UNIVERSITY OF CALIFORNIA
SANTA CRUZ

# DIRECT RNA SEQUENCING OF *E. coli* INITIATOR tRNA USING THE MinION SEQUENCING PLATFORM

A thesis submitted in partial satisfaction of the
requirements for the degree of

Master of Science

in

BIOMOLECULAR ENGINEERING & BIOINFORMATICS

by

**Vinay C. Poodari**

June 2019

The Thesis of Vinay C. Poodari
is approved:

Professor David Bernick, Chair

Professor Mark Akeson

Professor Rebecca DuBois

Lori Keltzer
Vice Provost and Dean of Graduate Studies

# Table of Contents

# List of Figures

# List of Tables

**Abstract**

Direct RNA Sequencing of *E. coli* initiator tRNA Using the MinION Sequencing Platform

by

Vinay C. Poodari

Transfer RNAs (tRNAs) are small RNA molecules responsible for decoding the genetic code into an amino acid sequence. tRNAs bring amino acids to the ribosome, playing an active role in translation. tRNAs are vital to living systems and can impact human health, making them clinically relevant. This motivates the development of high throughput methods to screen and sequence tRNAs. Nanopore sequencing, specifically the commercialized MinION sequencing platform presents an opportunity to directly sequence tRNAs. I have developed a framework for direct RNA sequencing of both biological and synthetic canonical *E. coli* initiator tRNA that will be expanded to other tRNAs. I discuss the methods developed to sequence tRNAs on the MinION. This includes improving coverage on the terminal ends of the tRNAs, in vitro production of tRNAs, and the detection of modified nucleosides using the MinION platform. The work presented here is intended to be a launching point for further sequencing of tRNAs and modification discrimination on the MinION. This work culminates in presenting distinct differences between the alignment profiles of biological *E. coli* fmet tRNA and a canonical version due to the presence of modified nucleosides on the biological tRNA. This has implications for both direct sequencing of tRNAs and the broader goal of detecting modified nucleosides.

To all my friends.

Their support and encouragement means the world to me.

## Acknowledgments

This project would not have been possible without the help of many of my lab members. I would like to thank Dr. Robin Abu-Shumays who assisted me with lab work, mass spectrometry, and helped me develop a set of skills I didn't think I would have a chance to learn. I also want to thank Dr. Miten Jain, a great mentor, he is always willing to listen, provide scientific input, and crack a few jokes. I thank Logan Mulroney for taking time out his busy days to help me edit my writing and his wealth of information on all things scientific. I thank Dr. Hugh Olsen for his words of encouragement and his sense of humor that always brightened up a long day in the lab.

Finally I would like to thank Dr. David Bernick and Dr. Mark Akeson. Since my undergraduate career, Dr. Bernick provided support for my research interests and introduced me to this project. A special thanks to Dr. Akeson, for his advice and his full support for this work.

# Chapter 1

# Introduction

tRNAs, are essential to the central dogma of biology and are one of the most abundant nucleic acids in cells [13]. During translation, tRNAs form codon-anticodon pairs with mRNA in the ribosomal complex. tRNAs also bring with them the corresponding amino acids encoded by the mRNA strand to initiate or extend the nascent polypeptide chain. While tRNAs are small in sequence length, they are very complex. tRNAs have a strong secondary structure, cloverleaf in shape, made up of 4 loop regions [12]. It is estimated that there are 595 tRNA genes in the human genome and in simpler prokaryotes such as *E. coli*, 89 tRNA genes, both more than the 64 possible codon-anticodon pairs in the translation paradigm [23]. The narrow distribution of tRNA sizes and sequence conservation have lead researchers to believe that primordial tRNAs were the basis of translation with the modern ribosome appearing later in evolution [12].

The complexity of tRNAs is not limited to their abundance or their role in translation; tRNAs are also heavily modified. There are more than 70 different known modifications on tRNA molecules [6]. These modifications are necessary for accurate translation, maintaining tRNA structure and charging tRNA with the proper amino acid [6]. Improper modification states can lead to a variety

of disease phenotypes often in mitochondria [1]. In eukaryotes the majority of tRNAs are nuclear encoded, however mitochondria also have a small set of tRNAs (mt-tRNAs) encoded in their genomes. The number of mt-tRNAs varies among species with humans having 22 mt-tRNAs, but others less relying on nuclear encoded tRNAs [16]. In humans mutations in mt-tRNA genes can have impacts on health and fitness. Many ailments such as cardiomyopathy, aural and visual impairments, and cancers are due to sequence mutations in mt-tRNA[1]. The vital role of tRNAs and their impacts on human health necessitates methods to better screen and categorize tRNAs. These screening methods need to be scalable, low cost, and require rudimentary lab skills for accessibility in various clinical environments.

There are multiple methods for detecting and categorizing tRNAs, the most commonly used are mass spectrometry, hybridization arrays, and next generation sequencing, NGS [23]. These methods have shown promise, but suffer from scalability or information loss. Mass spectrometry identifies tRNA by digesting the tRNA into either fragments or nucleosides. The mass spectra is then used to separate species of tRNA from one another [23]. Mass spectrometry is very sensitive however sequence determination of the tRNA by mass spectrometry is difficult. Mass spectrometry is also not practical in all clinical settings. The machines require a large area of space and the technical skills needed to run experiments and interpret data requires additional training. Other methods such as oligo arrays rely on tRNA molecules to anneal to probes. Annealing can be inhibited by strong secondary structure and modifications on the tRNA, both prevent binding to oligo probes [7]. In addition arrays need to be very specific as the difference between two tRNAs can be a single base and without the primary sequence it is impossible to separate two similar tRNA [23]. Next Generation Sequencing techniques are

the logical choice and many methods to sequence RNA have been developed. One of the most popular methods is sequencing by synthesis i.e. Illumina. Sequencing by synthesis requires generating cDNA copies of the tRNA and sequencing the cDNA. Sequencing by synthesis has many problems as the reverse transcription step can be stymied by the secondary structure of the tRNA and modifications can halt the enzyme, leading to inefficient full length cDNA production [24]. Multiple methods to circumvent these problems have been developed. This includes removing methylation modifications or using more robust reverse transcriptase enzymes, such as TGIRT, to increase the yield of full length cDNA [23, 24]. The problem with sequencing by synthesis is that the original tRNA is not sequenced. Information regarding the modification states of nucleosides is not represented in the cDNA. This information is lost. The modification information could differentiate tRNAs, and more importantly directly recognizing modification presence can be important in recognizing diseases.

I show that MinION sequencing is a viable method to sequence tRNAs directly. Nanopore sequencing has been developed over the last decade for single molecule sequencing of DNA. The system uses an applied voltage to thread nucleic acid fragments through a single biological nanopore embedded in a lipid membrane. As the nucleic acid translocates through the pore, there is a measurable change in current creating blockade events [8]. An enzyme such as a helicase or polymerase regulates the rate that nucleic acids move through the pore so measurements can be made as each nucleotide enters the pore. The measured current can then be segmented into discrete events or states and converted into a nucleotide sequence [8]. There is a variety of research focusing on the applications of nanopore sequencing including detection of nucleoside modifications which has shown significant promise [8, 17, 19].

The UCSC Nanopore group has translocated tRNA through a nanopore channel [18]. These experiments used a single $\alpha$-hemolysin pore. These experiments did not produce sequence level data as the $\alpha$-hemolysin pore used reads 10-15 nucleotides at a time, making discrimination of each nucleotide difficult and beyond the scope of the work [3, 11]. Smith et al. showed the blockade events of *E. coli* initiator (fmet) tRNA translocating through a nanopore are different than those of lysine tRNA [18]. This work showed that tRNAs can be differentiated by the current signal. While the single channel experiments did not provide sequencing data, translocation and differentiation of tRNAs on a nanopore based system was validated. Nanopore sequencing has been commercialized by Oxford Nanopore Technologies (ONT). Their flagship device, the MinION, uses 2048 pores to directly sequence nucleic acids providing base level resolution. The MinION platform was developed for long read DNA sequencing and can directly sequence RNA molecules. The UCSC Nanopore group adapted the methods of tRNA capture and translocation developed by Smith et al. to the MinION. This was done by producing a set of splint adapters that ligate to the NCCA overhang on the $3'$ end of the tRNA. The splint adapters also ligate to the conventional sequencing adapters developed by ONT. The Nanopore group sequenced synthetic bovine tRNA constructs on the MinION. These constructs were comprised of only canonical bases, however the group was able to demonstrate that these constructs can translocate through the MinION nanopore, and are basecalled into nucleotide sequences. The initial goal was to produce a set of synthetic molecules and train a classifier to classify bovine mt-tRNAs, based on sequence. There were unanticipated complications with T7 transcribed tRNAs for nanopore sequencing. This resulted in constraining the problem to sequencing biological and a synthetic canonical *E. coli* initiator tRNA and determine if reported modifications could be detected on

the MinION. To better sequence these two tRNA molecules, the set of adapters originally used by the Nanopore group were modified resulting in better $3'$ and $5'$ coverage. I show that there is a significant difference between the biological and synthetic tRNA alignments due to the presence of modified nucleosides which has been confirmed with mass spectrometry testing.

# Chapter 2

# Methods

The methods section details materials, oligos, and general protocols used. The reference sequences for designing tRNA constructs are available on Modomics [2]. The bioinformatic tools used to produce and visualize alignments are also discussed.

**Oligos:**

In vitro transcription (IVT) template:

Template sequence:

TGGTAGTACGGGAAGGATATAAACCAACATTTTCGGGGTATGGGCCC
GATAGCTTAATTAGCTGACCTTACTAATAGTGAGTCGTATTAAATGAT
GATG

The DNA template used to produce the bovine initiator mt-tRNA was synthesized by Integrated DNA Technologies (IDT), rehydrated in NF $H_2O$ to produce a 100 µM stock and kept at -20°C. This template is based on the reference sequence available at Modomics [2]:

AGUAAGGUCAGCUAAUUAAGCUAUCGGGCCCAUACCCCGAAAAUGUU
GGUUUAUAUCCUUCCCGUACUACCA

T7 RNA polymerase will include an initial adenine in the transcript, the tRNA

encoded in this template also starts with an adenine. In order to avoid consecutive adenines at the start of the tRNA, the region encoding the tRNA in the template was adjusted to start encoding from the second position of the tRNA onwards.

