UNIVERSITY OF CALIFORNIA
SANTA CRUZ

MODIFICATION DETECTION USING NANOPORE SEQUENCING

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

DOCTOR OF PHILOSOPHY

in

BIOMOLECULAR ENGINEERING AND BIOINFORMATICS

by

Andrew D. Bailey IV

December 2021

The Dissertation of Andrew D. Bailey IV
is approved:

Professor Benedict Paten, Chair

Professor Mark Akeson

Professor David Haussler

Professor Manuel Ares Jr.

Peter Biehl
Vice Provost and Dean of Graduate Studies

Table of Contents

List of Figures

List of Tables

Abstract

Modification Detection using Nanopore Sequencing

by

Andrew D. Bailey IV

Both DNA and RNA modifications play critical roles in cell regulation. Traditionally, a chemical selection process alters base pairing or sequencing coverage is used to sequence modified nucleotides. Therefore, a new chemical 2'O labeling process needs to be created for each modification. Currently, we do not have methods for sequencing the majority of the over 150 RNA and over 40 DNA modifications. However, with nanopore sequencing, we can directly detect modifications on native DNA or RNA reads without any selection or chemical 2'O labeling techniques. Nanopore sequencing measures the change in current across a nanopore as a polynucleotide threads through the nanopore and we can use this signal to identify modifications.

In chapter 2, we present a framework for the unsupervised determination of the number of nucleotide modifications from nanopore sequencing readouts. We demonstrate the approach can effectively recapitulate the number of modifications, the corresponding ionic current signal levels, as well as mixing proportions under both DNA and RNA contexts. We further show, by integrating information from multiple detected modification regions, that the modification status of DNA and RNA molecules can be inferred. This method forms a key step of de novo characterization of nucleotide modifications.

In chapter 3, we present a graph convolutional network-based deep learning framework for predicting the mean of kmer distributions from corresponding chemical structures.

We show such a framework can generalize the chemical information of the 5-methyl group from thymine to cytosine by correctly predicting 5-methylcytosine-containing DNA 6mers.

In chapter 4, using a combination of yeast genetics and nanopore direct RNA sequencing, we have developed a reliable method to track the modification status of single rRNA molecules at 37 sites in 18S rRNA and 73 sites in 25S rRNA. We use our method to characterize patterns of modification heterogeneity and identify concerted modification of nucleotides found near functional centers of the ribosome. Distinct undermodified subpopulations of rRNAs accumulate when ribosome biogenesis is compromised by loss of Dbp3 or Prp43-related RNA helicase function. Modification profiles are surprisingly resistant to change in response to many genetic and environmental conditions that affect translation, ribosome biogenesis, and pre-mRNA splicing. The ability to capture complete modification profiles for RNAs at single-molecule resolution will provide new insights into the roles of nucleotide modifications in RNA function.

To love of my life,

Jordan Mravca-Bailey

thank you.

Acknowledgments

First and foremost I want to thank my parents Lauren M. Purcell and Andrew D. Bailey III. Without their love and support I would not have tried to do something as emotionally and financially reckless as getting a Ph.D. I would also like to thank my wife, Jordan Mravca-Bailey, for all that she does for us. We are with each other for most hours of most days and without her positive attitude and endless support I would not have graduated. I would also like to thank my brother Egan Bailey for always making me laugh. Finally, I would like to thank the rest of my family who have supported me through the ups and downs of my work.

I also want to thank Benedict Paten who has given me the space, resources, patience and mentorship to work on challenging problems. Thank you to Manny Ares for trusting me with your data and editing our paper together with a fine tooth comb. Jason Talkish and Manny Ares have been excellent collaborators and their advice and knowledge of yeast biology was indispensable to the final chapter of this thesis. I also want to thank the committee members Mark Akeson and David Haussler for their time and insight into new ways of looking at my problems.

Other graduate students and post-docs at University of California, Santa Cruz have done tremendous amounts of work and been wildly supportive through these many years. Specifically, I want to thank Yanni Anastopoulos, Ryan Lorig-Roach, Colleen Bosworth, Miten Jain, Art Rand, Jordan Eizenga, Kishwar Shafin, Trevor Pesout, Marina Haukness, David Parks and Jon Akutagawa.

Last but certainly not least, I want to thank Hongxu Ding. Ding came to UCSC

while I was languishing in my 3rd year. I cannot understate how important Ding was to helping me see how to tackle hard problems and take ideas from the whiteboard all the way to publication. Ding is diligent and generous with his time. He is always ready to talk about technical problems until we come to a solution. Without his support and frank conversations I would still be a graduate student. Thank you Ding.

Hongxu Ding, Andrew Bailey and Benedict Paten conceived the idea for the project in Chapter 2 (Gaussian Mixture Model-Based Unsupervised Nucleotide Modification Number Detection Using Nanopore Sequencing Readouts). Hongxu Ding and Andrew Bailey performed the analysis. Miten Jain and Hugh Olsen collected and pre-processed the data. Benedict Paten supervised the project. Hongxu Ding, Andrew Bailey and Benedict Paten wrote the manuscript.

Hongxu Ding conceived the idea for the project in Chapter 3 (Towards Inferring Nanopore Sequencing Ionic Currents from Nucleotide Chemical Structures). Ioannis Anastopoulos performed deep learning framework modeling, optimization and analysis. Andrew Bailey and Hongxu Ding performed the nanopore sequencing data analysis. Hongxu Ding, Joshua Stuart and Benedict Paten supervised the project. All authors prepared the manuscript.

All authors contributed to conception and experimental design. JT, MA and HI grew yeast, extracted RNA, prepared libraries, ran the nanopore sequencer and performed base calling. AB performed most of the nanopore data analysis, including model building, clustering, and correlation measurements. HD provided feedback and advice throughout the development and analysis process. AD and SM built some of the clustering analysis scripts.

AD and SM's participation in this research took place under the auspices of the Science Internship Program at the University of California Santa Cruz. MA and BP supervised the project. All authors assisted in manuscript preparation.

All authors contributed to conception and experimental design for the project in Chapter 4 (Single-molecule modification profiling of Saccharomyces cerevisiae ribosomal RNA reveals concerted modification at functional locations in the ribosome). JT, MA and HI grew yeast, extracted RNA, prepared libraries, ran the nanopore sequencer and performed base calling. AB performed most of the nanopore data analysis, including model building, clustering, and correlation measurements. HD provided feedback and advice throughout the development and analysis process. AD and SM built some of the clustering analysis scripts. AD and SM's participation in this research took place under the auspices of the Science Internship Program at the University of California Santa Cruz. MA and BP supervised the project. All authors assisted in manuscript preparation.

Chapter 1

Introduction

1.1 Biological Importance of Modifications

Both RNA and DNA modifications are dynamically regulated and play important roles in cell function. 5-methylcytosine (5mC) is the most studied and abundant DNA modification [19]. 5mC modifications help regulate histone binding, chromatin structure, transcription factor binding, transcription start sites, transposition, recombination and overall genome stability [130, 19]. Irregular methylation patterns of 5mC on CpG islands, CpG island shores, and repeat regions are linked to several cancers, neurological disorders and autoimmune diseases [130, 133]. A table of diseases related to epigenetic modifications can be found in Figure [130]. While there has been extensive research into 5mC, we are learning more about other important but less frequent DNA modifications[33]. N6-methyladenine (6mA) modulates transcription and causes resistance against host immune responses in several bacteria[133]. 6mA also seems to effect nucleosome positioning and seems to play a role in mammalian development[133, 97]. 5-hydroxymethylcytosine (5hmC) in DNA is

maintained at enhancers and genes, can recruit specific binding proteins and is an epigenetic signal for neuronal development[19, 33, 83]. For example, mice brains have up to 10 times more 5hmC than most average cells[19, 33, 83]. Although there are over 40 verified DNA modifications, most modifications do not have a sequencing based detection method[155]. Therefore, we know very little about many of these modifications and their role within cell regulatory networks[155]. It is still an open question how many DNA modifications effect cell regulatory systems and until we have accurate sequencing and detection of all DNA modifications, it will be difficult to make progress towards a complete understanding of the epigenetic landscape.

RNA post-transcriptional modifications (PTMs) are also crucial for cell function [19, 74]. There are over 150 RNA PTMs and the majority of which have not been identified using sequencing[15]. Over the past several years, we have been discovering new RNA modifications and the variety of roles they play in all types of RNA including mRNA, tRNA, rRNA, snRNA and snoRNA[74]. RNA modifications have been linked to development of cognitive functions, neurological defects, breast cancer, genetic birth defects and diabetes[6, 186, 70, 38, 37, 10, 74] Several tRNAs and rRNAs have specific modifications which are required for the RNAs to function[74]. 6mA, the most abundant mRNA modification in mammals, has binding proteins that signal degradation of transcripts or increase translational efficiency[174]. The regulation of 6mA also directly effects neuronal signaling pathways[69]. Although RNA modifications are an important part of the cell regulatory system, we cannot resolve the majority of modifications at a per read level which limits our ability to understand of the regulatory importance of the RNA modification landscape[74].

For example, back in 2010, we thought that RNA modifications were irreversible [76]. However, recently we discovered that many modifications are reversible, evolutionarily conserved and required for correct function of mRNA and ncRNA[145, 94]. Improving modification detection techniques an important step if we want to discover the role of modifications in the regulatory networks of the cell.

1.2 Conventional Modification Detection

Classical approaches to modification detection were thin-layer chromatography and capillary electrophoresis[74]. However, most non-sequencing based modification detection is done by liquid chromatography-tandem mass spectrometry (LC-MS/MS) or cryogenic electron microscopy (cryo-EM)[74, 23, 164]. Although all of these techniques made it possible to identify the presence of modified nucleotides, they do not have the resolution for single-molecule modification detection[23, 74]. LC-MS/MS is the most accurate way that we currently have to determine the chemical signature of modifications[162]. LC-MS/MS separates already digested DNA or RNA fragments, often according to their polarity, by high-performance liquid chromatography (HPLC)[23, 164]. These fragments are then ionized into sub-fragments which in turn are selected by their mass to charge ratio (m/q) and analyzed by mass spectrometry[23, 164]. However, this technique requires a fairly extensive laboratory protocol, expensive equipment and expertise in the lab[23, 164]. More importantly, a pure sample of several identical molecules are required to be analyzed via mass spectrometry which removes the ability to detect modified nucleotides on an individual molecule basis[23, 164]. Cryo-EM can resolve 3D structures of molecules at 2-4

Angstroms[114]. Cryo-EM starts by freezing a purified sample within a non-crystalline structure. Then transmission electron microscopy takes pictures of the several identical molecules of which have froze in different orientations[136]. These 2D representations are then aggregated to form a 3D model of the molecule of interest[136]. Although recent advances in cryo-EM have made this technique able to detect modifications, is time consuming, expensive, cannot process a large number of samples efficiently[114, 136]. Therefore, we need to develop high throughput sequencing techniques in order to identify modifications across thousands or millions of individual molecules[114, 136].

1.3 Next Generation Sequencing

In order to understand the underlying cell state, it is often preferred to identify modifications via sequencing so that we can understand the heterogeneity of modifications within a sample. Traditional chain termination sequencing has been overtaken by the massively parallel sequencing by synthesis paradigms of Illumina, Ion Torrent and pyrosequencing [141, 137, 138, 11, 170]. These platforms have made dramatic progress in accuracy, throughput and price per base over the past 20 years[141, 137, 138, 11, 170]. Sequencing by synthesis (SBS) measures byproducts of polymerase nucleotide incorporation reactions and uses Watson-Crick base pairing to infer the target sequence[170]. SBS techniques also need a DNA amplification step in order to generate enough signal from the sequencing byproducts and have relatively short read lengths (150-800bp)[137, 138, 11]. For example, Ion Torrent measures the change in pH of a small well which contains a bead covered in amplified DNA sequences[138]. A change in pH corresponds to the release of

protons during nucleotide incorporation events which is then correlated to the expected nucleotide[138]. Pyrosequencing converts the pyrophosphate byproduct of a nucleotide incorporation event into light via a cascade of enzymatic reactions which culminate in the oxidization of luciferin and light generation[137]. Illumina, the current industry leader, measures the light emitted from fluorescently labeled reversible-terminator nucleotides as they incorporate into clusters of amplified DNA (see Figure 1.3) [11, 150, 170]. The past 20 years have been dominated by these sequencing by synthesis platforms[150]. Therefore, modification detection protocols require added information to make predictions because modified bases to not change the Watson-Crick base pairing mechanism[74]. As outlined in Figure 1.2, RIP-seq, Chem-Seq and mismatch signatures are the three main sequencing approaches to modification detection[74]. RIP-seq uses a chemical enrichment to isolate reads with a specific modification before sequencing[42]. Often the enrichment is performed with an antibody[42]. After the enrichment and sequencing, read coverage peaks over the targeted modified nucleotide[42]. This technique can generate accurate per-reference position modifications, but it only works for nucleotides with highly specific antibodies targeting the modified nucleotide of interest[42]. For RNA modification detection, there were antibodies for just 6mA, 1mA, 5mC, and 5hmC in 2017[74]. Another approach to modification detection, chem-seq, uses differences in reactivity of modified bases compared to the canonical nucleotides to extract information about their position. The classic example of chem-seq is bisulfite sequencing. Bisulfite deaminates cytosine to uracil and does not effect 5mC[53]. Therefore, all called cytosines in a bisulfite treated sample are 5mC and 5mC can be detected by comparing against an untreated bisulfite sample[53]. Another example

of Chem-seq is pseudouridine-seq which uses 1-cyclohexyl-(2-morpholinoethyl)carbodiimide metho-p-toluene sulfonate (CMCT) to modify guanosine, uridine, and pseudouridine with carbodiimide (CMC)[146]. After CMCT treatment, an alkaline treatment at pH 10.3 removes CMC from all other sites except the N3 of pseudouridine[146]. Reverse-transcription is blocked by CMC so there is a pileup of reads terminating at pseudouridine sites[146]. Although Chem-seq also offers per-reference position, it is dependent upon the discovery of a highly targeted and efficient chemical reaction for every unique modified nucleotide. One last approach to modification detection is mismatch-signature based analyses. Mismatch-signature analysis relies on modified nucleotides either entirely changing the Watson-Crick base-pairing or generating non-random errors. For example, a group used known modifications in tRNA to create a model of mismatch signatures to detect modifications in other datasets[139]. Since SBS technology requires amplification and incorporation byproducts to sequence, none of the detection methods outlined are able to identify modified nucleotides without a highly specialized chemical treatment[137, 138, 11]. However, with the development of real-time, single-molecule (RTSM) sequencing, there are now platforms which can generate relatively accurate sequencing data from the native polynucleotide and detect modifications without chemical treatment or amplification steps.

1.4 Real-time Single-molecule Sequencing

Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT) are currently two most prevalent RTSM sequencing platforms[150]. PacBio sequencing measures the fluorescence from incorporation of fluorescently tagged nucleotides by a DNA poly-

merase [91]. As seen in Figure 1.4, PacBio is able to detect single nucleotide incorporations by fixing the DNA polymerase to the bottom of a small hole, limiting light exposure from surrounding reactions and focusing the detection of nucleotide incorporation to an individual polymerase[91, 150]. Figure 1.4 shows that PacBio sequencing can detect modifications by correlating the change in the rate of incorporation by a DNA polymerase to detect N6-methyladenine, 5-methylcytosine and 5-hydroxymethylcytosine[51]. It has also been shown that the same sequencing technique can be used using a reverse transcriptase to detect m6A in RNA[140]. PacBio sequencing has random errors at a rate of around 10% but allows for multiple readings of the same nucleotides because DNA templates are circularized during library preparation[150]. Each subsequent prediction of an individual base increases the accuracy of PacBio reads[51]. However, as the template strand gets longer, it takes longer to process one complete loop around the template strand[51]. So, there is a trade off between very long reads and improved accuracy via several passes around the template[51]. Also, PacBio sequencers cost between $350,000 and $700,000[1]. In comparison, ONT nanopore sequencing can also detect modified nucleotides on a per read basis with no read length limitation and the upfront cost of a Minion is only $1000`https://nanoporetech.com/products/minion` [57]. As seen in Figure 1.3, nanopore sequencing measures the current across a parallel array of nanopores as an enzyme controls the rate of translocation of the polynucleotide through each pore[35, 57]. The signal recorded from each nanopore corresponds to the nucleotides within the pore. Therefore, nanopore sequencing has the ability to directly detect modified nucleotides on the native polynucleotide and has already been shown to identify 6mA, inosine (I), 7-methylguanine

[1] https://allseq.com/knowledge-bank/sequencing-platforms/pacific-biosciences/

(7mG), pseudouridine (Q) in RNA, and 6mA, 5mC, 5hmC, and thymidine analogs including EdU, FdU, BrdU and IdU in DNA[135, 152, 154, 112, 184, 99, 115, 108]. Although, we can detect some modifications within some contexts, we know very little about the location and function of most modifications.

1.5 Data Analysis of Nanopore sequencing

1.5.1 Understanding Nanopore Sequencing Signal

The current through the pore is recorded via an analog to digital converter (ADC) which records 13 bits of information at 4000Hz. The current is directly related to the nucleotides within the constriction site of the nanopore[77]. The constriction site of the CsgG pore used for the R9/R9.4 sequencing chemistry has as height of about 0.9nm which corresponds to about 3 nucleotides[60, 20]. Often, in order to correlate signal to specific nucleotides within the pore (kmer), the raw signal is first segmented and summarized using the mean, standard deviation, event start time and event duration (Figure 1.5)[34]. Event detection or segmentation algorithms determine where there are significant changes in the current level. The most common event detection algorithm uses a sliding Welch's t-test to determine event boundaries (Equation 1.5) [2]. The segmented events correspond to a set of nucleotides (kmer) within the pore. In theory, given any sequence context, all events associated with a kmer of optimal length k should have the same corresponding event signal. However, this is not the case so people often model the distribution of signals for a given kmer as Gaussian [135, 108, 152]. Many modification detection algorithms are dependent

[2] https://github.com/UCSC-nanopore-cgl/signalAlign/commit/e468812eb0604562ac049828e4faf1eb046e48aa#diff-60dfcf04ab7bcc7174c8a2fc19d7f83a

upon the event segmentation of signal and the subsequent alignment of basecalled sequence to the events[135, 108, 152]. Therefore, it is important to understand basecalling before moving on to modification detection algorithms.

1.5.2 Basecalling Nanopore Reads

Over the past few years there have been several basecallers; Nanocall, BasecRAWller, DeepNano, Chiron, Metrichor, Albacore, Guppy, Scrappie [3], Flappie [4], and Runnie [5] to just name a few[34, 158, 17, 166]. DeepNano, BasecRAWller and Chiron showed that using machine learning based approaches to basecalling were much more accurate than hidden Markov models[181]. Figure 1.8 shows the general structure of the basecalling model used by Chiron. The Chiron model takes in normalized current readings, $\{x_1...x_{n_j}\}$, and feeds them through three stacked residual layers which feed into three stacked bidirectional long short-term memory layers (BLSTM)[71, 144]. The BLSTM's are important for integrating information from before and after the current time step into the prediction of the current time step[71, 144]. However, the key difference insight from Chiron was the inclusion of the residual layers which increased accuracy from around 80% to 90%[158, 17, 166, 66]. Residual layers are feature extracting tools which allows the model to find valuable information in the raw signal data without the vanishing gradient problem[182, 66, 166]. The final BLSTM outputs are fed through a fully connected layer with 5 output nodes representing $\{A, T, G, C, b\}$. These outputs are decoded by a connectionist temporal classification (CTC) decoder and finally converted into predicted nucleotide sequence[61]. Chiron was

[3] `https://github.com/nanoporetech/scrappie`
[4] `https://github.com/nanoporetech/flappie`
[5] `https://github.com/nanoporetech/flappie/blob/master/RUNNIE.md`

the best open source basecaller at it's time so all of the subsequent basecallers have very similar architectures[181]. However, there are some key differences between the Chiron model architecture and the ONT models like Flappie and Runnie. For example, ONT uses GRU recurrent networks instead of LSTMs and has played with the size and number of layers[181, 28]. Flappie uses a "flip-flop" decoding layer which allows for very specific transitions between "flip" and "flop" states which is then decoded as a linear conditional random field [6]. Runnie encodes nucleotide run lengths as a discrete Weibull distributions [7]. Although there are some differences in basecalling networks, all downstream analysis of nanopore sequencing comes after the basecalling step. Ideally, we would have basecallers directly identify modified nucleotides. Flappie can detect 5mC in DNA but detecting infrequent modifications becomes very difficult for basecallers[132]. This is because modification detection in a basecalling framework creates an extremely unbalanced dataset classification problem. Therefore, it is often beneficial to incorporate reference information to correctly detect modifications.

1.5.3 Supervised Nanopore Modification Detection Algorithms

Supervised modification detection algorithms all rely on labelled training data and information regarding the alignment to the reference sequence. Given the basecalled sequence, we use the basecall to reference alignment to determine if the canonical nucleotides or modified nucleotides generated the signal. This approach adds a significant amount of information because we now know the expected nucleotide sequence of the signal as well

[6] `https://github.com/nanoporetech/flappie`
[7] `https://github.com/nanoporetech/flappie/blob/master/RUNNIE.md`

as information about the canonical and expected modified nucleotides. Megalodon[8] is currently under development at ONT as the next iteration of Tombo as a variant/modification detection caller and it uses the output of the basecalling neural network to determine a score for the reference and proposed alternative sequence. However, it currently does not support RNA or any de-novo detection algorithms so we have not explored using Megalodon for our RNA modification detection work. Instead we use some of the more established modification detection algorithms. Both Nanopolish eventalign and signalAlign use an hidden Markov Model (HMM) to generate an event to reference alignment[152, 135]. First, an adaptive banded alignment is generated between the basecalled sequence and the events so that each event is mapped to a predicted kmer. The adaptive banded alignment is similar to the Smith-Waterman alignment but only computes a fraction of the total alignment matrix[55, 160, 45]. Then a guide alignment between the basecalled sequence and the reference is produced using a long read mapper, such as minimap2[92]. This event to kmer to reference alignment is used anchor the HMM alignment, only include informative signal and to decrease cost of computation by ignoring large stretches of perfect matches[152, 135]. Both of these algorithms use kmer models to produce emission probabilities and have a hard coded set of transition probabilities. However, there are two main differences between signalAlign and eventalign. First, eventalign uses the viterbi algorithm to determine the best path through the events to generate the reference sequence. On the other hand, signalAlign uses the forward-backward algorithm to determine the posterior probability of an alignment between an individual event and a kmer. Second, signalAlign has the option to use a non-parametric hierarchical Dirchlet process (HDP) to model the emission distributions whereas

[8] https://github.com/nanoporetech/megalodon

eventalign uses the Gaussian distribution to model the event mean and the Inverse-Gamma distribution to model the standard deviation[152, 135]. The HDP uses Gibbs sampling to generate a countably infinite set of shared mixture components which allows for more flexibility in modelling kmer distributions[165]. As seen in Figure 1.7, in order to detect modifications, signalAlign uses a pair-HMM that which allows branch points at specified reference nucleotides to be modelled as a modified nucleotide and eventalign computes the log-likelihood ratio between alignment of target positions to a modified reference compared to canonical reference. Besides these two HMM approaches, Tombo is the other kmer model based modification detection approach. Tombo[9] is based off of the idea that there is no need to generate an intermediate alignment to the basecalled sequence so instead you just generate an alignment directly between events and the reference sequence[108]. Once the alignment is generated, an outlier robust likelihood ratio is calculated across all positions covering a target base[108]. The outlier robust likelihood ratio decreases once the event means fall too far outside the canonical or modified kmer distributions[108]. All of these kmer model alignment based approaches have used labelled modified nanopore reads as training data for generating the kmer distributions of non-canonical nucleotides. The other class of supervised modification detection algorithms are neural network based.

DeepMod[10], DeepSignal[11] and DNAscent v2[12] are a few neural network based modification detection algorithms and they each have several similarities [115, 99, 16]. DeepMod and DeepSignal both classify 5mC in the CpG context of DNA where DNAscent identifies Bromodeoxyuridine (BrdU) in DNA [115, 99]. While all three have unique net-

[9] `https://nanoporetech.github.io/tombo/resquiggle.html`
[10] `https://github.com/WGLab/DeepMod`
[11] `https://github.com/bioinfomaticsCSU/deepsignal`
[12] `https://github.com/MBoemo/DNAscent/`

work architectures, they all have similar tasks and underlying data quality. The network architectures are similar to basecalling networks, relying on either recurrent neural networks, convolutional neural networks or a combination of both[115, 99, 16]. These neural network based approaches require a significant number of well labelled reads with the single target modification covering a significant number of different sequence contexts [115, 99, 16]. The neural network based modification detection algorithms perform well when given well labelled data but given the difficulty of producing high quality labelled sequencing data for many modifications, de novo detection techniques may be required to identify rare or less abundant modifications.

1.5.4 De novo Nanopore Modification Site Detection Algorithms

Nanoraw/Tombo was the first program to de novo identify modified nucleotides[108, 89]. In order to test the de-novo detection of modified nucleotides, the authors generated control data by whole genome amplifying E. coli DNA and created test data by introducing various types of methylases to the PCR amplified E. coli DNA. A summary of the methylases and prediction accuracy can be found in Figure 1.6. Tombo initially normalizes the signal and then generates an alignment between the segmented events and the reference sequence using a banded alignment algorithm [108]. The match probabilities are generated using the z-score of an event coming from the corresponding kmer distribution [108]. After alignment, Tombo resolves skipped bases by re-segmenting signal around deletes in the reference and generating alignments without the option for skips [108]. Once there is a kmer to event mapping, we can do de-novo modification detection or sample compare modification detection[108]. De-novo modification detection calculates the fraction of

events which fall confidently into the expected kmer distribution compared to the fraction of events which fall outside of the kmer distribution [108]. Z-scores are calculated using the following equation where x is the event mean, μ and σ are the mean and standard deviation of the kmer distribution respectively $z = abs(\frac{x-\mu}{\sigma})$ [108]. P-values are calculated using the following equation $\mathrm{p} = 1 + \mathrm{erf}\left(\frac{z}{\sqrt{2}}\right)$ and can be aggregated using Fisher's method. One p-values are calculated, the algorithm classifies reads as modified if the p-value is less than 0.05 or canonical if the p-value is above 0.4. The fraction modified is is calculated per reference position j the following equation $fraction_j = \frac{\sum_{j=0}^{n} p_j \leq t_2}{\sum_{j=0}^{n} p_j \leq t_2 + \sum_{j=0}^{n} p_j \geq t_1}$ where t_1 is the minimum canonical p-value threshold and t_2 is the maximum modified p-value threshold. The problem with this approach is that it is highly dependent upon an accurate kmer model and relies on the assumption that a non-canonical distribution is far away from the canonical distribution. Therefore, in order to look for modified nucleotides, we use Tombo's sample compare method. Sample compare uses an control experiment to generate the expected per reference position kmer distribution which is then compared against a test experiment's distribution. This is done by computing the cumulative distribution function of event means for each reference position using the equation $F_n(x) = \frac{1}{n} \sum_{i=1}^{n} I_{[-\infty,x]}(X_i)$. Then the KS-test is calculated using the following equation $D_n = \sup_x |F_n(x) - F(x)|$. Although the sample compare framework requires more sequencing, the method is more accurate than de-novo and can quickly identify changes between control and experiment sequencing runs[108]. Tombo's sample compare method is extremely similar to the recently published Nanocompore[89]. Nanocompore includes an option to use event duration information along with event means to compute the probability that a position is modified[89].

The original authors did not do a comparison with Tombo so it as of right now it is unclear if there is a significant improvement using the Nanocompore detection framework[89]. The key drawbacks from current de novo detection tools is that there is no way to determine the underlying modification and it is very difficult to identify the specific nucleotide which has been modified. However, there is some evidence that the underlying modification could be determined using information from chemical structures[40].

1.6 Research Outline

Chapter 3, *Towards Inferring Nanopore Sequencing Ionic Currents from Nucleotide Chemical Structures.* The characteristic ionic currents of nucleotide kmers are commonly used in analyzing nanopore sequencing readouts. We present a graph convolutional network-based deep learning framework for predicting kmer characteristic ionic currents from corresponding chemical structures. We show such a framework can generalize the chemical information of the 5-methyl group from thymine to cytosine by correctly predicting 5-methylcytosine-containing DNA 6mers, thus shedding light on the de novo detection of nucleotide modifications.

Chapter 2, *Gaussian Mixture Model-Based Unsupervised Nucleotide Modification Number Detection Using Nanopore Sequencing Readouts.* We present a framework for the unsupervised determination of the number of nucleotide modifications from nanopore sequencing readouts. We demonstrate the approach can effectively recapitulate the number of modifications, the corresponding ionic current signal levels, as well as mixing proportions under both DNA and RNA contexts. We further show, by integrating information

from multiple detected modification regions, that the modification status of DNA and RNA molecules can be inferred.

Chapter 4, *Single-molecule Modification Tracking of Saccharomyces cerevisiae 18S and 25S Ribosomal RNA using Nanopore Sequencing.*

Nucleotides in both RNA and DNA are subject to numerous enzymatic activities that chemically modify them, altering their functional characteristics. Aberrant modification patterns are linked to several cancers, neurological disorders, and autoimmune diseases. Eukaryotic ribosomal RNA is modified at more than 100 locations, in particular at highly conserved and functionally important nucleotides. During ribosome biogenesis, modifications are added at various stages of assembly. The precise timing, order, dependencies or existence of differently modified classes of ribosomes are unknown because no method for evaluating modification status at all sites within a single rRNA molecule is available. Using a combination of yeast genetics and nanopore direct RNA sequencing, we have developed a reliable method to track the modification status of single rRNA molecules at 37 sites in 18S and 73 sites in 25S rRNA. We use our method to identify the presence of long-range correlated modifications in wild type yeast and clear modification subpopulations within several genetic and environmental conditions that affect yeast ribosome biogenesis and pre-mRNA splicing.

1.7 Equations

$$w = N_1 = N_2 \tag{1.1}$$

$$\bar{X}_1 = \frac{1}{w} \sum_{i=j-w}^{j} x_i \tag{1.2}$$

$$\bar{X}_2 = \frac{1}{w} \sum_{i=j+1}^{j+w} x_i \tag{1.3}$$

$$\sigma^2 = E[\bar{x^2}] - E[\bar{x}^2] = \frac{\sum_{i=1}^{N} x_i^2 - (\sum_{i=1}^{N} x_i)^2/N}{N} \tag{1.4}$$

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}}} \tag{1.5}$$

$$a = \text{scale} \tag{1.6}$$

$$b = \text{shift} \tag{1.7}$$

$$c = \text{drift} \tag{1.8}$$

$$d = \text{variance} \tag{1.9}$$

$$t = \text{time} \tag{1.10}$$

$$\mu_k = \text{event mean} \tag{1.11}$$

$$\sigma_k = \text{event standard deviation} \tag{1.12}$$

$$t_i = \text{start time} \tag{1.13}$$

$$\text{normalized event standard deviation} = (d \cdot \sigma_k)^2 \tag{1.14}$$

$$\text{normalized event mean} = a \cdot \mu_k + b + c \cdot t_i \tag{1.15}$$

$$\sigma(x) = \frac{1}{1 + e^{-x}} \tag{1.16}$$

$$tanh(x) = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}} \tag{1.17}$$

$$f_t = \sigma(W_f[x_t, h_{t-1}] + b_f) \tag{1.18}$$

$$i_t = \sigma(W_i[x_t, h_{t-1}] + b_i) \tag{1.19}$$

$$c_t^* = tanh(W_c[x_t, h_{t-1}] + b_c) \tag{1.20}$$

$$o_t = \sigma(W_o[x_t, h_{t-1}] + b_o) \tag{1.21}$$

$$c_t = f_t \circ c_{t-1} + i_t \circ c_t^* \tag{1.22}$$

$$h_t = o_t \circ tanh(c_t) \tag{1.23}$$

$$L(y, \hat{y}) = -\frac{1}{N} \sum_{i=0}^{N} (y * log(\hat{y}_i) + (1 - y) log(1 - \hat{y}_i)) \tag{1.24}$$

1.8 Figures

Table 1 Epigenetic modifications in human diseases

Aberrant epigenetic mark	Alteration	Consequences	Examples of genes affected and/or resulting disease
Cancer			
DNA methylation	CpG island hypermethylation	Transcription repression	*MLH1* (colon, endometrium, stomach[11]), *BRCA1* (breast, ovary[11]), *MGMT* (several tumor types[11]), *p16^{INK4a}* (colon[11])
	CpG island hypomethylation	Transcription activation	*MASPIN* (pancreas[92]), *S100P* (pancreas[92]), *SNCG* (breast and ovary[92]), *MAGE* (melanomas[92])
	CpG island shore hypermethylation	Transcription repression	*HOXA2* (colon[20]),*GATA2* (colon[20])
	Repetitive sequences hypomethylation	Transposition, recombination genomic instability	*L1* (ref. 11), *IAP*[11], *Sat2* (ref. 107)
Histone modification	Loss of H3 and H4 acetylation	Transcription repression	*p21^{WAF1}* (also known as *CDKN1A*)[11]
	Loss of H3K4me3	Transcription repression	*HOX* genes
	Loss of H4K20me3	Loss of heterochromatic structure	*Sat2*, *D4Z4* (ref. 107)
	Gain of H3K9me and H3K27me3	Transcription repression	*CDKN2A*, *RASSF1* (refs. 115–116)
Nucleosome positioning	Silencing and/or mutation of remodeler subunits	Diverse, leading to oncogenic transformation	*BRG1*, *CHD5* (refs. 127–131)
	Aberrant recruitment of remodelers	Transcription repression	*PLM-RARa*[103] recruits NuRD
	Histone variants replacement	Diverse (promotion cell cycle/destabilization of chromosomal boundaries)	H2A.Z overexpression/loss
Neurological disorders			
DNA methylation	CpG island hypermethylation	Transcription repression	Alzheimer's disease (*NEP*)[135]
	CpG island hypomethylation	Transcription activation	Multiple sclerosis (*PADI2*)[135]
	Repetitive sequences aberrant methylation	Transposition, recombination genomic instability	ATRX syndrome (subtelomeric repeats)[135,143]
Histone modification	Aberrant acetylation	Diverse	Parkinson's and Huntington's diseases[135]
	Aberrant methylation	Diverse	Huntington's disease and Friedreich's ataxia[135]
	Aberrant phosphorylation	Diverse	Alzheimer's disease[135]
Nucleosome positioning	Misposition in trinucleotide repeats	Creation of a 'closed' chromatin domain	Congenital myotonic dystrophy[151]
Autoimmune diseases			
DNA methylation	CpG island hypermethylation	Transcription repression	Rheumatoid arthritis (*DR3*)[154,155]
	CpG island hypomethylation	Transcription activation	SLE (*PRF1*, *CD70*, *CD154*, *AIM2*)[6]
	Repetitive sequences aberrant methylation	Transposition, recombination genomic instability	ICF (*Sat2*, *Sat3*), rheumatoid arthritis (*L1*)[152,155]
Histone modification	Aberrant acetylation	Diverse	SLE (*CD154*, *IL10*, IFN-γ)[6]
	Aberrant methylation	Diverse	Diabetes type 1 (*CLTA4*, *IL6*)[159]
	Aberrant phosphorylation	Diverse	SLE (NF-κB targets)
Nucleosome positioning	SNPs in the 17q12-q21 region	Allele-specific differences in nucleosome distribution	Diabetes type 1 (*CLTA4*, *IL6*)
	Histone variants replacement	Interferes with proper remodeling	Rheumatoid arthritis (histone variant macroH2A at NF-κB targets)[157]

Figure 1.1: Table of diseases associated with epigenetic modifications. All references to cited papers can be found in original paper where this figure was found [130].

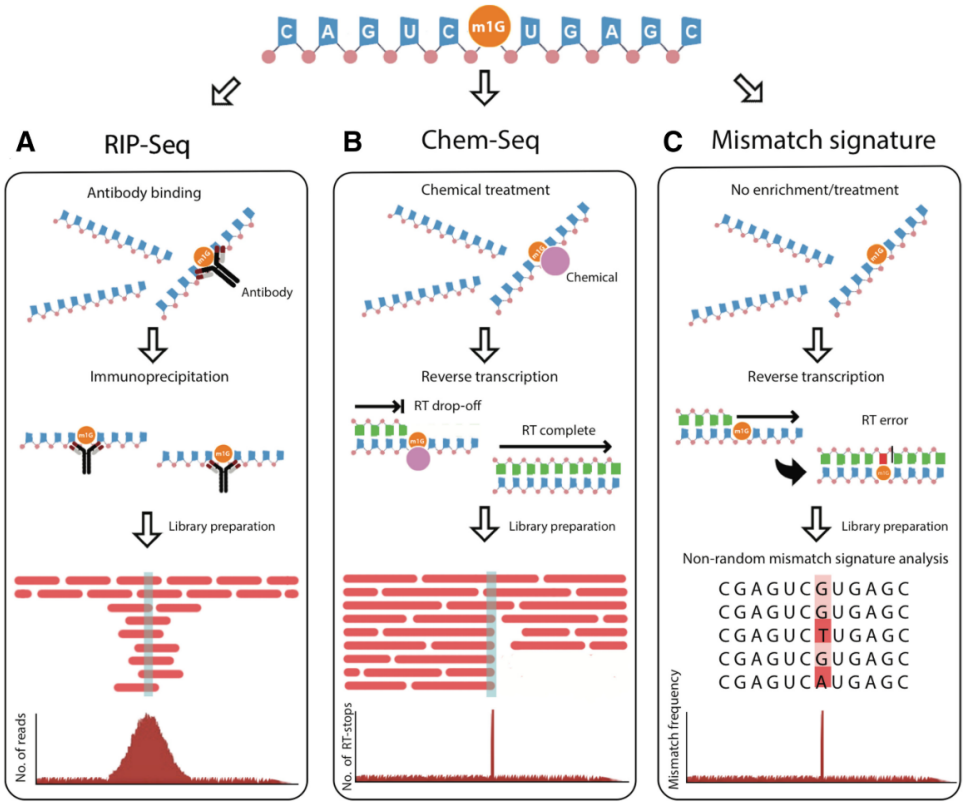

Figure 1.2: (Original Caption) Current genome-wide detection methods used to identify RNA modifications. (A) In the left panel, antibody-based methods (RIP-seq) show how RNA-modification enriched fragments are selected using pool-down, and compared to a total fragmented sample (input), which is used for normalization, obtaining genome-wide maps with peak resolution. (B) In the middle panel, RNA samples are pretreated with chemical reagents (Chem-seq), which inhibit the reverse transcription reaction beyond the chemically modified position. (C) In the right panel, mismatch signature-based methods, which are based on the increased mismatch rates that occur upon reverse transcription at certain RNA-modified positions, are depicted. [74]

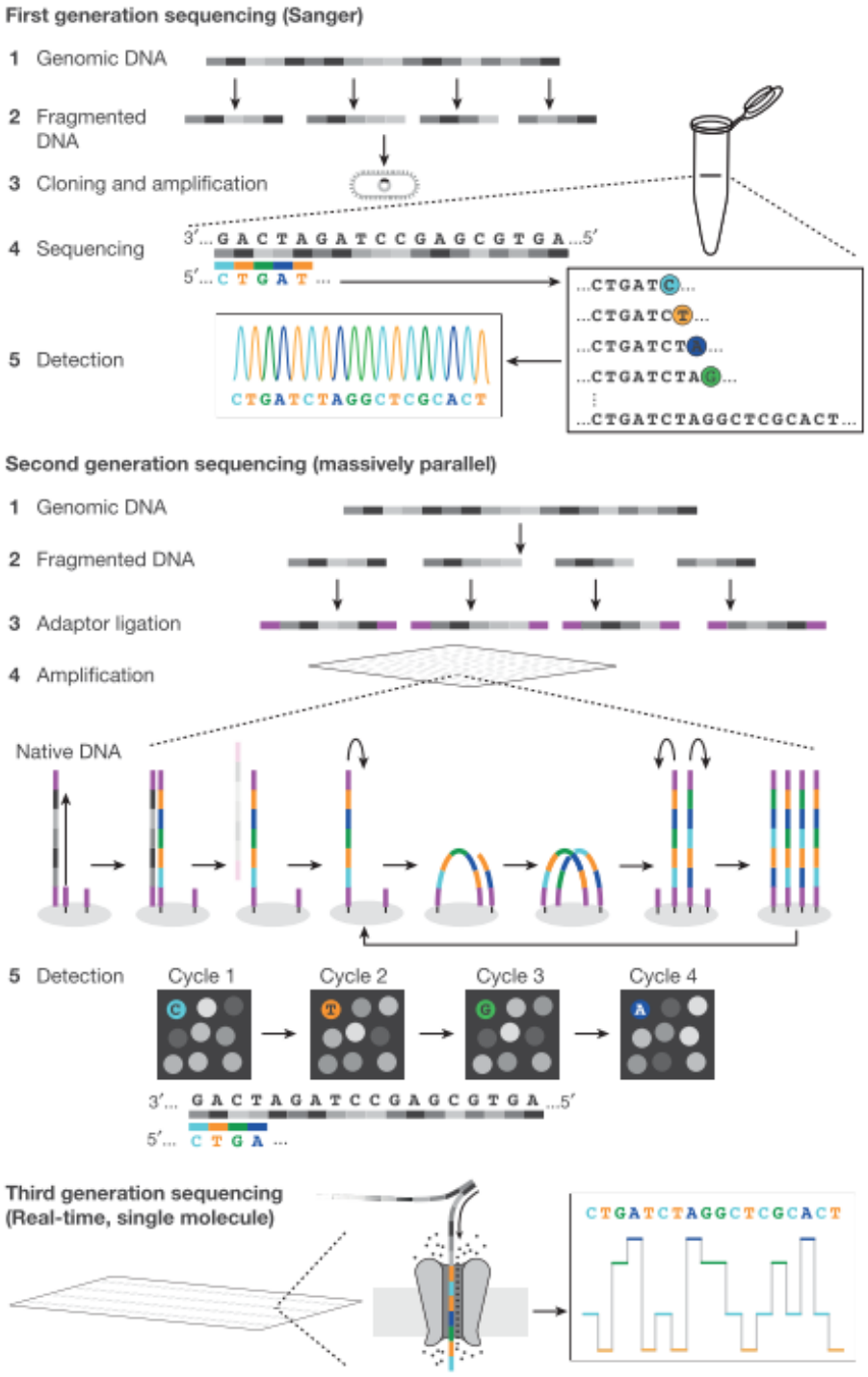

Figure 1.3: (Original Caption) Schematic examples of first, second and third generation sequencing are shown. Second generation sequencing is also referred to as next-generation sequencing (NGS) in the text.[150]

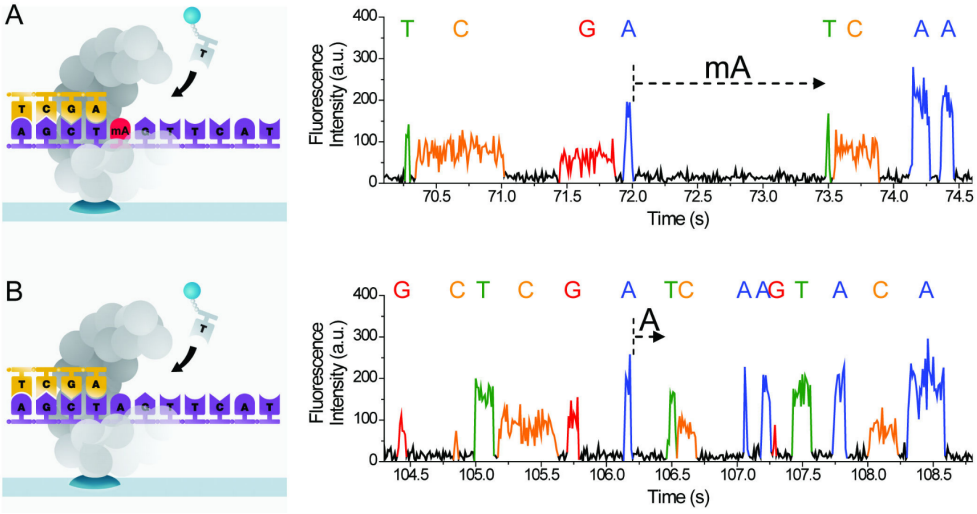

Figure 1.4: (Original Caption) Principle and corresponding example of detecting DNA methylation during SMRT sequencing. (a) Schematics of polymerase synthesis of DNA strands containing a methylated (top) or unmethylated (bottom) adenosine. (b) Typical SMRT sequencing fluorescence traces from these templates. Letters above the fluorescence trace pulses indicate the identity of the nucleotide incorporated into the growing complementary strand. The dashed arrows indicate the IPD before incorporation of the cognate T, and, for this typical example, the IPD is ∼5x larger for mA in the template compared to A. [51]

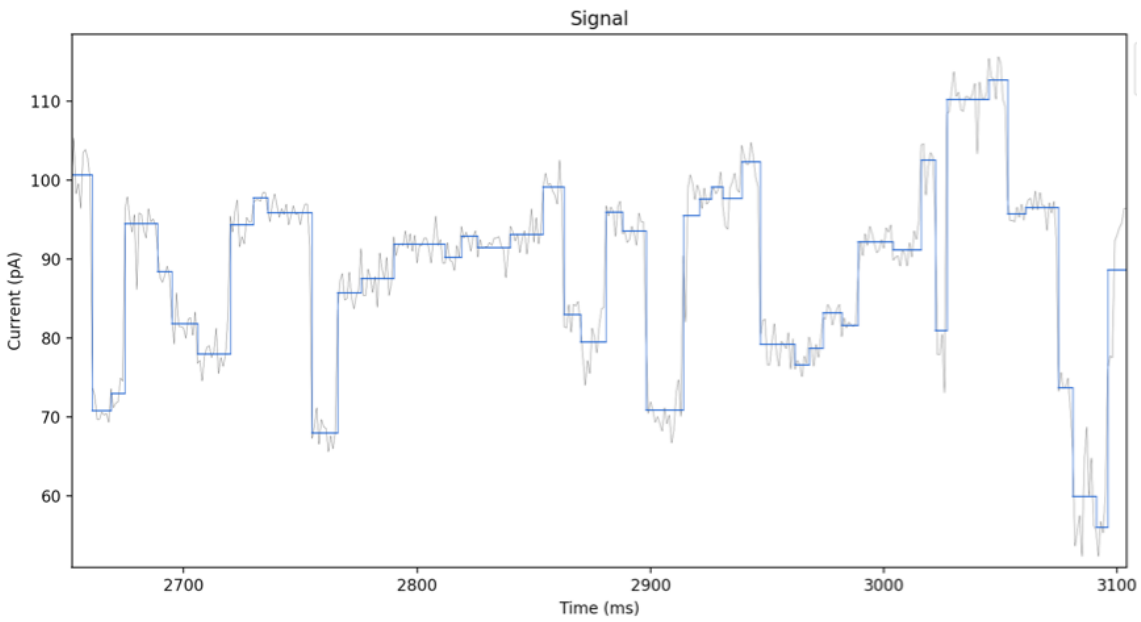

Figure 1.5: The grey line is a standard ONT nanopore sequencing DNA signal trace and the blue lines denote the events detected using the standard t-test event detection algorithm.

