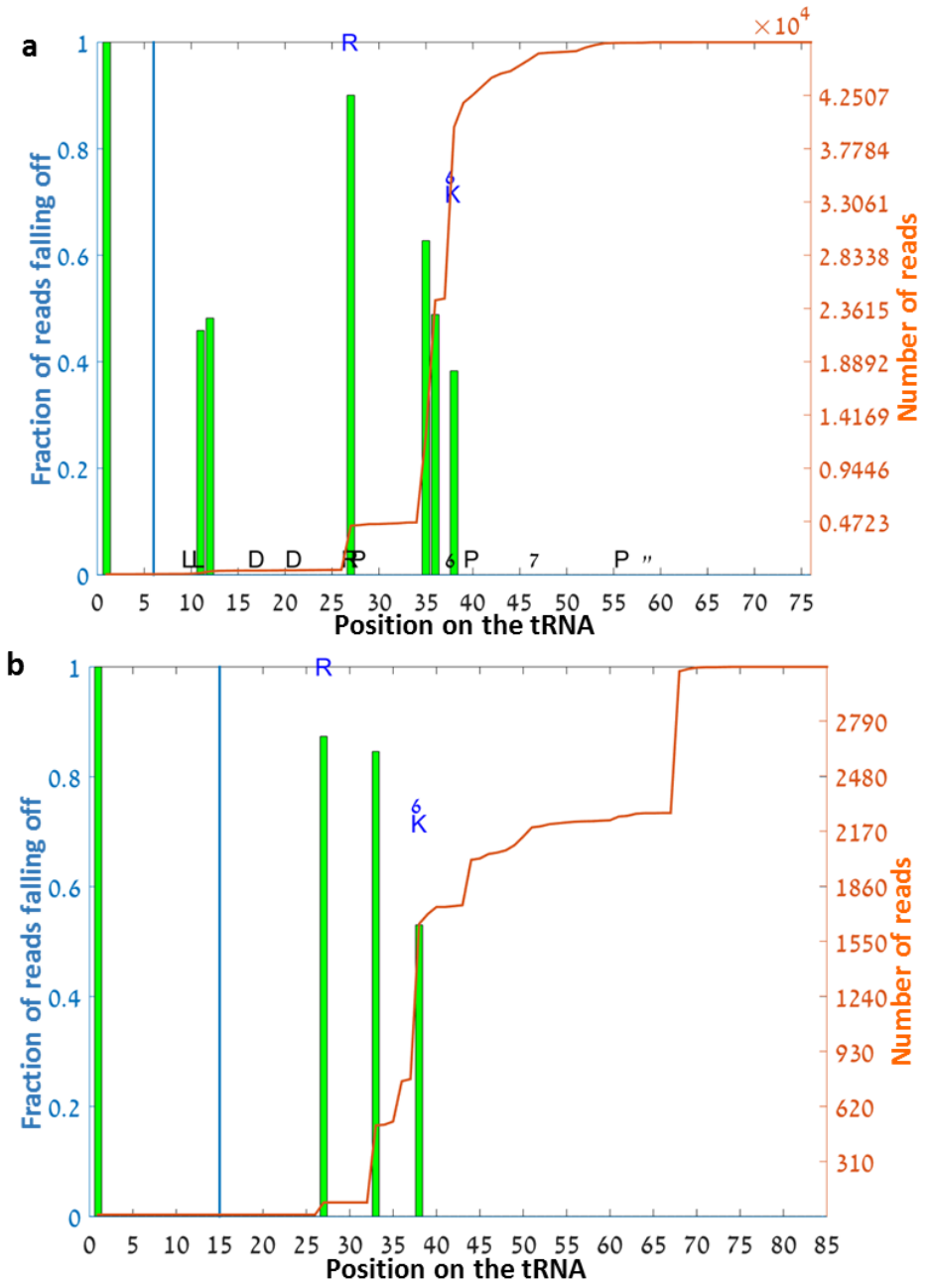

Figure 17- Prediction of modification along with falling off peaks above threshold. The bars are falling off chance above the threshold (0.36). The blue line indicate the length of the sequencing assuming the read started at the 3' end of the tRNA. The orange plot shows the number of remaining reads, The known modification (if known) are shown in black at the bottom of the y-axis and the prediction is written in blue and the y-axis position indicate the probability of correct identification. **(a)** shows the annotated Leu-CAA-1-1 yeast1, **(b)** shows the unannotated Ser-GCT-1-1 yeast1

4. Discussion

4.1. tRNA count

The Advanced tRNA-seq method is relatively short in time (about 3 days), reproducible, uses standard enzymes and equipment, and is able to detect all tRNA species, although with different efficacy. The method is also applicable to all organisms as it based on RNA sequencing, and we demonstrated it on yeast and mouse cells (not shown here).

