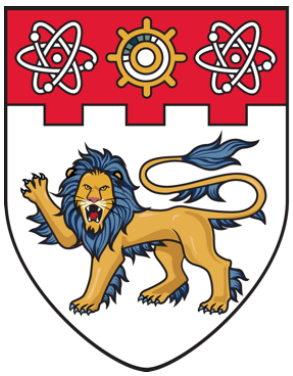

NANYANG TECHNOLOGICAL UNIVERSITY
SINGAPORE

# RNA structure probing using long-read Nanopore sequencing technology

**AW JONG GHUT, ASHLEY**

**SCHOOL OF BIOLOGICAL SCIENCES**

**2022**

# RNA structure probing using long-read Nanopore sequencing technology

**AW JONG GHUT, ASHLEY**

**SCHOOL OF BIOLOGICAL SCIENCES**

A thesis submitted to the Nanyang Technological University in partial fulfilment of the requirement for the degree of Doctor of Philosophy

2022

# Statement of Originality

I hereby certify that the work embodied in this thesis is the result of original research done by me except where otherwise stated in this thesis. The thesis work has not been submitted for a degree or professional qualification to any other university or institution. I declare that this thesis is written by myself and is free of plagiarism and of sufficient grammatical clarity to be examined. I confirm that the investigations were conducted in accord with the ethics policies and integrity standards of Nanyang Technological University and that the research data are presented honestly and without prejudice.



16/08/2021

. . . . . .
Date

. . . . . .
Aw Jong Ghut Ashley

# Supervisor Declaration Statement

I have reviewed the content and presentation style of this thesis and declare it of sufficient grammatical clarity to be examined. To the best of my knowledge, the thesis is free of plagiarism and the research and writing are those of the candidate's except as acknowledged in the Author Attribution Statement. I confirm that the investigations were conducted in accord with the ethics policies and integrity standards of Nanyang Technological University and that the research data are presented honestly and without prejudice.

17/08/2021

. . . . . .
Date

. . . . . .
Wan Yue

# Authorship Attribution Statement

Please select one of the following; *delete as appropriate:

*(B) This thesis contains material from 2 paper(s) published in the following peer-reviewed journal(s) / from papers accepted at conferences in which I am listed as an author.

Chapters 3 and 4 are published as Aw, J.G.A., et al., Determination of isoform-specific RNA structure with nanopore long reads. Nat Biotechnol, 2021. 39(3): p. 336-346.

The contributions of the co-authors are as follows:

- Dr Shaun W Lim and Finnlay R P Lambert performed most of the experiments.
- Dr Jia Xu Wang cultured and provided the H9 cell lines.
- Wen Ting Tan did the experiments to validate the result of structural and transcription efficiency changes.
- Dr Yang Shen provided the analysis of the TRip-seq data.
- Dr Yu Zhang provided the analysis for the SNV and RBP data.
- Dr Chenhao Li provided the initial advice on coding and analysis of data.
- Dr Leah A Vardy did the work on Trip-seq.
- Prof Meng How Tan, Dr Pornchai Kaewsapsak and Dr Sarah B Ng were involved in the discussion of the project.
- Prof Niranjan Nagarajan guided our work and the analysis of data.
- Prof Yue Wan provided the initial project direction and wrote the drafts of the manuscript.
- I had developed the analytical pipeline for PORE-cupine, analyzed all the data generated by nanopore sequencing and generated the data for the initial work.

Chapter 5 is published as Siwy Ling Yang, L.D., Danielle E. Anderson, Yu Zhang, Ashley J Aw, Su Ying Lim, Xin Ni Lim, Kiat Yee Tan, Tong Zhang, Tanu Chawla, Yan Su, Alexander Lezhava, Andres Merits, Lin-Fa Wang, Roland G. Huber, Yue Wan1, Comprehensive mapping of SARS-CoV-2 interactions in vivo reveals functional virus-host interactions. Nature Communications, 2021.

The contributions of the co-authors are as follows:

- The first and corresponding authors worked on the direction of the project.
- Prof Yue Wan wrote the drafts of the manuscript.
- Su Ying Lim did the experiment on SPLASH.
- Xin Ni Lim and Kiat Yee Tan infected the cells with the virus and did the structure probing modifications.
- Prof Roland G and Danielle E predicted the secondary structure of the sgRNAs with the aid of PORE-cupine reactivity
- Prof Lin-Fa Wang provided the virus used in the project.
- I have sequenced, processed and analysed the sgRNA reads sequenced by direct RNA sequencing and generated the results based on PORE-cupine

16/08/2021

. . . . . .

Date

. . . . . .

Aw Jong Ghut Ashley

# Acknowledgements

I would like to thank Prof Wan Yue for providing me with her time and effort to guide me through my post-graduate study.

I would like to also thank Prof Francesc Xavier Roca Castella for taking the time to provide me with the feedback to improve my writings.

I would also like to thank my colleagues in Wan Yue's lab, Shaun, Finn, Jiaxu, Natalie, Wen Ting, Siwy Ling, Zhang Yu, Zhang Tong, Xin Ni, Antson and all the interns for all the discussion and the work that they have done.

# Table of Contents

# Summary

RNA can fold into complex structures that can perform specific biological functions and the ability to characterise the structures of the RNA is the key in understanding its functions. Therefore, various groups have developed several methods that utilizes enzymatic or chemical structure probes and coupling with high-throughput short-read sequencing to obtain the structural information from in-vitro and in-vivo samples. However, high-throughput short-read sequencing has its limitation, where it is difficult to accurately generate the structural information for gene with multiple isoforms and the structural information between reads are lost. Therefore, we developed PORE-cupine, a method that uses direct RNA sequencing, a high-throughput long-read sequencing technology to detect the modifications caused by NAI-N3, a single-stranded chemical probe. Using PORE-cupine, we showed that shared sequences in different transcript isoforms of the same gene can fold into different structures. We also demonstrate that structural differences between transcript isoforms of the same gene lead to differences in translation efficiency. We also apply PORE-cupine to investigate the structural difference of the RNA of the wild type and Δ382 strain of SARS-CoV2. Similarly, we found that shared sequences in different subgenomic RNA shows structural differences, highlighting the importance of long-read sequencing for obtaining phase information. Lastly, we have shown that besides NAI-N3, PORE-cupine can detect other structure probing compounds, like 1AI and CMCT.

# Chapter 1 Introduction

Ribonucleic acid (RNA) is a biopolymer that consists of a nitrogenous base and a ribose sugar and a phosphate group that is synthesised in a single-stranded state, unlike its closely related biopolymer, deoxyribonucleic acid which primarily exists as a double-stranded chain. It has been long thought that RNAs would first fold into themselves to form stable secondary structures, such as double helix, bulge, hairpin loops and junctions, which can further interact to form intricate less stable tertiary structures[1] (**Figure 1.1**). However, with the various number of studies done on the RNA structure in the past decades, it was seen that the folding of the RNA not just depends on the base pairing of the nucleotides, environmental and cellular factors can have a role in determining its structure[2,3]. With the increased understanding of RNA, t also found that RNA can play an important role in the regulation of biological processes inside cells [4-8]. For example, a single mutation in the 5' untranslated region in RB1 can result in increased structure heterogeneity, and the shift in the proportion of structure is highly associated with retinoblastoma and breast cancer[5,9]. Additionally, a mutation in the Tau gene destabilizes a stem-loop to result in a change of the alternative splicing patterns of the transcript, and frontotemporal dementia and Parkinsonism[10]. These examples highlight the importance of the roles that the RNA structures can have in the regulations of cellular processes to maintain the normal functions of the cells. As such, there is now great interest to determine the structures of the RNAs to understand RNA function ad regulation in various cellular processes.

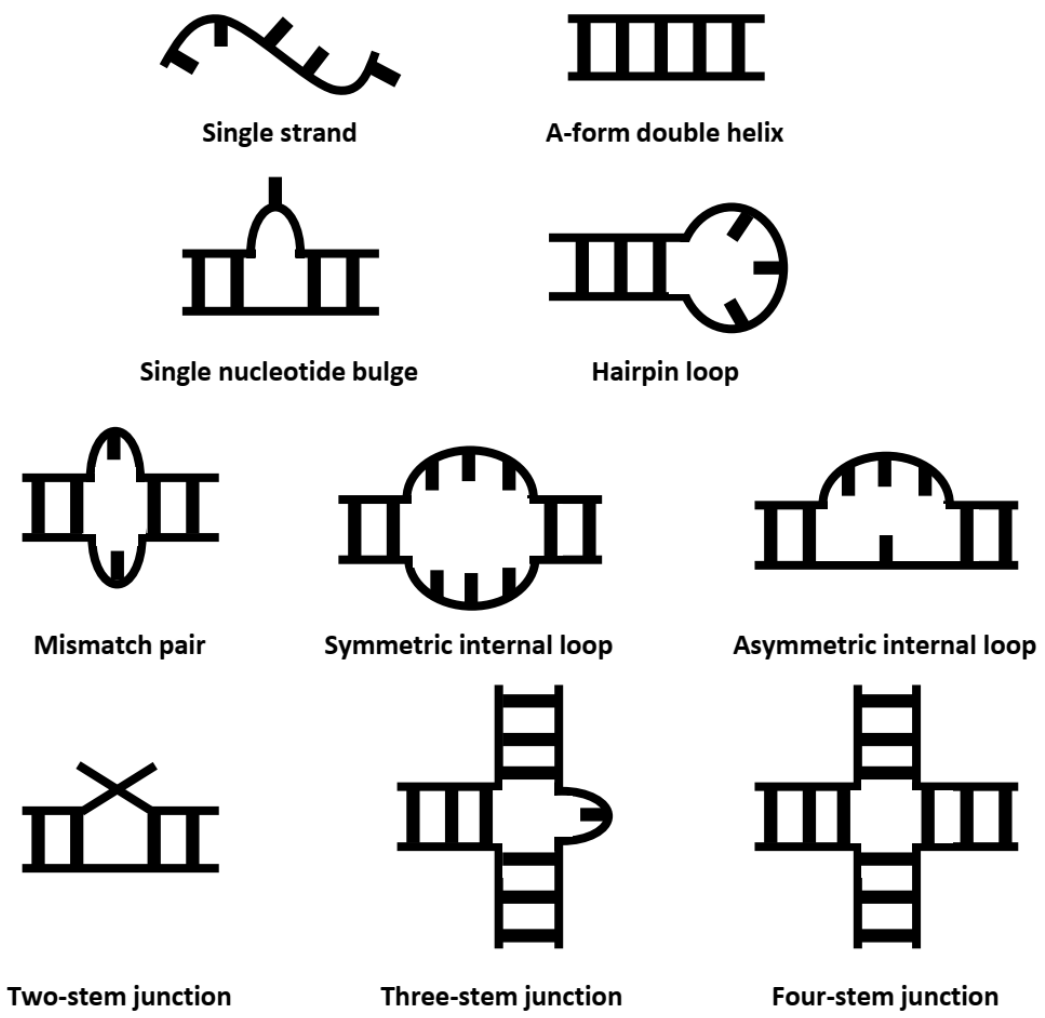


**Figure 1.1. Secondary structures of RNAs.**

## 1.1 Methods to study RNA structures

Different methods have been adapted or developed to determine the structures if the RNA[7,11,12]. These methods can be categorised into either computational, experimental or a combination of both methods to predict or determine RNA structures, with varying accuracy and limitations. In computational approaches, primary sequence from the RNAs are used to predict its secondary structures and there are three general categories, phylogenetic, thermodynamic and machine learning approaches.

In phylogenetic approaches, such as RNAalifold[13], Tfold[14] and CaCoFold[15], RNA secondary structures are predicted by measuring the covariations derived from multiple homologous genes across different species. This approach assumes that RNA structures that are essential for the transcript, have a selection pressure to preserve the RNA base-pairing through evolution, resulting in the preservation of base pairing by either preventing the mutations or having compensatory mutations to maintain the structure. Therefore, by identifying regions in the RNA with or without compensatory mutations, RNA base paring and its secondary structure can be predicted. However, this approach is limited by the need of having multiple homologous sequences across organisms, as it requires the use multiple sequence alignment to detect the conserved structures [13-16].

Thermodynamic approaches, such as Mfold[17] and RNAfold[18], predict RNA structures by searching for structures that have the lowest predicted free energies. It is assumed that structures with the lowest predicted free energies, will be the most stable and the most probable structure. However, these methods are usually used for shorter RNA, as the accuracy of the predicted structures decreases and the time required increases exponentially with increasing RNA length[17,18].

Machine learning approaches, such as CONTRAfold[19] and algorithm that combines with thermodynamic approach[20], aims to predict RNA structures by using models trained on known RNA structure data. However, the accuracy of machines learning approaches is dependent on the dataset used for training, and current dataset have limited RNA structures are limited.

In addition to computational predictions of RNA structures, experimental biophysical methods such as X-ray crystallography[21,22], nuclear magnetic resonance (NMR)[23,24], cryogenic electron microscopy (Cryo-EM)[25,26] and low angle X-ray scattering

(SAXS)[27,28], have been used to study the structures of RNAs. X-ray crystallography captures the three-dimensional structures of a crystallized sample by using the diffraction pattern of the X-rays. Although this method can generate three-dimensional structure models at the sub-atomic resolution, only a few RNAs can be crystallized in solution[21,22]. While NMR can detect the dynamics of the structure of RNA across time, in the range of milliseconds to seconds, there is a size constraint to the samples it can model. In theory, an RNA molecule with a length of up to 1,000 nucleotides can be studied with this method. However, the largest RNA structure modelled by NMR so far is around 200 nucleotides[23,24]. Cryo-EM can reconstruct the three-dimensional structures of a sample frozen in a thin sheet of ice. Using an electron microscope, thousands or millions of images of an RNA are taken and a three-dimensional model can be constructed. Cryo-EM does not require complicated sample preparation, allowing for a vast number of samples to be imaged. However, samples smaller than 50 kDa and flexible regions of samples are not easily imaged at high resolution[25,26]. In addition to Cryo-EM, SAXS can also generate structure models by measuring the scattering of the X-ray beam through the sample. SAXS can be used to visualize a large variety of samples, from frozen to near-native states. However, this method can only generate low-resolution models[27,28].

To probe the secondary structures of RNAs, biochemical methods utilizing probes to modify or cleave double-, single-stranded or solvent-accessible bases regions in a RNA[29-34]. Thus by determining the regions that are affected by the probes, the secondary structure of the sample can be inferred. The main advantage of using biochemical approaches is in the versatility of the protocols, whereby they can be applied to most RNAs, unlike the biophysical methods whereby only certain sample types can be visualized. To probe RNA structures across many RNAs simultaneously,

several groups have combined biochemical structure probing with high throughput sequencing[35-48].

## 1.2 Structure probing of RNA structures

The principle behind structure probing, requires the incubation of the structure probes with the folded RNA, that will label the RNA by either cleaving or modifying the bases of the RNAs[29-34]. Various groups have identified and used probes that can be categorised into three types, nucleases, small molecules or hydroxyl radicals to determine double-, single-stranded or solvent-accessible regions of an RNA. Single stranded RNA nucleases, like nuclease S1 or RNase I, and double stranded RNA nuclease, like RNase V1, are used separately to label the RNA by digestion, and base-paring status of the unaffected regions can be determined. Small molecules, like carbodiimide metho-p-toluenesulfonate (CMCT), dimethyl sulfate (DMS) or kethoxal, labels the RNAs by modifying specifics bases of the nucleotides (**Figure 1.2**). These modifications would affect the reverse transcriptase which results in the truncation of the products. Traditionally, small molecules probes are limited by the specific nucleotides that it can modify, however Weeks' lab was able to solve this issue by developing a novel class of compounds, selective 2'-hydroxyl acylation and primer extension (SHAPE), N-methylisatoic anhydride (NMIA), that can modify all single stranded bases[2]. Hydroxyl radical labels solvent-accessible regions by cleaving the backbone of the RNA, and the protected regions can be determined. The positions of the labels on the RNA can then be detect by performing reverse transcription (RT) and resolving the RT products on a sequencing gel, where the bands can be quantitated to determine the positions[29-34] (**Figure 1.3**). Although structure probing by footprinting have been used to study the structures from various classes of RNA, like tRNA and

rRNA, it remains a low throughput, time-consuming process as structure probing by footprinting is limited by the short regions that it can probe at a single time, around 150-200 nucleotides in a single lane, thus the need for multiple runs to obtain structure information for the whole transcript. Therefore, to increase the efficiency of structure probing, various groups have coupled high-throughput short-read sequencing with structure probing[35-48].

The process of using high-throughput short-read sequencing is similar to traditional structure probing, where various enzymatic or chemical probes can be used for different protocols[35-46]. For example, Parallel Analysis of RNA Structure (PARS)[29] uses short-read sequencing to detect the undigested regions after incubating with the nuclease V1 or S1. DMS-seq[36], icSHAPE[37] and CIRS-seq[46] uses the sequencing data to detect truncated reads caused by chemical probes on the RNA during RT. Whereas SHAPE mutational mapping (SHAPE-MaP)[38], DMS-MapSeq[47] and icSHAPE-MaP[48], detects the modifications by observing the mutations found in the sequenced reads. With the detection of the labels, the structures of the RNA can then be predicated with software such as ViennaRNA[49], RNAstructure[50] or SHAPEMapper2[51]. By using high-throughput sequencing to detect the modifications, we can probe for the *in-vivo* structures of the transcriptome, across thousands of transcripts simultaneously, allowing us to answer questions that were not previously possible. For example, it was found that RNA is less structured in cells when compared to *in-vitro* folded RNA[36-38].

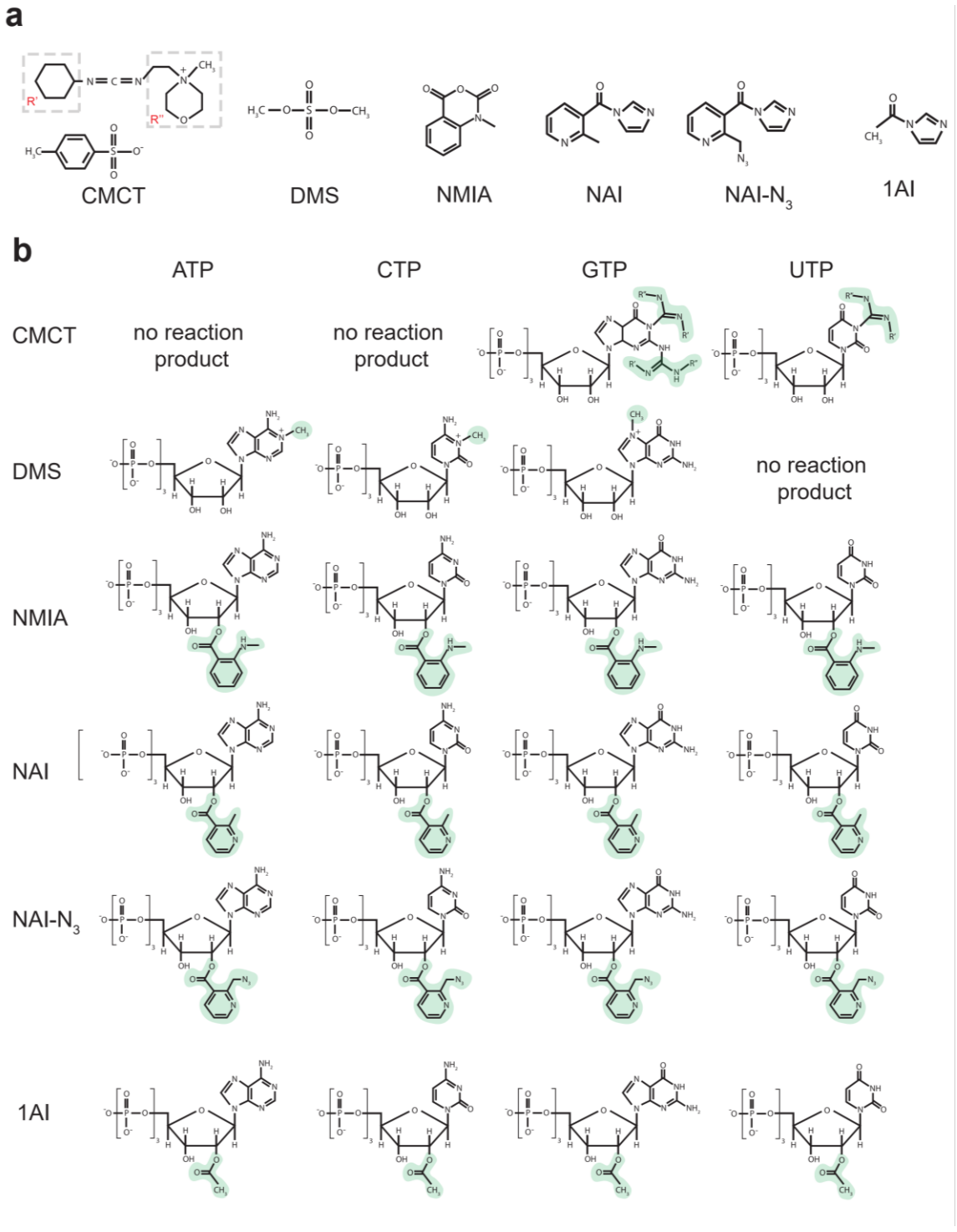


**Figure 1.2. Chemical structures of RNA structure probing compounds, associated reaction products. a,** Chemical structures of RNA structure probing compounds. Side chains for the carbodiimide of CMCT are highlighted and abbreviated as R' and R'' for part (**b**). **b,** RNA nucleotide triphosphates with chemical adducts formed from reaction with structure probing compounds. Adducts are highlighted in green. Figures are adapted from *Aw. at el*[52].

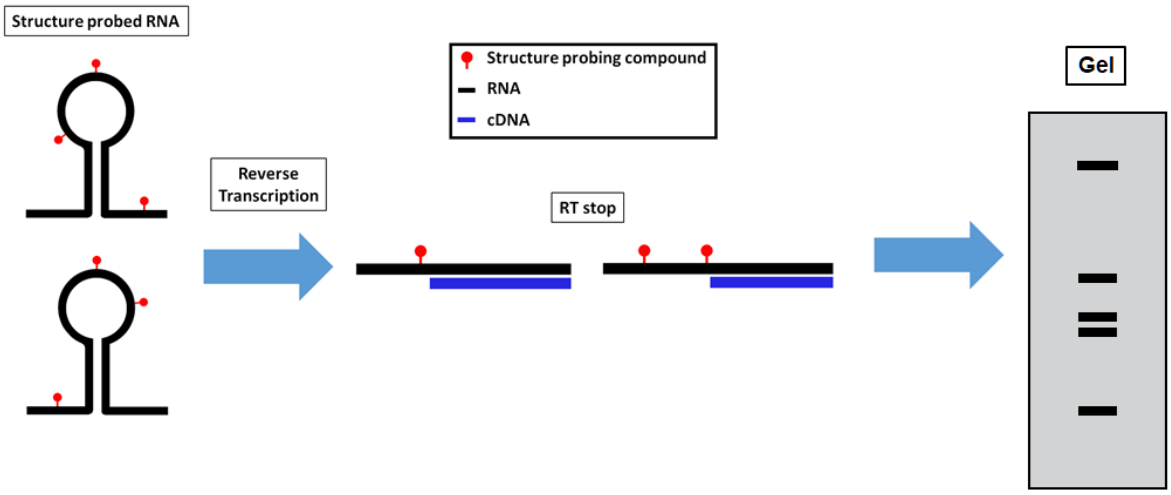


**Figure 1.3. Workflow of RNA structure probing.** The RNAs that are modified with structure probing compounds are amplified with reverse transcription PCR. During reverse transcription, the modifications on the RNA would cause the reverse transcriptase to stop, resulting in the truncation of RT products. The truncated product will be amplified and visualized on a gel, where the locations of the modifications can be determined[29-31].

## 1.3 Limitations of short-read high-throughput sequencing resulting in inaccurate structure profile

Although high-throughput structure probing has advanced the field of studying RNA structures, most of the initial studies were carried out using second generation, short-read sequencing, such as Illumina or SOLID sequencing, which has its limitations[53]. During the preparation of short-read sequencing libraries, the RNAs are first fragmented before being sequenced, resulting in the loss of connectivity between the sequenced reads. Such loss causes two major issues: (1) it increases the ambiguity of the sequenced reads that are mapped to genes with multiple isoforms, and (2) modifications between sequenced reads cannot be phased.

With the fragmentation of the RNAs, the alignment of reads that do not span across unique sequences of the genes that have multiple isoforms may not be assigned to the correct one but assigned randomly to one of them. Currently. there are no easy solution to this issue, thus current high-throughput structure probing bypasses this limitation by aggregating the mutational results of all isoforms to generate an average RNA structure at the gene level. However, using the aggregate signals causes noise and results in a more inaccurate structure.

Although the phasing of the structure information between reads does not affect the accuracy of determining the structure at the transcript level, having this ability allows us to obtain structural information at the individual strand level, where we can begin to examine the structural heterogeneity of a transcript, and potentially allowing us to filter for structures that are functionally important. We hypothesized that replacing short-read sequencing with long-read sequencing will address the issues stated above and enable us to obtain more accurate structure data.

## 1.4 Isoforms in the transcriptome

Most RNA in eukaryotic cells undergoes extensive splicing that removes the introns before becoming mature and it is estimated that around 35% of human genes are alternatively spliced[54]. Alternatively splicing results in generating a pool of isoforms that originate from the same gene but having a slightly different combination of exons or introns (**Figure 1.4**)[55]. The limitation of using short-read sequencing to study transcriptome-wide data becomes more evident when the data contains multiple isoforms, as the reads do not span across the whole transcripts making it difficult to determine the isoforms where they had originated from. Therefore, various statistical

methods had been developed to quantify the expression of the isoforms[56,57]. However, the methods developed for the quantification of expression are not suitable for obtaining the reactivity from the isoforms and there is no available analytical pipeline that could achieve it, current strategies require the accurate assignment of each sequenced reads to the transcripts. Therefore, we explore the use of long-read sequencing to detect structure probing modifications on RNA.

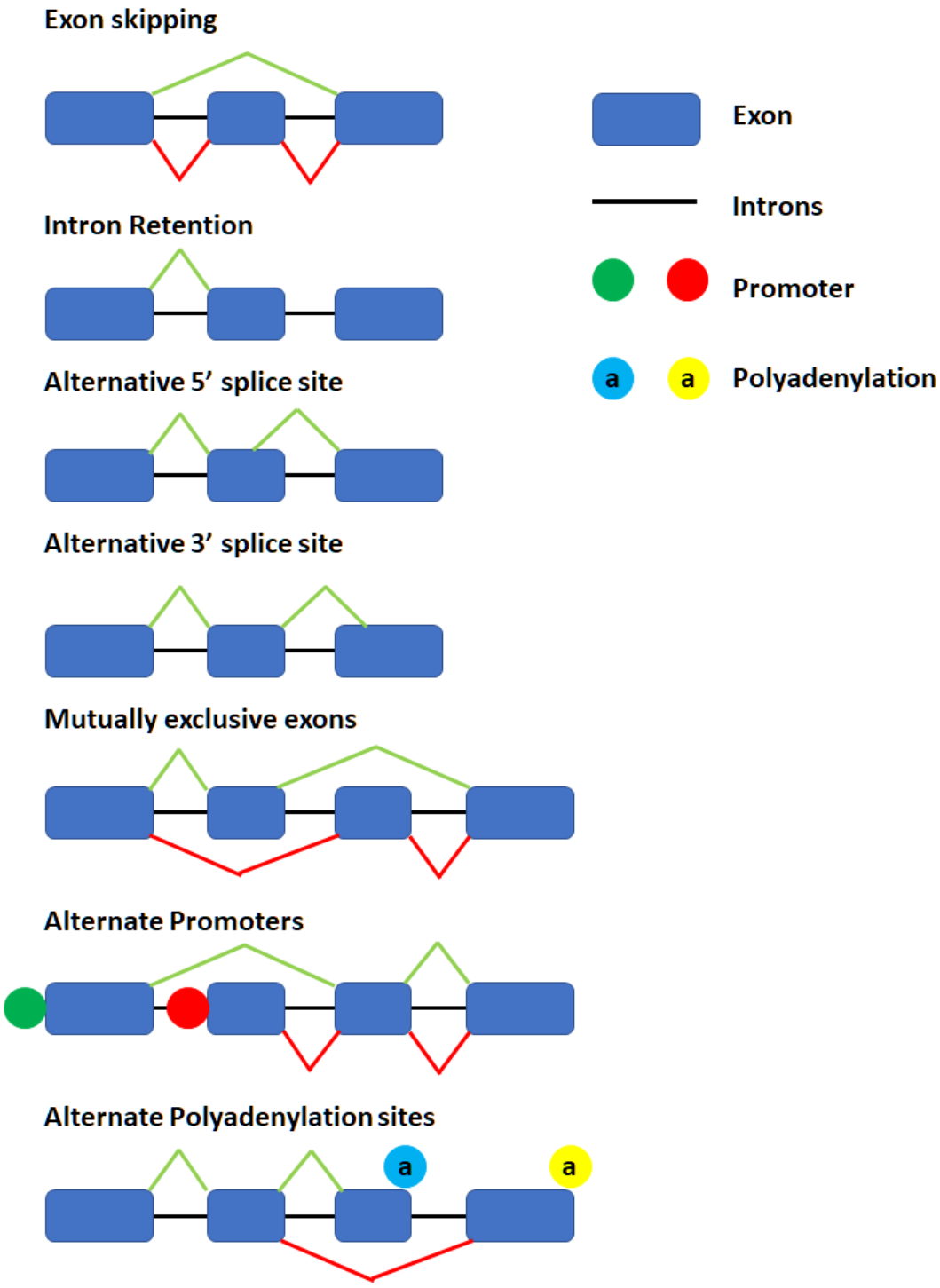


**Figure 1.4. Different types of alternative splicing**

## 1.5 High throughput long-read sequencing

Currently, two sequencing technologies are capable of long-read sequencing, developed by Pacific Biosciences[58] and Oxford Nanopore Technologies[59] respectively. The principles used by the two long-read sequencing technologies differ greatly from each other and each system has its strengths and weaknesses.

### 1.5.1 Pacific Biosciences single-molecule, real-time (SMRT)

Pacific Biosciences has developed SMRT, where it uses the properties of the synthesis of DNA to sequence the reads. A DNA polymerase located in a sequencing unit uses the sequencing read as the template. The DNA polymerase synthesizes a new strand with labelled nucleotides, that have a distinct fluorophore attached to each of the four bases. When the nucleotide is incorporated into a new strand, the fluorophore would be released and detected by the sequencer and the detected signals would be converted to the sequences (**Figure 1.5a**). Theoretically, SMRT can sequence a read up to 25 kb in length and its sequencing accuracy is in between nanopore sequencing and Illumina sequencing, where Illumina sequencing have the highest accuracy[58]. SMRT is also capable of sequencing samples without any amplification, but it cannot sequence RNA directly unlike the technology from Oxford Nanopore Technologies.

### 1.5.2 Oxford Nanopore Technologies nanopore sequencing

Oxford Nanopore Technologies has developed nanopore sequencing, where it uses the perturbation of the current caused by a strand of nucleic acids threading through

a protein pore for sequencing[59,60]. In the sequencer, a constant current is applied across the pore. When a nucleic acid is threading through, it blocks the flow of current which causes a change in the measured current. This blockage of current creates a distinct signal, where it can be converted to the sequence of the nucleic acid (**Figure 1.5b**). Theoretically, nanopore sequencing does not have an upper limit to the sequencing length and a team at the University of Nottingham have managed to sequence a DNA strand longer than 2 Mb[61,62]. Although nanopore sequencing is the only technology that can sequence RNA directly, it has the lowest accuracy among all sequencing methods.

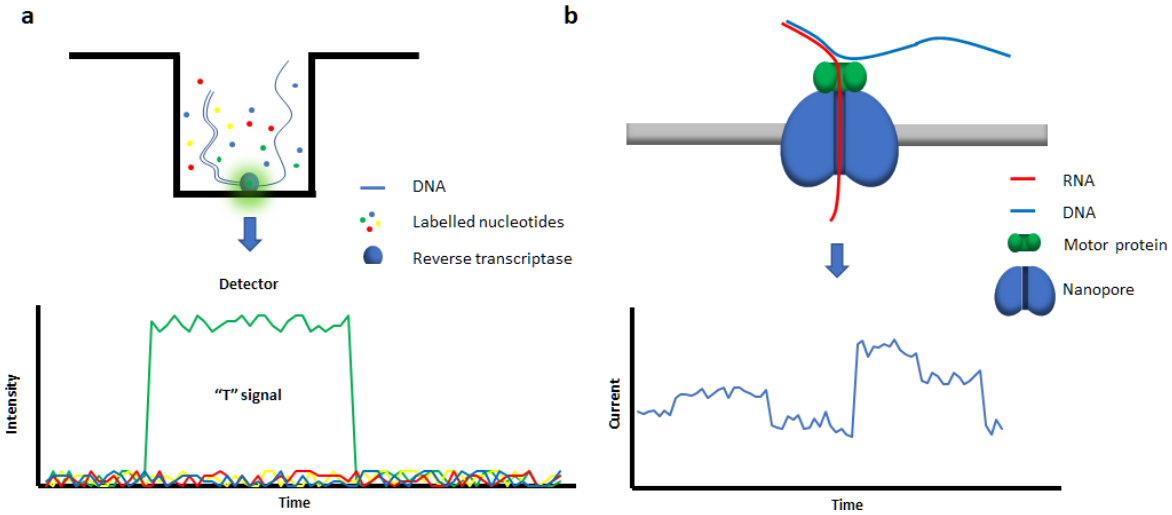


**Figure 1.5. Principles of SMRT and nanopore sequencing. a.** Upper, a diagram of how SMRT from Pacific Biosciences sequences a DNA sample by synthesis of a new DNA strand. Each nucleotide is labelled with a unique fluorophore, indicated by blue, green, red and yellow, respectively for A, T, C, and G[59,60]. When a nucleotide is being incorporated into a new strand of DNA by the RT enzyme, it would be close to the bottom of the well and will be detected by the sequencer. Bottom, shows an example of the detected fluorescence during sequencing. The detected fluorescence will be converted to the sequences of the template strand. **b.** Upper, a diagram of how direct RNA sequencing from Oxford Nanopore Technology works. Using a motor protein to slow down the speed of the RNA treading through the pore, the sequencer can measure the change in current flowing through the pore. The cDNA strand in direct RNA sequencing is to stabilise the RNA during sequencing. Bottom, shows an example of the current detected during sequencing. The measured current can be converted in sequences by using a basecalling software[61,62].

## 1.5.3 Selection criteria for long-read sequencing

Our selection criteria were based on the ability to resolve the issues that were stated in the previous section. Both SMRT and nanopore sequencing were shown to have the ability to achieve high-throughput long-read sequencing[58,59], which will allow us to obtain reads that span across the isoforms, solving the first issue. However, structure probing modifications on the RNA is known to cause truncations during

reverse transcription, and the library preparation of SMRT requires the samples to undergo reverse transcription[58], resulting in a truncated read that only contains one modification. Due to the ability of nanopore sequencing to sequence RNA strands directly, we hypothesized that it can sequence long reads with multiple modifications, allowing us to examine the heterogeneity of the transcripts. Therefore, we have selected direct RNA sequencing from Oxford Nanopore Technologies for the development of our protocol.

## 1.6 Selection of the RNA transcripts used for the development

The selection of the RNA transcripts is one of the key factors in determining the success of the development of our protocol. Therefore, we looked for RNAs that fulfil either of the criteria: (1) Having a well-defined secondary structure, based on either in vivo or in vitro data. (2) Able to fold into different structures under specific conditions. Among the various classes of RNA, Ribozyme[63-70], ribosomal RNA[38,71-73] and Riboswitch[74-82] was able to fit our criteria.

Ribozyme is a class of RNA that can fold into a specific structure that recognises specific sites on an RNA to perform a catalytic function. Among the ribozyme, there are a few well-known types of RNA, like the self-splicing group I and II introns, RNase P, spliceosome and Hammerhead ribozyme. We focused on the self-splicing group I intron from *Tetrahymena thermophila*, where studies have shown that it has a stable structure that can be folded in-vitro, making it an excellent candidate for the initial evaluation of the development of our protocol with the aid of the crystal structure[67] (**Figure 1.6**).