Methylase	Known Meth. Site	Average Depth	Methylase Class	Motifs in Genome	Detection AUC	1D Accuracy
TaqI	TCG**A**	22	6mA	30914	0.82	87.8
BamHI	GGAT**C**C	36	4mC	988	0.84	87.9
EcoRI	GA**A**TTC	27	6mA	1290	0.88	87.8
HhaI	G**C**GC	50	5mC	65566	0.97	87.3
MpeI	**C**G	39	5mC	693340	0.62	86.4
SssI	**C**G	19	5mC	693340	0.78	83.9
dam	G**A**TC	33	6mA	38240	0.66	89.6

Figure 1.6: (Original Caption) Tested methylases with known recognition site (methylated base underlined), depth of sequencing, methylation class, and other sample statistics.[108]

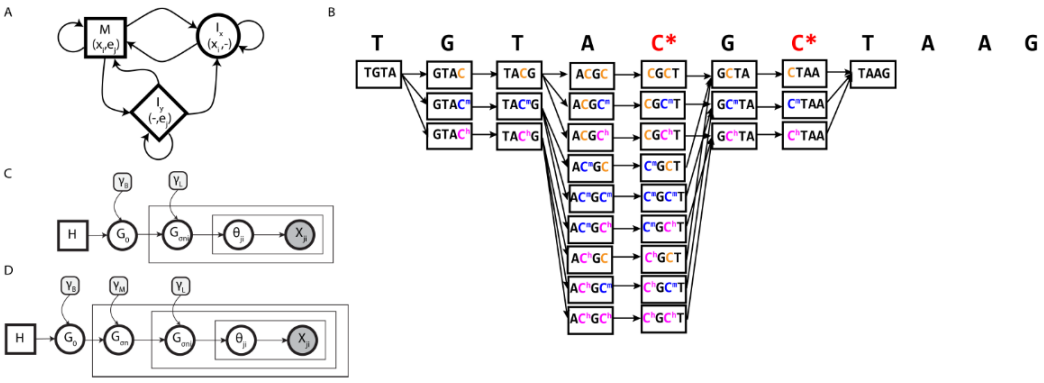

Figure 1.7: (Original Caption) Overview of models. A. Architecture of hidden Markov model used in this study. The match state, M (square), emits an event-k-mer pair and proceeds along the reference and the event sequence, Insert-Y, I_y (diamond), emits a pair and proceeds along the event sequence but stays in place with respect to the reference, and Insert-X, I_x (circle), proceeds along the reference but does not emit a pair and stays in place with respect to the event sequence. B. Variable-order HMM meta-structure over an example reference sequence containing ambiguous methylation variants. Each C* in the reference represents a potentially methylated cytosine. The structure expands around the C* to accommodate all possible methylation states (in this case, C, 5-mC, and 5-hmC). Each cell contains the three states shown in A, and transitions span between cells. The transitions are restricted so that methylation states are labeled consistently within a path. The match states are drawn with 4-mers for simplicity, but the model is implemented with 5-mers and 6-mers. Two-level (C) and three-level (D) hierarchical Dirichlet process shown in graphical form. Circles represent random variables. The base distribution H is a normal inverse-gamma distribution for both models. The Dirichlet processes G_0, $G_{\sigma n}$, and $G_{\sigma ni}$ are parameterized by their parent distribution and shared concentration parameters γB, γM, and γL. The factors Θ_{ji} specify the parameters of the normal distribution mixture component that generates observation x_{ji}.[135]

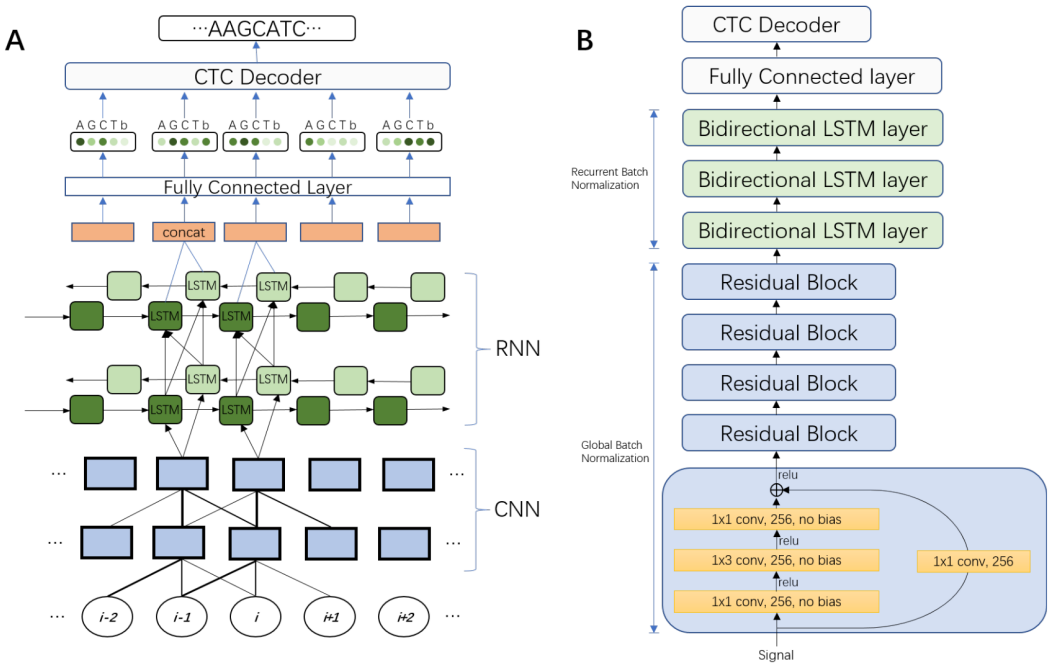

Figure 1.8: (Original Caption) A) An unrolled sketch of the neural network architecture. The circles at the bottom represent the time series of raw signal input data. Local pattern information is then discriminated from this input by a CNN. The output of the CNN is then fed into a RNN to discern the long-range interaction information. A fully connected layer is used to get the base probability from the output of the RNN. These probabilities are then used by a CTC decoder to create the nucleotide sequence. The repeated component is omitted. B) Final architecture of the Chiron model. We explored variants of this architecture by varying the number of convolutional layers from 3 to 10 and recurrent layers from 3 to 5. We also explored networks with only convolutional layers or recurrent layers, 1×3 conv, 256, no bias means a convolution operation with a 1×3 filter and a 256 channels output with no bias added. [166]

Chapter 2

Gaussian Mixture Model-Based Unsupervised Nucleotide Modification Number Detection Using Nanopore Sequencing Readouts

Hongxu Ding[1,2,*], Andrew D. Bailey IV [1,2] Miten Jain[1], Hugh Olsen[1] and Benedict Paten[1,*]

[1]Department of Biomolecular Engineering and Genomics Institute, University of California, Santa Cruz, Santa Cruz, California, USA. [2]These authors con-

tributed equally to this work. *Correspondence should be addressed to H.D. (hding16@ucsc.edu) or B.P. (benedict@soe.ucsc.edu).

2.1 Abstract

We present a framework for the unsupervised determination of the number of nucleotide modifications from nanopore sequencing readouts. We demonstrate the approach can effectively recapitulate the number of modifications, the corresponding ionic current signal levels, as well as mixing proportions under both DNA and RNA contexts. We further show, by integrating information from multiple detected modification regions, that the modification status of DNA and RNA molecules can be inferred. This method forms a key step of de novo characterization of nucleotide modifications, shedding light on the interpretation of various biological questions.

2.2 Introduction

Modified nucleotides play critical roles in diverse biological processes [94, 103]. Oxford Nanopore Technologies (ONT) nanopore sequencing monitors ionic current signal shifts caused by various chemical structures of the nucleotides [35], which opens up the possibility of routinely identifying DNA/RNA modifications [80]. Up to now, modification calling softwares has been shown to identify 6mA [99, 135, 115], 5mC [99, 135, 115, 152], 5hmC [115] as well as the thymidine analogs EdU, FdU, BrdU, IdU [112] in DNA, and 6mA [184], inosine (I) [184], 7-methylguanine (7mG)

[154], pseudouridine (Q) [154] in RNA. All of these softwares require some models of the expected signals for given modifications. For instance, nanopolish [152, 101], signalAlign [135] and DNAscent [112] perform modification calling based on a priori kmer models, which keep track of ionic current signals associated with all native and modified kmers. DeepMod [99] and DeepSignal [115] are deep learning based modification detection algorithms, which identify modifications based on neural networks trained on control datasets. However, these algorithms can only analyze modifications appeared in labelled training data, thereby considered as supervised methodologies. Meanwhile, for unidentified modifications, potential sites can be inferred using unsupervised approaches, e.g. tombo [108] and nanocompore [89]. However these unsupervised modification techniques do not include more detailed characterizations, such as modification numbers, corresponding signal levels and proportions. Understanding the number of modifications under specific sequence contexts can provide critical biological insights. For instance, during DNA demethylation, 5mC is sequentially converted into 5hmC, 5-flucytosine (5fC), 5-carboxylcytosine (5caC) and finally C. Therefore the number and corresponding proportion of modifications would be indicator for DNA demethylation dynamics [13]. Meanwhile, from a technological perspective, understanding the number of modifications is a crucial part of de novo modification characterization, which is considered as one of the most important topics in the nanopore sequencing community.

2.3 Materials and Methods

2.3.1 Data collection and preprocessing.

Nanopore sequencing datasets included here were composed of fast5 files, which contain raw ionic current readouts from the sequencer, together with fastq files, which contain sequences basecalled from corresponding fast5 records. The fast5 and fastq files are considered to be the "raw data" to be collected and preprocessed. Specifically, in cases where fastq files were embedded in fast5 records, nanopolish extract (0.11.1) [101], followed by porechop demultiplexing (0.2.4) [180] were used to recover the fastq files. We used a Zymo native synthesized oligo nanopore sequencing dataset which was provided by authors of the original study [135]. We also used a thymidine analogs-containing primer extension and native yeast genomic DNA nanopore sequencing datasets which are available at GEO with accession number GSE121941 [112]. Specifically, for the thymidine analogs-containing primer extension dataset, EdU, FdU, BrdU and IdU were incorporated in the synthesized "head" oligo (GAATTGGGCCCGCTCAGCAGACACAGAGCCTGAGCATCGCCGCGGAC, underscore denotes positions where thymidine analogs were incorporated). For a specific read, incorporated thymidine analog bases of the two positions are the same. And the portions of EdU, FdU, BrdU, IdU and T were the same. Then primer extension was performed, adding different extended "tail" sequences to different modifications, such that these reads with different modifications can be separated by alignment. Our RNA control dataset is a NA12878 cell line mRNA dataset (UCSC Run1 of Oxford

Nanopore Human Reference Datasets) is available at: `https://github.com/nanopore-wgs-consortium/NA12878/tree/master/nanopore-human-transcriptome` [184]. The RNA modification dataset is a E.coli 16S rRNA knockdown experiment provided by authors of the original study [154]. Three sub-datasets were sequenced in this study, containing reads from native, pseudouridine-deficient (Psi516) and m7G-deficient (m7G) strains. For m7G strain, m7G at position 527 is substituted with G, while for Psi516 strain, Q at position 516 is substituted with U [154]. For the m7G and Psi516 strains, mutations only affect m7G at position 527 and Q at position 516, respectively, and such mutation will cause 100% of the reads to be aberrantly modified.

2.3.2 Alignment, quality filtering and event table generation.

For the Zymo native synthesized oligo, thymidine analogs-containing primer extension, native yeast genomic DNA and NA12878 cell line mRNA nanopore sequencing datasets, in total 38685, 3173426, 121266 and 1291028 reads were obtained. Such reads in fastq files were first indexed using nanopolish index (0.11.1) [101], to establish one-to-one correspondence between sequences and ionic current records. The indexed fastq files were then aligned using minimap2 (2.16-r922) [92], followed by samtools view, sort and index (1.9) [93], yielding sorted and indexed bam files. Specifically, without loss of generality, for yeast genomic DNA and NA12878 cell line mRNA datasets, only reads mapped to the first chromosome were used for downstream analysis. During the alignment, for ZYMO, primer extension, yeast genomic

DNA and NA12878 cell line mRNA datasets, 35280, 17216, 117970 and 1269076 reads were aligned, respectively. After alignment, reads with MAPQ score equal to 60 and without secondary and supplementary alignments were kept for downstream analysis. Specifically, for the thymidine analogs-containing primer extension dataset, only reads mapped to the forward strand, where thymidine analogs reside, were kept. After such data filtering, for ZYMO, primer extension, yeast genomic DNA and NA12878 cell line mRNA datasets, 30241, 8450, 496 and 8640 reads were kept for downstream event level analysis, respectively. The event tables were generated using nanopolish eventalign, by taking fast5 files, bam files, and indexed fastq files as described above. Event tables contain kmer sequences and statistics of corresponding ionic current signals, e.g. mean and standard deviation values. Here, we modified nanopolish eventalign so that it can also output per read event tables containing the position of each kmer from the fastq sequence. We used these event tables to retrieve corresponding CIGAR strings and Q-scores. Quality control results were shown in Supplementary Figure 2.5, 2.6, 2.10 and 2.13-2.16. Specifically, filtered event tables for the 16S rRNA dataset were provided by authors of the original study, therefore the above-mentioned procedures were not applied.

2.3.3 Optimal kmer length determination

We first explored the kmer length that affects the signal (effective length). So, in Figure 2.1, we analyzed both the native yeast genomic DNA and the NA12878 cell line mRNA datasets. Kmers with various lengths (4-8 for DNA, 3-7 for RNA) were

generated based on the event tables (see previous section) and reference sequences. The event tables contain mapping positions of kmers, based on which sequences covering +2 to -2 positions (prolonged kmers) were retrieved from the references. These prolonged kmers were then trimmed centering around the original kmer. For instance, for native yeast genomic DNA read 001082a7-d27b-418c-85f6-a0297adb346b, the first signal event corresponded to ACGATT and mapped to position 11571, based on which the prolonged kmer was determined as ATACGATTGC. This prolonged kmer was further trimmed into, e.g. {ATACGATT, TACGATTG, ACGATTGC} for length=8, annotated by the corresponding trimming strategy as {8_2_0, 8_1_1, 8_0_2}. Since {ATACGATT, TACGATTG, ACGATTGC} were trimmed from the same signal event, they were corresponded to the same signal event level, in this case 71.89 pA. Following the same principle, we constructed other kmer trimming strategies including {4_2_0, 4_1_1, 4_0_2, 5_1_0, 5_0_1, 7_1_0, 7_0_1}. Such kmers, together with the above mentioned {8_2_0, 8_1_1, 8_0_2} and original kmer {6_0_0}, were all corresponded signal event level 71.89 pA. Then, for each trimming strategy, across all kmers included, we calculated the distribution of single event MAD (median absolute deviation). As described in the main text, such MAD distributions were used for determining optimal k for both DNA and RNA contexts.

2.3.4 Assessing the contributions of kmer positions to the ionic current shifts.

We then determined the effect of kmer positions on the signal, as shown in Supplementary Figure 2.8 and 2.9. We analyzed the same datasets from the previous section. Pairwise signal differences within kmer pth position quadruplet {Np-1ANk+1-p, Np-1TNk+1-p, Np-1GNk+1-p, Np-1CNk+1-p} were analyzed to assess the contribution of position p, where k equals 6 (DNA) or 5 (RNA), integer p ranges from 1 to k, Ns range in A, T, G, C and are identical for the same position, the subscripts indicates number of independently varying Ns. For instance, for DNA 6mer 1st position quadruplet {ATGCAT, TTGCAT, GTGCAT, CTGCAT}, 6 pairwise absolute value differences of kmer event signal medians (A-T, A-G, A-C, T-G, T-C, G-C) were calculated. Together with distance values generated from all other included DNA 6mer 1st position quadruplets, the contribution of 1st position can then be assessed. We then performed this analysis across all positions (1-6 for DNA, 1-5 for RNA) and used the distributions of absolute distance values as representations of kmer positional contributions. We further assessed the contribution of different nucleotides. For each nucleotide, e.g. A, at a given position p (Np-1ANk+1-p), we calculated the average pairwise distance of event signal medians from the corresponding 3 other nucleotides (Np-1TNk+1-p, Np-1GNk+1-p, Np-1CNk+1-p). The distributions of positional average signal shift for each nucleotide were presented as quantification of nucleotide-specific contributions.

2.3.5 Skewness and Kurtosis determination.

Skewness and kurtosis values were calculated using skewness() and kurtosis() functions in the CRAN R package e1071. As shown in Figure 2.2A-C, empirical signal distribution of kmers usually have long tails caused by outlier events, which will bias the determination of skewness and kurtosis. Therefore, for this specific analysis, we filtered out the following kmer event signal data points:

$s_i < medians - 3 * MAD_s$, or $s_i > medians + 3 * MAD_s$.

Subscript i denotes one specific signal event, and s denotes all corresponding events of a specific kmer.

2.3.6 Gaussian mixture model order determination.

The order (number of components) of Gaussian mixture models were determined by the statistical test reported in [25, 26], with the implementation of the emtest.norm() function in the CRAN R package MixtureInf. The statistical test was performed with the null hypothesis as order equals m0, against an alternative hypothesis where order equals 2m0. We search across various null hypotheses (m0 equals 1-9 and 1-4 for primer extension and rRNA datasets, respectively) for empirical kmer signal distributions, denoting the number of underlying Gaussian components of a certain empirical kmer signal distribution. To ensure correct inference, we used a more stringent filter to remove the following data points:

$s_i < medians - 2 * MAD_s$, or $s_i > medians + 2 * MAD_s$.

as outliers, before performing the fitting, considering they might account

for the "long tails" of the empirical kmer signal distribution, further introducing additional Gaussian components as artifacts for determining number of modifications. Subscript i denotes one specific signal event, and s denotes all corresponding events of a specific kmer. Elbow points on order-p-value curves were used to determine the number of components. P-values quantify significant levels of fitting performance gained by modeling with 2m0 as opposed to m0 components. Elbow points on the order-p-value curves denote marginal fitting performance gaining by including more components, therefore considered as optimal number of components. Following such principle, for both modified sites in the primer extension dataset, 7 was considered as the optimal number of components. By filtering out components whose proportions were less than 10%, for both sites 4 components remained, corresponding to T, FdU, EdU and BrdU-IdU containing kmers. Removed Gaussian components usually account for noises. For instance, the 1st, 2nd and 7th components of GCCTGA fitting were removed, and comparison between red (with all 7 components) and green (with remaining 4 components) curves in Figure 2.3A showed such filtering majorly affected the "tails" of the signal distribution. Actually signal levels of the removed Gaussian components were 88.897, 89.692 and 97.499, which were in the range of the "tails". For CATCGC (Figure 2.3E), the signal levels of the three removed Gaussian components were 110.597, 111.911 and 126.550, which were also in the range of the "tails". Specifically, BrdU- and IdU-containing kmers were considered as the same component due to close signal levels, which was further quantified by U-test (Supplementary Figure 2.17A-F). For both modified sites in the rRNA dataset, 2 was considered as the

optimal number of components, corresponding to the canonical and modified kmers (Supplementary Figure 2.17G and H). Reads were annotated based on their modification status in the original studies for both primer extension and rRNA datasets. Therefore for every analyzed modification site, we took the original annotation of reads covering this specific site, and calculated the mixing proportions of modified kmers. We further calculated the median values (after filtering by equation 3 and 4) of these modified kmers. Such proportion and median values were further used as gold standard in evaluating the performance of Gaussian mixture model.

2.3.7 Clustering nanopore sequencing reads.

Only nanopore sequencing reads covering all targeted positions (position 25-36 in the reference oligo sequence for primer extension dataset; position 511-515 and 522-526 in the reference transcript sequence for 16S rRNA dataset) were used for the analysis. Nanopore sequencing positional kmer signal events were then represented in read-position matrices, where reads in rows, targeted positions in columns and corresponding signals as elements. Clustering analysis was performed based on such read-position matrices.

2.3.8 Code availability.

Modified nanopolish is available at: `https://github.com/adbailey4/nanopolish/tree/cigar_output`. All other codes used to reproduce the results are available at: `https://github.com/hd2326/ModificationNumber`.

2.4 Results

2.4.1 Determining effective length for kmers

Shifts in ionic current (signal events) can be associated with nucleotide sequences (kmers) during their translocation through nanopores [152]. For multiple methods, characterizing such kmer-current relationships is essential to interpreting nanopore sequencing readouts. We first demonstrate that for our purpose an effective k for kmers (effective length for associating with the ionic current) equals 6 and 5 for DNA and RNA respectively, consistent with the information provided by Oxford Nanopore Technologies. To determine an effective k for our datasets we associated signal events to kmers of various lengths (4-8 for DNA, 3-7 for RNA). We chose a k that minimizes variation in current observations between different instances of the kmer while maximizing the numbers of distinct observations of each kmer. Specifically, for every kmer, we used the event signal fluctuation (quantified by median absolute deviation, MAD) as the criterion for determining the optimal k (see section 2.3). Here we analyzed a native yeast genomic DNA dataset [112] and one NA12878 cell line mRNA dataset [184] (see section 2.3), as examples for DNA and RNA scenarios, respectively. Genomic and transcriptomic sequences were used in order to make sure abundant sequence contexts could be included. As shown in Figure 2.1, the MAD ecdf (empirical cumulative distribution function) curve started to dramatically shift rightwards when k became smaller than 6 (DNA) or 5 (RNA). On the other hand, marginal differences were observed among MAD distributions when k exceeded

6 (DNA) or 5 (RNA). Taken together, these indicate the effective sequence length for shifting ionic current during nanopore sequencing equals 6 and 5 for DNA and RNA. We also quantified pairwise Kolmogorov-Smirnov d-values between the ecdf curves of different kmer constructing strategies, as confirmation of the effective kmer lengths (Supplementary Figure 2.7). We further assessed the contributions of kmer positions to the ionic current shifts by measuring the difference in signal among constructed sets of 4 kmers that are only different in 1 base at the examined position, e.g. ATGCAT, TTGCAT, GTGCAT, CTGCAT (see section 2.3). Results suggested for DNA 6mers, the 3rd position contributes the most, followed by the 4th position. The 2nd and 5th positions have minor contributions and the 1st and 6th positions have least contributions. For RNA 5mers, the 2nd position contributes the most, followed by the 3rd and 4th positions, and the 1st and 5th positions have least contributions (see Supplementary 2.8 and 2.9).

2.4.2 Empirical signal event distribution follows Gaussian

Gaussian has been widely used to model signal distribution. For instance kmer models provided by ONT, as well as several widely-acknowledged modification analysis algorithms [152, 112, 101, 108], assume the kmer signal distribution follows Gaussian. Here, we further demonstrated, using quantitative measurements, that the empirical distribution of nanopore sequencing kmer signal event means can generally be modeled by a normal distribution N(median, MAD). Specifically, we analyzed a Zymo synthesized oligo dataset [135] (see section 2.3), in order to make

sure sequenced nucleotide molecules were well-controlled. Median and MAD are calculated from all signal events of the corresponding kmer (Figure 2.2A). Compared to N(median, MAD), N(mean, standard deviation) fittings tend to be "widened" and "skewed" compared to the empirical distributions (Figure 2.2A). Such "widened" and "skewed" fittings can be explained by deviated means and increased standard deviations (Figure 2.2D and E), which are caused by "long tails" of kmer signal event empirical distributions. We argue such "long tails" are outlier kmer signal events well modeled by accounting for low sequencing quality and compromised alignment, rather than being due to the nature of an underlying kmer signal event distribution. We used the z-score computed from the kmer signal event median and MAD as a measurement of the likelihood of being an outlier. As shown in Figure 2.2B, C and Supplementary Figure 2.11, the likelihood of being an outlier is correlated with sequencing quality (quantified by Q-score) and affected by alignment status (quantified by the number of matches in the CIGAR string), indicating that the "long tails" are caused by outliers. Indeed most of the analyzed kmers can be well modeled by a normal distribution, suggested by absolute kurtosis and skewness (see section 2.3): as shown in Figure 2.2F, for 90.4% of analyzed kmers, such values ranged in the interval [0, 0.5].

2.4.3 Gaussian mixture model-based modification number inference

Considering that the signal event distribution for a given kmer can be reasonably modeled as normal, we can use a Gaussian mixture model to determine the

number of "isoforms" for a specific kmer. If there's no sequence variation, such as a single nucleotide variation (SNV), then we can consider such "isoforms" as different base modifications. The number of modifications correspond to the order (number of components) of the Gaussian mixture model, determined by the statistical test reported in [25] [26] (see section 2.3). As a proof of concept, we analyzed a thymidine analog DNA primer extension dataset reported in [152]. Thymidines in the sequence GAGCCTGAGCATCGCCG were substituted with EdU, FdU, BrdU or IdU, therefore we analyzed kmers GCCTGA and CATCGC (Figure 2.3A-D and E-H). For both kmers, 4 components were detected (Supplementary Figure 2.17A and D, see section 2.3), corresponding to T, FdU, EdU and BrdU-IdU containing kmers. BrdU and IdU were considered as one component by the Gaussian mixture model, due to the similar kmer event signal levels (Figure 2.3A and H, see section 2.3). As negative controls, we analyzed those non-modified sites, and 23 out of 26 sites were modeled by a single Gaussian component (Supplementary Figure 2.18). We further quantified the performance of a Gaussian mixture model in recapitulating signal event median and MAD values, as well as mixing proportion, for each kmer. As shown, median values were well recapitulated (Figure 2.3B and F); mixing proportions were in general recapitulated (Figure 2.3D and H); while inference on MAD values were unsatisfactory (Figure 2.3C and G). Such biases were caused by the "long tails" of the kmer signal event empirical distribution (Figure 2.3A and H), as previously discussed in Figure 2.2. Although such unsatisfactory performance on MAD inference can be considered as a limitation of the method, we argue MAD values are not very informative

for describing kmer signal events. As shown in Supplementary Figure 2.12 for kmer signal events >95% of the MAD values fall into the range of [1, 3], with no significant correlation with the corresponding median values. We speculate the variation of kmer signal events is largely caused by the noise associated with the nanopore sequencing platform itself, rather than an inherent characteristic of individual kmer signal events.

We further applied the Gaussian mixture model approach in analyzing RNA modifications. Specifically, we analyzed the dataset reported in [154], where E.coli 16S rRNAs from native, pseudouridine-deficient (Psi516) and m7G-deficient (m7G) strains were profiled. Compared to a native strain, in the Psi516 strain pseudouridine in UCCGUGCCA site is substituted with U, while in the m7G strain m7G in AGCCGCCGU site is substituted with G, therefore we analyzed kmers UGCCA and GCCGC. Following the same analytical pipeline as previously discussed for the DNA analysis (Supplementary Figure 2.17G and H, see section 2.3), we recapitulated the signal event median values, as well as the mixing proportions of the corresponding kmers (Supplementary Figure 2.19).

To further explore the detection limit of our approach, we performed down-sampling as well as remixing analysis. As a proof of concept, we focused on the RNA 5mer UGCCA and corresponding counterpart QGCCA (Q stands for pseudouridine), where we analyzed in Figure 2.3I. Specifically, we down-sampled to 100, 1000 and 2000 observations, at various QGCCA fractions, including 0.01, 0.05, 0.1, 0.25 and 0.5. We performed such down-sampling as well as remixing 10 times. It's clearly shown in Supplementary Figure 2.20 that with as few as 100 observations, our approach

can accurately recapitulate signal level and proportion of QGCCA component that accounts for 25% of total observations. If we have 2000 observations, such detection limits can further go down to only 1%. Taken together these suggested the high robustness and sensitivity of our approach.

2.4.4 Associating identified modifications

Now that we characterized the per sequence site modification pattern, the next question is how these modifications associate with each other. Therefore we then performed sequencing read-level analysis to assess the association, e.g. the co-occurrence, mutual-exclusiveness or independence, of the detected modifications. Reads covering all the modified regions were represented in read-position signal matrices, based on which hierarchical clustering was performed (see section 2.3). As shown in Figure 2.4A, for the primer extension dataset, four major clusters (Cluster2-5, account for 96% of total reads) were detected. We further quantified the composition of the four clusters, and as shown in Figure 2.4B and C, Cluster2-5 were majorly composed by T, FdU, EdU and Br/IdU reads, respectively. These results further suggested the expected co-occurrence of the T, FdU, EdU and Br/IdU (same modification at both T-sites), consistent with the experimental design. As shown in Figure 2.4D, for the 16S rRNA dataset, three major clusters (Cluster1-3) were detected, which were majorly composed of native, Psi516 and m7G reads, respectively. These results further suggested the mutual-exclusiveness of the U and G (in Psi516 strain pseudouridine is substituted with U, and in m7G strain m7G is substituted with G),

again consistent with the experimental design.

2.5 Discussion

Nanopore sequencing has the potential to detect every canonical and modified nucleotide accurately. Without improved de novo detection techniques, progress in modification detection will be dependent upon generating accurate labelled datasets for every modification. Currently, there are over 40 known DNA modifications [155] and over 150 known RNA modifications [15]. Also considering there might be modifications that have never been identified, generating labelled training datasets would be extremely challenging. Therefore, there is a pressing need for a de novo modification analysis pipeline. Such a pipeline can further be divided into three sequential steps, including de novo identification of modification sites, then de novo determination of modification numbers, and finally de novo inference on the corresponding chemical structures. The first step has been successfully implemented by Tombo [108] and Nanocompore [89], and our study focuses on the second step, providing a novel algorithm for the community. Specifically, we first confirmed the effective length of kmers for shifting ionic current signals during their translocation through nanopores equals 6 and 5 for DNA and RNA respectively. We then demonstrated the distributions of such kmer signals are mostly normal. A Gaussian mixture model can therefore be used for unsupervised modification number determination. Such a Gaussian mixture model-based approach can effectively recapitulate the number

of modifications, the corresponding kmer signal event median values, as well as the mixing proportions, in both DNA and RNA contexts. By integrating information from multiple regions, we further assessed the association between the corresponding modifications, which will shed light on modification status of DNA/RNA molecules, allowing for insights into various biological questions. Now that we can accurately determine number, signal levels and proportions of modifications, the next question is what are the corresponding chemical structures for each determined modification component. Answering this question would complete the pipeline for de novo modification analysis, which should be one future direction to pursue. One major limitation of the method, however, would be how to handle kmers with non-Gaussian signal distributions. For instance, as shown in Supplementary Figure 2.21, kmer TGATCC appeared in 3 different sequence contexts of the Zymo dataset [135], and in all cases a secondary peak was observed. Please note such secondary peaks were unlikely to be caused by quality issues, thus they cannot be removed by excluding low-quality reads. Such non-Gaussianity will introduce artifacts when analyzing modification numbers. We speculate the non-Gaussianity could be something related to the biophysical characteristics of the nanopores, which are largely unknown. Therefore, another future direction would be to find appropriate mixture models, for instance the extreme value mixture model as reported in [104], to model the modifications on non-Gaussian kmers.

2.6 Acknowledgements

Research reported in this publication was supported by the National Institutes of Health under Award Numbers U54HG007990, U01HL137183, 2U41HG007234 and 5R01HG010053. The research was also made possible by the generous financial support of the W.M. Keck Foundation (DT06172015). The authors would like to thank Dr. Mark Akeson's valuable comments on the study.

2.7 Author Contributions

H.D., A.B. and B.P. conceived the idea. H.D. and A.B. performed the analysis. M.J. and H.O. collected and pre-processed the data. B.P. supervised the project. H.D., A.B. and B.P. wrote the manuscript.

2.8 Figures

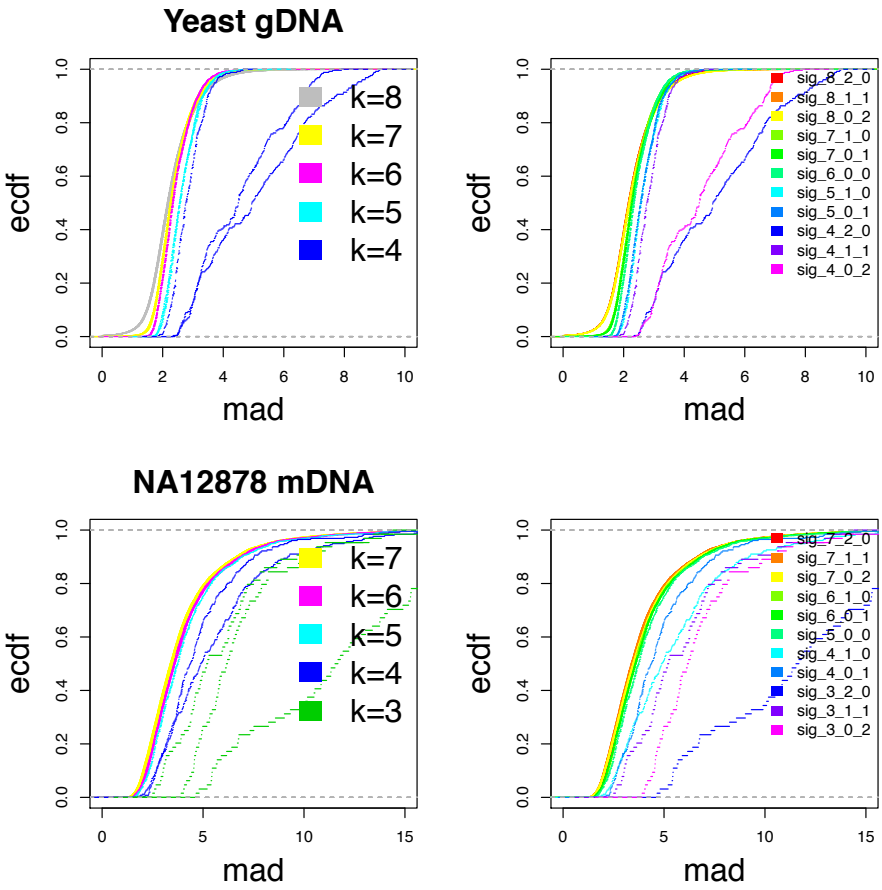

Figure 2.1: We analyzed native yeast genomic DNA and NA12878 cell line mRNA datasets, in both cases focusing on reads mapped to the first chromosome[135, 184]. Kmers with various lengths (4-8 for DNA, 3-7 for RNA) were generated based on the event tables and reference sequences. The event tables contain mapping positions of kmers, based on which sequences covering +2 to -2 positions (prolonged kmers) were retrieved from the references. Such prolonged kmers were then trimmed, centering around the original kmer, into desired lengths. For each kmer, we calculated the mad of signal events. For kmers with various lengths (4-8 for DNA, 3-7 for RNA) corresponding event signal mad (median absolute deviation) the ecdf (empirical cumulative distribution function) curve are shown. (A, B) Yeast genomic DNA and (C, D) NA12787 cell line mRNA datasets were analyzed as examples for DNA and RNA scenarios. The mad ecdf distributions as opposed to kmer lengths and constructing strategies are shown in (A, C) and (B, D).

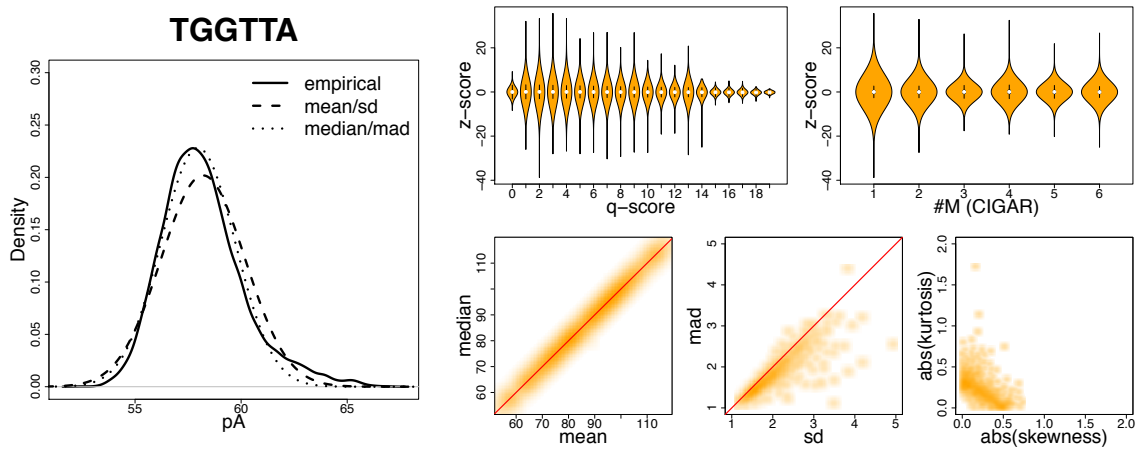

Figure 2.2: (A). Signal event distribution for an example 6-mer TGGTTA from the Zymo dataset[135]. Solid curve, empirical distribution; dashed curve, normal distribution fitted using mean and standard deviation (sd) of signal event; dotted curve, normal distribution fitted using median and median absolute deviation (mad) of signal event. (B) Violin plot showing z-score distribution under different q-score categories. Z-scores were computed using median and mad of signal events. (C) Violin plot showing z-score distribution under different CIGAR-string categories. Z-scores were computed using median and mad of signal events. #M denotes number of matches in CIGAR strings. (D, E) Smoothscatter plots showing signal event mean-median and sd-mad relationship of kmers. Red dashed line, slope equals 1. (F) Smoothscatter plot showing signal event empirical distribution skewness-kurtosis relationship of kmers.

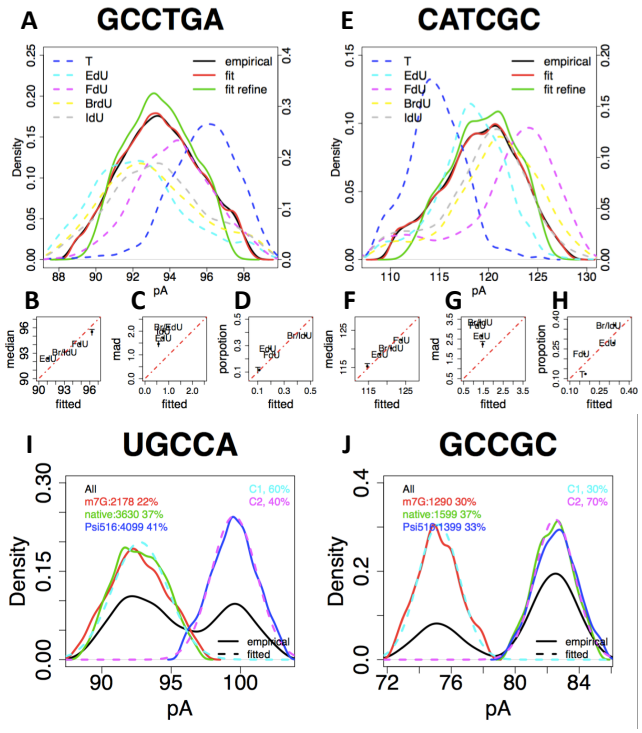

Figure 2.3: (A, B) Signal event distribution for the two modified kmers (GCCTGA and CATCGC) from the primer extension dataset [152]. Solid black curve, empirical distribution of all kmer signal events mapped to the specific position; solid red curve, fitted distribution with all Gaussian components of the mixture model; solid green curve, fitted distribution with Gaussian components that passed the mixing proportion threshold; dashed curves, empirical distribution of T (blue), EdU (cyan), FdU (purple), BrdU (yellow) and IdU (grey) kmer signal events. (A1, B1) -log10(p-value) of the fitting. #Components, numbers of Gaussian components as the null hypothesis (see section 2.3). Accepted null hypotheses were colored as red. (A2, B2) Mixing proportion of each Gaussian component. Removed components were colored as red. (A3, B3) The -log10(p-value) of a pairwise two-sided U-test among T, EdU, EdU BrdU and IdU kmer signal events. (A4-6, B4-6) Relationship between empirical and fitted kmer signal event medians values, kmer signal event mads and mixing proportions, respectively. Red dashed line, slope equals 1. (C, D) Signal event distribution for the two modified kmers (UGCCA and GCCGC) from the 16S rRNA dataset [115].