In vitro transcription promoter:

CATCATCATTTAATACGACTCACTATTA

T7 RNA polymerase requires the T7 promoter region to be double stranded. This DNA oligo hybridizes to the template oligo to form the duplex promoter region. It was also synthesized by IDT, and kept at -20°C after being rehydrated in NF $H_2O$ to produce a 100 µM stock.

AS102 and M221:

The AS102 and M221 adapters were synthesized by IDT. They were gel purified by colleagues during their early attempts at sequencing tRNAs and stored at -80°C.

Modified AS102_18 and M221_6:

AS102 and M221 adapters were modified such that both adapters are DNA/RNA hybrid oligos. They are 5′ phosphorylated and ordered from IDT. The adapters were modified so that there are ribonucleotide regions flanking the tRNA on both ends. AS102_18 consists of 18 ribonucleotides flanking the tRNA at the 5′ side and M221_6 has 6 ribonucleotides adjacent to the 3′ overhang of the tRNA. M221_6 was purified using HPLC purification by IDT. Oligos were rehydrated in NF $H_2O$ to make a 100 µM stock. Adapter oligos were stored at -80°C.

*E. coli* biological fmet construct:

Biological *E. coli* fmet tRNA was recovered from storage. It was manufactured by Subriden RNA, no longer in business, and was previously used for functional tests. In older mass spectrometry data, data not shown, this tRNA showed the presence of modifications. This tRNA was gel purified and then aliquoted. The

biological tRNA was phosphorylated using PNK (NEB: M0201S) before being used for ligation and libraries. The purified tRNA was kept at -80°C.

*E. coli* synthetic fmet construct:

This phosphorylated oligo was synthesized by Dharmacon. The following sequence from Modomics was ordered [2]:

CGCGGGGUGGAGCAGCCUGGUAGCUCGUCGGGCUCAUAACCCGAAGG
UCGUCGGUUCAAAUCCGGCCCCCGCAACCA

It was rehydrated in NF $H_2O$, aliquoted, and stored at -80°C.

**Urea Gels:**

TBE urea gels were made in 75 mL batches with 7.5 mL 10x TBE buffer, 15 mL 40% bis-acrylamide, 37.7 grams of urea, and filled up to 75 mL with Milli-Q $H_2O$. The solution was mixed using a magnetic stirrer until clear and then 488 µL 10% APS and 48.8 µL TEMED were added to polymerize the mixture. While the solution was mixing, glass plates were washed with Milli-Q $H_2O$ and EtOH. The plates were then dried. The gel mix was poured between two glass plates and allowed to polymerize for 30-45 minutes. After polymerization the gel was prerun for at least 20 minutes at 25 watts while the samples were prepared in 2-9x sample volume of urea load (7M urea and 0.1x TBE) depending on the maximum capacity of the wells. After adding urea load, samples would be mixed by pipetting or flicking, then heated to 90°C for 5 minutes. In many gels NEB 2x RNA dye, B0363S, rather than urea load was used. The NEB 2x RNA dye contains formamide which resolved smaller fragment bands better. When using 2x RNA dye NEB instructions were followed. The ladder used in all the PAGE results shown is NEB low range ssRNA ladder (NEB: N0364S), unless otherwise specified. After loading samples, gels were run for 50-60 minutes at 23 watts. If the loading dyes, bromophenol blue and xylene cyanol, were running unevenly

across the gel a metal plate was placed on top of the glass plates to improve heat distribution. Gels were stained with 6x-8x SYBR gold stain in 50 mL 1X TBE solution. Staining was done in the dark for 15 minutes, before gels were visualized.

**In Vitro Transcription Reaction:**

In order to produce the bovine fmet tRNA at a low cost, a commercial T7 IVT kit was used. The kit used was the Hiscribe T7 Quick High Yield kit from NEB (E2050s). The template DNA oligo and promoter oligo from IDT were hybridized together, using a heat and slow anneal process. T7 requires the promoter region be double stranded. 33 pmols of the template was mixed with 165 pmols of the complementary promoter oligo and allowed to hybridize in TNE buffer. The total volume was 7 µL. Hybridization was done by heating the mixture to 75°C for 15 seconds and slowly cooling to 25°C at 1.6°C/minute. After hybridization, NEB's standard RNA synthesis protocol for the kit was followed. 1 µL RNasin (Promega: N2611) was added to prevent RNA degradation. The total reaction volume was kept at 20 µL. IVT reactions ran overnight, 16 hours minimum, at 37°C. The next morning 2 units DNAse I (NEB: M0303) and the supplied 10x reaction buffer was added raising the reaction volume to 30 µL with the addition of NF $H_2O$. DNAse I treatment took 15 minutes at 37°C. The next step is to extract and precipitate the tRNA produced.

**Phenol/Chloroform Extraction:**

The IVT reaction was prepared for a phenol/chloroform extraction by bringing up the volume of the reaction to 200 µL with NF $H_2O$. 20 µL of 3M sodium acetate was added to the reaction, followed by a 200 µL 1:1 mixture of acidic phenol, pH 4.3, to chloroform. DNA LoBind tubes are recommended for all work. The sample was vortexed and well mixed before being spun at 10,000 RCF for 5 minutes in a microcentrifuge. This centrifuge was kept at 4°C. The organic layer was removed

with gel loading tips, and 200 µL of chloroform was added. The sample was then vortexed again and spun at the same parameters. The chloroform wash and spin was repeated once more. After the last spin most of the organic layer was removed and the sample was spun for the last time to separate out the organic and aqueous layer. The aqueous layer was collected and placed in a new tube.

**Ethanol (EtOH) Precipitation:**

After the phenol/chloroform extraction, the volume of collected aqueous layer was measured and 2.5x volume of 100% EtOH was added to the sample. The reaction was then stored at -80°C until it could be purified, usually overnight. The mixture was taken out the freezer and spun down for 30 minutes at maximum RCF in a microtube centrifuge kept at 4°C. The supernatant would be removed after confirming a pellet was visible and 500 µL of 80% EtOH was added to the pellet. The tube was respun for 20 minutes and the supernatant removed. The same ethanol wash and spin was repeated. After the second wash the supernatant was removed and the pellet was air dried. The pellet was then rehydrated with NF $H_2O$ and RNA concentration was measured with a NanoDrop 1000 using the default ssRNA mode, aiming for A260/280 and A260/230 ratios of 2.0. If at any times a pellet was not visible the tube would be respun. Free NTPs may still be present after EtOH precipitation, affecting quantification. Bio-Rad P6 (7326221) columns were used to remove any excess NTPs after EtOH precipitation. The recommended protocol from Bio-Rad was followed. The P6 columns come packaged in buffer, which was replaced with NF $H_2O$ following the additional steps Bio-Rad recommends for buffer exchange. 40 µL of the precipitated IVT reaction was loaded onto a P6 spin column and RNA concentration was measured on a NanoDrop 1000.

**Preparing IVT Product for Ligation with RppH:**

One major limit of IVT products is additional processing is needed for efficient ligation to occur. IVT products may contain 3′ non-templated ends and have a 5′ triphosphate. Both result in inefficient ligation of the splint adapters. A loss was taken on molecules with non-templated nucleotides as those are expected to be the minority product. The 5′ end of the tRNA was made amenable for ligation by using RNA 5′ Pyrophosphohydrolase (RppH) purchased from NEB, catalog #M0356. The NEB decapping eukaryotic mRNA with RppH protocol was used with some modifications. The amount of IVT produced tRNA substrate was increased to 3 µg and the reaction was run for 6 hours with 25 units of enzyme. This reaction was then purified using a phenol/chloroform extraction and ethanol precipitation, similar to the methods above. 3M NaOAc was added twice, once before the addition of phenol and before adding EtOH with the belief that it would improve recovery. In addition 1 µL molecular biology grade glycogen at 20mg/ml (ThermoFisher/Fermantas) was added to help precipitate the tRNA, after adding EtOH. The extracted sample would be placed at -80°C and purified in 1 to 3 days. The EtOH precipitation used a similar protocol mentioned previously with the wash steps using 150-200 µL of 80% EtOH.

**Phosphorylation of *E. coli* fmet tRNA:**

The biological tRNA was phosphorylated using T4 PNK (NEB: M0201S). 400 pmols of tRNA was mixed with 5 µL of T4 DNA Ligase 10x reaction buffer and 20 units of enzyme. NF $H_2O$ was added to bring the reaction volume to 50 µL. The mix was heated to 37°C for 30 minutes. The tRNA is then purified using a phenol/chloroform extraction. The method above was used, however 3M NaOAc was added at the end. The amount of NaOAc added was 10% of the recovered volume. 1 µL molecular biology grade glycogen, 20mg/ml, was added to help precipitate the tRNA, before placing the sample at -80°C. EtOH precipitation used

two 200 µL washes of 80% and 100% EtOH. The sample was eluted in 11 µL of NF $H_2O$.

**Library Preparation:**

All libraries were run as similar to each other as possible, there are variances in SPRI cleanup and elution volumes for the following runs, the two biological runs on 11/14/18 and 10/17/18 used 4.4x SPRI wash as these libraries were made before experimenting with the SPRI cleanup procedure. The 02/04/19 run used the SQK-RNA002 kit, ONT phased out the older SQK-RNA001 kits, and had an elution volume of 21 µL.

**Splint Ligation:**

All libraries were prepared using ONT's SQK-RNA001 or SQK-RNA002 protocol, except with the following changes. The RTA was replaced with either AS102 and M221 or AS102_18 and M221_6 adapters. The adapters are prepared by hybridizing 100 picomol of each adapter in 10 µL volume containing NF $H_2O$ and 1X TNE, creating a 10 µM stock of hybridized adapter. Adapters were hybridized by heating the reaction up to 75°C for 15 seconds and slowly cooling to 25 °C at 1.6°C/minute. This process takes approximately 31 minutes.