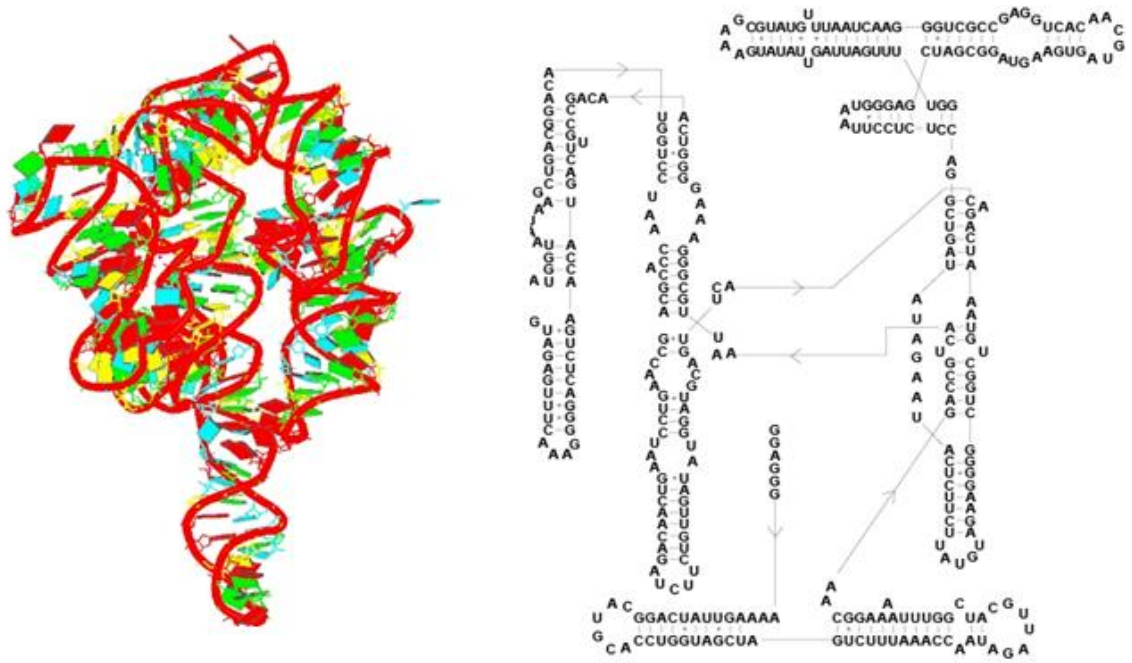

**Figure 1.6 Structure of Tetrahymena ribozyme RNA.** Crystal structure (left) and secondary structure (right) of Tetrahymena ribozyme RNA. The crystal structure is extracted from Guo at el[67].

Ribosomal RNA (rRNA) is the most abundant RNA found in the cell and it is essential for the synthesis of protein. It has also one of the most well-defined in-vivo folded RNA, where multiple groups have generated the secondary structure map and crystal structures[73,83]. rRNA consist of 5s, 5.8s, 18s and 28s in eukaryotic cells and 5s, 16s and 23s in prokaryotic cells, where the rRNAs binds to the ribosomal protein to form the ribosome. Due to its abundance and relativity long length, we selected 16s rRNA to evaluate for the accuracy of in vivo structure probing.

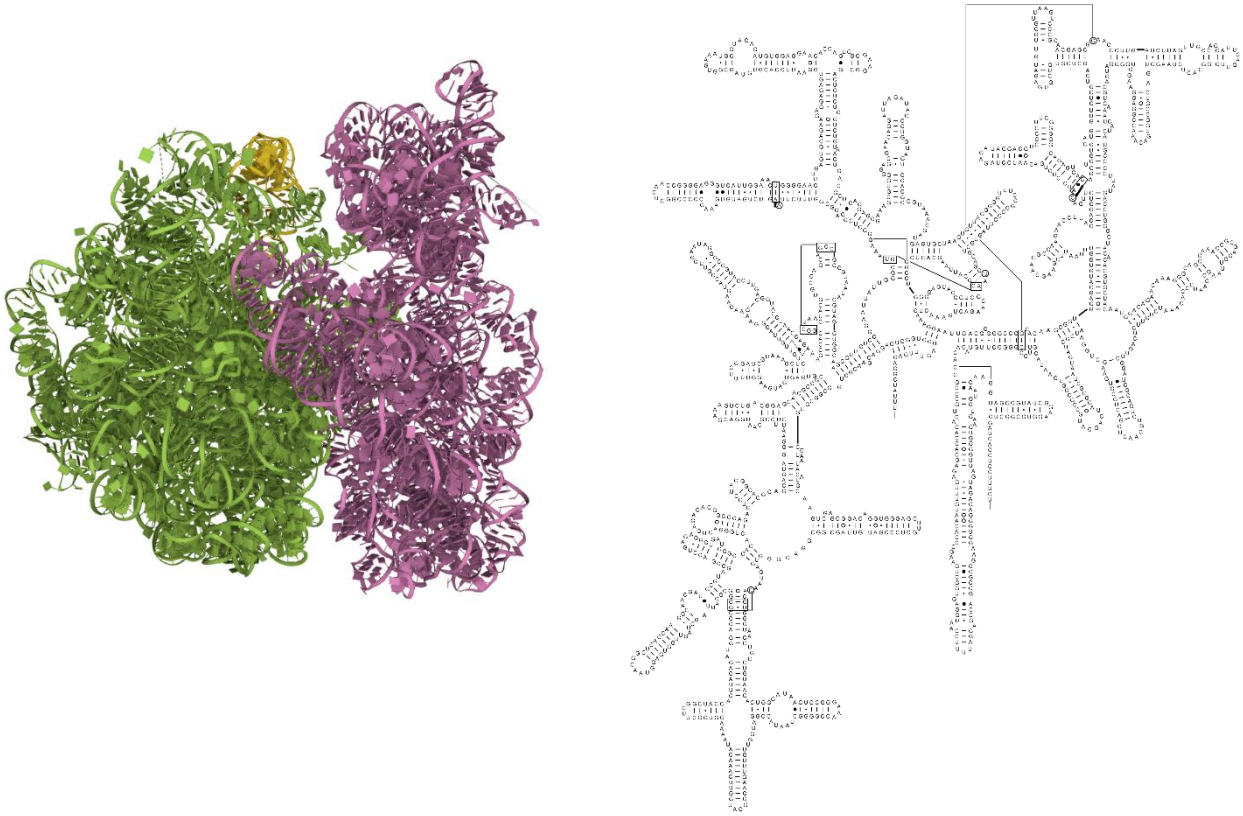

**Figure 1.7. Crystal structure of rRNA.** The model shows the crystal structure (left) of rRNA, 23s (green), 16s (pink) and 5s (yellow) and the secondary structure of 16s rRNA (right). The crystal structure and the secondary structure is extracted from Zhang at el[73] and Cannone at el[83] respectively.

Lastly, we looked at Riboswitch, a class of RNA regulatory element of an mRNA, where it can recognise and bind to a small molecule, at the aptamer region, that results in a change of structure leading to a change in the regulation of the expression of a gene. Various groups have identified multiple Riboswitches and we have select thiamine pyrophosphate (TPP) riboswitch from *Bacillus subtilis*, which is known to have a change in the structure in the presence of TPP. This will allow us to evaluate the accuracy of the detection of the structural changes in the RNA.

## 1.7 Objectives

I hypothesized that using direct RNA sequencing would allow us to obtain transcriptome-wide RNA structures, allowing us to increase our understanding of how the differences in RNA structures might affects the RNA. The main aim of this thesis is to develop a protocol that utilizes long-read sequencing technology and apply the protocol to answer questions that were not possible. My aim for this thesis consists of the following objectives:

1. To develop an analysis pipeline that can detect modifications using Oxford Nanopore direct RNA sequencing
2. To obtain transcriptome-wide isoform-specific structural profiles in a human embryonic stem cell line
3. To determine the structural differences between the different sub-genomic sequences within or across two SARS-CoV-2 strains
4. To further optimise our pipeline

## 1.8 ORGANISATION OF THIS THESIS

This thesis will address the topic of the importance of having the ability to detect RNA structures.

The second chapter lists the methods and equipment used for the experiments done in the thesis.

The third chapter will describe the development of a novel pipeline of For chemical utilized probing interrogated using Nanopores (PORE-cupine), a method that uses one-class support vector machine (SVM) to detect biochemical structure probing modifications by using the raw signals from direct RNA sequencing.

The fourth chapter will summarize the structure differences between isoforms found in a human embryonic stem cell line (hESC), H9. Where the structural data of the different isoforms are obtained with PORE-cupine.

The fifth chapter will describe the work of using PORE-cupine to study the structural differences of SARS-CoV-2 RNA isolated from patients in Singapore, wild type and Δ382 strain.

The sixth chapter will summarise the work of further optimization of PORE-cupine, where we explore the detection accuracy of different chemical probes and an alternative method of processing the raw signal to detect the modifications.

The final chapter will conclude the work that I have done in this thesis, discuss on results and the potential future direction of my work.

# Chapter 2 Methods

## 2.1 RNA modifying reagents

CMCT, NAI, NMIA and DMS were purchased from Sigma Aldrich (**Figure 1.3**). NAI-N3 was synthesized as previously described from ethyl 2-methylnicotinate in 4 steps, as in Spitale et al[37].

## 2.2 *In vitro* transcription, folding and *in vitro* structure probing

RNA was transcribed from PCR-amplified inserts using the Hiscribe T7 high yield synthesis kit (NEB). The RNA of interest was folded, and structure probed in the presence or absence of ligand (TPP). Depending on the solvent for the RNA-modifying chemical, DMSO or water was added to the negative control. CMCT, NAI and NAI-N3 were added to final concentrations of 100mM, while NMIA and DMS were used at final concentrations of 20mM and 5%(v/v) respectively. DMS reactions were quenched with 30% β-mercaptoethanol in 0.3M sodium acetate. Reactions were column-purified (Zymo research) and re-suspended in nuclease-free water.

## 2.3 RNA footprint analysis

To determine sites of modifications along an RNA, an RT primer (IDT) was designed around 20bp downstream of the region of interest. Primers were radiolabeled with $P^{32}$and purified on a 15% TBE-Urea PAGE gel. The purified labelled primer (1µl) was then incubated with 500ng of RNA (in 5.5µl) for reverse transcription and running on a 8% TBE-Urea PAGE sequencing gel. Gels were dried for 2 hours on a vacuum gel drier before exposure to a phosphorimager plate for 24 hours. The phosphorimager plates were imaged on the Amersham Typhoon 5 Biomolecular Imager (General Electric) (**Figure 2.1**). Gel images were quantified using the software Semi-automated footprinting analysis (SAFA)[31].


**Figure 2.1 Gel image of a portion from the structure probing experiment of Tetrahymena ribozyme.** The intensity of the bands would be quantitated with SAFA.

## 2.4 Human and bacterial cell culture with *in vivo* SHAPE modification

hESCs (H9) were obtained from labs in the Genome Institute of Singapore and cultivated in feeder-free conditions with mTESR basal media supplemented with mTESR supplement (Stemcell Technologies).

For *in vivo* SHAPE modification, cells were rinsed once on the plate with room temperature PBS (Thermo Fisher Scientific) before being incubated with Accutase (Stemcell Technologies) for 10 mins at 37$^{o}$C to dissociate cells. The cells were washed with PBS and spun down at 400g for 5 min before resuspension in 950µl of PBS in a 1.7ml Eppendorf. NAI-N3 (2M in 50µl) in DMSO(+) or DMSO(-) was then added to two separate cell suspensions and immediately mixed by inversion, before incubation at 37$^{o}$C at 10rpm (Model 400 hybridization incubator, Scigene) for 5 mins. After the incubation period, the cell suspensions were immediately spun down at 400g for 5 mins at 4$^{o}$C. The supernatant was removed before total RNA was isolated with the addition of TRIzol reagent (Thermo Fisher Scientific).

*Bacillus subtilis* strain 168 was grown in LB media at 37ºC to an $OD_{600}$ = 0.6. Cells were harvested by centrifugation at 3000rpm for 5 mins, pelleted and washed in PBS, before being treated with 100mM NAI-N3 as in the H9 example above for 5 mins at 37ºC. The cells were pelleted after incubation and then re-suspended in bacterial lysis buffer (1% SDS, 8mM EDTA, 100mM NaCl) and lysozyme (final concentration: 15mg/ml), incubating for 15 mins at 37ºC. Total RNA was then isolated with the addition of TRIzol reagent.

## 2.5 Direct RNA sequencing library preparation

Direct RNA sequencing libraries were constructed using an input of 500ng of RNA for single templates or 1µg of mRNA isolated from H9 hESC total RNA. For the training and test set of RNAs, a total of 14 RNAs were structure probed individually *in vitro*, and then pooled together and sequenced as a mix. Out of these 14 RNAs, 11 were used as the training set for the SVM and the remaining 3 were used as a test set.

H9 mRNA was isolated using the Poly(A) purist MAG kit (Thermo Fisher Scientific). Oxford Nanopore Technologies' direct RNA library preparation kit SQK-RNA001 was used for all sequencing runs in this study. All preparation steps were followed according to the manufacturer's specifications, except for the omission of a single reverse transcriptase step. The libraries were loaded onto R9.4.1 or R9.5 flow cells and sequenced on a MinION device. For the *in vitro* and 16S rRNA templates, the RTA DNA adapter from the sequencing kit in the first step was replaced with a DNA adapter complementary to the 3' end of the RNA. The sequences of these adapters are detailed in **Table 2.1**. The sequencing parameters were modified for all runs, and the specific changes can be found described in the document "Modified Minkown parameters" found in our GitHub repository: http://github.com/awjga/PORE-cupine.

| Primer name | Primer Sequence (5'->3') | Comments |
|---|---|---|
| Oligo A | /5PHOS/GGCTTCTTCTTGCTCTTAGGTAGTAGGTTC | General primer for annealing with Oligo B to form adapter replacing RTA |
| Oligo B TPP | GAGGCGAGCGGTCAATTTTCCTAAGAGCAAGAAGAAGCCACGGATGGAC | Oligo B for annealing to Oligo A |
| Oligo B Tetrahymena | GAGGCGAGCGGTCAATTTTCCTAAGAGCAAGAAGAAGCCCGAGTACTCC | Oligo B for annealing to Oligo A |
| Oligo B Lysine riboswitch | GAGGCGAGCGGTCAATTTTCCTAAGAGCAAGAAGAAGCCAGAAAGAGCG | Oligo B for annealing to Oligo A |
| Oligo B 16S | GAGGCGAGCGGTCAATTTTCCTAAGAGCAAGAAGAAGCCAGAAAGGAGG | Oligo B for annealing to Oligo A |
| Tetrahymena Kinase R273 | CTTCCCCGACCGACATTTAG | Primer for kinasing for structure probing of Tetrahymena RNA |
| Lys Kinase R191 | AGAAAATGATTTCTTGACAGCC | Primer for kinasing for structure probing of Lysine riboswitch RNA |
| 16S Kinase R274 | ACGCATCGTCGCCTTGGTGA | Primer for kinasing for structure probing of 16S RNA |
| 16S Kinase R488 | CTTTCTGGTTAGGTACCGTC | Primer for kinasing for structure probing of 16S RNA |
| 16S Kinase R871 | GCGGAGTGCTTAATGCGTTA | Primer for kinasing for structure probing of 16S RNA |
| RPS29 Kinase R209 | CCGGATAATCCTCTGAAGGA | Primer for kinasing for structure probing of RPS29 RNA |
| AdoCbl Kinase R218 | TACGACGGTAAAGATGCTGT | Primer for kinasing for structure probing of AdoCbl RNA |

**Table 2.1 Primers for direct RNA sequencing and structure probing**

# 2.6 TrIP sequencing sample preparation

Samples were prepared according to the published protocol[84] (**Figure 2.2**). H9 cells from a 15cm plate were treated for 10 min with 100µg/ml cycloheximide at 37°C. The cells were next washed with warm PBS, dissociated with trypsin, and neutralized with ice cold media containing FBS (all supplemented with 100µg/ml cycloheximide). Next, they were pelleted before re-suspending in 1x RSB buffer (10mM Tris-HCl pH 7.4, 150mM NaCl, 15mM $MgCl_2$) with 200µg/ml cycloheximide, lysed in lysis buffer (10mM Tris-HCl pH 7.4, 150mM NaCl, 15mM $MgCl_2$, 1% Triton-X, 2% Tween-20, 1% Deoxycholate), and incubated on ice for 10 mins. Centrifugation at 12,000*g* for 3 mins removed the nuclei, and the supernatant was removed for a subsequent centrifugation.

Equal OD units were loaded onto a linear 10%-50% sucrose gradient that was made using the 107 Gradient Master Ip (BioComp Instruments). The gradients containing the cell lysate were centrifuged in SW41 bucket rotors (Beckman Coulter) at 36,000 rpm at 8°C for 2 hours. Twelve fractions were separated and collected from the top of the gradient using the PGF Ip piston gradient fractionator (BioComp Instruments) and the Fraction Collector (FC-203B, Gilson). The absorbance readings were collected at 260nm with a Econo UV Monitor (EM-1

220V, Biorad). After fractionation, 110µl of 20% SDS and 12µl of Proteinase K (Thermo Fisher Scientific) were added to each fraction for a 30 min incubation at 42°C, after which 10µl of GeneChip Eukaryotic Poly-A RNA controls (Affymetrix, final concentration 1:120000) were added to each fraction. Total RNA from each fraction was extracted using phenol-chloroform-isoamyl alcohol (25:24:1, Sigma), poly(A) selected and made into a cDNA library using Ultra Directional RNA Library Prep Kit (NEB) following manufacturer's instructions.

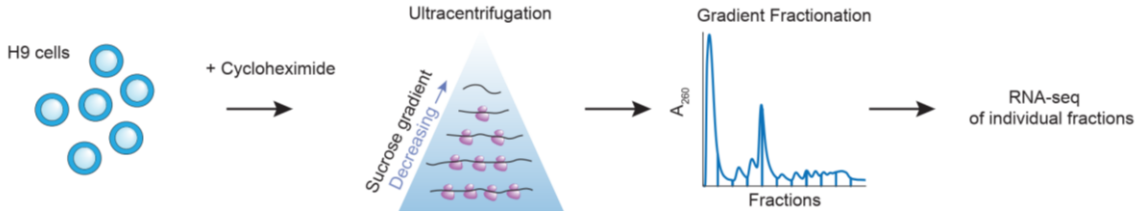


**Figure 2.2. Schematic of the TrIP-seq workflow.** Figure was published in *Aw. at el*[52].

## 2.7 Processing and quantification of TrIP-seq data

Raw paired-end reads were first trimmed and then mapped to the human transcriptome (Ensembl version GRCh38.93) using Salmon[85] options: -l A --seqBias --gcBias –posBias**)**. Relative abundances of each isoform were estimated as Transcripts Per Million (TPM) by Salmon and corresponding values were used for downstream analysis.

## 2.8 Basecalling and mapping of nanopore reads

Reads were basecalled with Albacore version 2.3.3 or Guppy 3.1.5 without filtering. Basecalled sequences were aligned with GraphMap version 0.5.2[86]. For single gene mapping, references for the individual genes were used. For H9 transcriptome mapping, cDNA and non-coding reference sequences obtained from Ensemble were used (GRCh38).

## 2.9 Determination of single-stranded positions

To evaluate the performance of various structure probing methods, single-stranded positions were determined as those having a value greater than one standard deviation above the

median of the SAFA value within a gel. Single-stranded positions were then used to evaluate various methods and determine ROC curves.

## 2.10 Calculation and comparison of error rates

Mismatch, deletion and insertion rates were calculated using custom Python scripts from aligned BAM files. The Wilcoxon test was used to compare error rates across modified and unmodified samples. Fold-changes in mismatch, deletion and insertion rates per position were calculated by dividing the error rate for modified samples by the rate for unmodified samples. The fold change was Winsorized where values $\geq 99^{th}$ percentile were set to the value at $99^{th}$ percentile and values <1 was set to 1. AUC-ROC values for mismatch, deletion and insertion rate-based prediction of single-stranded bases was calculated and compared between modified and unmodified libraries using the Wilcoxon test.

## 2.11 Alignment of nanopore signals

Current measurements above 200 pA and below 0 pA were considered as outlier values and are removed from the raw nanopore sequencing files. The current signal was aligned with Nanopolish (version 0.10.2). As the current mean drifts with increasing sequencing time, we normalized the current per strand to that of the expected model current mean in Nanopolish. Multiple events from the same read and position were collapsed into a single value by calculating the weighted average of event mean and event standard deviation and taking the sum of event lengths.

## 2.12 Training of parameters, determination of thresholds and calculation of reactivity profiles

A one-class support vector machine (SVM) was used to determine the percentage of modifications per position. Specifically, the current mean and current standard deviation were used as features for each base.

To determine the number of unmodified and modified reads needed for robust analysis, we subsampled reads to various depths (25 to 500, 100 iterations) and compared the reactivities of the subsampled strands to that of the full dataset using Pearson correlation. Reactivity scores were determined per position by calculating the percentage of modified bases detected using one-class SVM. For the hESC H9 transcriptome, transcripts that were present in both replicates of the modified libraries with a minimum of 100× coverage and transcript length >200 nucleotides, were retained for analysis. To compare the reactivity across isoforms, a 5-nucleotide moving average was applied to the transcript reactivity profile followed by Z-score normalization.

## 2.13 Reactivity near RBP binding sites

HITS-CLIP libraries for the RNA binding protein Lin28 (SRR531463 and SRR531464) were downloaded from SRA and mapped to the human genome (Homo_sapiens.GRCh38) using BWA[87]. Binding peaks were enriched using the CLIP Tool Kit (CTK)[88] package and binding motifs were detected using HOMER[89]. We analysed the reactivities -50bp to +50bp around Lin28 motifs (G{GU}

AG{C}A) that have CLIP binding peaks and randomly selected the same number of motifs outside of CLIP peaks as controls (compared using a t-test). 552 HITS-CLIP binding sites on 316 transcripts were used for the analysis of Lin28 binding in our data. We randomly sampled 552 sites with the same motif sequence on 351 transcripts as control.

## 2.14 SNV structure analysis

Illumina RNA-sequencing reads from the libraries RHN1291 and RHN1295 were mapped to the human genome (Homo sapiens.GRCh38) using BWA. Single nucleotide variants (SNVs) were called using bcftools with default parameters[90]. From the identified SNVs positions, nanopore mapped reads were separated into three categories, corresponding to matching (reads that match the annotation), mutated (mutated reads based on the variant calling results) and unclassified (reads that are not found in the previous two groups) by Biostar214299 from Jvarkit. Matching and mutated reads were used for further SNV analysis.

For each of the SNV pairs, both transcripts were filtered for having >200 unmodified reads and >100 modified reads. SNV pairs with at least one changing region were considered as having a change in structure.

## 2.15 Determination of structure changing regions

To determine whether a base changes structure significantly between two reactivity profiles, we used Fisher's exact test to compare the number of unmodified and modified reads at each position between two transcripts. A 5-nucleotide sliding window was applied across the transcripts and regions with two or more positions with p-value<0.05 across a transcript pair and that were not significant between biological replicates, were identified as being structurally changing. Hommel's method was used for FDR correction of the p-value. As the structure changing region cannot be structurally different across biological replicates, this allows us to filter off regions that fluctuate in coverage across biological replicates, reducing the amount of noise that is called as structurally significant.

## 2.16 Analysis of isoforms pairs

Transcript coordinates were converted into genomic coordinates to allow ease of comparison across isoforms. Two transcripts were considered to be a gene-linked isoform pair if they have overlapping genomic positions and >100bp of unique positions. Reactivity values from shared positions for each isoform pair were retained for comparisons. For global analysis, all shared positions were used to calculate Pearson correlation. For local analysis, 100 nucleotides to the left and right of differential splice sites were used (sites with fewer adjacent bases were excluded). Similarity of the isoform pairs were calculated by taking the positions that are overlapping in each pair divided by the maximum length of the bases found in the genome.

Translation efficiency (TE) for each transcript was determined by TrIP-seq. To calculate the significance in TE differences between two isoform pairs, the raw counts for both alleles across low polysome fractions (sum of fractions 6 and 7) and high polysome fractions (sum of fractions 9, 10) were compared using Fisher's exact test to assess if the reads derived from the reference allele are significantly enriched/depleted in high polysome fractions (p-value<0.05). Hommel's method was used for FDR correction of p-values.

## 2.17 Cells and viruses infection

African green monkey kidney, clone E6 (Vero-E6) cells (ATCC# CRL-1586) were maintained in Dulbecco's modified Eagle Medium (DMEM) supplemented with 5% fetal bovine serum (5% FBS). SARS-CoV-2 wild-type (WX56) and Δ382 mutant (CA001) were isolated from COVID-19 patients in Singapore, as reported previously[91].

## 2.18 Structure probing of SARS-CoV-2 virus in Vero E6 cells

Vero E6 cells were infected with SARS-CoV-2 viruses (WT and Δ382) at a multiplicity of infection (MOI) = 0.01 for 1 h at 37°C. Following 1 h infection, virus inoculum was removed and replaced with DMEM-5% FBS. Flasks were incubated for 48 h at 37°C, 5% $CO_2$.

At 48 hpi, cells were washed once with PBS and trypsin was added to detach the cells from the flask. The cells were collected and centrifuged at 2000 rpm. The pellet was resuspended in PBS and the cells were then separated into three reactions: (1) added 1:20 volume of 1M NAI (03-310, Merck, 25 µl of NAI in 500 µl of infected cells) and incubated for 15 min at 37 °C for structure probing; (2) added 1:20 volume of dimethyl sulfoxide (DMSO) and incubated for 15 min at 37 °C, as negative control; and (3) set aside a third portion of the infected cells without any treatment, for the denaturing control in the downstream library preparation process. The total RNA was extracted from the cells using E.Z.N.A. Total RNA Kit (Omega bio-tek) according to the manufacturer's instructions. We then performed library preparation following the SHAPE-MaP protocol to generate cDNA libraries compatible for Illumina sequencing[38].

## 2.19 Interactome mapping of SARS-CoV-2 virus in Vero E6 cells.

Vero E6 cells were infected with SARS-CoV-2 viruses (WT and Δ382) at a multiplicity of infection (MOI) = 0.01 for 48 h. The cells were washed once with PBS and trypsin was added

to detach the cells from the flask. The cells were collected and centrifuged at 2000 rpm. The pellet was resuspended in PBS and the cells were then incubated with 200 µM biotinylated psoralen and 0.01% digitonin in PBS for 10 min at 37 °C. The cells were spread onto a 10cm dish and irradiated at 365 nm of UV on ice for 20 min. The cells were collected, and the total RNA was then extracted using E.Z.N.A. Total RNA Kit (Omega bio-tek) according to the manufacturer's instructions. We performed SPLASH libraries similarly to the published protocol[92,93].

## 2.20 Direct RNA sequencing using Nanopore sequencing

Sequencing reads obtained from two replicates of unmodified and NAI-treated total RNA from WT (EPI_ISL_407987) and Δ382 (EPI_ISL_414378) SARS-CoV-2 infected Vero E6 cells were obtained using Oxford Nanopore direct RNA sequencing 002 kit. The concentration of samples was sequenced and aligned according to the method used by Kim et al[94]. Subsequently, the aligned reads were sorted into its individual subgenomic RNA by first filtering for the full-length reads, by searching for the presence of the leader sequence. Next, the full-length reads are sorted into the individual subgenomic RNA by its aligned positions after the leader sequences, based on the location of known subgenomic RNAs.

## 2.21 Modelling of individual subgenomic RNA structure from PORE-cupine data

The distributions of mutation rates obtained from Nanopore RNA sequencing were compared with the distribution of SHAPE reactivity values from short-read based experiments. We found that a scale factor of 100 brings the mutation rate distribution from the Nanopore experiment in line with the distribution observed for conventional SHAPE experiments. Hence, we applied this scaling factor and then employed these as SHAPE data in the same 'Superfold' protocol as for the full-length models with a maximum base-pairing distance of 600 bases, and default SHAPE slope (1.8) and intercept (-0.6) parameters[95].

## 2.22 PORE-cupine analysis of direct RNA sequencing data

Filtering for full-length sub-genomic RNA: To separate full length aligned reads into their sub-genomic transcripts, we used two filtering conditions for all sgRNAs except for ORF6: 1, The aligned reads need to contain the leader sequence, 2. The aligned positions after the leader sequence must fall within ± 100 of the annotated sub-genomic sequences. For ORF6, we had to extend the second filter to -300 of the annotated sub-genomic sequence.

We calculated the reactivity for each subgenomic transcript by using PORE-cupine, with two adjustments to the analysis. 1. The length filter was removed, as only full-length transcripts were used for the analysis. 2. To reduce the amount of computing resources required, 20000 strands from each subgenomic transcript in the unmodified libraries were randomly selected and used for the generation of models.

To determine the differences between reactivity, Wilcoxon Rank Sum test was applied to a 101 bases window with a step size of 25 nucleotides. Reactivity differences were compared across shared sequences between the different subgenomic transcripts within each strain (WT or Δ382), and across the two different strains. For the comparison between WT and Δ382 genome, the region in the WT strain that was not present in the Δ382 strain was masked. The *p*-values are corrected with Hommel's method. In addition to using the statistical test to determine the differences, we added a second criteria of Pearson correlation < 0.7.

## 2.23 Software and scripts

Source code for all scripts (R version 3.4.1) and commands used for analysis can be found at http://github.com/awjga/Pore-cupine.

## 2.24 Data availability

Raw sequencing data and reactivity profiles can be downloaded from https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE133361

.

# Chapter 3 Development of high-throughput long-read structure probing, PORE-cupine.

During the initial phase of development, the direct RNA sequencing kit was recently released by Oxford Nanopore technology. Therefore, we validated the results from the published literature with our sequencing results[59,96] before proceeding to the development of the protocol. For our initial evaluation of the sequencing and development of the protocol, we have used *in-vitro* transcribed Tetrahymena ribozyme RNA, a relatively short RNA of 420 bp, that has a well-defined secondary structure (**Figure 1.6**). We have also looked at the numerous structure probing compounds that have been used in various short-read structure probing methods and selected five probes that modify single-stranded bases for determining how these modifications would affect the sequencing results (**Figure 1.3a**). We selected chemical probes that modify all bases: N-methylisatoic anhydride (NMIA), 2-methylnicotinic acid imidazolide (NAI) and 2-methylnicotinic acid imidazolide azide (NAI-N3), as well as base-specific chemical probes: dimethyl sulfate (DMS) and 1-cyclohexyl-3-(2-morpholinoethyl) carbodiimide metho-p-toluenesulfonate (CMCT) (**Figure 1.3a**)[37,40,97,98]. DMS alkylates single-stranded bases of As and Gs, while CMCT primarily reacts with single-stranded Us and Gs (**Figure 1.3b**). We treated the *in-vitro* folded Tetrahymena ribozyme RNA with each of the compounds and analysed the data to determine the suitability of using direct RNA sequencing to detect the modifications caused by the structure probing compounds. Most of the work in this section has been published in the journal of Nature Biotechnology[52].

## 3.1 Direct RNA sequencing statistics of Tetrahymena ribozyme RNA

We treated Tetrahymena ribozyme RNA with NMIA, CMCT, DMS, NAI and NAI-N3 as described in the methods section of the thesis and individually sequenced the unmodified and modified samples with direct RNA sequencing. We sequenced around 20k to 70k reads and 13% to 86% of the sequenced reads were able to be aligned to the Tetrahymena ribozyme RNA reference sequence. The result for the unmodified sample is similar to the published data[59,96], where around 85% of the unmodified reads can be aligned. However, we noticed a decrease in the percentage of reads from the modified samples that can be aligned, indicating that RNA structure probing modifications affect the alignment of the sequenced reads.

In addition to looking at the alignment difference between the unmodified and structure probed samples, we looked at the length distribution of the aligned. The modified reads from the Tetrahymena ribozyme RNA are similar to that of the unmodified samples, except for DMS modified mapped sequences which are slightly shorter(**Figure 3.1)**.

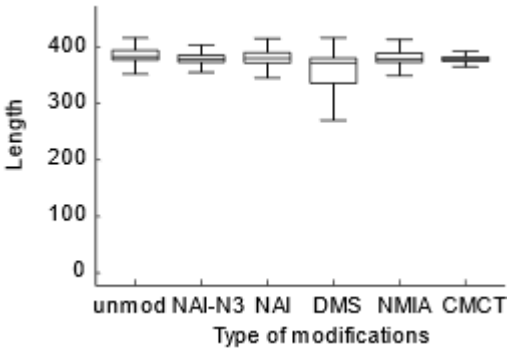


**Figure 3.1. Boxplot of length distribution of aligned unmodified and modified Tetrahymena ribozyme RNA reads.** The figure is published in *Aw. at el*[52]

We also observed a higher error rate in the modified samples when compared to the unmodified samples. The mismatch rate (modified: 6.5%-11.7%, unmodified: 5.3%), deletion rate (modified: 8.7-13.2%, unmodified: 8.7%), and insertion rate (modified: 3.4-4.1%, unmodified: 3%) in modified Tetrahymena ribozyme RNA sequences for the five chemical compounds are significantly higher (**Figure 3.2**).

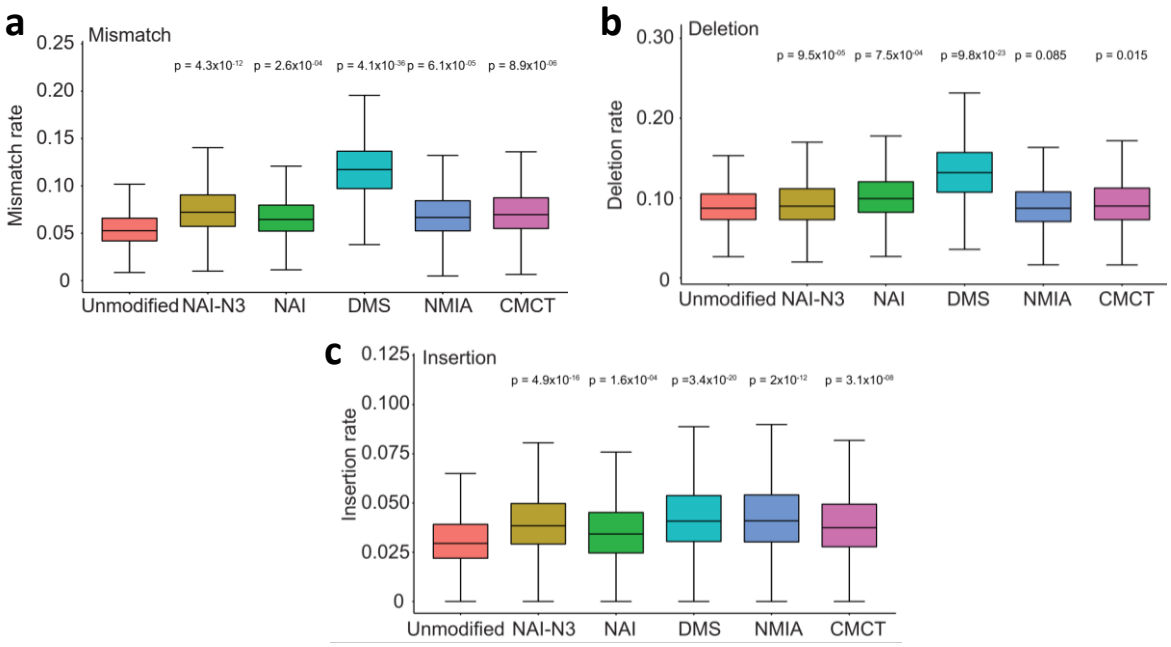


**Figure 3.2. Boxplot of error rates in aligned unmodified and modified Tetrahymena ribozyme RNA reads.** Boxplots showing the frequency of mismatch (**a**), deletion (b) and insertion (c) rates for different structure probing chemicals on Tetrahymena ribozyme RNA, as compared to unmodified RNA. *P*-values were calculated using the two-sided Wilcoxon Rank Sum test. Figures were adapted from *Aw. at el*[52].

Overall, our results indicate that modifications caused by structure probing compounds in Tetrahymena ribozyme RNA affect the sequencing performance in direct RNA sequencing, suggesting that we can use this technology for the detection of modifications.

## 3.2 Detecting modifications with alignment error rates.

Having determined that direct RNA sequencing is affected by the structure probing modifications, we proceed to the next phase of the protocol development, which is to determine the accuracy of using direct RNA sequencing to detect structure probing modifications on the RNA. Using the error rates, we calculated the fold differences between the mismatches, insertion and deletions rates between the unmodified control and for each of the five compounds (**Figure 3.3**). We evaluated the performance by using area-under-curve (AUC-ROC) analysis based on footprinting signals for the Tetrahymena ribozyme RNA (**Figure 3.4**). Overall, we observed that NAI, NAI-N3 and NMIA were able to detect the single-stranded bases over the other compounds. DMS modifications resulted in the worst performance, which might be due to its ability to modify Gs in a structurally independent manner. NAI-N3 modifications had the best performance among the compounds that we have tested. To focus our effort on developing the protocol, we choose to further optimize the detection method with the structure compound NAI-N3.