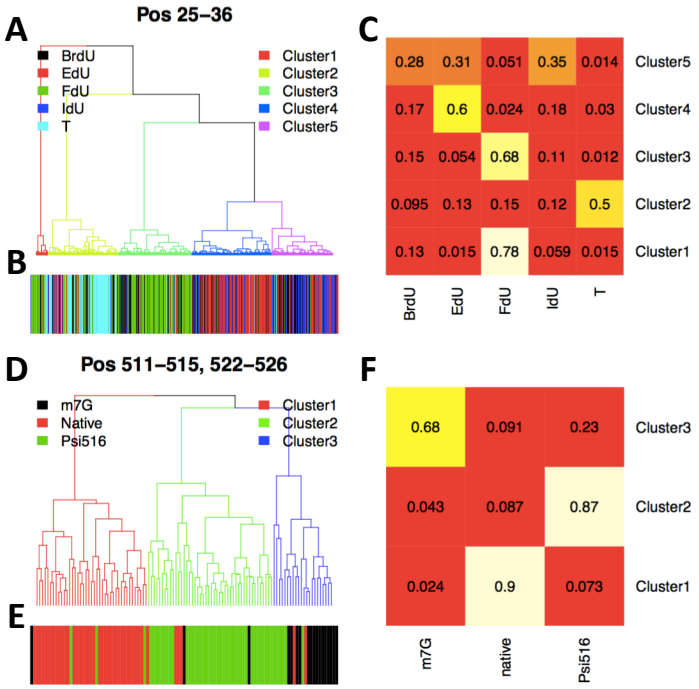

Figure 2.4: (A) Hierarchical clustering analysis on primer extension reads covering reference position 25-36 (see section 2.3). Branches of dendrogram were color-coded according to the cluster assignments. (B) Corresponding read annotation, including T- (cyan), IdU- (blue), FdU- (green), EdU- (red) and BrdU-containing reads (black). (C) Read composition of each cluster. (D) Hierarchical clustering analysis on 16S rRNA reads covering reference position 511-515 and 522-526 (see section 2.3). Branches of dendrogram were color-coded according to the cluster assignments. (B) Corresponding read annotation, including Psi516 (green), Native (red) and m7G reads (black). (C) Read composition of each cluster.

2.9 Supplementary Information

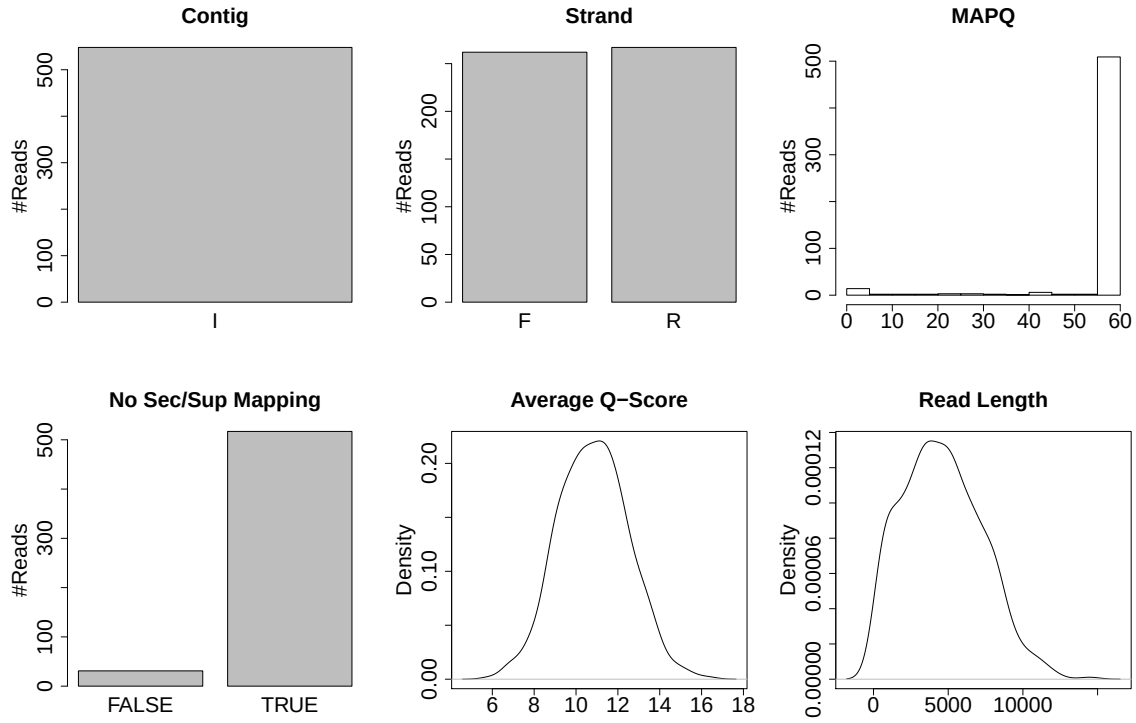

Figure 2.5: Quality control plots of the yeast genomic DNA dataset

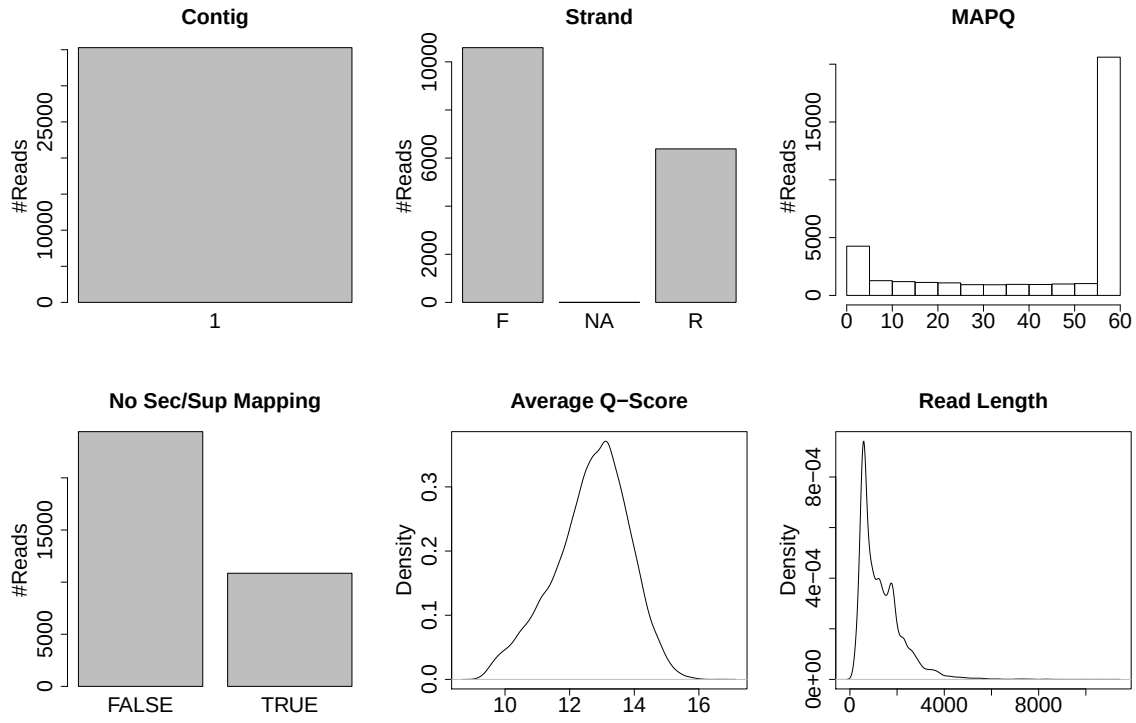

Figure 2.6: Quality control plots of the NA12787 cell line mRNA dataset.

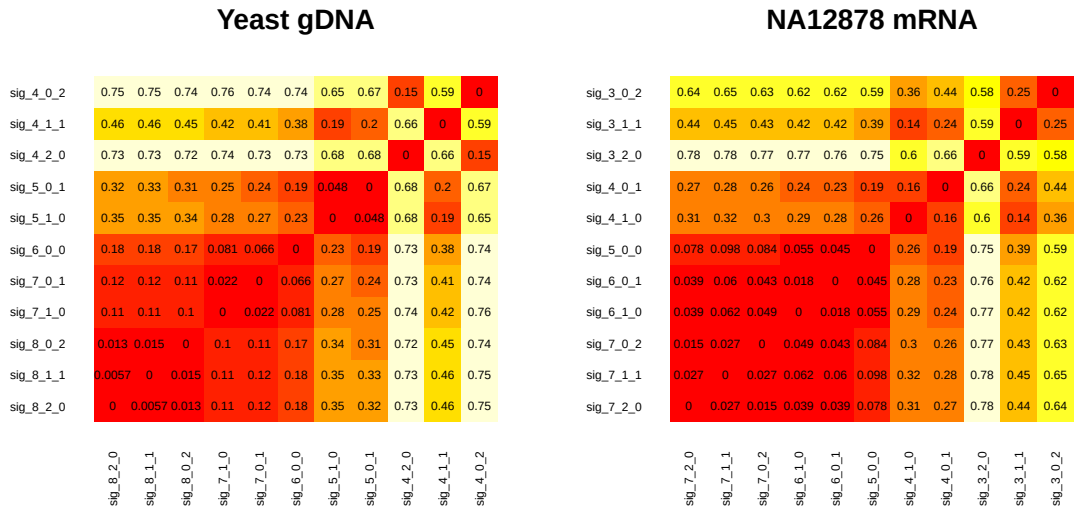

Figure 2.7: Pairwise Kolmogorov-Smirnov d-values between the ecdf curves of different kmer constructing strategies.

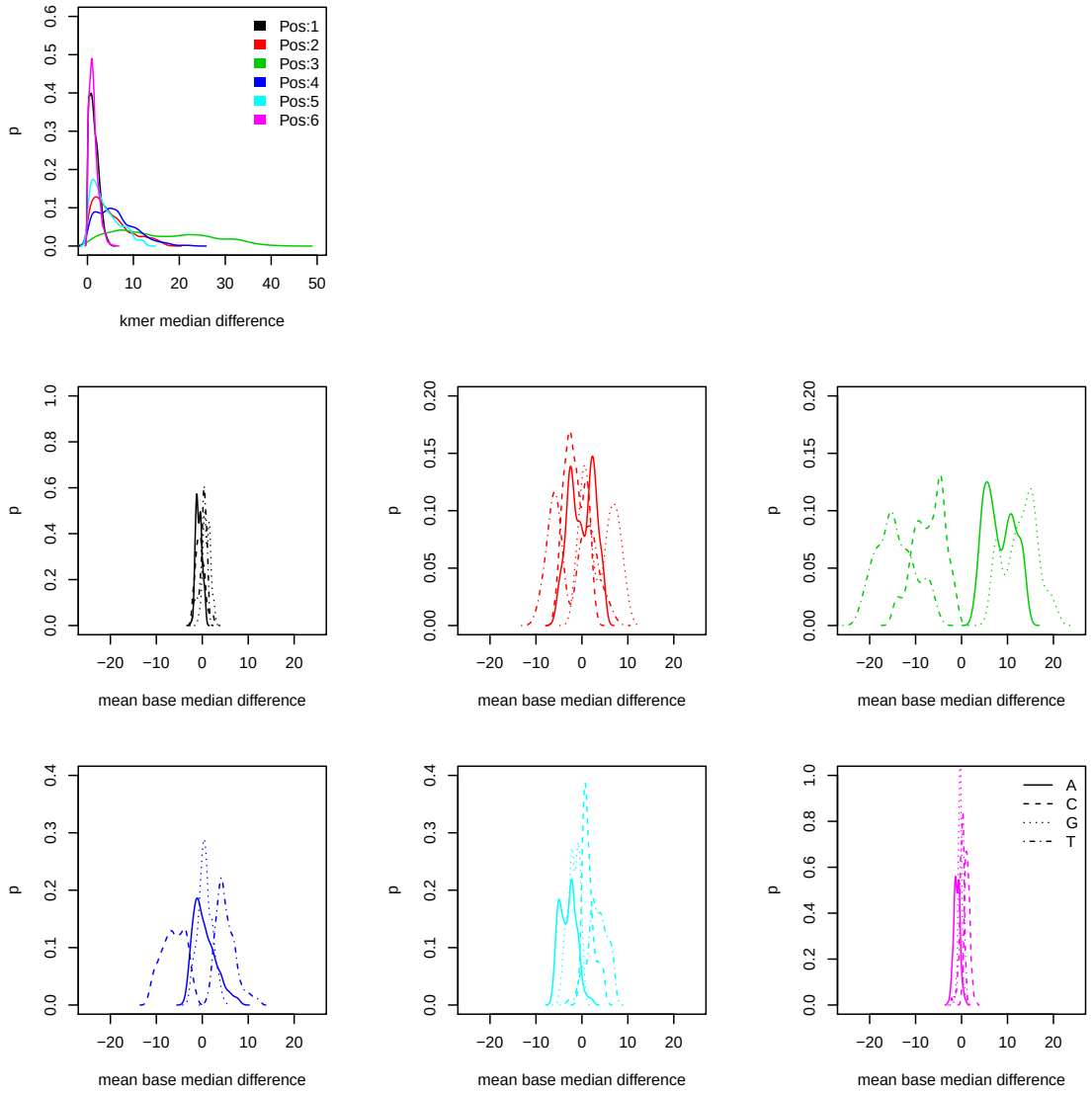

Figure 2.8: Assessing the contributions of DNA 6mer positions to the ionic current shifts. (A) Positional contribution. For every 6mer, median of all corresponding events were considered as 6mer-specific event signal level, as described in Figure 3.7A. 6mers that are different only at the examined position were collected into the same group. For every group, the 6 pairwise absolute value differences (A-T, A-G, A-C, T-G, T-C, G-C) were measured. Density distribution of such differences across groups was then visualized (see section 2.3). (B-G) Nucleotide-specific contribution of position 1-6. Same as in (A), for every 6mer, median of all corresponding events were considered as 6mer-specific event signal level, and 6mers that are different only at the examined position were collected into the same group. Then, for each nucleotide, e.g. A, the average pairwise distance of event signal level from the corresponding 3 other nucleotides, e.g. T, G and C, were calculated. Density distribution of such differences across groups was then visualized (see section 2.3).

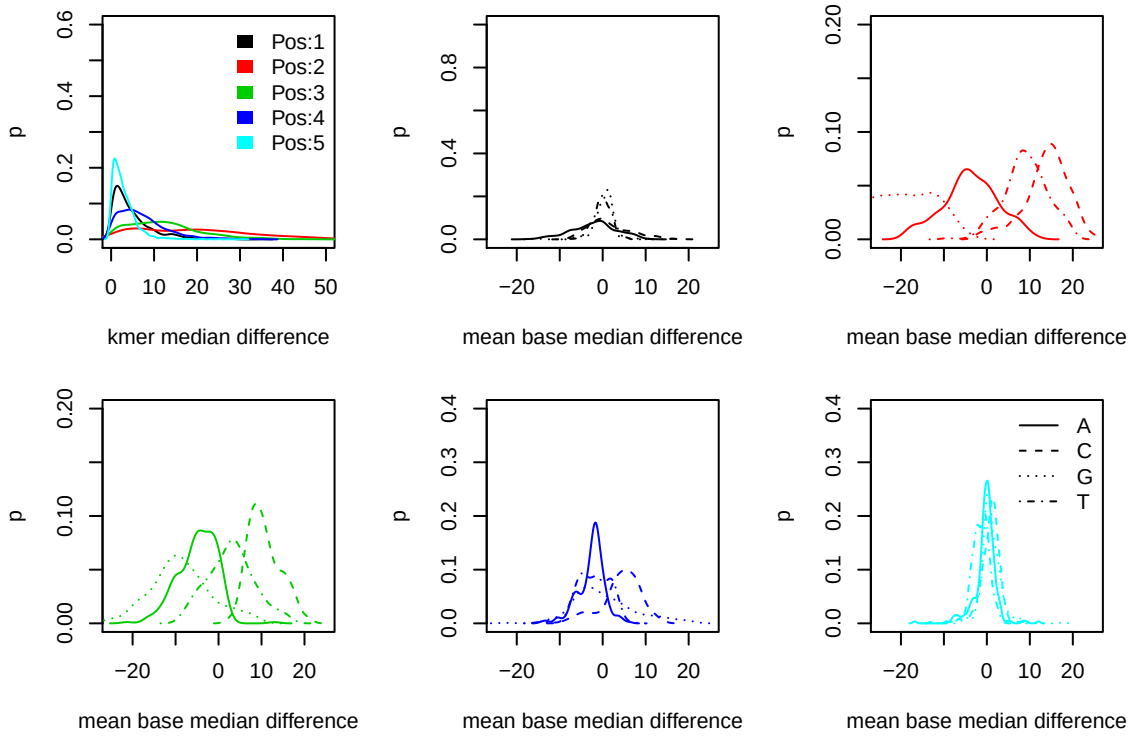

Figure 2.9: Assessing the contributions of RNA 5mer positions to the ionic current shifts. (A) Positional contribution. (B-F) Nucleotide-specific contribution of position 1-5. Same as Supplementary Figure 2.8, but in RNA context.

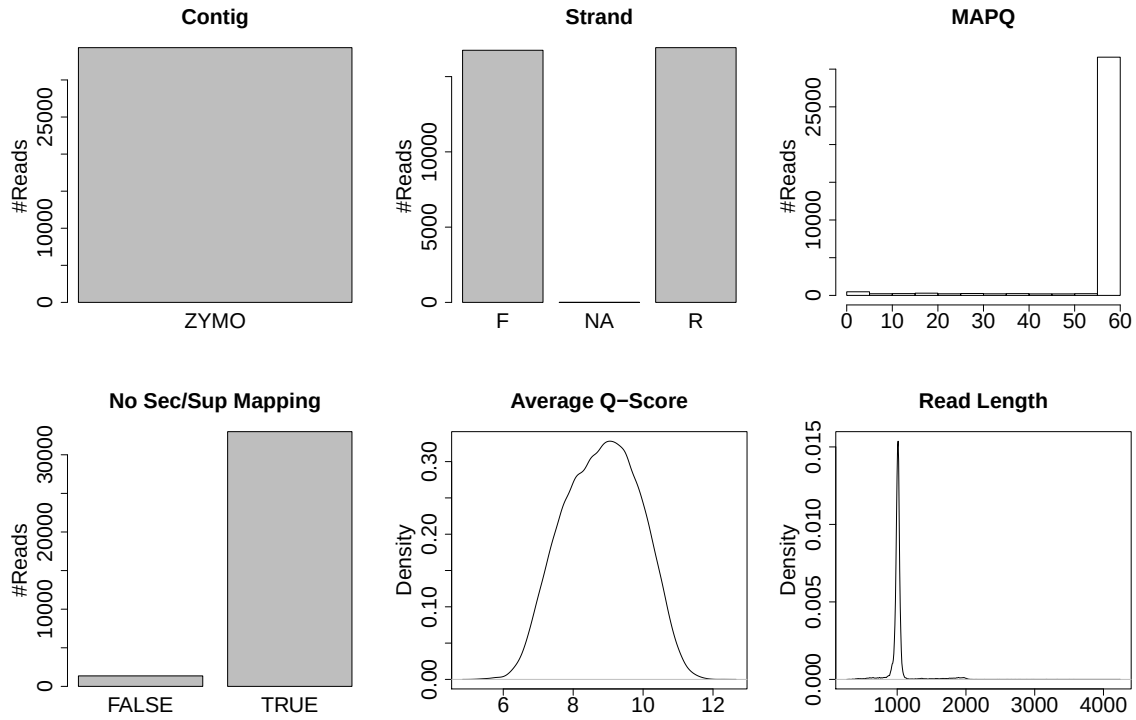

Figure 2.10: Quality control plots of the Zymo dataset.

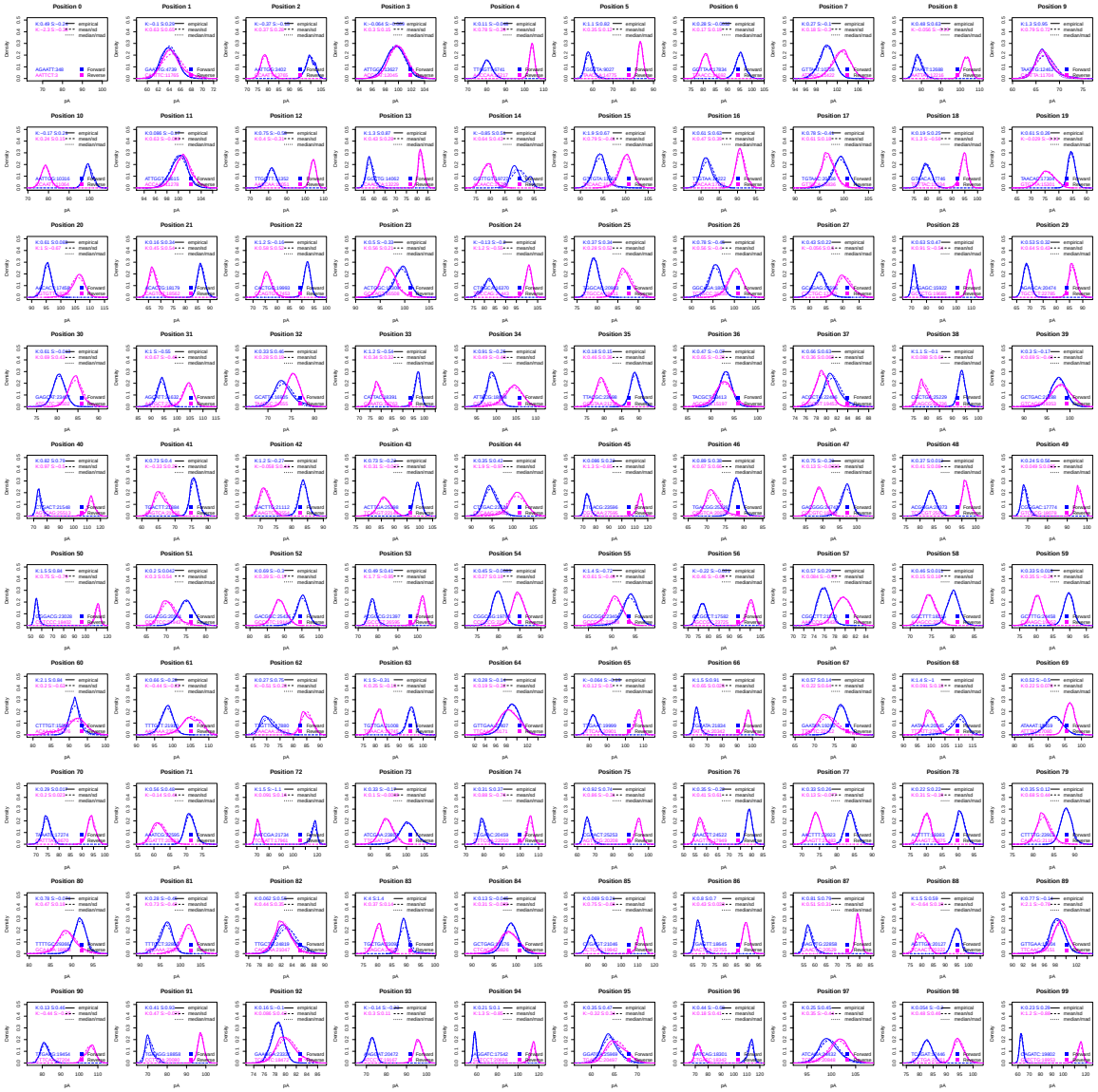

Figure 2.11: Basic statistics of kmer signal event distribution. Same as Figure 3.1B-F, but visualized in a strand-specific way. F, forward strand (blue); R, reverse strand (purple).

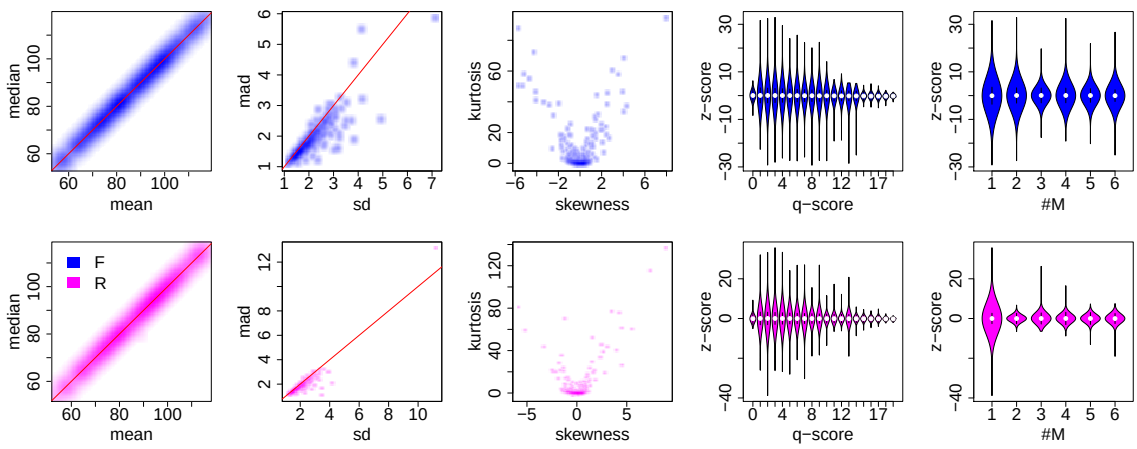

Figure 2.12: Signal event median-mad relationship. F, forward strand (blue); R, reverse strand (purple).

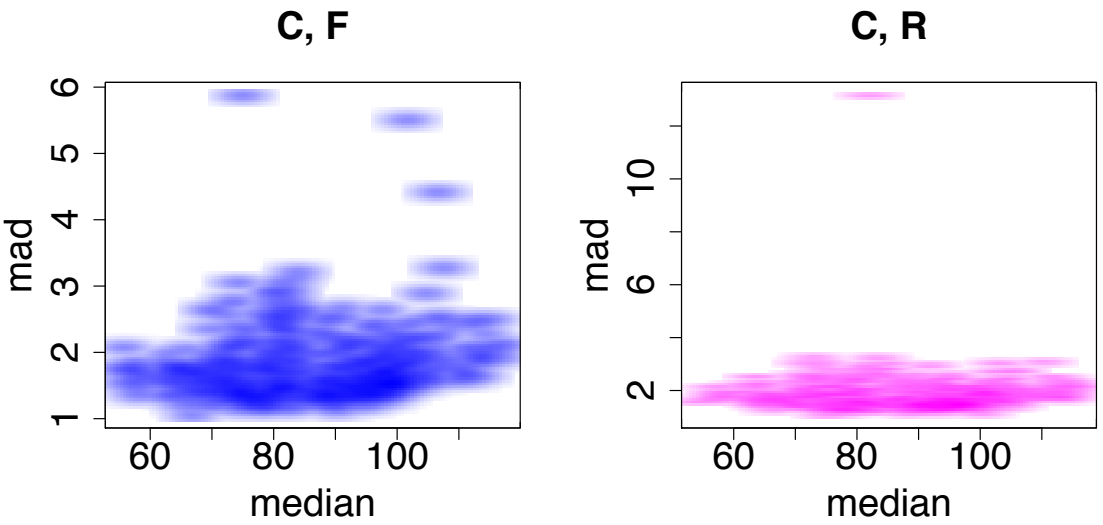

Figure 2.13: Quality control plots of the primer extension dataset.

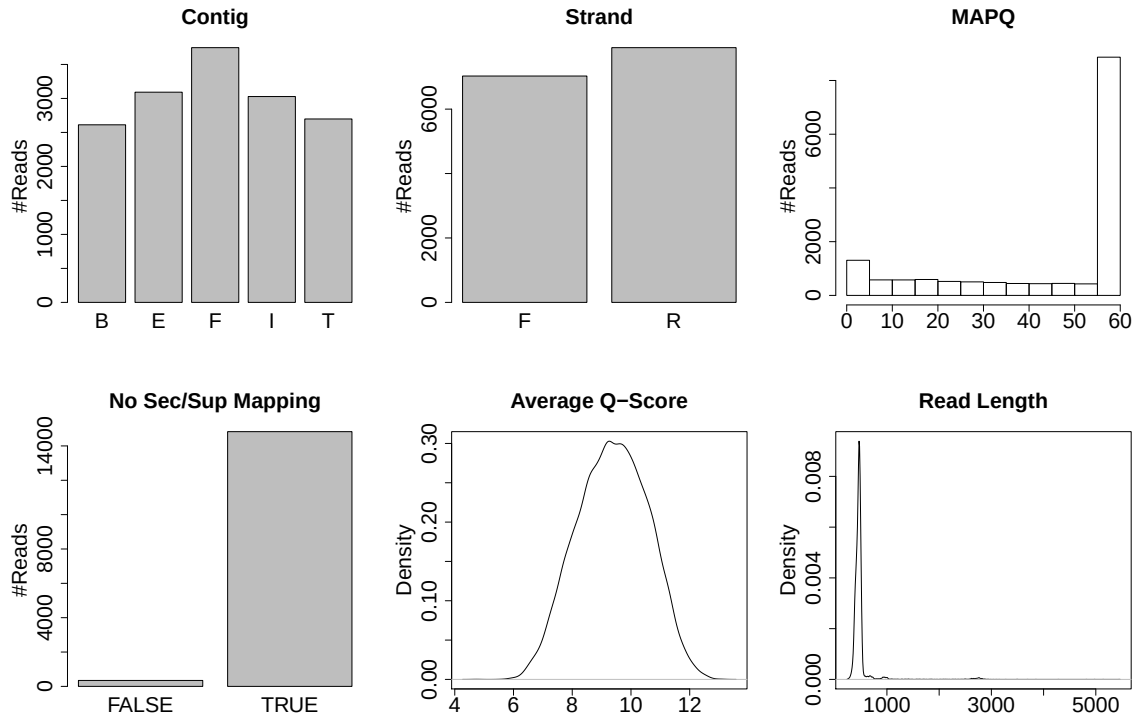

Figure 2.14: Quality control plots of the pseudouridine-deficient 16S rRNA dataset.

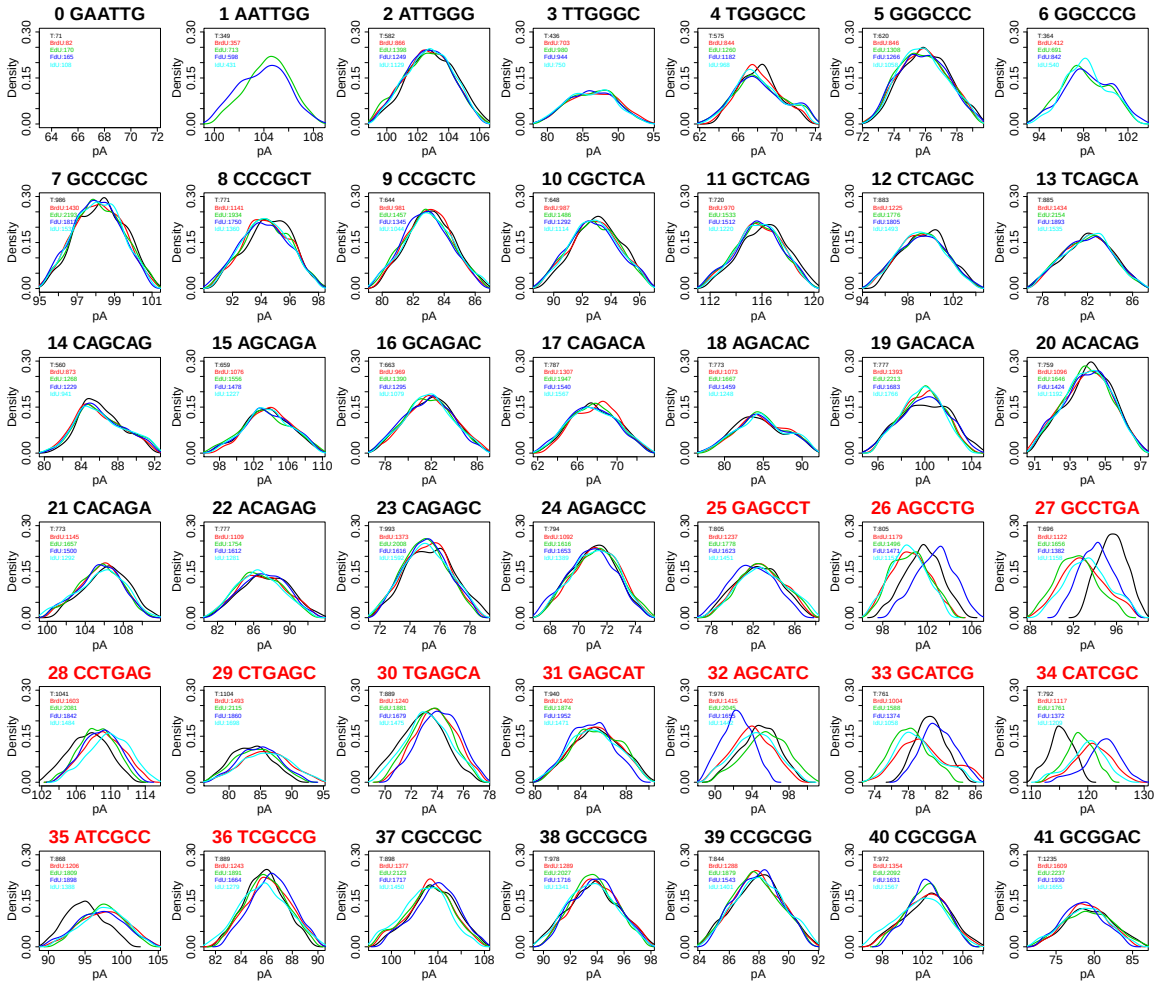

Figure 2.15: Quality control plots of the native 16S rRNA dataset.

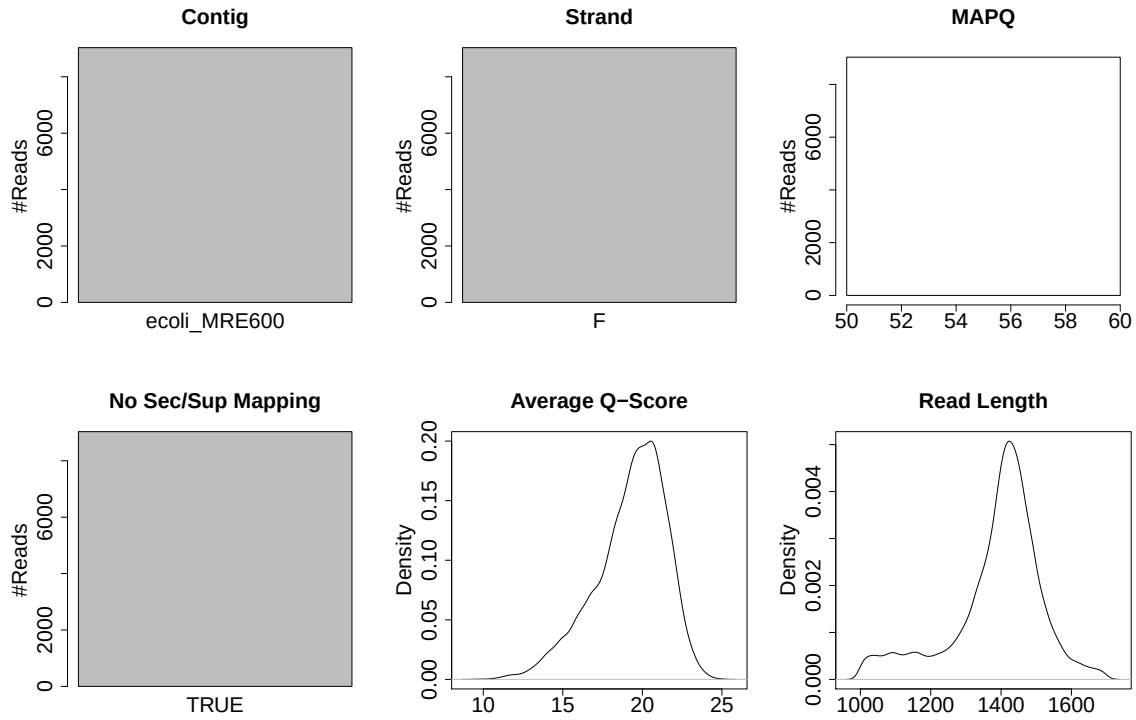

Figure 2.16: Quality control plots of the m7G-deficient 16S rRNA dataset.

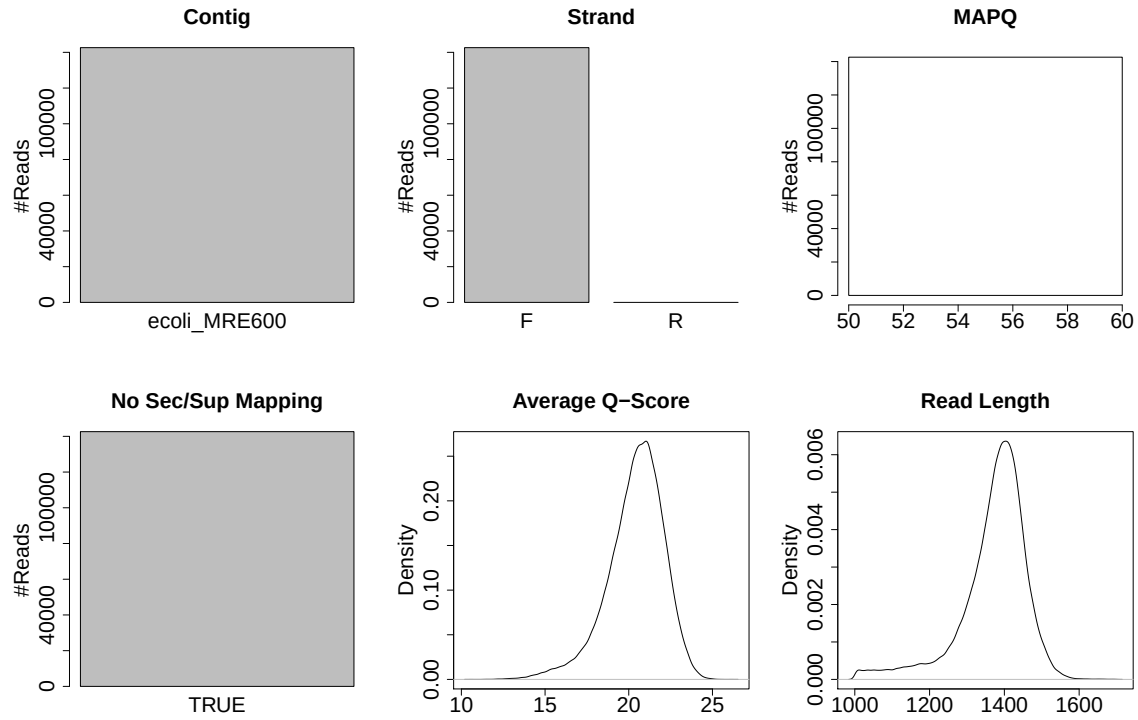

Figure 2.17: Determining optimal number of Gaussian mixture components. (A, D) Order-p-value curves for the two modification sites in the primer extension dataset. For both sites, 7 (marked as red) were considered as the optimal number. (B, E) Proportion of each predicted Gaussian component. Components that were less than 10% were filtered out (marked as red). (C, F) BrdU- and IdU-containing kmers were considered as the same component due to close signal levels, quantified by pairwise u-test. As shown, for both sites, BrdU-IdU pair gave the highest p-value. (G, H) Order-p-value curves for the two modification sites in the rRNA dataset. For both sites, 4 (marked as red) were considered as the optimal number.

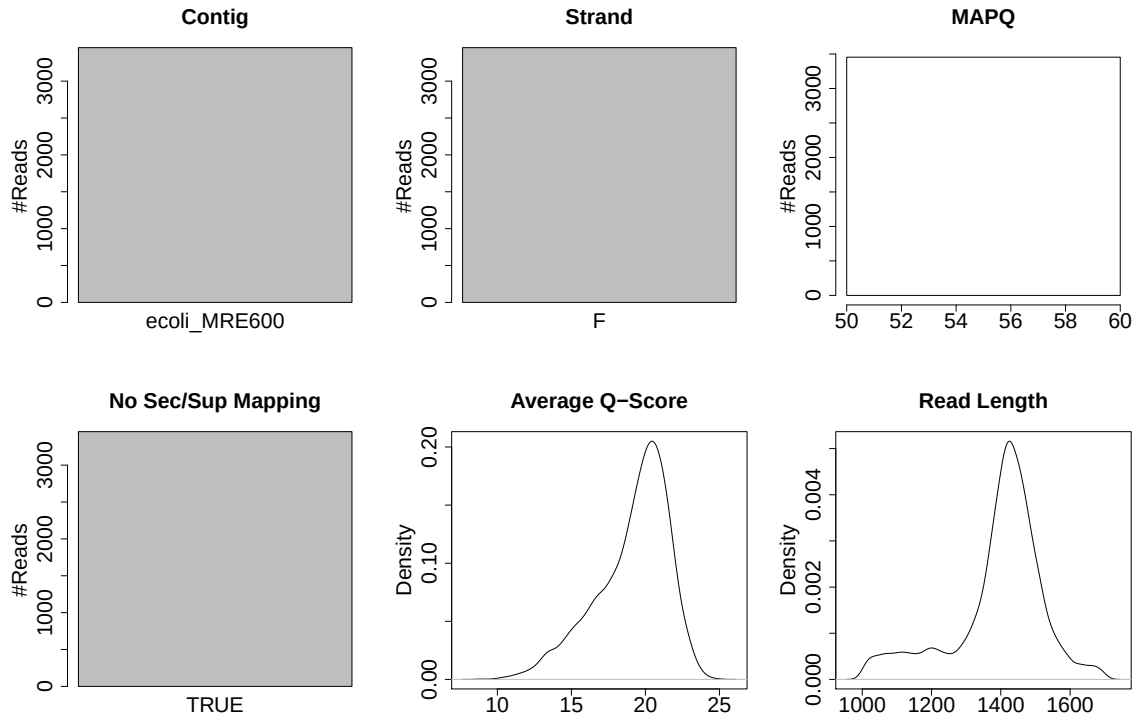

Figure 2.18: Unsupervised modification number detection for un-modified sites in 16S rRNA dataset. Consistent with modified sites, elbow point on order-p-value curves were to determine the optimal number of components for unmodified sites, as negative controls. All 26 non-modified sites in the "head oligo" (see "Data collection and preprocessing" subsection of Materials and Methods for detail) weer analyzed, and 3 out of 26 were considered as false positive by showing a big decent as order increased from 1 to 2.