The yield of the method can be significant improved, as now the yield of the preparation and cleaning steps until the first ligation (methods 3.4.-3.6.3.) is around 5% and the yield of the library preparation (methods 3.6.4.-3.6.10.) have a 0.1% yield. Although, after this step there is a PCR enrichment step, but this could reduce the complexity of the library and introduce biases.

The Advanced tRNA-seq method is not highly correlative to gene copy number which is the gold standard in tRNA expression in yeast (figure 7). The method has a better correlation to the microarray method, as seen in figure 8, but still some tRNA types seem to be highly enriched while others are under-representative, skewing the results. The DM-tRNA-seq method (13) reported poor correlation (R^2=0.2) to the gene copy number (figure 18), and the same trend (number not given) when comparing to tRNA microarray.

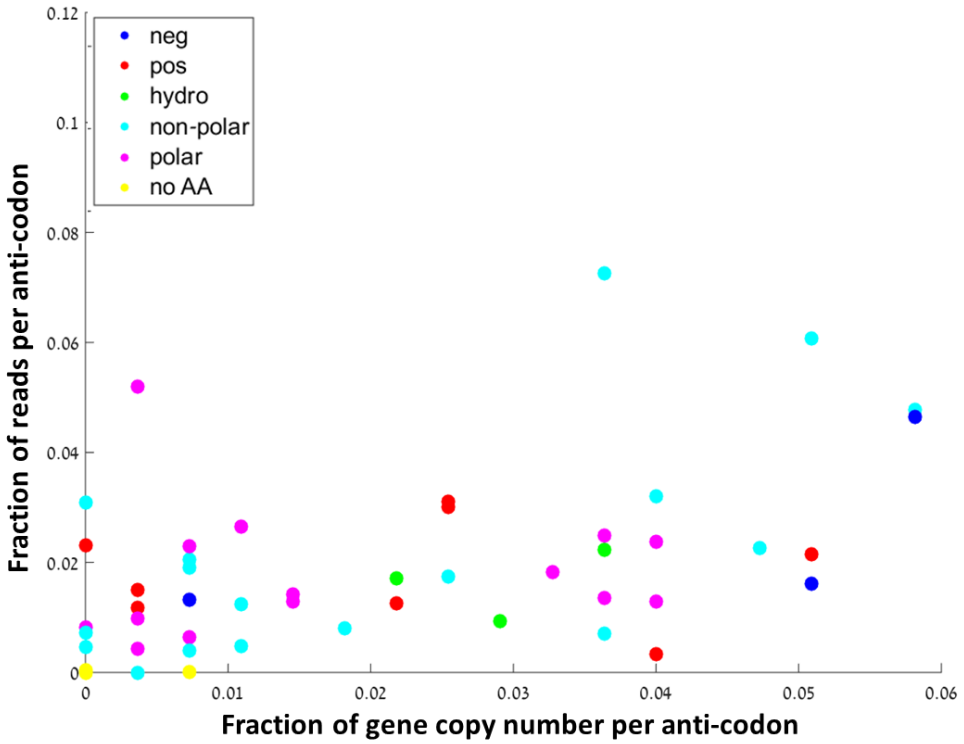

Figure 18- Scatter plot of fraction of anti-codons separated tRNA against fraction of gene copy number of tRNA in mapped reads of published DM-tRNA-seq method (13). This is a sample which was treated with special enzymes to remove certain modifications. The x-axis is the fraction of gene copy number and the y-axis is the fraction of the reads of the published DM-tRNA-seq method. The colors of the anti-codons are according to their amino-acid group. This graph includes also undefined anti-codons and stop codons tRNA genes. The Pearson R^2 correlation of the reads separated to anti-codons and the gene copy number is only around 0.2.