10 pmols of tRNA and NF $H_2O$ are mixed and heated to 90°C for 20 seconds to denature the tRNA. This allows the adapters and tRNA to anneal. The tRNA is then allowed to cool for several minutes on ice. The next step is to add the following components: 8 pmols of hybridized adapter, 2 µL PEG 8K, 2 µL 10x T4RNL2 reaction buffer, 1 mM ATP, 6.25 mM DTT, and 6.25 mM MgCl2. 5 units T4 RNA Ligase 2 (NEB M0239S) are added last, bringing the total volume of the reaction to 20 µL. The ligation occurs at room temperature for 45 minutes. 1 µL of the crude ligation reaction was often saved for PAGE to verify proper ligation. The ligated material was purified using RNAClean XP SPRI beads (Beckman Coulter

Life Sciences: A63987) either 4.4x or 1.8x the volume of the reaction. The beads incubated with the ligation reaction for 20 minutes in a LoBind Eppendorf tube. The tube was placed on a magnetic rack for 10 minutes to pellet the beads. The supernatant was removed. The beads were washed with 150 µL of 80% EtOH wash followed by a 150 µL 100% EtOH wash. The tube was switched between alternate sides of the magnet 4 times to clean the beads, between each EtOH wash. This action forces the beads to move through the wash. The final reaction was eluted in 11.5 µL NF $H_2O$, over a 20 minute period with occasional flicking.

**RMX Ligation and Loading:**

The RMX ligation requires the following components: 6 µL of the RMX adapter, 5 µL of the NEB 5x Quick Ligation Reaction buffer (NEB: B6058S), 11 µL of the splint ligation, and 3 µL of 2,000,000 units/mL T4 DNA ligase (NEB: M0202S/M0202M). RMX adapter is included in ONT's RNA sequencing kits. If needed, NF $H_2O$ was added to increase the reaction volume to 25 µL. The reaction was carried out at room temperature for 30 minutes. 1 µL of this reaction was often saved for a gel, before starting the SPRI cleanup. The reaction was then purified with 2.5x or 1.5x SPRI beads. ONT's wash protocol for the second ligation was followed. The reaction was eluted in 12.5 µL elution buffer, using a 20 minute elution time. 1 µL of the reaction was used for NanoDrop quantification, to confirm presence of RNA. ONT protocol was followed for final preparation of the sample and the flow cell before loading. Sequencing runs were less than 24 hours of runtime due to events that lead to flow cell deterioration. Continuous data was collected for the first 1-2 hours of sequencing. When nanopores in the flow cell were clogged, the experiment was restarted; most sequencing runs would be restarted 4 or 5 times.

**Bioinformatics Work:**

All data was basecalled using Guppy version 2.1.3. Reads were aligned to a tRNA reference sequence that included the ribonucleotide portions of the adapters. BWA-MEM, version 0.7.17-r1188, with the following settings, "-W 13 -k 6 -x ont2d" was used for all alignments [9]. The SAM files were filtered for primary alignments using SAMtools version 1.6 and viewed using the Broad Institute's Integrative Genomics Viewer (IGV) version 2.4.14 [10, 15]. Full length reads were determined using SAMtools to filter reads that aligned to the terminal 3 bases AS102/AS102_18 adapters at the 3′ side and within the subset of those reads filtered down to reads that also aligned to the CCA tail of the tRNA. In addition these regions were used to calculate coverage on either ends of the tRNA. These regions were chosen based on two criteria, they are both close to the ends of the tRNA and the coverage of the region is fairly high regardless of sample. See supplements for IGV screen captures.

**Liquid Chromatography/Mass Spectrometry Analysis for RNA Modifications of *E. coli* fmet tRNA:**

Mass spectrometry was performed using modified ribonucleoside standards and on tRNA digested to ribonucleosides using three different enzymes. The method used was derived from Crain [5]. The standards are 4-thiouridine, 2′-O-methylcytidine, 7-methylguanosine, and 5-methyluridine. 60 ng of each standard was mixed into 0.1% formic acid in NF $H_2O$, to be run either individually or as a mix. This was used to find retention times of the expected modifications. The *E. coli* fmet tRNA is reported to have dihydrouridine, as well. A commercial standard was unavailable making it difficult to unequivocally determine the presence of this modification. Both the biological tRNA and canonical version which is used as a negative control were digested. 4 µg of each tRNA was first digested with 1 unit of nuclease P1 (Sigma Aldrich: N8630-1VL) in a 10 µL solution of 10 mM $NH_4OAc$

and NF $H_2O$. This mix was incubated in a thermocycler at 45°C for 2 hours. The tRNA is then digested with 0.004 units Phosphodiesterase 1 (Thermo Fisher Scientific: AAJ20240EXR) in 50 mM $NH_4HCO_3$, and 2.5 mM $MgCl_2$ with $H_2O$ up to 20 µL. The reaction was heated to 37°C for 3 hours. The final digest used Antarctic Phosphatase (NEB: M0289S). The enzyme was mixed with the supplied buffer as this dilution is easier to pipette. The enzyme mix is then added to the reaction such that final concentration of Antarctic Phosphatase buffer is 0.2x and 0.5 units of Antarctic Phosphatase are in the reaction. The reaction volume totaled 25 µL adding NF $H_2O$ as needed, and was incubated at 37°C for 1 hour. 50 µL of 0.1% formic acid solution in NF $H_2O$ was added to the digested sample. The sample is then purified by loading the sample onto a Nanosep 3k column and then spun for 10 minutes at 14,000 RCF. The flowthrough is collected and placed into vials. LC-MS/MS was done at UC Santa Cruz, with the LTQ-Orbitrap Velos Pro MS from ThermoFisher, in positive ion mode. The column used was the Synergi 4 µm Fusion-RP 80Å C18 column manufactured by Phenomenex. Two solvents were used A, which is composed of 0.1% formic acid in NF $H_2O$ and B, which is 0.1% formic acid in acetonitrile. The software used to control the LC-MS/MS is Thermo Fisher's Xcalibur software and was also used for data analysis.

# Chapter 3

# Results and Discussion

**IVT production of bovine fmet tRNA:**

In vitro transcription with T7 RNA polymerase was done to produce bovine mitochondrial initiator (bovine fmet) tRNA for libraries. NanoDrop measurements of the IVT product estimated an abundance of tRNA was being produced, however initial gel electrophoresis experiments conflicted with these estimates. In order to confirm the production of bovine fmet tRNA different masses of the IVT product were loaded onto a gel for visualization. These masses are based on NanoDrop measurements and all values are expected to be visible on PAGE. In Figure 3.1, lanes 2-5 contain IVT product loaded from 5 picomoles, (115 ng) to 150 picomoles (3.4 µg) of tRNA. The IVT reaction was treated with DNAse I. Any visible bands are indicative of partial or complete RNA product. Lane 1 contains the DNA template, 99 bases used in the IVT reaction. There is no evidence of any residual DNA template in lanes 2-5. The IVT tRNA is 72 bases, and in lanes 2, 3, and 4 there is a band present between 50 and 80 bases. Lane 2 contains 150 picomoles of tRNA in this lane. There are bands below 50 bases. These faint bands suggest incomplete product is being produced. Lane 5 contains

5 picomoles of tRNA, however there are no visible bands. This is surprising as 5 picomoles of the tRNA is more than a 100 ng of material. This amount of mass is expected to be visible on PAGE stained with SYBR Gold. Lane 4 contains 3 times the amount material as lane 5, however the band in lane 4 is faint. This suggests that the NanoDrop is overestimating the concentration of the sample, potentially due to free NTPs that are not removed during the phenol/chloroform extraction and EtOH precipitation. To remove smaller fragments and NTPs, Bio-Rad P6 columns were used to further purify the IVT reaction. The NanoDrop measurements of the P6 purified tRNA were more consistent. This leads to the conclusion that the IVT production process requires more stringent purification. The current method is an extraction followed by EtOH precipitation and additional P6 column filtration. In order to avoid tandem purification methods, future work will likely use gel purification methods as the band of interest can be excised out. In addition a secondary method of quantification such as Qubit may provide more accurate measurements.

The following 8% PAGE, stained with SYBR gold, shows IVT produced bovine tRNA loaded onto a gel at various quantities based on NanoDrop quantification. This was diagnostic to determine if the NanoDrop was accurate and visualize the production of any partial products.

The gel to the right is both labeled at the top and bottom. All samples were in buffer of 7M urea and 0.1x TBE.
L) NEB low range ssRNA ladder
1) Template oligo: DNA template for the mitochondrial bovine initiator tRNA.
2) 150 pmols, 3.4 µg, of IVT produced bovine fmet tRNA. There is evidence of some partial product.
3) 50 pmols, 1.16 µg, of IVT produced tRNA
4) 15 pmols, 347.5 ng, of IVT produced tRNA
5) 5 pmols, 115.8 ng, of IVT produced tRNA. There are no bands visible in this lane.
L) NEB low range ssRNA ladder

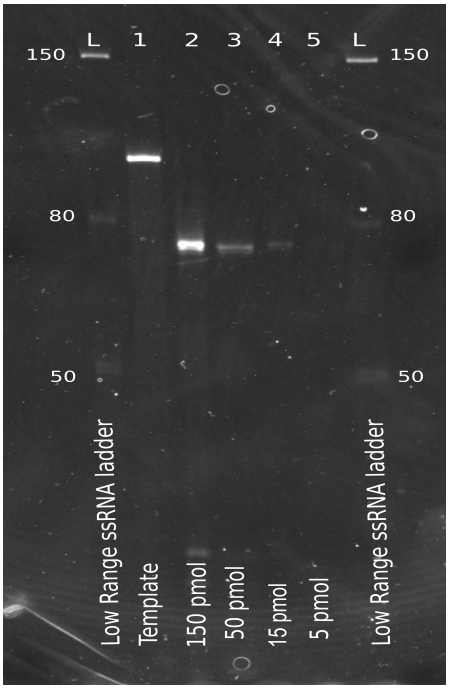



**Figure 3.1:** IVT Production of Bovine fmet tRNA

**Determining Efficacy of RppH:**

After producing the IVT construct, the splint adapters and ONT RMX adapters were ligated to the tRNA and sequencing runs were carried out. Sequencing runs had low throughput. The number of basecalled reads was much lower than expected for the MinION. The initial hypothesis was that RppH was inefficient in converting the 5′ triphosphate into a monophosphate for ligation. To test this hypothesis tRNA treated with RppH and untreated tRNA were ligated with the splint adapters. In addition, tRNA treated with RppH at 42°C, rather than the NEB recommended 37°C was also ligated. Since tRNA has a strong secondary

structure, heating the reaction to a higher temperature may better allow RppH access to the tRNA to cleave the 5′ triphosphate. As a positive control, *E. coli* fmet tRNA was ligated with M221 and AS102 adapters. The biological *E. coli* tRNA is expected to have the proper 5′ monophosphate end for ligation.