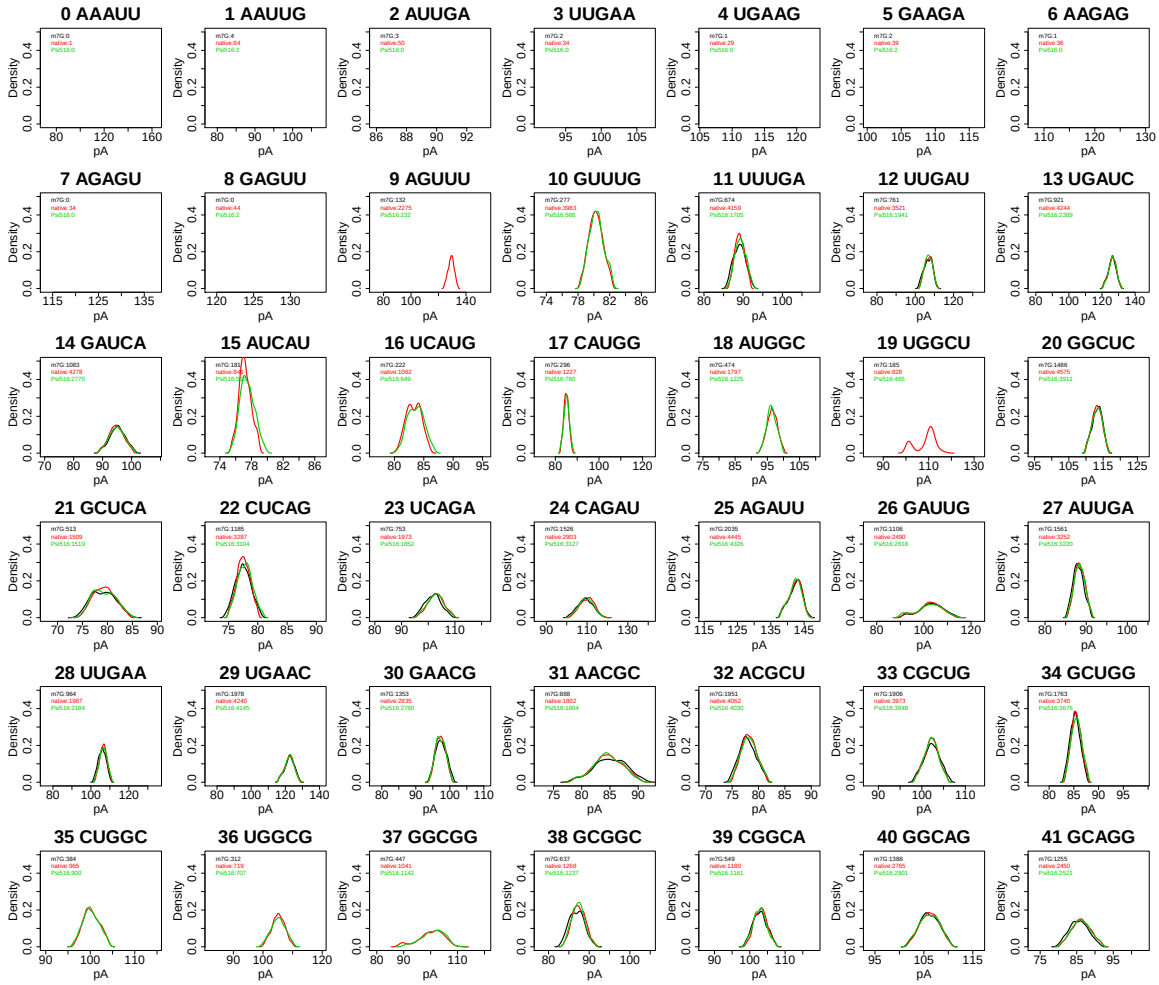

Figure 2.19: Unsupervised modification number detection for modified sites in 16S rRNA dataset. Signal event distribution for the modified kmers UGCCA (A) and GCCGC (B) from the 16S rRNA dataset. Solid black curve, empirical distribution of all kmer signal events mapped to the specific position; solid red curve, empirical distribution of kmer signal events from the m7G-deficient sample; solid green curve, empirical distribution of kmer signal event from native sample; solid blue curve, empirical distribution of kmer signal event from the pseudouridine-deficient sample; dashed curves, Gaussian mixture model-fitted distributions. Numbers in red, green and blue denote sample-wise number of events and percentages for corresponding samples. Numbers in cyan and purple denote the fitted proportion of each component.

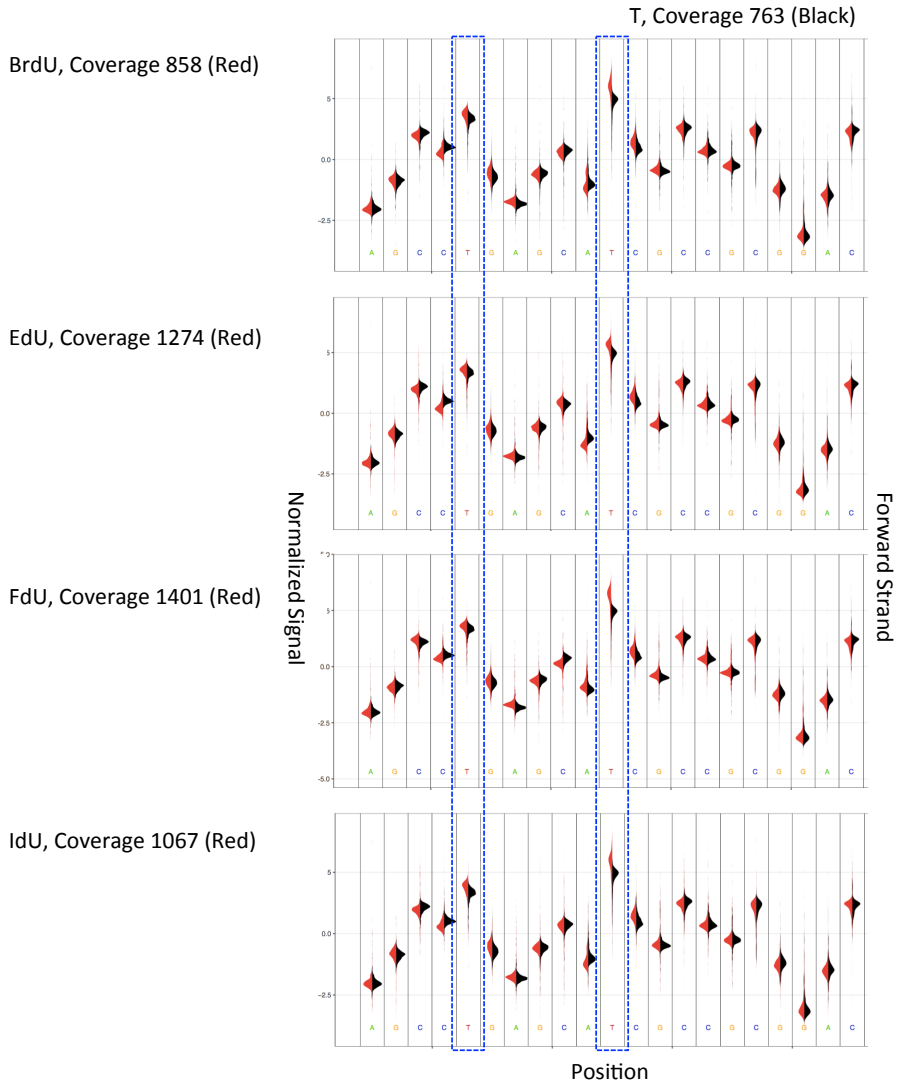

Figure 2.20: Robustness and sensitivity analysis. (A) Boxplots of predicted QGCCA fractions. Actual fractions were shown by horizontal red dashed lines. (B) Boxplots of predicted UGCCA (blue) and QGCCA (black) signal levels (pAs). pAs determined from all observations were shown by horizontal dashed lines.

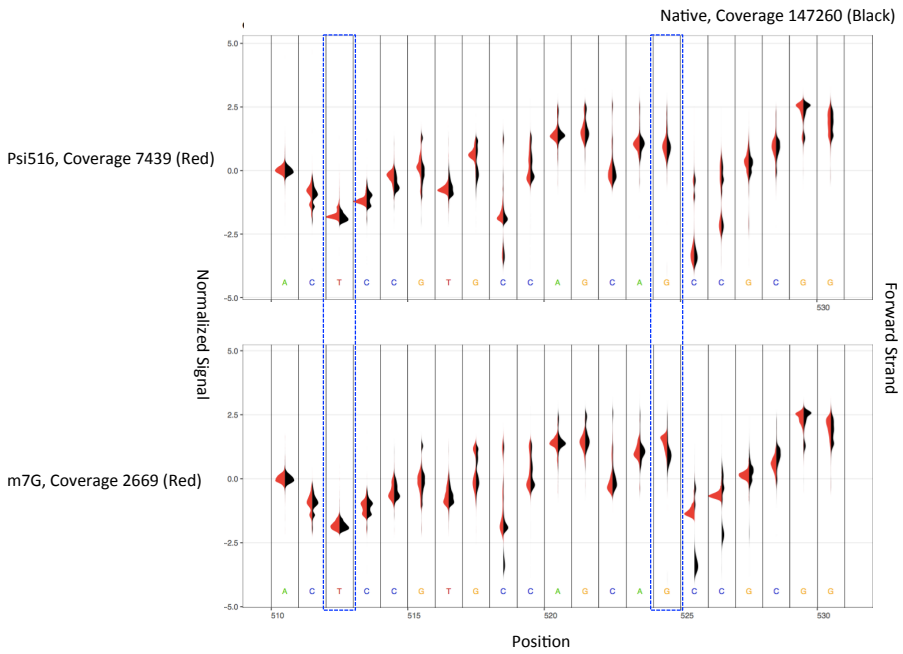

Figure 2.21: Signal distribution of example kmer TGATCC. In the Zymo dataset, TGATCC appears in 3 different sequence contexts, including position 95, reverse strand (black), position 444, forward strand (red) and position 504 reverse strand (blue). Solid curves shows the signal distribution from all reads, and dashed curves shows the signal distribution from high-quality reads (top 50% Q-score).

Chapter 3

Towards Inferring Nanopore Sequencing Ionic Currents from Nucleotide Chemical Structures

Hongxu Ding[1,2,3,*], Ioannis Anastopoulos [1,2,3], Andrew D. Bailey IV [1,2,3], Joshua Stuart[1,2,*] and Benedict Paten[1,2,*]

[1]Department of Biomolecular Engineering, UC Santa Cruz, Santa Cruz, California, USA. [2]UC Santa Cruz Genomics Institute, Santa Cruz, California, USA. [3]These authors contributed equally to this work. [*]Correspondence should be addressed to H.D. (hding16@ucsc.edu), B.P. (bpaten@ucsc.edu) or J.S. (jstuart@ucsc.edu).

3.1 Abstract

The characteristic ionic currents of nucleotide kmers are commonly used in analyzing nanopore sequencing readouts. We present a graph convolutional network-based deep learning framework for predicting kmer characteristic ionic currents from corresponding chemical structures. We show such a framework can generalize the chemical information of the 5-methyl group from thymine to cytosine by correctly predicting 5-methylcytosine-containing DNA 6mers, thus shedding light on the de novo detection of nucleotide modifications.

3.2 Glossary

Kmer, DNA or RNA sequence with length of k. **Canonical kmer**, kmer sequences purely composed of non-modified nucleotides, including A, T, G, C for DNA and A, U, G, C for RNA. **Characteristic ionic current**, ionic currents yielded by a specific kmer are usually modeled by a Gaussian distribution, the mean of which is referred to as the characteristic ionic current. **Kmer model**, a table recording kmers and their corresponding nanopore sequencing characteristic ionic currents. To avoid confusion, the "deep learning model" will be referred to as "framework" throughout the paper. **Framework**, in this paper "framework" specifically refers to the deep learning model used to predict the characteristic ionic current from kmer chemical structures. **GCN**, Graph Convolutional Network. **CNN**, Convolutional Neural Network. **NN**, Neural Network. **RMSE**, Root Mean Square Error. **R**, Pearson corre-

lation. **BA**, Balanced accuracy. **5mC**, 5-methylcytosine. **6mA**, N6-methyladenine. **I**, Inosine. **SMILES**, Simplified Molecular Input Line Entry System for annotating chemical structures using character strings. **Atom**, specifically refers to non-hydrogen atoms throughout the paper.

3.3 Introduction

During nanopore sequencing, consecutive nucleotide sequence kmers block the pores sequentially, producing ionic currents [35]. Chemical modifications on nucleotides additionally alter the ionic currents measured during nanopore sequencing [135, 152, 100, 99, 111, 112, 115, 59, 81, 116, 89, 96, 154, 172, 184, 102, 106, 120, 157, 4, 56]. The characteristic ionic currents of kmers, which are represented in kmer models, are used in interpreting nucleotide modifications [135, 152, 112, 184]. Up to now, 29 [135, 152, 100, 99, 111, 112, 115, 59, 81, 116] and 30 [89, 96, 154, 172, 184, 102, 106, 120, 157, 4, 56] modifications have been successfully characterized in DNA and RNA, respectively. To date, most modification analysis algorithms are based on kmer models [135, 152, 101, 41]. However, such learning strategies struggle to generalize knowledge between related kmers. For example, our previous hierarchical Dirichlet process approach could be structured to learn associations between kmers with specific shared properties, e.g. by numbers of pyrimidine bases, but could not generally learn relationships between arbitrary chemical similarities [135]. Moreover, such approaches necessarily represent base modifications as distinct, unrelated char-

acters. The upshot being that such kmer character-based models require extensive training data and are unable to de novo predict the impact of a chemical modification. Given that the number of possible kmers increases polynomially with the number of modifications being modeled, it is extremely challenging to generate sufficient control data for such models, especially considering that more than 50 and 160 nucleotide modifications have been verified in DNA and RNA respectively [155, 15].

To start to tackle this problem, we propose a graph convolutional network (GCN)-based deep learning framework [46, 79] for predicting kmer characteristic ionic currents from corresponding kmer chemical structures. We confirm that the proposed framework is able to represent individual kmer chemical modules, such as the phosphate group, the sugar backbone, as well as the nucleobase methyl and amine groups. We further demonstrate that this framework can infer full kmer models even when the training data does not include all possible kmers. This opens up the possibility of modeling kmers that are under-represented in control datasets. We also show the framework can generalize the 5-methyl group in thymine to cytosine, thereby accurately predicting the characteristic ionic currents of 5-methylcytosine (5mC)-containing DNA 6mers. Such generalization of chemical information is a reason for optimism about the potential for *de novo* detection of nucleotide modifications.

3.4 Results

3.4.1 Architecture of the deep learning framework

Our deep learning framework consists of three groups of layers, including GCN layers, convolutional neural network (CNN) layers, and one fully connected neural network (NN) layer. As shown in Figure 3.1A, the kmer chemical structures are first represented as graphs, with atoms as nodes and covalent bonds as edges. The atom chemical properties are then assigned as node attributes. Based on such graphs, GCN layers extract one chemical feature vector for every atom, by visiting its immediate graph neighbors. By this means, after several GCN layers, atom feature vectors will contain chemical information for all atoms within a certain graph distance. Specifically, this distance equals the number of GCN layers applied. Considering the small encoding distance of each layer of a GCN, to improve the encoding efficiency of the framework, CNN layers are then applied to summarize relatively long-range chemical information above the GCN layers. The output matrices of the final CNN layer are then "flattened" as feature vectors. Such feature vectors are then passed to the final fully connected NN layer to summarize kmer-level information and finally predict the kmer characteristic ionic currents (see section 2.3). For DNA and RNA, the corresponding best-performing architecture in hyper-parameter tuning was selected for downstream analysis (see 3.6).

3.4.2 Kmer-level generalization

We first confirmed the proposed framework can accurately predict characteristic ionic currents of kmers from their chemical structures. To do so, we performed a down-sample analysis on the canonical DNA 6mer model provided by Oxford Nanopore Technologies (ONT, see section 3.6), by randomly partitioning canonical DNA 6mers with various train-test splits. For each train-test split group, we performed 50-fold cross-validation and used root mean square error (RMSE) and Pearson correlation (r) to quantify the goodness-of-fit (see section 3.6). As shown in Figure 3.1B, Supplementary Figure 3.6 and Supplementary Table 3.11.1, the performance stabilized as more than 40% of DNA 6mers were included in the training. Specifically, for DNA 6mers only used in the test, average RMSE and Pearson correlation reached 1 and 0.995, respectively. Such a result indicated on average 40% of randomly selected DNA 6mers contain sufficient information to recapitulate the full DNA 6mer model.

We next explored how training specific kmer subsets influence the ionic current predictions. Specifically, we trained the framework using either the DNA 6mers that a) do not contain a given nucleotide (base-dropout), b) do not specify a nucleotide at a given position (position-dropout) and c) that are combined from different base-dropouts (for instance using the union of A-dropout and T-dropout kmers, such that kmers containing both A and T would be excluded, but not kmers containing either A or T, noted as A-T model combination, see section 3.6 and Supplementary Note 3.11.2 for details). As with the down-sample analysis for each group in a-c) 50 independent

repeats were performed, and goodness-of-fit was used to evaluate the performance. As shown in Figure 3.1B and Supplementary Figure 3.6, base and position-dropouts significantly decreased the prediction power. Moreover, dropouts in 3rd and 4th positions contributed the most to prediction power decrease, followed by 2nd and 5th positions, consistent with [41]. Model combinations, on the other hand, in general had a minor influence on the prediction power.

The above-mentioned analyses together suggest, once properly trained with sufficient and diverse 6mers, the kmer-level generalizability of the framework. To further validate and extend our framework we performed all the above-mentioned analyses using RNA, switching to using 5mers instead of 6mers to match the available training data. Considering the significantly smaller amount of training data (1/4th the number of distinct RNA 5mers vs DNA 6mers), the prediction power of the RNA architecture is compromised. However, once trained with a similar number of kmers, the RNA architecture yielded comparable prediction power. For instance, the RNA 0.95-0.05 (972 training kmers) and DNA 0.25-0.75 (1024 training kmers) train-test splits yielded comparable performance on test data. Such a result suggests the validity of our proposed architecture (see section 3.6, Supplementary Figure 3.7 and Supplementary Note 3.11.3).

Such kmer-level generalizability could facilitate nucleotide modification detection by greatly reducing the required control data to generate reliable full modification-containing kmer models. As a proof-of-concept, we trained the DNA deep learning architecture with all canonical 6mers plus 1%, 5%, 10%, 30%, 50%, 70%, 90% of

randomly selected 5mC-containing 6mers. The characteristic ionic current signals of such 5mC-containing DNA 6mers were obtained from the nanopolish model as reported in [152, 101]. For each training group 50 independent repeats were performed (see section 3.6). As shown in Figure 3.1C and Supplementary Figure 3.8, decent goodness-of-fit could be obtained when as few as 5% of 5mC-containing DNA 6mers were used as training data. Specifically, for test DNA 6mers, average RMSE and Pearson correlation reached 1.2 and 0.995, respectively. Furthermore, models trained with knowledge of 50% 5mC-containing DNA 6mers performed about as well as models trained with 90%.

3.4.3 Chemical group-level generalization in DNA 5mC de novo prediction

We noted that performance of the model on held out 5mC kmers trained with just 1% of 5mC kmers was better than chance. This raised the question of if chemical group-level information was being usefully generalized among nucleotides by our framework, potentially allowing the 5mC to be predicted de novo, without ever having been seen by the model. As a chemical derivative of cytosine, 5mC contains an additional methyl group at the 5th position (5-methyl) of the pyrimidine ring. This 5-methyl group is shared between 5mC and thymine. We thus hypothesized that 5mC can be generalized by combining the pyrimidine ring from cytosine and 5-methyl group from thymine. As a proof-of-concept, we trained the framework with all canonical DNA 6mers to make de novo predictions on 5mC-containing DNA 6mers. Similar to

previous analyses, 50 independent repeats were performed, and the prediction power was first quantified by goodness-of-fit against the above-mentioned nanopolish model. As shown in Figure 3.1D and Supplementary Figure 3.8, although goodness-of-fit of 5mC-containing DNA 6mers were significantly worse than canonical counterparts, decent performance could still be obtained (average RMSE and Pearson correlation reached 1.8 and 0.993, respectively). We also compared the goodness-of-fit between canonical and 5mC-containing DNA 6mers, and as shown in Supplementary Figure 3.9, a positive correlation trend could be observed. Such a result confirmed that no overfitting was introduced during architecture-training with canonical DNA 6mers, and further suggested 5-methyl generalization.

3.4.4 Predictive analysis

We next performed "predictive analysis" to test whether the DNA 6mer models inferred by our deep learning framework could be used to correctly predict DNA C/5mC status at a per-read, per-site resolution from ionic currents ("predictive accuracy", see section 3.6). C/5mC-sites to be predicted were confirmed by bisulfite sequencing (see section 3.6). We also quantified the predictive accuracy with the above-mentioned nanopolish model as a baseline control (see section 3.6). As shown in Figure 3.1E, average predictive accuracy, quantified by balanced accuracy, became comparable with baseline control with 50% of imputed 5mC-containing 6mers. Taken together, these results confirmed the kmer-level generalizability of our framework, as well as suggesting that reliable modification-containing kmer models can be built

with significantly less control data once facilitated by our methodology. Such a result confirmed the successful 5-methyl generalization. More confusion matrix-based prediction evaluations can be found in Supplementary Figure 3.10.

3.4.5 The encoding of chemical structures

To better understand how chemical structures were encoded we visualized DNA 6mer atom similarity matrices. Specifically, we trained the proposed framework with all canonical DNA 6mers. We then calculated and visualized the Pearson correlations of the feature vectors derived by the final GCN layer as atom-level similarities. As shown in Supplementary Figure 3.11, we visualized 10 randomly chosen canonical DNA 6mers. Taking CGACGT as an example, as shown in Figure 3.2A and 3.2B, atoms were in general aggregated by chemical contexts. For instance, for the first cytidine monophosphate in CGACGT, atoms #0-4 were tightly clustered with average r>0.9, recapitulating the phosphate group. Atoms #5-8 and #17-18 also clustered with average r>0.9, denoting the deoxyribose backbone. Among cytosine atoms #9-16, #9 nitrogen atom connected the nucleobase to the deoxyribose backbone, atoms #10-11 denoted the C=O group, and atoms #12-16 composed the C=C-C=N conjugation system and the covalently bonded amine group. Similarly, atoms in other nucleotides can also be clustered into phosphate groups, deoxyribose backbones and nucleobases. Within the nucleobases, chemical modules including chemical groups and conjugation systems can further be dissected. Such a phosphate-deoxyribose-nucleobase pattern repeated and constituted DNA 6mers.

We also examined the inter-nucleotide similarities of different components. As shown in Figure 3.2A and 3.2B, in general high similarities (average r>0.9) were observed among phosphates, as well as deoxyriboses from different nucleotides. Meanwhile, chemical modules sharing similar structures, e.g. the conjugation systems of adenines, cytosines and guanines were more similar to each other. On the other hand, low similarities (average r<0.5) were observed between chemical modules with distinct structures, e.g. the cytosine C=O group and the thymine methyl group. Taken together, these results suggest that the GCN layers in the proposed framework can effectively capture features interpretable as individual chemical modules.

We further visualized the atom-level similarity matrices of 5mC-containing DNA 6mers, aiming to understand the generalization of methyl group among thymine and 5-methylcytosine. We thus trained our deep learning framework with all canonical DNA 6mers, calculated the Pearson correlations of the feature vectors derived by the final GCN layer, and further visualized such atom-level similarity matrices of 10 randomly selected 5mC-containing DNA 6mers (Supplementary Figure 3.12). Taking GT(5mC)AGA as an example (Figure 3.2C and D), the phosphate-deoxyribose-nucleobase repetitive pattern was recapitulated. Within nucleobases, high similarities (average r>0.9) were again observed among chemical modules with similar structures. Specifically, strong similarities (average r>0.9) were observed between thymine (#37-38) and 5mC (#57-58) methyl groups (Me). In addition, such methyl groups were uniquely encoded as they were less correlated with any other DNA 6mer chemical modules (average r<0.5). We also quantified the atom-level similarity between

GT(5mC)AGA and corresponding canonical counterpart GTCAGA. As shown in Supplementary Figure 3.13, strong similarities (average r>0.9) were observed between GT(5mC)AGA and GTCAGA thymine methyl groups, as well as the 5mC methyl groups from GT(5mC)AGA and thymine methyl groups from GTCAGA. These observations together suggested the successful chemical information generalization. Noticeably, the methyl groups were encoded with the pyrimidine backbone C=C modules. Such a result suggests that the GCN-encoding is driven by chemical context, which further implies when generalizing one specific chemical group among different nucleotides, the corresponding chemical contexts in which such chemical group resides should be the same.

Finally, we projected kmer atom feature vectors into the tSNE space, in order to summarize the atom-level similarity matrices further providing a global visualization of kmer atoms. As shown in Figure 2B and E, atoms under the same chemical context clustered together, e.g. phosphate group phosphate atoms (#1, #20, #42, #63, #82, #104 in B and #1, #23, #43, #63, #84, #106 in F), deoxyribose ring oxygen atoms (#7, #26, #48, #69, #88, #110 in B and #7, #29, #49, #69, #90, #112 in E), as well as NH3 group nitrogen atoms (#14, #35, #55, #76, #97 in B and #16, #56, #76, #99, #119 in E). Specifically, as shown in E, in 5mC-containing DNA 6mer GT(5mC)AGA, T-methyl group carbon atom #38 and 5mC-methyl group carbon atom #58 clustered together, along with pyrimidine backbone C=C module atoms #37 and #39 in T, as well as #57 and #59 in 5mC. Taken together, these results confirm that GCN could properly encode chemical structures

based on the corresponding chemical contexts.

3.4.6 Analyzing the 2mG site in E.coli 16S rRNA

Our deep learning framework could potentially shed light on previously understudied, less prevalent nucleotide modifications. As a proof-of-concept, we analyzed 2mG, which can be represented as the purine ring in guanine with the N2-methyl group in 6mA. Specifically, we generated a RNA 5mer model using canonical and 6mA-containing kmers (see section 3.6). We then predicted the characteristic ionic current signals of 2mG-containing RNA 5mers (see section 3.6). To test our predictions, we analyzed nanopore sequencing reads of E.coli 16S rRNA transcript J01859.1, which contains an annotated 2mG at position 1206 (see section 3.6). As shown in Supplementary Figure 3.14, our predictions recapitulated the characteristic ionic current signals of 2mG-containing and pairing canonical RNA 5mers (see section 3.6). Moreover, we confirmed that such predicted characteristic ionic current signals could be used to correctly determine the G/2mG modification status (see section 3.6).

3.5 Discussion

We propose a GCN-based deep learning framework for associating kmer chemical structures with corresponding characteristic ionic currents. We show that such a framework can recapitulate full kmer models from partial training data, thus greatly facilitating modification analysis by reducing the amount of required control data. Specifically, for cases where a small proportion of random kmers are under-

represented in control data, we can apply the same principle as the down-sample analysis to learn around these training deficiencies. For cases where comprehensive control datasets are available only for single modifications, we could apply model combination (as we showed for individual nucleotides) to model kmers containing multiple modifications simultaneously.

We further demonstrated that our framework can represent novel modifications by generalizing encoded chemical groups between nucleotides, thus shedding light on de novo modification detection. However, the current model is not without its limitations. For example, the proposed framework encodes chemical groups, e.g. the methyl groups in thymine and 5mC, as well as the amine groups in cytosine, guanine and adenine, with covalently bonded "backbone atoms", showing a strong chemical context-specificity (Figure 3.2 and Supplementary Figure 3.11, 3.12). Thus, the current framework cannot properly handle "stacked" chemical groups. For instance, the methylamine group in N6-methyladenine (6mA) cannot be correctly encoded by simply stacking methyl with amine. As shown in Supplementary Figure 3.15, substituting A with 6mA was predicted to decrease characteristic ionic currents, which is the opposite of a previous study [111]. Therefore the extensibility of the framework is largely limited. To overcome such a limitation, controlled nanopore sequencing profiles of diverse nucleotide modifications are needed, in addition to the modeling of other chemical interactions.

Deep learning-based approaches have emerged as powerful tools for detecting nucleotide modifications from nanopore sequencing readouts. Compared to kmer

model-based counterparts, deep learning-based approaches are reported to have better accuracy and less computational resource consumption [99] [115]. Recently, ONT released the megalodon algorithm [1], which can drastically increase the accuracy for 5mC identification (Supplementary Figure 3.10, see section 3.6). Thus, one potential future extension of the paper would be using the learned models as components of a larger, recurrent deep neural network.

Another potential future direction would be generalizing the proposed framework to handle both DNA and RNA kmers. Due to different translocation speed, the nanopore sequencing ionic currents of DNA and RNA are not directly comparable [39]. Therefore, advanced deep learning frameworks which can take both kmer chemical structures and nanopore sequencing experimental setups are needed. Considering DNA and RNA share several non-canonical nucleobases, e.g. Inosine (I) [1], we might combine the ribose in RNA and I in DNA to reconstruct I-containing RNA 5mers, and vice versa for I-containing DNA 6mers. By this means, required RNA control nanopore sequencing reads, which are usually challenging to obtain, can be largely compensated. Meanwhile, such generalization would largely diversify the chemical contexts that can be represented, further facilitating the de novo modification analysis.

[1] https://github.com/nanoporetech/megalodon

3.6 Methods

3.6.1 Methods summary

The deep learning framework proposed here aims to associate kmer chemical structures with corresponding characteristic ionic currents. The chemical structure refers to the chemical properties of kmer atoms and how these atoms are covalently bonded. Characteristic ionic current, on the other hand, refers to the average ionic current that a specific kmer produces during nanopore sequencing.

Thus, in the following sections, we first describe how the chemical structures were represented ("graph representation of kmer chemical structures"). We then describe the deep learning framework used in the study ("architecture of the deep learning framework", "training procedure" and "hyper-parameter tuning"). We further describe analyses performed to evaluate the performance of the proposed framework ("down-sample, base-dropout, position-dropout and combination analysis", "predicting modification-containing kmers", "human genome C/5mC-status predictive analysis" and "E.coli 16S rRNA 2mG-site analysis"). Finally we describe all required resources for the study ("kmer models", "data availability" and "code availability").

3.6.2 Graph representation of kmer chemical structures

Following the workflow described in [46], kmer chemical structures were first described by SMILES (Simplified Molecular Input Line Entry System) strings,

which were assembled by concatenating SMILES strings of individual nucleotides, as summarized in Table 3.1. Each nucleotide base can be described by several SMILES strings. The SMILES strings presented in the table below were selected due to the ease of combining them into complete kmers. Based on information provided by Oxford Nanopore Technologies, as well as a previous study [41], DNA and RNA is represented by 6mer and 5mer, respectively. An "O" was then added to the end of each concatenation to represent the residual unbonded hydroxyl group on the sugar backbone.

We then represent the SMILES string of each kmer as a graph noted as $G(A, X)$. Specifically, the topology (atom order is determined by SMILES string) of each kmer chemical structure was represented by an adjacency matrix A, with $A_{i,j}$ equals 1 iff the ith and jth atoms were covalently bonded. Meanwhile, for every atom in A, the corresponding chemical properties were represented by feature matrix X, with X_i recording the chemical property vector for the ith atom. Atom chemical properties included in the study were summarized in Table 3.2. Therefore, the GCN has encoded as input a chemical feature matrix X with the guide of chemical topology matrix A, representing kmer chemical structures. Notably, for convenient GCN implementation, the size of A and X is kept constant. Due to the variable number of atoms across kmers, A and X were thereby padded with zeros based on the largest kmers. Specifically, the A matrix was padded at the end of its rows and columns, with $\dim(A)$ is 133, 133 and 116, 116 for DNA and RNA, respectively. While the X matrix was padded at the end of its rows, with $\dim(X)$ is 133, 8 and 116, 8 for DNA

and RNA, respectively. Note that the kmer representation is guided by the non-zero elements (covalent bonds) in A, thus such padding will not affect the GCN encoding.

3.6.3 Architecture of the deep learning framework

The Graph Convolutional Network (GCN) layers of our framework were built based on the procedure described by [46]. Fast approximate convolutions on G were used to create a graph-based neural network $f(X, A)$, following the propagation rule:

$$H^{l+1} = \sigma\big(\tilde{U}^{\frac{1}{2}}\tilde{A}\tilde{U}^{\frac{1}{2}}H^{l}W^{l}\big) \tag{3.1}$$

$\sigma()$ is the activation function applied to each layer. Here, the activation function used was Exponential Linear Unit (ELU). $\tilde{U}_{i,j} = \sum_j A_{i,j}$ the degree matrix for each atom in the graph. $\tilde{A} = A + I$ adds self edges to each of the atoms. The $\tilde{U}^{\frac{1}{2}}\tilde{A}\tilde{U}^{\frac{1}{2}}$ transformation prevents changes in the scale of the feature vectors [79] and constructs filters for the averaging of neighboring node features. H and W denote the output (activation vectors) and weights of each GCN layer, respectively. The corresponding superscript represents the layer index. $H^0 = X$, however subsequent H represent the GCN derived features.

The intuition of the graph convolution process is described as follows. For every kmer, chemical properties of atoms, together with their covalently bonded neighbors, will be convoluted with the guidance of G. Such graph convolution yields an activation matrix H, following the aforementioned propagation rule. H is an atom-

by-feature matrix, with dimension 133, N and 116, N for each of the DNA and RNA kmers, respectively. Here N equals the number of nodes of the GCN layer, which determines the number of features to be derived. The selection rule for N is described in the following section. As more GCN layers are stacked, the graph convolution process is repeated. The H matrix will thus contain chemical information of all atoms within a certain graph distance, which equals the number of GCN layers applied. By this means, "chemical modules" composed of several atoms linked by covalent bonds are encoded.

Considering the small encoding distance of a GCN, for a better encoding efficiency we wanted additional layers that can quickly summarize atom information. We thus applied standard 1-D CNN layers with Rectified Linear Unit (ReLU) activation right after the GCN layers. Average Pooling [87] was applied on the output of each 1-D CNN layer. Average Pooling takes the average of each 2x2 patch of the CNN output matrix. Specifically, output dimension of the first CNN layer equals 133-K+1, N' and 116-K+1, N' for DNA and RNA kmers, respectively. Here K is the CNN kernel size and N' is the node number of the final GCN layer. Output dimensions of subsequent CNN layers equals m-K+1-2+1, n-2+1, where m, n denotes the output dimension of the previous layer, and 2 denotes the Average Pooling patch size. The output from the final 1-D CNN layer, after Average Pooling, was passed to a Flatten layer, which converts the final 1-D CNN output matrix to a 1-D feature vector in a row-wise fashion. The NN layer then takes the flattened vector as input, thereby summarizing information about the entire kmer, and producing a highly informative

representation. Elements of the NN layer output vector are linearly combined as the final pA value.

3.6.4 Training procedure

Our framework was trained with the Keras [30] framework with TensorFlow backend using the Adam [78] optimizer for gradient descent optimization. The framework was allowed to train for a maximum of 500 epochs. To control for overfitting, EarlyStopping [189] was used by monitoring the increase in validation loss. Early termination of training was reached if the validation loss was increasing for 10 consecutive epochs, indicating that the framework had reached maximum convergence. Mean Squared Error (MSE) was used as the loss function during the training process. Meanwhile, a 10% random dropout was applied after each layer, to further prevent overfitting [156]. In the following experiments the exact same training routine was used.

3.6.5 Hyper-parameter tuning

In order to determine the optimal architecture, we performed hyperparameter grid search. The search involved the hyperparameters shown in Table 3.3. We used the following scaling factor to determine the number of nodes in each GCN/CNN layer of our framework:

$$n = 16 * 2^{l-1} \tag{3.2}$$

where l is the layer index of the GCN, CNN, and NN layer groups. For instance, the number of GCN layers determined to yield the best performance for DNA were 4. The number of nodes for each GCN layer was therefore 128, 64, 32, and 16. The same logic was applied to all other layer groups. We performed 10-fold cross validation for each hyper-parameter combination. The combination that produced the lowest average RMSE across all folds was adopted as the optimal architecture. The optimal DNA framework has 4 GCN layers, 3 CNN layers with a kernel size of 10 and 8192 nodes in the NN layer. The optimal RNA framework has 4 GCN layers, 5 CNN layers with a kernel size of 10 and 8192 in the NN layer.

3.6.6 Down-sample, base-dropout, position-dropout and combination analysis

For down-sample analysis, we performed random train-test splits in 5% intervals, noted as 0.95-0.05, etc. For base-dropout analysis, we created training sets by removing certain bases. Such train-test split creates 729/4096 (18%) training kmers and 3367/4096 (82%) test kmers for DNA, and 243/1024 (24%) training kmers and 781/1024 (76%) test kmers for RNA. It is important to note that everytime a base is dropped from the training set it is retained in the test set. Similar to base-dropout, the position-dropout adds one more dimension, which is the position of the nucleotide base. For a given position-dropout, the testing kmers are all kmers with the dropout nucleotide covering the target position, and the training kmers are the remaining kmers. Such position-dropout creates 3072/4096 (75%) training kmers and 1024/4096

(25%) test kmers for DNA, and 768/1024 (24%) training kmers and 256/1024 (25%) test kmers for RNA. It is important to note that bases dropped in a specific position in the training appear in the same position in testing. For combination analysis, we trained the framework by combining any of the two base-dropout kmer sets. For instance, all G and C-dropout DNA 6mers, which was noted as G-C. Such analysis creates 1394/4096 (34%) training kmers and 2702/4096 (66%) test kmers for DNA, and 454/1024 (44%) training kmers and 570/1024 (56%) test kmers for RNA. For each above-mentioned train-test split, in order to perform statistical analyses, we produced 50 independently trained frameworks for each experiment. Specifically, we performed 50-fold cross validation in the down-sample analysis, considering for each fold the train kmers were randomly selected. As for other analyses, we performed 50 independent repeats using the same training kmer sets. The variability among repeats came from the stochasticity of the training process. To confirm the robustness of our architecture, we further performed two independent replicates (run-1 and run-2) of 50.

3.6.7 Predicting modification-containing DNA 6mers

For the 5mC imputation experiment, the framework was trained on all 4096 A, T, C, G DNA 6mers, plus 1%, 5%, 10%, 30%, 50%, 70%, 90% of randomly selected 5mC-containing DNA 6mers, following the training process as described above. In order to perform statistical analyses, we produced 50 independently trained frameworks (50 independent repeats) for each category, with in total two independent replicates

(run-1 and run-2) of 50. Such frameworks were then applied on all 15625 possible A, T, C, G, 5mC DNA 6mers. For the chemical group-level generalization experiment, the framework was trained on all 4096 A, T, C, G DNA 6mers following the training process as described above. In order to perform statistical analyses, we produced 50 independently trained frameworks (50 independent repeats), with in total two independent replicates (run-1 and run-2) of 50. Such frameworks were then applied on all 15625 possible DNA 6mers, including those composed of A, T, C, G, 5mC and A, T, C, G, 6mA.

3.6.8 Predictive analysis of predicted kmer models

Overview

To test whether the generated kmer models can be used to correctly interpret C-5mC status from nanopore readouts, we performed predictive analysis by using signalAlign to make per-read per-base predictions [135]. For a given reference position, signalAlign can produce posterior probabilities for all possible bases based on a provided kmer model. Thus, for DNA 6mer models generated as described in "predicting modification-containing DNA 6mers", the empirical nanopolish [152] [101] model obtained as described in "kmer models", we allowed signalAlign to predict between C and 5mC. Considering no significant goodness-of-fit differences were observed between run-1 and run-2, only models generated in run-1 were used here. All predictive analyses performed in this paper were within the human NA12878 cell line.

Selecting prediction sites

The prediction sites were selected among the entire human genome. To avoid artifacts caused by ambiguous genomic DNA modification status, we only focused on confident 5mC sites and canonical genomic regions in our analysis. Besides 5mC, other modifications exist in genomic DNA. Considering extremely low fractions of other modifications, e.g. only 0.05% are modified as 6mAs in the human genome [187], we define "non-5mC" sites as "canonical regions" during predictive analysis. Among these canonical regions, we used the Poisson process with lambda equals 50 to randomly select genomic sites for signalAlign to predict. Such selected sites were at least 12 nucleotides apart, avoiding potential interference by the neighbors. We thus obtained confident 5mC and C sites for signalAlign prediction. The genomic DNA C-5mC status was determined by analyzing two independent NA12878 cell line bisulfite sequencing datasets [44]. A C-site was determined as confidently methylated if, for both bisulfite sequencing datasets, 95% of reads were methylated with at least 10x coverage. On the other hand, a C-site was considered confidently unmodified if, for both bisulfite sequencing datasets, at most 1% of reads were methylated with at least 10x coverage. Such analysis covered 3367/3367 canonical C-containing DNA 6mers, and 3950/6144 single 5mC-containing DNA 6mers.

Selecting nanopore sequencing reads

We then ran signalAlign with reads reported in the nanopore consortium NA12878 cell line native genomic DNA datasets [73] covering the above-mentioned

prediction sites. Considering the computational complexity of signalAlign, we performed the following filtering steps to use the fewest reads to cover the most kmers. First, we calculated read-level kmer coverage. For example, the center 5mC-site of DNA read CAGAT(5mC)ACAGA was selected for signalAlign prediction. 6mers CAGAT(5mC), AGAT(5mC)A, GAT(5mC)AC, AT(5mC)ACA, T(5mC)ACAG and (5mC)ACAGA span such 5mC-site, therefore considered as being covered. Based on such read-level kmer coverage, we iteratively selected reads that covered the least frequently covered kmers. Thus, building a read set which covers as many kmers as possible as often as possible with the fewest number of reads. We included two biological replicates of NA12878 cell line native genomic DNA sequencing experiments (FAB39088 and FAF01169) in the C-5mC predictive analysis. For such analysis, our final FAB39088 set contained 1706 reads, which covered 2625/3367 C-only DNA 6mers with an average 61.52x coverage as negative control, and 3105/3950 possible single-5mC DNA 6mers with an average 5.01x coverage. The final FAF01169 set contained 1396 reads, which covered 2610/3367 C-only DNA 6mers with an average 63.26x coverage as negative control, and 3140/3950 single-5mC DNA 6mers with an average 4.76x coverage. Combining the two sets, in total 2792/3367 C-only DNA 6mers were covered with an average 58.49x coverage, and 3481/3950 single-5mC DNA 6mers were covered with an average 4.38x coverage.