The published tRNA seq method (14) data processed for the same format we used (in figure 7) for better comparison and was compared to gene copy number (figure 19), that comparison wasn't available in the paper. The Pearson correlation we get in the Advanced tRNA-seq method (R^2=0.4) is much better than this paper's (R^2=0.1), which indicate an improvement in the accuracy on our part .The published tRNA-seq method (figure 19) also, seems to have highly enriched tRNA species, but they had enriched Gly-GCC, Gly-UCC, His-GUG, Ser-AGA and Cys-GCA, while we had Gly-GCC, Leu-CAA, Glu-TTC, Glu-CTC and Asp-GTC. So it seems that the root cause of this is shared by those protocols.

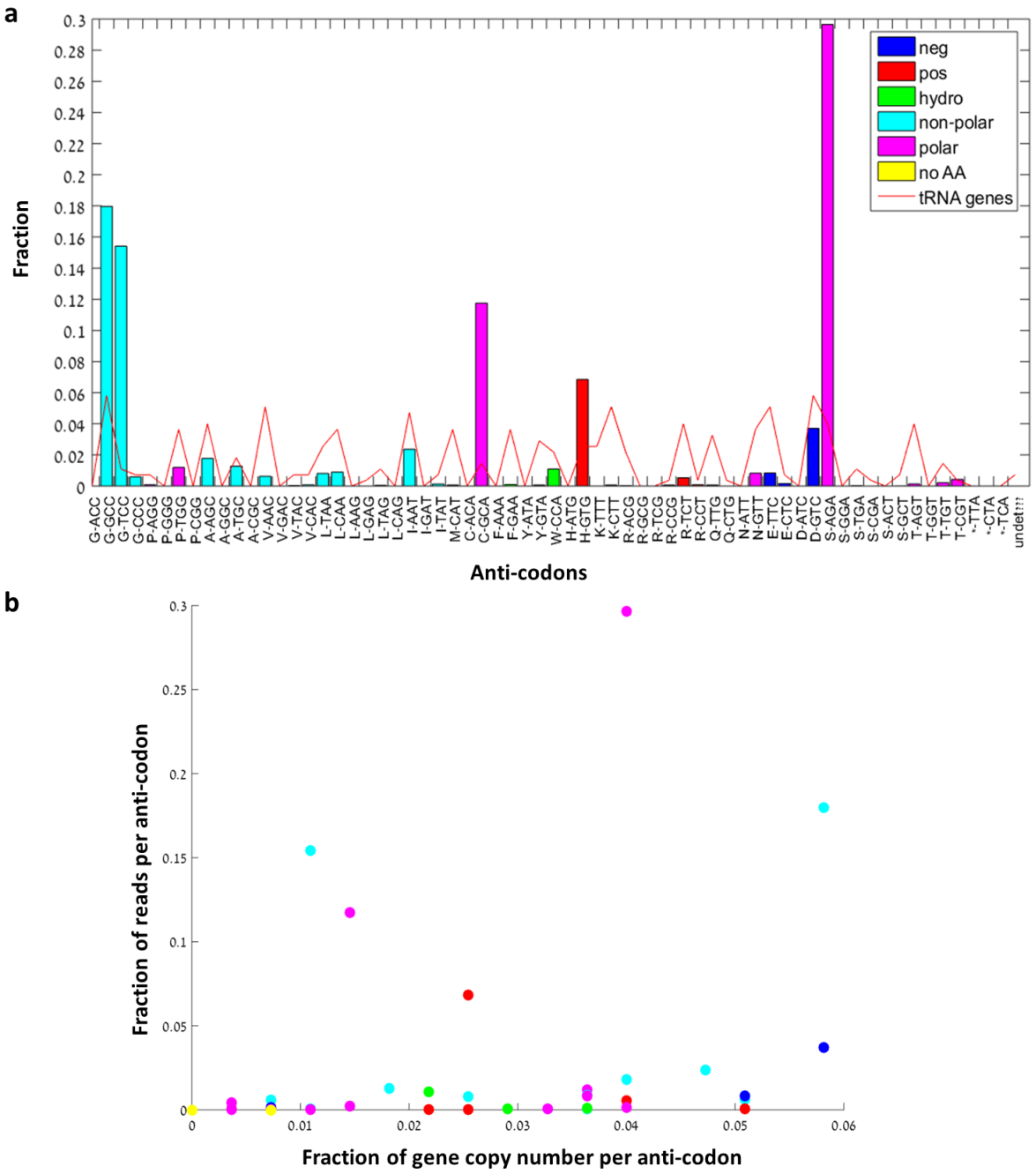

Figure 19- fraction of anti-codons separated tRNA against fraction of gene copy number of tRNA in mapped reads of published tRNA-seq method (14). This is an average of 6 samples of untreated yeast, HPLC purified tRNA that ran on HiSeq 2000 sequencing machine. **(a)** The bars indicate the fraction of anti-codon from the total mapped reads. **(b)** Scatter of fraction of anti-codons separated tRNA against fraction of gene copy number of tRNA.

The colors of the anti-codons are according to their amino-acid group. The red plot shows the fraction of the specific genes among the tRNA genes. This graph includes also undefined anti-codons and stop codons tRNA genes. The Pearson R^2 correlation of the reads separated to anti-codons and the gene copy number is around 0.1.

Interestingly when we grouped the different enriched tRNAs based on the chemical properties of their respective amino acid we found that the tRNAs with negatively charged AA or non-polar AA are more highly abundant in the reads (figure 7). This surprising finding may partly account for the medium correlation with gene copy number and microarray results. This bias can occur from different efficiency in removing AA causing different ligation efficiency of the 3' adaptor between various tRNAs. This can also be explained by the fact that as different groups of tRNAs have different modification pattern, there can be modification that cause falling off of RT before a length which is sufficient for alignment, causing that read to be discarded. Another explanation is different polymerase efficiency. Yet another hypothesis we had was that the SPRI beads are adhering to the tRNA in different affinity because of the AA, we disproved this by changing the order of the tRNA-seq protocol instead of removing AA after separation by SPRI beads to the other way around with no difference.