Figure 3.2 shows all ligation reactions. Lane 1 contains the IVT produced tRNA, 72 bases, below the 80 base marker as expected. Lanes 2 and 3 are ligations with tRNA treated by RppH at 37°C and 42°C. Visually it is apparent they are very similar. Ligation of the adapters to the tRNA has 3 possible configurations. The first is ligation of both ends to the tRNA getting a product of 126 bases in length. The other two configurations are ligation to only the 3′ side of the tRNA with M221, 102 bases, or ligation of the 5′ side of the tRNA with AS102, 96 bases. In both the 37°C and 42°C reaction lanes all 4 bands are visible. The IVT produced tRNA, is a visible band below the 80 base marker. There are 3 bands between 80 and 150 bases, corresponding to each of the possible ligations. The top band is the complete ligation product. When looking at the negative control in lane 4 there is only one band above the 80 base marker. The negative reaction was not treated with RppH therefore, only 3′ ligation is possible. In the -RppH lane there is only single band which migrates down between 80 and 150 bases corresponding to the expected size of 102 bases. In lanes 2 and 3 there is a size shift of the tRNA band compared to lane 1, this is likely due to the RppH treatment. The tRNA band in -RppH migrates to same position as in lane 1.

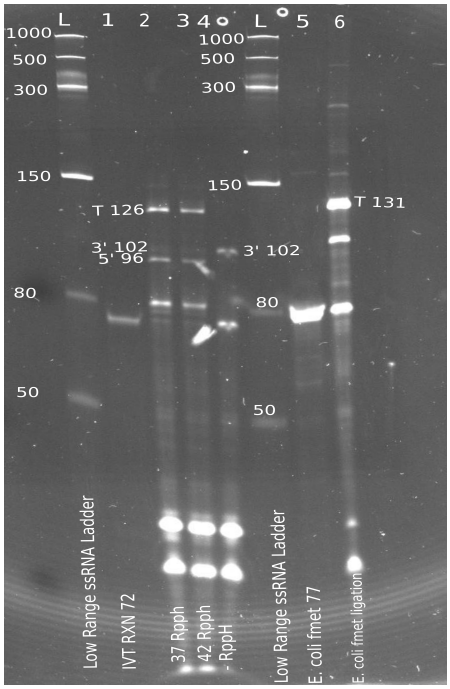
Lanes 5 and 6 show ligation of the *E. coli* fmet tRNA. Lane 5 contains the tRNA by itself and is 77 bases long. It runs very close to the 80 base marker as expected. In lane 6 there are two distinct bands above the tRNA. One is the complete ligation and the second is a partial ligation. Without another marker it is difficult to tell whether it is a partial ligation due to just 5′ ligation or 3′

ligation, however the complete product is formed. In addition to the major bands corresponding to ligation products, there are also faint bands higher up which may be material that has not denatured.

Figure 3.2 shows that RppH does cleave the 5′ triphosphate to make it amenable for ligation. An important caveat is that ligation efficiency is relatively low, and needs to be improved. RppH inefficiency is unlikely to be the cause of the clogs faced in early experiments. The splint ligation is not efficient as there is still unligated tRNA, even in the biological *E. coli* fmet tRNA. In the RppH negative lane the 3′ ligation band is the dimmest. This suggests that M221 ligation may help stabilize the AS102 adapter facilitating complete ligation. This is based on the fact that 3′ ligation, 102 bases, in the 37°C and 42°C is faint in both lanes. However, the dynamics appear to be more complicated as the 5′ ligation by itself happens as often as the complete ligation as well. This gel also confirms that the *E. coli* tRNA was able to ligate as efficiently as the IVT product; this was a pivotal moment leading to a switch between working with IVT constructed tRNAs and sequencing this biological sample. The IVT preparation of tRNAs takes a significant amount of time, to produce and purify. In addition a bulk scale RppH reaction is needed to purify enough material before ligation can even occur. Modomics reports the presence of modifications on the *E. coli* fmet tRNA; this lead to shifting the focus of the work to sequencing biological tRNA and to understand how modifications affect MinION basecalling [2]. For others who wish to use IVT to generate material for sequencing, it is beneficial to do a bulk RppH reaction, followed by a bulk splint ligation. Then use any method of choice to preferentially select for completely ligated material such as gel purification.

**Figure 3.2:** Verification of RppH Activity and Splint Ligation

This 8% PAGE, stained with SYBR gold, shows ligations of IVT product after RppH treatment at various temperatures along with a negative control. The RppH treated lanes show that the RppH is necessary to get complete ligation of the splint adapter, 126 bases. Lanes 5 and 6 are *E. coli* fmet tRNA followed by ligation. The *E. coli* tRNA is a positive control. All samples are in NEB 2x RNA dye.




L) NEB low range ssRNA ladder
1) IVT bovine fmet tRNA, expected to be 72 bases
2) Ligation using RppH treated IVT tRNA. The RppH reaction was carried out at 37°C. Two partial ligations with each half of the adapter are believed to have occurred, the bands are labeled 5′ 96 and 3′ 102. The complete ligation is believed to be the band labeled T 126.
3) Ligation using RppH treated IVT tRNA. The RppH reaction was carried out at 42°C. The bands are very similar to lane 2.
4)The negative control lane; the IVT tRNA was not treated with RppH. There is evidence of partial ligation due to M221 being able to ligate to the 3′ side of the tRNA.
L) NEB low range ssRNA ladder
5) Biological *E. coli* fmet tRNA, 77 bases.
6) Ligation of the *E. coli* fmet tRNA, the complete ligation is expected to be at the band labeled T 131.

**Optimizing SPRI Cleanup:**

After confirming the ligation of the *E. coli* fmet tRNA and activity of RppH, Figure 3.2, I attempted to sequence *E. coli* fmet tRNA. These sequencing runs also had low throughput, this lead to a new hypothesis that free adapter and unligated

tRNA are unable to completely translocate through the nanopores. When these analytes get stuck in the nanopore an electronic feedback system attempts to eject them. When sequencing tRNAs, these small molecules fail to get ejected.

The SPRI cleanup is a key part of library preparation. The beads bind the nucleic acid, separating the tRNA from the enzymes and the reaction buffer. The original protocol developed by the Nanopore group required 4.4x the volume of the ligation reaction for the SPRI cleanup to maximize recovery of the tRNA. With such a high ratio of SPRI beads the unligated tRNA and free splint adapter is also collected as well. There was a need to optimize the SPRI bead conditions used in the cleanup steps to maximize the amount of ligated tRNA recovered while minimizing other extraneous RNA. I did a bulk scale ligation using IVT produced bovine fmet tRNA purified with a P6 column. The reaction was split up evenly for 5 SPRI cleanups. The SPRI cleanups were done at 1.8x, 2.0x, 2.25x, 2.5x and 3.0x, Figure 3.3. The SPRI cleanups have made a marginal difference in material recovery. All washes recovered some amount of adapter boxed red and partially ligated material in green. The 2.25x cleanup recovered the most amount of RNA, 85.8 ng, however there is a significant amount of adapter remaining. At the 2.0x and 1.8x SPRI concentrations, unligated adapter and partially ligated tRNA populations are recovered proportionally less. It is important to note the parabolic nature of the recovery, going above or below 2.25x SPRI decreases recovery. In the 1.8x, 2.5x, and 3.0x cleanups, the amount of material recovered was 66 ng, 68.2 ng and 66 ng respectively. From this gel, the 2.5x and 3.0x lanes have significantly more adapter as the bands are brighter, than the 1.8x lane. The NanoDrop results in conjunction with the gel suggest that a more stringent SPRI wash removes a greater proportion of extraneous adapter at the cost of recovery. The hypothesis is that either small adapters or unligated tRNAs are causing clogs; decreasing the

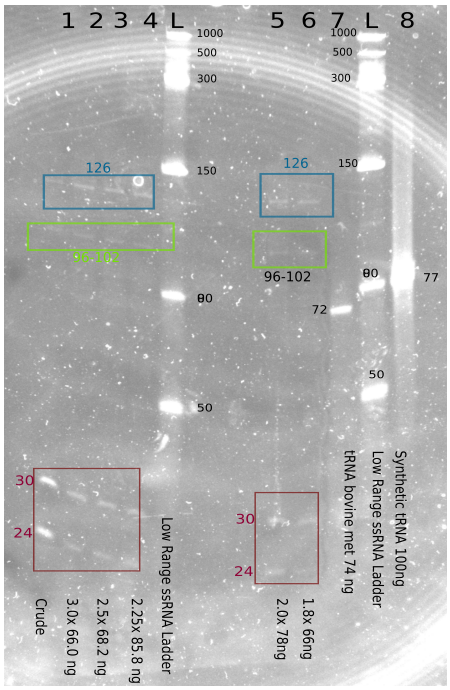
volume of the SPRI cleanup was the best option.