Performing signalAlign prediction

Based on the selected prediction sites and nanopore sequencing reads as described above, per-read per-site predictive analysis was performed by signalAlign. The signalAlign analysis was performed with default parameters, except for internal read-level quality filtering. Such quality filtering removes reads with poor kmer to ionic current correspondence. During signalAlign analysis, kmer-to-ionic current correspondence probability matrices (event tables) are first generated. Based on such event tables, signalAlign will remove reads with low average probabilities (<10-5). Additionally, reads with >50 consecutive ionic current signals that cannot be corresponded to kmers (probability equals 0) will be discarded. Considering the event table generation is based on the provided kmer model, therefore after the above-mentioned default quality filtering, the number of remaining reads varies when different kmer models are supplied during predictive analysis. To ensure the statistical soundness, we deactivate the default quality filtering, such that reads to be analyzed by different supplied kmer models will be the same.

Performing megalodon prediction

We also performed predictive analysis using the deep learning-based modification calling algorithm megalodon[2] as an additional baseline control. The megalodon (version 2.3.1) analysis was performed with tags "fast5 --outputs mod_mappings mods --reference reference –processes 1 –overwrite --guppy-server-path guppy_basecall_se

[2]https://github.com/nanoporetech/megalodon

rver –output-directory output_dir –guppy-timeout 1000 --guppy-concurrent-reads 1 –guppy-params '–num_callers 7 –cpu_threads_per_caller 10 –chunks_per_runner 100' "

Considering the extraordinary performance of megalodon (Supplementary Figure 3.10), we further used megalodon predictions as an additional ground truth for the C/5mC status for every nanopore sequencing read at every prediction site. Please see Supplementary Note 3.11.4 for more information.

Quantifying predictive accuracy

signalAlign quantifies the probability of being C or 5mC for every prediction. We used probability threshold 0.7 to ensure only confident predictions were included in predictive accuracy quantification. Together with the megalodon 5mC calling results, we further created confusion matrices (2x2 for 5mC predictive analysis with 5mC as "positive" class and C as "negative" class) to quantify predictive accuracy. Specifically, we calculated the true positive rate (TPR), true negative rate (TNR), positive predictive value (PPV), negative predictive value (NPV), F1-score (F1) and balanced accuracy (BA) as predictive accuracy quantifications. BA was presented in Figure 3.1E as representative quantification, and the full predictive performance can be found in Supplementary Figure 3.10.

3.6.9 E.coli 16S rRNA 2mG-site analysis

Ionic current signal distributions

We first downloaded the nanopore sequencing fast5 reads of E.coli 16S rRNA nanopore sequencing reads reported in [154]. We then performed nanopolish extract analysis [152, 101]to retrieve the fastq records, with tags "-v -r -q -t template". The fastq records were then aligned using minimap2 [92] with flags "-ax map-ont", further sorted and indexed by samtools [93]. Per-read event tables were generated using nanopolish eventalign with flag "–scale-events", by taking fast5 reads, alignment files, and retrieved fastq records as described above. The yielded event tables contain RNA 5mer sequences and corresponding ionic current signals. We then quantified the distributions of RNA 5mer ionic current signals.

Predictive analysis

We also performed predictive analysis for the A, 6mA, T, G, 2mG, C RNA 5mer model described in "predicting modification-containing kmers". Specifically, we tested whether the predicted RNA 5mer model could be used to correctly identify the 2mG site in E.coli 16S rRNA (position 1206, see `https://www.ncbi.nlm.nih.gov/nuccore/J01859` for details). We thus ran signalAlign with nanopore sequencing reads reported in [14], following the same steps as described in "human genome C/5mC-status predictive analysis". We also used probability threshold 0.7 to select confident predictions.

3.6.10 Kmer models

Canonical DNA 6mer and RNA 5mer models are available at: `https://github.com/nanoporetech/kmer_models`. The Nanopolish 5mC-containing DNA 6mer model is available at: `https://github.com/nanoporetech/nanopolish/tree/master/etc/r9-models`. The GSE124309 model, which contains the union of {A, U, C, G} and {6mA, U, G, C} RNA 5mers, was constructed by the following steps. We first downloaded the nanopore sequencing fast5 reads of modified and non-modified "curlcake constructs" replicate 1 with GEO accession code GSE124309 [96]. We then performed nanopolish extract analysis [152, 101] to retrieve the fastq records, with tags "-v -r -q -t template". The fastq records were then aligned using minimap2 [92] with flags "-ax map-ont", further sorted and indexed by samtools [93]. Per-read event tables were generated using nanopolish eventalign with flag "–scale-events", by taking fast5 reads, alignment files, and retrieved fastq records as described above. The yielded event tables contain RNA 5mer sequences and corresponding ionic current signals. For every RNA 5mer, we averaged ionic current signals of all instances recorded in the event tables to build the GSE124309 model. Please note that for more recent nanopore sequencing chemistries, e.g. R10 where ONT kmer models are no longer available, empirical kmer models could be trained instead as above-mentioned. Please see Supplementary Note 3.11.5 for details.

3.6.11 Data availability

The FAB39088 and FAF01169 NA12878 cell line native genomic DNA nanopore sequencing datasets were downloaded from `https://github.com/nanopore-wgs-consortium/NA12878/blob/master/Genome.md`. The two independent NA12878 bisulfite datasets were downloaded from `https://www.encodeproject.org/experiments/ENCSR890UQO/`.

3.6.12 Code availability

Codes for constructing, training and running the deep learning framework are available at `https://github.com/ioannisa92/Nanopore_modification_inference`. Codes for nanopore sequencing data analysis are available at `https://github.com/adbailey4/functional_model_analysis`. Codes for reproducing all figures are available upon request to the corresponding authors.

3.7 Acknowledgements

Research reported in this publication was supported by the National Institutes of Health under Award Numbers R01-HG010053-02, U01HG010961, U41HG010972, R01HG010485, 2U41HG007234, 5U54HG007990, 5T32HG008345-04 and U01HL137183. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. The authors would thank Jordan Eizenga, Dr. Jonas Sibbesen, Dr. Mark Akeson and Dr. Miten Jain for

critical insight and help with drafting the manuscript.

3.8 Author Contributions

H.D. conceived the idea. I.A. performed deep learning framework modeling, optimization and analysis. A.B. and H.D. performed the nanopore sequencing data analysis. H.D., J.S. and B.P. supervised the project. All authors prepared the manuscript.

3.9 Figures

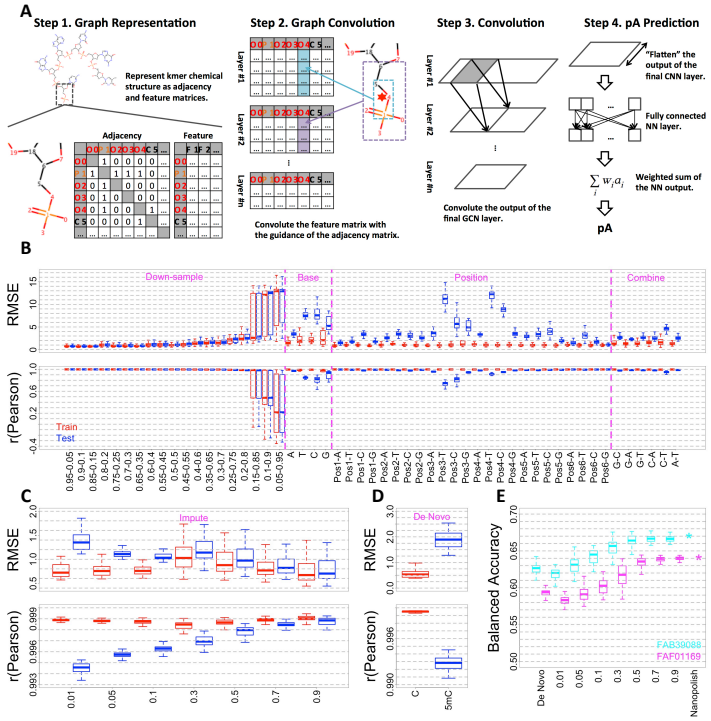

Figure 3.1: **Predicting kmer characteristic ionic currents from chemical structures.** (A) Graphic overview of the proposed deep learning framework for DNA analysis. (B) Goodness-of-fit of DNA canonical random down-sample, base-dropout, position-dropout and model combination analyses. (C) Goodness-of-fit of 5mC-containing DNA 6mer imputation analysis. (D) Goodness-of-fit of de novo 5mC-containing DNA 6mer prediction. C and 5mC refer to goodness-of-fit of canonical DNA 6mers and 5mC-containing DNA 6mers, respectively. In panel B-D, Train (red) and Test (blue) refer to goodness-of-fit of the training and test DNA 6mers, respectively. (E) Predictive accuracy of C/5mC status quantified by balanced accuracy.

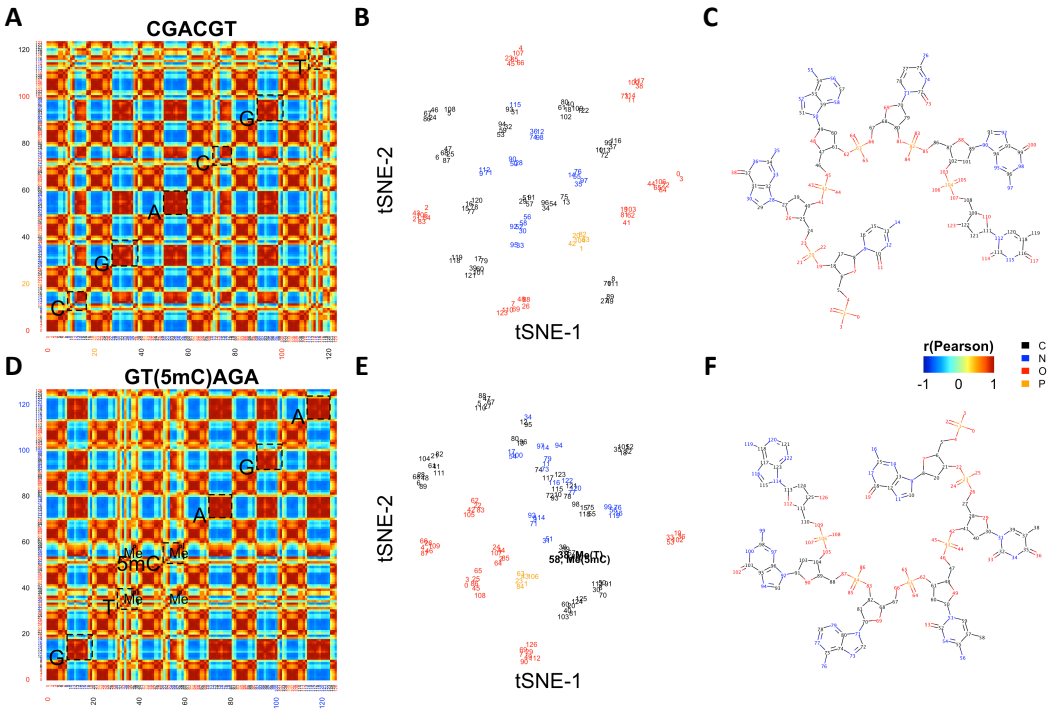

Figure 3.2: **Visualizing the encoding of chemical structures.** (A-C) Atom similarity matrix, tSNE visualization and chemical structure of the example canonical DNA 6mer CGACGT. In (A) and (B), atoms were numbered and colored based on the chemical structure in (C). Carbon, nitrogen, oxygen and phosphorus were colored as black, blue, red and orange, respectively. Specifically, in (A), nucleobases were highlighted by dashed boxes. (D-F) Atom similarity matrix, tSNE visualization and chemical structure of the example 5mC-containing DNA 6mer GT(5mC)AGA. In (D) and (E), atoms were numbered and colored based on the chemical structure in (F). Carbon, nitrogen, oxygen and phosphorus were colored as black, blue, red and orange, respectively. Specifically, in (D) and (E), methyl group carbon atoms (#38 in T and #58 in 5mC) were highlighted.

3.10 Tables

Table 3.1: SMILE String Encoding

Nucleotide	**SMILES string**
A (DNA)	OP(=O)(O)OCC1OC(N3C=NC2=C(N)N=CN=C23)CC1
T (DNA)	OP(=O)(O)OCC1OC(N2C(=O)NC(=O)C(C)=C2)CC1
C (DNA)	OP(=O)(O)OCC1OC(N2C(=O)N=C(N)C=C2)CC1
G (DNA)	OP(=O)(O)OCC1OC(N2C=NC3=C2N=C(N)NC3=O)CC1
5mC (DNA)	OP(=O)(O)OCC1OC(N2C(=O)N=C(N)C(C)=C2)CC1
6mA (DNA)	OP(=O)(O)OCC1OC(N3C=NC2=C(NC)N=CN=C23)CC1
A (RNA)	OP(=O)(O)OCC1OC(N3C=NC2=C(N)N=CN=C23)C(O)C1
U (RNA)	OP(=O)(O)OCC1OC(N2C(=O)NC(=O)C=C2)C(O)C1
C (RNA)	OP(=O)(O)OCC1OC(N2C(=O)N=C(N)C=C2)C(O)C1
G (RNA)	OP(=O)(O)OCC1OC(N2C=NC3=C2N=C(N)NC3=O)C(O)C1

Table 3.2: Features in Feature matrix X

Feature	**Description**
Carbon	1 if the atom is carbon, 0 otherwise (boolean)
Nitrogen	1 if the atom is nitrogen, 0 otherwise (boolean)
Oxygen	1 if the atom is oxygen, 0 otherwise (boolean)
Phosphorus	1 if the atom is phosphorus, 0 otherwise (boolean)
Atom degree	Total number of covalent bonds around an atom (integer)
Implicit valence	It equals the valence of the atom minus the valence calculated from the bond connections (integer)
Number of hydrogens	Total count of hydrogens (integer)
Aromaticity	1 if atom in an aromatic ring, 0 otherwise (boolean)

Table 3.3: Hyper-parameter tuning grid search parameters

Parameters	**Space Searched**	**Best Parameters (DNA)**	**Best Parameters (RNA)**
The number of GCN layers	{2, 3, 4, 5, 6}	4	4
The number of CNN layers	{2, 3, 4, 5, 6}	3	5
The kernel size for the CNN layers	{2, 4, 10, 20}	10	10
The number of nodes in dense (NN) layer	{32, 128, 512, 2048, 8192}	8192	8192

3.11 Supplementary Information

3.11.1 Supplementary Table 1

	RMSE		r(Pearson)	
0.05–0.95	12.93	12.92	0.2142	0.215
0.1–0.9	12.11	12.51	0.4823	0.4832
0.15–0.85	2.591	2.899	0.9874	0.9833
0.2–0.8	2.401	2.533	0.9875	0.9846
0.25–0.75	2.049	2.243	0.9918	0.9895
0.3–0.7	1.505	1.633	0.9964	0.9948
0.35–0.65	1.468	1.595	0.9959	0.9944
0.4–0.6	1.285	1.441	0.9968	0.9956
0.45–0.55	1.128	1.217	0.9979	0.9971
0.5–0.5	0.9553	1.05	0.9985	0.9978
0.55–0.45	0.9976	1.071	0.998	0.9974
0.6–0.4	0.9814	1.06	0.9983	0.9977
0.65–0.35	0.7809	0.8686	0.999	0.9985
0.7–0.3	0.8658	0.9145	0.9989	0.9985
0.75–0.25	0.7468	0.8009	0.9989	0.9987
0.8–0.2	0.9185	0.9629	0.9988	0.9984
0.85–0.15	0.6881	0.7183	0.9992	0.999
0.9–0.1	0.6765	0.7212	0.9991	0.9989
0.95–0.05	0.724	0.7601	0.9992	0.999
	Run–1	Run–2	Run–1	Run–2

Figure 3.3: **Median RMSE and Pearson correlation values of the down-sample analysis.**

3.11.2 Supplementary Note 1. Goodness-of-fit of the canonical DNA analysis.

We first evaluated whether the proposed framework could generalize information to nucleotides that were not present in the entire training data. We thus trained the framework using the DNA 6mers that do not contain each nucleotide (base-dropout, See section 3.6). Such training sets retain 18% of the total 6mers. Therefore we used the 0.2-0.8 train-test split as the baseline null model. As shown in Figure 3.1B and Supplementary Figure 3.6, base-dropouts significantly decreased the prediction power compared to the baseline null model. Such a result suggests that the four DNA nucleotides provide orthogonal information during training. In addition, the prediction power was more impaired by excluding T and C, which suggests that the four nucleotides have unequal importance.

We also evaluated the framework's generalizability to nucleotides that were not present in particular DNA 6mer positions (position-dropout, see section 3.6). Such position-dropout retains 75% of total 6mers for training, so we used the 0.75-0.25 train-test split as the baseline null model. As shown in Figure 3.1B and Supplementary Figure 3.6, in general the prediction power was significantly impaired by excluding T and C, consistent with the nucleotide importance evaluated by base-dropout analysis. Meanwhile, dropouts in 3rd and 4th positions contributed the most to prediction power decrease, followed by 2nd and 5th positions. The positional importance suggested here was further consistent with [41].

We further explored whether full DNA 6mer models can be generalized by

combining complementary base-dropout training sets, e.g. G-dropout and C-dropout that contains instances of C and G containing kmers, but no kmers containing both C and G (noted as G-C, see section 3.6). Such training sets contain 34% of total DNA 6mers, thus 0.35-0.65 train-test split was used as the baseline null model. As shown in Figure 3.1B and Supplementary Figure 3.6, in general the prediction power was comparable with the baseline null model, suggesting the validity of such model combination.

3.11.3 Supplementary Note 2. Goodness-of-fit of the canonical RNA analysis.

Following the same pipeline as in DNA, down-sample, base-dropout, position-dropout and model combination analyses were also performed under RNA context. Meanwhile, RMSE and r were also used for prediction power evaluation for RNA analysis (see section 3.6). Compared to DNA analysis, two major differences were observed. First, for RNA analysis as shown in Supplementary Figure 3.7, in general the prediction power was lower. For instance random down-sample analysis with 0.95-0.05 train-test split (best-performing random down-sample group), average RMSE values were 0.8 and 2.4 for DNA and RNA, respectively. We speculate that such prediction power difference was majorly caused by the number of training data points. As mentioned in the main text, with the currently most prevalent Oxford Nanopore Technologies R9.4 nanopore sequencing chemistry, DNA is modeled with in total 4096 6mers. On the other hand, RNA is modeled with in total 1024 5mers, only

25% as opposed to the DNA scenario. Such fewer possible training data points might strongly compromise the prediction power of our framework. However, once trained with a similar amount of kmers, the RNA architecture could yield comparable prediction power. For instance, the RNA 0.95-0.05 (972 training kmers) and DNA 0.25-0.75 (1024 training kmers) train-test splits yielded comparable performance. Such a result suggested the validity of our proposed architecture.

The other major difference between DNA and RNA analysis is, the four canonical DNA bases (A, T, G, C) are "orthogonal" to each other (Figure 3.1B and Supplementary Figure 3.6). In contrast, base-dropout will not cause statistically significant decrease in prediction power, suggesting that the four canonical RNA bases (A, U, G, C) can complement each other in terms of their chemical properties (Supplementary Figure 3.7). Here "orthogonal" means base-dropouts will significantly decrease the prediction power as opposed to the corresponding random down-sample null model. Notably, such an "orthogonality effect" was particularly strong for T and C. We speculate that such a difference can be explained by the additional methyl group in T. Among the four DNA canonical bases, methyl only appears in T, thus cannot be compensated by other combining A, G and C. Similarly, considering such methyl is encoded with pyrimidine backbone (Figure 3.7A and Supplementary Figure 3.11), the representation of the other pyrimidine nucleobase, C is also affected. Thus T and C were more "orthogonal" compared to A and G. As for RNA, without the additional methyl, the four canonical RNA bases complement each other in terms of their chemical structures. Further, the chemical information generalization

among bases guarantees the proper representation of RNA 5mers under base-dropout scenario, thus producing statistical insignificant prediction power as opposed to the corresponding null model.

3.11.4 Supplementary Note 3. Benchmarking human genome C/5mC-status predictive analysis with the megalodon algorithm.

As described in Supplementary Figure 3.10, the recently released megalodon algorithm (https://github.com/nanoporetech/megalodon) could drastically increase the accuracy for NA12878 cell line C/5mC status prediction, therefore could be used as an additional ground truth for benchmarking our "human genome C/5mC-status predictive analysis". Compared to the ground truth established from bisulfite sequencing datasets [44], using megalodon predictions as ground truth has two prominent advantages: 1) the megalodon algorithm yields per-read per-site predictions, which could provide higher resolution as opposed to bisulfite sequencing ground truth when evaluating predictive accuracy. 2) The bisulfite sequencing ground truth was established from separate experiments thus potential biological/technical batch effects could be the concerns. We therefore adopted the megalodon predictions as the additional per-read per-site C/5mC status ground truth. We yielded the following predictive accuracy, which is comparable with the result in Supplementary Figure 3.10.

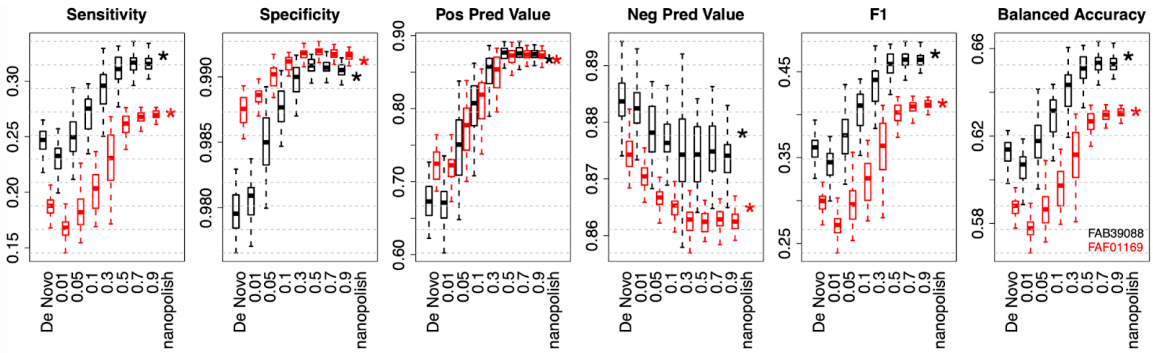

Figure 3.4:

However, please note that the undermining limitation of the megalodon ground truth is the reliability of the megalodon predictions. Considering the algorithm has not been peer reviewed, as well as there are no available results benchmarking the performance under various scenarios, e.g. biological/technical replicates, different species, etc., we adopted the bisulfite sequencing ground truth throughout the study.

3.11.5 Supplementary Note 4. Building empirical kmer models.

For nanopore sequencing chemistries after R9.4, the "official" ONT kmer models are no longer available. To solve such a problem, users could build empirical kmer models by the nanopore sequencing of synthesized control oligos. The first question regarding building empirical kmer models is determining the effective kmer length (k). This could be done by following the procedures reported in our previous study [41]. The second question would be to determine the sequences of synthesized control oligos. Specifically, users need to decide 1) depth of nanopore sequencing,

e.g. number of reads, and 2) kmers to be covered, e.g. the "minimal ideal kmer set". These need to be done to 1) make sure full kmer models could be recapitulated from partial training using the deep learning architecture, and 2) save oligo synthesis and sequencing cost. Please note that it's crucially important that the control oligos cover all possible kmers. Otherwise, modification calling might be compromised. As shown in the following figure, we quantified the predictive accuracy of C/5mC status in a sequence context-specific manner. Specifically, with the same set of nanopore sequencing reads described in "human genome C/5mC-status predictive analysis" in section 3.6, we quantified the balanced accuracy under **C**A, **C**T and **C**G contexts (**C**C motif is rare in human genome and was not covered by the selected reads). As shown in the following figure, in general the predictive power of DNA kmer models inferred by our deep learning architecture (De Novo, 0.01, 0.05, 0.1, 0.3, 0.5, 0.7, 0.9) was comparable to the control nanopolish model. Meanwhile, kmer model-based signalAlign algorithm and deep learning-based megalodon algorithm yielded comparable predictive accuracy for CG context. However, signalAlign predictive accuracy was significantly compromised under CA and CT contexts. Moreover, predictive accuracy for CA and CT motifs was in general lower compared to the CG motif, even with the control nanopolish model. This is because the nanopolish model was constructed with the human genome, in which CA and CT motifs are less prevalent, and CC motif is rare. Therefore, CA, CT and CG contexts are less confidently represented in the nanopolish model.

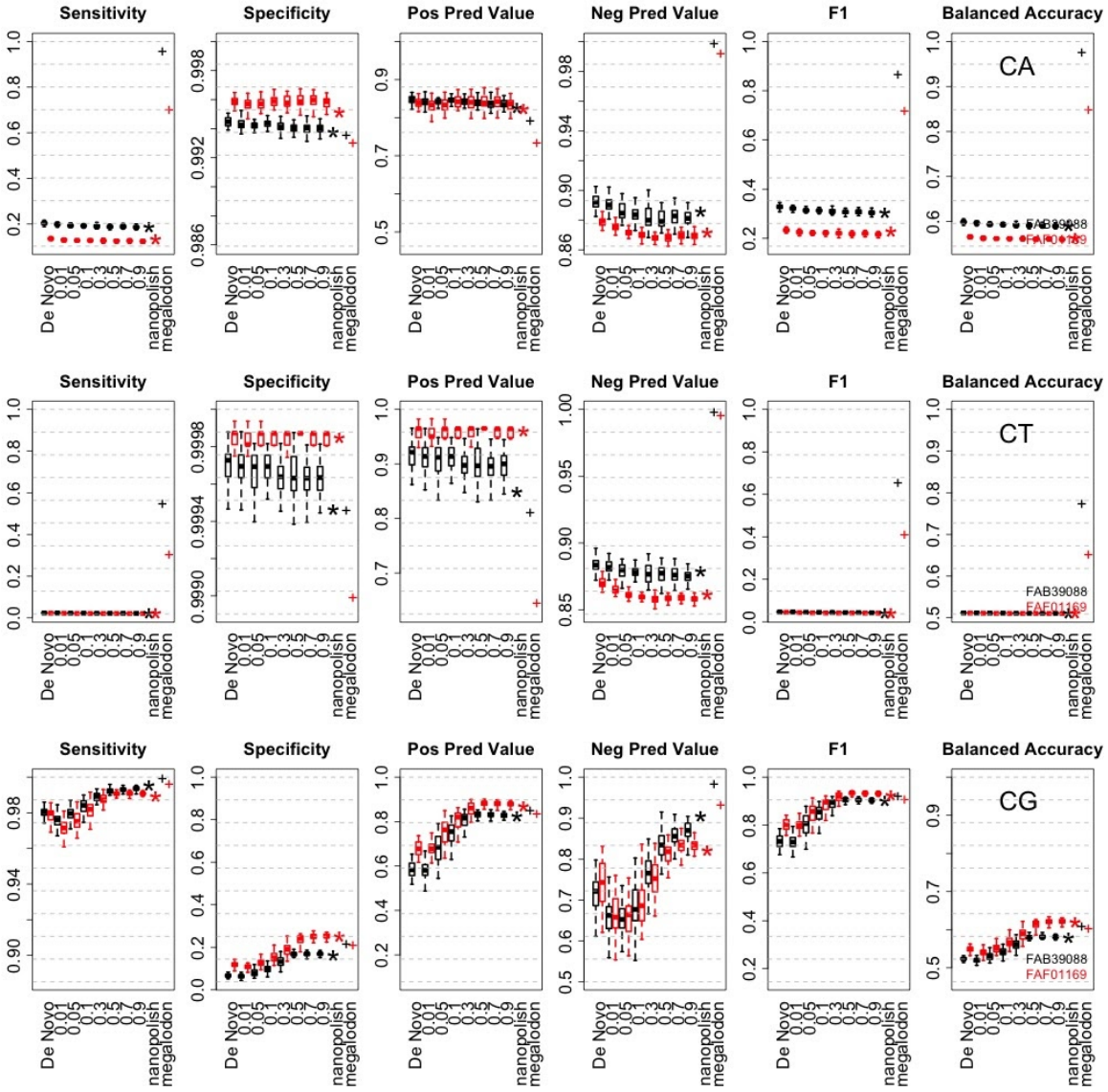

Figure 3.5:

As for determining the proper depth of nanopore sequencing, we find it to be unnecessary: based on previous work, we believe that only modest sequencing is needed for building robust kmer models. Specifically, as reported in [96], the authors performed two parallel nanopore sequencing of synthesized RNA molecules,

each with only a single MinION flowcell, further constructed corresponding 6mA, U, G, C RNA 5mer models (two technical replicates). It was demonstrated that the two technical replicate RNA 5mer models are highly comparable, suggesting that robust kmer models could be generated from relatively small scale nanopore sequencing experiments. As for finding the "minimal ideal kmer set", although it might be a valuable idea to pursue, we find it to be less feasible. The reason being that the stochastic deep learning framework training process will introduce stochasticity in the prediction performance. That being said, the best prediction performance achieved by a certain training set could just be an effect of stochasticity, rather than the actual kmer composition. We also find the idea of finding the "minimal ideal kmer set" to be unnecessary, as kmers could be covered with relatively short sequences. As reported in the above-mentioned study [96], all possible 6mA, U, G, C RNA 5mers could be covered with 4 sequences with average length ∼2.5kb (2329, 2543, 2678, and 2795bp, respectively). Taken together, we believe that robust full kmer models could be built with affordable cost, following procedures reported in [96]. The last question would be to generate kmer models from nanopore sequencing readouts on the synthesized control oligos. We provide detailed procedures in section 3.6.

3.11.6 Supplementary Figures

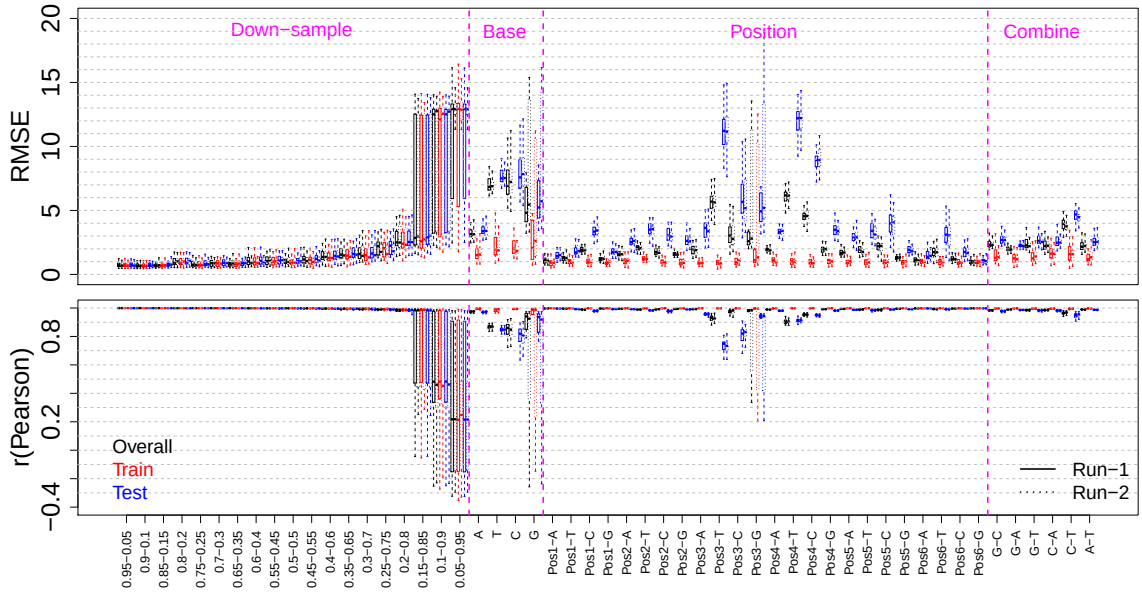

Figure 3.6: **Goodness-of-fit of the canonical DNA analysis.** Root Mean Square Error (RMSE) and Pearson correlation (r) values of DNA down-sample, base-dropout, position-dropout and model combination analyses. Run-1 (solid boxes) and Run-2 (dashed boxes) refer to two independent replicates. RMSE and r values for the predictions of all DNA 6mers (Overall), DNA 6mers in training set only (Train) and DNA 6mers in test set only (Test) were marked as black, red and blue, respectively. The median, minimum/maximum (excluding outliers) and first/third quartile values were shown by the boxplots. See section 3.6 for details.

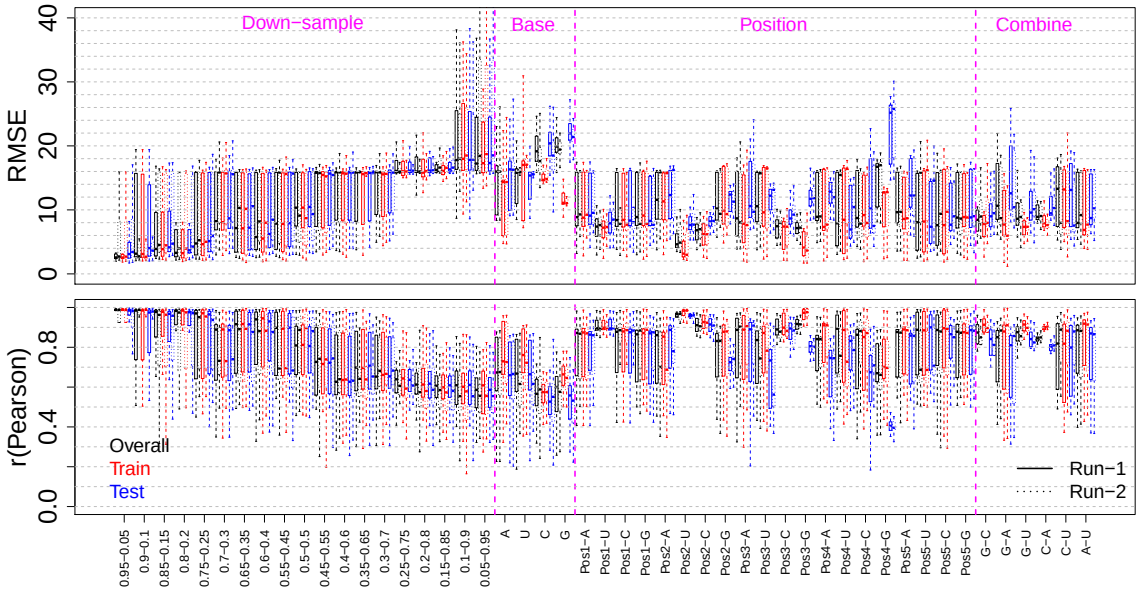

Figure 3.7: **Goodness-of-fit of the canonical RNA analysis.** Root Mean Square Error (RMSE) and Pearson correlation (r) values of DNA down-sample, base-dropout, position-dropout and model combination analyses. Run-1 (solid boxes) and Run-2 (dashed boxes) refer to two independent replicates. RMSE and r values for the predictions of all DNA 6mers (Overall), DNA 6mers in training set only (Train) and DNA 6mers in test set only (Test) were marked as black, red and blue, respectively. The median, minimum/maximum (excluding outliers) and first/third quartile values were shown by the boxplots. See section 3.6 for details.

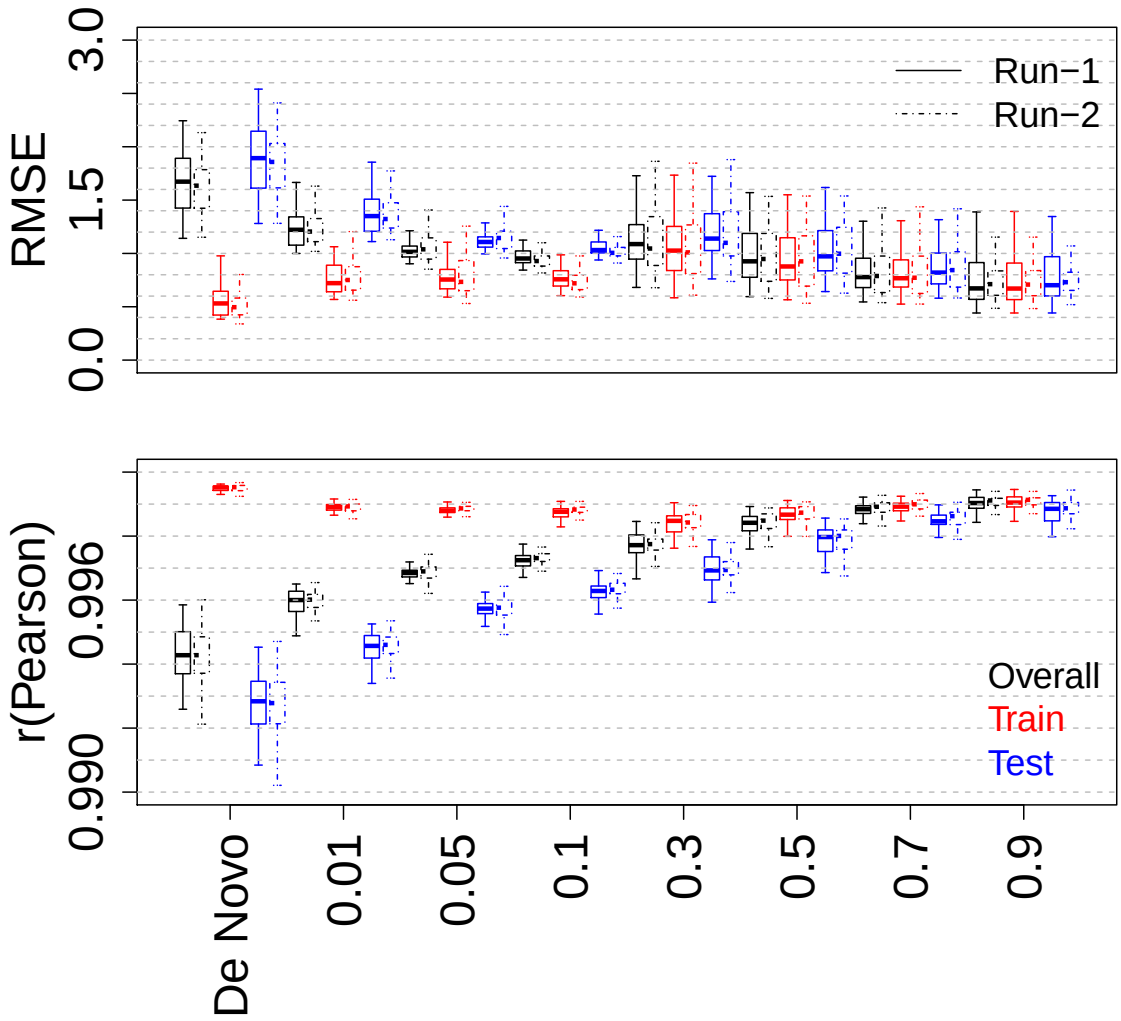

Figure 3.8: **Goodness-of-fit of the DNA 5mC analysis.** Root Mean Square Error (RMSE) and Pearson correlation (r) values of DNA 5mC-imputation analysis. These values were quantified against the nanopolish model [152, 101]. Run-1 (solid boxes) and Run-2 (dashed boxes) refer to two independent replicates. RMSE and r values for the predictions of all DNA 6mers (Overall), DNA 6mers in training set only (Train) and DNA 6mers in test set only (Test) were marked as black, red and blue, respectively. The median, minimum/maximum (excluding outliers) and first/third quartile values were shown by the boxplots. See section 3.6 for details.

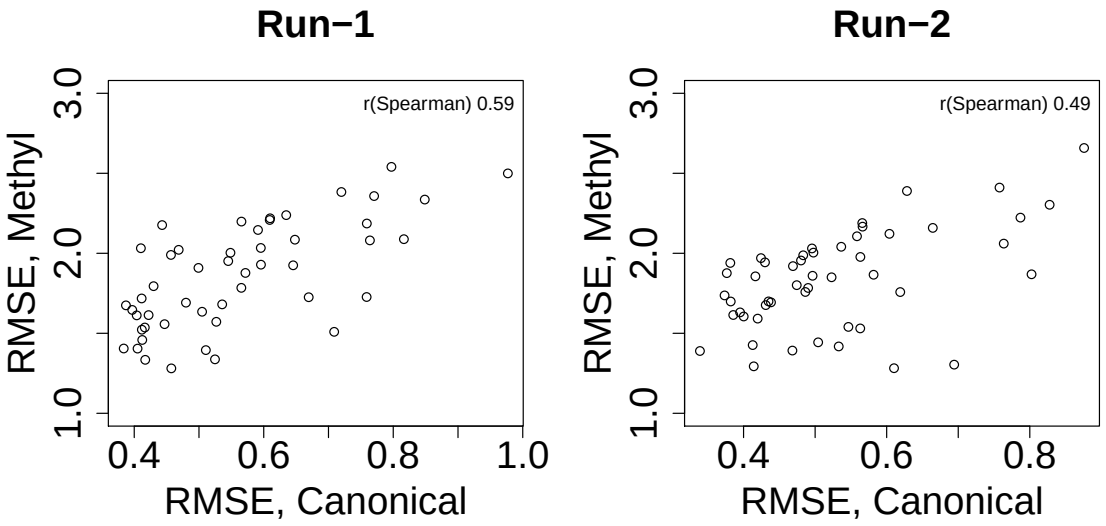

Figure 3.9: **RMSE correlation in DNA 5mC-de novo analysis.** For both Run-1 and Run-2, RMSE values obtained from canonical and 5mC-containing DNA 6mers were compared. Dots on the scatter-plots represent training-prediction repeats.

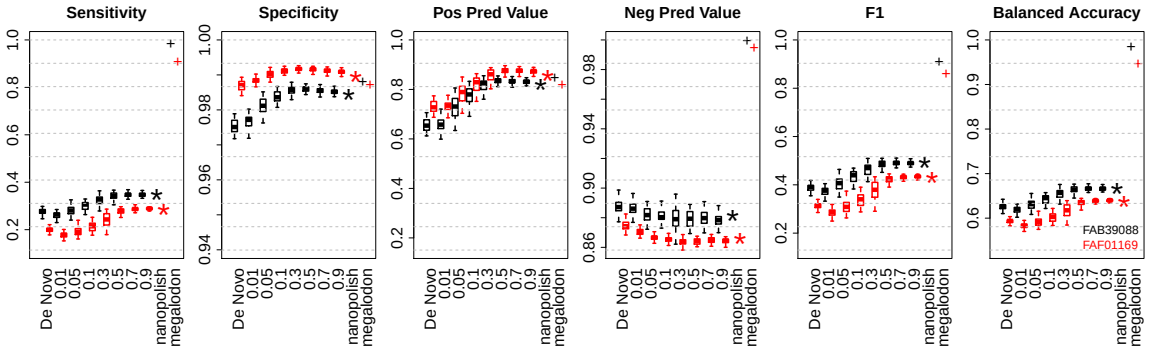

Figure 3.10: **Predictive accuracy of DNA 5mC analysis.** Predictive accuracy was quantified by true positive rate (TPR), true negative rate (TNR), positive predictive value (PPV), negative predictive value (NPV), F1-score (F1) and balanced accuracy (BA). FAB39088 (black) and FAF01169 (red) refer to two independent NA12878 cell line native genomic DNA nanopore sequencing datasets [73]. Nanopolish refers to predictive analysis using the nanopolish model [152, 101]. Megalodon refers to predictive analysis performed using the deep learning-based megalodon algorithm `https://github.com/nanoporetech/megalodon`. The median, minimum/maximum (excluding outliers) and first/third quartile values were shown by the boxplots. See section 3.6 for details.