While during my project I was able to establish a working protocol for tRNA deep-sequencing that overcome some of the problems associated with these molecules, there is still room for improvement. The additional steps that can be added to the protocol in the future in order to improve our ability to accurately measure tRNA levels are:

- Use UMI's (Unique Molecular Identifier's) which can help normalize the reads to biases originating from different polymerase efficiency during the PCR amplification step of library construction.
- Change RT enzymes in order to overcome some modification causing falling off.
- Add enzymes that remove some of the modification as was done in this paper (13).
- Use primers for all known tRNAs sequences and gene base prediction of tRNAs for the RT step in order to bypass the ligation to the 3' end which contain the AA.
- Use RT primer for the CCA, this will allow to separate the mature tRNA from the non-mature which can skew the data for some tRNAs.
- Separate the tRNA in the lab to AA charged and not charged in order to learn about the role of the AA in the low correlation and also in order to learn about the tRNA life-cycle.

There is also room to improve the percentage of tRNA molecules out of the entire RNA pool that is sequenced which, currently, reaches around 30%. The reason for this is mainly due to having too broad size selection that let other medium size RNA get sequenced. More than half of the un-aligned reads were mapped to genes that are smaller than 200bp and from those 80% were snRNA genes.
This can be done by refining the SPRI beads separation of tRNA, or by using gel extraction in order to remove the reads which are medium size. Yet another way, that was mention in other context, is to reverse-transcribe only tRNA by either using primer for CCA or use primers for all the known tRNAs.

Another reason for the low alignment percentage may lay in the alignment software not able to detect all the tRNA reads. If this is caused by having too many mutations, caused maybe by modifications, we can address that by finding the tRNAs which slip through the cracks of the alignment software by creating a reference with expected mutations, or lowering threshold on sequence identity.

When comparing the alignment percentage to the other methods, we see that in the tRNA-seq method (14) the total small RNA (<200nt) had between 17-23% mapped reads while with those that had been purified by gel electrophoresis followed by HPLC, the alignment rose to around 70%. In the DM-tRNA-seq method (13) the alignment percentage after purifying with gel electrophoresis was between 75 to 82%.