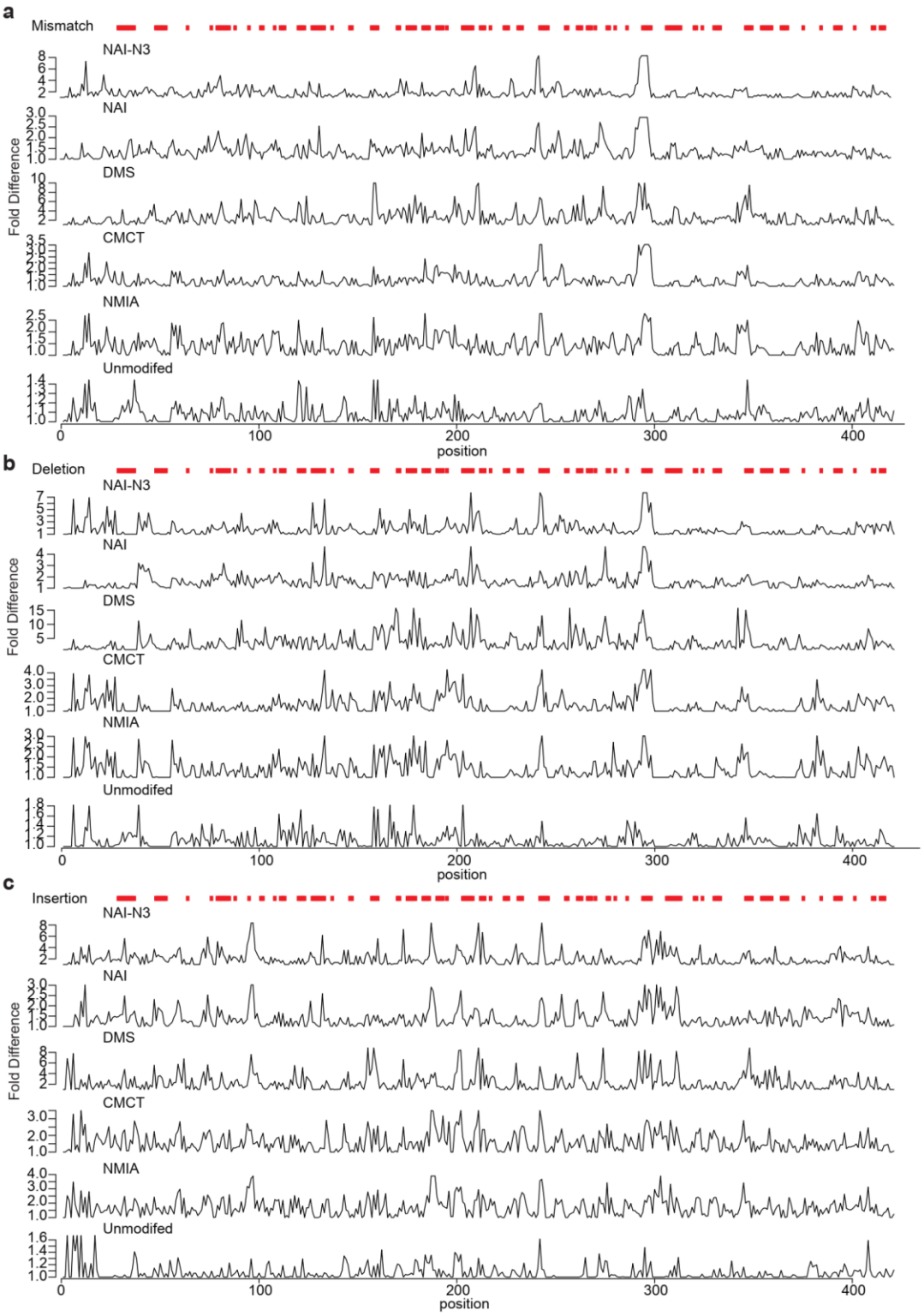


**Figure 3.3. Distribution of mismatches, insertions, deletions along Tetrahymena ribozyme RNA sequence.** Line plots of the normalized number of mismatches (a), deletions (b) and insertions (c) caused by the different compounds and unmodified, along the length of the Tetrahymena ribozyme RNA sequence. The red bars on top of the plots indicate the location of single-stranded bases in the secondary structure. Figures were published in *Aw. at el*[52].

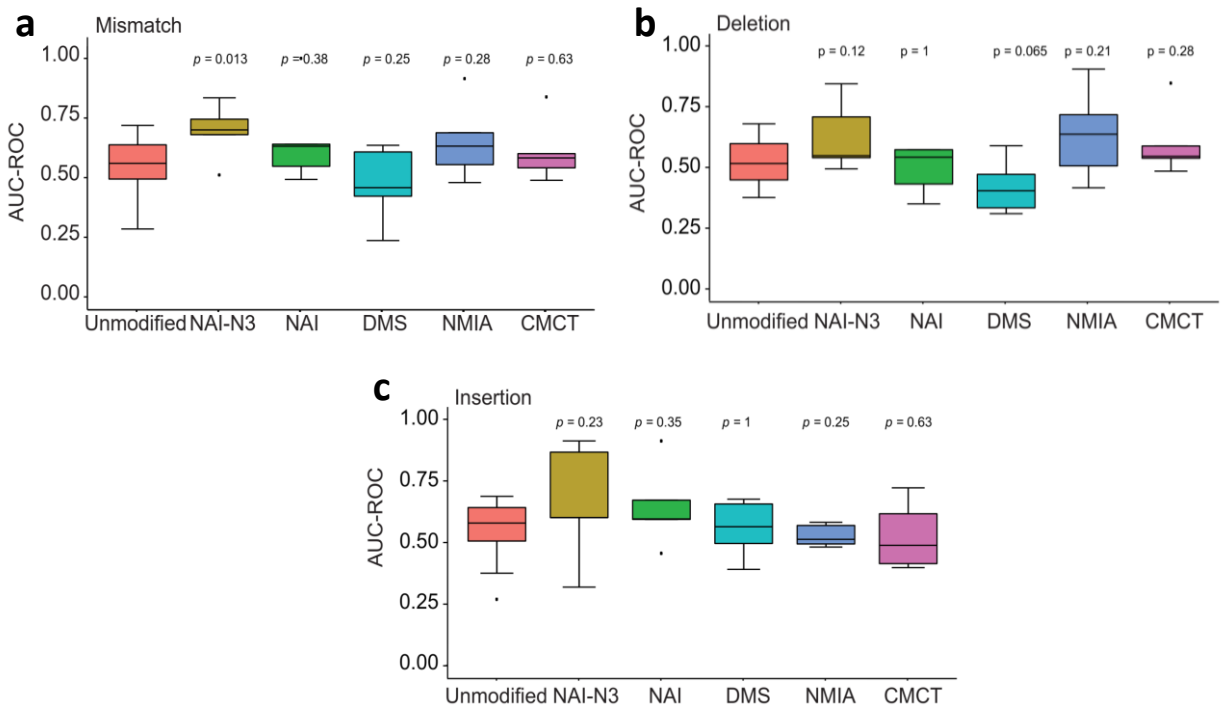


**Figure 3.4. Accuracy of detecting the various structure probing modifications.** Boxplots showing the AUC-ROC performance of mismapping (a), deletion (b) and insertion (c) rates for the different compounds on the Tetrahymena ribozyme RNA secondary structure. P-values were calculated using two-sided Wilcoxon rank-sum test. Figures were adapted from *Aw. at el*[52].

## 3.3 Using raw signal of direct RNA sequencing as features

Due to the inability of error rates to detect the modifications at the individual strand level, we investigated the potential of using the raw current signal to detect the modifications. The raw signal of direct RNA sequencing originates from the measurement of the current caused by the perturbations in a protein pore during sequencing. A constant current is applied across a protein pore. When a strand of RNA is threading through, the amount of current flowing through the pore varies with the different combinations of sequences, giving rise to the raw current signals. The raw current signal can be used for basecalling by a neural network software released

by Oxford Nanopore. However, the basecalling software does not indicate the corresponding locations of the raw signal to the bases in the sequence, which is an essential step for us to utilise the raw signal to detect the modifications. Therefore, we used the program Nanopolish[86] to align the current signals to its basecalled sequences at the individual strand level (**Figure 3.5**). From the output of Nanopolish, it is evident that multiple data points contribute to a position in the sequenced reads, resulting in a huge amount of data, making it difficult to process and analyze the data. Therefore, to condense the amount of data, the current signal for each position is condensed into three factors, the current mean, current standard deviation and dwell time.

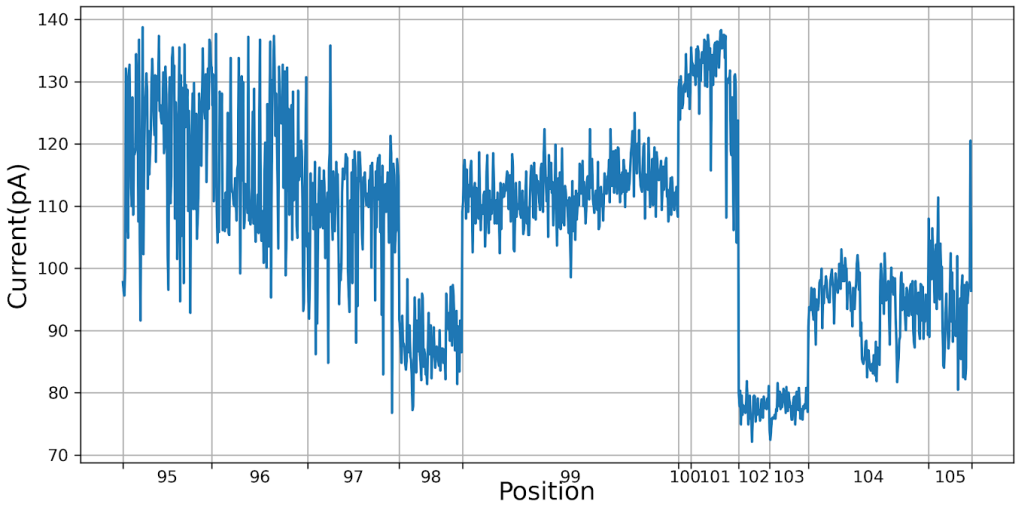


**Figure 3.5. A section of the raw current signal from a single RNA strand.** The raw current signal was extracted and aligned with Nanopolish.

Using the condensed features that we have extracted from the raw signals, we compared the distribution of modified versus unmodified single-stranded and double-stranded bases along the Tetrahymena ribozyme RNA, to determine if these features have the potential to detect the modifications. From the distribution, we see a shift in

the modified single-stranded bases in current mean and current standard deviation, but not in their dwell time, as compared to unmodified bases (**Figure 3.6)**. This suggests that we could use the current mean and current standard deviation as features to determine if the bases are modified.

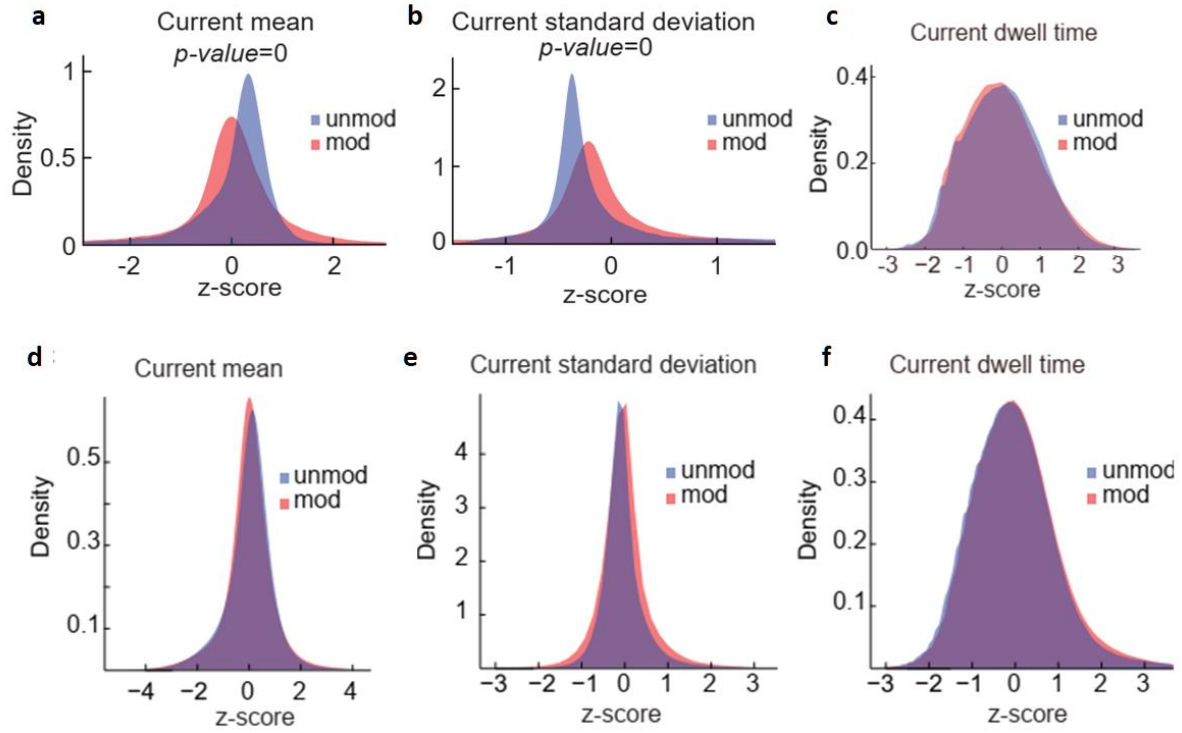


**Figure 3.6. Density plots of normalized extracted features from unmodified and NAI-N3 treated Tetrahymena ribozyme RNA.** a-c, Using footprinting gels as a guide, the top 10% of single-stranded regions on the Tetrahymena ribozyme RNA were chosen for these plots. Normalized current mean (a), standard deviation (b) and dwell time (c) distributions for single-stranded positions on unmodified (blue) and modified (red). d-f, Normalized current dwell time for unmodified regions of Tetrahymena ribozyme RNA. Normalized current mean (e), standard deviation (f) and dwell time (g) distributions for all positions on unmodified Tetrahymena ribozyme RNA and RNA modified with NAI-N3. Figures were adapted from *Aw. at el*[52].

## 3.4 Detecting modifications with one-class support vector machine is accurate, reproducible and comparable to existing methods

Having determined the two features that are suitable for identifying the positions that are modified, we looked for a suitable machine learning algorithm to determine the modifications based on the unmodified signals. We have selected and evaluated the machine learning algorithm, one-class support vector machine (SVM), to detect modified bases[99]. To identify modified bases from the sequenced reads, the features extracted from the raw current signal from the unmodified sample were used to train a model for each position. The trained models were then used to detect the anomaly from the signals in the NAI-N3 modified samples. Using AUC-ROC analysis, based on footprinting signals for the Tetrahymena ribozyme RNA, we optimized the parameters of SVM to best differentiate signals from modified versus unmodified bases (Methods). Having optimized the parameters, we were able to generate the extent of modified outliers per base that could be calculated as a “reactivity score”, whereby the higher the score, the higher the probability that a base is single-stranded. For example, the double-stranded base 182 in the Tetrahymena ribozyme RNA is less likely to be modified by NAI-N3, and the current signal with or without chemical modification upon sequencing are shown to be similar (**Figure 3.7**). However, the single-stranded base 129, has a higher chance of being modified by NAI-N3 and this is reflected by the “comet tail” indicating deviations in current for the modified base (**Figure 3.7**).

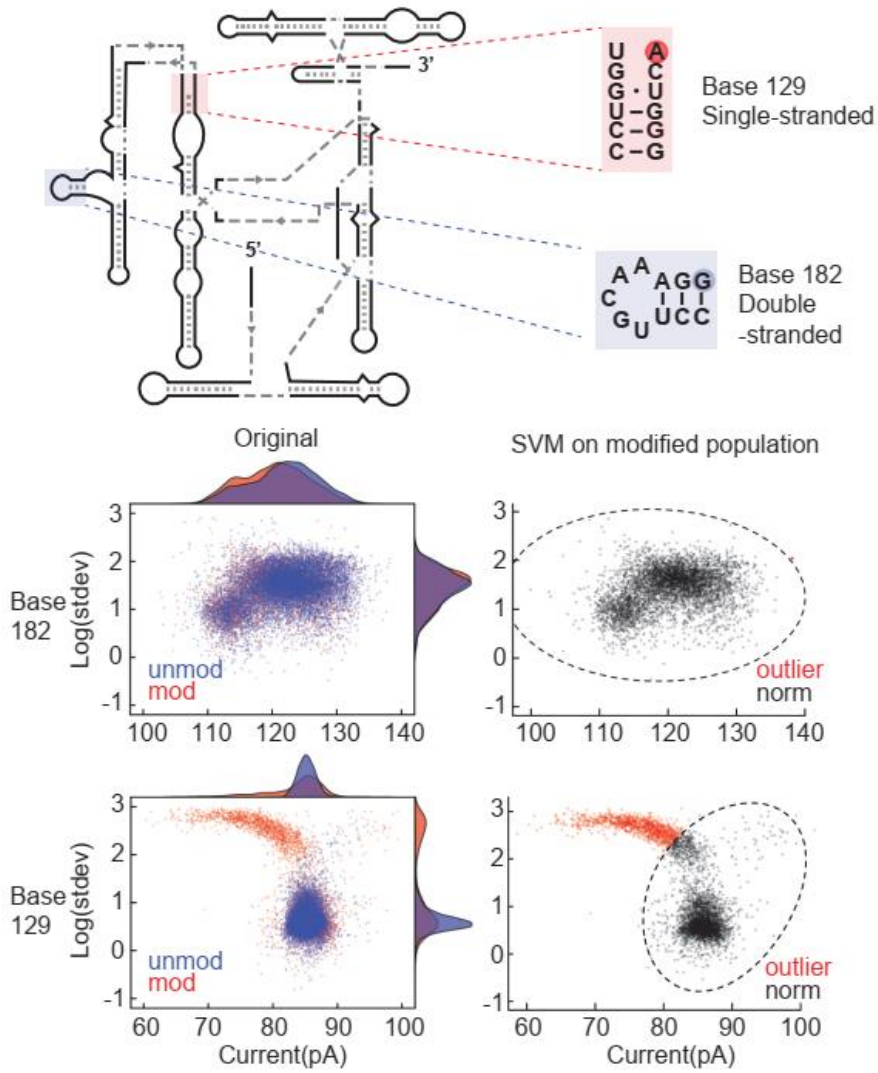


**Figure 3.7 Scatter plots of single-stranded and double-stranded positions.** *Upper panel*, schematic of the secondary structure of Tetrahymena ribozyme RNA and the location of a representative single- and double-stranded base. *Lower panel*, Current mean and $\log_{10}$ (standard deviation of current) of the highlighted bases in the *upper panel*. Each data point is from a base in a single RNA strand. Plots on the left show the distribution for non-structure probed (blue) and structure probed (red) bases before SVM classification. Plots on the right show only the distributions for structure probed bases, but with the SVM boundary drawn (dotted lines). Outliers are in red and points within the boundary are in black. Figures were published in *Aw. at el*[52].

Next, we examined the reproducibility of SVM to detect the modifications. The reactivity score from two replicates of the Tetrahymena ribozyme RNA was highly correlated, indicating that our data is reproducible (R=0.97, **Figure 3.8**). As a precaution to confirm that modifications detected by SVM are due to structure specific

modifications, we examined the reactivity of a biological replicate of unmodified and denatured control treated with NAI-N3. The reactivity scores from the unmodified samples were observed to be relatively low when compared to the reactivity scores from the *in-vitro* folded samples and the reactivity score was evenly distributed in our denatured control, indicating that the reactivity scores represent real structure modifications (**Figure 3.9**).

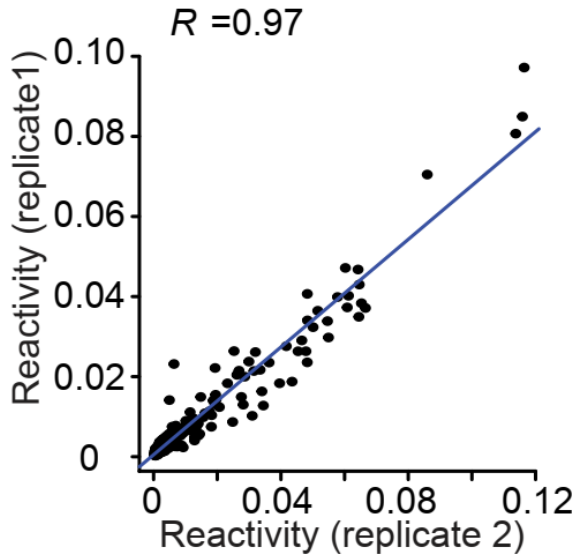


**Figure 3.8 Scatterplot of reactivity between two biological replicates of Tetrahymena ribozyme RNA.** The reactivity of the two replicates was normalised with Z-score, and Pearson correlation was used to calculate the similarity. R=0.97, $CI_{95\%}$ = [0.97,0.98] (Pearson correlation). P-value=2.5x10-262, two-tailed Student's T-test. Figure was published in *Aw. at el*[52].

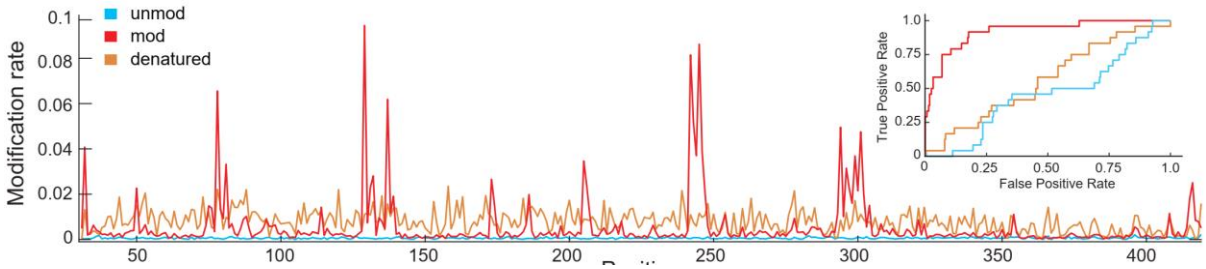


**Figure 3.9. Line plots of NAI-N3 modification rates along the entire length of the Tetrahymena ribozyme RNA.** The modification rate of folded RNA (red), randomly modified denatured RNA (brown) and another unmodified replicate (blue) is shown. The y-axis indicates the modification rate per base, whereas the x-axis indicates the position along the RNA. Inset, ROC curves for unmodified, modified and denatured Tetrahymena ribozyme RNA sequences. Figure was published in *Aw. at el*[52].

During the evaluation of the parameters, we noticed that the reactivity scores generated by our protocol have a two-base frameshift relative to footprinting signals (**Figure 3.10**), and correcting for this frameshift results in the highest Pearson correlation coefficient (**Figure 3.11**). This might be due to how nanopore sequences the samples. As five bases of an RNA strand would occupy the nanopore channel at a given time, the two-base shift seems to indicate that when the modification is located in the middle of the channel, it will result in the largest current difference in our study. Thus, the two-base shift is applied for all our downstream analyses.

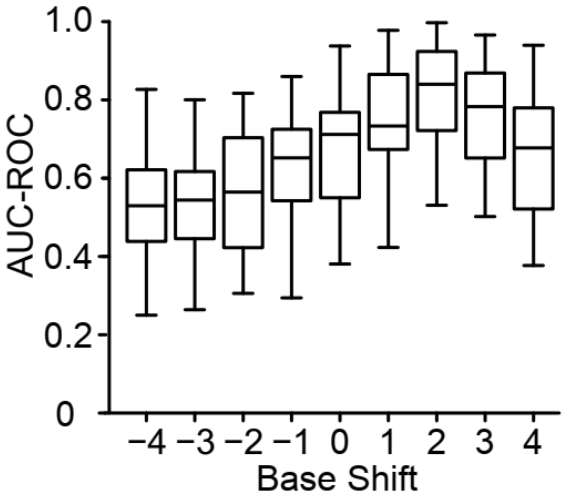


**Figure 3.10. Boxplots of AUC-ROC performance of the correlation of NAI-N3 reactivities to footprinting results with the different shifts in the bases.** The reactivity was shifted from -4 to 4 bases, and AUC-ROC was calculated after each shift. Figure was published in *Aw. at el*[52].

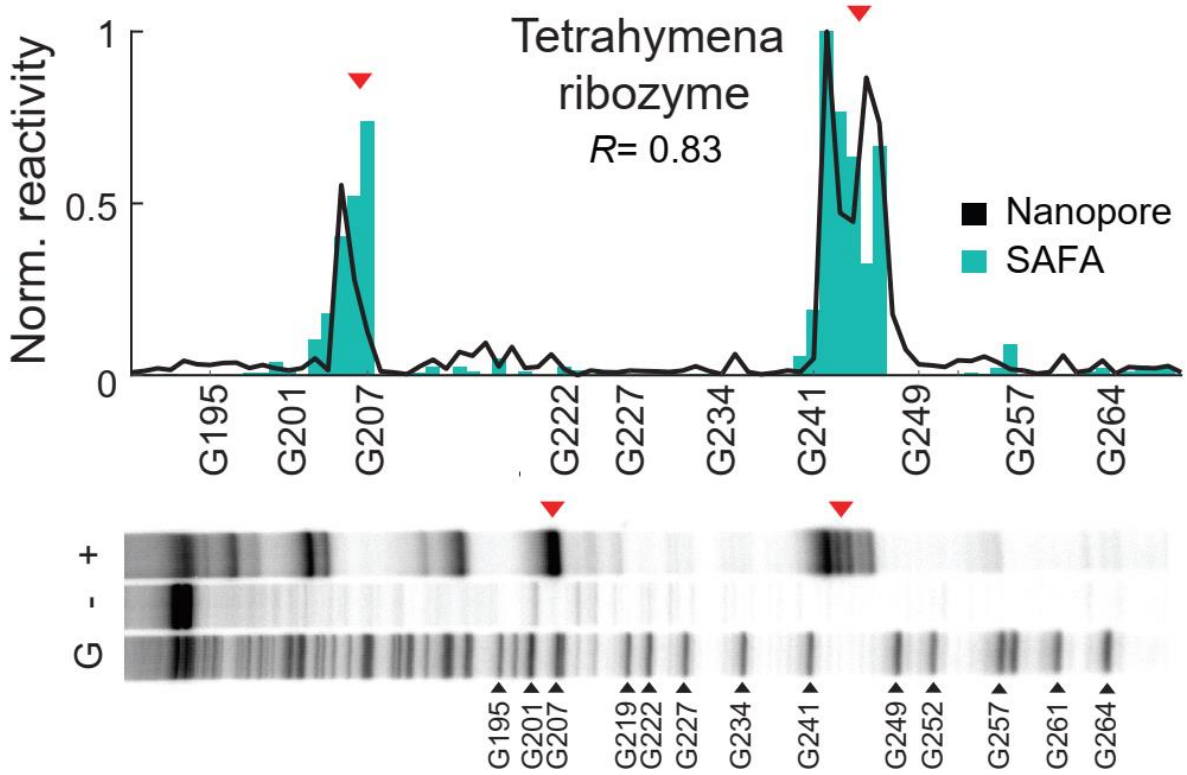


**Figure 3.11. Plots reactivity of NAI-N3 modifications of structure probing and Nanopore sequencing on Tetrahymena ribozyme RNA.** Comparison of NAI-N3 modification reactivity between detected modifications (black) and the average SAFA footprinting signals from n= 2 biological replicates (teal). The footprinting is from base 189 to 269 along the Tetrahymena ribozyme RNA. Lane 1 is G ladder, lane 2 is unmodified RNA and lane 3 is NAI-N3 modified RNA. Correlations were quantified using the Pearson correlation coefficient. Figures were published in *Aw. at el*[52].

Using a sole transcript to determine the parameters used in SVM might result in the model over-fitting. Therefore, we expanded our training and test set to 14 RNAs (2663 bases including two human mRNAs) and used 11 transcripts for training and three transcripts for the evaluation of the optimization (**Figure 3.11, Figure 3.12, Table 3.1, Appendix**). Except for 16S rRNA, where we performed *in vivo* structure probing, the rest of the 13 RNAs were *in-vitro* transcribed, and structure probed *in vitro*. We sequenced the RNAs as a pool to obtain 0.5-2M reads using direct RNA sequencing. Similar to our previous results, our analysis showed high reproducibility in reactivity between different biological replicates for RNAs in the test set (**Figure 3.13, 3.14**). The optimized SVM parameters performed similarly to the initial SVM parameters based on the Tetrahymena ribozyme RNA but exhibited a slightly higher median AUC-ROC

score of 0.79 on the test set (**Figure 3.15**). To rule out the possibility of overfitting of our data, we repeated the training of our parameters 20 times, each time using a random set of transcripts and overall performance was similar to our current parameters (**Figure 3.16**).

| No. | Training and test RNA |
|---|---|
| 1 | Tetrahymena ribozyme |
| 2 | 16S rRNA |
| 3 | ykoK (Magnesium riboswitch) |
| 4 | ydaO (ATP riboswitch) |
| 5 | ribD (FMN riboswitch) |
| 6 | ypaA (FMN riboswitch) |
| 7 | xpt (Purine riboswitch) |
| 8 | lysC (Lysine riboswitch) |
| 9 | yvrC (AdoCbl riboswitch) |
| 10 | Dengue 3' UTR |
| 11 | Yeast SCR1 |
| 12 | Yeast Hac1 3'UTR |
| 13 | Human RPS12 |
| 14 | Human RPS29 |

**Table 3.1. List of RNAs used for training and test.**

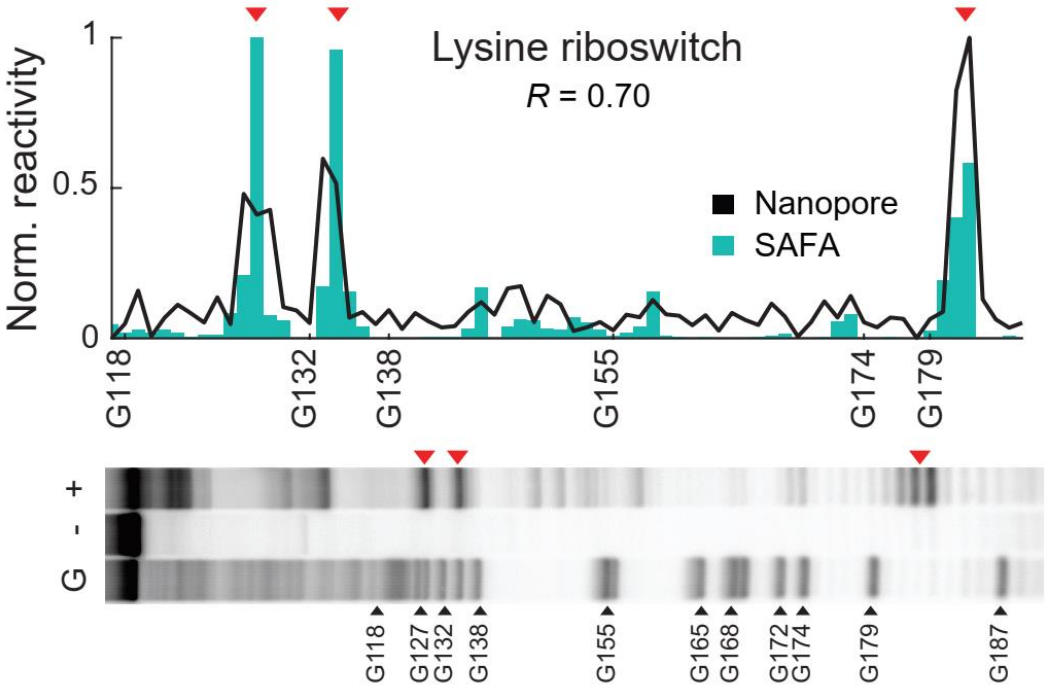


**Figure 3.12 Plots reactivity of NAI-N3 modifications of structure probing and Nanopore sequencing of lysine riboswitch.** Comparison of NAI-N3 modification reactivity between detected modifications (black) and the average SAFA footprinting signals from n= 2 biological replicates (teal). The footprinting is from base 117 to 186 along the lysine riboswitch. Lane 1 is G ladder, lane 2 is unmodified RNA and lane 3 is NAI-N3 modified RNA. Correlations were quantified using the Pearson correlation coefficient. Figures were published in *Aw. at el*[52].

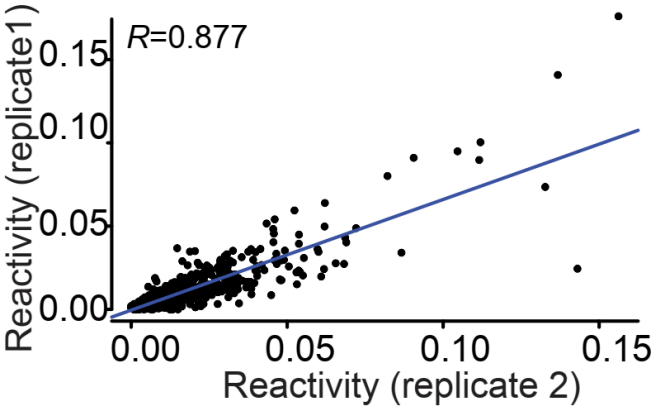


**Figure 3.13. Scatterplot of per-base reactivity in two replicates of the three test RNAs**. Reactivity from two replicates of 16S rRNA, yvrC and RPS29 was plotted and Pearson correlation was used to calculate the similarity. P-value = 0 using two-tailed Student's T-test. R= 0.877, $CI_{95\%}$ = [0.87, 0.89] Figure was published in *Aw. at el*[52].

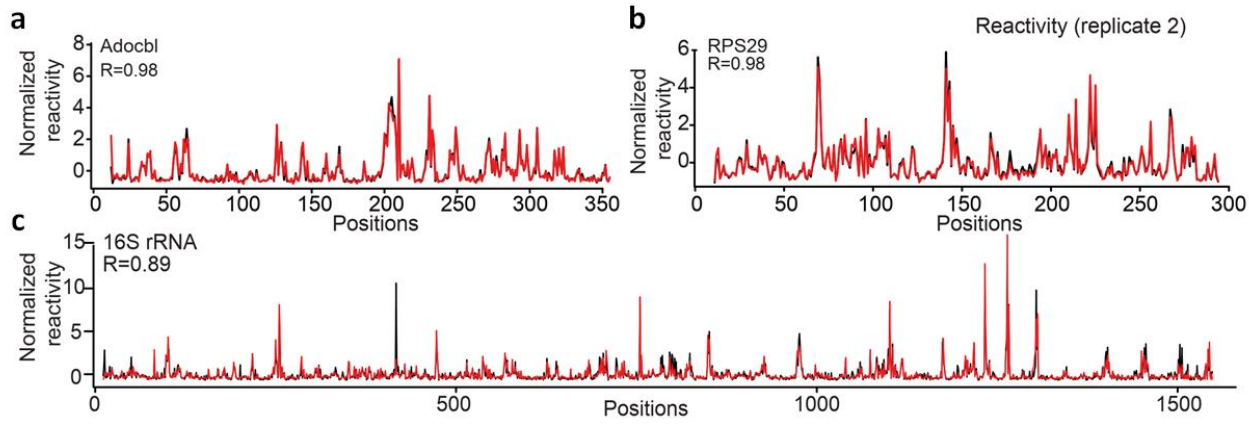


**Figure 3.14. Line plots showing the overlay of reactivity from two replicates of the test RNAs.** Reactivity from two replicates of 16S rRNA, yvrC or RPS29 are overlaid on each other. R>=0.89, using Pearson correlation. Figures were published in *Aw. at el*[52].

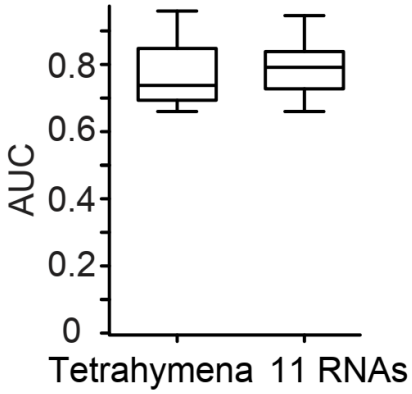


**Figure 3.15. Boxplot showing the performance of the SVM parameters on the 3 test RNAs**. Results from the detection of structure probing modifications from 16S rRNA, yvrC and RPS29 based on training on the Tetrahymena ribozyme RNA (left) or on 11 RNAs (right, **Table 3.1.**). Figure was published in *Aw. at el*[52].

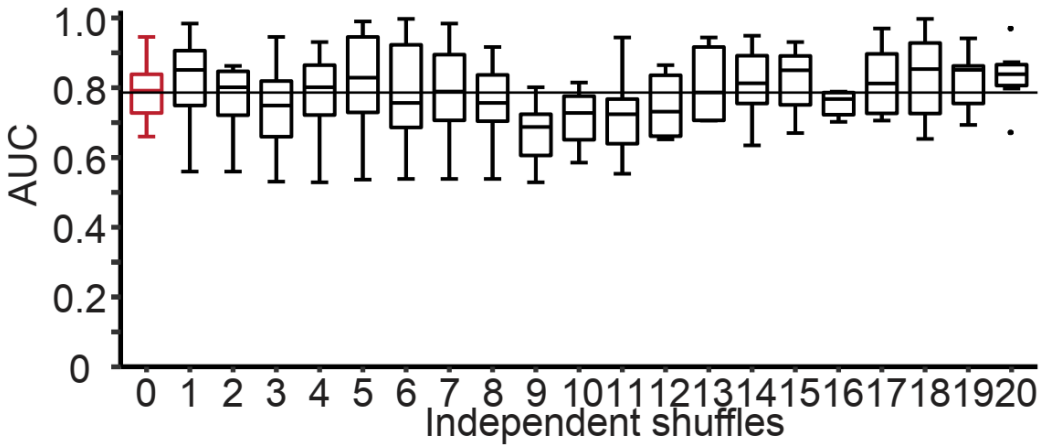


**Figure 3.16. Accuracy of randomizing the transcripts used for training and test.** AUC-ROC performance of SVM parameters on 3 test RNAs (red, based on our current 11 training RNAs) versus test RNAs after random selection of 11/14 RNAs as training, for 20 times.