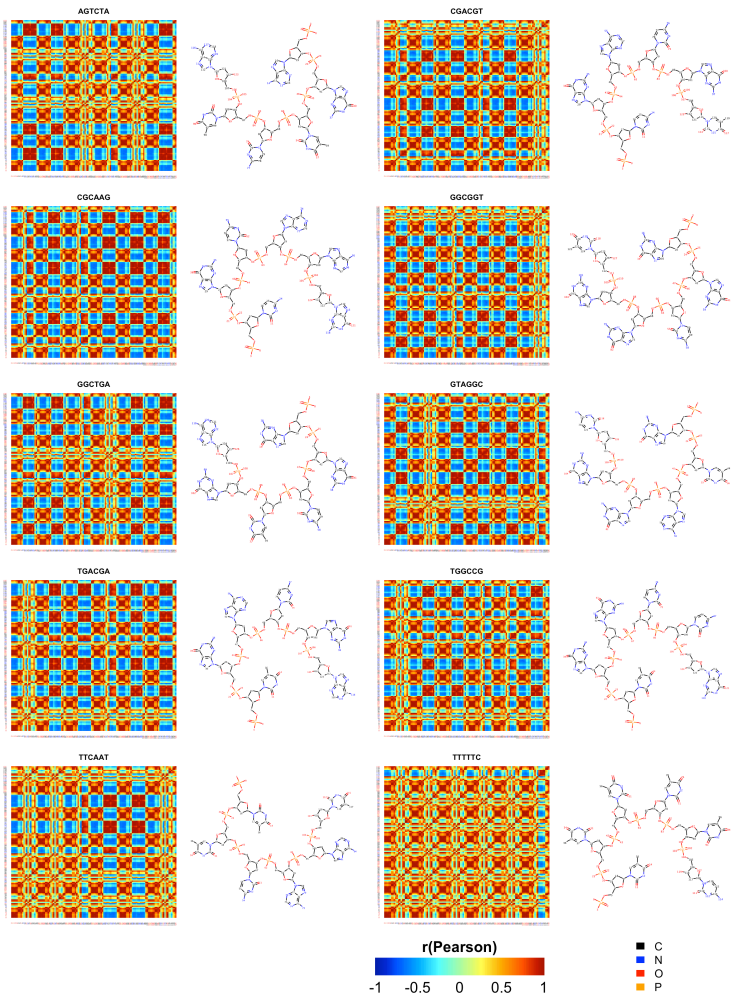

Figure 3.11: **Visualizing canonical DNA 6mer atom similarity matrices.** Without losing generality, we visualized the atom similarity matrices of 10 random canonical DNA 6mers. Similarity matrices were calculated using the Pearson correlation of the state vectors outputted by the final GCN layers. Corresponding chemical structures of analyzed DNA 6mers were shown side-by-side of the similarity matrices, based on which atoms were numbered and colored. Carbon, nitrogen, oxygen and phosphorus were colored as black, blue, red and orange, respectively.

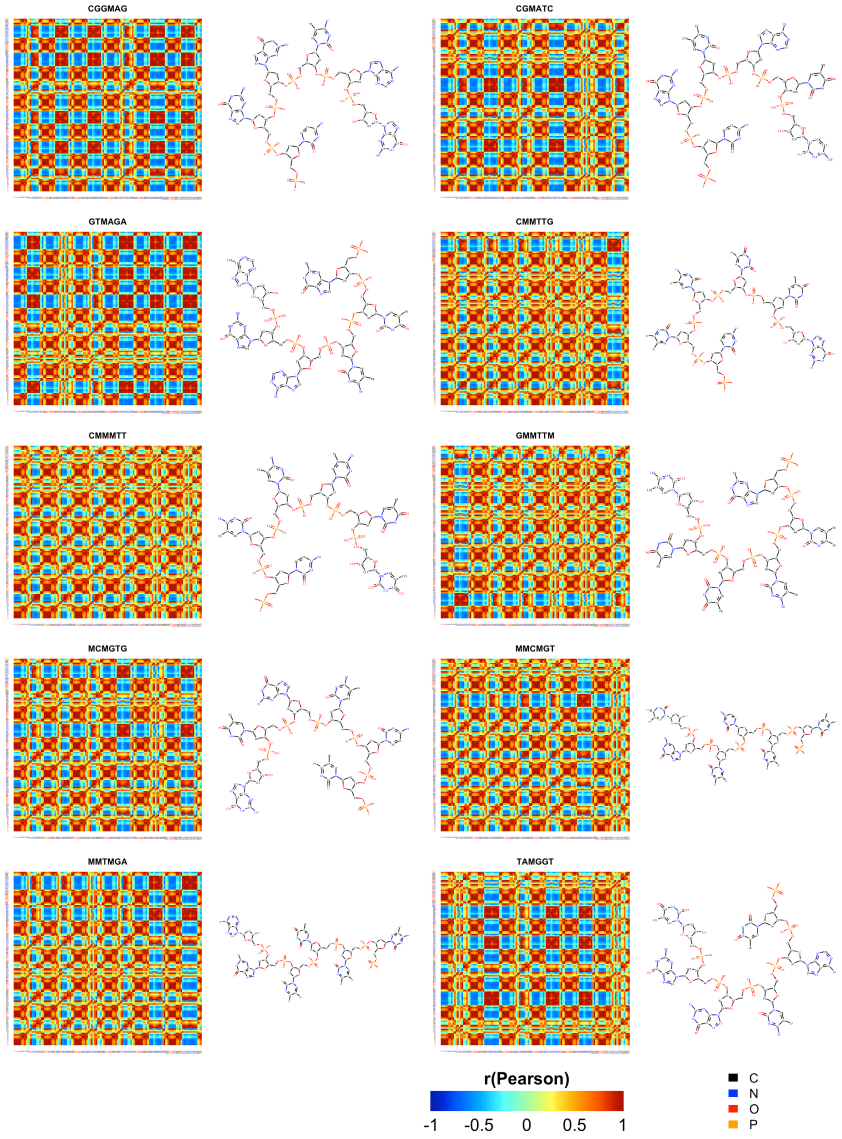

Figure 3.12: **Visualizing 5mC-containing DNA 6mer atom similarity matrices.** Without losing generality, we visualized the atom similarity matrices of 10 random 5mC-containing DNA 6mers. 5mC was abbreviated as M for simplicity. Similarity matrices were calculated using the Pearson correlation of the state vectors outputted by the final GCN layers. Corresponding chemical structures of analyzed DNA 6mers were shown side-by-side of the similarity matrices, based on which atoms were numbered and colored. Carbon, nitrogen, oxygen and phosphorus were colored as black, blue, red and orange, respectively.

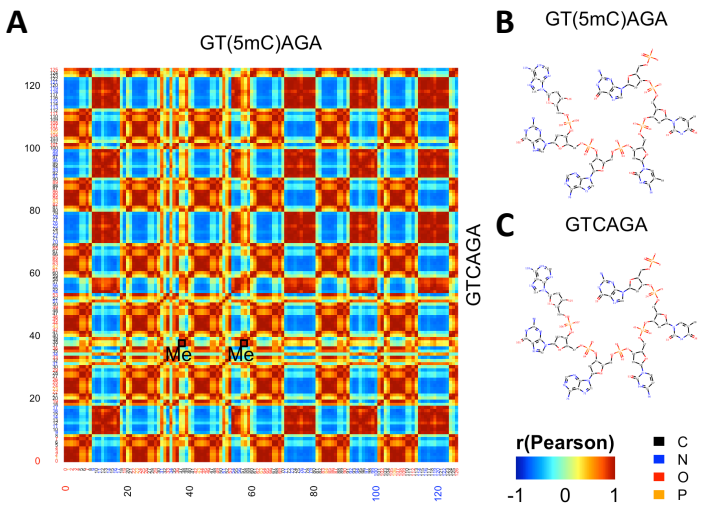

Figure 3.13: **Visualizing inter-kmer atom similarity matrices.** Without losing generality, we analyzed the inter-kmer atom similarity between modified DNA 6mer GT(5mC)AGA and corresponding canonical counterpart GTCAGA. (A) Visualizing the inter-kmer similarity matrix, which was calculated using the Pearson correlation of the state vectors outputted by the final GCN layers. (B) The chemical structure of DNA 6mer GT(5mC)AGA. (C) The chemical structure of DNA 6mer GTCAGA. Based on chemical structures in (B) and (C) atoms were numbered and colored. Carbon, nitrogen, oxygen and phosphorus were colored as black, blue, red and orange, respectively.

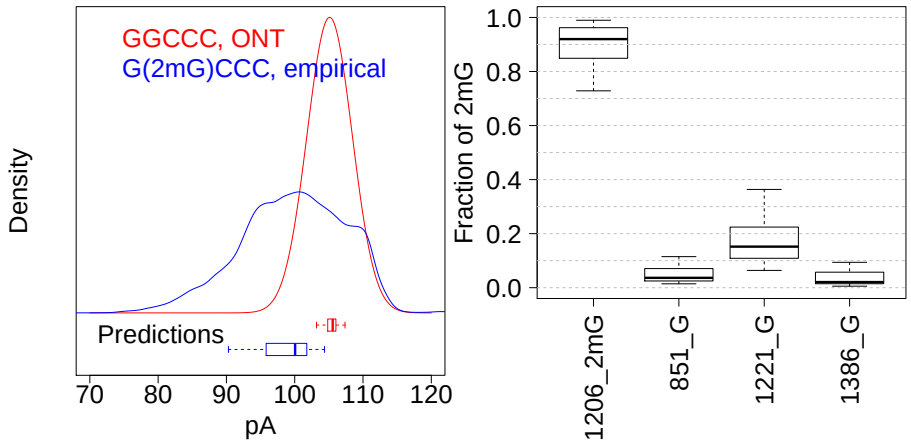

Figure 3.14: **RNA 2mG analysis.**(A) The empirical ionic current signal distribution of RNA 5mer G(2mG)CCC, as well as the ONT ionic current signal distribution of pairing canonical RNA 5mer GCCCC were visualized in red and blue curves, respectively. Characteristic ionic current signals of G(2mG)CCC and GGCCC predicted by the deep learning framework were visualized in red and blue boxes, respectively. (B) For E.coli 16S rRNA transcript J01859.1 position 1206, the fraction of modified (2mG) reads determined by signalAlign with predicted RNA 5mer ionic current signals was quantified. For boxplots in (A) and (B), the median, minimum/maximum (excluding outliers) and first/third quartile across the 50 prediction repeats were shown.

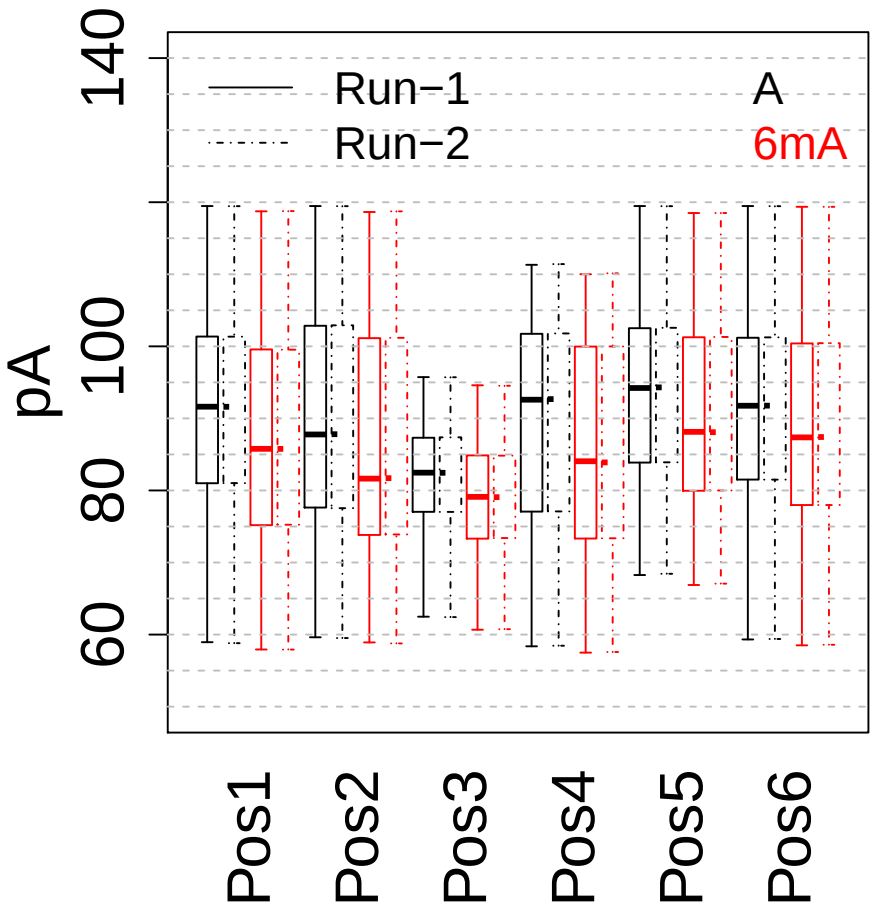

Figure 3.15: **Chemical group stack analysis.**Framework trained with all possible canonical DNA 6mers was used to predict 6mA-containing 6mers. 6mA-containing kmers were grouped by the positions of 6mAs. Signal distributions of 6mA-containing kmers and their canonical counterparts were shown in the boxplot. The median, minimum/maximum (excluding outliers) and first/third quartile values were shown by the boxplots. See section 3.6 for details.

Chapter 4

Single-molecule modification profiling of Saccharomyces cerevisiae ribosomal RNA reveals concerted modification at functional locations in the ribosome

Andrew D. Bailey IV[1,6,*], Jason Talkish [2,6] Hongxu Ding[1,3], Haller Igel[2], Alejandra Durán[4], Shreya Mantripragada[5], Benedict Paten[1,*], and

Manuel Ares, Jr.[2,*]

[1]Department of Biomolecular Engineering and Genomics Institute, University of California, Santa Cruz, Santa Cruz, California, USA. [2]RNA Center and Department of Molecular, Cell and Developmental Biology, UC Santa Cruz, Santa Cruz, California, USA. [3]Department of Pharmacy Practice and Science, College of Pharmacy, University of Arizona, Tucson, Arizona USA. [4]Colegio Santa Francisca Romana, Bogotá, Colombia. [5]Monta Vista High School, Cupertino, California. [6]These authors contributed equally to this work. [*]Correspondence should be addressed to A.B. (andbaile@ucsc.edu), B.P. (bpaten@ucsc.edu) or M.A. (ares@ucsc.edu).

4.1 Abstract

Nucleotides in RNA and DNA are subject to numerous enzymatic activities that chemically modify them, altering their functional characteristics. Eukaryotic ribosomal RNA is modified at more than 100 locations, particularly at highly conserved and functionally important nucleotides. During ribosome biogenesis, modifications are added at various stages of assembly. The existence of differently modified classes of ribosomes is unknown because no method for simultaneously evaluating modification status at all sites within a single rRNA molecule is available. Using a combination of yeast genetics and nanopore direct RNA sequencing, we have developed a reliable method to track the modification status of single rRNA molecules at 37 sites in 18S rRNA and 73 sites in 25S rRNA. We use our method to characterize pat-

terns of modification heterogeneity and identify concerted modification of nucleotides found near functional centers of the ribosome. Distinct undermodified subpopulations of rRNAs accumulate when ribosome biogenesis is compromised by loss of Dbp3 or Prp43-related RNA helicase function. Modification profiles are surprisingly resistant to change in response to many genetic and environmental conditions that affect translation, ribosome biogenesis, and pre-mRNA splicing. The ability to capture complete modification profiles for RNAs at single-molecule resolution will provide new insights into the roles of nucleotide modifications in RNA function.

4.2 Introduction

In addition to the four standard nucleotides, there are more than 160 distinctly modified ribonucleotides and more than 50 distinctly modified deoxyribonucleotides found in the RNA and DNA of cells[74, 15, 155]. Many of these modified nucleotides provide extra regulatory information and are crucial for cell function. Irregular DNA methylation patterns are linked to several cancers, neurological disorders and autoimmune diseases[130, 133]. RNA modifications have been linked to the development of cognitive functions, neurological defects, breast cancer, genetic birth defects and diabetes [185, 6, 70, 37, 10, 38]. In ribosomal RNA (rRNA), extensive and highly conserved modifications are vital for correct ribosome structure and function [128]. Modifications on rRNA have been generally considered to be constitutive in support of fine tuning function [76] rather than mediating specific regulatory changes

in ribosome function. However, the fraction of rRNA molecules modified at many specific positions can change in association with changes in environment, disease and during development [163, 14, 153]. Recently modification status at an adenosine near the 3' end of yeast 18S rRNA has been implicated in controlling sulfur metabolism [98]. It seems possible that subtle alterations in modification status at particular locations in the ribosome could be used to control translation by creating functional heterogeneity in the cell's pool of ribosomes.

One technical challenge in analyzing the effect of modification on the function of rRNA is that it has not yet been possible to capture modification states at all positions of single RNA molecules. Traditional modification detection approaches examine ensembles of molecules and estimate the extent of modification at individual sites independently. For example, non-sequencing based techniques such as liquid chromatography-tandem mass spectrometry (LC-MS/MS) and cryogenic electron microscopy (cryo-EM) can identify the presence of many modified nucleotides [114, 136, 163, 23, 164]. Some methods such as immunoprecipitation-seq or mismatch-seq, aggregate information from several reads to detect modifications at a specific site [65, 134, 147, 184, 36, 154, 42, 43], but do not capture associations between modification status at distant sites in large RNA molecules. Other approaches such as bisulfite-seq, pseudouridine-seq, and RiboMeth-seq [53, 146, 14, 150] are highly specific for a single modified nucleotide, but also require fragmentation, preventing capture of modification status at multiple distant sites in an RNA. Such whole molecule information would be necessary for assessing the relationship between function and

modification status of individual ribosomal sub-units.

New advances in direct single-molecule sequencing of RNA using nanopore technology may circumvent many of these limitations. Direct nanopore sequencing of full-length RNA molecules[35, 57] can report modification status across entire RNA molecules without chemical treatment or amplification steps. Modified nucleotides produce changes in electrical current that are distinct to canonical nucleotides, permitting modification detection algorithms to identify modifications in both DNA and RNA. Given enough training data, basecalling algorithms like Guppy can predict modifications directly from the signal along with canonical nucleotides [181]. However, training data for most modifications is limited, and thus many detection algorithms rely on aligning basecalled reads to a reference sequence and comparing modified signal to canonical signal [54]. Current signals can be modelled using secondary features like quality scores and base miscalls [98, 7] or directly using the underlying signal [108, 135, 152]. To model the underlying signal, most algorithms summarize the stream of signal into segments (events), align events to subsequences (kmers) of a reference, and compare aligned events to models of canonical or modified nucleotides. Thus far, no currently available method captures combinations of distinctly modified nucleotides at multiple distant sites in RNA.

Here we demonstrate accurate, single-molecule modification profiling of 13 distinct types of modified nucleotide at 110 positions across full transcripts of 18S and 25S rRNA from *S. cerevisiae*. We preserve long-range associations between modification status at distant positions on single RNA molecules, allowing us to identify

highly correlated positions and explore heterogeneity in ribosomal RNA modification. Clustering analysis identifies distinct populations of differently modified ribosomes in wild type yeast as well as yeast depleted for various components of the modification machinery, providing evidence that modification status of groups of nucleotides are established in a concerted manner, especially at functional centers of the ribosome. Further application of single-molecule modification profiling will enable dissection of the contributions of nucleotide modification to the function of large RNAs.

4.3 Results

4.3.1 Profiling rRNA modifications at single-molecule resolution

To investigate the overall modification status of yeast rRNA on a single-molecule level, we sought to use the nanopore current traces from Oxford Nanopore flow cells (see section 4.5) of complete rRNA transcripts to capture modification status at each modified position along individual molecules. To create these single-molecule profiles, we trained signalAlign [135] by modeling wild type rRNA reads as modified and in vitro transcribed (IVT) reads as unmodified to detect all 110 annotated modifications in S. cerevisiae 18S and 25S rRNA [163, 108](See section 4.5, Supplemental Fig. 4.7). For each rRNA read, the model estimates the probability of modification at each annotated modified site and outputs a list of modification probabilities for the entire read, regardless of modification type (Supplemental Fig. 4.8A). We anticipated that rRNA modification profiles obtained from different conditions could

be searched for subpopulations of distinctly modified ribosomes using hierarchical clustering, or to quantify the extent of correlation between modified positions under different conditions of ribosome function.

To test the ability of the trained model to capture single-molecule modification profiles in yeast catastrophically depleted of rRNA modifications, we isolated RNA under conditions in which either of two classes of snoRNA-guided modifications is blocked. In S. cerevisiae, 34 of the 37 18S and 66 of the 73 25S rRNA annotated modifications are guided by the C/D box and H/ACA box snoRNPs, respectively [167, 188, 127]. To ablate these modifications en masse we used strains in which expression of Nop58 (core component of C/D snoRNPs) or Cbf5 (H/ACA snoRNP pseudouridylase) are under control of a GAL1 promoter [178, 85]. Thus, in cells shifted to glucose medium, Nop58 (or Cbf5) expression will be repressed, leading to depletion of C/D box (or H/ACA box) snoRNPs, and widespread loss of modification [84, 85]. As expected, single-molecule modification profiles produced by our model reveal accumulation of large numbers of rRNA molecules lacking most 2'O-methyl (Nm) (Nop58-depleted) or pseudouridine (Ψ) (Cbf5-depleted) modifications at snoRNA-guided positions in 18S (Supplemental Fig. 4.9) or 25S rRNA (Fig. 4.1).

To examine subpopulations of modified rRNA molecules in these cells, we pooled single-molecule modification profiles from IVT, wild type, and cells depleted of Nop58 or Cbf5, and performed hierarchical clustering. We observe clear separation of 4 clusters representing wild type rRNA, IVTs, and molecules arising from the Nop58 or Cbf5-depleted cells, respectively (Fig. 4.1A, Supplemental Fig. 4.9A).

Dimension reduction UMAP visualization [110] of 18S and 25S rRNA modification profiles confirms the separation of these distinct molecular populations (Fig. 4.1C and Supplemental Fig. 4.9C). Comparing the two clusters derived from the snoRNP-depleted cells (Fig 4.1A and C, clusters 1 and 3) suggests that 2'O methylation is largely independent from pseudouridylation. Some molecules from snoRNP-depleted cells appear in cluster 2 with wild type rRNA (Fig. 4.1A and C), more often for 18S rRNA than for 25S rRNA, consistent with the finding that 18S rRNA accumulation is more sensitive to snoRNP depletion than is 25S rRNA [84, 85](Supplemental Fig. 4.9D). We conclude that clustering of single-molecule rRNA modification profiles reveals two large but distinct classes of undermodified rRNA molecules induced by depletion of one or the other class of snoRNPs.

To test for correlation of modification at two different sites, we measure the change in Spearman correlation for each pair of modified sites in an experimental sample as compared to their correlation in wild type profiles (Fig. 4.1D and E, Supplemental Fig. 4.9E-F, Supplemental Table S1). We then test if the Spearman correlation in the mutant is significantly different from that in wild type, correcting for multiple testing via the Benjamani-Hochberg procedure (two sided t-test and Fisher z-transform test, see section 4.5, Supplemental Table S1). This test indicates that changes in correlation between each of the snoRNA-guided pseudouridine positions in the Cbf5-depleted cells are highly significant (p-value=5.5e-05, Brown's method). Likewise, changes in correlation between each of the snoRNA guided 2'O-methyl positions in the Nop58-depleted cells are highly significant (Fig. 4.1D-E and Supplemental

Fig. 4.8C-D, p-value=1.5e-16, Brown's method). A comparison of the pairwise tests for all combinations of modified positions (Fig. 4.1D and 4.1E) confirms that to a large extent 2'O-methylation is independent of pseudouridylation in yeast rRNA.

Although the majority of molecules in either depletion experiment lacked the expected modifications and clustered together (Fig. 4.1A-B and Supplemental Fig. 4.9A-B), several subpopulations of molecules displayed concerted patterns of modification loss. The sites of methylation guided by the C/D box snoRNA U24 (Cm1437, Am1449, Gm1450 within the nascent polypeptide exit tunnel (PET) in 25S rRNA) appear to be modified together, or not, in a concerted fashion. Almost half of the molecules from the Nop58 depletion remain methylated at all three sites, splitting cluster 3 into two nodes (Fig. 4.1A). Furthermore a fraction of molecules in cluster 1 (formed by depletion of Cbf5) are also unmodified at all three sites, suggesting that concerted methylation at these positions may be partly dependent on pseudouridine modification elsewhere. Particularly striking is the highly concerted modification at 25S rRNA positions Um2921, Gm2922, and Ψ2923 in the peptidyl transfer center (PTC). These appear refractory to loss of modification in both depletion experiments, remaining modified on a large fraction of molecules otherwise lacking multiple other modifications (Fig. 4.1A). Modification of Um2921 is guided by C/D box snoRNA snR52, and Gm2922 is modified by the non-snoRNP methyltransferase Sbp1, which can also methylate Um2921 in the absence of snR52 [86]. The extremely low numbers of ribosomes unmodified at these important positions suggests their modification may be essential for rRNA stability.

We also observe correlation between the two N4-acetylcytidines (ac4C1280 and ac4C1773) in 18S and the 2'O-methyl positions in the Nop58 depletion (p-value = 2.3e-05, Brown's method) (Supplemental Fig. 4.9A and E). N4-acetylcytidine modification depends on the C/D box snoRNAs snR4 and snR45, which do not guide methylation, but instead bring the cytidine acetylase Kre33 to positions C1280 and C1773, respectively [171, 149]. These atypical C/D box snRNAs also require Nop58, explaining the coordinate loss of cytidine acetylation and 2'O-methylation. We confirmed that our model recognizes ac4C modified sites by knocking out each snoRNA (Supplemental Fig. 4.11A, B). The N1-methyl-N3-aminocarboxypropyl-pseudouridine (m1acp3Φ1191) residue in 18S is significantly correlated with pseudouridine positions in the Cbf5 depletion (p-value = 5.4e-07, Brown's method, Supplemental Fig. 4.9A and F), as expected given that snR35-guided pseudouridylation of U1191 is the first step to generate this complex modification[18]. We conclude that single molecule modification profiling allows identification of subpopulations of individual rRNA molecules and captures modification correlation between chemically distinct modifications.

The model we have developed assesses signal for 5 sequential overlapping 5-mers centered on the annotated position of interest, and when there is tight clustering of modification sites, the evaluation of modification at one position may be influenced by the modification status at the nearby position. In both locations above, we were concerned that the co-modification patterns we observe might arise from limited training of the model to resolve closely spaced partly modified regions. To

examine whether the modification patterns observed match the patterns found in the underlying signal, we clustered single molecule signal event means covering a subset of modifications (see section 4.5) [41]. This test reveals that, in general, patterns of modifications found in the modification profile clustering match with the underlying event means clustering (Supplemental Fig. 4.10B/D). However, upon close inspection of three highly concerted modifications in the PTC (Supplemental Fig. 4.10E), event mean clustering reveals a slight partitioning of depletion reads. Given that both Um2921 and Gm2922 are expected to be present in both depletion experiments and wild type, the slight variation in signal found in the clustering of event means indicates that the true level of Φ2923 in the Cbf5 depletion is lower than estimated (Supplemental Note 4.11.1).

4.3.2 Resolving subpopulations of ribosomes that differ at a single modified site

The global loss of modification by depletion of snoRNPs creates catastrophically undermodified rRNA molecules that are easily distinguished by profiling. To test the ability of the method to resolve classes of ribosomes with modification profiles that differ at one or a few sites, we first estimated variation in our wild type rRNA profiles arising from experimental noise, model uncertainty, or true variation in modification levels in wild type. We calculated the variance in the predictions for each annotated modification for three wild type replicates. Based on the largest variance (position 18S:562, 9%), we chose a conservative cutoff of a 10% change in modification

frequency as a sign that modification at a site was affected by a given experimental perturbation (mutation or treatment, Supplemental Fig. 4.11B). We also compare the predicted modification frequency at a given site in an experiment to its predicted frequency in wild type using a chi-square test (see section 4.5, Supplemental Table S2). We sequenced rRNAs from strains containing individual snoRNA knockouts (snR80, snR83, snR87, snR4, and snR45) expected to completely lack modification at one or a few annotated sites in each case. There are significant decreases in modification frequency at the appropriate site for each snoRNA knockout (p-values¡1e-04, chi-square test, Supplemental Fig. 4.11B) and snoRNA knockout kmer distributions match the model's canonical kmer distributions (Supplemental Fig. 4.12). This experiment confirms our ability to identify undermodifcation at single locations with high confidence.

To test if we could deconvolute a mixture of heterogeneously modified rRNAs, recognizing those missing just 1-2 modifications against a background of other differently modified ribosomes, we pooled equal amounts of total RNA from three snoRNA knockout strains (snR80, snR83 and snR87) and wild type, and acquired single molecule modification profiles from the mixed sample (Fig. 4.2D). The idea was to create a sample that might mimic a cellular population of heterogeneously modified ribosomes. As seen in Fig. 4.2A and C, hierarchical clustering of the profiles obtained from the reconstructed sample reveals four similarly sized main clusters of differently modified 18S rRNA (see section 4.5). In this experiment we would expect to see positive correlation changes between positions Φ1290 and Φ1415, since their

loss arises from a shared dependence on snR83, and this expectation is fulfilled (Fig. 4.2B, p-value= 6.4e-16, Fisher z-transform test). Furthermore, since loss of modification in this reconstruction occurs independently at M436, Φ759, and [Φ1290+Φ1415], we expect negative correlation changes between these pairs relative to wild type, and this is what we observe (Fig. 4.2B, p-value=3.5e-08, Brown's method). These significant changes in correlations between long range modifications demonstrate clear and accurate partitioning of known subpopulations of differentially modified rRNA.

4.3.3 Correlated modification at distant sites on rRNA from wild type yeast

Previous studies have shown that alternatively modified yeast rRNA leads to changes in translational patterns of specific mRNAs [143, 148, 98]. Analysis of the modification status of the ensemble of wild type rRNAs reveal positions that are nearly completely modified as well as others that are only partly modified (Supplemental Table S3). For most annotated positions, our estimates largely agree with previous efforts to quantify percent modification in total yeast rRNA (Supplemental Fig. 4.13A-B) [14, 107, 163, 188]. Examination of the relationship between extent of modification and location in the ribosome shows that positions around the functional centers (decoding site, PTC, and intersubunit bridge) of the ribosome are overall less modified than those that lie in the periphery (Supplemental Fig. 4.14A). However, our analysis also shows a number of positions in the functional centers that are less than 95% modified (Supplemental Fig. 4.14B-D, Supplemental Table S3) suggesting

their modification status could have a larger impact on control of ribosome function.

To evaluate patterns of heterogeneity in the modification status of normal yeast ribosomes, we searched for subpopulations of rRNA in wild type yeast that might carry distinct modification profiles not expected by chance. Hierarchical clustering of wild type reads shows that there are no large and distinct classes of differently modified ribosomes in wild type cells, however some smaller (¡10% of total) subpopulations appear to cluster on the basis of correlated unmodified status between pairs of positions (Fig. 4.2E and Supplemental Fig. 4.15A). To identify correlated modification status at pairs of positions we applied our correlation change method by comparing wild type to IVT (See section 4.5). As seen in Fig. 4.2F and Supplemental Fig. 4.15B, wild type ribosomes have significant correlation changes in modification at distant positions in rRNA. One pair of significantly correlated positions in 18S, Φ632 and Φ766, are guided by the same snoRNA (snR161, p-value=1.3e-04, Fisher z-transform test), possibly explaining the basis for this correlation. We also observe a significant correlation between Am100 and Am436 (p-value=3.1e-04, Fisher z-transform test) as well as between Cm1639 and ac4C1773 in 18S (p-value=4.5e-06, Fisher z-transform test, Fig. 4.2F). None of these sites share a snoRNA or modification enzyme that could account for these correlations, however the correlated pairs lie close to each other (15-22Å) in three-dimensional structure of the mature ribosome (Fig. 4.2G), suggesting a structural or functional basis for their coordinated modification status. In 25S, consistent with our observations in the depletion experiment, 25S positions Cm1437, Am1449 and Gm1450 are all significantly more correlated with each other

than expected (Supplemental Fig 4.15, p-value=1.0e-44, Brown's method). Several of the significant long range correlations in wild type show up at significant levels in many other experiments (see below), indicating that the concerted modification status relationships at those positions are features of normal yeast ribosomes.

4.3.4 Loss of different RNA helicase-related functions results in distinct subpopulations of differently modified rRNA molecules

Previous studies have connected helicase activity required for ribosome biogenesis with changes in 2'O methylation at single positions in ensembles of rRNA molecules[2]. To explore how RNA helicases may affect correlated patterns of rRNA modification, we profiled cells compromised for Dbp3 or Prp43 helicase functions, both known to contribute to ribosome biogenesis[32, 88, 179]. Using a Dbp3 knockout strain (dbp3Δ) or a cold-sensitive Prp43 Q423N mutant (prp43-cs) grown at nonpermissive temperature (see section 4.5), we observed loss of 2'O methylation at specific locations in 18S and 25S rRNAs (Fig. 4.3A and Supplemental Fig. 4.16A) consistent with previous ensemble studies (Supplemental Figure 4.13C-D)[2]. Despite the numerous locations at which modification is compromised, hierarchical clustering of 25S rRNA single molecule profiles reveals that just 2-3 distinct but related sets of modification profiles describe the vast majority of ribosomes in both experiments. The triad of 2'O methylations guided by the snoRNA U24 at 25S positions 1437, 1449 and 1450 are often left unmodified in a highly concerted manner (Fig. 4.3B), as observed in wild type 25S (Supplemental Fig 4.15), and the snoRNP depletion experi-

ments (Fig 4.1). The pairwise correlations within this triad are significantly higher in both the dbp3Δ and the prp43-cs mutants relative to wild type (Fig4, Supplemental Fig. 4.17D-E, dbp3Δ p-value = 3.6e-68, prp43-cs 3.5e-11), Brown's method). To confirm the pattern seen in the probability clustering of the U24 positions, we clustered the underlying raw signal event means from the dbp3Δ and prp43-cs mutants (see section 4.5) [41]). As seen in Supplemental Fig. 4.18, there are two clear subpopulations of reads distinguished by the signal means at positions Cm1437, Am1449, Gm1450 in both the dbp3Δ and prp43-cs mutants, supporting the profile clustering results generated using the trained model. This suggests that if the U24 snoRNP is unable to guide modification of any of these positions, then all 3 positions are very likely to be left unmodified in a concerted fashion.

Prp43 interacts with a number of G-patch proteins that direct it to either the ribosome or the spliceosome [67, 124, 109, 161, 168, 27]. Two of these, Pxr1 and Sqs1, are important for correct pre-rRNA processing[124, 5, 63]. To test how the loss of Pxr1 or Sqs1 might affect ribosome modification profiles, we sequenced libraries from strains deleted for each. Although deletion of Sqs1 had little effect on modification (Fig. 4.5), loss of Pxr1 produced an extreme alteration in modification profiles resembling the more mild pattern produced by the prp43-cs mutant (Fig. 4.3A). All modifications affected by prp43-cs and all but two 18S 2'O methylations affected by dbp3Δ (Supplemental Fig. 4.16A) are also observed in pxr1Δ. This suggests that loss of Prp43 activity guided by Pxr1, but not that guided by Sqs1, is responsible for the concerted changes in modification pattern observed in the prp43-cs

strain.

Despite the similarities in modification patterns in the different mutants (Fig. 4.4, Supplemental Fig. 4.17), they are not identical. For example in 25S, positions Am817 and Gm908 (both guided by snR60) and Gm2619 and Um2724 (both guided by snR67) are significantly more correlated in dbp3Δ relative to wild type (817-908 p-value = 2.4e-33, 2619-2724 p-value = 3.5e-22, Fisher z-transform tests, Supplemental Fig. 4.17D), however, these positions are not significantly more correlated in the prp43-cs mutant relative to wild type (817-908 p-value=0.34, 2619-2724 p-value=0.919, Fisher z-transform test). Pxr1 and (more significantly) Dbp3 appear to promote efficient modification of positions guided by snR60 and snR67, however the contribution of Prp43 is less clear. It is possible that the conditional Prp43 mutation is not severe enough to produce a strong block to modification at those sites, or alternatively that Pxr1 has functions that do not require Prp43. Together our data show that loss of Dbp3 and Prp43 activity leads to loss of an overlapping but not identical set of rRNA modifications that create distinct classes of ribosomes in the cell in these mutants.

We have summarized the network of correlation changes observed in each mutant relative to wild type, displaying nucleotide positions as nodes and correlation changes as edges (Fig 4.4). In addition to the overlapping changes described above, this analysis highlights the connected nature of these modifications as well as their association with the functional centers of the ribosome. For example, loss of Prp43 and Pxr1 induce a concerted loss of modification of a set of nucleotides in the decoding

site of the small subunit (Fig 4.4B). Loss of Pxr1 leads to concerted loss of a set of modifications in the peptidyl transfer center of the large subunit Fig 4.4A. And all three mutants create a complex set of correlated modification changes in the triad Cm1437, Am1449, and Gm1450 near the protein exit tunnel of the large subunit (Fig 4.4A). Concerted modification of this triad is observed in wild type ribosomes (Supplemental Fig 4.15) as well as in the snoRNP depletion experiments (Fig 4.1). As discussed above, a shared snoRNP (e. g. snR60, snR67) may explain part of the concerted modification phenomenon, however in the majority of cases the elements that underlie concerted modification are not obvious.

4.3.5 Resilience of rRNA modification profiles to other genetic mutations and environmental treatments

Prp43 has a separate function in the disassembly of spliceosomes [109]. In addition some snoRNAs are encoded within introns and their synthesis can be compromised by mutations that affect splicing [173, 127, 125, 118]. To disentangle effects on modification by factors like Prp43 and intronic snoRNAs that arise from regulatory crosstalk between ribosome biogenesis and RNA splicing in yeast, we acquired single profile modification profiles for ribosomes from additional yeast mutants. Spp382 (also called Ntr1) is a G-patch protein that specifically mediates Prp43 interactions with the spliceosome [168, 161, 31, 52, 119]. In addition we employed a cold sensitive mutant of Prp16 (prp16-302) that accumulates splicing intermediates, as well as a deletion of Dbr1 that prevents debranching of the intron lariat [24, 169, 82, 47], a

reaction that promotes processing of some intronic snoRNAs, in particular U24 [118].

Using the threshold of ¿10% change in modification relative to wild type established above (Fig. 4.3A and Supplemental Fig. 4.11), we examined splicing-related perturbations for effects on rRNA modification that might be mediated through loss of one or more intronic snoRNAs (Fig 4.5). We observe a 36.8% reduction in modification for 18S Ψ106 (guided by snR44 from intron 2 of RPS22B) and an 11.0% reduction in modification frequency for 18S Am974 (guided by snR54 from intron 1 of IMD4). Although these are the only modification changes that pass the threshold, modification of the nucleotides in the 25S triad Cm1437, Am1449, and Gm1450 (all guided by U24 from ASC1) is detectably reduced, as is modification of Ψ2258 and Ψ2250 (both guided by snRN191 from intron 1 of NOG2). There are alternative snoRNA maturation pathways that are independent of splicing, for example via Rnt1 [62] and partially processed U24 still can guide modifications at its corresponding locations [118, 2, 125], consistent with our observation that loss of function in splicing is not sufficient to greatly impact rRNA modification through either Prp43 or by virtue of the intronic origin of some snoRNAs. Only one modification event, the pseudouridylation of 18S U106 by snR44, seems substantially affected by the loss of Dbr1 (Fig 4.5).

To test whether single molecule modification profiles were altered by environmental and growth conditions known to affect ribosome function and biogenesis[58, 176], we isolated and acquired profiles of rRNA isolated from cells at stationary phase [75], after a 1 hour shift to potassium acetate to induce starvation, treated with ra-

pamycin (TOR kinase inhibitor) for 1 h to block nutrient signalling [131, 22, 64, 68], treated with cycloheximide to block translational elongation [117] and create ribosome collisions [151], or after cold shock. In none of these treatments did we detect substantial changes in modification profile (Fig 4.5, Supplemental Tables S1 and S2). These observations indicate that in general the annotated modification patterns on rRNA are refractory to rapid alterations by dramatic changes in the physiological conditions we tested. At this time there are no known enzymes that would reverse either pseudouridylation or 2'O methylation of ribose in RNA, as would be required if modifications added during ribosome biogenesis were removed as part of a regulatory response to changes in the environment.

4.4 Discussion

A central goal of this research was to capture single molecule modification profiles of S. cerevisiae 18S and 25S rRNA, in order to understand the coordination of modification across the ribosome during ribosome biogenesis, and to discover relationships and dependencies between distant modifications. Using a catastrophic disruption of modification by depletion of the two main classes of snoRNPs responsible for the bulk of ribosome pseudouridylation and 2'O methylation, we validate the framework for our method and find that to a large extent these two classes of modification are not dependent on each other (Fig 4.1). Using a mixture of rRNA from wild type cells and cells deleted for different individual snoRNAs, we show we can

resolve populations of ribosomes that differ by a single modification, and apply this to characterize modification heterogeneity and identify instances of concerted modification of sets of nucleotides in the wild type ribosome population (Fig 4.2). We then characterized the single molecule modification profiles that result from loss of two distinct helicase activities provided by Dbp3 and Prp43, finding that a complex set of concerted effects on modification arise from these disruptions, with implications for ribosome biogenesis and the important functional centers of the ribosome (Figs 4.3 and 4.4). Finally, we examine the effect of other mutations, changes in physiological conditions, or inhibitors of ribosome function on the annotated modifications across the ribosome and find that they are refractory to change (Fig 4.5). These results provide a new perspective on ribosome heterogeneity as represented by RNA modification patterns, and open a path to whole molecule analysis of RNA modification for other classes of RNA.

rRNA modifications are thought to fine-tune and regulate rRNA folding and ribosome function [153]. Many rRNA modifications cluster around the functional centers of the ribosome and recent studies have illuminated the role that different individual modifications play during translation of specific sets of mRNAs under different physiological conditions [143, 148, 98]. Our results reveal a number of instances where RNA modifications in the functional domains of wild type ribosome are heterogeneous, and their presence or absence occurs in a concerted manner.