In lane 8, 100 ng of the synthetic fmet tRNA, 77 bases, based on *E. coli* fmet sequence purchased from Dharmacon was loaded. There is a single bright band at the 80 base marker and no other minor bands. Upon confirming the ligation of the biological fmet tRNA in previous work, the project oriented towards sequencing both modified and canonical versions of *E. coli* fmet tRNA. The current library preparation protocol uses SPRI washes at 1.8x and 1.5x the volume of the first and second ligation reaction respectively.

**Figure 3.3:** Optimizing SPRI Cleanup

The 8% PAGE, stained with SYBR gold, shows how varying the stringency of the SPRI wash on IVT constructed tRNA affects recovery. In these lanes, bands believed to be splint adapters, 24 and 30 bases are boxed red, partial ligations, 96-102 bases, green and complete ligations, 126 bases, blue. The last lane contains the synthetic fmet tRNA purchased from Dharmacon to confirm its purity. All samples are in NEB 2x RNA dye.

1) Crude ligation reaction, to confirm ligation occurred.
2) SPRI cleanup, 3.0x SPRI beads, 66.0 ng loaded into lane.
3) SPRI cleanup, 2.5x SPRI beads, 68.2 ng loaded into lane.
4) SPRI cleanup, 2.25x SPRI beads, 85.8 ng loaded into lane.
L) NEB low range ssRNA ladder.
5) SPRI cleanup, 2.0x SPRI beads, 78.0 ng loaded into lane.
6) SPRI cleanup, 1.8x SPRI beads, 66.0 ng loaded into lane.
7) IVT produced Bovine fmet tRNA, expected size of 72 bases.
L) NEB low range ssRNA ladder
8) Synthetic *E. coli* fmet tRNA manufactured by Dharmacon, 77 bases long.



**fmet Biological and Synthetic Ligations:**

Figure 3.4 shows the ligation of the biological *E. coli* fmet tRNA and its canonical copy. The ligations shown are sampled from library preparations. Lanes 1-3 focus on the biological tRNA and lanes 4-6 are the synthetic tRNA. The ligation of the M221/AS102 adapters to the biological tRNA was successful. Lane 1 shows the ligation of the biological *E. coli* fmet tRNA, 77 bases, to the splint. The tRNA is visible at the 80 base marker in lane 1. There is a partial ligation and a complete ligation, both between 80 and 150 bases. The splint ligation of synthetic tRNA, lane 4, also occurs successfully. In lane 4 the tRNA is visible as a faint band at 80 bases followed by the complete ligation below 150 bases. When the gel is exposed for a longer period there is a partial ligation visible as well, see supplements. The ligation of the ONT RMX adapters to the tRNA is not visible in either lanes 2 or 5.

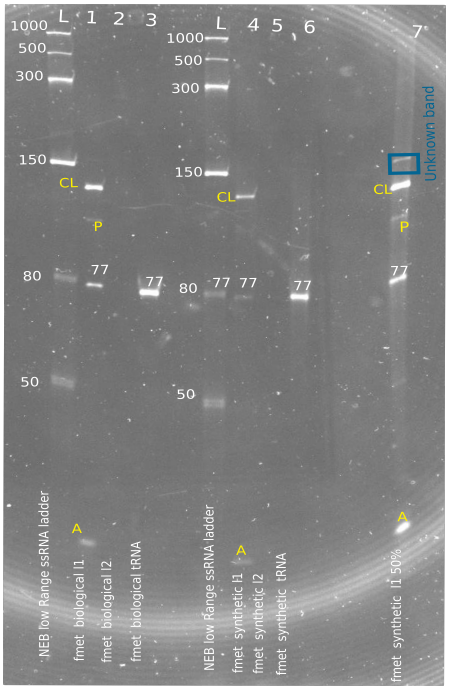
The banding patterns of the splint ligation, lane 1 and lane 4, show that ligation is inefficient with a significant amount of unligated and partially ligated tRNA. In order to better understand how much tRNA is being ligated, lane 7 was loaded with 50% of a standard ligation reaction using synthetic fmet tRNA. This is approximately 125 ng of tRNA. In this lane there is a bright adapter band, followed by tRNA at 77 bases. Between the 80 base and 150 base marker, there are two bands, corresponding to the partial ligation and the complete ligation. In addition there is a band above the 150 base marker that appears to be an unknown product, at higher exposure this product is also visible in other lanes (see supplements, page 5). This band may be material that has not fully denatured or dimerized. To test this, the band will be excised from the gel and run in a second denaturing gel. If the band migrates lower this would suggest incomplete denaturation of the sample and more loading buffer and heat is necessary. If

it is the formation of other ligation products, the relative intensity of the band suggests it is a minority product and size exclusion based purification such as gel purification can be used to remove these products.

The gel shows that the splint ligation did occur for the sequencing runs of the synthetic and biological tRNA, however it is not efficient. The RMX ligations in lanes 2 and 5 are not visible, however tRNA was sequenced, reported in Table 3.1. The lack of visible bands is likely caused by the amount of mass loaded into the gel. The SPRI cleanup recovers <100 ng for any of the SPRI concentrations used, see Figure 3.3. Assuming a best case scenario of 50% recovery, the concentration of tRNA in ligation 2 is 5 ng/µL. These lanes contain 1 µL of the ligation reaction. In the second ligation the concentration of nucleic acid is too low to see on PAGE stained with SYBR gold. While these sequencing runs were able to generate some data, Table 3.1, the flow cells also failed to sequence for a full day as the pores would eventually get clogged. The gel shows that the splint ligation is inefficient as a significant amount of material is partially ligated or not ligated at all. If unligated tRNA is carried forward it may not translocate through the pore as ligation of the RMX adapter is hindered. In addition without both ends of the splint adapter attached there is an effect on coverage, Table 3.2. A major future goal is improving the efficiency of splint ligation and selection for fully ligated tRNA.

**Figure 3.4:** Ligation of *E. coli* Synthetic and Biological fmet tRNA

Below is an 8% PAGE, stained with SYBR gold, showing the ligations of the *E. coli* biological tRNA and the synthetic fmet tRNA. Lane 7 is half the splint ligation for synthetic fmet tRNA. This figure also has companion in the supplements with a longer exposure showing additional bands. Adapter bands are labeled with an A. Partial ligations are labeled with a P. CL is used to denote complete ligation. All samples are in NEB 2x RNA dye.




L) NEB low range ssRNA ladder
1) Ligation of splint adapters to *E. coli* fmet biological tRNA.
2) Ligation of the RMX adapters to the product of the first ligation, lane 2. No bands are visible.
3) Biological *E. coli* fmet tRNA, expected size is 77 bases.
L) NEB low range ssRNA ladder
4) Ligation of splint adapters to *E. coli* fmet synthetic tRNA. A partial band is present when exposed for longer amount of time.
5) Ligation of the RMX adapters to the product of the first ligation, lane 4. No bands are visible.
6) Synthetic *E. coli* fmet tRNA, size is 77 bases.
7) 50% of the splint ligation reaction using synthetic *E. coli* fmet tRNA as the substrate. The blue box indicates the presence of an unknown product to large to be the ligation.

**Confirming AS102_18 and M221_6 Splint Ligation:**

Early alignments showed low coverage at the terminal edges of the tRNA, specifically at the 3′ end, see Table 3.2. The hypothesis is that addition of more RNA on either side of the tRNA would improve coverage. The logic for additional ribonucleotides flanking that tRNA is to provide the basecaller more raw current signal. This allows the basecaller to start calling the adapter region first before entering the tRNA and improve basecalling accuracy and coverage. The basecaller then has more data before and after the tRNA to more accurately determine the ends. The simplest adjustment to make to the adapters was replacing regions adjacent to the tRNA with corresponding ribonucleotide bases. These new adapters AS102_18 and M221_6 have 18 and 6 ribonucleotide bases to flank the tRNA on both ends. Direct RNA sequencing on the MinION works by translocating RNA 3′ to 5′. The presence of ribonucleotides prior to the 3′ CCA tail should improve coverage at the CCA tail by providing additional information for basecalling.

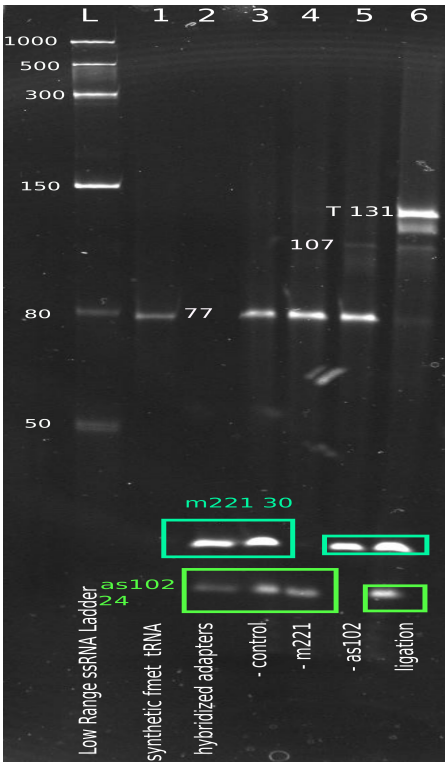
After purchase of the new adapter oligos they were hybridized and tested on the synthetic *E. coli* fmet tRNA to confirm ligation. Figure 3.5 shows ligation of the tRNA to the new adapters, along with a set of controls. Lanes 1 and 2 contain the tRNA and splint adapters respectively. Lane 3 is the negative control, no T4RNL2 enzyme, the tRNA and adapters are visible as clear bands. Ligation products are expected to be above the 80 base marker, in the negative there are no ligation products as expected. Lane 4 is the ligation, with the AS102_18 adapter only. In this lane there are bands corresponding to the tRNA and the AS102_18 adapter but no evidence of partial ligation. Lane 5 is ligation with the M221_6 adapter only and there is a faint band above the tRNA band, not present in the negative control. This evidence suggests that partial ligation of M221_6 at the 3′ end of the tRNA is possible, but happens very inefficiently. In lane 6 all

components of the ligation are included. In this lane there is evidence of partial ligation most likely with M221_6 as the band migrates to the same position as the ligation band in lane 5. The complete ligation product, 131 bases, is visible under the 150 base marker in lane 6. There is also a smear under this band as well. It is unknown if the smear is the complete ligation or other fragments still migrating down. Lanes 6 shows that the new adapters are able to ligate to the tRNA and lane 5 shows that M221_6 is able to ligate to the tRNA by itself. I have confirmed that the modified RNA/DNA hybrid adapters are able to ligate to the tRNA using the same T4RNL2 enzyme.