The length analysis (figure 9) shows that before alignment the majority of reads are the full length but after the alignment this change dramatically and the reads are spread more evenly in sizes. This is to be expected as the reads before alignment contain a very high percentage of non-tRNA RNAs which are subject to fewer modifications, causing it to reach full length and the aligned reads are tRNAs which have modified bases that can cause RT fall-off. When comparing this result to the published tRNA-seq method (14, figure 2 therein), we see that when examining the read length distribution in bins of 10nt of length we get that before alignment they had between 0.05-0.25% of reads in each bin and that the bins of 41-50nt and 51-60nt had the lowest percentage, while after alignment the picture changes and most of the reads (50%) are at full length (61-70), the 21-30 and 31-40nt bins have 18% and 41-50 and 51-60 have around 10%. The lack of reads in the

lower lengths (1-20) after alignment is due to actively screening for those as they perceive them to be too short for accurate tRNA detection. The difference from our method in length distribution before the alignment can be attributed to them using small RNA (<200nt) while we used medium size RNA (50-200nt) causing them to have more small reads derived from non-tRNA while in our method the smallest length detected (15nt) correspond to the first modification in most tRNAs. The reason for the discrepancy between the methods in the length after the alignment can be derived from different RT used (Primescript against AffintyScript), or different alignment software and screening parameters.

4.2. Analyzing modifications

The coverage analysis (figure 11) shows that some positions cause a falling off of the RT. The coverage also shows that most of the mature tRNAs (with CCA) are of type 2 which means that most of them don't reach full length sequence. This is an indication that the modified bases can interfere with the RT passage in most of the tRNA species. There is also some tRNA species that show an increase at coverage after a certain position (type 3). The likely reasons for it are a fragmented tRNA which got into the library; an alignment problem in some areas; another RNA that have similar sequence; or because there is a non-mature version of the tRNA in the reference for alignment which aligns better and thus causing the 3' end position to seem to have low coverage. Some tRNAs coverage map show an increase after the red vertical line at the right of the figure, which indicates the proper end of the read, meaning the position that the read would have ended if read from the adaptor at the CCA. This can result from a tRNA without a CCA, or a tRNA which was fragmented during the library preparation or, of course, an alignment error.

The falling off graph (figures 13) show that some modification are likely to cause falling off of the RT, while other do not. There is also a difference in the percentage of falling off (the height of the peak) in different modification and sometimes across different tRNAs. Also, some peaks of fall off don't correspond to a known modification. It need to

be considered that the tRNA pool consists of mature and non-mature tRNA, this can skew the effect of the modification. The comparison between a certain modification in different tRNAs can possibly inform us on the abundance of non-mature tRNA among them (this can also be inferred from the nucleotide composition when comparing the nucleotide change in a position that contain a modification that cause a nucleotide change consistently, then it can be assumed that the percentage that didn't change are non-mature tRNAs). The peaks can also be explained by secondary structure, this can cause difficulty for the transition of the RT and causing it to fall. While the treatment that had been performed was supposed to minimize it, we can't dismiss this possibility. Another explanations for the peaks are unmapped modifications, or low processivity of the RT.

It is interesting to notice that the falling off of the RT (figure 14) can have different positions relative to the location of the modification, in some cases this happens before the position of the modification, indicates falling off when encountering the modification, while in other this happens in the position itself, and in some others falling off occurs right after the modification.

The nucleotide composition at different positions of the known tRNA (figure 15) shows us that there are several modifications that cause some kind of change in the nucleotide composition, while most of the modifications don't influence the RT in this regard (figure 16). The nucleotide changing modifications causes a very distinct change, different for each one. It is also interesting to notice that none of the modifications shows a mixture of nucleotides, as would be expected if the RT would have inserted a random base.

The most prominent changes are Am which change A to T in 100% of cases, Inosine which change A to G in more than 90% of cases, and m1G which change G to C in more than 80% of cases. In those cases the un-mutated sequences can stem from non-mature tRNA. This change in Inosine is known from previous studies (15), while m1G and Am are not known.

In the modified bases m1I and m22G we also observe some nucleotide changes, in m1I around 25% from A to G and in m22G around 5% from G to A. This can occur from

having a small but reasonable chance of change in base, or from the modification presence in only this small fraction of the tRNA species.