In addition to evaluating the reactivity scores with footprinting signals and comparing the similarity between biological replications, we further validated our method by comparing our results with existing short-read structure probing methods. We performed structure probing on two transcripts, Tetrahymena ribozyme and 16s RNA, with our methods, icSHAPE and SHAPE-MaP, and compared their accuracy based on the footprinting signal. We found that the accuracy of the method we developed was comparable with the existing methods ($AUC_{PORE\text{-}cupine}$: 0.92, $AUC_{SHAPE\text{-}MaP}$: 0.93, $AUC_{icSHAPE}$: 0.91, **Fig. 3.17**).

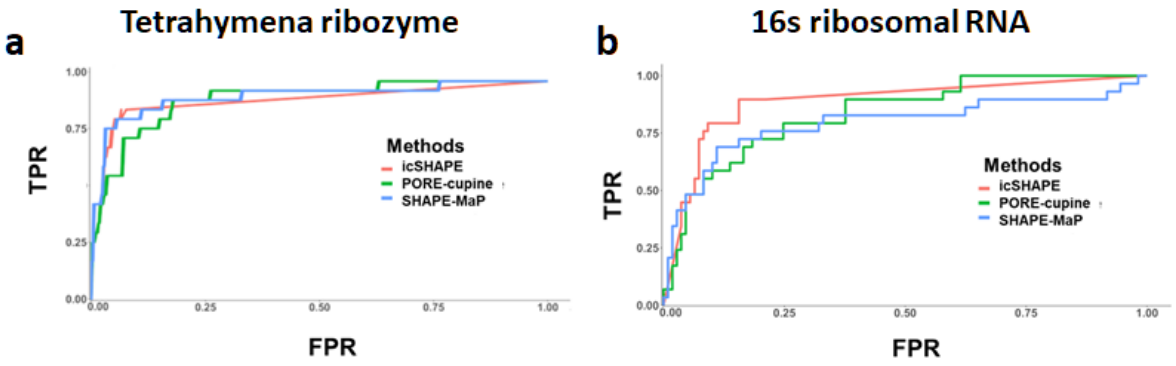


**Figure 3.17. Comparison of PORE-cupine with high throughput short-read structure probing techniques. a,b,** Comparison of ROC from three different techniques, *PORE-cupine* (green), *icSHAPE* (red) and SHAPE-MaP (blue). The calculation of ROC was based on the footprinting results. **a**, Line plot of ROC for *in-vitro* probed Tetrahymena ribozyme RNA. **b**, Line plot of ROC for *in-vivo* probed 16s ribosomal RNA.

With the evaluation of our protocol, we have shown that we can accurately and reproducibly detect structure-specific NAI-N3 modifications with direct RNA sequencing. The results generated with our protocol are also comparable to current methods. We named the protocol that we developed For chemical utilized probing interrogated using Nanopores (PORE-cupine).

## 3.5 *PORE-cupine* can characterize the dynamics of RNA structures under different conditions

Having developed PORE-cupine, we proceeded to evaluate the ability of PORE-cuipine to detect structural changes. By using a known riboswitch, TPP riboswitch, where it folds into different structures with ligand binding and without[81]. We applied PORE-cupine to the TPP riboswitch folded under the two different conditions and obtained 5-64k sequenced reads. By comparing the reactivity scores from the two conditions, we detected structural differences in the aptamer region due to the binding of TPP ($R$=0.3 between vehicle and 10mM TPP in the aptamer region versus $R$=0.9 in non-aptamer regions (**Figure 3.18a**). In addition, we could observe a gradated change in reactivity score under increasing concentrations of TPP *in vitro*, indicating

that PORE-cupine could detect gradual changes in RNA secondary structure (**Figure 3.18b**).

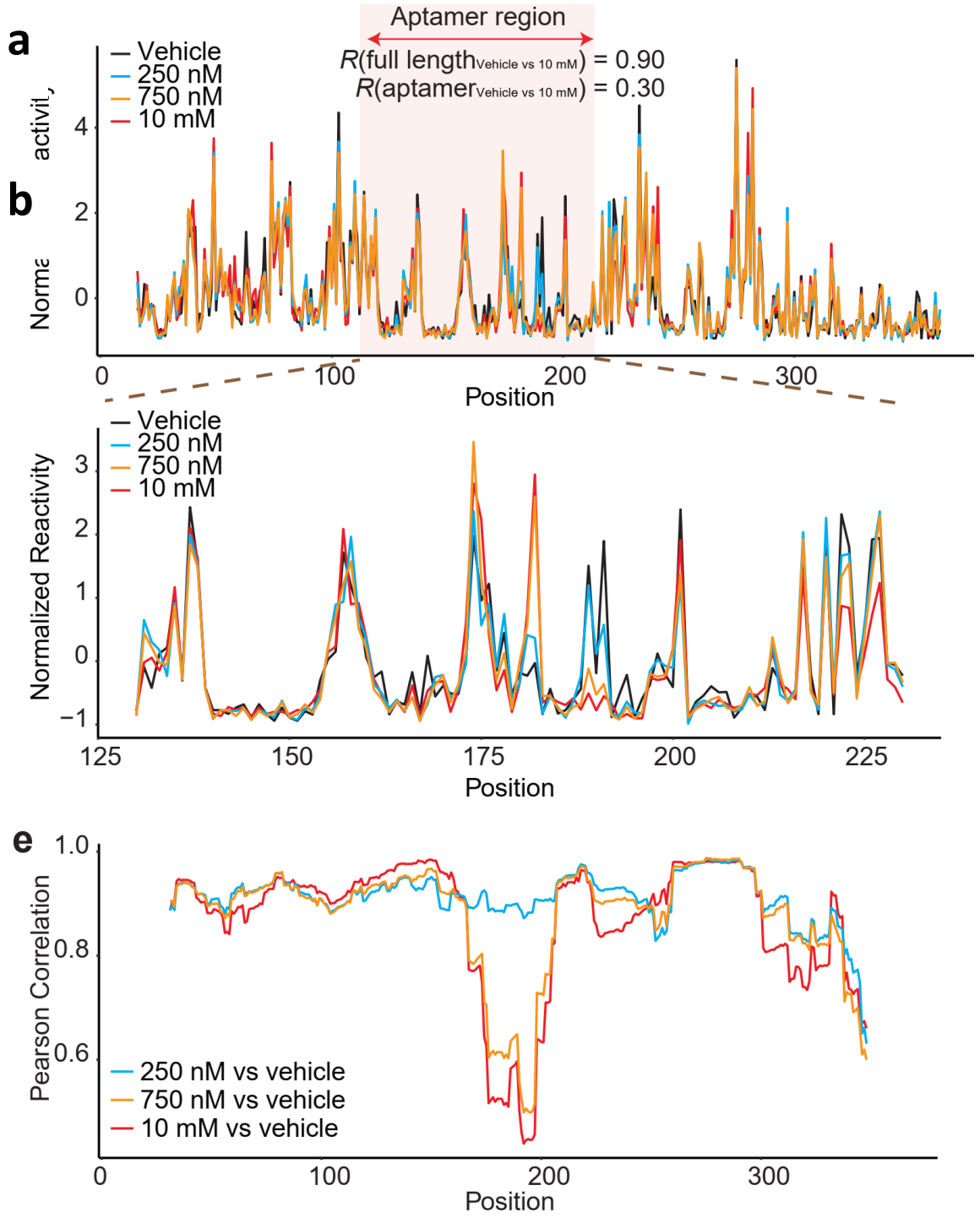

**Figure 3.18. *PORE-cupine* captures riboswitch structural dynamics.** a) *PORE-cupine* captures TPP riboswitch dynamics. 20-nucleotide sliding window Pearson correlation (upper) and normalized reactivity profiles of the aptamer region (lower) of TPP riboswitch RNA folded in the presence and absence of 10μM TPP. b) *PORE-cupine* detects a gradual change in TPP structure. 20-nucleotide sliding window Pearson correlation profiles of TPP riboswitches folded in the presence of increasing concentrations of ligand. Figures were published in *Aw. at el*[52].

# 3.6 Conclusion

During our initial phases of development, we have evaluated five different structure probing compounds as we were uncertain how different modifications along RNA might affect nanopore sequencing. We observed that all modifications cause an increase in the error rates, and we could use the error rates to detect the modifications. However, due to the difficulty of using the error to detect modifications at the single strand level, we investigated the potential of using the raw current signals. Together with the extracted raw current signal that is error corrected and machine learning strategies such as SVM, we were able to identify NAI-N3 modified bases at a single strand level.

We have set out to develop a protocol that couples structure probing with high-throughput sequencing, and we have successfully developed PORE-cupine. By using direct RNA sequencing from Oxford Nanopore Technologies, and the current signal extracted from the sequencing of unmodified and modified samples, we can accurately detect NAI-N3 modifications with SVM (**Figure 3.19**). We showed that PORE-cupine is highly reproducible and its results are also comparable to other short-read high-throughput sequencing methods[36,37,100,101]. In addition to detecting structures, we could detect structure differences with the results from PORE-cupine, by comparing the reactivity scores from RNA with different structures. Thus, we are confident in applying PORE-cupine to address the biological questions that we are interested in.

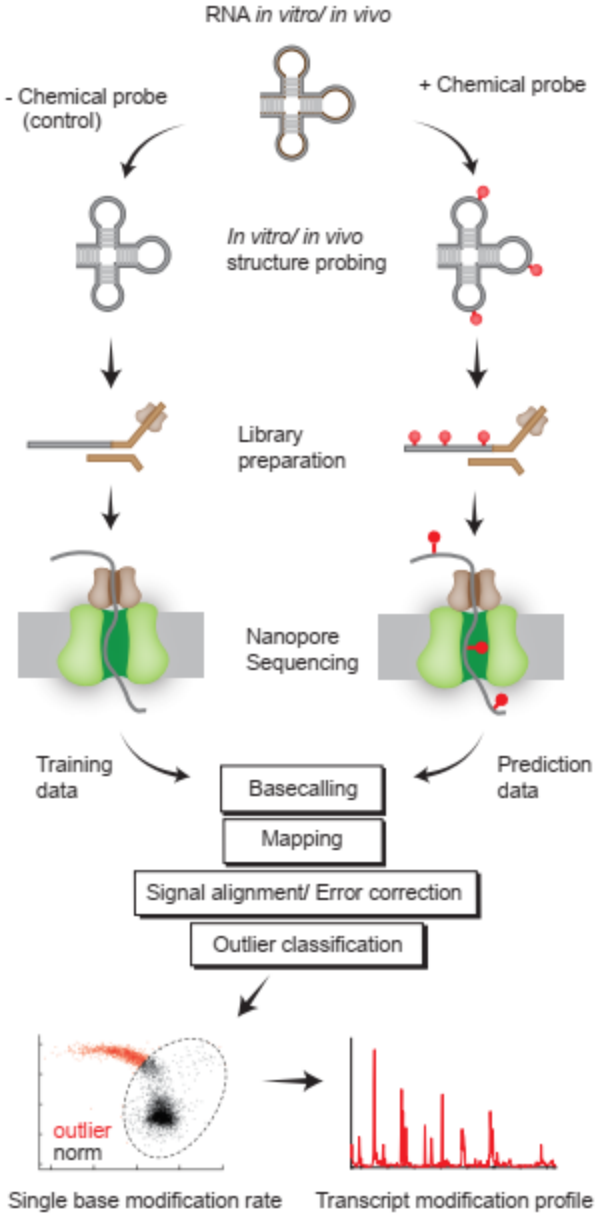


**Figure 3.19. Workflow of PORE-cupine.** Figures were published in *Aw. at el*[52].

# Chapter 4 Determination of isoform-specific RNA structure with nanopore long reads in human embryonic stem cells (hESCs)

The human transcriptome is extensively spliced resulting in genes having multiple isoforms[102,103]. Current methods for high-throughput structure probing use short-read sequencing and are unable to accurately assign the reads to the individual isoforms, which results in obtaining the aggregate structural information at the gene level making it difficult to study the RNA structures at the transcriptome level[37,104]. With PORE-cupine, however, we hypothesized that we will be able to study the individual structures for each RNA isoform in the hESCs. Where we can compare RNA structures of isoforms within the same gene to identify regions with structural differences. By coupling the differences in structures with other datasets, we might be able to find an association between changes in RNA structures and biological functions.

We prepared four biological replicates of NAI-N3 modified and two biological replicates of unmodified hESC transcriptomes. With PORE-cupine, we obtained the reactivity scores for 1751 transcripts. We filtered for genes that have more than one isoform, as we are interested in comparing the pair-wise differences between the isoforms. By coupling the structural differences with TrIP-seq[84], a method to measure the translation efficiency of the transcripts, we were able to correlate the translation efficiency with structural differences. The results in this section have been published in Nature Biotechnology[52].

# 4.1 Genome-wide analysis of RNA structures in hESCs using PORE-cupine

We sequenced around 10 million reads for both unmodified control and NAI-N3 treated samples (**Figure 4.1**). The percentage of aligned reads from the unmodified and modified samples were 86.1% and 59.6% respectively (**Figure 4.2**), with most reads having a modification rate of 1-2% (**Figure 4.3**). We observed a decrease in the read coverage at both ends of the transcripts, indicating that these could be blind spots in obtaining the reactivities scores (**Figure 4.4**).

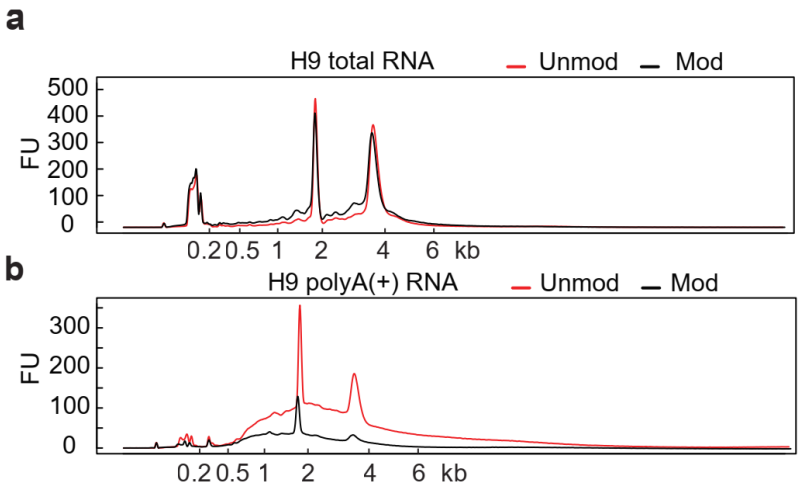


**Figure 4.1. Bioanalyzer traces of unmodified and modified RNA from hESC.** Overlapping traces of the size distribution of unmodified (red) and NAI-N3 (black) RNA from (**a**) total RNA and (**b**) polyA selected RNA. Samples were prepared according to the methods sections. Figures were published in *Aw. at el*[52].

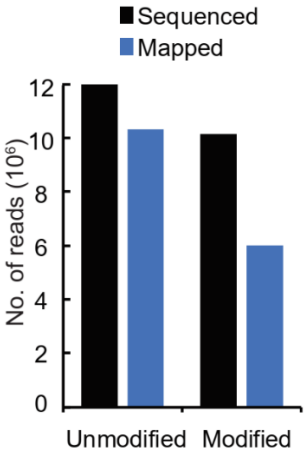


**Figure 4.2. Barplots representing the number of reads after basecalling and mapping in unmodified and modified hESC samples.** From the total number of sequenced reads, 86% of 12007032 and 60% of 10118432 reads from the unmodified and modified hESC samples were aligned to the human transcriptome respectively. Figure was published in *Aw. at el*[52].

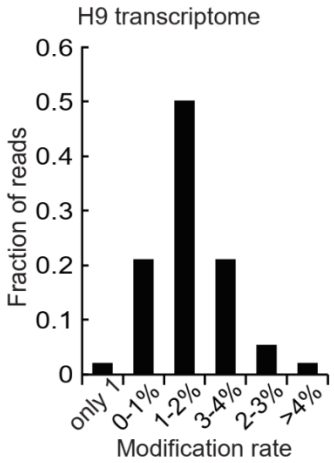


**Figure 4.3. Histogram showing the distribution of modifications on a single read in the hESC samples.** The percentage of modifications per strand was extracted and plotted in the histogram. Figure was published in *Aw. at el*[52].

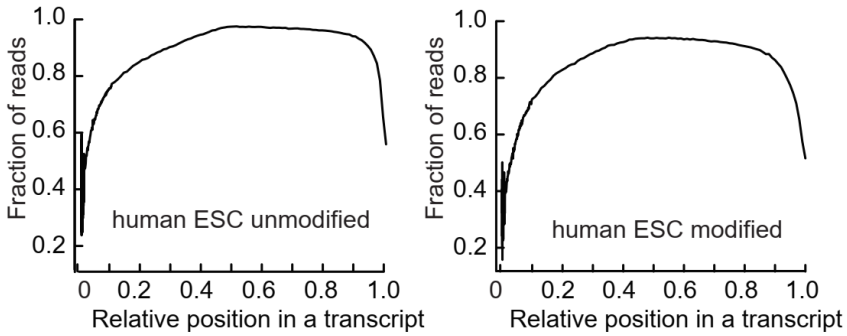


**Figure 4.4. Average coverage of reads along the hESC transcriptome using direct RNA sequencing.** The coverage from each transcript from unmodified (left) and modified (right) reads, were normalised by its length and an average value across all transcripts was calculated. Figures were published in *Aw. at el*[52].

Unlike using PORE-cupine to study single transcripts, where the number of reads was not limited, we needed to determine the minimum numbers of reads required for accurate structure determination in hESCs. From the sequenced Tetrahymena ribozyme RNA, we subsampled a range of unmodified and modified reads and compared the PORE-cupine results with the full dataset. The correlation increases with the number of reads used and begins to plateau at around $R$=0.8 with 200 reads of unmodified and 100 reads of modified RNA (**Figure 4.5**). At this threshold, we observed that transcript abundances and reactivity profiles of hESC RNAs were highly correlated across biological replicates (**Figure 4.6**), indicating that our data is reproducible (**Figure 4.7**). We obtained structural information for 1582 coding genes, 98 noncoding genes, 67 pseudogenes and 4 rRNAs across the hESC transcriptome after filtering for abundance (**Figure 4.8, 4.9**). The median length of the mapped reads from the hESCs transcriptome was 772 and 752 bases for unmodified and modified libraries respectively (**Figure 4.10**), with 37.9% and 42.8% of the unmodified and modified transcripts having greater than 90% of the annotated gene length respectively (**Figure 4.11**). As we are concerned about the accuracy of using truncated reads, we filtered for the sequenced reads that are near full length and compared the transcripts above the abundance threshold (>99% of annotated length, n=83, median, **Figure 4.12**). We observed that the results are highly correlated to the full dataset.

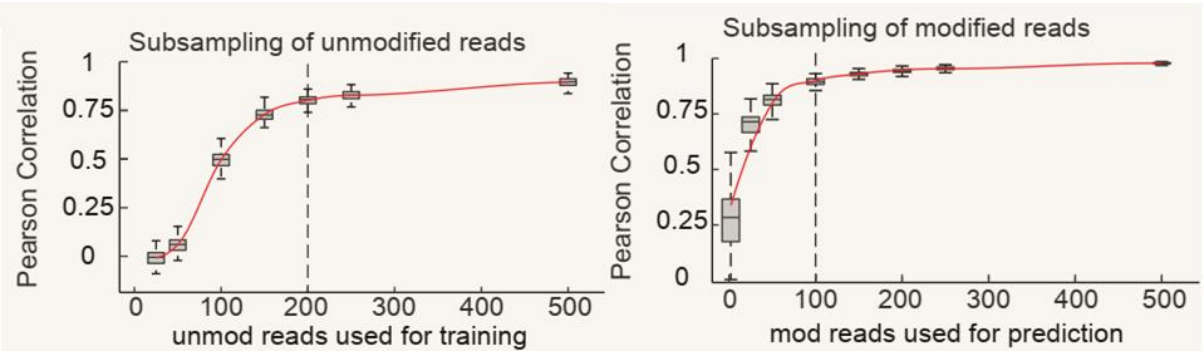


**Figure 4.5. Determining the number of reads required for accurate structure probing modification detection.** Left, Pearson correlation of Tetrahymena ribozyme RNA structural profiles obtained via subsampling of unmodified reads in comparison to the structural profile obtained using the full data set. The minimum number of unmodified reads used for downstream analysis was set to 200 (dotted line in grey). Right, Pearson correlation of Tetrahymena ribozyme RNA structural profiles obtained via subsampling of modified reads in comparison to the structural profile obtained using the full data set. The minimum number of modified reads used for downstream analysis was set to 100 (dotted line in grey); we subsampled 100× for each abundance. Figure was published in *Aw. at el*[52].

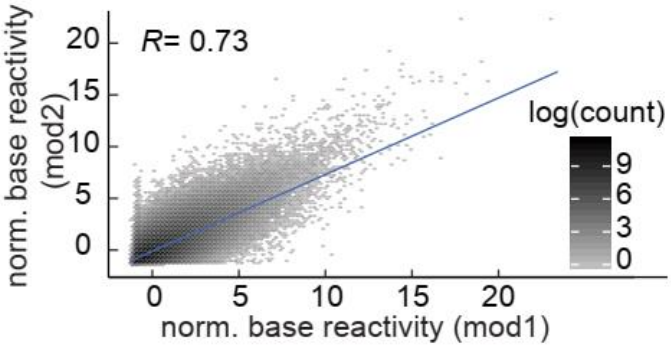


**Figure 4.6. Scatter plot of normalized reactivity between two biological replicates hESC libraries.** **P**earson correlation was used to calculate the similarity between the two signals; P = 0 and $CI_{95\%}$ = (0.73, 0.73) using two-tailed Student's t-test. n= 2 biological replicates were performed. Figure was published in *Aw. at el*[52].

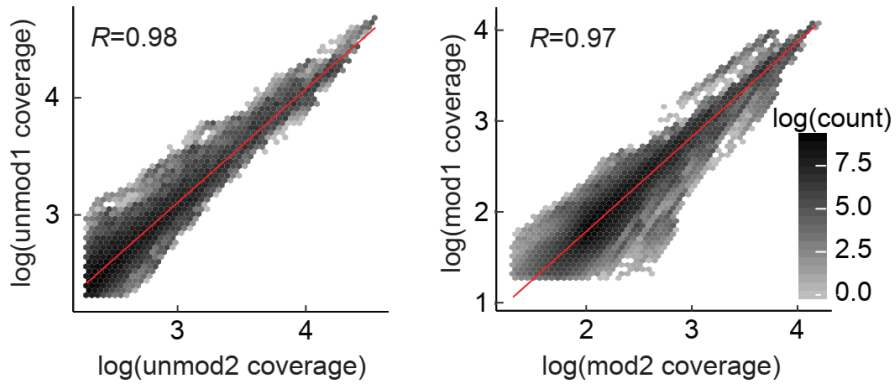


**Figure 4.7. Scatterplot showing the log coverage of aligned reads from the unmodified and modified hESC samples.** Transcripts from two biological replicates of unmodified (Left) and modified (right) from hESC samples were compared with Pearson correlation. Unmodified (R= 0.98, p-value =0, using two-tailed Student T-test, $CI_{95\%}$ = [0.98,0.98], 1613 transcripts) and modified transcripts (right, R=0.97, p-value =0 using two-tailed Student T-test, $CI_{95\%}$ = [0.97,0.97], 1751 transcripts). Figure were published in *Aw. at el*[52].

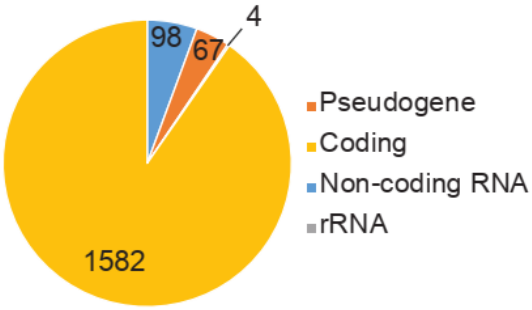


**Figure 4.8. Pie chart showing the number of genes belonging to different classes of transcripts captured in the hESC data set.** Figure was published in *Aw. at el*[52].

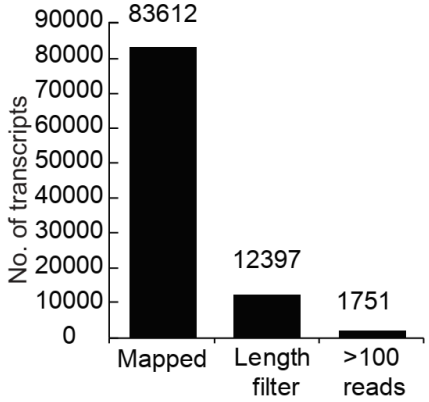


**Figure 4.9. Barplot showing the number of transcripts left after abundance and length filter.** The number of transcripts in each group is shown above the plot. Figure was published in *Aw. at el*[52].

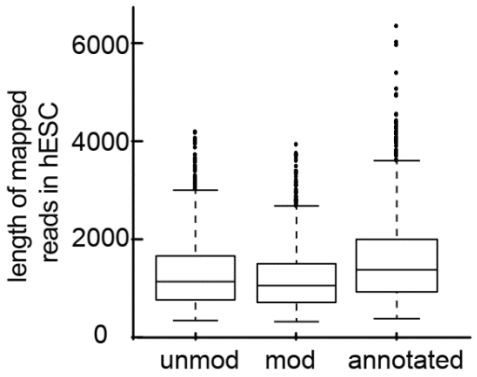


**Figure 4.10. Boxplots showing the distribution of median mapped lengths of mapped hESC reads and their expected length.** Aligned reads from unmodified (left) and NAI-N3 modified (middle) hESC mRNAs are shown (1751 transcripts). Annotated refers to the distribution of expected lengths for each transcript based on ENSEMBL GRCh38 annotation (right). Figure was published in *Aw. at el*[52].

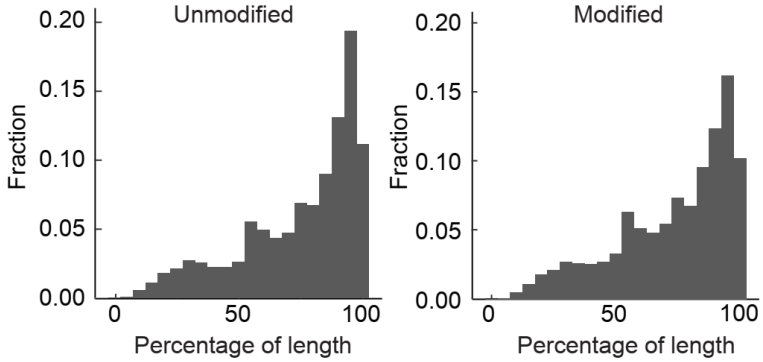


**Figure 4.11. Histogram showing the distribution of transcripts having different fractions of annotated length in unmodified and modified samples.** Figures were published in *Aw. at el*[52].

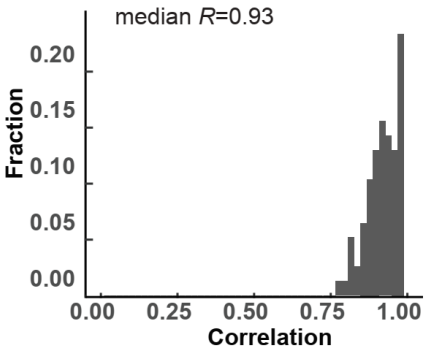


**Figure 4.12. Distribution of Pearson correlations between full-length (>99% of known length) and partial transcripts in hESC.** (83 transcripts from N=2 two biological replicates were used). The y-axis shows the fraction of transcripts with a particular correlation. The x-axis depicts Pearson correlation coefficients. Figure was published in *Aw. at el*[52].

# 4.2 Comparison of PORE-cupine with existing short-read methods

As we have just developed PORE-cupine, we had done extensive testing with mainly *in-vitro* transcribed RNAs, which might serve as a strong indication that PORE-cupine is accurate. Therefore, to further validate the results from PORE-cupine, we compared our results with other short-read high-throughput structure probing methods such as icSHAPE and SHAPE-MaP[37,38]. We compared the results of icSHAPE, SHAPE-Map and PORE-cupine, by filtering for the overlapped positions with high reactivity, where peaks above median reactivity was deemed as high reactivity. From 2684 genes, we observed that 38% of PORE-cupine's high reactivity positions overlapped with icSHAPE or SHAPE-MaP sites, while 36% of SHAPE-MaP high reactivity positions overlapped with icSHAPE or PORE-cupine sites, and 39% of icSHAPE high reactivity positions overlapped with SHAPE-MaP or PORE-cupine sites (**Figure 4.13, 3.14, Methods**). While the overlapping sites are low, they are consistent with previous observations on read-through versus RT stop methods[105] and point to the complementary range of various genome-wide structure probing methods in capturing different populations of single-stranded bases.

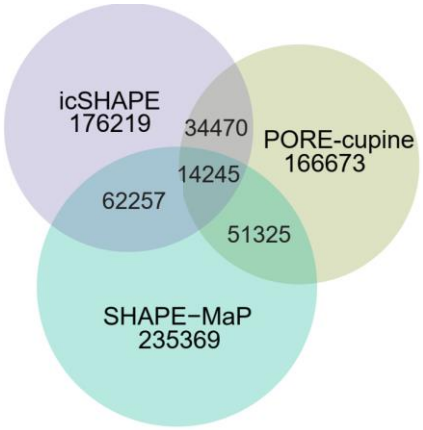


**Figure 4.13. Comparison of PORE-cupine to icSHAPE and SHAPE-MaP in the hESC transcriptome.** Venn diagram showing the overlap of high-reactivity sites identified by PORE-cupine, SHAPE-MaP and icSHAPE. Figure was published in *Aw. at el*[52].

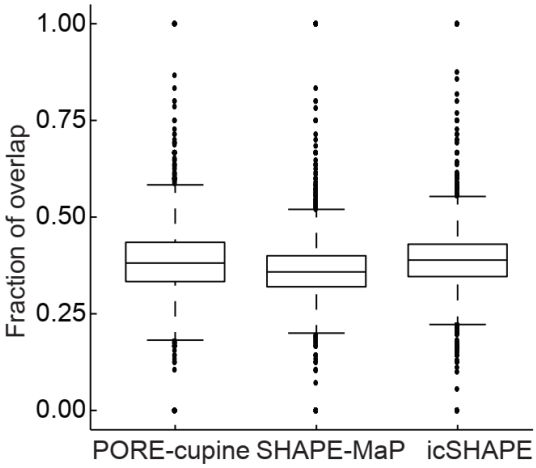


**Figure 4.14. Box plot showing the fraction of high-reactivity bases that were identified in PORE-cupine, SHAPE-MaP and icSHAPE and that also had high signals in at least one other method.** The comparison was performed at the gene level, using 3,037 genes and 1,617,397 positions. All signals were taken from the average of n= 2 biological replicates. The middle line of the box plot indicates the median, whereas the lower and upper boundary of the box plot corresponds to the first and third quartiles. Figure was published in *Aw. at el*[52].

We also compared the ability of PORE-cupine to capture global structural properties seen in other high-throughput structure-probing datasets by calculating the average reactivity signal in different RNA classes. As expected, we observed that rRNAs are the most structured, followed by lncRNAs and mRNAs, in agreement with the importance of structure for noncoding RNAs[104] (**Figure 4.15**). In addition, metagene

analysis of the reactivity of the mRNAs aligned by their translational start and stop sites showed the classic three nucleotides periodicity in their coding sequences (CDS), and not in their 5' and 3' UTRs[100,101,106] (**Figure 4.16**), highlighting PORE-cupine's ability to recapitulate known structural patterns in other datasets.

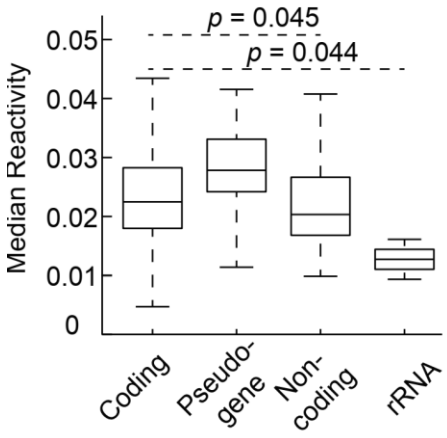


**Figure 4.15. Boxplot showing PORE-cupine reactivity of different classes of transcripts.** P-values were calculated using two-sided Wilcoxon Rank Sum test. 1584 coding genes, 67 pseudogenes, 81 non-coding genes and 4 rRNAs were used. reactivity between the profiles shown above. P-value was calculated using two-sided Student's t-test. Figure was published in *Aw. at el*[52].

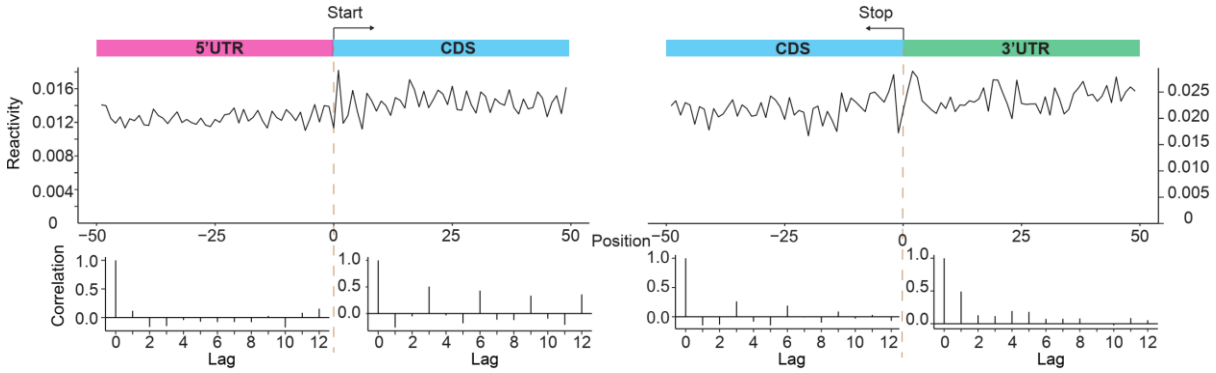


**Figure 4.16. Metagene analysis of PORE-cupine-derived mean reactivities.** Top, PORE-cupine-derived mean reactivities aligned according to start (Upper) and stop (Lower) codons for all 559 transcripts. Bottom, Metagene autocorrelation function (ACF) plot for the 5' UTR, CDS and 3' UTR. reactivity between the profiles shown above. P-value was calculated using two-sided Student's t-test. Figures were published in *Aw. at el*[52].

icSHAPE and SHAPE-MaP can also identify RNA binding protein (RBP) sites by detecting different reactivities inbound versus unbound positions[107]. Therefore, we determined if PORE-cupine could also detect RBP binding sites in our hESC data. We examined the reactivity profiles of potential Lin28 binding sites and used the results from High-throughput sequencing of RNA isolated by crosslinking immunoprecipitation (HITS-CLIP) as evidence for binding in hESCs[108] (**Methods**). We observed that Lin28 binding sites with HITS-CLIP evidence showed an increase in reactivity in the bases flanking the binding motif and a decrease in reactivity within the binding motif when compared with sites without HITS-CLIP evidence (**Figure 4.17).** This indicates that real Lin28 binding sites are more structurally accessible around the motif and that Lin28 binding likely prevents NAI-N3 from modifying the RNA sequence. Our results were consistent with the result from icSHAPE.

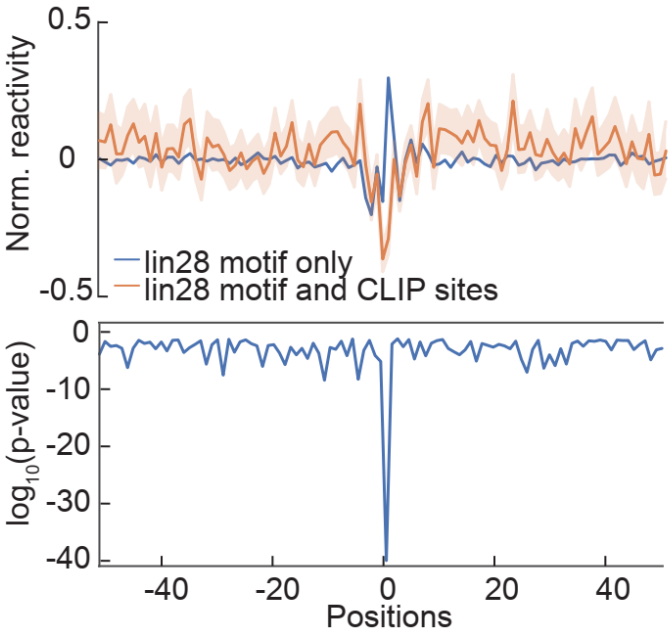


**Figure 4.17. Metagene analysis of reactivity profiles for transcripts with and without HITS-CLIP evidence for Lin28 binding**. Top, Mean reactivity profiles for transcripts with (orange) and without (blue) HITS-CLIP evidence, centred at the binding motifs. Bottom, $\log_{10}$(P-value) of the difference in reactivity between the profiles shown above. P-value was calculated using two-sided Student's t-test. Figures were published in *Aw. at el*[52].