**Figure 3.5:** Ligation of AS102_18 and M221_6

This 8% gel, stained with SYBR gold, shows the ligation of synthetic fmet tRNA to the AS102_18 and M221_6 ribonucleotide heavy adapters and controls. All samples are in NEB 2x RNA dye. The results of this gel show that splint ligation to the tRNA is possible even with the increase in ribonucleotide bases of each adapter. In addition M221_6 is able to ligate to the tRNA by itself but the 5′ AS102_18 ligation by itself is unlikely.

L) NEB low range ssRNA ladder
1) Synthetic fmet tRNA, 77 bases.
2) Hybridized mixture of M221_6, 30 bases, and AS102_18, 24 bases in green boxes.
3) Negative control with no enzyme.
4) Ligation with AS102_18 portion of adapter only. No product formed.
5) Ligation with M221_6 only. Faint band, labeled at 107 bases, is visible. This suggests partial ligation is possible.
6) Ligation reaction with all components. There is evidence of both a partial and complete ligation.



### Summary of Sequencing Runs

Tables 3.1 and 3.2 provide information about the sequencing runs, including the number of aligned reads and number of full length reads. Tables 3.1 and 3.2 show the effects of longer flanking ribonucleotide regions on coverage of the tRNA at its terminal ends. The effects of switching from the original AS102/M221 adapters to AS102_18/M221_6 for each construct are discussed separately.

For the synthetic construct there were two sequencing runs, one with the original M221/AS102 adapter mix and the second with the AS102_18/M221_6 adapter. The results of the sequencing run can be seen in the table below. The first run with the original adapters had 40,353 reads and the second run used AS102_18/M221_6 had 36,234 reads basecalled. While the number of reads decreased between the first and second run, the number of aligned reads to the reference sequence increased. The run with the original adapters only had 5.87% of the reads align, increasing to 14.18% with the new adapters. The number of full length reads increased from 808 for the original adapter to 3,127. A major goal was improving coverage at the terminal ends. With the M221_6 adapter, the basecaller has signal to call before entering the 3′ end of the tRNA. Sequencing runs of both biological and synthetic tRNA show improvement in coverage at terminal ends of the tRNA when using the ribonucleotide containing adapters. The 5′ coverage increases from 64.89% to 72.21% and 3′ coverage increases from 55.94% to 86.82%. While the sample space for the synthetic construct is small, a similar trend in coverage also occurs for the biological tRNA.

For the biological *E. coli* fmet tRNA, 5 sequencing runs were done of which 2 used the AS102_18/ M221_6 adapters. The 01/23/19, original adapters, and 01/30/19, new adapters, runs are the most comparable biological runs, as they used the same SPRI purification ratios and same ONT library kits. With the

original adapters, 2.84% of the reads aligned, with AS102_18 and M221_6 that rose to 14.74%. The number of full length reads increased as well. The number of full length reads increased to 79.82% from 45.12% when using SAMtools to filter full length reads. In addition coverage at the 5′ end improved. With the original adapters 85.23% of the reads covered the 5′ region improving to 90.39% when using AS102_18/M221_6. The 3′ end also saw a large increase in coverage. The jump was from 53.05% to 89.04%, similar numbers to the synthetic.

When comparing all runs in aggregate there is an increase in the number of alignments when using the AS102_18/M221_6 adapter pair as compared to the original adapters. The coverage also improves at the terminal ends of the tRNAs. The 3′ coverage had a drastic improvement, with the original adapters less than 60% of the reads covered the CCA tail increasing up to 89.04% for the highest biological tRNA run. The 5′ end of the tRNA was flanked by ribonucleotides with the original AS102 adapter, however there is small increase in coverage with AS102_18. It is less than 10% but this increase is noticeable and consistent across all biological runs. The average coverage at the 5′ terminal for the 3 biological runs with the original AS102/M221 adapters is 84.39% was boosted to 91% in the two runs with AS102_18/M221_6. For the synthetic it was an 8% increase.

**Table 3.1:** Summary of fmet Biological and Synthetic Sequencing Runs

| Date | Sample | Adapter | Reads | % Aligned | % Full Length |
|---|---|---|---|---|---|
| 10/17/18 | biological | AS102/M221 | 42681 | 3.12 | 38.95 |
| 11/14/18 | biological | AS102/M221 | 111238 | 3.43 | 42.99 |
| 12/03/18 | synthetic | AS102/M221 | 40353 | 5.87 | 34.14 |
| 01/09/19 | synthetic | AS102_18/M221_6 | 36234 | 14.18 | 60.86 |
| 01/23/19 | biological | AS102_18/M221_6 | 80893 | 14.74 | 79.82 |
| 01/30/19 | biological | AS102/M221 | 51895 | 2.84 | 45.12 |
| 02/04/19 | biological | AS102_18/M221_6 | 137553 | 20.28 | 80.26 |

**Table 3.2:** Summary of Coverage for Sequencing Runs

| Date | Sample | Adapter | % 5′ coverage | % 3′ coverage |
|---|---|---|---|---|
| 10/17/18 | biological | AS102/M221 | 83.46 | 46.92 |
| 11/14/18 | biological | AS102/M221 | 84.49 | 51.59 |
| 12/03/18 | synthetic | AS102/M221 | 64.89 | 55.94 |
| 01/09/19 | synthetic | AS102_18/M221_6 | 72.21 | 86.82 |
| 01/23/19 | biological | AS102_18/M221_6 | 90.39 | 89.04 |
| 01/30/19 | biological | AS102/M221 | 85.23 | 53.05 |
| 02/04/19 | biological | AS102_18/M221_6 | 91.61 | 88.23 |

The data suggests that flanking the tRNA with more ribonucleotides has an improvement on coverage at the terminal regions around the tRNA. This can qualitatively be seen in the IGV screen captures included in the supplements, pages 2-4. Coverage of the 3′ CCA tail is relatively lower in the alignments without the modified adapters. In addition the percentage of alignments increases and among those alignments the number of full length reads also increases. The percentage of reads that cover either the 5′ or 3′ end of the tRNA stays relatively consistent throughout the biological samples regardless of library preparation or kit used, only changing as a function of the adapter used. It is unknown if this trend will be true for synthetic reads. The improvement in coverage was similar to the biological runs however more sequencing runs with the synthetic will need to be carried out.

ONT's basecaller uses a recurrent neural network paradigm, which relies on a notion of past and future data points existing. This allows the basecaller to consider how previous bases will affect the next base. The addition of ribonucleotide regions flanking the tRNA on either side provides more raw current signal and thus more data points for the basecaller to more accurately determine the terminal bases on the tRNA. Using the original M221 adapter there were no ribonu-

cleotide bases preceding the tRNA to provide information for basecalling. After redesigning the splint adapters to have more RNA sequence, there is an increase in coverage on the 3′ of the tRNA. In addition while the original AS102 adapter did have ribonucleotide bases there was still an improvement when switching to the AS102_18 adapter, this suggests that more ribonucleotide bases flanking the tRNA could further improve basecalling at terminal ends of the tRNA, up to the limits of nanopore error.

**Table 3.3:** Summary of Mass Spectrometry

| Modification | Mass (amu) | $[M+H]^+$ (m/z) | Breakdown product (m/z) | Visible |
|---|---|---|---|---|
| 4-thiouridine (4) | 260.0467 | 261.00545 | 129 | unknown |
| 2′-O-methylcytidine (B) | 257.1912 | 258.1082 | 112 | Y |
| 7-methylguanosine (7/m7g) | 299.123 | 298.1151 | 166 | Y |
| 5-methyluridine (T) | 258.0852 | 259.093 | 127 | Y |
| pseudouridine (P) | 244.0695 | 245.0773 | 209/179 155 | Break-down products visible |

**Difference in Alignment Profiles of Canonical and Biological tRNA:**

The motivation for running the biological and synthetic version of the same tRNA was to determine the sensitivity of the MinION to modifications. If modifications perturb the measured current it will have an effect on basecalling. The basecaller, Guppy, is designed report canonical bases, but the nanopore itself can process analytes that have been modified. This disparity between what occurs in biology and the limitations of basecalling software may cause modified nucleosides to present themselves as mismatches and indels when aligned to the reference