As we saw the modifications can explain 80% of the falling off peaks at high enough peaks. This is while only around 22% of modifications are at a peak. This makes sense as not all the modifications cause falling off. When we examine table 4 we see that there are several modifications that have a high probability to predict a falling off. Some of them seem to cause falling off but at lower rates, so they are not counted in this analysis. These findings are in agreement with other studies (16). It is interesting to notice that some modifications have specific locations along the molecule, as is known in the literature (17).

The prediction of the modification in the unmapped tRNA is based on the position and chance to fall off above the threshold. These predictions are currently limited to only three modifications and sometimes it can't distinguish between them. A future improvement in the predication power can be obtained by including the information we obtained regarding nucleotide changes at modified sites, reducing the cut-off used, such as to include the lower falling off peaks. Another way to increase its range is to add a larger dataset of known modification, this will allow to infer for more modifications their response. Another way to improve the data is, as was mentioned before, to try and have only mature sequences, this will reduce the noise.

There are several approaches to validate the modifications predictions: use enzymes that remove a modification or use a strain without some modification gene, and compare the data; use other RNA with known modification, hopefully without a secondary structure (like mRNA), in order to increase the sample size and also to remove concern for secondary structure causing falling off. Another approach is to use MS (mass spectroscopy) methods that can deduce the modification as was done to most of the known modification in tRNA.

In this project we demonstrate the ability of using RNA-seq based method to detect all tRNA types by capturing even tRNA fragments. We then demonstrate the reproducibility and reasonable similarity to other methods. The method is used to characterize the ability

of different modified bases to cause RT falling off or mis-incorporation of nucleotides. We manage to try and predict the location and type of modification in other unknown tRNAs.

5. Literature

1. Bloom-Ackermann, Z., Navon, S., Gingold, H., Towers, R., Pilpel, Y., & Dahan, O. (2014). A Comprehensive tRNA Deletion Library Unravels the Genetic Architecture of the tRNA Pool. PLoS Genet, 10(1), e1004084.
2. Dittmar, K. A., Goodenbour, J. M., & Pan, T. (2006). Tissue-specific differences in human transfer RNA expression. PLoS genetics, 2(12), e221.
3. Kanaya, S., Yamada, Y., Kudo, Y., & Ikemura, T. (1999). Studies of codon usage and tRNA genes of 18 unicellular organisms and quantification of *Bacillus subtilis* tRNAs: gene expression level and species-specific diversity of codon usage based on multivariate analysis. Gene, 238(1), 143-155.
4. Zhou, Y., Goodenbour, J. M., Godley, L. A., Wickrema, A., & Pan, T. (2009). High levels of tRNA abundance and alteration of tRNA charging by bortezomib in multiple myeloma. Biochemical and biophysical research communications, 385(2), 160-164.
5. Percudani, R., Pavesi, A., & Ottonello, S. (1997). Transfer RNA gene redundancy and translational selection in Saccharomyces cerevisiae. Journal of Molecular Biology, 268(2), 322-330.
6. Tuller, T., Carmi, A., Vestsigain, K., Navon, S., Dorfan, Y., Zaborske, J., Pan, T., Dahan, O., Furman, I., & Pilpel, Y. (2010). An evolutionarily conserved mechanism for controlling the efficiency of protein translation. Cell, 141(2), 344-54.
7. Bloom-Ackermann, Z., Navon, S., Gingold, H., Towers, R., Pilpel, Y., & Dahan, O. (2014). A Comprehensive tRNA Deletion Library Unravels the Genetic Architecture of the tRNA Pool. PLoS Genet 10(1): e1004084.
8. Wang, Z., Gerstein, M., & Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. Nature Reviews Genetics, 10(1), 57-63.
9. Gustilo,E.M, Vendeix,F.A.P., & Agris, P.F.(2008). tRNA's modifications bring order to gene expression. Current Opinion in Microbiology, 11(2), 134-140.
10. Phizicky, E. M. (2005). Have tRNA, will travel. PNAS, 102(32), 11127-11128.
11. Phizicky, E. M., & Alfonzo, J. D. (2010). Do all modifications benefit all tRNAs?. FEBS letters, 584(2), 265-271.