During protein synthesis, the nascent polypeptide chain moves from the PTC and exits the ribosome through the polypeptide exit tunnel (PET). Numerous studies

have shown that interactions between the nascent polypeptide chain and the PET can lead to ribosome pausing/stalling resulting in regulation of translation and protein folding [29]. Our work reveals distinct clusters of 25S rRNAs missing 2'O-methyl modifications at positions Cm1437, Am1449, and Gm1450 in Pxr1, Prp43, and Dbp3 mutants (Fig. 4.3-4.4). Importantly, we also observe correlation and clustering of rRNAs that do not contain these modifications in wild type cells (Supplemental Fig. 4.15) suggesting a potential regulatory mechanism. Cm1437, Am1449, and Gm1450 all line the PET of the 60S subunit and appear to interact with conserved internal loops of ribosomal proteins L4 and L17[8]. These loops insert into the PET to form the constriction site and is thought to act as an "exit gate" by interacting with the nascent polypeptide chain during translation (Fig. 4.6 and [183, 191, 113]). Furthermore, these three positions are in domain 0 of ribosomal rRNA, which acts as a central hub around which the other six 25S rRNA domains fold [126]. Regulation of these modifications could impact how each domain of rRNA folds upon each other during ribosome biogenesis and exit tunnel formation. Thus, in the absence of Cm1437, Am1449, and Gm1450, the rRNA and the loops of L4 and L17 may not be properly positioned, affecting the structure and chemistry of the PET, translation, and protein folding.

We also observe a number of correlated modifications in wild type 18S rRNA within the decoding center and the intersubunit bridge of the 40S subunit. Cm1639 (snR70) in the P-site exhibits correlation with Φ999 (snR31, E-site), ac4C1280 (Kre33), and ac4C1773 (Kre33, intersubunit bridge) (Fig. 4.2). Furthermore, ac4C1773 and

m26A1782 (Dim1, intersubunit bridge) are correlated. Together, our correlation data using single-molecule profiling suggests a functional relationship among groups of modifications in wild type ribosomes that could impact how these functional regions form as well as their activity during translation.

Recent work from [2] demonstrated the role that RNA helicases play in regulating the dynamics of snoRNPs during rRNA modification and ribosome biogenesis. Their data suggests a model in which Dbp3 and Prp43 function by releasing snoRNAs from the pre-ribosome to allow subsequent modification of adjacent sites that are otherwise occluded due to overlapping basepairing positions of adjacent snoRNPs. Here, by profiling full-length rRNAs, we extend this model by revealing concerted changes in modifications over long distances when the activity of Dbp3 or Prp43 is compromised. Furthermore, our work shows that Pxr1, but not Sqs1, is the main G-patch protein important for Prp43 function during rRNA modification.

In the absence of these helicases we observe a large set of overlapping but not identical changes in modified positions for each of the mutants tested. Analysis of pairs of nucleotides that change in a concerted fashion in each mutant, across the entire length of the rRNA, reveals distinct hubs of correlated modifications, many of which reside in the functional centers of the ribosome. These hubs of correlated positions might reflect critical points in ribosome biogenesis and function such that each This suggests a dependency among them during ribosome biogenesis or ribosome function (Fig. 4.4).

We developed a hidden Markov model-based approach that allows 1) single

molecule profiling and clustering of RNAs to visualize high-level relationships within a population, 2) the ability to test for changes in correlations between any given pair of modifications on the same molecule, and 3) a way to estimate the fraction of modification of each site. The model training paradigm we have developed to profile modifications can easily be applied to other nucleic acids of interest such as other non-coding RNAs and messenger RNAs, provided unmodified molecules (IVT) and fully modified molecules are available as reference for training. Here we used wild type rRNA as our fully modified training example, with the clear understanding that not all wild type molecules are fully modified. In several instances we confirmed that this had little or no effect on performance of the model (Supplemental Fig. 4.8). A second limitation arises when the training samples do not have enough information to learn to resolve dense clusters of modifications. In cases where this was a concern, we were able to validate the predictions of our model by clustering the raw signal means and showing that closely spaced modifications that shared overlapping k-mers were called correctly (Supplemental Fig. 4.10). While there is some evidence that unknown modified kmer distributions can be estimated using known kmer distributions [40], generating more specific modification training data sets that contain all combinations of partially modified closely spaced clusters of nucleotides may be required to produce more accurate and general modification detection algorithms. This is especially true if de-novo detection of modifications within complex sequences is the goal[108, 89].

4.5 Methods

4.5.1 Growth of yeast strains

Yeast strains GAL-NOP58 and GAL-CBF5 are described in (Lafontaine and Tollervey 1999)(Lafontaine et al. 1998). Cells were grown at 30 °C in YEPgal liquid medium (2% galactose, 2% peptone, 1% yeast extract) or shifted to liquid YEPD (2% dextrose, 2% peptone, 1% yeast extract) to mid-log phase (OD600 = 0.25-0.5) for 16 hours to repress expression of Nop58 or Cbf5. Cells were harvested by centrifugation and RNA was isolated. Unless indicated, all other strains were grown in YEPD at 30 °C to mid-log phase. Cells exposed to various environmental conditions were treated as follows: 1% KOAc (1 hr, 30 °C), cycloheximide (1 ug/ml for 1 hour), rapamycin (200 ng/ml for 1 or 5 hours), and pladienolide B (5 uM for 1 hour). Stationary phase cells were grown to an OD600 = 10. Strains carrying prp16-302 [105] and prp43 Q423N [88] mutations, and wild type [142] were grown to mid log phase at 30 °C and shifted to 18 °C for 1 hour by addition of an equal volume of 6 °C YEPD. The spp382-1 strain is described in [119]). The strains deleted for the SNR80 (YWD448a), SNR83 (YWD451a) or SNR87 (YWD452a) genes are described in [142]. Yeast strains deleted for the SNR4 and SNR45 genes are described in [121]. All yeast strains and genotypes can be found in Supplemental Table S4.

4.5.2 RNA isolation

RNA was extracted from approximately five total OD600 of cells (usually 10 ml culture at OD600 = 0.5 for mid log cells, 0.5 ml of stationary cells at OD600 = 10) using a hot phenol protocol 1 described in [3].

4.5.3 In vitro synthesis of 18S and 25S rRNA

Unmodified yeast 18S and 25S rRNAs were transcribed in vitro from plasmids encoding T7-18S and T7-25S sequences using T7 RNA polymerase. PCR products encoding 18S and 25S rDNA were amplified from the plasmid pWL155 which contains the RDN1-1 gene fused with the GAL promoter at the 5' end ([95] a kind gift from Jelena Jakovlievic) and cloned into a T7 promoter-containing plasmid digested with EcoRI and HindIII using Gibson Assembly (NEB). The resulting plasmids were then digested with HindIII and run-off transcription was performed using the MEGAscript T7 kit (Invitrogen) following the manufacturer's instructions. T7-18S and -25S in vitro transcription reactions were evaluated by gel electrophoresis for bands of correct size that correspond to 18S and 25S rRNAs. Transcription reactions were extracted and purified with phenol:chloroform:isoamyl alcohol (25:24:1), ethanol precipitated and resuspended in nuclease-free H2O. Purified T7-18S and -25S rRNA transcripts were then quantified on a NanoDrop spectrophotometer and pooled in equimolar ratios for sequencing library preparation. The T7 run-off transcription reactions terminate in a 3' end generated by HindIII digestion and thus include an additional AAGCU sequence not present in endogenous 18S and 25S rRNAs. There-

fore, T7-18S and T7-25S splint oligonucleotides were used to capture the 3' end of T7 transcribed rRNAs (see below, Supplemental Table S5).

4.5.4 Sequencing library preparation

Direct RNA sequencing libraries were constructed using the SQK-RNA002 (Oxford Nanopore Technologies) kit following the manufacturer's protocol with the following modifications. Briefly, 750 ng of total yeast RNA was used as input material. To facilitate ligation of sequencing adapters to endogenous yeast 18S and 25S rRNA, 1 ul of 10 pmol/ul custom oligonucleotide duplexes complementary to the 3' ends of 18S and 25S rRNA and the 5' end of the ONT RMX sequencing adapter were used instead of the kit provided RTA adapter (Supplemental Table S5). To create duplexes, 100 pmol of either 18S or 25S splint oligo was incubated with 100 pmol of sequencing adapter and nuclease free H20 in a total volume of 10 ul. Reactions were heated to 95 °C for 2 minutes and gradually cooled at 65 °C for 10 minutes, 48 °C for 10 minutes, room temperature for 10 minutes and then placed on ice. Annealed oligonucleotide duplexes targeting 18S and 25S rRNAs were then pooled in equimolar ratio and 1 ul of the pool was used for sequencing library preparation. In the case of T7 rRNA sequencing libraries, T7-18S splint and T7-25S splint oligos were used to capture the 3' end generated by HindIII digestion and run-off transcription. To enhance ligation efficiency during library preparation, the first and second ligation steps were increased from 10 minutes to 15 minutes and performed at room temperature. Reverse transcription was omitted. Sequencing-adapted libraries were eluted in 21 ul of elution

buffer.

4.5.5 Nanopore sequencing

RNAs extracted from GAL-NOP58 and GAL-CBF5 strains, and in vitro transcribed RNA were sequenced on the MinION Mk1B sequencer using MinION FLO-MIN106D R9.4.1 flow cells (Oxford Nanopore Technologies) following the manufacturer's instructions. 20 ul of Sequencing libraries was mixed with 17.5 ul of H20 and 37.5 ul of RRB buffer. 75 ul of the prepared sequencing library was loaded onto a flushed and primed flow cell and sequenced for 12-48 hour depending on the lifetime of active pores. RNAs extracted from all other strains and growth conditions were sequenced on the MinION Mk1B sequencer using Flongle FLO-FLG001 R9.4.1 flow cells. Flongle flow cells were flushed and primed with 120 ul of flush buffer mix (117 ul FLB and 3 ul FLT). 30 ul of prepared sequencing library (described above) was loaded onto the flow cell and sequenced for 8-24 hours. Sequencing experiments were controlled using the MinKNOW software (Oxford Nanopore Technologies).

4.5.6 Data preprocessing

The following preprocessing steps were applied to all of the sequencing experiments. Basecalling was done by Guppy v3.1.5+781ed57. In order to analyze specific subsets of reads more efficiently, we split the multi-fast5 reads into individual reads using the multi_to_single_fast5 command from `https://github.com/nanoporetech/ont_fast5_api`. We then created an index file matching a fast5 to a fastq entry us-

ing nanopolish index from https://github.com/jts/nanopolish (Simpson et al. 2017). The reference sequence for the S. cerevisiae 18S and 25S rRNA came from (Engel et al. 2014). Initial basecalled sequence to reference alignment was done via minimap2 version 2.17-r943-dirty from `https://github.com/lh3/minimap2` using the –MD flag which speeds up processing of signalAlign[92]. Alignment files were sorted and filtered using samtools version 1.9 by flag -F 2308 which filters out unmapped reads, non-primary alignment reads and supplemental alignment reads [93]. Given that nanopore sequencing with RNA is 3'-5', in order to filter for 'full length' reads we used samtools view to select for reads that covered the first 15 bases of both 18S and 25S rRNAs[93]. Read information and quality control metrics in Supplemental Table S6 were gathered using pycoQC version v2.5.0.23 [90].

4.5.7 SignalAlign Pipeline

Model Definition

We initialized the transition probabilities from previous signalAlign r9.4 models. The initialized kmer distributions were defined in r9.4_180mv_70bps_5mer_RNA from ONT `https://github.com/nanoporetech/kmer_models`. Unlike previous kmer model modification detection algorithms, we chose to model modifications independently from other modifications of the same class in order to maintain the same informational inputs to each modification position. So, we iteratively redefined shared kmers with unused kmers from the model until all modifications were covered by unique kmers (see Code availability). For all kmers outside of mod-

ification branch points, we used the default RNA kmer distributions from ONT (r9.4_180mv_70bps_5mer_RNA).

Training Configuration

SignalAlign uses a variable-order hidden Markov model (HMM) which allows the number of paths through the HMM to be correctly constructed when ambiguous positions are defined [135]. Recent updates to signalAlign allow for relatively easy model definition and variant site selection which allows a user to define modifications, set prediction site locations and train a model. We defined all positions in the IVT sample as canonical and all positions in the wild type as modified. The locations of ambiguous positions are determined by the presence of ambiguous characters in the reference sequence[163]. In this experiment, ambiguous characters represent two possible nucleotides, a canonical nucleotide and the most prevalent modified nucleotide. The ambiguous characters were defined in a small model file. The annotated modified nucleotides in 18S and 25S S. cerevisiae rRNA were defined as ambiguous during all inference steps. For supervised training using IVT and wild type sequencing data, all potential ambiguous positions were defined as either canonical or modified respectively. We used 500 18S and 25S wild type reads and 500 18S and 25S IVT reads and ran 30 rounds of training. For each round of training, we generated alignments between events and the reference sequence. Then, we generated new event Gaussian distributions for all kmers covering modified positions. The mean of the Gaussian distribution was defined as the median of the empirical kmer distribution and the

standard deviation was defined as the median absolute deviation of the empirical distribution. Similar to another study, we have seen that the median is less susceptible to being influenced by outliers[41]. To train the model, we used trainModels.py from signalAlign.

Inference and Accuracy Metrics

In order to validate our results, we used 'runSignalAlign.py' and a trained model to predict modification status on all positions of 500 hold out IVT reads and 500 hold out wild type reads. We placed ambiguous characters at modified positions in the reference for both IVT and wild type reads and signalAlign produced the posterior probability of event to kmer alignments given the trained model. We use embed_main sa2bed to decode the posterior probabilities from the signalAlign output into the probability of a position being modified (Rand et al. 2017). These probabilities are used for the receiver operating characteristic curve, precision-recall curve, and calibration curve of Supplemental Fig 4.8. A probability cutoff of 0.5 is used for the confusion matrix as well as the quantification of percent modified for any position. We also compared accuracy on our test set to several snoRNA knockouts. Again, assuming snoRNA knockouts completely ablate target modifications and modifications are 100% present at all other positions, the average balanced accuracy over the snoRNA knockout positions is 82.8% and the expected balanced accuracy is 87.1% (Supplemental Table S2-3). Average balanced accuracy is calculated by getting the average of all balanced accuracies across all snoRNA knockout positions. Balanced

accuracy for one position is calculated by adding the specificity to the sensitivity and dividing by two.

Percent Modification Change

For every experiment and each modification position, we perform a chi-square two sample test comparing the wild type's modification frequency to the experiment's modification frequency[122]. We then correct for multiple tests using the Benjamani-Hochberg procedure [9]. We also control for batch effects by filtering out reads which fall below the maximum change in modification frequency between the replicates of the snR48 KO. Percent modified, chi-square two sample test between wild type and all other samples p-value, Benjamani-Hochberg corrected p-value can be found in Supplemental Table S2.

4.5.8 Hierarchical Clustering Analysis

Dendrogram creation procedure

In order to determine any subclusters of reads based on a modification profile, we used hierarchical clustering on the per-read per-site modification probabilities we generated from the inference step [177, 123]. We generated the dendrogram using Ward's method as the hierarchical clustering method and euclidean distance as the distance metric [175]. UMAP dimension reduction was done using the umap python package and visualization using matplotlib [110, 72]. Before clustering analysis, we filter out reads which do not cover every modification site.

Cluster Partitioning

To determine the number of reads in a set of N clusters we simply cut the dendrogram to create N subclusters and calculated the fraction of reads within each branch.

4.5.9 Modification Correlations

To calculate correlations between modified positions, we first filter out reads which did not cover all modifications and select the set of probabilities associated with each position. We then calculate the Spearman correlation between all pairwise combinations of modification positions on the same molecule. P-values were calculated using a two sided t-test and multiple tests corrected via the Benjamani-Hochberg procedure[159, 9].

To compare correlations between experiments, we used Fisher's z-transformation to convert correlations into z-scores and then performed a z-test to obtain p-values[50, 48, 49, 190]. We then correct for multiple tests using the Benjamani-Hochberg procedure [9]. These p-values represent the confidence that, between two samples, there is a significant difference between the two correlations. All correlation plots have stars for positions which are both significantly different from a comparison experiment (wild type or IVT) and are significantly different from zero (p-value ¡ 0.05). To account for variation in experimental repeats, we plot the minimum difference and highest corrected p-value for all pairwise comparisons between experimental repeats and wild type repeats.

For higher order claims which require aggregating information from several hypothesis tests we use Empirical Brown's method [129, 21]. The Empirical Brown's method uses empirical data to calculate the covariance matrix which is used to extended Fisher's method to the dependent case by using a re-scaled χ^2 distribution (see Code availability). Spearman correlation values, original two sided t-test p-values, corrected two sided t-test p-values, Fisher z-transform test comparison p-values, and corrected Fisher z-transform tests p-values can be found in Supplemental Table S1.

4.5.10 Event Cluster Visualization

Using almost the exact same procedure outlined in a previous study[41], we gather the kmer to reference mapping generated from signalAlign and extract the most probable event to kmer alignment path using the maximum expected accuracy alignment [135, 45]. For each read, we standardize the raw signal and calculate event means. Prior to clustering and visualization, we combine all reads together and standardize events by column. We generate the dendrogram using the same procedure as hierarchical clustering of modification profiles, Ward's method and euclidean distance [175].

For kmer distributions seen in Supplemental Figure 4.11, we plot the kernel density estimate of all events aligning to the corresponding kmer with a probability greater than 0.5. We then simply plot the corresponding kmer distributions from the final trained kmer model.

4.5.11 Sample Compare Site Detection

Tombo Pipeline

Using Tombo version 1.5.1, initial embedding of fastq data into the raw fast5s was done with the tombo preprocess annotate_raw_with_fastqs and signal to reference alignment with tombo resquiggle [108]. Finally, tombo detect_modifications level_sample_compare was used to generate windowed means of individual position Kolmogorov–Smirnov tests comparing the IVT sample position signal distributions to the wild type sample (WT_YPD) position signal distributions [108]. For a given position i, the windowed mean D-statistic is $w_i = \frac{\sum_{i-1}^{i+1} d_i}{3}$ where d is the D-statistic for a given position and w is the final reported statistic plotted in Supplemental Fig. 4.7.

Accuracy of Modification Site Prediction

In order to get a general view of how all of the modifications are affecting the current signal we analyzed the signal shift between in vitro transcribed (IVT) and one wild type sample (WT_YPD) using Tombo [108]. The signal difference of 18S and 25S strands using Tombo is shown in Supplemental Fig. 4.7A and 4.7B respectively. There is a clear correlation between annotated modified positions and signal deviation but in order to quantify the relative accuracy of both approaches, we naively labeled the per-position deviations with the corresponding windowed mean D-statistic. As shown in Supplemental Fig. S1C-D, the per-position modification

calling detection AUROC (Area Under the Receiver Operating Characteristic) was 0.924 for 18S and 0.934 for 25S. However, if a canonical position is directly next to a modified position, it is very likely the underlying current is going to be shifted for that position. Also, the uncertainty of which specific nucleotide in the pore gives rise to the most significant signal shift makes site selection for kmer based sample compare frameworks very difficult [41, 89, 108]. Therefore, instead of evaluating Tombo on the per-position modification calling accuracy, we used a less stringent metric of modification window calling accuracy. We looked to see if a peak was within a window of a specific modification and disregarded large differences in signal in the neighboring 2 bases of a modification. Specifically, for each modification, we took the maximum corresponding statistic value of a window of 5 positions covering that modification. For example, if pos 20 was modified, the corresponding statistic for position 20 was the maximum value for positions 18, 19, 20, 21 and 22. Then, we removed the 2 upstream and downstream values from being classified. So, positions 18, 19, 21 and 22 will not be classified as true negatives or false positives. This approach allows for uncertainty of where the modification is within a small window of 5 positions and greatly reduces the false positive rate. As seen in Supplemental Fig. 4.7C and 4.7D, by decreasing the stringency of our accuracy metric we see a marked improvement of modification detection to an AUROC of 0.984 for 18S and 0.986 for 25S.

4.5.12 Modification Labels and Frequency

Underlying labels for modification and frequency for the S. cerevisiae 18S and 25S rRNA came from Taoka et al [163]. Expected changes in modification frequency in the Dbp3 deletion experiment came from Aquino et al [2]. SnoRNA modification sites on yeast rRNA come from the UMASS Amherst Yeast snoRNA database [127].

4.5.13 Data availability

Fastq files from all direct RNA sequencing runs and signalAlign modification calls are publicly available in NCBI's Gene Expression Omnibus (GEO) and are accessible through GEO Series accession number GSE186634 `https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE186634`. Fast5 files for all direct RNA sequencing are available in the European Nucleotide Archive (ENA) at EMBL-EBI under accession number PRJEB48183 `https://www.ebi.ac.uk/ena/browser/view/PRJEB48183`. A detailed description of the datasets used and sequenced in this work with their corresponding ENA, GEO, and SRA IDs can be found in Supplementary Table S7.

4.5.14 Code availability

Documentation, install requirements, and analysis scripts can be found at `https://github.com/adbailey4/yeast_rrna_modification_detection`. SignalAlign can be found at `https://github.com/UCSC-nanopore-cgl/signalAlign` and

embed_fast5 can be found `https://github.com/adbailey4/embed_fast5`.

4.6 Acknowledgements

We would like to thank our colleagues for graciously sharing strains and plasmids: Skip Fournier and Wayne Decatur (snR80, snR83, and snR87 deletions), David Tollervey (GAL-CBF5 and GAL-NOP58), Raymond O'Keefe (snR4 and snR45 deletions), Jon Staley (Prp43 Q423N), Jelena Jakovljevic and John Woolford (rDNA plasmid originally provided by Skip Fornier). We also thank Jordan Eizenga, John Paul Donohue, Miten Jain, and Logan Mulroney for technical assistance and advice. This research was supported by NIH grants R01 HG010053 (to M. Akeson, Ares and Paten Co-PIs) and R01 GM040478 (M. Ares), as well as U41HG010972, R01HG010485, U01HG010961, OT3HL142481, OT2OD026682, U01HL137183, and 2U41HG007234 to B. Paten.

4.7 Author Contributions

All authors contributed to conception and experimental design. JT, MA and HI grew yeast, extracted RNA, prepared libraries, ran the nanopore sequencer and performed base calling. AB performed most of the nanopore data analysis, including model building, clustering, and correlation measurements. HD provided feedback and advice throughout the development and analysis process. AD and SM built some of the clustering analysis scripts. AD and SM's participation in this research took place

under the auspices of the Science Internship Program at the University of California Santa Cruz. MA and BP supervised the project. All authors assisted in manuscript preparation.

4.8 Competing Financial Interests

The authors declare no competing financial interests.

4.9 Figures

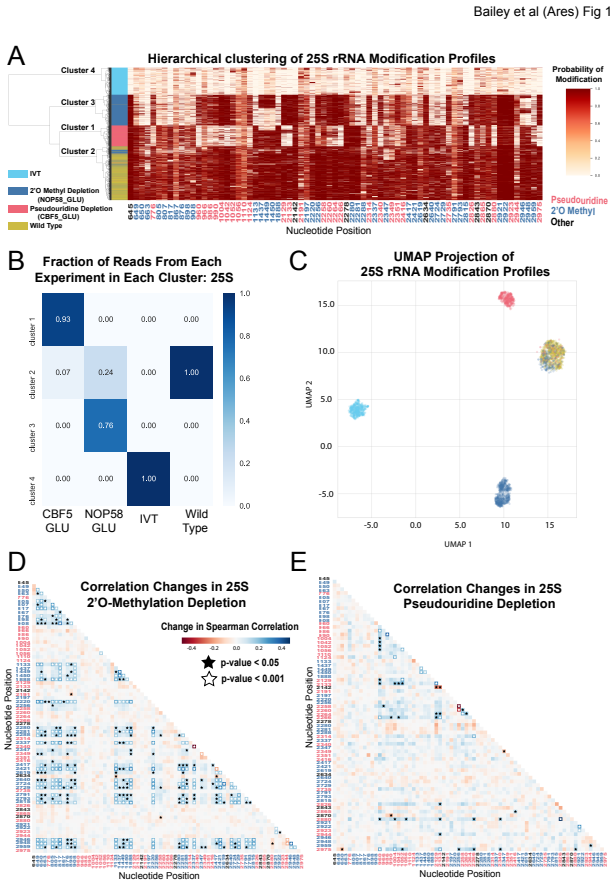

Figure 4.1: Clustering and correlation analysis of depletion experiment modification profiles in 25S. (A) Hierarchical clustering of 25S yeast rRNA modification profiles of IVT, wild type, and both pseudouridine and 2'O methyl depletion experiments. (B) Fraction reads from IVT, wild type and both depletion experiments in each cluster of 25S rRNA. (C) UMAP visualization of 25S yeast rRNA modification profiles of IVT, wild type, and both pseudouridine and 2'O methyl depletion experiments. (D/E) Change in Spearman correlations of 25S reads in 2'O methyl depletion (D) and pseudouridine depletion (E) when compared to wild type. Stars represent significant changes when compared to wild type correlation and significantly different from zero correlation. All nucleotide positions are color coded where blue positions are 2'O-methyl, red positions are pseudouridine, and black positions are neither 2'O-methyl nor pseudouridine.

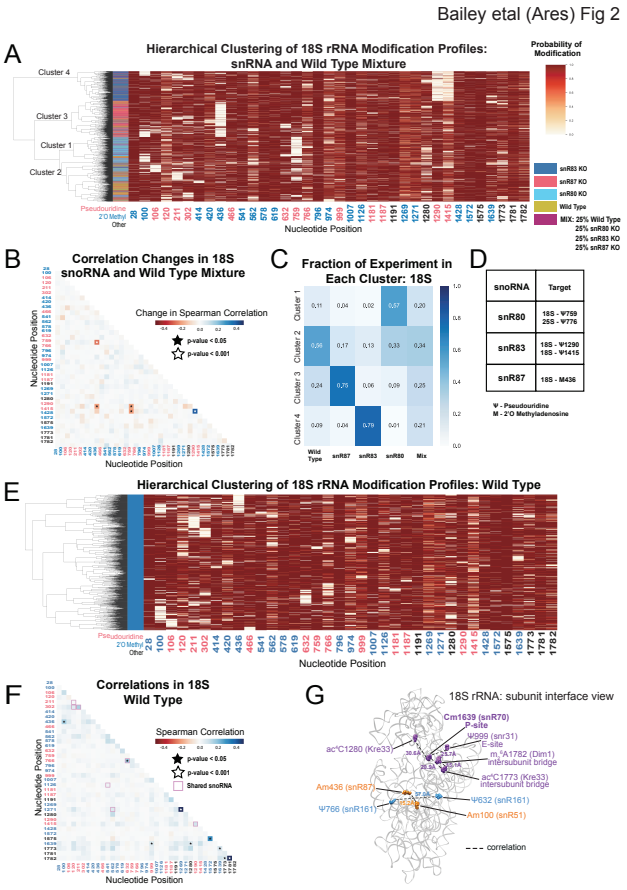

Figure 4.2: Clustering of 18S rRNA modification profiles and correlation analysis from the mixture experiment and wild type rRNA. (A) Hierarchical clustering of 18S modification of profiles from wild type, mixed, snR80 KO, snR83 KO, and snR87 KO samples. (B) Change in Spearman correlations of 18S reads in the mixture experiment when compared to wild type. Stars represent significant changes when compared to wild type correlation and significantly different from zero correlation. (C) Fraction of wild type, mixed sample, snR80 KO, snR83 KO, and snR87 KO in each cluster of 18S rRNA. (D) Table of snoRNAs knocked down with the corresponding expected knocked down modifications. (E) Hierarchical clustering of 18S yeast rRNA modification profiles from wild type yeast. (F) Wild type Spearman correlation of 18S wild type reads. Stars represent significantly different to IVT correlations and significantly different from zero correlation. (G) Crystal structure model of wild type S. cerevisiae 18S rRNA highlighting significant correlated positions. PDB: 4V88 [8]

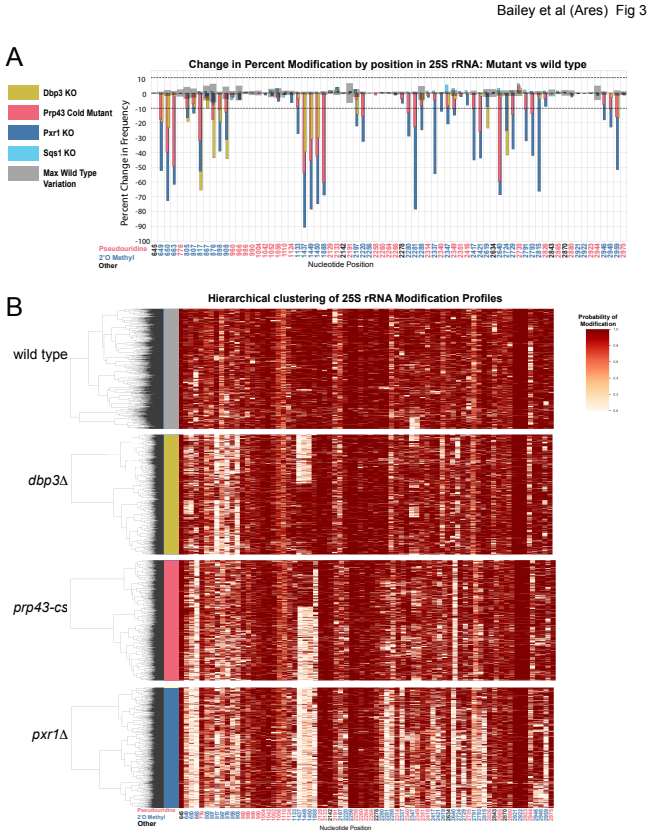

Figure 4.3: Clustering of 25S rRNA modification profiles and percent change in modification frequency of helicase mutants Dbp3 and Prp43 and G-patch proteins Pxr1 and Sqs1. (A) Barplots of the difference between wild type modification frequency and Dbp3 KO, Prp43 cold mutant, Pxr1 KO, and Sqs1 KO modification frequencies in 25S yeast rRNA. Grey bars indicate the variance of wild type rRNA modification at each position and the black dotted lines represent the maximum variance observed at any site. (B) Hierarchical clustering of 25S yeast rRNA modification profiles from wild type, Dbp3 knockout, Prp43 cold mutant, and Pxr1 KO.

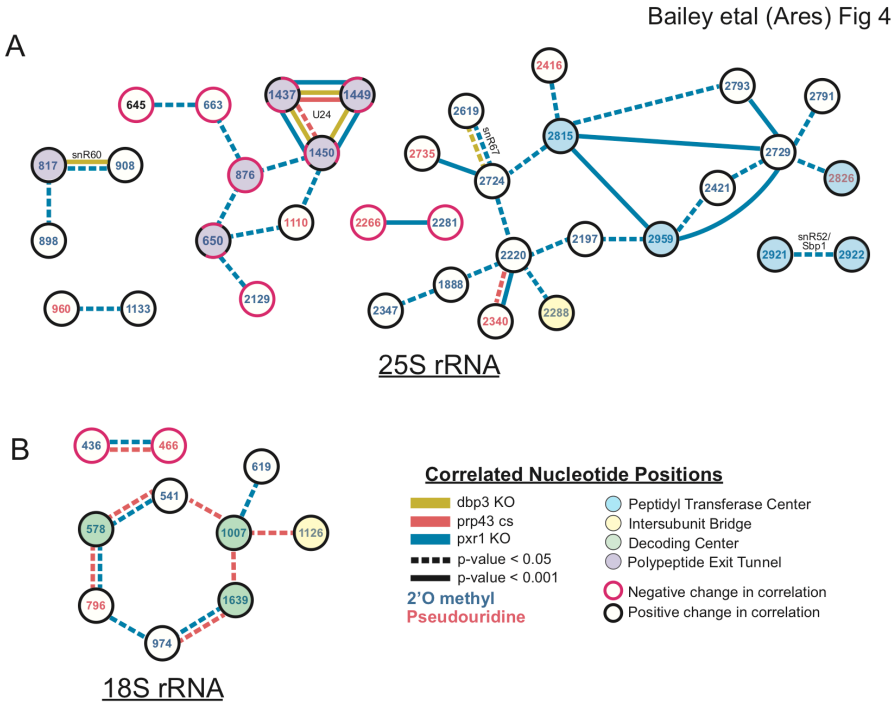

Figure 4.4: Figure 4: Changes in correlated nucleotide positions in dbp3Δ , prp43-cs, or pxr1Δ mutants. Pairs of correlated nucleotide changes (nodes) are shown for each mutant (edges) relative to wild type yeast 25S rRNA (A) and 18S rRNA (B). In cases where correlated pairs show differential changes in correlation in different mutants (eg. U24 modifications), node color rings are fragmented with the appropriate mutant edge connecting to either the magenta (negative change in correlation) or black (positive change in correlation) portion of the node.

Figure 4.5: Resilience of yeast rRNA modifications to a variety of splicing mutants and experimental conditions. Barplots of the difference between wild type modification frequency and Dbr1 KO, Spp382 KO, Prp16 cold mutant, KOAc treated, cycloheximide treated, stationary, rapamycin treated and cold shock yeast modification frequencies in yeast 18S (A) and 25S (B) rRNA. Grey bars indicate the variance of wild type rRNA modification at each position and the black dotted lines represent the maximum variance across sites.

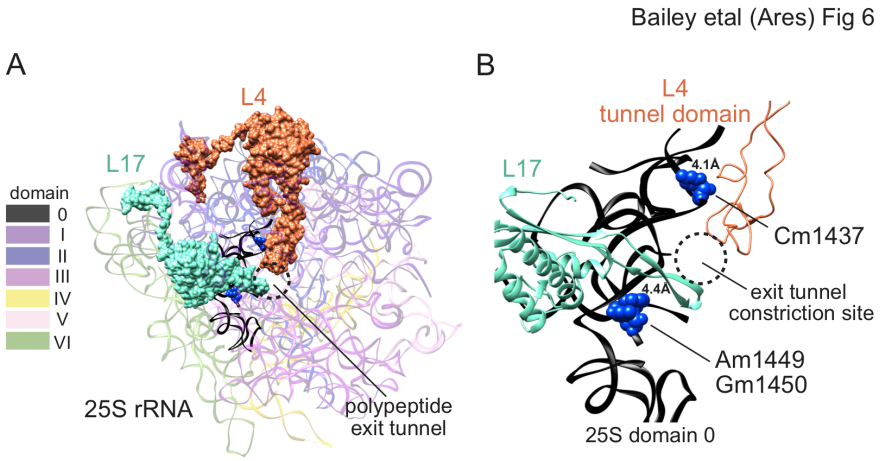

Figure 4.6: 2'O-methyl modifications guided by U24 line the polypeptide exit tunnel and interact with ribosomal proteins L4 and L17. (A) Crystal structure model of yeast 25S rRNA and ribosomal proteins L4 and L17 in surface view (PDB:4V88)[8]. rRNA domains are color coded according to the RiboVision Suite [12]. The distal end of the polypeptide exit tunnel is indicated. U24-guided modified nucleotides Cm1437, Am1449, and Gm1450 are shown in blue. (B) Focused view of the L4 tunnel domain and the internal loop of L17 forming the exit tunnel constriction sites. 25S rRNA domain 0 is shown in black.

4.10 Tables

All tables for this chapter have been included in a supplementary file `yeast_rrna_supplemental_tables.xlsx`.

4.11 Supplementary Information

4.11.1 Supplementary Note 1

In order to better resolve tight clusters of modification we need training data with labelled reads for all possible permutations of modifications. However, we do not have that information. So, we rely on prior information, experimental design expectations and signal comparisons to determine confidence in signalAlign predictions of modification clusters. Specifically, for high interest modification clusters, we validate modification profile clusters found using signalAlign using nanopore signal patterns by clustering the underlying event means (see section 4.5)[41]. In Supplemental Figure 4.10, we noticed two interesting patterns of modifications positions located in the peptidyl transfer center (PTC) (Um2921, Gm2922, Ψ2923) and positions targeted by U24 (Cm1437, Am1449, Gm1450).

Prior to running our depletion experiments, we were uncertain if inhibiting box C/D snoRNP function would alter the modification status of Um2921 because both Um2921 and Gm2922 can be methylated with the non-snoRNP methyltransferase Sbp1 [86]. However, we did expect the Cbf5 depletion would create a high proportion of reads with a modification pattern unseen by the model (only missing

the Ψ2923). Thus, prior to analysis, we were uncertain on the number of high proportion modification patterns across these three positions. After analysis by signalAlign, we see similarly modified wild type and Nop58 depletion reads with a slight decrease in frequency of all three modifications in the Cbf5 depletion (Supplemental Figure 4.10A and Supplemental Table S2). Thus, our initial hypotheses are that Um2921 modification is not altered by inhibiting box C/D snoRNP function and that the altered signal caused by missing Ψ2923 manifests as a slight (¡5%) decrease in modification frequency of Um2921 and a larger (¡10%) decrease of modification frequency of Gm2922 and Ψ2923. To determine if our hypothesis is correct, we used the underlying event means to identify the number modification patterns through the PTC modifications. At a high level look (2917-2922) we see two clear clusters; one cluster of IVT reads indicating three unmodified positions and one cluster with wild type and both depletion experiments (Supplemental Figure 4.10C). Upon closer inspection of the most informative kmers (2921, 2922, 2923, and 2924), we see clustering of event means partition Cbf5 depletion reads and 2'O-methyl depletion reads (Supplemental Figure 4.10E). Given that we only see two main clusters outside of the IVT cluster leads us to believe that the 2'O-methyl depletion had little to no effect on modification status on Um2921 and Cbf5 depletion experiment most likely causes an unknown decrease in modification at Ψ2923.

For the U24 positions, our model shows a high level of correlation between each position, unexpected missing 2'O-methyls in the Cbf5 depletion and unexpected presence of 2'O-methyls in the Nop58 depletion (Supplemental Figure 4.10A). Given

the isolation of Cm1437, we are confident that the unexpected predicted status of Cm1437 in both depletion experiments are accurate. The high level of correlation is not necessarily surprising given the shared snoRNA and we confirm the patterns found in modification profile clustering with event means clustering (Supplemental Figure 4.10B). However, to confirm the results at the pair of 2'O-methyls (Am1449 and Gm1450), we clustered the most informative kmers (1448, 1449, and 1450) and saw only two clusters of events, corresponding with IVT and wild type (Supplemental Figure 4.10D). Given we see only two clusters and no partitioning between the two depletion experiments leads us to believe that there are only two primary modification patterns for Am1449 and Gm1450, either both modified or both unmodified. Given the underlying clustering of the U24 positions, we believe that there is a high level of correlation between the U24 positions, Cbf5 depletion leads to a decrease in U24 modification efficiency and rRNAs with U24 modifications maybe preferentially selected or modified in the Nop58 depletion.

4.11.2 Supplementary Figures

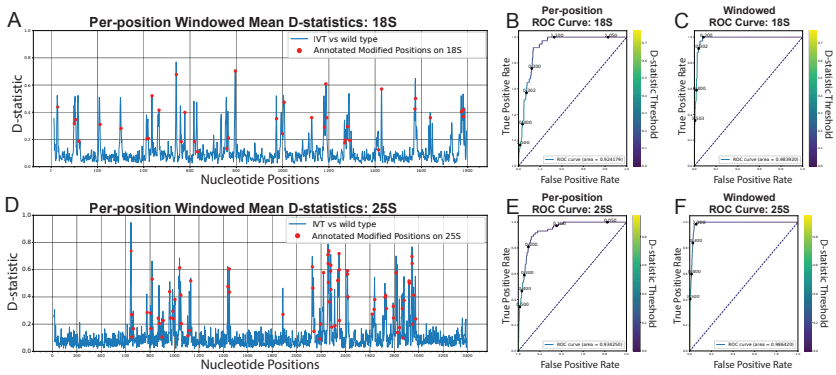

Figure 4.7: De-novo detection of modifications using Tombo. (A-B) Per position, window averaged D-statistic plots from Tombo's sample compare method for yeast 18S (A) and 25S (B) rRNA [108]. The blue line represents the difference between the per-position distributions of the IVT sample vs the wild type sample. The red markers are the location of each annotated modification on the corresponding rRNA4.5).

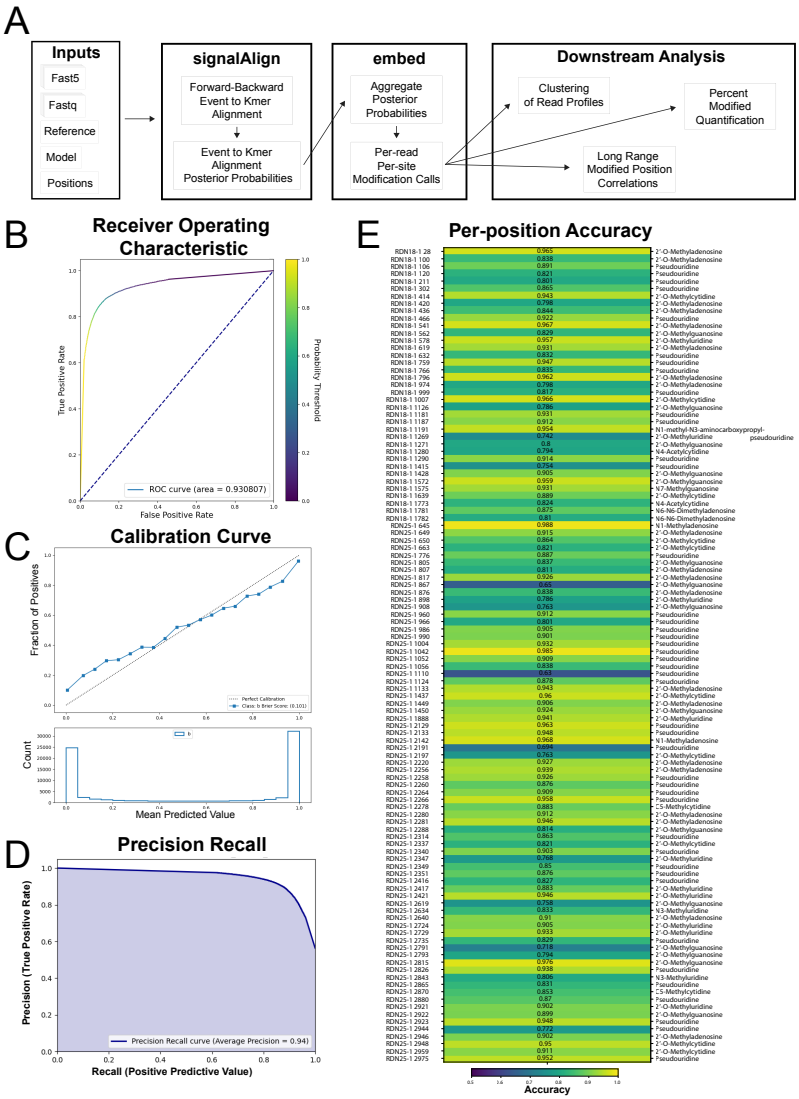

Figure 4.8: SignalAlign pipeline overview, overall accuracy metrics from testing data and per-position model accuracy. (A) Analysis pipeline. (B-E) Testing accuracy metrics of the final model of supervised training. Both training protocol and testing metrics are described in detail in section 4.5. (B) Receiver operating characteristic (ROC) curve and area under the ROC (0.93). (C) Calibration curve showing the fraction of true positives for several ranges of probabilities. The brier score (0.101) is a metric for determining how well a model is calibrated. (D) Precision-recall curve. (E) Per-position accuracy with corresponding modification annotation for each position [163].