## 4.3 Detecting structural differences in between single nucleotide variations SNVs

Besides RBP binding, previous studies have also shown that single nucleotide variations (SNVs) can result in structural changes along an RNA[104]. To determine whether PORE-cupine could identify structural changes in transcripts with SNVs, we first identified SNVs in the hESC transcriptome using Illumina RNA sequencing data (**Methods**). We then separated mapped direct RNA sequencing reads based on the different alleles observed. We identified 90 transcripts with two or more SNVs and sufficient coverage for reactivity analysis: 10/90 SNVs were observed to result in statistically significant reactivity changes (11.1%, Fisher's exact test, **Methods**). Metagene analysis of the reactivity profiles across alleles showed that the largest reactivity differences occur locally and extend up to 25 bases upstream and downstream of the SNV location (**Figure 4.18, 4.19**).

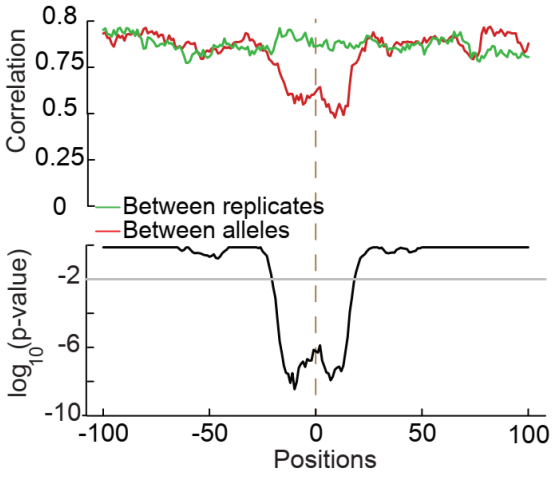


**Figure 4.18. Metagene analysis of correlation of reactivity profiles centred around SNV positions.** Top, Correlation of reactivity profiles +/- 100 bases SNV positions for reads from different alleles from the same biological replicate (red) and reads from the same allele but different biological replicates (green). SNVs result in local reactivity differences up to 25 bases upstream and downstream of the SNV site. Bottom, $\log_{10}$(p-value) of the difference in metagene profiles between red and green lines on top. Pearson correlation values were computed by taking 30-nucleotide sliding windows across each transcript, and p-values were calculated using two-sided Fisher's exact test. Figures were published in *Aw. at el*[52].

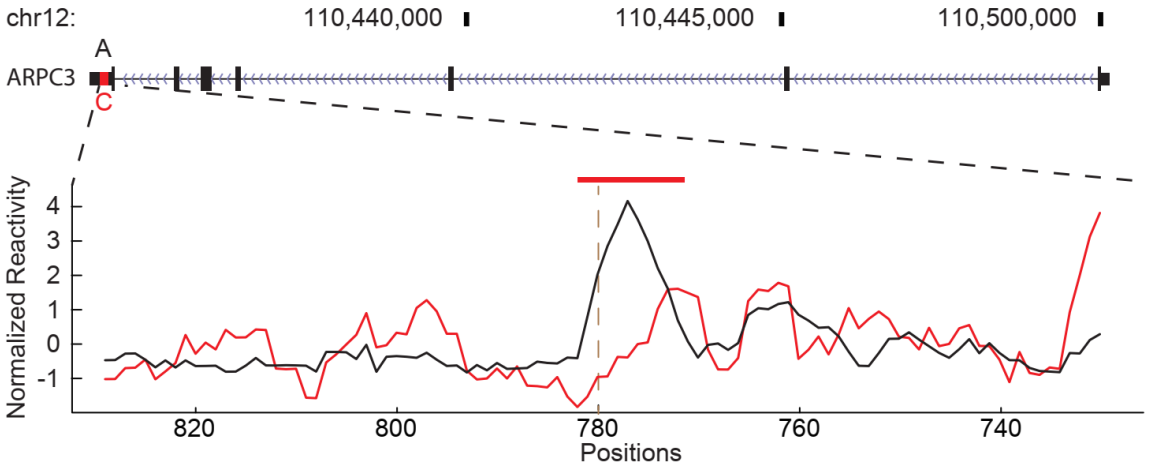


**Figure 4.19. An example of the reactivity difference due to A to C allele change in ARC21 transcript.** The black and red lines represent normalized reactivity of the A and C allele respectively around the SNV. The dotted brown line on the line plot represents the location of the SNV, red bars represent above the reactivity profiles the positions that have significant change between the two profiles (**Methods**). Figures were published in *Aw. at el*[52].

# 4.4 Detecting structural differences in shared exons from alternative isoforms

Using PORE-cupine, we obtained structural information of individual isoforms from the hESCs. We obtained 104 genes (corresponding to 204 pair-wise transcript comparisons) that had two or more isoforms for downstream analysis (**Figure 4.20**). We observed that the majority of the isoforms pairs (178 out of 204, 87%) showed reactivity differences in shared regions (**Methods**). Globally, this is reflected in lower reactivity similarities in shared sequences across the isoforms pairs as compared to biological replicates of the same transcripts (**Figure 4.21** ). In general, there is a weak positive correlation between the sequence similarity and their structures among the isoform pairs and the correlation is stronger when there are two or more alternative splice sites along a transcript (**Figure 4.22** ). While the biggest reactivity differences occur around the alternative splice site, 70% of isoforms contain both local and distal (>200bp away) reactivity changes relative to splice sites (**Figure 4.23**). We confirmed this distal impact on structures by showing that identical sequences that are far away from an alternative splice site show a lower reactivity correlation between isoforms

pairs between the same replicate than between identical transcripts across biological replicates (**Figure 4.2**).

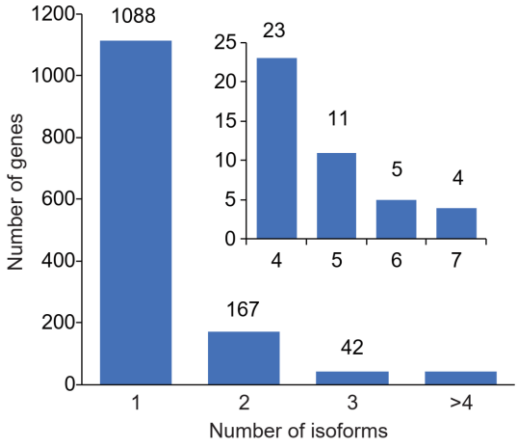


**Figure 4.20. Histogram showing the distribution of structure-probed genes according to the number of isoforms present.** The values for each group are shown above. Figures were published in *Aw. at el*[52].

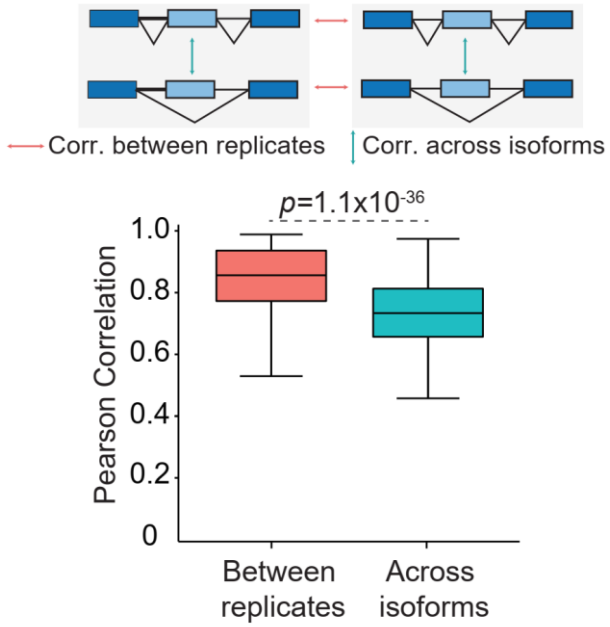


**Figure 4.21. Metagene analysis of correlation of reactivity between replicates and across isoforms.** Top, schematic showing pairwise structure comparisons between 1) biological replicates of the same transcript and 2) different isoforms of the same gene. Bottom, box plot showing the distribution of structural similarity for the same isoform across biological replicates (salmon) or between different isoforms within the same biological replicate (teal). In total, 204 transcripts were compared. Structure similarity was calculated using the Pearson correlation. P values for comparison between the two distributions were calculated using the two-sided Wilcoxon rank-sum test Figures were published in *Aw. at el*[52].

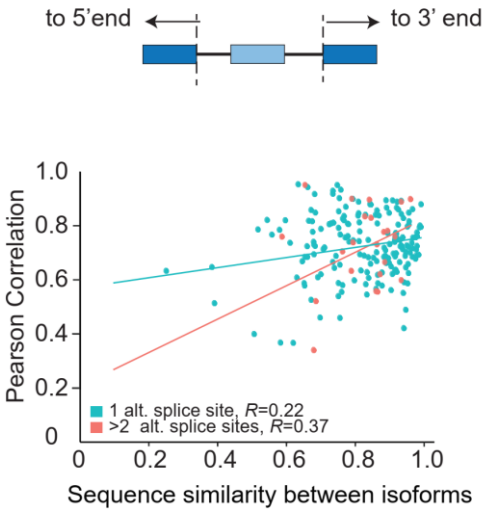


**Figure 4.22. Scatter plot of structural similarity between shared exons of different isoforms versus their sequence similarity.** The shared exons and the sequence similarity was based on the published human transcriptome (Ensembl version GRCh38.93). Pearson's R is shown (1 alt P = 0.003 using two-tailed Student's t-test, $CI_{95\%}$ = (0.074, 0.35); >2 alt P = 0.089 using two-tailed Student's t-test, $CI_{95\%}$ = (−0.06, 0.68)). In total, 182 transcripts with one alternate splice site and 22 transcripts with more than two alternate splice sites are compared to each other. Figures were published in *Aw. at el*[52].

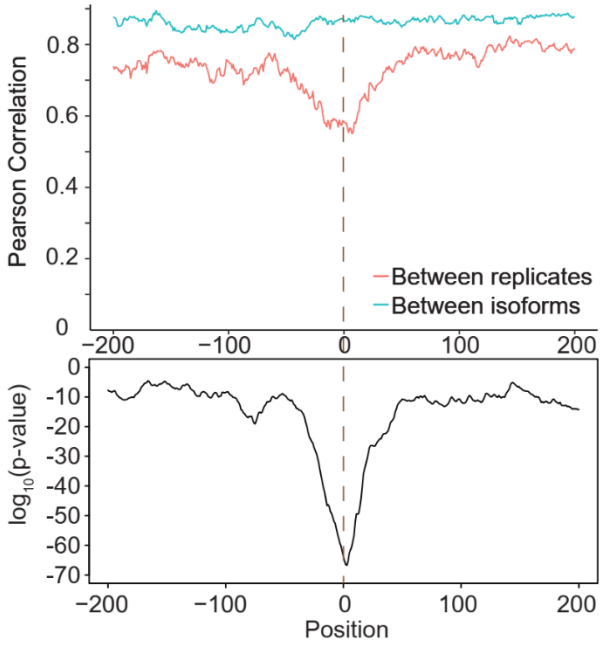


**Figure 4.23. Metagene analysis of reactivity similarity between alternatively spliced isoforms centred at the alternative splice site.** Top, The line plot for the correlation of reactivity similarity between alternatively spliced isoforms (blue), as well as reactivity similarity between biological replicates (red) centred at the alternative splice site. Bottom, $\log_{10}$(P-value) of the difference in the blue and red lines above. P values were calculated using two-sided Fisher's exact test. Figures were published in *Aw. at el*[52].

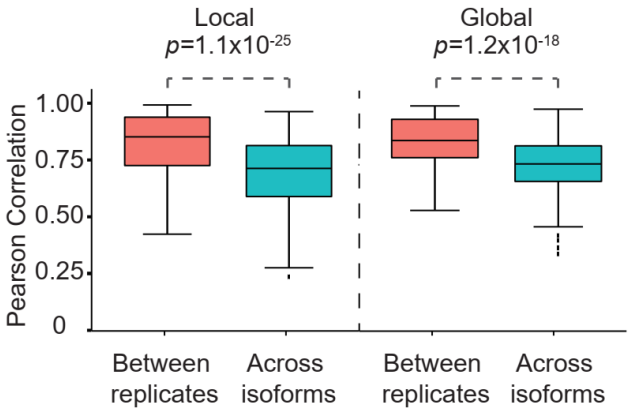


**Figure 4.24. Box plot showing the distribution of reactivity similarity for the same isoform across biological replicates or between the biological replicate of the same transcripts.** The distribution of correlation between isoforms of biological replicates (salmon) and across different isoforms (teal), for local contexts (left) and global excluding local contexts (right). Global excluding local contexts are defined as regions extending from both sides of the alternative splice site to the 5′ and 3′ ends of the transcript, excluding 200 bp to the right and left of the alternative splice site. In total, 204 transcripts were compared globally. Local contexts are defined as 50 bp to the left and right of the alternative splice site; 226 transcripts were compared at the local level. P-values were calculated using two-sided Wilcoxon rank-sum test. Figures were published in *Aw. at el*[52].

## 4.5 PORE-cupine can phase structures along isoforms

The presence of two or more alternative structures that reside and span identical sequences makes it particularly challenging for short-read sequencing to determine which combinations of RNA structures co-exist in an isoform (**Figure 4.25, 4.26**). An example of this is RPS8, which is alternatively spliced into two isoforms that share identical sequences for three exons near the 3’ end but are alternatively spliced near the 5’ end (**Figure 4.25a**). PORE-cupine analysis shows that the two isoforms contain different structures (A1 versus A2, and B1 versus B2) that are separated by ~400 bases from each other in the shared sequences. In short-read sequencing, the lack of connectivity between structures A and B makes it difficult to know whether A1 is linked to B1 or B2 in the blue isoform and vice versa. Conversely, PORE-cupine enables us to link and correctly assign structural information to their individual isoforms in shared regions (**Figure 4.25b**). Globally, 36.4% of the transcripts contain two structure-changing regions that are more than 200 bases apart (**Figure 4.27**), demonstrating

the importance of PORE-cupine for phasing structures along long isoforms by providing connectivity in RNA structure information across the transcriptome.

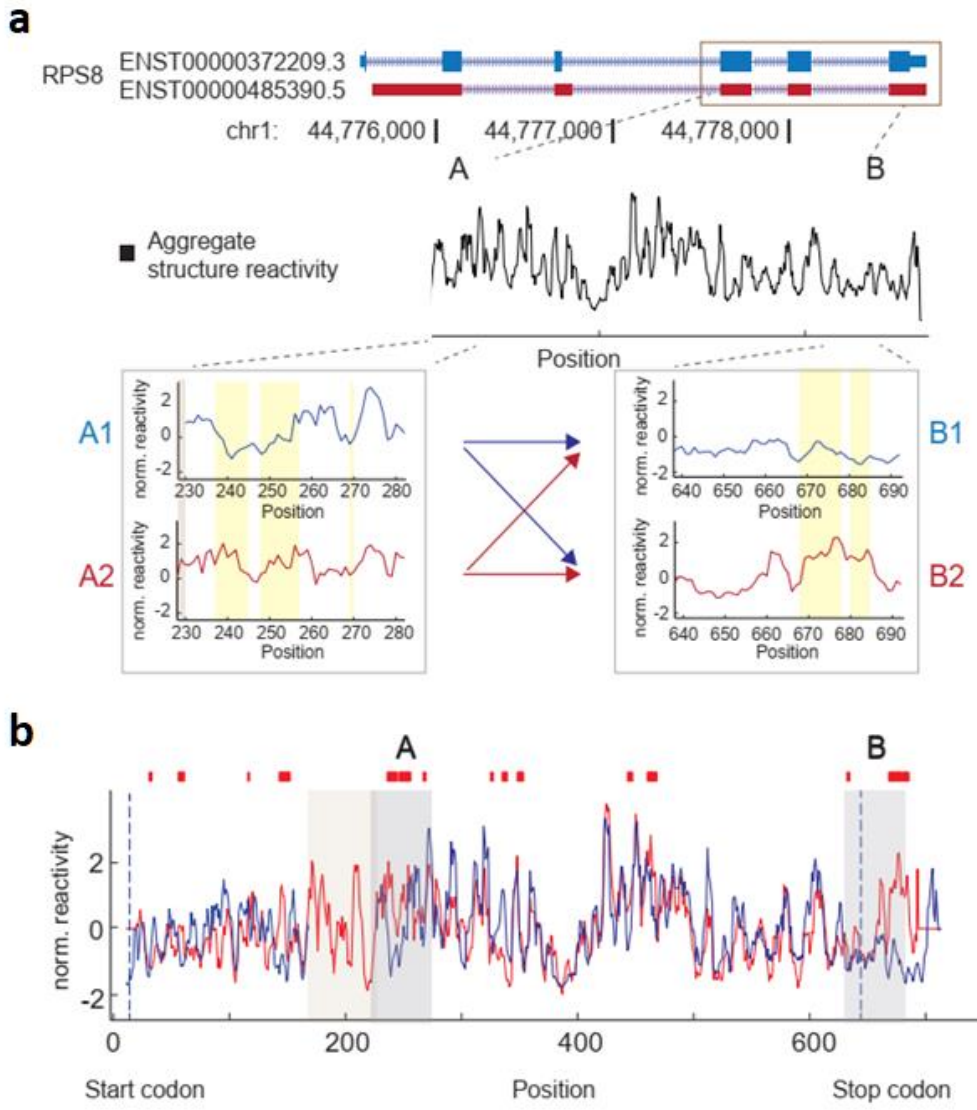


**Figure 4.25. Structural information from two isoforms of RPS8. a**, (Top) exon and intron organization displayed with their respective Ensembl spliced transcript IDs. An alternative exon seen in our structural data is coloured in brown. The three shared exons near the 3′ end of the gene are boxed. (Bottom) normalized reactivity profiles for the aggregate signal of the isoforms in the three shared exons. The two structure-changing regions (A and B) between the isoforms identified by PORE-cupine are boxed. **b**, Normalized reactivity profiles across the entire length of the different isoforms. PORE-cupine could assign the correct structures to the different isoforms due to long-read sequencing. Red bars represent above the reactivity profiles the positions that have significant change between the two profiles (Methods). The grey filled-in boxes indicate the boxed structure regions (A and B) in **a**. The beige filled-in bar indicates the location of the retained intron. Figures were published in *Aw. at el*[52].

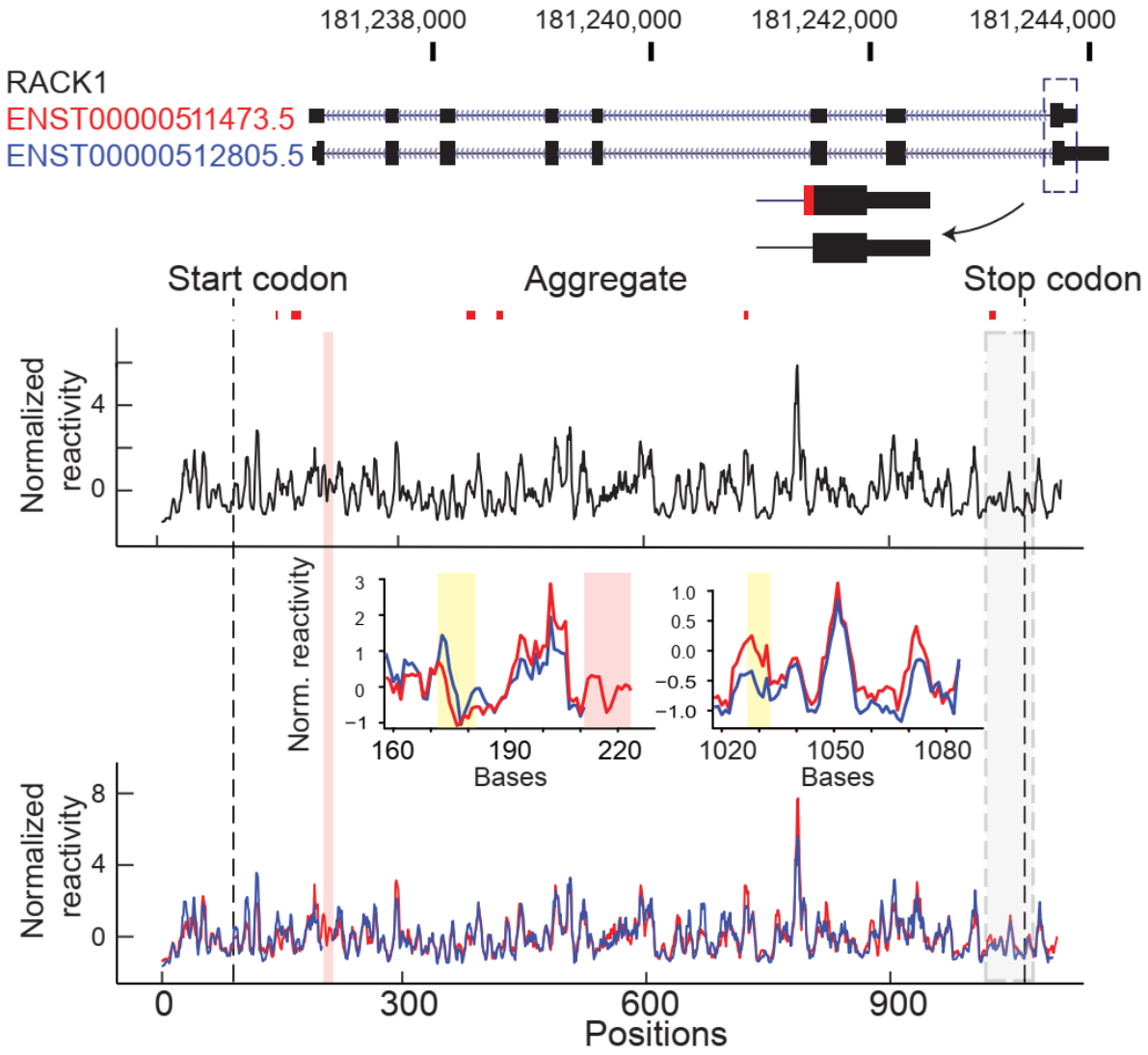


**Figure 4.26. Structural information from two isoforms of RACK1.** Upper, Transcript organization of different RACK1 isoforms. Alternative exon is shown in red (also in inset). Lower, Line plots for the aggregate reactivity signal between the two isoforms are shown (Top). Middle, Line plots showing the expanded view of the reactivity difference between the isoforms. Bottom, Line plots showing the individual reactivity information for each isoform along its length. Figures were published in *Aw. at el*[52].

.

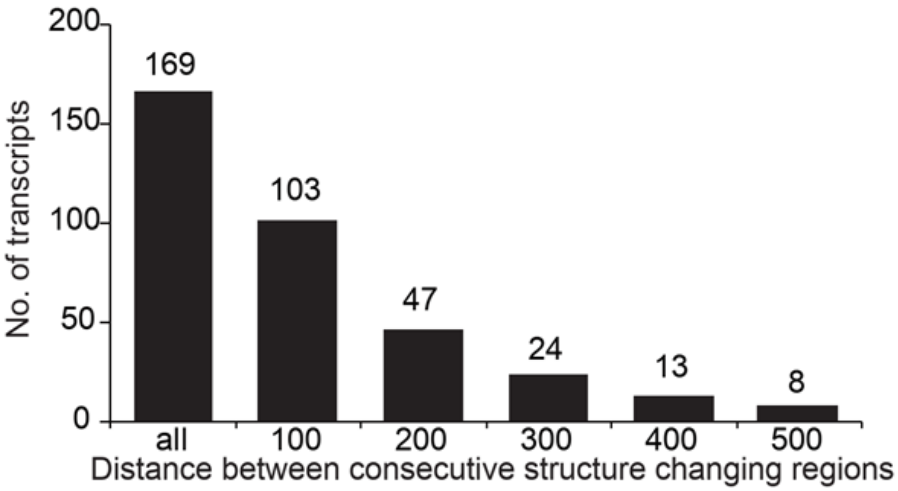


**Figure 4.27. Bar plot showing the distance between structural changing bases.** The number of transcripts with two structure changing regions that are more than 100, 200, 300, 400 or 500 bases apart. The value for each group is shown above the bar. Figure was published in *Aw. at el*[52].

## 4.6 Isoforms with structural differences show differences in translation efficiency

Different RNA structures are used to regulate gene expression, during translation, splicing and decay[29]. To determine whether structural differences between isoform pairs could regulate translation, we performed TrIP-seq on hESCs to analyze the distribution of isoforms across a polysome gradient[84]. Isoforms that are found predominantly in higher polysome fractions are typically associated with more ribosomes and have higher translation rates, although RNAs could also be associated with other high molecular weight complexes in high polysome fractions (**Figure 2.2, 4.27**). We obtained a high degree of correlation between two biological replicates across polysome fractions (**Figure 4.28**) and observed that highly translated transcripts are found in high polysome fractions, while poorly translated transcripts are found in low polysome fractions as expected (**Figure 4.29**), indicating that our TrIP-seq data provides an accurate reflection of mRNA-polysome association.

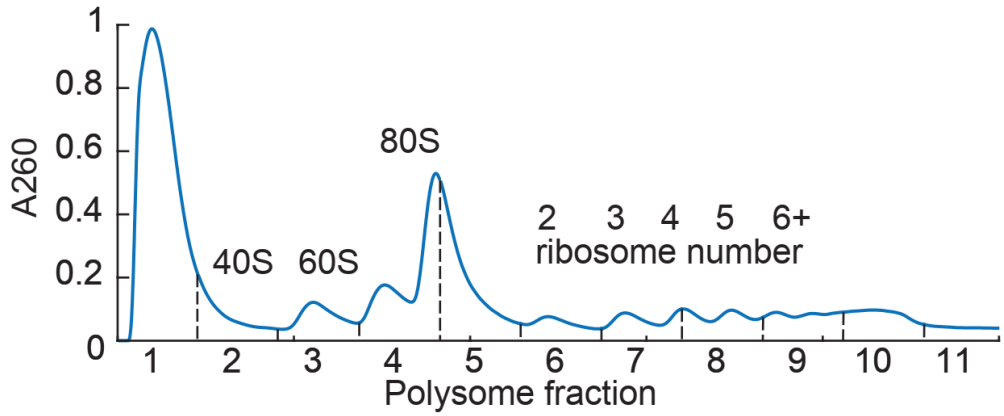


**Figure 4.27. Line plot showing the absorbance A260 of each fraction (2-12) after polysome fractionation.** Figure was published in *Aw. at el*[52].

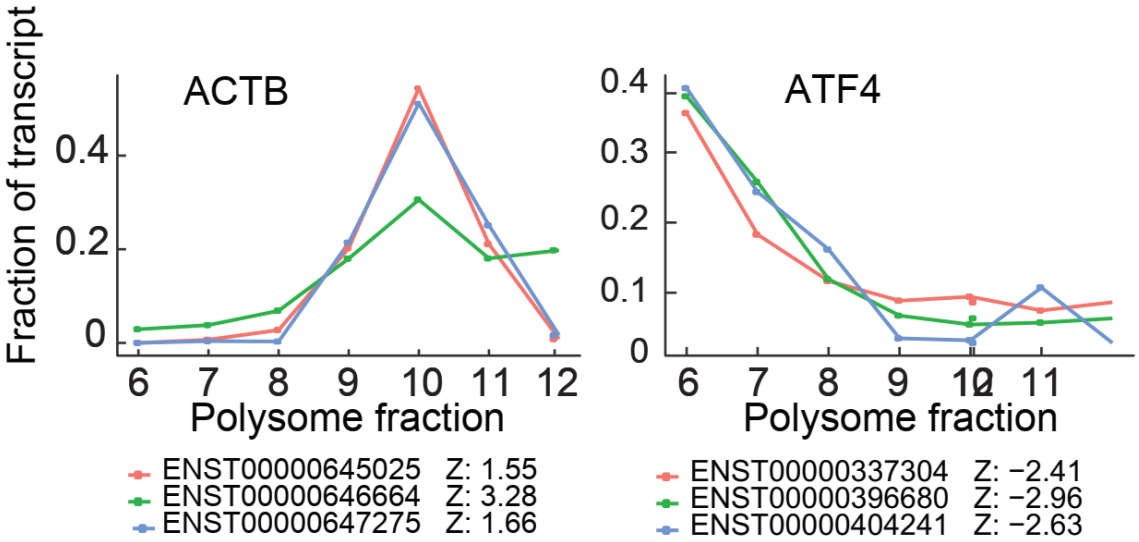


**Figure 4.28. Distribution of read-counts across different polysome fractions for two biological replicates.** Line plots for Actin B (left) and Activating transcription factor 4 (right) are shown. Figures were published in *Aw. at el*[52].

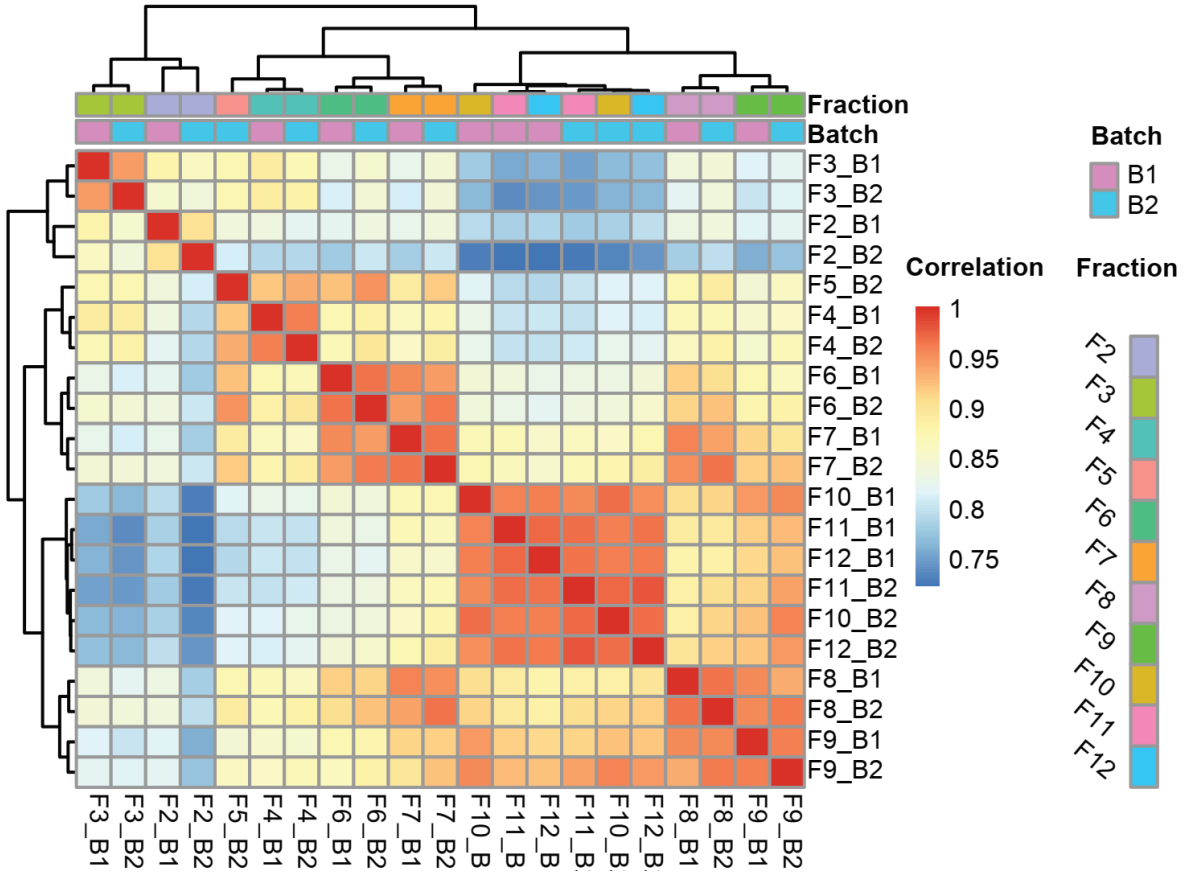


**Figure 4.29. Heatmap of the pair-wise correlations between different replicates and fractions.** Spearman correlation was applied of the read-counts/transcript for each fraction between two biological replicates. Fractions and batches are denoted as F2-12 and B1-2 respectively. Figure was published in *Aw. at el*[52].

Out of 178 structure changing isoform pairs, 153 pairs have polysome fractionation data. We observed that 28 pairs showed changes in translation efficiency by TrIP-seq (18.3%, Fisher's Exact test, **Methods**). Structurally similar isoform pairs are translated at similar rates while structurally more divergent pairs show greater differences in their translation, suggesting that isoform-specific structures could impact isoform-specific translation (**Figure 4.30**). We observed that one of the isoform pairs of RPL17 showed reactivity differences in shared regions based on PORE-cupine analysis, as well as translation efficiency differences. The transcript ENST00000618619.4 (RPL17_1) is highly translated while ENST00000579408.5 (RPL17_2) is poorly translated and contains a retained intron of 161 bases in the 5' UTR (**Figure 4.31, 4.32**). To study how this retained intron resulted in structural changes in the 5' UTR and translation repression, we examined pair-wise RNA interactions in this region from a previously

published dataset that uses proximity ligation sequencing (*S*equencing of *P*soralen crosslinked, *L*igated, and *S*elected *H*ybrids, SPLASH)[35]. SPLASH reads showed strong interactions between the retained intron and sequences upstream and downstream to it, resulting in an extensively structured environment around the start codon (**Figure 4.31**)**.** In the absence of the retained intron (RPL17_1), the isoform folds into a simpler structure, allowing the start codon to be more accessible for translation.

To experimentally validate that the poor translatability of RPL17_2 is indeed due to extensive structures formed by the retained intron around the start codon, we cloned the 5' ends of RPL17_1 and RPL17_2 in front of a luciferase reporter and performed mutagenesis experiments on RPL17_2 (**Figure 4.33, Methods**). We confirmed that RPL17_1 indeed translates much better than RPL17_2, as shown by greater than 10-fold increase in luciferase units upon RNA transfection (**Figure 4.34**). Mutations that disrupt the pairwise interactions of 3 different stems (1.1 or 1.2, 2.1 or 2.2, 3.1 or 3.2, **Figure 4.33**) open the structures around the start codon and increase the translatability of RPL17_2 while compensatory mutations that restore the helical structures partially rescue the poor translatability of RPL17_2 (**Figure 4.34**). These results confirm that structure plays an important role in regulating isoform-specific translation of RPL17.

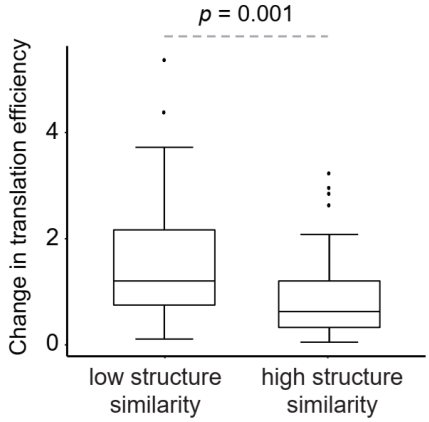


**Figure 4.30. Box plot showing that gene-linked isoforms with greater structural similarity show smaller differences in translation efficiency.** P-value was calculated using the two-sided Wilcoxon rank-sum test (P= 0.001). In total, 43 transcripts with low structure similarity and 48 transcripts with high structure similarity were used for the comparison. Figure was published in *Aw. at el*[52].