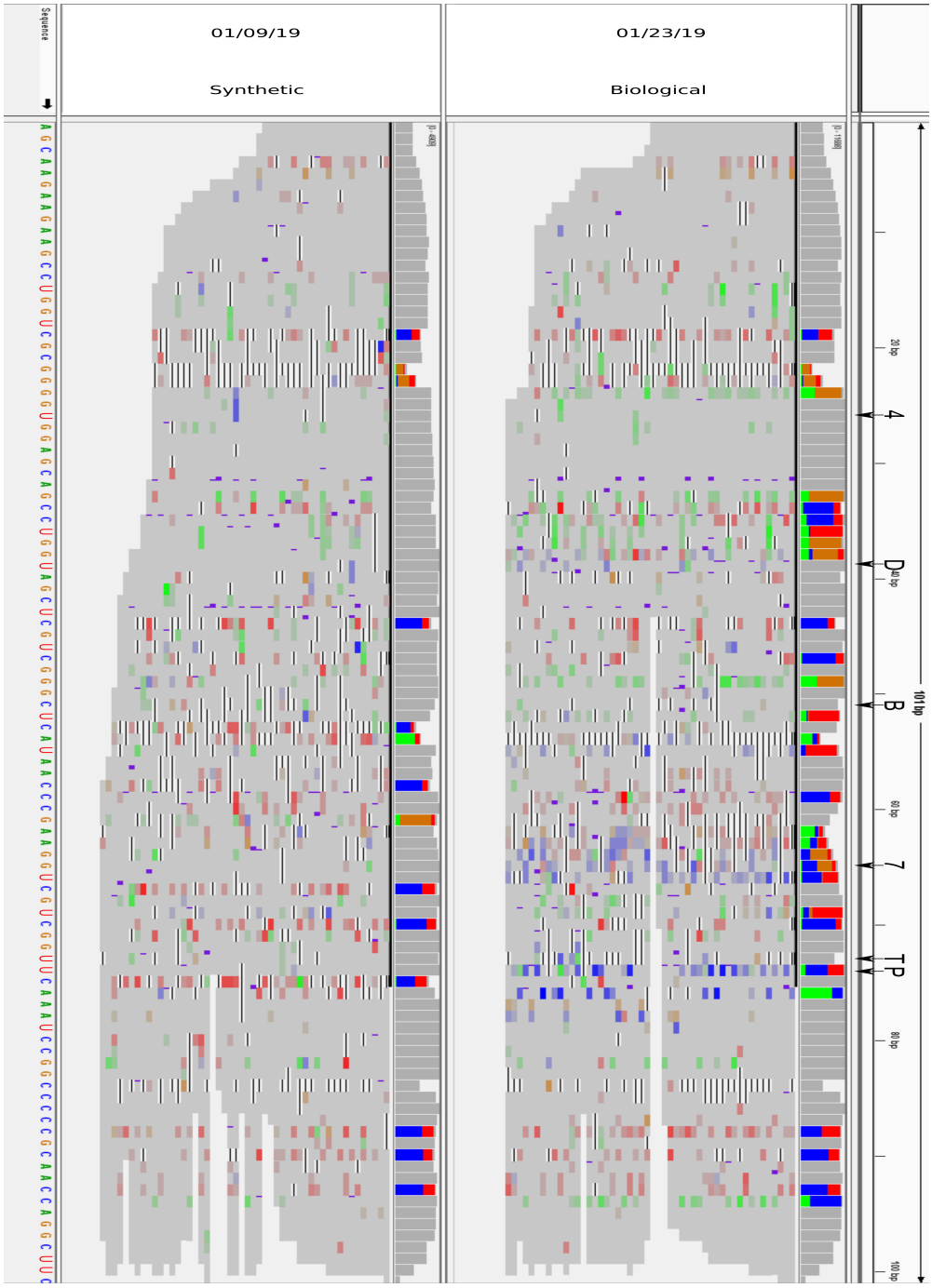



**Figure 3.6:** *E. coli* fmet tRNA is shown against the synthetic equivalent, there are regions of mismatches shown in the coverage plot, colored bars, that are different between the biological, top, and synthetic, bottom. IGV was used to visualize a downsampled set of alignments.

sequence. This difference in alignment space should be consistent and errors attributed to modifications should be higher than what is expected of the MinION. The canonical synthetic provides a baseline to differentiate expected nanopore basecalling error from error due to modifications. Detection of modifications via nanopore is not novel. Smith et al. showed that 7-methylguanosine perturbs current such that a systematic difference can be seen in MinION current traces [19]. In addition to sequencing both the synthetic and biological fmet tRNA, mass spectrometry of the biological sample was carried out by a member of the UCSC Nanopore Group. The results are summarized in Table 3.3. The values for mass charge ratios were taken from Pomerantz et al. and Modomics [14, 2]. Of the six modifications reported to be present on *E. coli* fmet tRNA, four were confirmed to be present.

Figure 3.6 is the alignments of a biological fmet tRNA (top) sequencing run against a synthetic tRNA run (bottom). The specific samples are 01/23/19, biological, and 01/09/19, synthetic. These two samples were prepared using the same library preparation protocol. There is a significant difference in alignment profiles, looking at the coverage plots in the IGV browser; there are more areas (colored bars) in which mismatches occur for the biological. This is believed to be caused by modifications, as many regions of mismatches unique to the biological are proximal to the positions of modifications.

The first modification 4-thiouridine, 4, occurs at base 26 in the reference sequence. In both the synthetic and biological this base and proximal positions do not show any form of indel or systematic mismatch. Most of the reads that aligned at this position did not have deletions providing high coverage, and there is very little indication of systematic mismatches. The mass spectrometry results for this modification was inconclusive as there was no evidence of a product ion

peak in either the biological or standard. This modification will require future exploration, but it is unlikely to be present in the biological sample.

The next modification is dihydrouridine, D, which occurs at position 39. In the biological sample there is a region from positions 33-38 having mismatches. At position 39 the reference base, uridine is called more than 80% of the time, when looking at all biological runs it is only bases 35 and 38 that suffer from mismatches not the whole intervening region, regardless of adapter. One thing to note is that nanopore reads multiple bases at a time, rather than a single nucleotide so modified nucleosides can affect basecalling of proximal bases not just the base where the modification occurs. If the modification is truly present in significant levels it may be that the modification affects bases at its $5'$ proximal position rather than the base itself. In the synthetic runs the mismatches do not occur as often at either positions 35 or 38. This suggests that this state is intrinsic to the biological tRNA itself.

At position 51 is $2'$-O-methylcytidine, B. This modification was verified to be present with mass spectrometry data, and has strong breakdown product peak. The evidence suggests that it exists in the biological sample. Here positions 47, 49 and 52, have mismatches and it is consistent among all biological samples regardless of adapters. Neither of the two synthetic runs suffer from mismatches at the same frequency as the biological in this area. In the biological sample the reference base is called more than 80% of the time at the position of the modification. Much like dihydrouridine it may be possible that effects of the modification are proximal to the modification rather than the site of the modification itself. Both the mass spectra and the alignments point towards the modification occurring in the biological.

7-methylguanosine, 7, occurs at position 65. This modification has a very

prominent peak on the mass spectra digest of the biological tRNA. In the alignment profile it appears that the modification exists and causes a significant effect on the alignments as positions 62-66 suffer from a wide variety of mismatches. This pattern is consistent across all of the biological samples and does not occur at all in the synthetic. In the synthetic runs, reads that called a base at position 65 called the reference more than 80% of the time. The alignment profile having a large perturbed region in the area of the modification combined with the mass spectrometry results are indicative of the biological tRNA having the 7-methylguanosine modification. Clinically this is very useful, as one of the enzymes responsible for 7-methylguanosine modification in humans, WDR4 when mutated, results in the 7-methylguanosine modification to be lost from tRNAs. This mutation is also known to cause microcephalic primordial dwarfism [21]. This is an opportunity in which tRNA sequencing may provide clinical information, bypassing the need for DNA sequencing and RNA-expression analysis. 7-methylguanosine causes a significant perturbation in the signal, as the whole region proximal to the site of modification significantly differs from the reference sequence.

The last two modifications occur consecutively, at position 73 is 5-methyluridine, T, and at 74 is pseudouridine, P. There is a small peak in mass spectrometry for the presence of 5-methyluridine along with presence of its product ion. For pseudouridine the mass spectrometry data shows peaks corresponding to breakdown products that are not expected for uridine but for pseudouridine. Mass spectrometry alone suggests the presence of both these modifications, however the alignments are more difficult to decipher. At position 73 there is a drop in coverage, rather than mismatches. There is a clear mismatch pattern that occurs at position 74 where pseudouridine is located. For the biological reads that spanned position 73, many of the reads have a deletion at that position. For the biological

sequencing runs, the frequency of reads with a deletion where 5-methyluridine occurs ranges from 16%-25% for each run. In the synthetic the number of reads with a deletion at that position stays below 10%. Pseudouridine displays an interesting state as it most often interpreted as a cytosine rather than a uracil ranging from 44% to 51% at its highest in the biological runs. For the synthetic runs, position 74 is properly interpreted as a uracil more than 90% of the time. In addition the coverage is very high at this position. For 5-methyluridine the mass spectra plot shows a low abundance of the product ion. If 5-methyluridine is only present in low abundance in this biological sample it may not have a significant effect in alignment space.

There are common errors throughout the synthetic and biological alignments that stem from homopolymeric repeats and at points close to positions of ligation. At the start of the tRNA, position 19, there is a drop in coverage and a mismatch occurs in both the biological and synthetic. The 5′ start of the *E. coli* fmet tRNA does not base pair with the stem as it does in other tRNA. This may create a bubble during the ligation of the 5′ adapter. Another consistent error is at positions 22-24 where there is a stretch of guanosine with an intervening cytosine that likely confuses the basecaller as the tRNA enters the pore 3′ first. There is dip in coverage at position 54. There are a large number of deletions at this position across all runs. This region falls within the region of the anticodon loop. This area has very small homopolymeric runs, that may affect basecalling accuracy. The last place there is common error is at positions 88 and 90. This is at the 3′ end of the tRNA before the CCA tail. In this case there are small homopolymer repeats, specifically there is a repeat of Cs that is interspersed with a consecutive set of As. The mismatch at these positions are consistent across all runs, the mismatched base is usually called as uracil instead of the reference

cytosine. In one of the runs for the biological sample using the M221_6 adapter, position 90 is not colored by IGV in the coverage plot. When this run is inspected by hand it does show a similar error profile at this position with 25% of the reads that call a base at this position calling uracil as well. In the synthetic there are positions that show mismatches that are not present in the biological, this may be the affects of modifications on basecalling affecting the signal such a way that certain regions are more accurately called.