12. Jühling, F., Mörl, M., Hartmann, R.K., Sprinzl, M., Stadler, P.F., & Pütz, J. (2009). tRNAdb 2009: compilation of tRNA sequences and tRNA genes. NAR, 37 (suppl 1), D159-D162
13. Zheng, G., Qin, Y., Clark, W.C., Dai, Q., Yi, C., He, C., Lambowitz, A.M. & Pan, T. (2015). Efficient and quantitative high-throughput tRNA sequencing. Nature Methods 12, 835–837.
14. Pang, Y.L., Abo, R., Levine, S.S., Dedon, P.C. (2014). Diverse cell stresses induce unique patterns of tRNA up- and down-regulation: tRNA-seq for quantifying changes in tRNA copy number. NAR, 42(22):e170.
15. Findeiß, S., Langenberger, D., Stadler, P.F., & Hoffmann, S. (2011). Traces of post-transcriptional RNA modifications in deep sequencing data. Biological chemistry 392, 305-313.
16. Motorin, Y.,Muller, S.,Behm-Ansmant,I., & Branlant, C. (2007). Identification of Modified Residues in RNAs by Reverse Transcription-Based Methods. Methods in Enzymology 425, 21-53.
17. Jackman, J.E., Montange, R.K., Malik, H.S., & Phizicky, E.M. (2003). Identification of the yeast gene encoding the tRNA m1G methyltransferase responsible for modification at position 9. RNA 9, 574-585

6. Acknowledgements

I would like firstly to thank my two advisors: Dr. Ido Amit and Prof. Yitzhak (Tzachi) Pilpel. Ido was very supportive, shared a lot of his experience and helped me think like a scientist. Tzachi was really helpful, full of great ideas and gave me a lot of motivation.

Dr. Orna Dahan, also contributed a great deal to this project, be it with help in the wet lab or with the analysis.

I couldn't made the sequencing analysis without the help of Eyal David who assisted me whenever needed.

From the Amit lab members I would like to thank all them for being great people and scientists, they gave support whenever asked for. I would especially want to thank Dr. Hadas Keren-Shaul, Chamutal Borenstein, Dr. Ronnie Blecher-Gonen, and David Lara that helped me with different protocols and equipment and Liran Valadarsky and Dr. Erika Lorenzo for good advices.

The Pilpel lab members, has an important part in this project success as they aided me whenever needed and made each day in the lab fun. I had great many talks and discussions with them that got my cogs moving. I would like to thanks especially to Dr. Ruth Towers that in addition to helping in some protocols gave great logistic support. Dr. Hila Gingold also shared with me her experience in analysis data and gave good advises.

I would also like to thank Yoach Rais from Yacob Hanna lab for his help with growing mouse ES cells.

7. Abbreviations

RT- reverse transcriptase/reverse transcription.

Modified bases

.- unknown nucleotide

"- (m1A) 1-methyladenosine

+- (i6A) N6-isopentenyladenosine

6- (t6A) N6-threonylcarbamoyladenosine

:- (Am) 2'-O-methyladenosine

I- (I) inosine

O- (m1I) 1-methylinosine

^- (Ar(p)) 2'-O-ribosyladenosine (phosphate)

<- (?C) unknown modified cytidine

B- (Cm) 2'-O-methylcytidine

M- (ac4C) N4-acetylcytidine

?- (m5C) 5-methylcytidine

'- (m3C) 3-methylcytidine

K- (m1G) 1-methylguanosine

L- (m2G) N2-methylguanosine

#- (Gm) 2'-O-methylguanosine

R- (m22G) N2,N2-dimethylguanosine

7- (m7G) 7-methylguanosine

Y- (yW) wybutosine

N- (?U) unknown modified uridine

J- (Um) 2'-O-methyluridine

&- (ncm5U) 5-carbamoylmethyluridine

1- (mcm5U) 5-methoxycarbonylmethyluridine

3- (mcm5s2U) 5-methoxycarbonylmethyl-2-thiouridine

)- (cmnm5Um) 5-carboxymethylaminomethyl-2'-O-methyluridine

D- (D) dihydrouridine

P- (psi) pseudouridine