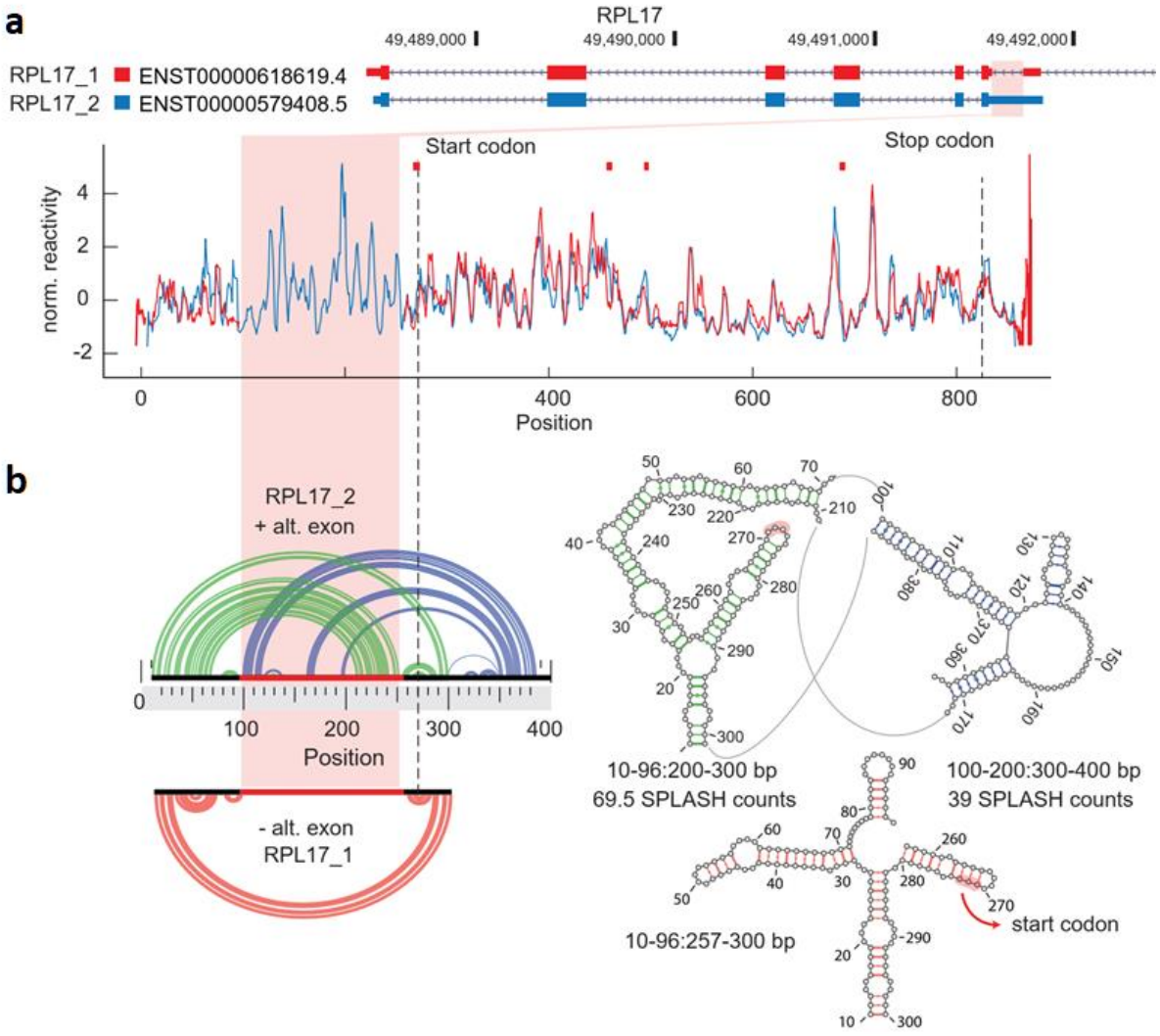


**Figure 4.31. Structural information from a pair of isoforms from RPL17 that show structure and translational differences. a**,Top, exon and intron organization displayed with their respective Ensembl spliced transcript IDs. The alternative exon seen in our structural data is highlighted in red. Bottom, normalized reactivity profiles for the two gene-linked isoforms. The red bars on top of the line lots indicate positions of significant structure changes ($P < 0.05$, Fisher's exact test; Methods). **b**, Left, the upper structure (blue and green) is based on RNAcofold-calculated interactions for 10–400 bp, whereas the lower structure (red) is based on the RNAcofold-calculated interactions between regions 10–96 bp and 257–300 bp. Right, RNAcofold-derived structures for the region around the alternative exon. The SPLASH read counts are indicated below the structures. The start codon is highlighted (red) on the structures. Figures were published in *Aw. at el*[52].

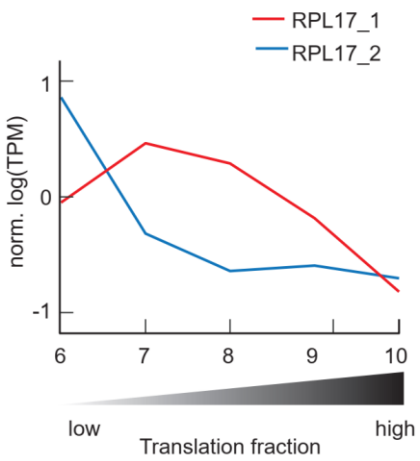


**Figure 4.32. Line plot for the normalized log(TPM) expression level for the two RPL17 isoforms in fractions 6–10.** P-value = 4 × 105 (Fisher's exact test; Methods).

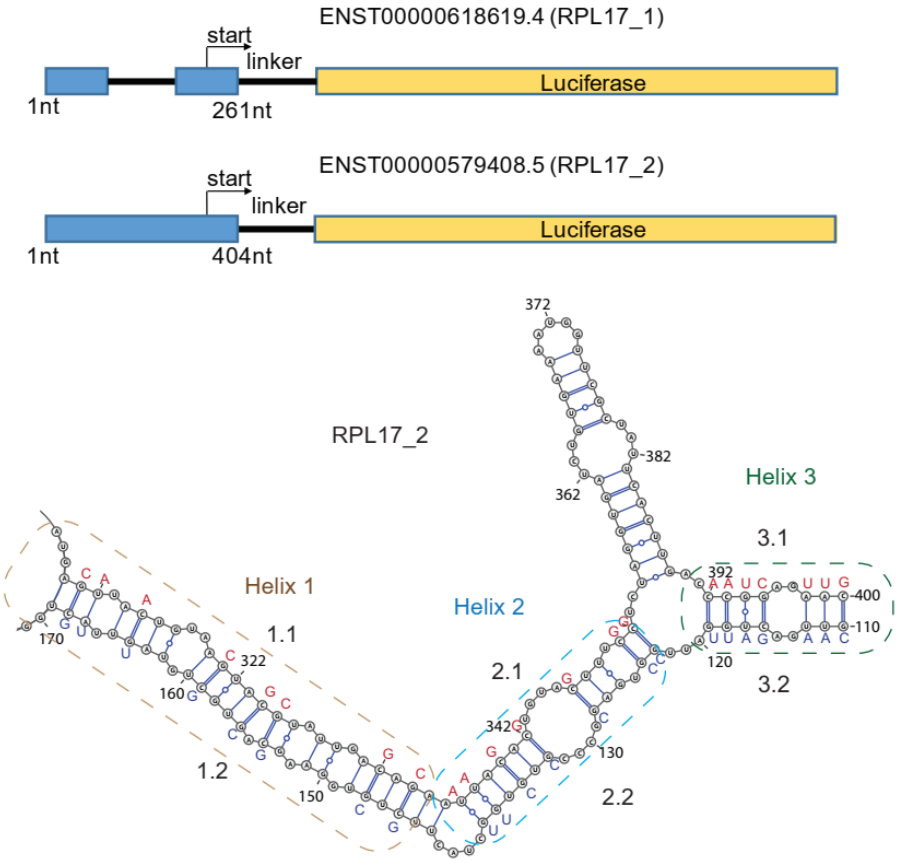


**Figure 4.33. Schematic of the design of the fusion RPK17 RNAs.** Top, design of the fusion RNAs used in the luciferase assay. Bottom, expanded view of the predicted secondary structure of RPL17_2, with the mutations shown next to the original base. We performed mutations along three different stems (helixes 1–3) of the structure, first on each side of the structure (1.1/1.2, 2.1/2.2 and 3.1/3.2) to disrupt the helixes and then on both sides of the structure to restore the helixes (compensatory mutations 1.1 + 1.2, 2.1 + 2.2 and 3.1 + 3.2). Figures were published in *Aw. at el*[52].

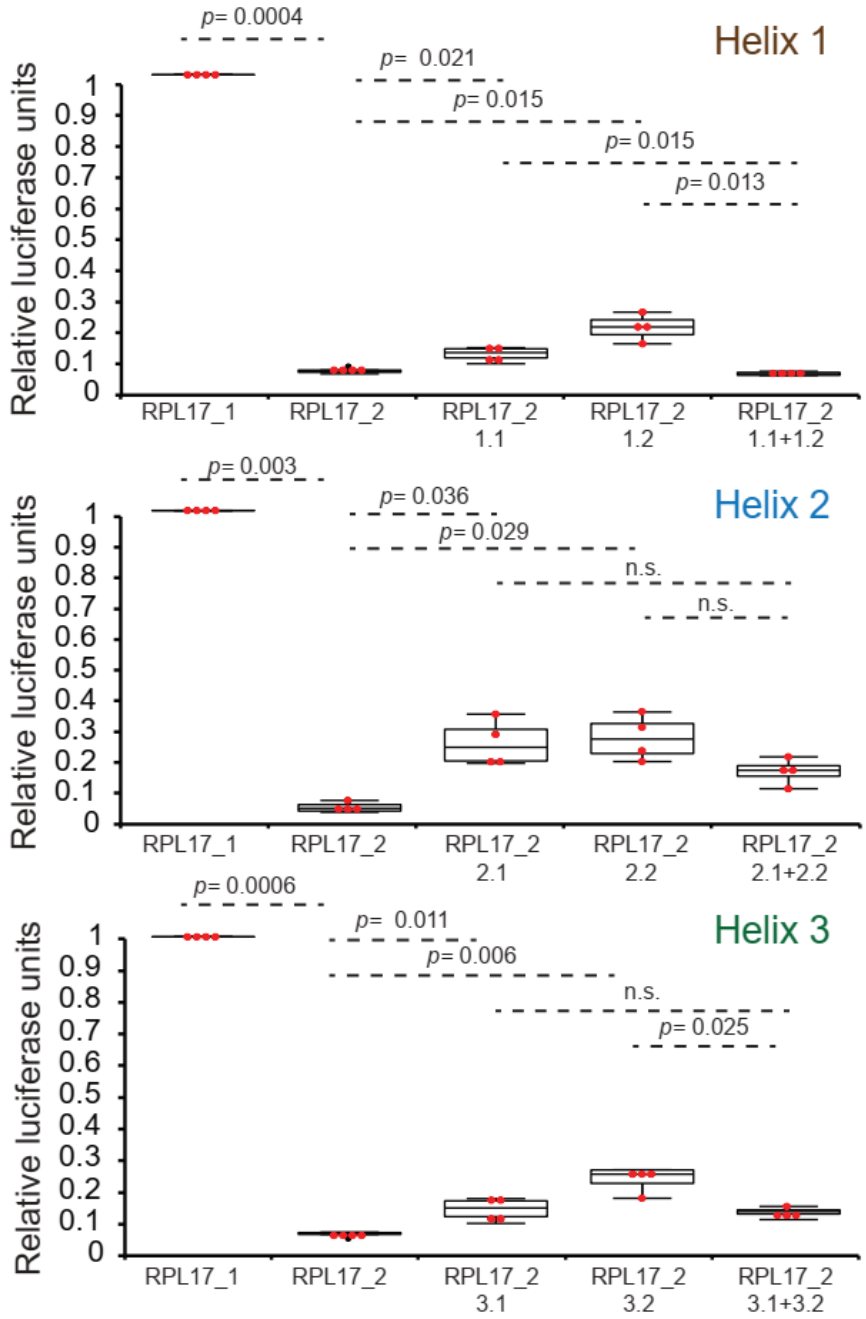


**Figure 4.34. Box plots showing the luciferase activity of fusion RNAs containing 5′ ends of RPL17 isoforms and their mutants.** The results were obtained from 8 or 16 h after transfection in 293T cells. We performed four biological replicates for each experiment. Red dots represent the average of technical replicates for each biological replicate. P values were calculated using the two-sided Student's t-test with n= 4 for each condition. RPL17_1 shows greater than 10-fold increase in luciferase activity as compared to RPL17_2. Structure mutations that disrupt the helical stems of RPL17_2 increase luciferase activity, whereas compensatory mutations that restore the structure partially rescue the low luciferase activity. Figures were published in *Aw. at el*[52].

## 4.7 Conclusion

Using PORE-cupine, we have obtained the structure information of 1751 transcripts from around 20 million sequencing reads. We observed that for the sequenced reads from both test and hESC, the median lengths in unmodified samples are longer and was able to align to the reference with less error when compared to the modified samples. We also compared the structure information obtain from PORE-cupine, icSHAPE and SHAPE-Map, and observed that overlapping results across all methods were low. In addition, PORE-cupine had captured the global structural properties of the difference classes of RNA, which were consistent with previous observations on read-through versus RT stop methods. The structural information obtained from PORE-cupine allows us to determine and phase RNA reactivities in isoform pairs. However, isoforms that only differ at the extreme 5' or 3' ends of the transcripts limits the ability to uniquely map reads to individual gene-linked isoforms due to 5' and 3' end decay. We focused our analysis on mRNA gene-linked isoforms due to controversies on whether lncRNAs could be translated. We observed that many isoform pairs exhibit reactivity differences in shared regions and that this is associated with changes in translational efficiency using polysome profiling. While polysome profiling data is not a perfect proxy for translation as RNAs could also be associated with other RBPs that reside in different polysome fractions, we do observe that highly translated RNAs such as ACTB and poorly translated RNAs such as ATF4 are in high and low polysome fractions respectively in our data. Furthermore, our luciferase experiments on a poorly translated verses highly translated isoform validated the polysome fractionation results, suggesting that the rate of translation largely correlates with the fraction number. Lastly, our mutational experiments on RPL17 further demonstrate the importance of isoform-specific structure in regulating translation. PORE-cupine expands our current repertoire of RNA structure probing strategies to deepen our understanding of the role of RNA structure in isoform-specific gene regulation[37,100,101,109].

# Chapter 5 The structural differences between the different subgenomic sequences within or across two SARS-CoV-2 strains

Coronaviruses (CoVs) are enveloped viruses with positive-sense single-stranded RNA genomes. They can infect animals and people, and usually cause mild respiratory and intestinal symptoms[110]. However, in these past two decades, highly pathogenic strains have emerged and led to three major outbreaks. SARS-CoV and MERS-CoV caused the 2002-2004 and 2012 outbreaks respectively, and the most recent strain SARS-CoV-2 caused the current pandemic[111]. Since SARS-CoV-2 was first reported in Wuhan, China in December 2019, it has resulted in over 160 million infections and more than 3 million deaths as of 20$^{th}$ May 2021, according to WHO. Throughout the SARS-CoV-2 pandemic, different variants of the virus have been emerging from patients. Notably, multiple strains that contain deletions of various sizes in the ORF8 region have been found circulating globally, including in Singapore, Taiwan, Bangladesh, Australia and Spain[91]. In particular, a 382-nucleotide deletion (Δ382) of the SARS-CoV-2 genome that truncates ORF7 and deletes ORF8 was found in patients in Singapore[112]. While patients infected with the Δ382 virus showed less severe symptoms than those infected with wild-type (WT) viruses, the molecular mechanisms behind virus attenuation in patients are unclear[91]. Hence, we would like to understand how the SARS-CoV-2 and their variants differ in their RNA structures.

CoV genomes are among the largest of the RNA viruses, with lengths of 26-32 kb[113]. Upon entry into the cell, the positive-sense genome is translated from two open

reading frames (ORF1a and ORF1b) and the resulting polyproteins are cleaved into non-structural proteins. A unique feature for the coronaviruses is that it generates numerous sub-genomic RNA (sgRNA) species that are translated to its accessory proteins, where each sgRNA contains a common leader sequence, followed by a nested set of sequences[94,114,115] (**Figure 5.1**).


**Figure 5.1 Schematic of the SARS-CoV-2 subgenomic RNAs.** The figure was published in Siwy at el[116].

We applied PORE-cupine to investigate the secondary structures along with the WT and Δ382 SARS-CoV-2 genomes to identified sgRNA-specific structures as well as the differences in the structures between the two strains. Due to the nested nature of the sgRNA transcripts, large portions of their sequence overlaps with each other. Thus, making it difficult to obtain the structures of the individual sgRNAs structures with short-read structure methods, as the major ity of the sequenced reads cannot be accurate be assigned to the correct sgRNA transcript. Therefore, using PORE-cupine would enable us to accurately identify the sequenced reads, allowing us to look at the individual sgRNAs structures. Together with the interaction data generated from SPLASH[92], a method to probe for long-range interactions, we hoped to have a comprehensive understanding of the structural differences. All data in this section will be published in the journal of Nature Communications[116].

## 5.1 SARS-CoV-2 sgRNAs are found to be structurally different

In addition to the synthesis of the full-length SARS-CoV-2 genome, a nested set of 3' co-terminal sgRNAs are transcribed in SARS-CoV-2 infected cells using discontinuous RNA synthesis[94] (**Figure 5.1**). These sgRNAs ranges from 2-8 kb long, contain a leader sequence and have a different level of transcription efficiency. Although methods such as SHAPE-Map or ic-SHAPE provides the reactivity score across the genome, short-read sequencing makes it difficult to map structure information unambiguously to individual sgRNAs. As such, it is unclear if there are structures that are unique to the individual sgRNA transcripts that could be important for sgRNA-specific regulations.

To address this issue, we utilize PORE-cupine to allow us to determine the structures of individual sgRNAs, as we can filter and assign the full-length reads to the individual sgRNAs[19]. We sequenced two biological replicates of RNAs extracted from NAI-treated, WT and Δ382 SARS-CoV-2 infected Vero cells. We filtered and assigned the sequenced reads to the individual sgRNAs and determined the reactivity scores along individual ORF3a, E, M, ORF6, ORF7a, ORF7b, ORF8 (WT only) and N transcripts (**Methods**, **Figure 5.2, 5.3**). We compared the reactivity obtained from the two replicates and found that most sgRNAs were highly correlated, with ORF3a and E of moderate correlation (**Figure 5.3** R: 0.57-0.98), indicating that the results that we obtained are consistent across replicates.

Among the sgRNAs, we observed that ORF7b transcript has the highest average reactivity scores for both WT and Δ382 strains, implying that it has a more open structure among the sgRNAs of SARS-CoV-2 (**Figure 5.4**). Sun et al. have previously shown that structures of leader sequences for each sgRNA were shown to have a weak positive correlation with gene expression[117]. Therefore, we calculated the correlation between PORE-cupine reactivity of the TRS-B sites for each sgRNA and their relative abundance and our results showed a similar trend, for both WT and Δ382 strains (**Figure 5.5**).

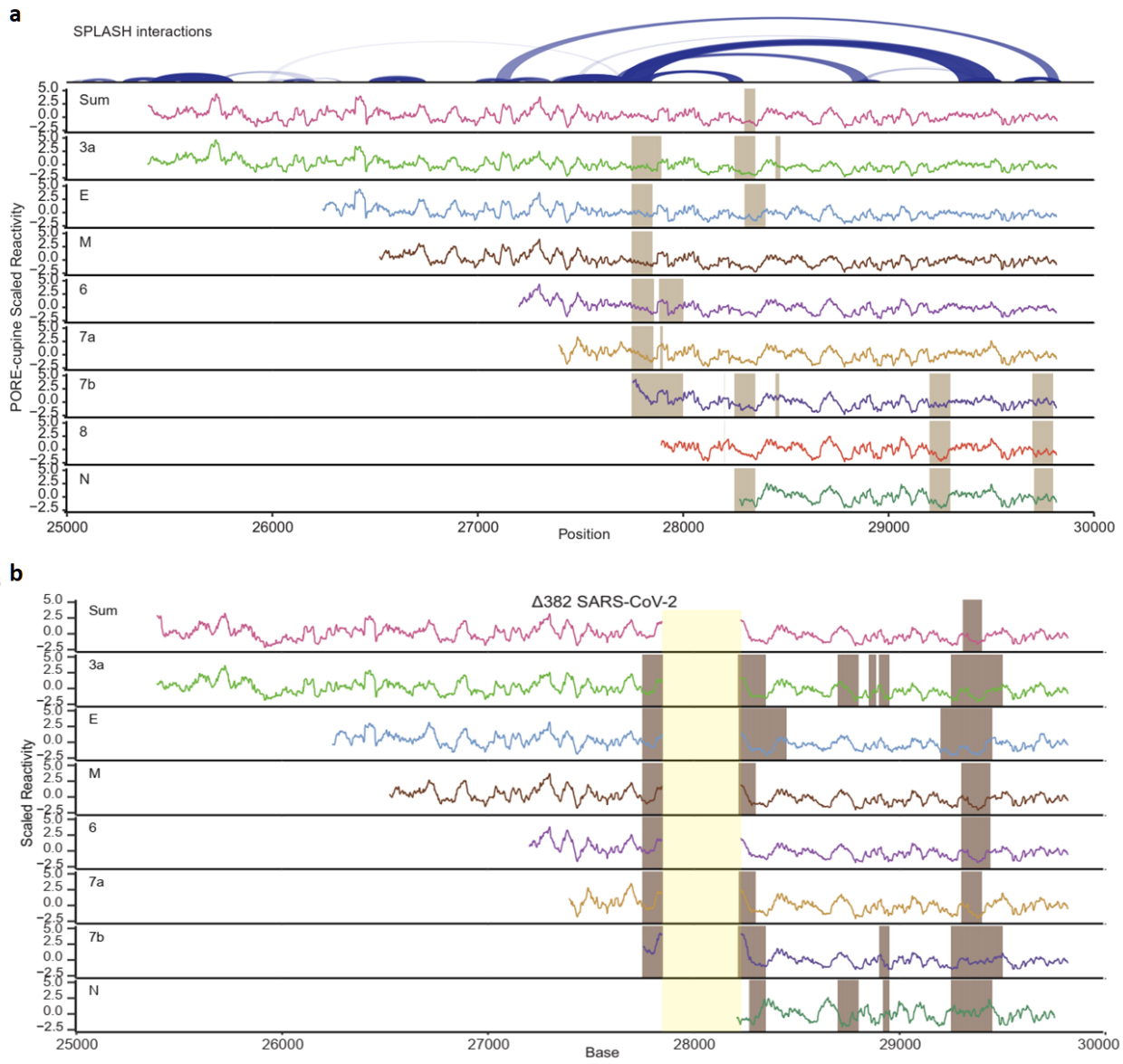


**Figure 5.2. PORE-cupine reveals subgenomic RNA-specific structures in wild type and Δ382 SARS-CoV-2 strains. a,** Results generated from wild type strains. **b,** Results generated from Δ382 strains. **a**,**b,** PORE-cupine reactivity signals from **a**, WT and **b,** Δ382 are averaged across all the signals from the subgenomic RNAs (Sum). PORE-cupine reactivity signals are also shown for 3a (green), E (blue), M (brown), 6 (purple), 7a (light brown), 7b (navy), 8 (red) and N (dark green), ORF8 are not present in Δ382 strain. PORE-cupine reactivity signals for each sgRNA are filtered for full-length sequences that contain leader sequences for each sgRNA. Reactivities for the leader sequence are not shown. Regions with significant differences are highlighted, (**Methods**). Figures were published in Siwy at el[116].

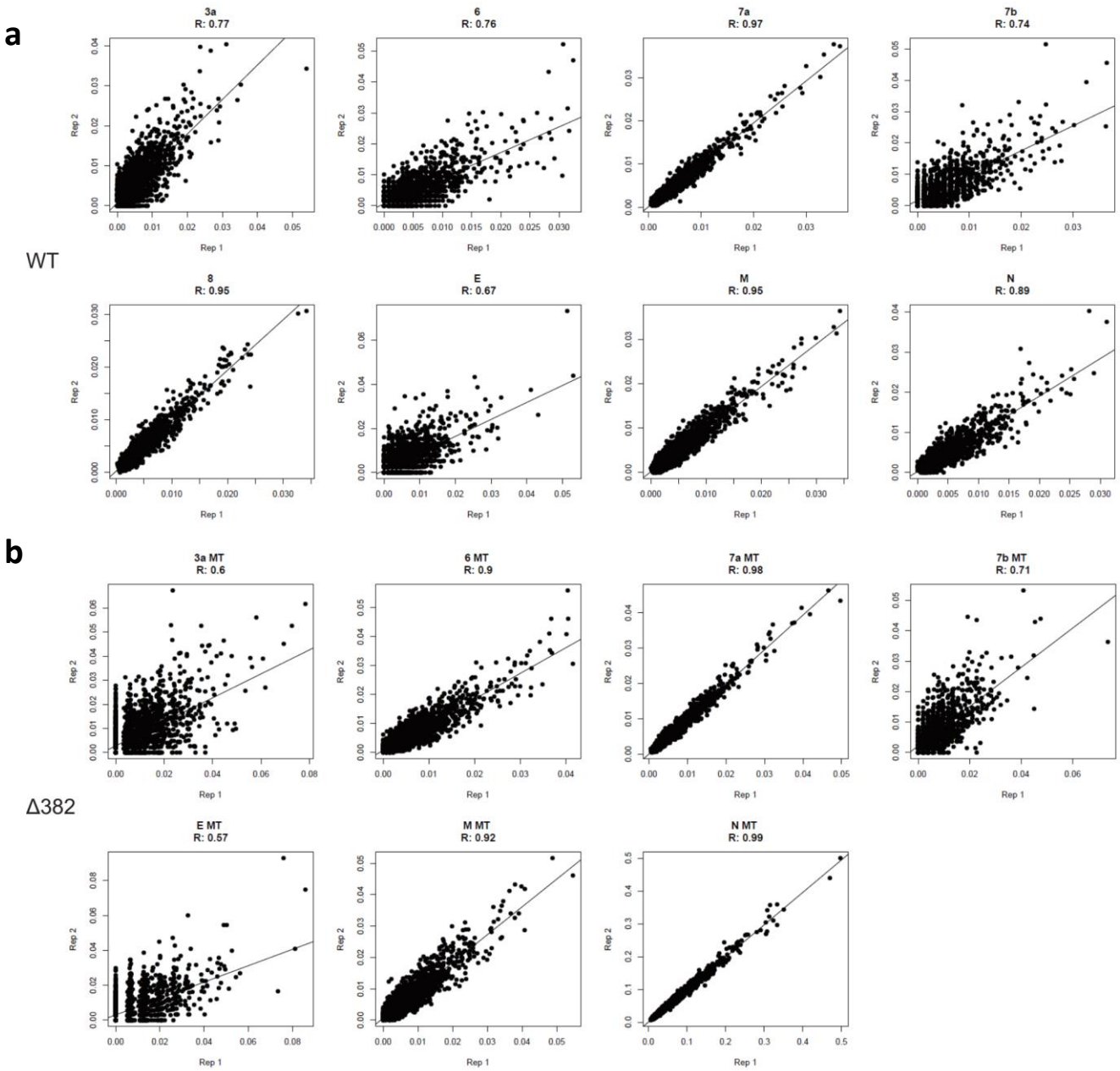


**Figure 5.3. Scatter plots of the PORE-cupine reactivities between 2 biological replicates for each SARS-CoV-2 strain.** The reactivity from two biological replicates of WT (**a**) and Δ382 (**b**) are shown. The R, calculated with Pearson correlation, ranges from 0.67-0.97 (WT) and 0.57-0.98 (Δ382). Figures were published in Siwy at el[116].

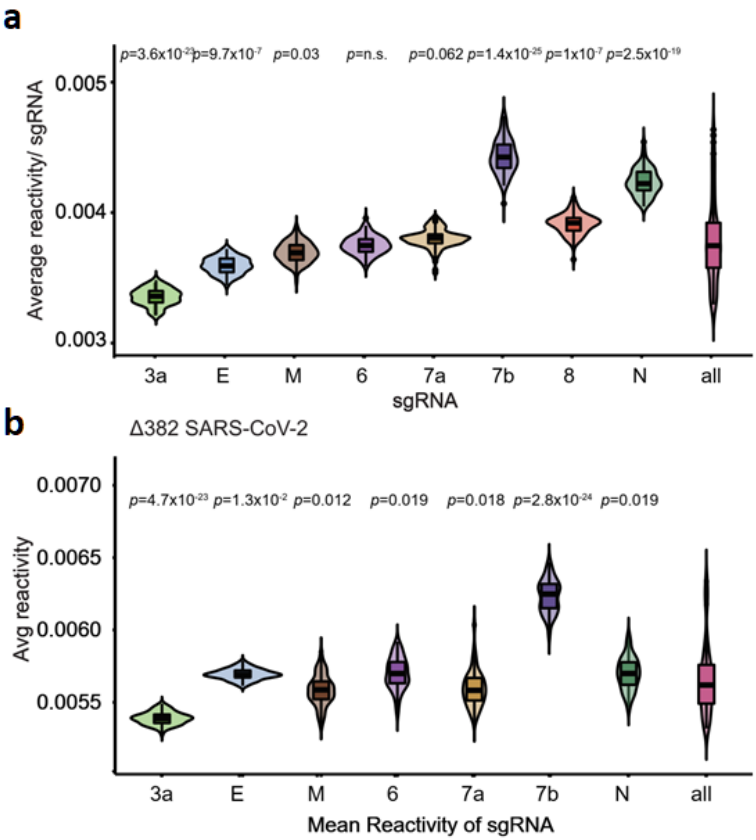


**Figure 5.4. Violin plots of the distribution of average reactivities for each sgRNA. a,** Results generated from wild type strains. **b,** Results generated from Δ382 strains. **a**,**b**, Violin plots from **c**, WT and **d,** Δ382 showing the distribution of average reactivities for each sgRNA. Each sgRNA is subsampled for 500 strands before calculating its mean, n=100. P-values are calculated by comparing the distribution of the reactivities in each sgRNA against all of the sgRNAs. Figures were published in Siwy at el[116].

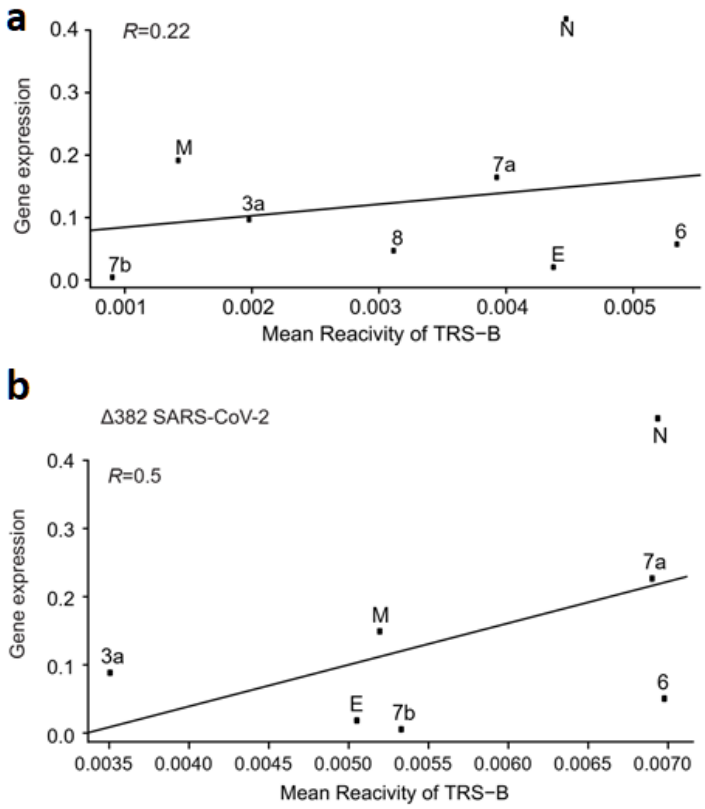


**Figure 5.5. Scatter plots of the correlation between the PORE-cupine reactivity around TRS-B. a,** Results generated from wild type strains. **b,** Results generated from Δ382 strains. **a**,**b**, Scatterplot from **a**, WT and **b,** Δ382 showing the correlation between the PORE-cupine reactivity around TRS-B for each sgRNA (x-axis) against transcript levels inside cells (y-axis). Figures were published in Siwy at el[116].

We compared the reactivity scores of overlapping sequences among the sgRNAs to identify structures unique to each sgRNA (**Figure 5.2a, Methods**). We detected four regions in the sgRNAs of WT SARS-CoV-2 that showed structural differences between different sgRNAs. Three of which are also seen in the sgRNAs of Δ382 SARS-CoV-2 (**Figure 5.2b**). While two regions centred around bases 27800 and 28250 correspond to the leader sequences of sgRNAs of ORF7b and N respectively, two other structurally different regions (centred around 29300 and in 3'UTR) are present within all sgRNAs (**Figure 5.6, 5.7**).

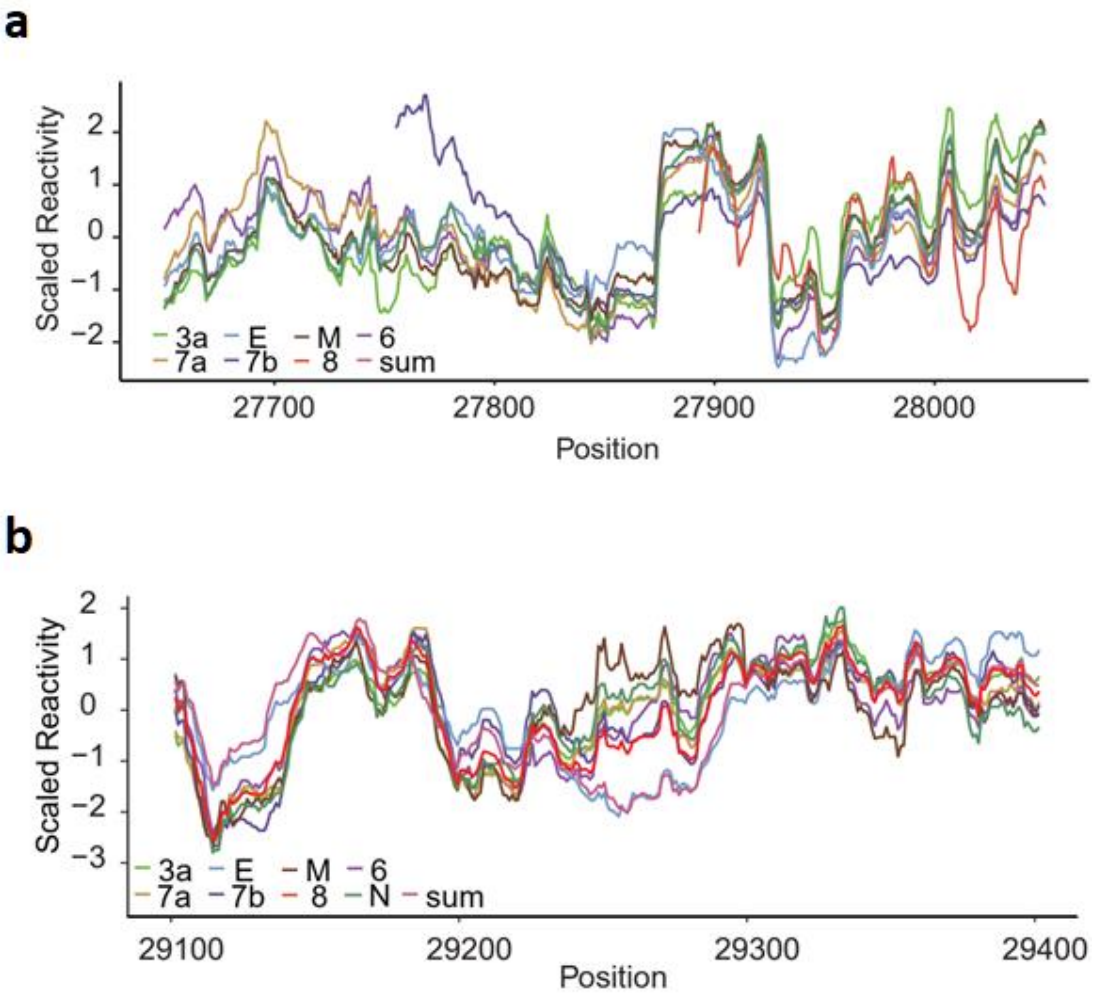


**Figure 5.6. PORE-cupine reveals subgenomic RNA-specific structures in WT strain.** Reactivity plots of regions that show significant structural differences between the sgRNAs from positions (**a**) 27600-28500 and (**b**) 29050-29450. Figures were published in Siwy at el[116].

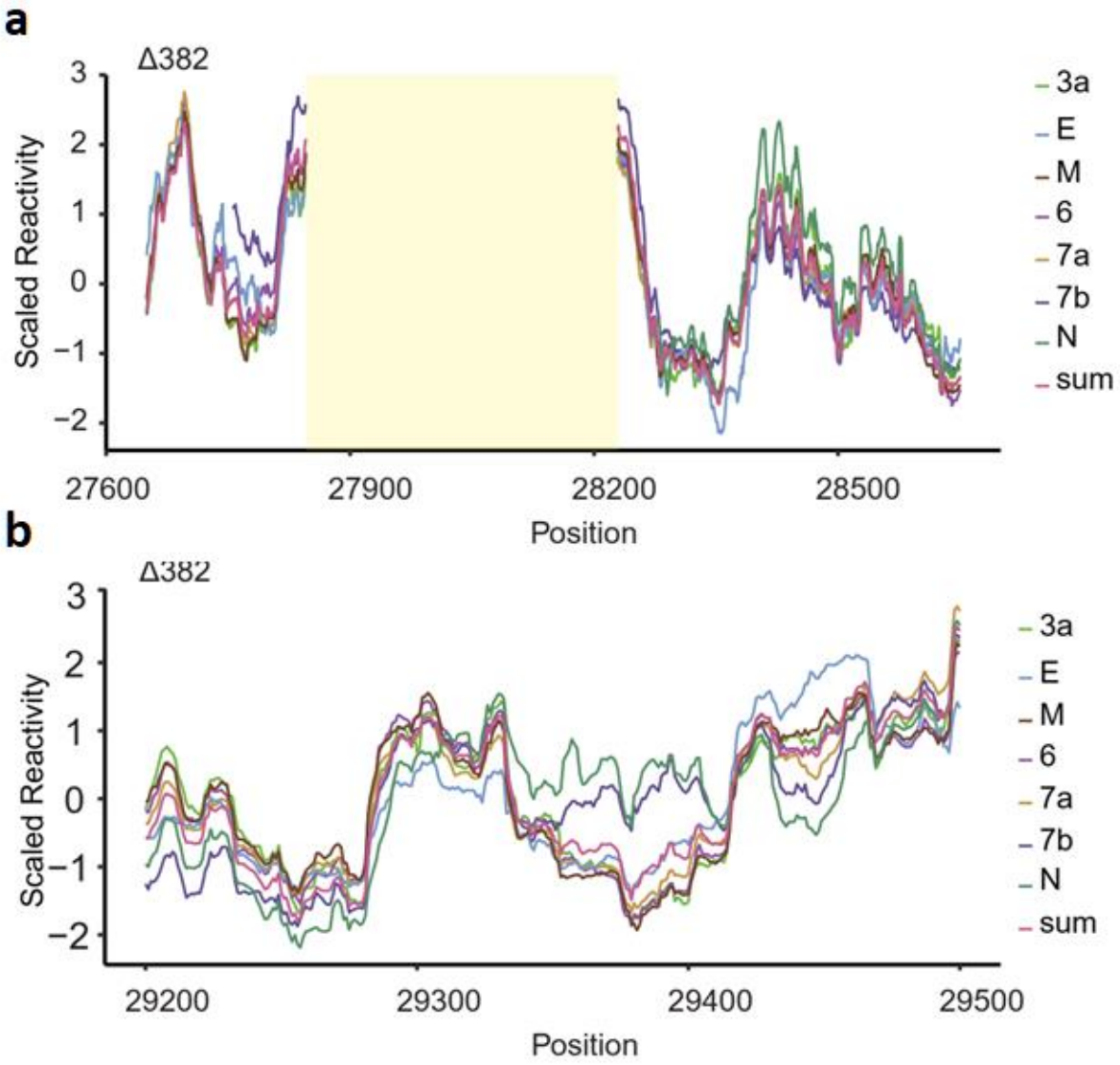


**Figure 5.7. PORE-cupine reveals subgenomic RNA-specific structures in Δ382 strain.** Reactivity plots of regions that show significant structural differences between the sgRNAs from positions (**a**) 27600-28600 and (**b**) 29200-29550, with reference to the WT strain. Figures were published in Siwy at el[116].