Comparing the alignments of both the synthetic and biological run suggests that the presence of modifications affects nanopore basecalling. Of the six modifications that are known to occur on this tRNA, mass spectrometry is inconclusive about 4-thiouridine and alignments suggest that modification is not there or it does not perturb current enough to affect basecalling. Dihydrouridine could not be confirmed with mass spectrometry however there is a difference in the alignment profiles of the synthetic and biological at bases proximal to the location of this modification. This difference only occurs in biological samples and not the synthetic suggesting that modification may have an effect on proximal bases. All other modifications on the tRNA were confirmed with mass spectrometry, with 7-methylguanosine having the strongest effect on basecalling accuracy. 2′-O-methylcytidine creates errors at regions proximal to modification itself, rather than the position of the modification itself. The last two modifications pseudouridine and 5-methyluridine occur consecutively making it difficult to determine whether it is the effect of both modifications that causes the profile seen. At position 73 there are more deletions in the biological than the synthetic. Without additional molecules that contain 5-methyluridine independent of pseudouridine it is difficult to conclude if 5-methyluridine has an affect on basecalling. In addition the mass spectra plot suggests that 5-methyluridine is not in high abundance in the

sample. 5-methyluridine modifications are controlled by the TRMT2A gene in humans which may have a role in cancer [22]. The results of TRMT2A expression studies are mixed but the gene may be overexpressed in certain breast cancers making it a valuable marker for sequencing based diagnosis [4]. If it is possible to link enzymes that modify tRNAs with disease phenotypes it may be possible to relate tRNA modification rates with diseases as well. The alignments show the potential of nanopore sequencing to recognize modifications. The work so far focuses on the effects of modifications on basecalling. Working at the signal level it may be these modifications could be detected and called natively, with progressive updates to the basecaller itself. Modification detection on the nanopore is still in its infancy, and the long range of effects of modifications on neighboring bases will be a focus of future research.

# Chapter 4

# Conclusion and Future Work

The results shown builds on the previous work of my colleagues. I have shown it is possible to translocate tRNAs through nanopores on the ONT MinION device. In addition this work provides a paradigm to detect modifications on tRNA. By sequencing biological *E. coli* initiator tRNA and a synthetic canonical construct, I have shown a difference in the alignment profiles. Many of the regions that differed between the biological tRNA and synthetic are likely due to modifications, confirmed with mass spectrometry, the most striking of which is 7-methylguanosine.

**Future Work:**

While I have made progress on sequencing tRNAs there are improvements to be made. The biggest issue is fixing subpar throughput. The MinION platform is designed to generate hundreds of thousands of reads. The best sequencing run had less than 200,000 reads. Throughout the runs, nanopores ended up clogged. The culprit is likely unligated or partially ligated tRNA entering the pores and getting stuck. The system attempts to eject the material but fails to do so, ultimately causing the pore to go out of commission. tRNAs are heavily modified and many of the modifications reported for *E. coli* fmet tRNA were verified,

however quantification of the modifications in the sample still has yet to be done. The current method is focused on determining the presence of modifications, but not how often they are present.

Future work will focus on the following: improving throughput, modification detection, signal-level analysis, and working with mixed sets of tRNAs. The biggest issue is improving throughput. If low throughput is caused by inefficient ligation, this could be solved by adjusting the library preparation steps. This includes more stringent SPRI washes to remove smaller material, adjusting the stoichiometry of the ligation, or looking at other methods such as gel purification of fully ligated material. In addition methods such as biotinylated adapters are being looked at as this presents an opportunity to selectively target tRNA in an impure sample. If a combination of these methods improves the ligation efficiency of the tRNA the next step is to sequence a mixed sample of tRNA. Many strains of *E. coli* are well documented in addition their tRNAs can be ordered commercially (Sigma Aldrich: 10109541001). The current method developed provides a framework in which tRNAs could potentially be sequenced in bulk. The alignments can the be combined with databases such as Modomics that provide information about the positions of modifications on tRNAs, to filter for modifications that may potentially perturb the signal. This could be used to develop a guided training set for newer basecallers in an efficient manner as tRNAs are modification dense. A more near term goal is a simple binary classification of modification presence. In regards to the *E. coli* fmet tRNA there is no proof of the 4-thiouridine modification, nor is it possible conclude the individual effects of pseudouridine and 5-methyluridine on basecalling accuracy. A future test would be to sequence this modifications independently. This will provide data regarding the effects of sequence context and the effects of consecutive modifications on the

current signal.

The rise of next generation sequencing has resulted in an abundance of data for diagnostic purposes. Direct RNA-sequencing is one of the newest methods, however the MinION has primarily been focused on long poly-A transcripts leaving small RNA out of the picture. tRNAs are essential to life and small mutations could have a large impact on overall health and quality of life. Sequencing tRNAs and the detection of their modifications could have a huge impact not just on sequencing technology applications but also healthcare. The MinION is a simple sequencing platform that requires very little desk space and only a laptop for sequencing. It does not take a special analyst to run the hardware or software components. A test case with an *E. coli* tRNA has shown that the MinION can directly sequence tRNAs and potentially be used to detect modifications. Continuing to sequence additional tRNAs may provide new insights into one of the most fundamental elements of biology.

# Chapter 5

# Supplementary Data

supplemental_data.pdf

# Bibliography


[1] J. A. Abbott, C. S. Francklyn, and S. M. Robey-Bond. Transfer RNA and human disease. *Frontiers in Genetics*, 5, Jun 2014.

[2] P. Boccaletto, M. A. Machnicka, E. Purta, P. Piatkowski, B. Baginski, T. K. Wirecki, V. de Crécy-Lagard, R. Ross, P. A. Limbach, A. Kotter, and et al. MODOMICS: a database of RNA modification pathways. 2017 update. *Nucleic Acids Research*, 46(D1):D303–D307, Jan 2018.

[3] D. Branton, D. W. Deamer, A. Marziali, H. Bayley, S. A. Benner, T. Butler, M. Di Ventra, S. Garaj, A. Hibbs, X. Huang, and et al. The potential and challenges of nanopore sequencing. *Nature Biotechnology*, 26(10):1146–1153, Oct 2008.

[4] Y.-H. Chang, S. Nishimura, H. Oishi, V. P. Kelly, A. Kuno, and S. Takahashi. TRMT2A is a novel cell cycle regulator that suppresses cell proliferation. *Biochemical and Biophysical Research Communications*, 508(2):410–415, Jan 2019.

[5] P. F. Crain. Preparation and enzymatic hydrolysis of DNA and RNA for mass spectrometry. *Methods in Enzymology*, 193:782–790, 1990.

[6] E. M. Gustilo, F. A. Vendeix, and P. F. Agris. tRNA's modifications bring order to gene expression. *Current Opinion in Microbiology*, 11(2):134–140, Apr 2008.

[7] S. L. Hiley, J. Jackman, T. Babak, M. Trochesset, Q. D. Morris, E. Phizicky, and T. R. Hughes. Detection and discovery of RNA modifications using microarrays. *Nucleic Acids Research*, 33(1):e2, 2005.

[8] M. Jain, H. E. Olsen, B. Paten, and M. Akeson. The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome Biology*, 17, Nov 2016.

[9] H. Li. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv:1303.3997 [q-bio]*, Mar 2013. arXiv: 1303.3997.

[10] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, and . G. P. D. P. Subgroup. The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)*, 25(16):2078–2079, Aug 2009.

[11] A. Meller, L. Nivon, and D. Branton. Voltage-driven DNA translocations through a nanopore. *Physical Review Letters*, 86(15):3435–3438, Apr 2001.

[12] D. Pak, R. Root-Bernstein, and Z. F. Burton. tRNA structure and evolution and standardization to the three nucleotide genetic code. *Transcription*, 8(4):205–219, Jun 2017.

[13] E. M. Phizicky and A. K. Hopper. tRNA biology charges to the front. *Genes & Development*, 24(17):1832–1860, Sep 2010.

[14] S. C. Pomerantz and J. A. McCloskey. *Analysis of RNA hydrolyzates by liquid chromatography-mass spectrometry*, volume 193 of *Mass Spectrometry*, page 796–824. Academic Press, Jan 1990.

[15] J. T. Robinson, H. Thorvaldsdóttir, W. Winckler, M. Guttman, E. S. Lander, G. Getz, and J. P. Mesirov. Integrative Genomics Viewer. *Nature biotechnology*, 29(1):24–26, Jan 2011.

[16] T. Salinas-Giegé, R. Giegé, and P. Giegé. tRNA Biology in Mitochondria. *International Journal of Molecular Sciences*, 16(3):4518–4559, Feb 2015.

[17] J. T. Simpson, R. E. Workman, P. C. Zuzarte, M. David, L. J. Dursi, and W. Timp. Detecting DNA cytosine methylation using nanopore sequencing. *Nature Methods*, 14(4):407–410, Apr 2017.

[18] A. M. Smith, R. Abu-Shumays, M. Akeson, and D. L. Bernick. Capture, Unfolding, and Detection of Individual tRNA Molecules Using a Nanopore Device. *Frontiers in Bioengineering and Biotechnology*, 3, Jun 2015.

[19] A. M. Smith, M. Jain, L. Mulroney, D. R. Garalde, and M. Akeson. Reading canonical and modified nucleotides in 16S ribosomal RNA using nanopore direct RNA sequencing. *bioRxiv*, page 132274, Apr 2017.

[20] T. Suzuki and T. Suzuki. A complete landscape of post-transcriptional modifications in mammalian mitochondrial tRNAs. *Nucleic Acids Research*, 42(11):7346–7357, Jun 2014.

[21] C. Tomikawa. 7-Methylguanosine Modifications in Transfer RNA (tRNA). *International Journal of Molecular Sciences*, 19(12), Dec 2018.

[22] A. G. Torres, E. Batlle, and L. Ribas de Pouplana. Role of tRNA modifications in human diseases. *Trends in Molecular Medicine*, 20(6):306–314, Jun 2014.

[23] J. Zhao, B. Qin, R. Nikolay, C. M. T. Spahn, and G. Zhang. Translatomics: The Global View of Translation. *International Journal of Molecular Sciences*, 20(1), Jan 2019.

[24] G. Zheng, Y. Qin, W. C. Clark, Q. Dai, C. Yi, C. He, A. M. Lambowitz, and T. Pan. Efficient and quantitative high-throughput transfer RNA sequencing. *Nature methods*, 12(9):835–837, Sep 2015.