We also looked at the results from SPLASH to determine if it supports our results in identifying the structural differences. Using SPLASH, we can infer regions that have multiple interactions to have multiple structures. Our result from SPLASH showed similar regions that have multiple structures which agree with the results of PORE-cupine, indicating that those regions can exist in alternative conformations (**Figure 5.8**). We then visualized the sgRNA-specific structures by incorporating PORE-cupine reactivities into structure modelling and observed different structure models for the

same sequence region in different sgRNAs (**Figure 5.9**), further confirming that different sgRNAs could exist in different structures despite sharing the same sequences.

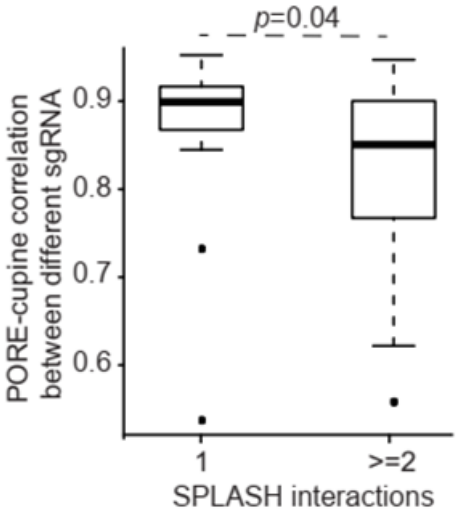


**Figure 5.8. Boxplots showing the distribution of correlation between reactivities of different sgRNAs based on the number of SPLASH interactions.** Correlation of regions that show unique SPLASH interactions (1) and regions that show alternative SPLASH interactions (>=2) between the reactivities of different sgRNAs. Regions that show alternative SPLASH interactions take on different conformations and show lower reactivity correlations between sgRNAs. Figure was published in Siwy at el[116].

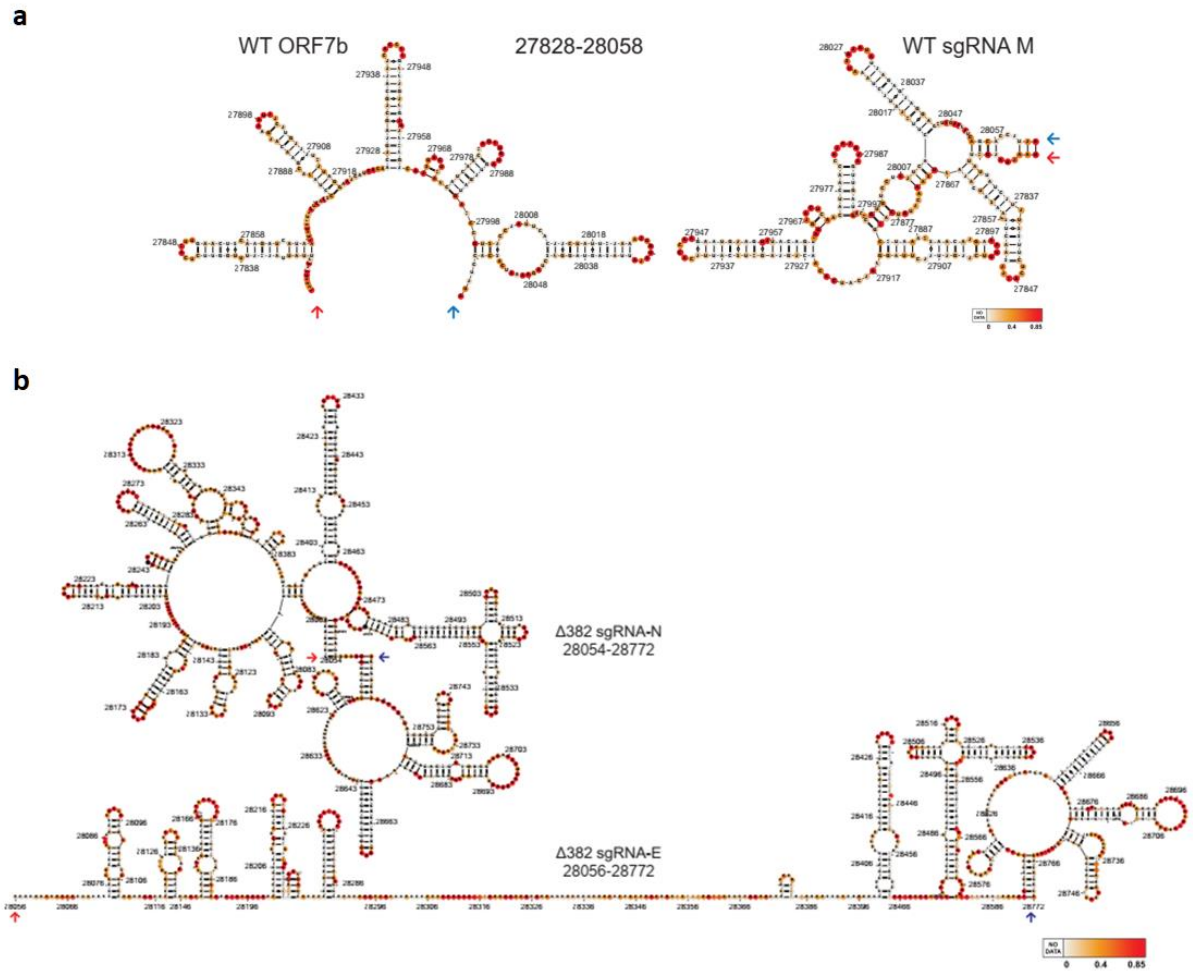


**Figure 5.9. Structure models of the two strains of SARS-CoV-2 sgRNAs. a,** Structure models of WT ORF7b and sgRNA M are generated using the program RNA Structure, using PORE-cupine reactivities as constraints[95,118]. PORE-cupine reactivities are mapped onto the secondary structure models. The red and blue arrows indicate the same positions (start for red and end for blue) in the structure models. **b,** Structure models of Δ382 ORF N and E are generated using the program RNA Structure, using PORE-cupine reactivities as constraints. PORE-cupine reactivities are mapped onto the secondary structure models. The red and blue arrows indicate the same start (red) and stop (blue) positions for the two structure models. Figures were published in Siwy at el[116].

## 5.2 sgRNAs of WT and Δ382 SARS-CoV-2 contain different RNA structures

Viruses that contain genomes with various ORF8 deletions have been found globally[91], however, the mechanism behind the difference in the severity of the symptoms are unknown. To better understand the differences between the two virus phenotypes, we compared the reactivity scores of their sgRNAs.

We compared their PORE-cupine reactivity scores of overlapping positions between the WT and Δ382 strains and we detected structural differences immediately before and after the deletion site when most of the WT and Δ382 sgRNA reads, except for ORF N, as the deletion occurs around position 28000, which is located at the upstream of ORF N. We also observed additional structure differences when individual WT and Δ382 sgRNAs are compared to each other (**Figure 5.10**). The largest structure differences between WT and Δ382 are observed in ORF3a and E sgRNAs (**Figure 5.10, 5.11, 5.12**). We also consistently observed a second structurally different region between WT and Δ382 sgRNAs at the bases 29200-29400 (**Figure 5.10, 5.12, 5.13**), indicating that the deletion could impact distal structures that are located more than 1kb away. As expected, we did not observe reactivity differences between sgRNAs N of WT and Δ382. Interestingly, when we compared the aggregate signals to replicate the similar reactivity profiles obtained by short-read structure probing, we could only detect very local reactivity differences immediately before and after the deletion site, and the differences in the reactivity between the 29300 regions in WT and Δ382 detected previously were absent (**Figure 5.10**). It is most likely due to the high abundance of sgRNA of N among all sgRNA of SARS-CoV-2 masking the differences. As such, using PORE-cupine to probe for RNA structures across sgRNAs can yield novel insights into sgRNA-specific RNA structures.

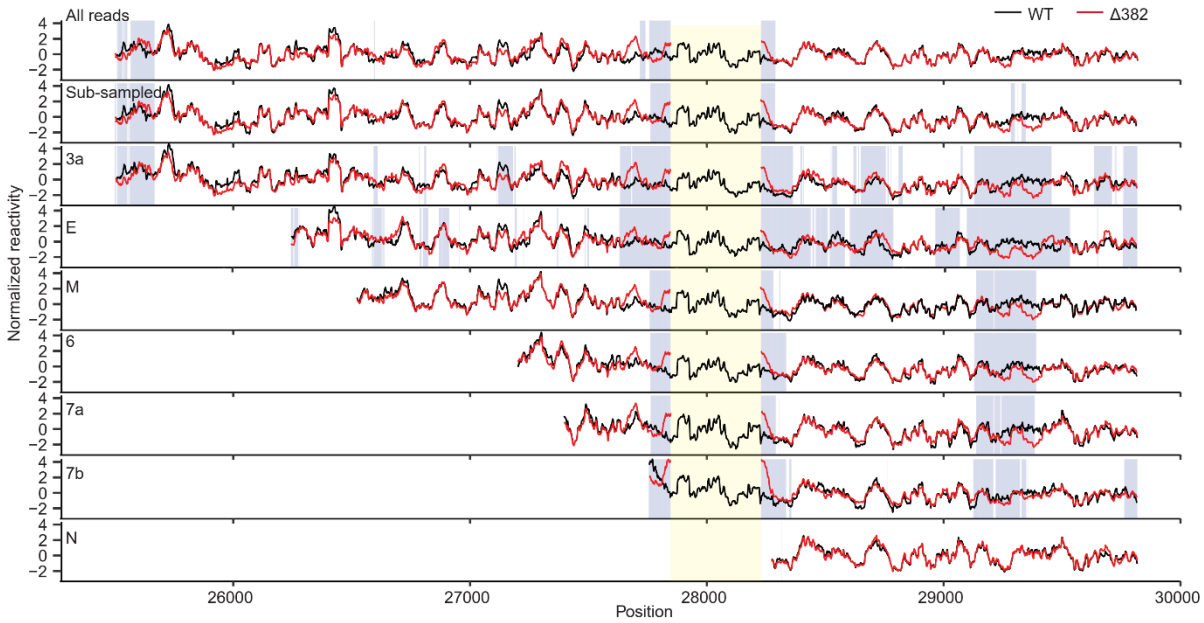


**Figure 5.10. PORE-cupine reactivities differences between WT and Δ382 SARS-CoV-2 genomes.** **a,** PORE-cupine reactivity signals for WT (black) and Δ382 (red) are averaged across all the signals from their respective subgenomic RNAs (All reads). Line plots representing the sub-sampled lane are the averaged WT (black) and Δ382 (red) signals across the different sgRNA after the sgRNAs have been subsampled to the same depth, and hence carry equal weightage to each other. PORE-cupine reactivity signals for WT (black) and Δ382 (red) are also shown for sgRNAs 3a, E, M, 6, 7a, 7b, and N. PORE-cupine reactivity signals for each sgRNA is filtered for full-length sequences that contain leader sequences for each sgRNA. Regions that show significant differences between WT and Δ382 sgRNAs are highlighted in blue, (**Methods**). Figures were published in Siwy at el[116].

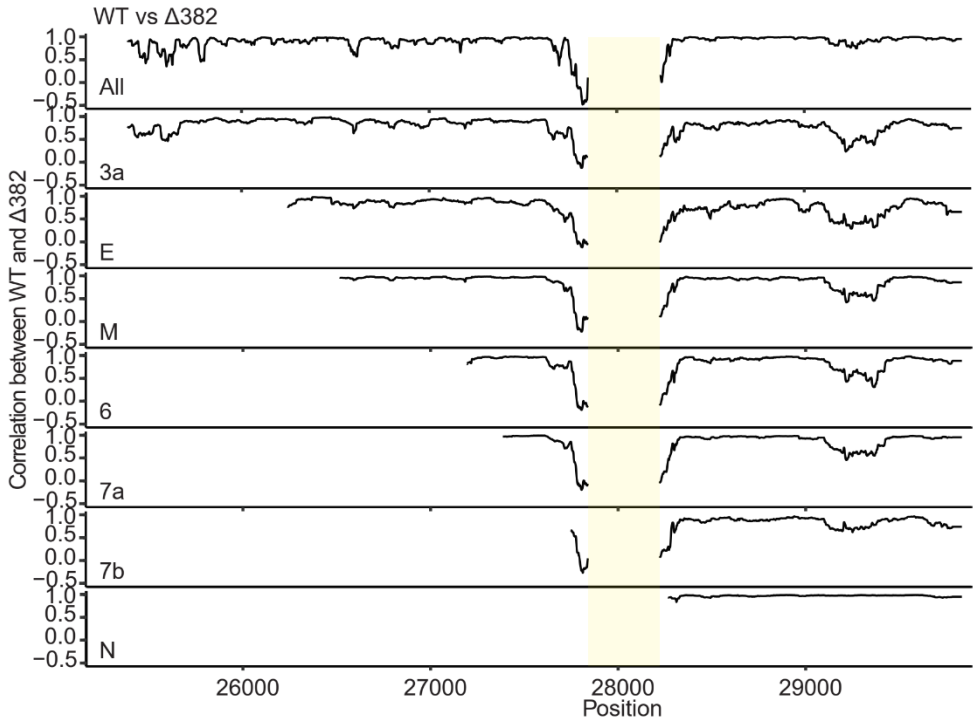


**Figure 5.11. Line plots of Pearson correlation between WT and Δ382 PORE-cupine reactivities for the sum of all sgRNAs for WT and Δ382 and between individual sgRNAs of WT and Δ382 genomes.** The position shown on the x-axis were based on the reference from WT genome. A sliding window of a size of 31 bases was used to calculate the Person correlation between both strains of PORE-cupine reactivity (**Figure 5.9**). Figures were published in Siwy at el[116].

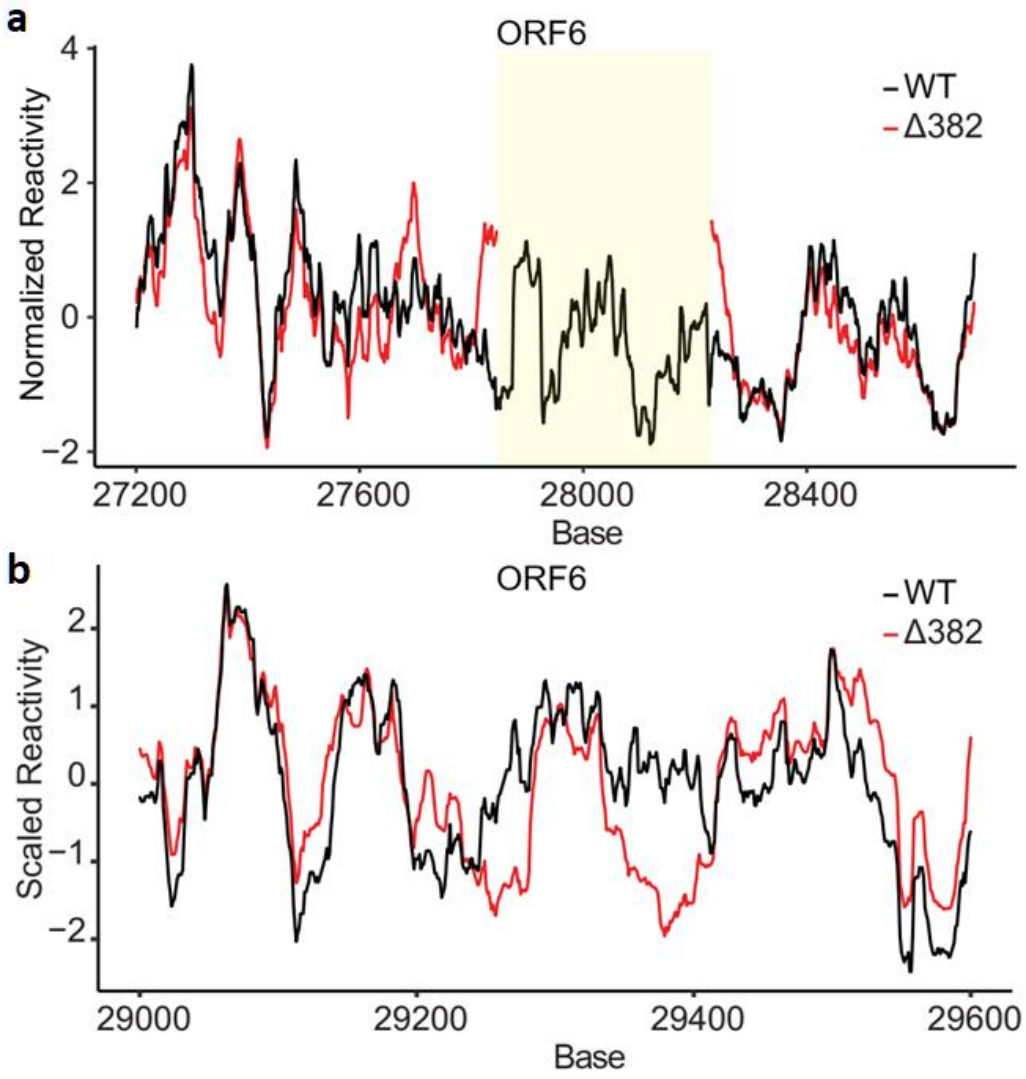


**Figure 5.12. Examples of PORE-cupine reactivity along WT and Δ382 SARS-CoV-2 strains.** Line plots showing the PORE-cupine reactivity of ORF6 along WT and Δ382 around the Δ382 deletion between (**a**) 27600-28600 and (**b**) 29000-29600. Figures were published in Siwy at el[116].

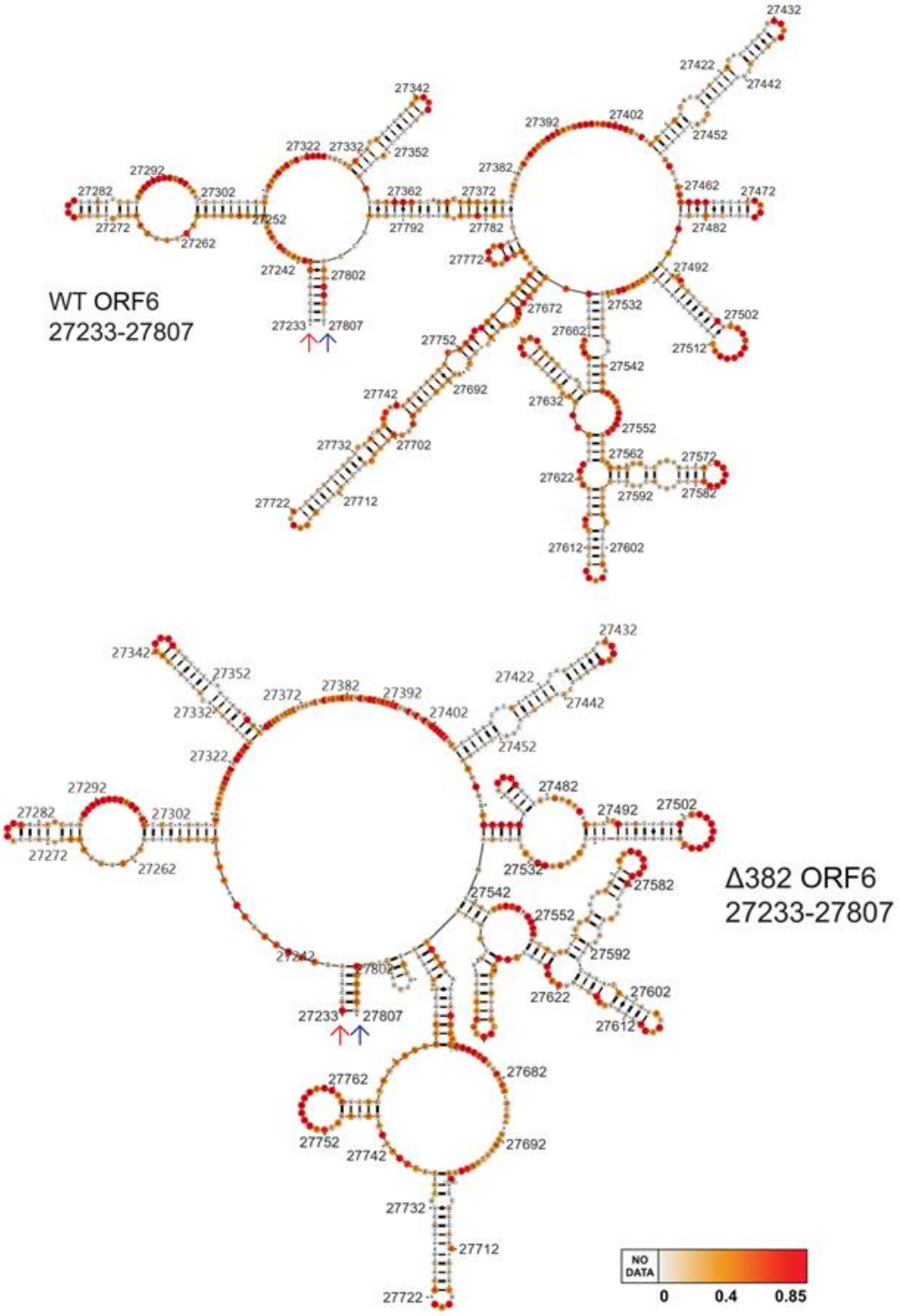


**Figure 5.13. Structure models of WT and Δ382 ORF6.** The models are generated using the program RNA Structure, with PORE-cupine reactivities as constraints. PORE-cupine reactivities are mapped onto the secondary structure models. Figures were published in Siwy at el[116].

## 5.3 Conclusion

Using PORE-cupine, we were able to determine the structural differences among the sgRNAs and between the WT and Δ382 SARS-CoV-2 strains, allowing us to have a better understanding of how sgRNAs differ from each other. This was previously not possible with short-read sequencing methods like ic-SHAPE or SHAPE-MaP due to extensive sequence similarity between the different sgRNAs. It is interesting to note that differences between the sgRNAs only occur after the 5' leader sequences and before its first TRS-L region or deletion of 382 bases, and these additional sequences lead to the differences in structures. We also identified the structural differences between sgRNAs of WT and Δ382 strains. Although more work needs to be done to determine if the differences are caused directly by the sequences or due to other factors, our results could serve as a basis for understanding sgRNA-specific functions in future.

# Chapter 6 Alternative methods to detect modification caused by small molecule structure probing compounds

Although we have developed and published our method PORE-cupine, we are constantly trying to improve the protocol. We are exploring ideas to make the protocol more versatile, increase the number of modifications it can detect, increase the efficiency of our pipeline and improve its accuracy in the detection of modifications. Therefore, we looked into the use of two additional single-stranded structure probing compounds, 1-cyclohexyl-3-(2-morpholinoethyl) carbodiimide metho-p-toluenesulfonate (CMCT) and 1-acetylimidazole (1AI)[45] (**Figure 1.3**), and compared their accuracy to NAI-N3. To increase the efficiency and accuracy, we explore a statistical method, kernel density estimation (KDE)[119] to replace SVM for the detection of modified bases and compared the accuracy with our current method.

## 6.1 Modifications

PORE-cupine can detect NAI-N3 modifications on RNA strands accurately, and it was shown that modifications caused by various structure probing compounds affect the quality of sequencing by direct RNA sequencing[52]. However, the accuracy of detecting those modifications is unknown. Therefore, to determine if PORE-cupine can detect other structure probing compounds accurately, we selected CMCT and 1AI and optimised the parameters, current mean and current standard deviation, used in

PORE-cupine with the same approach used for NAI-N3 modification detection (**Methods**).

The two results from the optimised parameters were compared to the accuracy of NAI-N3 modifications. Our results showed that the 1AI modified samples have similar accuracy and CMCT modified samples have lower accuracy when compared to the NAI-N3 modified samples, but the differences in accuracy are not significant (**Figure 6.2**). The results are within the range of our expectation, as both NAI-N3 and 1AI acylate the 2'-OH of the sugar of single-stranded nucleotides, where CMCT is a base-specific probe that modifies the Gs and Us single-stranded nucleotides.

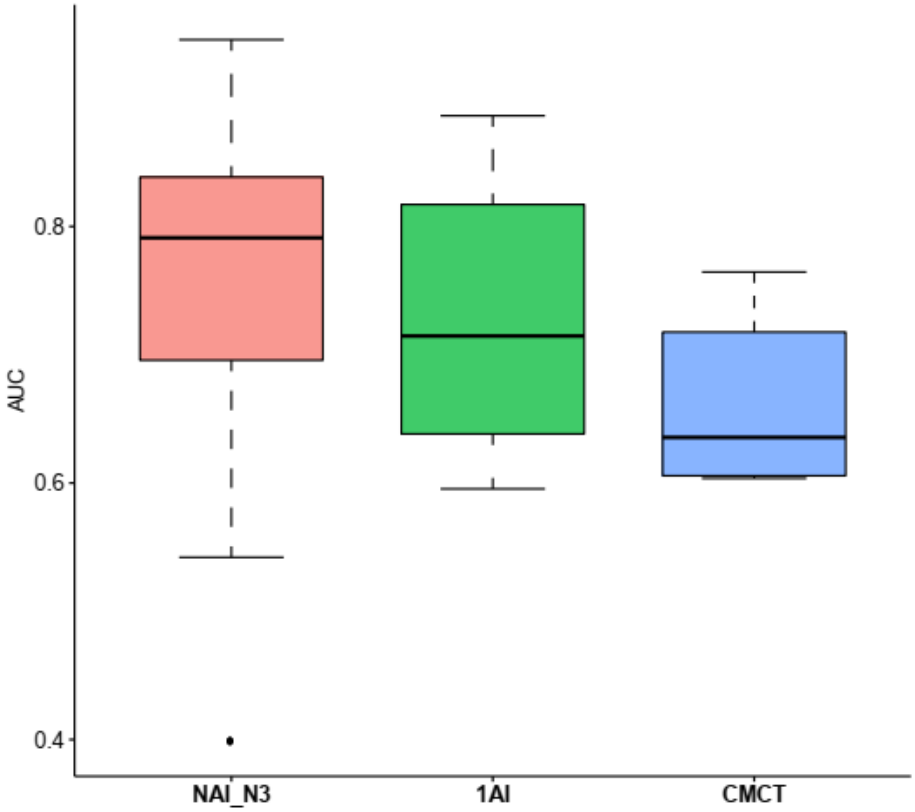


**Figure 6.2. Boxplots of ROC-AUC from different structure probing modifications**

The results suggest that PORE-cupine is a robust method, where it can detect different structure probing modifications accurately when the parameters are optimised for the specific structure probing compounds.

## 6.2 Alternate methods for detecting modifications

PORE-cupine uses a machine-learning algorithm SVM to detect abnormality in the current signal to identify the modified bases in the structure probed samples. During the development phase of PORE-cupine, we looked at a few algorithms to use the current signal for the detection of modifications and selected SVM for PORE-cupine due to the ease of implementation. As we did not explore the other methods extensively, we acknowledge that there might be other methods that have higher efficiency and accuracy in detecting the modifications. Therefore, we explored various methods and chose to evaluate the ability of a statistical method, kernel density estimation (KDE)[118], to detect modifications. We optimised the parameters in KDE by applying a similar approach used by SVM (Methods), for each of the three structure probing compounds that were used in the previous sections (1AI, CMCT and NAI-N3). To evaluate the performance of KDE, we compared the results that were generated with KDE and SVM, to determine if there is any improvement. The results showed that KDE and SVM have similar accuracy and CMCT has the lowest accuracy for both methods among the three chemical probes (**Figure 6.3**).

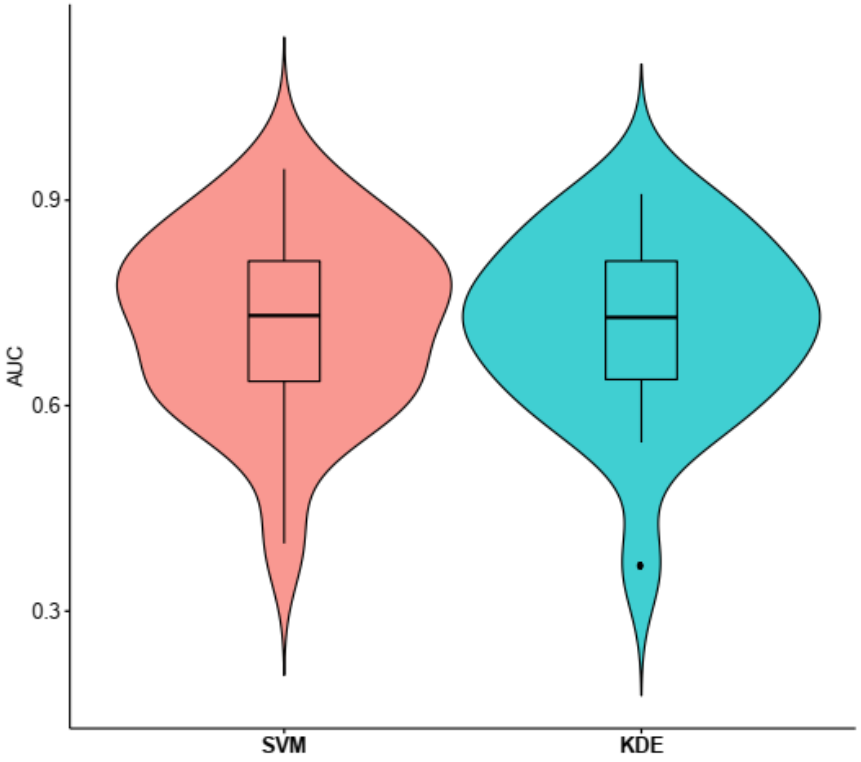


**Figure 6.3. Violin plots of ROC-AUC from the two detection methods.** The results were generated from all three structure probing modifications, 1AI, CMCT and NAI-N3.

## 6.3 Conclusion

In addition, to detect NAI-N3 modifications, PORE-cupine can detect 1AI and CMCT modifications on the RNA, showing that it is a versatile method that can be used to detect various structure probing modifications. Although its accuracy varies among the three compounds that we have tested, we believe that by testing through the different structure probing compounds, we might be able to obtain a compound that pushes PORE-cupine to higher accuracy.

Although the accuracy in detecting structure probing modifications was similar between KDE and SVM, KDE acts as an orthogonal approach to validate our results

from PORE-cupine. Although in most situations, SVM was slightly faster compared to KDE, there is a significant decrease in the time taken for KDE to complete its run for two of the transcripts, RPS12 and RPS29. Overall, the performance of SVM and KDE is very similar in their accuracy, suggesting that further optimization to PORE-cupine might require the use of additional features or making significant changes to the principles of the detection method.

# Chapter 7 Conclusion

## 7.1 Discussion

For my thesis, I have developed a pipeline that uses long-read sequencing to detect structure probing modifications on RNA transcripts. By coupling a long-read sequencing method, specifically direct RNA sequencing from Oxford Nanopore Technologies, with structure probing, I have developed PORE-cupine, a method that can detect NAI-N3 modifications on RNA strands. We have shown that PORE-cupine is an accurate and highly reproducible method and its results are comparable with current short-read high throughput structure probing methods. Thus, PORE-cupine can be used to study the structural dynamics and identify structural changes in the RNAs.

We have shown that PORE-cupine can identify the aptamer region of the TPP riboswitch by comparing the reactivity of the RNA folded with or without its ligand, indicating that it can serve as an orthogonal method to existing structure probing methods. In addition, PORE-cupine provides a rapid and cost-effective method to study a small number of transcripts when compared to existing short-reads structure probing.

We used PORE-cupine to look at the RNA structures of the hESC transcriptome and were able to identify isoform-specific structures. Due to the long-read sequencing capability of direct RNA sequencing, we can accurately distinguish and align the sequenced reads across multiple exons in the isoforms. We observed many isoform pairs that exhibit reactivity differences in shared sequence regions and observed a weak association with the change in structure and change in translation efficiency.

Similarly, applying PORE-cupine to study the sgRNA in SARS-CoV-2 allowed us to identify both sgRNA-specific structures, as well as structural differences between WT and Δ382 strains. We also observed a weak association between the structural difference and a change in translation efficiency for the isoform's pairs in the hESCs, suggesting that only a subset of changes affects the translation efficiency. Importantly, these structural results are supported using an orthogonal method, SPLASH, further demonstrating the power of Pore-cupine to determine RNA structures in shared regions.

One of the limitations of the study is that we utilized published human transcriptome as our reference for mapping of our sequencing reads. As such, sequenced reads could be misassigned leading to a less accurate reactivity. However, it is challenging to assemble our own transcriptome accurately due to the need for high sequencing depth [103]. Another challenge that we encountered during the projects is that it is difficult to identify the cause and effect of RNA structural changes and their biological consequences. While we can determine the link between structural and biological function differences, by coupling the results from PORE-cupine and functional genomics, for example between structural differences and hESC translation efficiency using TrIP-seq, further experiments are needed to determine whether the changes in structures were due to changes in translation or whether changes in translation was due to changes in structure.

Additionally, as direct RNA sequencing only performs well for longer transcripts, PORE-cupine is more suited to sequence reads that are longer than 400 nucleotides, making it difficult to analyse the structures of shorter transcripts. The sequencing yield of direct RNA sequencing is also significantly lower when compared to Illumina sequencing. This limits the number of transcripts that we can study, as we require a

minimum depth of 200 reads for each transcript. We also noticed that there may be a “blind spot” at the 5’ end of the transcripts, as Nanopore sequencing starts from the 3’end and there is a decay in signal particularly at the first 10% of bases of the transcript. This signal decay could be due to RNA degradation during the processing of the samples or during sequencing and makes identification of the true 5’end of the transcripts challenging. Existing strategies to identify the “real” 5’end utilizes a cap capture method that converts 5’capped intact RNA to a form that is ligation compatible, however, we found that 5’adapter ligation efficiency is poor in our hands.

To further expand the utilize of PORE-cupine, we are in the process of testing and expanding the number of structure probing compounds that can be used for structure probing with Nanopore sequencing. We also aim to improve PORE-cupine’s detection accuracy and efficiency. We have established that PORE-cupine is capable of capturing structure-based modification by two additional chemical probes, 1AI and CMCT, with 1AI having a similar accuracy as NAI-N3. Although we did not observe significant improvement in using KDE over SVM, the result from KDE suggests that other analytical methods, including statistical methods, could be used to analyse PORE-cupine data, expanding the repertoire of tools available to analyse our data.

## 7.2 Future directions

As PORE-cupine is a nascent technology, there are many different areas that we could continue developing the technology in, as well as apply it to understand different aspects of biology. First, we can potentially increase the sequencing depth of direct RNA sequencing by around fivefold if we were to replace the MinION flowcell (~1-2M reads per flow cell) with a PromethION (7M reads per flow cell), enabling us to

interrogate more transcripts in the transcriptome. While this is useful, we will need to characterise the current signals of PromethION to see if it can be adapted to PORE-cupine, as a different model of sequencing flowcell and hence different current characteristics might be present in the PromethION versus the MinION. Second, to identify the real 5' start sides of transcripts, we need to solve the issue of 5' end decay in the sequencing run. We first need to determine if the read was truncated during sequencing or it had been produced by the cell. This can be determined by ligating an adapter to the 5' end of our samples, so that we can filter for the full-length reads during the analysis. This is a similar procedure to how the full-length reads for the sgRNAs in SARS-CoV-2 were filtered. While our initial attempts at adapter ligation had low efficiency, we aim to test other cap capture methods to enrich for full-length transcripts. Third, we aim to combine PORE-cupine with other structure information, such as which bases are paired together in close or distal linear space[92], and the location of solvent-exposed bases along an RNA[120], to potentially generate an accurate three-dimensional model of the transcript. Fourth, we would also like to utilize the single-molecule sequencing feature of Nanopore direct RNA sequencing to test the ability of PORE-cupine to detect the presence of multiple conformations from the same RNA sequence. By clustering the pattern of modifications (generated by compounds such as NAI) along individual RNA strands, we aim to identify the major clusters that the RNAs can reside in. As structures in the RNAs are not static, multiple stable conformations can exist at a given time. We hypothesized that not all RNA structures are important for the regulation, therefore if we can separate the different structures in a transcript to their different major subpopulations, this will allow us to look at functionally relevant structures.

On the application/biology front, we are interested in extending beyond our association of RNA structure and translation to study the role of isoform-specific structures in isoform-specific decay. We are also interested in applying PORE-cupine to look at the structures of each isoform at the different organelles. By fractionating the cells, we would be able to purify and sequence the different organelles to obtain the reactivity, allowing us to determine if there are any differences between the same isoforms across the cellular compartments. Collectively, we believe that future advances and applications of PORE-cupine could deepen our understanding of RNA structure and function.

# References


1. Nowakowski, J. and I. Tinoco, *RNA Structure and Stability.* Seminars in Virology, 1997. **8**(3): p. 153-165.
2. Wilkinson, K.A., E.J. Merino, and K.M. Weeks, *RNA SHAPE chemistry reveals nonhierarchical interactions dominate equilibrium structural transitions in tRNA(Asp) transcripts.* J Am Chem Soc, 2005. **127**(13): p. 4659-67.
3. Rajkowitsch, L., et al., *RNA chaperones, RNA annealers and RNA helicases.* RNA Biol, 2007. **4**(3): p. 118-30.
4. Ganser, L.R., et al., *The roles of structural dynamics in the cellular functions of RNAs.* Nature Reviews Molecular Cell Biology, 2019. **20**(8): p. 474-489.
5. Halvorsen, M., et al., *Disease-associated mutations that alter the RNA structural ensemble.* PLoS Genet, 2010. **6**(8): p. e1001074.
6. Strobel, E.J., et al., *RNA systems biology: uniting functional discoveries and structural tools to understand global roles of RNAs.* Curr Opin Biotechnol, 2016. **39**: p. 182-191.
7. Larsen, K.P., et al., *Relating Structure and Dynamics in RNA Biology.* Cold Spring Harb Perspect Biol, 2019. **11**(7).
8. Meyer, M.M., *The role of mRNA structure in bacterial translational regulation.* Wiley Interdiscip Rev RNA, 2017. **8**(1).
9. Kutchko, K.M., et al., *Multiple conformations are a conserved and regulatory feature of the RB1 5' UTR.* RNA, 2015. **21**(7): p. 1274-85.
10. Liu, F. and C.X. Gong, *Tau exon 10 alternative splicing and tauopathies.* Mol Neurodegener, 2008. **3**: p. 8.
11. Strobel, E.J., A.M. Yu, and J.B. Lucks, *High-throughput determination of RNA structures.* Nature Reviews Genetics, 2018. **19**(10): p. 615-634.
12. Kenyon, J., L. Prestwood, and A. Lever, *Current perspectives on RNA secondary structure probing.* Biochem Soc Trans, 2014. **42**(4): p. 1251-5.
13. Bernhart, S.H., et al., *RNAalifold: improved consensus structure prediction for RNA alignments.* BMC Bioinformatics, 2008. **9**: p. 474.
14. Engelen, S. and F. Tahi, *Tfold: efficient in silico prediction of non-coding RNA secondary structures.* Nucleic Acids Res, 2010. **38**(7): p. 2453-66.
15. Rivas, E., *RNA structure prediction using positive and negative evolutionary information.* PLoS computational biology, 2020. **16**(10): p. e1008387-e1008387.
16. Knudsen, B. and J. Hein, *Pfold: RNA secondary structure prediction using stochastic context-free grammars.* Nucleic Acids Res, 2003. **31**(13): p. 3423-8.
17. Zuker, M., *Mfold web server for nucleic acid folding and hybridization prediction.* Nucleic Acids Res, 2003. **31**(13): p. 3406-15.
18. Denman, R.B., *Using RNAFOLD to predict the activity of small catalytic RNAs.* Biotechniques, 1993. **15**(6): p. 1090-5.
19. Do, C.B., D.A. Woods, and S. Batzoglou, *CONTRAfold: RNA secondary structure prediction without physics-based models.* Bioinformatics, 2006. **22**(14): p. e90-8.
20. Akiyama, M., K. Sato, and Y. Sakakibara, *A max-margin training of RNA secondary structure prediction integrated with the thermodynamic model.* J Bioinform Comput Biol, 2018. **16**(6): p. 1840025.
21. Cate, J.H., et al., *Crystal structure of a group I ribozyme domain: principles of RNA packing.* Science, 1996. **273**(5282): p. 1678-85.
22. Cate, J.H. and J.A. Doudna, *Solving large RNA structures by X-ray crystallography.* Methods Enzymol, 2000. **317**: p. 169-80.

23. Mittermaier, A. and L.E. Kay, *New tools provide new insights in NMR studies of protein dynamics.* Science, 2006. **312**(5771): p. 224-8.
24. Sun, A., et al., *Strategies for understanding RNA recognition by X-ray crystallography and NMR methods.* Methods Enzymol, 2019. **623**: p. 229-248.
25. Fica, S.M. and K. Nagai, *Cryo-electron microscopy snapshots of the spliceosome: structural insights into a dynamic ribonucleoprotein machine.* Nature Structural & Molecular Biology, 2017. **24**(10): p. 791-799.
26. Lyumkis, D., *Challenges and opportunities in cryo-EM single-particle analysis.* J Biol Chem, 2019. **294**(13): p. 5181-5197.
27. Chen, Y. and L. Pollack, *SAXS studies of RNA: structures, dynamics, and interactions with partners.* Wiley interdisciplinary reviews. RNA, 2016. **7**(4): p. 512-526.
28. Cantara, W.A., E.D. Olson, and K. Musier-Forsyth, *Analysis of RNA structure using small-angle X-ray scattering.* Methods (San Diego, Calif.), 2017. **113**: p. 46-55.
29. Wan, Y., et al., *Understanding the transcriptome through RNA structure.* Nature Reviews Genetics, 2011. **12**(9): p. 641-655.
30. Mailler, E., et al., *The evolution of RNA structural probing methods: From gels to next-generation sequencing.* Wiley Interdiscip Rev RNA, 2019. **10**(2): p. e1518.
31. Das, R., et al., *SAFA: semi-automated footprinting analysis software for high-throughput quantification of nucleic acid footprinting experiments.* RNA, 2005. **11**(3): p. 344-54.
32. Peattie, D.A. and W. Gilbert, *Chemical probes for higher-order structure in RNA.* Proc Natl Acad Sci U S A, 1980. **77**(8): p. 4679-82.
33. Ehresmann, C., et al., *Probing the structure of RNAs in solution.* Nucleic Acids Res, 1987. **15**(22): p. 9109-28.
34. Tullius, T.D. and J.A. Greenbaum, *Mapping nucleic acid structure by hydroxyl radical cleavage.* Curr Opin Chem Biol, 2005. **9**(2): p. 127-34.
35. Strobel, E.J., A.M. Yu, and J.B. Lucks, *High-throughput determination of RNA structures.* Nat Rev Genet, 2018. **19**(10): p. 615-634.
36. Rouskin, S., et al., *Genome-wide probing of RNA structure reveals active unfolding of mRNA structures in vivo.* Nature, 2014. **505**(7485): p. 701-5.
37. Spitale, R.C., et al., *Structural imprints in vivo decode RNA regulatory mechanisms.* Nature, 2015. **519**(7544): p. 486-90.
38. Siegfried, N.A., et al., *RNA motif discovery by SHAPE and mutational profiling (SHAPE-MaP).* Nat Methods, 2014. **11**(9): p. 959-65.
39. Ziehler, W.A. and D.R. Engelke, *Probing RNA structure with chemical reagents and enzymes.* Curr Protoc Nucleic Acid Chem, 2001. **Chapter 6**: p. Unit 6 1.
40. Sachsenmaier, N., et al., *Mapping RNA structure in vitro using nucleobase-specific probes.* Methods Mol Biol, 2014. **1086**: p. 79-94.
41. Flynn, R.A., et al., *Transcriptome-wide interrogation of RNA secondary structure in living cells with icSHAPE.* Nat Protoc, 2016. **11**(2): p. 273-90.
42. Mitchell, D., 3rd, S.M. Assmann, and P.C. Bevilacqua, *Probing RNA structure in vivo.* Curr Opin Struct Biol, 2019. **59**: p. 151-158.
43. Velema, W.A. and E.T. Kool, *The chemistry and applications of RNA 2'-OH acylation.* Nat Rev Chem, 2020. **4**(1): p. 22-37.
44. Tapsin, S., et al., *Genome-wide identification of natural RNA aptamers in prokaryotes and eukaryotes.* Nat Commun, 2018. **9**(1): p. 1289.
45. Stephenson, W., et al., *Direct detection of RNA modifications and structure using single molecule nanopore sequencing.* bioRxiv, 2020: p. 2020.05.31.126763.
46. Incarnato, D., et al., *Genome-wide profiling of mouse RNA secondary structures reveals key features of the mammalian transcriptome.* Genome Biol, 2014. **15**(10): p. 491.
47. Zubradt, M., et al., *DMS-MaPseq for genome-wide or targeted RNA structure probing in vivo.* Nat Methods, 2017. **14**(1): p. 75-82.

48. Luo, Q.J., et al., *RNA structure probing reveals the structural basis of Dicer binding and cleavage.* Nat Commun, 2021. **12**(1): p. 3397.
49. Lorenz, R., et al., *SHAPE directed RNA folding.* Bioinformatics, 2016. **32**(1): p. 145-7.
50. Reuter, J.S. and D.H. Mathews, *RNAstructure: software for RNA secondary structure prediction and analysis.* BMC Bioinformatics, 2010. **11**: p. 129.
51. Busan, S. and K.M. Weeks, *Accurate detection of chemical modifications in RNA by mutational profiling (MaP) with ShapeMapper 2.* RNA, 2018. **24**(2): p. 143-148.
52. Aw, J.G.A., et al., *Determination of isoform-specific RNA structure with nanopore long reads.* Nat Biotechnol, 2021. **39**(3): p. 336-346.
53. Wang, Z., M. Gerstein, and M. Snyder, *RNA-Seq: a revolutionary tool for transcriptomics.* Nat Rev Genet, 2009. **10**(1): p. 57-63.
54. Modrek, B. and C. Lee, *A genomic view of alternative splicing.* Nat Genet, 2002. **30**(1): p. 13-9.
55. Kim, H.K., et al., *Alternative splicing isoforms in health and disease.* Pflugers Arch, 2018. **470**(7): p. 995-1016.
56. Salzberg, S.L., et al., *GAGE: A critical evaluation of genome assemblies and assembly algorithms.* Genome Res, 2012. **22**(3): p. 557-67.
57. Zhang, C., et al., *Evaluation and comparison of computational tools for RNA-seq isoform quantification.* BMC Genomics, 2017. **18**(1): p. 583.
58. Eid, J., et al., *Real-time DNA sequencing from single polymerase molecules.* Science, 2009. **323**(5910): p. 133-8.
59. Smith, A.M., et al., *Reading canonical and modified nucleobases in 16S ribosomal RNA using nanopore native RNA sequencing.* PLoS One, 2019. **14**(5): p. e0216709.
60. Rhoads, A. and K.F. Au, *PacBio Sequencing and Its Applications.* Genomics Proteomics Bioinformatics, 2015. **13**(5): p. 278-89.
61. Payne, A., et al., *BulkVis: a graphical viewer for Oxford nanopore bulk FAST5 files.* Bioinformatics, 2019. **35**(13): p. 2193-2198.
62. Deamer, D., M. Akeson, and D. Branton, *Three decades of nanopore sequencing.* Nat Biotechnol, 2016. **34**(5): p. 518-24.
63. Zarrinkar, P.P. and J.R. Williamson, *The kinetic folding pathway of the Tetrahymena ribozyme reveals possible similarities between RNA and protein folding.* Nat Struct Biol, 1996. **3**(5): p. 432-8.
64. Russell, R. and D. Herschlag, *Probing the folding landscape of the Tetrahymena ribozyme: commitment to form the native conformation is late in the folding pathway.* J Mol Biol, 2001. **308**(5): p. 839-51.
65. Uchida, T., et al., *Multiple monovalent ion-dependent pathways for the folding of the L-21 Tetrahymena thermophila ribozyme.* J Mol Biol, 2003. **328**(2): p. 463-78.
66. Mitchell, D., 3rd and R. Russell, *Folding pathways of the Tetrahymena ribozyme.* J Mol Biol, 2014. **426**(12): p. 2300-12.
67. Guo, F., A.R. Gooding, and T.R. Cech, *Structure of the Tetrahymena ribozyme: base triple sandwich and metal ion at the active site.* Mol Cell, 2004. **16**(3): p. 351-62.
68. Doudna, J.A. and T.R. Cech, *The chemical repertoire of natural ribozymes.* Nature, 2002. **418**(6894): p. 222-8.
69. Guerrier-Takada, C., et al., *The RNA moiety of ribonuclease P is the catalytic subunit of the enzyme.* Cell, 1983. **35**(3 Pt 2): p. 849-57.
70. Pley, H.W., K.M. Flaherty, and D.B. McKay, *Three-dimensional structure of a hammerhead ribozyme.* Nature, 1994. **372**(6501): p. 68-74.
71. Newcomb, L.F. and H.F. Noller, *Directed hydroxyl radical probing of 16S rRNA in the ribosome: spatial proximity of RNA elements of the 3' and 5' domains.* RNA, 1999. **5**(7): p. 849-55.

72. Klaholz, B.P., et al., *Structure of the Escherichia coli ribosomal termination complex with release factor 2.* Nature, 2003. **421**(6918): p. 90-4.
73. Zhang, W., J.A. Dunkle, and J.H. Cate, *Structures of the ribosome in intermediate states of ratcheting.* Science, 2009. **325**(5943): p. 1014-7.
74. Winkler, W.C., S. Cohen-Chalamish, and R.R. Breaker, *An mRNA structure that controls gene expression by binding FMN.* Proc Natl Acad Sci U S A, 2002. **99**(25): p. 15908-13.
75. Mandal, M., et al., *Riboswitches control fundamental biochemical pathways in Bacillus subtilis and other bacteria.* Cell, 2003. **113**(5): p. 577-86.
76. Nahvi, A., J.E. Barrick, and R.R. Breaker, *Coenzyme B12 riboswitches are widespread genetic control elements in prokaryotes.* Nucleic Acids Res, 2004. **32**(1): p. 143-50.
77. Dann, C.E., 3rd, et al., *Structure and mechanism of a metal-sensing regulatory RNA.* Cell, 2007. **130**(5): p. 878-92.
78. Watson, P.Y. and M.J. Fedor, *The ydaO motif is an ATP-sensing riboswitch in Bacillus subtilis.* Nat Chem Biol, 2012. **8**(12): p. 963-5.
79. Wilson-Mitchell, S.N., F.J. Grundy, and T.M. Henkin, *Analysis of lysine recognition and specificity of the Bacillus subtilis L box riboswitch.* Nucleic Acids Res, 2012. **40**(12): p. 5706-17.
80. Sklyarova, S.A. and A.S. Mironov, *[Bacillus subtilis ypaA gene regulation mechanism involves FMN-binding sensor RNA].* Genetika, 2014. **50**(3): p. 364-8.
81. Winkler, W., A. Nahvi, and R.R. Breaker, *Thiamine derivatives bind messenger RNAs directly to regulate bacterial gene expression.* Nature, 2002. **419**(6910): p. 952-6.
82. Lang, K., R. Rieder, and R. Micura, *Ligand-induced folding of the thiM TPP riboswitch investigated by a structure-based fluorescence spectroscopic approach.* Nucleic Acids Res, 2007. **35**(16): p. 5370-8.
83. Cannone, J.J., et al., *The comparative RNA web (CRW) site: an online database of comparative sequence and structure information for ribosomal, intron, and other RNAs.* BMC Bioinformatics, 2002. **3**: p. 2.
84. Floor, S.N. and J.A. Doudna, *Tunable protein synthesis by transcript isoforms in human cells.* Elife, 2016. **5**.
85. Patro, R., et al., *Salmon provides fast and bias-aware quantification of transcript expression.* Nat Methods, 2017. **14**(4): p. 417-419.
86. Simpson, J.T., et al., *Detecting DNA cytosine methylation using nanopore sequencing.* Nat Methods, 2017. **14**(4): p. 407-410.
87. Li, H. and R. Durbin, *Fast and accurate long-read alignment with Burrows-Wheeler transform.* Bioinformatics, 2010. **26**(5): p. 589-95.
88. Shah, A., et al., *CLIP Tool Kit (CTK): a flexible and robust pipeline to analyze CLIP sequencing data.* Bioinformatics, 2017. **33**(4): p. 566-567.
89. Heinz, S., et al., *Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities.* Mol Cell, 2010. **38**(4): p. 576-89.
90. Li, H., *A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data.* Bioinformatics, 2011. **27**(21): p. 2987-93.
91. Young, B.E., et al., *Effects of a major deletion in the SARS-CoV-2 genome on the severity of infection and the inflammatory response: an observational cohort study.* Lancet, 2020. **396**(10251): p. 603-611.
92. Aw, J.G., et al., *In Vivo Mapping of Eukaryotic RNA Interactomes Reveals Principles of Higher-Order Organization and Regulation.* Mol Cell, 2016. **62**(4): p. 603-17.
93. Aw, J.G.A., et al., *Mapping RNA-RNA Interactions Globally Using Biotinylated Psoralen.* J Vis Exp, 2017(123).

94. Kim, D., et al., *The Architecture of SARS-CoV-2 Transcriptome.* Cell, 2020. **181**(4): p. 914-921 e10.
95. Busan, S. and K.M. Weeks, *Visualization of RNA structure models within the Integrative Genomics Viewer.* RNA, 2017. **23**(7): p. 1012-1018.
96. Workman, R.E., et al., *Nanopore native RNA sequencing of a human poly(A) transcriptome.* Nat Methods, 2019. **16**(12): p. 1297-1305.
97. Weeks, K.M., *Advances in RNA structure analysis by chemical probing.* Curr Opin Struct Biol, 2010. **20**(3): p. 295-304.
98. Spitale, R.C., et al., *RNA SHAPE analysis in living cells.* Nat Chem Biol, 2013. **9**(1): p. 18-20.
99. Chang, C.-C. and C.-J. Lin, *LIBSVM: A Library for Support Vector Machines.* ACM Transactions on Intelligent Systems and Technology, 2011. **2**: p. 27:1--27:27.
100. Kertesz, M., et al., *Genome-wide measurement of RNA secondary structure in yeast.* Nature, 2010. **467**(7311): p. 103-7.
101. Ding, Y., et al., *In vivo genome-wide profiling of RNA secondary structure reveals novel regulatory features.* Nature, 2014. **505**(7485): p. 696-700.
102. Pan, Q., et al., *Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing.* Nat Genet, 2008. **40**(12): p. 1413-5.
103. Wang, E.T., et al., *Alternative isoform regulation in human tissue transcriptomes.* Nature, 2008. **456**(7221): p. 470-6.
104. Wan, Y., et al., *Landscape and variation of RNA secondary structure across the human transcriptome.* Nature, 2014. **505**(7485): p. 706-9.
105. Sexton, A.N., et al., *Interpreting Reverse Transcriptase Termination and Mutation Events for Greater Insight into the Chemical Probing of RNA.* Biochemistry, 2017. **56**(35): p. 4713-4721.
106. Li, F., et al., *Global analysis of RNA secondary structure in two metazoans.* Cell Rep, 2012. **1**(1): p. 69-82.
107. Sun, L., et al., *RNA structure maps across mammalian cellular compartments.* Nat Struct Mol Biol, 2019. **26**(4): p. 322-330.
108. Wilbert, M.L., et al., *LIN28 binds messenger RNAs at GGAGA motifs and regulates splicing factor abundance.* Mol Cell, 2012. **48**(2): p. 195-206.
109. Mustoe, A.M., et al., *Pervasive Regulatory Functions of mRNA Structure Revealed by High-Resolution SHAPE Probing.* Cell, 2018. **173**(1): p. 181-195 e18.
110. Lau, S.K., et al., *Coronavirus HKU1 and other coronavirus infections in Hong Kong.* J Clin Microbiol, 2006. **44**(6): p. 2063-71.
111. Zhou, P., et al., *A pneumonia outbreak associated with a new coronavirus of probable bat origin.* Nature, 2020. **579**(7798): p. 270-273.
112. Su, Y.C.F., et al., *Discovery and Genomic Characterization of a 382-Nucleotide Deletion in ORF7b and ORF8 during the Early Evolution of SARS-CoV-2.* mBio, 2020. **11**(4).
113. Gralinski, L.E. and R.S. Baric, *Molecular pathology of emerging coronavirus infections.* J Pathol, 2015. **235**(2): p. 185-95.
114. V'Kovski, P., et al., *Coronavirus biology and replication: implications for SARS-CoV-2.* Nat Rev Microbiol, 2021. **19**(3): p. 155-170.
115. Hussain, S., et al., *Identification of novel subgenomic RNAs and noncanonical transcription initiation signals of severe acute respiratory syndrome coronavirus.* J Virol, 2005. **79**(9): p. 5288-95.
116. Siwy Ling Yang, L.D., Danielle E. Anderson, Yu Zhang, Ashley J Aw, Su Ying Lim, Xin Ni Lim, Kiat Yee Tan, Tong Zhang, Tanu Chawla3, Yan Su, Alexander Lezhava, Andres Merits, Lin-Fa Wang, Roland G. Huber, Yue Wan1, *Comprehensive mapping of SARS-CoV-2 interactions in vivo reveals functional virus-host interactions.* Nature Communications, 2021.
117. Lan, T.C.T., et al., *Insights into the secondary structural ensembles of the full SARS-CoV-2 RNA genome in infected cells.* bioRxiv, 2021: p. 2020.06.29.178343.

118. Darty, K., A. Denise, and Y. Ponty, *VARNA: Interactive drawing and editing of the RNA secondary structure.* Bioinformatics, 2009. **25**(15): p. 1974-5.
119. Mnatsakanov, R. and K. Sarkisian, *Varying kernel density estimation on R.* Stat Probab Lett, 2012. **82**(7): p. 1337-1345.
120. Zinshteyn, B., et al., *Assaying RNA structure with LASER-Seq.* Nucleic Acids Res, 2019. **47**(1): p. 43-55.

# Appendix

>yvrC (AdoCbl riboswitch)

GGTCATATAGTATAATTACTCCGAAAAACGGATACGAATGTCAAATAGGTGCCGGTCCGTGAACAACAGCCGGCTTAAAA
GGGAAACCGGTAAAAGCCGGTGCGGTCCCGCCACTGTAATTGGCCAAGCGCCAAGAGCCAGGATACCTGCCTGTTTGAT
CAGCACGAATTCTGCGAGGACAGATGATGTGTAAACAATAGGCTTTTTTGTGTTGTTTACAGCATCTTTACCGTCGTAGAG
ATGCTTTTTTAGTTCGTTTAGGAGGAAAAGATTATGAAAAAACGAGCCGGGATATGGGCTGCGCTTCTGTTGGCTGCTGT
TATGCTGGCAGGCTGCGGCAATCCTGCGGATCAGAAG

>Yeast SCR1

GAGGCTGTAATGGCTTTCTGGTGGGATGGGATACGTTGAGAATTCTGGCCGAGGAACAAATCCTTCCTCGCGGCTAGACA
CGGATTGCACGCCCTTTGGGCAAGGGATAGTTCTCTATTCCGCACCGTGCCCTGTTGTGGCAACCGTCTTTCCTCCGTCGT
AAATTTGTCCTGGGCAGAGCTGTCTGCCCGGAGGCGGGAGAGTCCGTTCTGAAGTGTCCCGGCTATAATAAATCGATCTT
TGCGGGCAGCCCGTTGGCAGGAGGCGTGAGGAATCCGTCTCTCTGTCTGGTGCGGCAAGGTAGTTCTGGGTCCTTAGGG
GCTCCACCTTCACCGCTGTTAGGGGAGTTTTATCCAGCGTCAGCAAAGGTGACCCGTGATGGAGGCGGCCGGGATAGCA
CATATCAGTCGGATAATTGTGCAAGTTGATCGCTTCGGCGGTTTAATTTGGCGGTGCCATCAGGATTTACTCGCACATTGT
GGCCGTTCCCTCGGGGATGGAGTGTGTCCTGAACCATATTTTT

>Yeast Hac1 3'UTR

GGCGTTATTATCGCTGTTGGTGGGTTTTTTCTTTTCATATATTTCTTTTTCGCTTAGTGGTTTCTACTGTTCTGTCTCCGGTTA
GTGTGTGCTACTTCAACCGAAGAAGAAGAGGCTTTTCAAGAATGCAAACGTGAGGTTGGCGCGCCCTCCTACAATTATTT
GTGGCGACTGGGCAGCGACACTGAACATAGCTCTTGAACAAGACCCTTTTTTGGCTGCAAGGAGCAAGACTGGCTGGGG
TTCCACCTCAAAGAGCCACGCTCTGC

>Human RPS12

GCCTCAATCCGGAAACAGAAACTCCCCAGGCCCGGCACTGCGGGAGCCACCTCTTTCGAGGGTGGGAGTGCACATGCGC
ACAGGGGACCTGAAAAACCCTTAAATACCGGCATGCGTGCGGATGAGGCCTCTTTCCCTGCCGCCGCCGAGTCGCGCGG
AGGCGGAGGCTTGGGTGCGTTCAAGATTCAACTTCACCCGTAACCCACCGCCATGGCCGAGGAAGGCATTGCTGCTGGA
GGTGTAATGGACGTTAATACTGCTTTACAAGAGGTTCTGAAGACTGCCCTCATCCACGATGGCCTAGCACGTGGAATTCG
CGAAGCTGCCAAAGCCTTAGACAAGCGCCAAGCCCATCTTTGTGTGCTTGCATCCAACTGTGATGAGCCTATGTATGTCAA
GTTGGTGGAGGCCCTTTGTGCTGAACACCAAATCAACCTAATTAAGGTTGATGACAACAAGAAACTAGGAGAATGGGTA
GGCCTTTGTAAAATTGACAGAGAGGGGAAACCCCGTAAAGTGGTTGGTTGCAGTTGTGTAGTAGTTAAGGACTATGGCA
AGGAGTCTCAGGCCAAGGATGTCATTGAAGAGTATTTCAAATGCAAGAAATGAAGAAATAAATCTTTGGCTCACA

>Human RPS29

CTTTTACCTCGTTGCACTGCTGAGAGCAAGATGGGTCACCAGCAGCTGTACTGGAGCCACCCGCGAAAATTCGGCCAGGG
TTCTCGCTCTTGTCGTGTCTGTTCAAACCGGCACGGTCTGATCCGGAAATATGGCCTCAATATGTGCCGCCAGTGTTTCCGT
CAGTACGCGAAGGATATCGGTTTCATTAAGTTGGACTAAATGCTCTTCCTTCAGAGGATTATCCGGGGCATCTACTCAATG
AAAAACCATGATAATTCTTTGTATATAAAATAAACATTTGAAAAAACCCTTCA

>Tetrahymena ribozyme

GGCTCTCTAAATAGCAATATTTACCTTTGGAGGGAAAAGTTATCAGGCATGCACCTGGTAGCTAGTCTTTAAACCAATAGA
TTGCATCGGTTTAAAAGGCAAGACCGTCAAATTGCGGGAAAGGGGTCAACAGCCGTTCAGTACCAAGTCTCAGGGGAAA
CTTTGAGATGGCCTTGCAAAGGGTATGGTAATAAGCTGACGGACATGGTCCTAACCACGCAGCCAAGTCCTAAGTCAACA
GATCTTCTGTTGATATGGATGCAGTTCACAGACTAAATGTCGGTCGGGGAAGATGTATTCTTCTCATAAGATATAGTCGGA
CCTCTCCTTAATGGGAGCTAGCGGATGAAGTGATGCAACACTGGAGCCGCTGGGAACTAATTTGTATGCGAAAGTATATT
GATTAGTTTTGGAGTACTCG

>xpt (Purine riboswitch)

GGAATATAATAGGAACACTCATATAATCGCGTGGATATGGCACGCAAGTTTCTACCGGGCACCGTAAATGTCCGACTATG
GGTGAGCAATGGAACCGCACGTGTACGGTTTTTTGTGATATCAGCATTGCTTGCTCTTTATTTGAGCGGGCAATGCTTTTTT
TATTCTCATAACGGAGGTAGACAGGATGGAAGCACTGA

>ribD (FMN riboswitch)

GGAAGGACAAATGAATAAAGATTGTATCCTTCGGGGCAGGGTGGAAATCCCGACCGGCGGTAGTAAAGCACATTTGCTT
TAGAGCCCGTGACCCGTGTGCATAAGCACGCGGTGGATTCAGTTTAAGCTGAAGCCGACAGTGAAAGTCTGGATGGGAG
AAGGATGATGAGCCGCTATGCAAAATGTTTAAAAATGCATAGTGTTATTTCCTATTGCGTAAAATACCTAAAGCCCCGAAT
TTTTTATAAATTCGGGGCTTTTTTGACGGTAAATAACAAAAGAGGGGAGGGAAACAAATGGAAGA

>Dengue 1 3' UTR

GGGTCAACACGTTTACAAAATAAAGGAAAATAAGAAATCAAACAAGGCAAGAAGTCAGGCCGGATTAAGCCATAGTACG
GTAAGAGCTATGCTGCCTGTGAGCCCCGTCTAAGGACGTAAAATGAAGTCAGGCCGAAAGCCACGGTTTGAGCAAACCG
TGCTGCCTGTAGCTCCATCGTGGGGATGTAAAAACCTGGGAGGCTGCAACCCATGGAAGCTGTACGCATGGGGTAGCAG
ACTAGTGGTTAGAGGAGACCCCTCCCGAAACACAACGCAGCAGCGGGGCCCAACACCAGGGGAAGCTGTACCCTGGTGG
TAAGGACTAGAGGTTAGAGGAGACCCCCCGCATAACAATAAACAGCATATTGACGCTGGGAGAGACCAGAGATCCTGCT
GTCTCTACAGCATCATTCCAGGCACAGAACGCCAGAAAATGGAATGGTGCTGTTGAATCAACAGGTTCT

>ypaA (FMN riboswitch)

GGAATTTCATATGATCAATCTTCGGGGCAGGGTGAAATTCCCTACCGGCGGTGATGAGCCAATGGCTCTAAGCCCGCGAG
CTGTCTTTACAGCAGGATTCGGTGAGATTCCGGAGCCGACAGTACAGTCTGGATGGGAGAAGATGGAGGTTCATAAGCG
TTTTGAAATTGAATTTTTCAAACGTTTCTTTGCCTAGCCTAATTTTCGAAACCCCGCTTTTATATATGAAGCGGTTTTTTTATT
GGCTGGAAAAGAACCTTTCCGTTTTCGAGTAAGATGTGATCGAAAAGGAGAGAATGAAGTGAAAGTAAAAAAATTAGTT
GTGGTCAGCATGCTGAGCAGCATTGCATTT

>lysC (Lysine riboswitch)

CCAAGTAATACGACTCACTATAGGAATTTCATAGTTAGATCGTGTTATATGGTGAAGATAGAGGTGCGAACTTCAAGAGT
ATGCCTTTGGAGAAAGATGGATTCTGTGAAAAAGGCTGAAAGGGGAGCGTCGCCGAAGCAAATAAAACCCCATCGGTAT
TATTTGCTGGCCGTGCATTGAATAAATGTAAGGCTGTCAAGAAATCATTTTCTTGGAGGGCTATCTCGTTGTTCATAATCAT
TTATGATGATTAATTGATAAGCAATGAGAGTATTCCTCTCATTGCTTTTTTTTATTGTGGACAAAGCGCTCTTTCT

>ydaO (ATP riboswitch)

CCAAGTAATACGACTCACTATAGAAAACAAATCGCTTAATCTGAAATCAGAGCGGGGGACCCAATAGAACGGCTTTTTGC
CGTTGGGGTGAATCCTTTTTAGGTAGGGCTAACTCTCATATGCCCGAATCCGTCAGCTAACCTCGTAAGCGTTCGTGAGAG
GAGATGAATGAAACCTGTGTTCGATGTTATGGCACAGGGGCATCCGTTTGCCTCTGTGTTTTTTGTTGTTCATTTTTGAATG
CAAATCGCATACAGAGACATCATAGCAGAA

>ykoK (Magnesium riboswitch)

CCAAGTAATACGACTCACTATAGTCAATTCAATTAGGAACTTCGTTAGGTGAGGCTCCTGTATGGAGATACGCTGCTGCCC
AAAAATGTCCAAAGACGCCAATGGGTCAACAGAAATCATCGACATAAGGTGATTTTTAATGCAGCTGGATGCTTGTCCTA
TGCCATACAGTGCTAAAGCTCTACGATTGAAGGCGCCCGCACGCTTTTTTTGCCGTGCTTCTTTCACCTTCAATCCCGAAGG
CTTTTTTTATGCCTTTAAAACGAAACCAATCAAAGGAGGTGCGGAGTATGGATTCTCCCCATTCGAGTCAATCTGAAGAA
GTACCCATATACTATGACAGCAAG

> Bacillus subtilis 16S rRNA

TTATCGGAGAGTTTGATCCTGGCTCAGGACGAACGCTGGCGGCGTGCCTAATACATGCAAGTCGAGCGGACA
GATGGGAGCTTGCTCCCTGATGTTAGCGGCGGACGGGTGAGTAACACGTGGGTAACCTGCCTGTAAGACTGG
GATAACTCCGGGAAACCGGGGCTAATACCGGATGGTTGTTTGAACCGCATGGTTCAAACATAAAAGGTGGCT
TCGGCTACCACTTACAGATGGACCCGCGGCGCATTAGCTAGTTGGTGAGGTAACGGCTCACCAAGGCGACGA
TGCGTAGCCGACCTGAGAGGGTGATCGGCCACACTGGGACTGAGACACGGCCCAGACTCCTACGGGAGGCA

GCAGTAGGGAATCTTCCGCAATGGACGAAAGTCTGACGGAGCAACGCCGCGTGAGTGATGAAGGTTTTCGG
ATCGTAAAGCTCTGTTGTTAGGGAAGAACAAGTGCCGTTCGAATAGGGCGGTACCTTGACGGTACCTAACCA
GAAAGCCACGGCTAACTACGTGCCAGCAGCCGCGGTAATACGTAGGTGGCAAGCGTTGTCCGGAATTATTGG
GCGTAAAGGGCTCGCAGGCGGTTTCTTAAGTCTGATGTGAAAGCCCCCGGCTCAACCGGGGAGGGTCATTGG
AAACTGGGGAACTTGAGTGCAGAAGAGGAGAGTGGAATTCCACGTGTAGCGGTGAAATGCGTAGAGATGTG
GAGGAACACCAGTGGCGAAGGCGACTCTCTGGTCTGTAACTGACGCTGAGGAGCGAAAGCGTGGGGAGCGA
ACAGGATTAGATACCCTGGTAGTCCACGCCGTAAACGATGAGTGCTAAGTGTTAGGGGGTTTCCGCCCCTTAG
TGCTGCAGCTAACGCATTAAGCACTCCGCCTGGGGAGTACGGTCGCAAGACTGAAACTCAAAGGAATTGACG
GGGGCCCGCACAAGCGGTGGAGCATGTGGTTTAATTCGAAGCAACGCGAAGAACCTTACCAGGTCTTGACAT
CCTCTGACAATCCTAGAGATAGGACGTCCCCTTCGGGGGCAGAGTGACAGGTGGTGCATGGTTGTCGTCAGC
TCGTGTCGTGAGATGTTGGGTTAAGTCCCGCAACGAGCGCAACCCTTGATCTTAGTTGCCAGCATTCAGTTGG
GCACTCTAAGGTGACTGCCGGTGACAAACCGGAGGAAGGTGGGGATGACGTCAAATCATCATGCCCCTTATG
ACCTGGGCTACACACGTGCTACAATGGACAGAACAAAGGGCAGCGAAACCGCGAGGTTAAGCCAATCCCACA
AATCTGTTCTCAGTTCGGATCGCAGTCTGCAACTCGACTGCGTGAAGCTGGAATCGCTAGTAATCGCGGATCA
GCATGCCGCGGTGAATACGTTCCCGGGCCTTGTACACACCGCCCGTCACACCACGAGAGTTTGTAACACCCGA
AGTCGGTGAGGTAACCTTTTAGGAGCCAGCCGCCGAAGGTGGGACAGATGATTGGGGTGAAGTCGTAACAA
GGTAGCCGTATCGGAAGGTGCGGCTGGATCACCTCCTTT