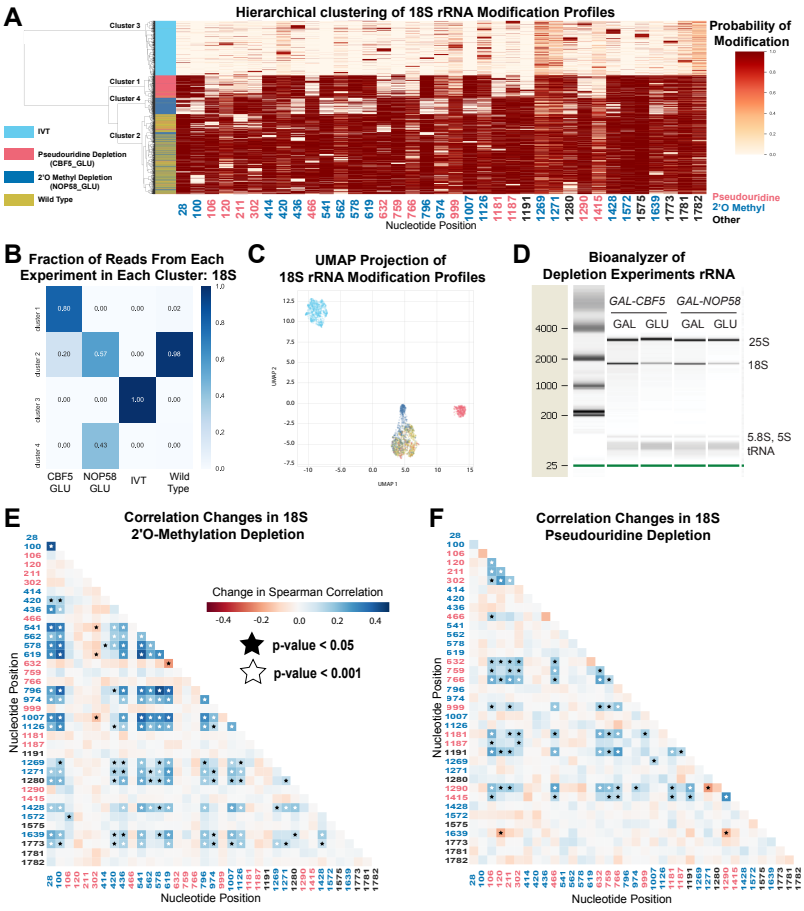

Figure 4.9: (related to Figure 4.1): Clustering and correlation analysis of depletion experiment modification profiles in 18S. (A) Hierarchical clustering of 18S yeast rRNA modification profiles of IVT, wild type, and both pseudouridine and 2'O methyl depletion experiments. (B) Fraction reads from IVT, wild type and both depletion experiments (`CBF5_GAL`, `NOP58_GAL`) in each cluster of 18S rRNA. (C) UMAP visualization of 18S yeast rRNA modification profiles of IVT, wild type, and both pseudouridine and 2'O methyl depletion experiments. (D) Bioanalyzer of comparing levels of 18S and 25S in galactose treated samples (`CBF5_GAL`, `NOP58_GAL`) compared to glucose treated samples (`CBF5_GLU`, `NOP58_GLU`). (E/F) Change in Spearman correlations of 25S reads in 2'O methyl depletion (E) and pseudouridine depletion (F) when compared to wild type. Stars represent significant changes when compared to wild type correlation and significantly different from zero correlation (see section 4.5).

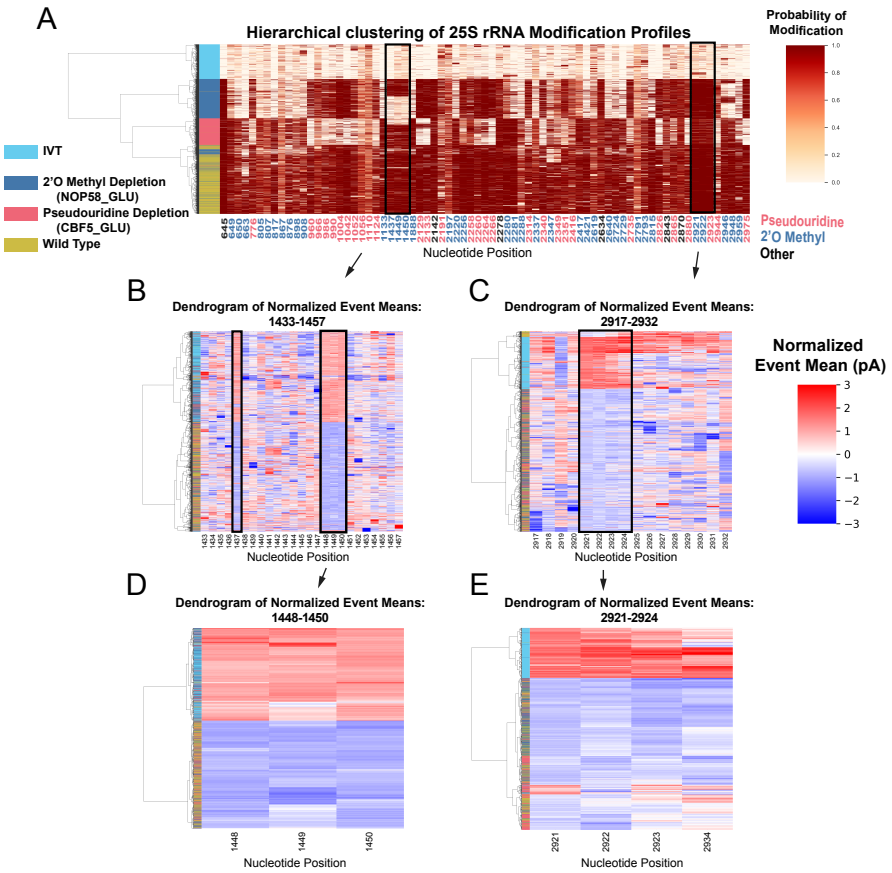

Figure 4.10: (related to Figure 4.1): Clustering of underlying events to search for patterns of modification in the pseudouridine and 2'O methyl depletion experiments. (A) Hierarchical clustering of 25S yeast rRNA modification profiles of IVT, wild type, and both pseudouridine and 2'O methyl depletion experiments. (B-E) Hierarchical clustering of normalized event means aligned to the reference sequence from IVT, wild type, and both pseudouridine and 2'O methyl depletion experiments covering positions 1433 to 1457 (B), 2917-2932 (C), 1448-1450 (D), and 2921-2924 (E) (see section 4.5 and Supplemental Note 4.11.1).

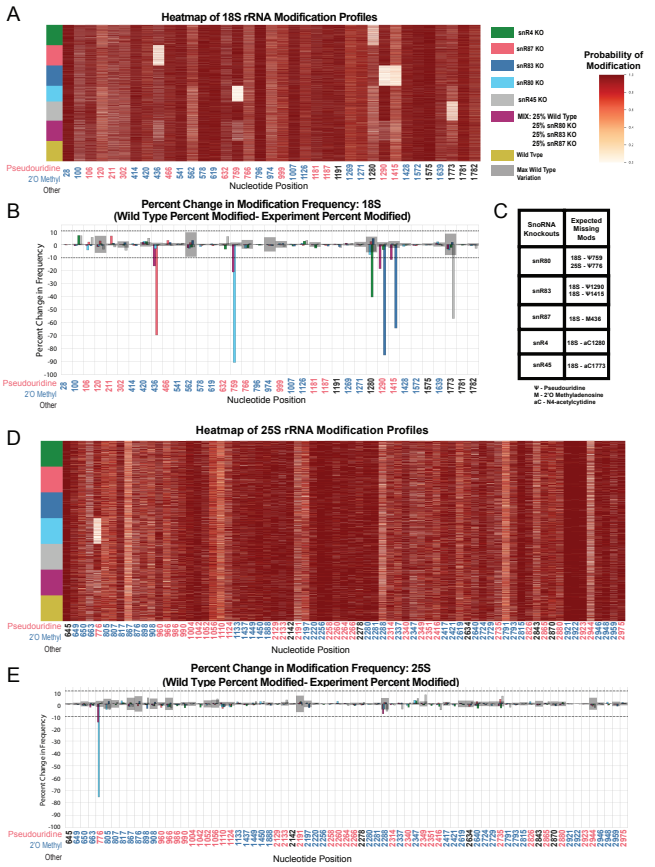

Figure 4.11: (related to Figure 4.2): Heatmaps and percent modification change of snoRNA knockout and mixture experiments. (A) Heatmap of wild type, mixed sample, snR80 KO, snR83 KO, snR87, snR45 and snR4 KO modification profiles of 18S. (B) Mixed sample, snR80 KO, snR83 KO, snR87, snR45 and snR4 KO 18S percent change in modification frequency when compared to wild type. Grey bars indicate the variance of wild type rRNA modification at each position and the black dotted lines represent the maximum variance found at any position. (C) Table of snoRNAs knocked down with the corresponding expected knocked down modifications. (D) Heatmap of wild type, mixed sample, snR80 KO, snR83 KO, snR87, snR45 and snR4 KO modification profiles of 25S. (E) Mixed sample, snR80 KO, snR83 KO, snR87, snR45 and snR4 KO 25S percent change in modification frequency when compared to wild type.

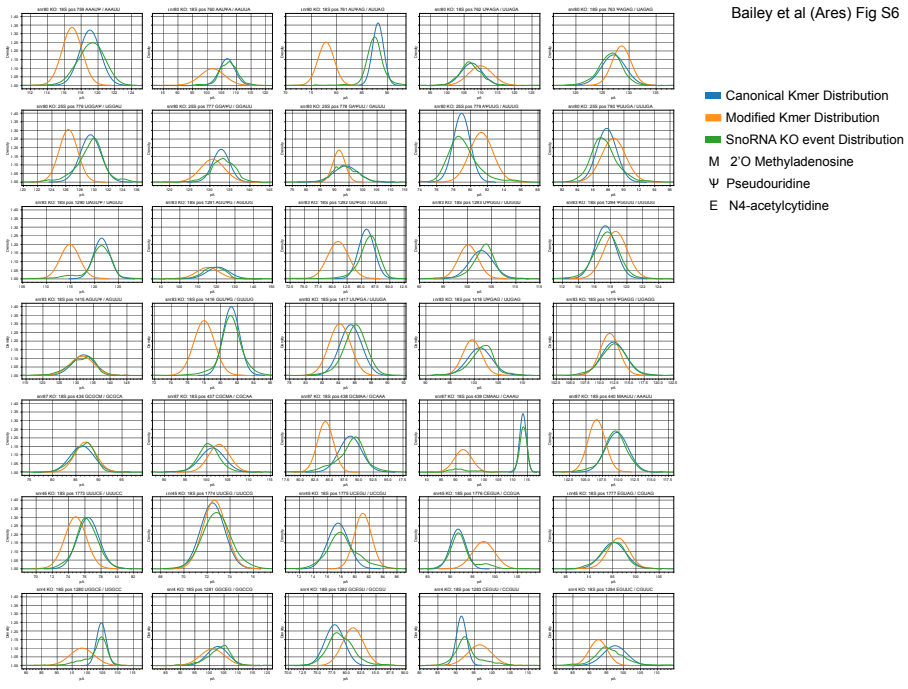

Figure 4.12: Kmer distribution comparison between snoRNA knockout kmer distributions and the trained model kmer distributions. Each figure has the model's canonical kmer distribution, the model's modified kmer distribution and the corresponding snoRNA knockout kernel density estimate (KDE) of all events aligned to that position (see section 4.5). The rows show kmers covering position 759 in 18S from snR80 KO, position 776 in 25S from snR80 KO, position 1290 in 18S from snR83 KO, position 1415 in 18S from snR83 KO, position 436 in 18S from snR87 KO, position 436 in 18S from snR87 KO, position 1773 in 18S from snR45 KO and position 1280 in 18S from snR4 KO.

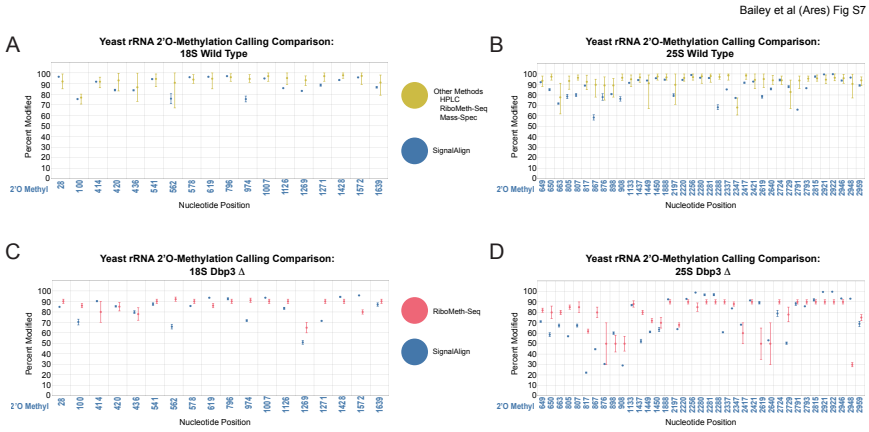

Figure 4.13: Comparison of rRNA 2'O-methylation calling from other modification detection techniques and signalAlign modification detection. (A-B) Comparison between the range of modification percentages called via mass spectrometry [163], HPLC [188], and two RiboMeth-seq approaches [14, 107] vs signalAlign modification percentages of wild type yeast in 18S (A) and 25S (B). (C-D) Comparison between RiboMeth-seq modification percentages [2] and signalAlign modification percentages for the Dbp3 knockout strain in 18S (C) and 25S (D) yeast rRNA. For the combination of several detection approaches, we calculated the minimum, maximum and mean modification percentage from the four papers. For all plots, error bars represent the minimum or maximum percent modification called and circles represent the mean modification percentage.

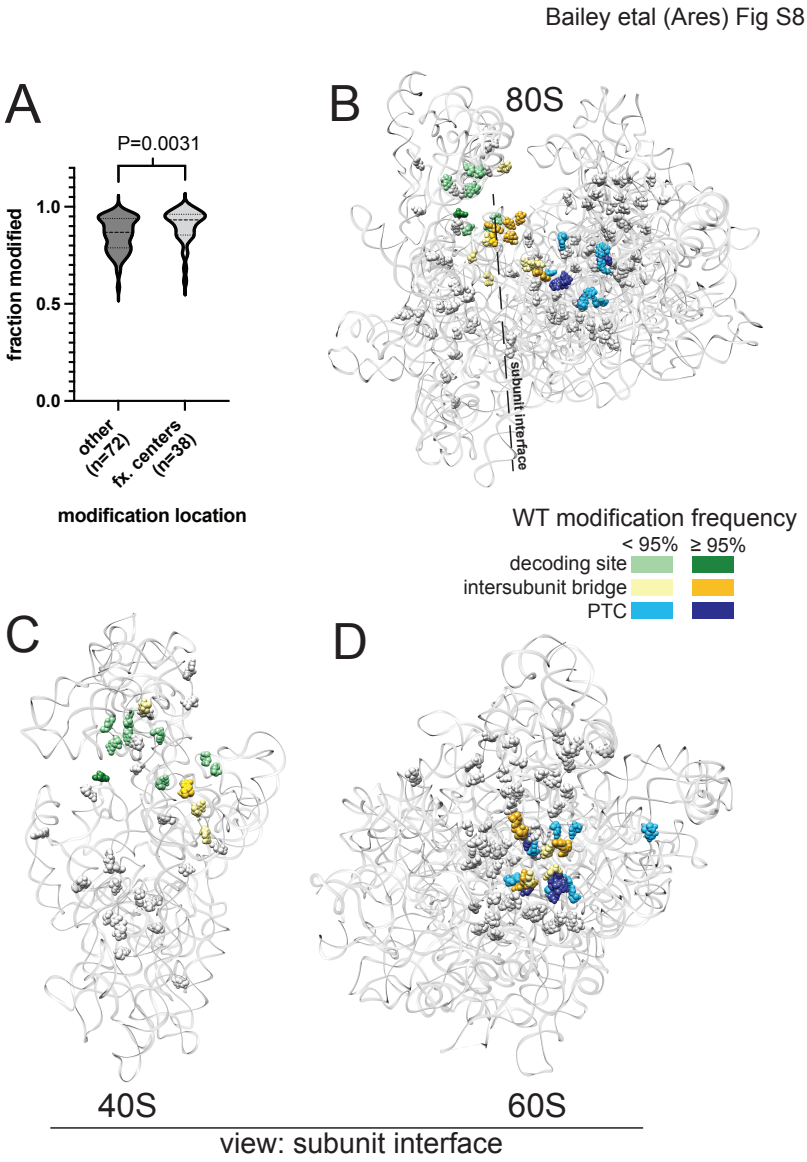

Figure 4.14: Analysis of yeast rRNA modification frequency in relation to functional centers of the ribosome. (A) Distribution of fraction modified for positions within or not within the functional centers of yeast rRNA. Distribution means are significantly (p-value=0.0031) different via a two-sided Mann-Whitney U-test. (B-D) Crystal structure model of wild type S. cerevisiae 80S (B), 40S (C) and 60S (D) rRNA highlighting modification frequency within functional centers. PDB: 4V88 [8].

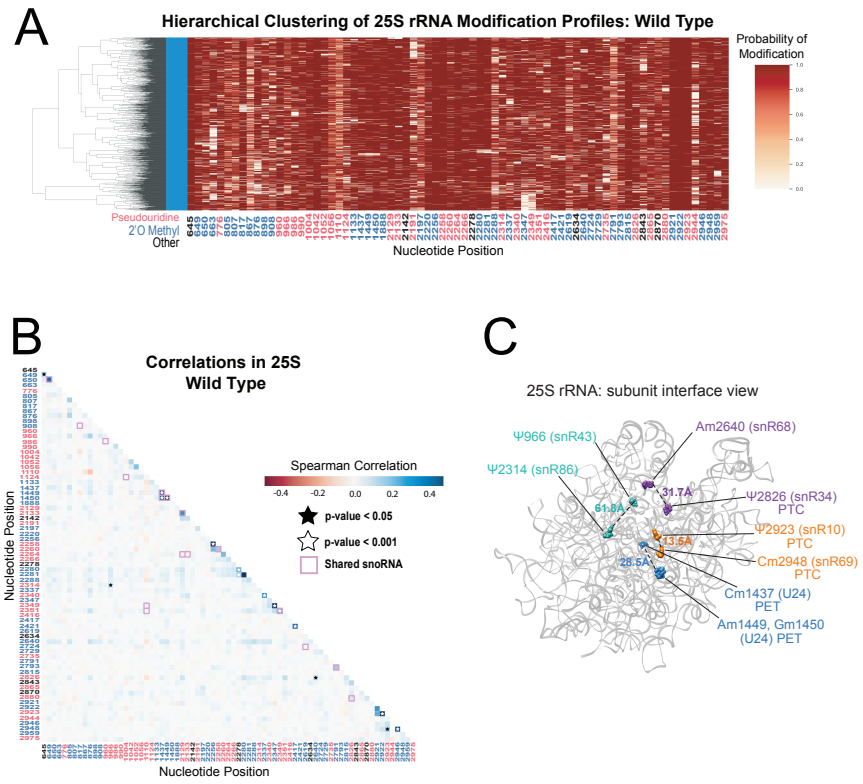

Figure 4.15: (related to Figure 4.2): Yeast 25S rRNA modification profile clustering and correlation analysis. (A) Hierarchical clustering of 25S yeast rRNA modification profiles from wild type yeast. (B) Wild type Spearman correlation of 25S wild type reads. Stars represent significantly different to IVT correlations and significantly different from zero correlation. (C) Crystal structure model of wild type S. cerevisiae 25S rRNA highlighting significant correlated positions. PDB: 4V88 [8]

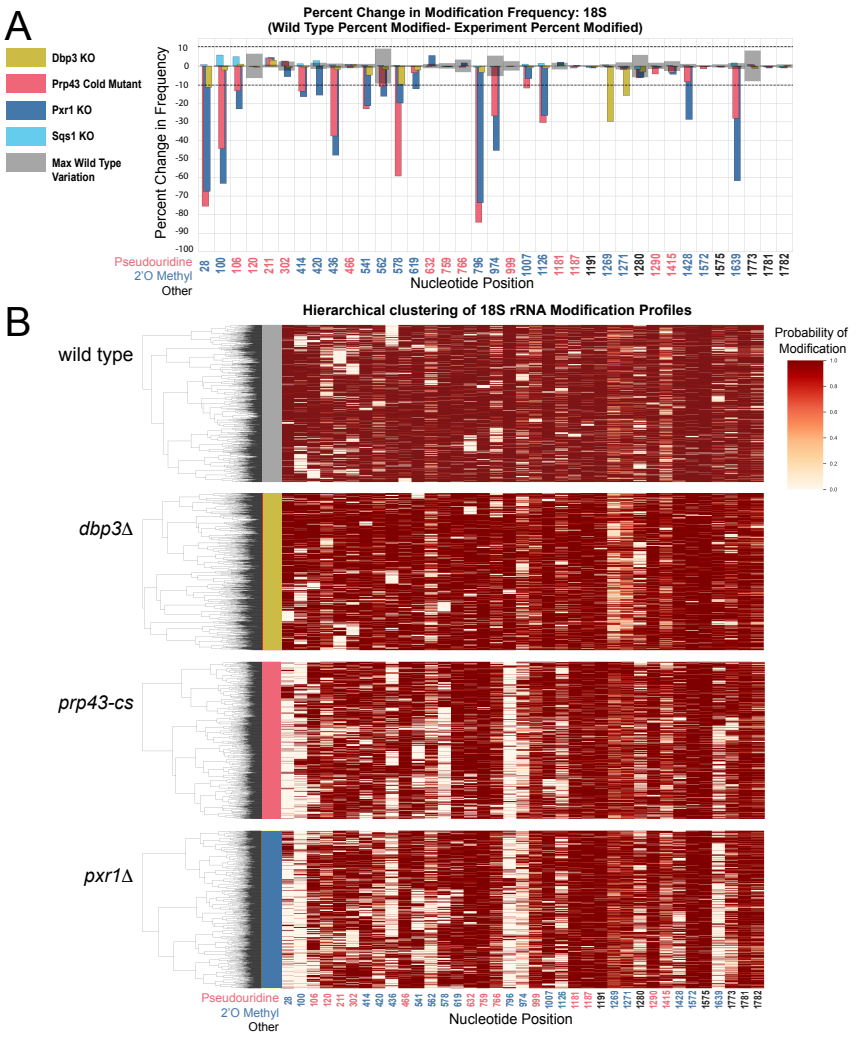

Figure 4.16: (related to Figure 4.3): Clustering of 18S rRNA modification profiles and percent change in modification frequency of helicase mutants Dbp3 and Prp43 and G-patch proteins Pxr1 and Sqs1. (A) Barplots of the difference between wild type modification frequency and Dbp3 KO, Prp43 cold mutant, Pxr1 KO, and Sqs1 KO modification frequencies in 18S yeast rRNA. Grey bars indicate the variance of wild type rRNA modification at each position and the black dotted lines represent the maximum variance. (B) Hierarchical clustering of 18S yeast rRNA modification profiles from wild type, Dbp3 knockout, Prp43 cold mutant, and Pxr1 KO.

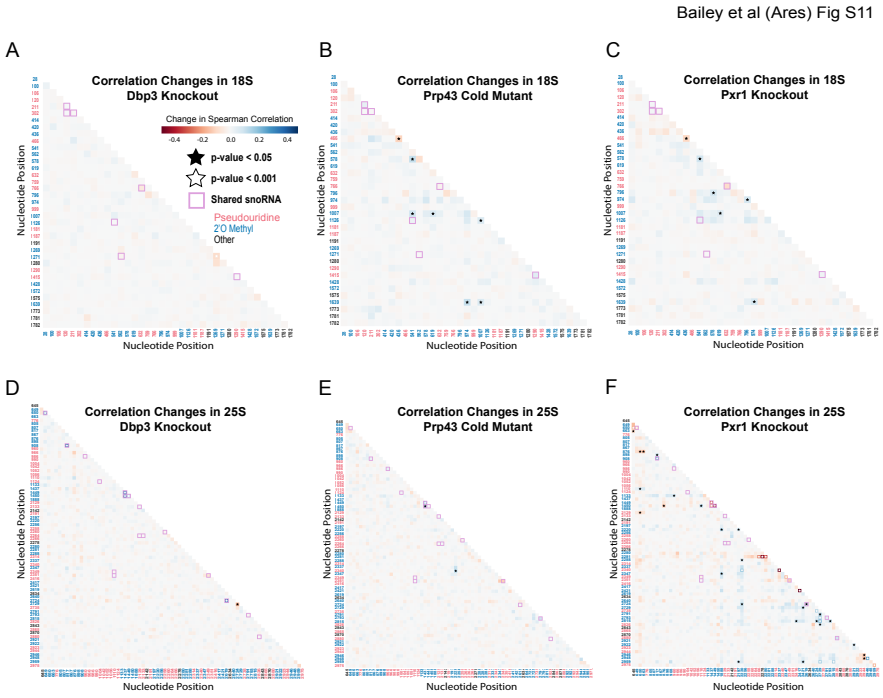

Figure 4.17: (related to Figure 4.3): Correlation analysis of Dbp3 knockout, Prp43 cold mutant Pxr1 knockout. Change in Spearman correlations of 18S (A-C) and 25S (D-E) reads in Dbp3 knockout (A/D), Prp43 cold mutant (B/E), and Pxr1 knockout (C/F) when compared to wild type. Stars represent significant changes when compared to wild type correlation and significantly different from zero correlation.

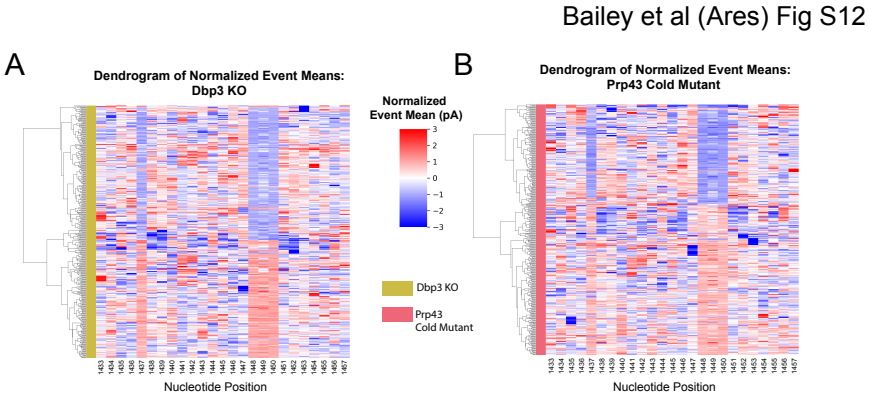

Figure 4.18: Clustering of underlying events to search for patterns of modification in the Dbp3 KO and Prp43 cold mutant. Hierarchical clustering of aligned standardized events from Dbp3 KO (A) and Prp43 cold mutant (B) covering the events from positions 1431 to 1455 (see section 4.5). These positions cover the 3 2'O ribose methylations guided by the snoRNA U24 at positions 1437, 1449 and 1450.

Chapter 5

Supplemental Files

yeast_rrna_supplemental_tables.xlsx

Bibliography

[1] Ingrun Alseth, Bjørn Dalhus, and Magnar Bjørås. Inosine in DNA and RNA. *Current Opinion in Genetics & Development*, 26:116–123, 6 2014.

[2] Gerald Ryan R Aquino, Nicolai Krogh, Philipp Hackert, Roman Martin, Jimena Davila Gallesio, Robert W. Van Nues, Claudia Schneider, Nicholas J Watkins, Henrik Nielsen, Katherine E. Bohnsack, and Markus T Bohnsack. RNA helicase-mediated regulation of snoRNP dynamics on pre-ribosomes and rRNA 2'- O -methylation. *Nucleic Acids Research*, 49(7):4066–4084, 4 2021.

[3] Manuel Ares. Isolation of total RNA from yeast cell cultures. *Cold Spring Harbor Protocols*, 7(10):1082–1086, 10 2012.

[4] Jong Ghut Ashley Aw, Shaun W. Lim, Jia Xu Wang, Finnlay R. P. Lambert, Wen Ting Tan, Yang Shen, Yu Zhang, Pornchai Kaewsapsak, Chenhao Li, Sarah B. Ng, Leah A. Vardy, Meng How Tan, Niranjan Nagarajan, and Yue Wan. Determination of isoform-specific RNA structure with nanopore long reads. *Nature Biotechnology*, 39(3):336–346, 3 2021.

[5] Daipayan Banerjee, Peter M. McDaniel, and Brian C. Rymond. Limited portability of G-patch domains in regulators of the Prp43 RNA helicase required for pre-mRNA splicing and ribosomal RNA maturation in saccharomyces cerevisiae. *Genetics*, 200(1):135–147, 5 2015.

[6] Andrea Bednářová, Marley Hanna, Isabella Durham, Tara VanCleave, Alexis England, Anathbandhu Chaudhuri, and Natraj Krishnan. Lost in Translation: Defects in Transfer RNA Modifications and Neurological Disorders. *Frontiers in Molecular Neuroscience*, 10(May):1–8, 5 2017.

[7] Oguzhan Begik, Morghan C. Lucas, Leszek P. Pryszcz, Jose Miguel Ramirez, Rebeca Medina, Ivan Milenkovic, Sonia Cruciani, Huanle Liu, Helaine Graziele Santos Vieira, Aldema Sas-Chen, John S. Mattick, Schraga Schwartz, and Eva Maria Novoa. Quantitative profiling of pseudouridylation dynamics in native RNAs with nanopore sequencing. *Nature Biotechnology*, 5 2021.

[8] Adam Ben-Shem, Nicolas Garreau de Loubresse, Sergey Melnikov, Lasse Jenner, Gulnara Yusupova, and Marat Yusupov. The Structure of the Eukaryotic Ribosome at 3.0 Å Resolution. *Science*, 334(6062):1524–1529, 12 2011.

[9] Yoav Benjamini and Yosef Hochberg. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300, 1995.

[10] Houda Benrahma, Hicham Charoute, Khaled Lasram, Redouane Boulouiz, Rym Kefi-Ben Ben Atig, Malika Fakiri, Hassan Rouba, Sonia Abdelhak, and Abdel-

hamid Barakat. Association Analysis of IGF2BP2, KCNJ11, and CDKAL1 Polymorphisms with Type 2 Diabetes Mellitus in a Moroccan Population: A Case–Control Study and Meta-analysis. *Biochemical Genetics*, 52(9-10):430–442, 10 2014.

[11] David R. Bentley, Shankar Balasubramanian, Harold P. Swerdlow, Geoffrey P. Smith, John Milton, Clive G. Brown, Kevin P. Hall, Dirk J. Evers, Colin L. Barnes, Helen R. Bignell, Jonathan M. Boutell, Jason Bryant, Richard J. Carter, R. Keira Cheetham, Anthony J. Cox, Darren J. Ellis, Michael R. Flatbush, Niall A. Gormley, Sean J. Humphray, Leslie J. Irving, Mirian S. Karbelashvili, Scott M. Kirk, Heng Li, Xiaohai Liu, Klaus S. Maisinger, Lisa J. Murray, Bojan Obradovic, Tobias Ost, Michael L. Parkinson, Mark R. Pratt, Isabelle M.J. Rasolonjatovo, Mark T. Reed, Roberto Rigatti, Chiara Rodighiero, Mark T. Ross, Andrea Sabot, Subramanian V. Sankar, Aylwyn Scally, Gary P. Schroth, Mark E. Smith, Vincent P. Smith, Anastassia Spiridou, Peta E. Torrance, Svilen S. Tzonev, Eric H. Vermaas, Klaudia Walter, Xiaolin Wu, Lu Zhang, Mohammed D. Alam, Carole Anastasi, Ify C. Aniebo, David M.D. Bailey, Iain R. Bancarz, Saibal Banerjee, Selena G. Barbour, Primo A. Baybayan, Vincent A. Benoit, Kevin F. Benson, Claire Bevis, Phillip J. Black, Asha Boodhun, Joe S. Brennan, John A. Bridgham, Rob C. Brown, Andrew A. Brown, Dale H. Buermann, Abass A. Bundu, James C. Burrows, Nigel P. Carter, Nestor Castillo, Maria Chiara E. Catenazzi, Simon Chang, R. Neil Cooley, Natasha R. Crake, Olubunmi O. Dada, Konstantinos D. Diakoumakos, Be-

len Dominguez-Fernandez, David J. Earnshaw, Ugonna C. Egbujor, David W. Elmore, Sergey S. Etchin, Mark R. Ewan, Milan Fedurco, Louise J. Fraser, Karin V. Fuentes Fajardo, W. Scott Furey, David George, Kimberley J. Gietzen, Colin P. Goddard, George S. Golda, Philip A. Granieri, David E. Green, David L. Gustafson, Nancy F. Hansen, Kevin Harnish, Christian D. Haudenschild, Narinder I. Heyer, Matthew M. Hims, Johnny T. Ho, Adrian M. Horgan, Katya Hoschler, Steve Hurwitz, Denis V. Ivanov, Maria Q. Johnson, Terena James, T. A. Huw Jones, Gyoung Dong Kang, Tzvetana H. Kerelska, Alan D. Kersey, Irina Khrebtukova, Alex P. Kindwall, Zoya Kingsbury, Paula I. Kokko-Gonzales, Anil Kumar, Marc A. Laurent, Cynthia T. Lawley, Sarah E. Lee, Xavier Lee, Arnold K. Liao, Jennifer A. Loch, Mitch Lok, Shujun Luo, Radhika M. Mammen, John W. Martin, Patrick G. McCauley, Paul McNitt, Parul Mehta, Keith W. Moon, Joe W. Mullens, Taksina Newington, Zemin Ning, Bee Ling Ng, Sonia M. Novo, Michael J. O'Neill, Mark A. Osborne, Andrew Osnowski, Omead Ostadan, Lambros L. Paraschos, Lea Pickering, Andrew C. Pike, Alger C. Pike, D. Chris Pinkard, Daniel P. Pliskin, Joe Podhasky, Victor J. Quijano, Come Raczy, Vicki H. Rae, Stephen R. Rawlings, Ana Chiva Rodriguez, Phyllida M. Roe, John Rogers, Maria C. Rogert Bacigalupo, Nikolai Romanov, Anthony Romieu, Rithy K. Roth, Natalie J. Rourke, Silke T. Ruediger, Eli Rusman, Raquel M. Sanches-Kuiper, Martin R. Schenker, Josefina M. Seoane, Richard J. Shaw, Mitch K. Shiver, Steven W. Short, Ning L. Sizto, Johannes P. Sluis, Melanie A. Smith, Jean Ernest Sohna Sohna, Eric J. Spence,

Kim Stevens, Neil Sutton, Lukasz Szajkowski, Carolyn L. Tregidgo, Gerardo Turcatti, Stephanie Vandevondele, Yuli Verhovsky, Selene M. Virk, Suzanne Wakelin, Gregory C. Walcott, Jingwen Wang, Graham J. Worsley, Juying Yan, Ling Yau, Mike Zuerlein, Jane Rogers, James C. Mullikin, Matthew E. Hurles, Nick J. McCooke, John S. West, Frank L. Oaks, Peter L. Lundberg, David Klenerman, Richard Durbin, and Anthony J. Smith. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, 456(7218):53–59, 2008.

[12] Chad R. Bernier, Anton S. Petrov, Chris C. Waterbury, James Jett, Fengbo Li, Larry E. Freil, Xiao Xiong, Lan Wang, Blacki L. R. Migliozzi, Eli Hershkovits, Yuzhen Xue, Chiaolong Hsiao, Jessica C. Bowman, Stephen C. Harvey, Martha A. Grover, Zachary J. Wartell, and Loren Dean Williams. RiboVision suite for visualization and analysis of ribosomes. *Faraday Discuss*, 169:195–207, 5 2014.

[13] Nidhi Bhutani, David M. Burns, and Helen M. Blau. DNA Demethylation Dynamics. *Cell*, 146(6):866–872, 9 2011.

[14] Ulf Birkedal, Mikkel Christensen-Dalsgaard, Nicolai Krogh, Radhakrishnan Sabarinathan, Jan Gorodkin, and Henrik Nielsen. Profiling of Ribose Methylations in RNA by High-Throughput Sequencing. *Angewandte Chemie International Edition*, 54(2):n/a–n/a, 11 2014.

[15] Pietro Boccaletto, Magdalena A. Machnicka, Elzbieta Purta, Pawek Pitkowski,

Blazej Baginski, Tomasz K. Wirecki, Valérie de Crécy-Lagard, Robert Ross, Patrick A. Limbach, Annika Kotter, Mark Helm, Janusz M. Bujnicki, Paweł Piątkowski, Błażej Bagiński, Tomasz K. Wirecki, Valérie de Crécy-Lagard, Robert Ross, Patrick A. Limbach, Annika Kotter, Mark Helm, and Janusz M. Bujnicki. MODOMICS: a database of RNA modification pathways. 2017 update. *Nucleic Acids Research*, 46(D1):D303–D307, 1 2018.

[16] Michael A Boemo. DNAscent v2: Detecting Replication Forks in Nanopore Sequencing Data with Deep Learning. *bioRxiv*, page 2020.11.04.368225, 1 2020.

[17] Vladimir Boza, Brona Brejova, and Tomas Vinar. DeepNano: Deep Recurrent Neural Networks for Base Calling in MinION Nanopore Reads. *arXiv*, pages 1–12, 2016.

[18] R C Brand, J Klootwijk, R J Planta, and B E H Maden. Biosynthesis of a hypermodified nucleotide ¡i¿in Saccharomyces carlsbergensis¡/i¿ 17S and HeLa-cell 18S ribosomal ribonucleic acid. *Biochemical Journal*, 169(1):71–77, 1 1978.

[19] Achim Breiling and Frank Lyko. Epigenetic regulatory functions of DNA modifications: 5-methylcytosine and beyond. *Epigenetics and Chromatin*, 8(1):1–9, 2015.

[20] Clive G Brown. No Thanks, I've Already Got One, 2016.

[21] Morton B. Brown. 400: A Method for Combining Non-Independent, One-Sided Tests of Significance. *Biometrics*, 31(4):987, 12 1975.

[22] M. E. Cardenas, N. S. Cutler, M. C. Lorenz, C. J. Di Como, and J. Heitman. The TOR signaling cascade regulates gene expression in response to nutrients. *Genes & Development*, 13(24):3271–3279, 12 1999.

[23] Clement T. Y. Chan, Madhu Dyavaiah, Michael S. DeMott, Koli Taghizadeh, Peter C. Dedon, and Thomas J. Begley. A Quantitative Systems Approach Reveals Dynamic Control of tRNA Modifications during Cellular Stress. *PLoS Genetics*, 6(12):e1001247, 12 2010.

[24] Karen B Chapman and Jef D Boeke. Isolation and characterization of the gene encoding yeast debranching enzyme. *Cell*, 65(3):483–492, 5 1991.

[25] Jiahua Chen and Pengfei Li. Hypothesis test for normal mixture models: The EM approach. *The Annals of Statistics*, 37(5A):2523–2542, 10 2009.

[26] Jiahua Chen, Pengfei Li, and Yuejiao Fu. Inference on the order of a normal mixture. *Journal of the American Statistical Association*, 107(499):1096–1105, 2012.

[27] Yan Ling Chen, Régine Capeyrou, Odile Humbert, Saïda Mouffok, Yasmine Al Kadri, Simon Lebaron, Anthony K Henras, and Yves Henry. The telomerase inhibitor Gno1p/PINX1 activates the helicase Prp43p during ribosome biogenesis. *Nucleic Acids Research*, 42(11):7330–7345, 6 2014.

[28] Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning Phrase Rep-

resentations using RNN Encoder–Decoder for Statistical Machine Translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Stroudsburg, PA, USA, 2014. Association for Computational Linguistics.

[29] Junhong Choi, Rosslyn Grosely, Arjun Prabhakar, Christopher P. Lapointe, Jinfan Wang, and Joseph D. Puglisi. How Messenger RNA and Nascent Chain Sequences Regulate Translation Elongation. *Annual Review of Biochemistry*, 87(1):421–449, 6 2018.

[30] François Chollet and others. Keras. https://keras.io, 2015.

[31] Henning Christian, Romina V Hofele, Henning Urlaub, and Ralf Ficner. Insights into the activation of the helicase Prp43 by biochemical studies and structural mass spectrometry. *Nucleic Acids Research*, 42(2):1162–1179, 1 2014.

[32] D Joshua Combs, Roland J Nagel, Manuel Ares, and Scott W Stevens. Prp43p Is a DEAH-Box Spliceosome Disassembly Factor Essential for Ribosome Biogenesis. *Molecular and Cellular Biology*, 26(2):523–534, 1 2006.

[33] Salvatore Cortellino, Jinfei Xu, Mara Sannai, Robert Moore, Elena Caretti, Madeleine Le Coz, Karthik Devarajan, Andy Wessels, Dianne Soprano, K Abramowitz, Marisa S Bartolomei, Florian Rambow, Maria Rosaria Bassi, Maurizio Fanciulli, Catherine Renner, Andres J Klein-szanto, Dominique Kobi, Irwin Davidson, Christophe Alberti, and Lionel Larue. Demethylation by Linked Deamination-Base Excision Repair. *Cell*, 146(1):67–79, 2011.

[34] Matei David, L. J. Dursi, Delia Yao, Paul C. Boutros, and Jared T. Simpson. Nanocall: An open source basecaller for Oxford Nanopore sequencing data. *Bioinformatics*, 33(1):49–55, 2017.

[35] David Deamer, Mark Akeson, and Daniel Branton. Three decades of nanopore sequencing. *Nature Biotechnology*, 34(5):518–524, 5 2016.

[36] Benjamin Delatte, Fei Wang, Long Vo Ngoc, Evelyne Collignon, Elise Bonvin, Rachel Deplus, Emilie Calonne, Bouchra Hassabi, Pascale Putmans, Stephan Awe, Collin Wetzel, Judith Kreher, Romuald Soin, Catherine Creppe, Patrick A Limbach, Cyril Gueydan, Véronique Kruys, Alexander Brehm, Svetlana Minakhina, Matthieu Defrance, Ruth Steward, and François Fuks. Transcriptome-wide distribution and function of RNA hydroxymethylcytosine. *Science*, 351(6270):282–285, 1 2016.

[37] Sylvain Delaunay and Michaela Frye. RNA modifications regulating cell fate in cancer. *Nature Cell Biology*, 21(5):552–559, 2019.

[38] Sylvain Delaunay, Francesca Rapino, Lars Tharun, Zhaoli Zhou, Lukas Heukamp, Martin Termathe, Kateryna Shostak, Iva Klevernic, Alexandra Florin, Hadrien Desmecht, Christophe J. Desmet, Laurent Nguyen, Sebastian A. Leidel, Anne E. Willis, Reinhard Büttner, Alain Chariot, and Pierre Close. Elp3 links tRNA modification to IRES-dependent translation of LEF1 to sustain metastasis in breast cancer. *The Journal of Experimental Medicine*, 213(11):2503–2523, 10 2016.

[39] Ian M. Derrington, Tom Z. Butler, Marcus D. Collins, Elizabeth Manrao, Mikhail Pavlenok, Michael Niederweis, and Jens H. Gundlach. Nanopore DNA sequencing with MspA. *Proceedings of the National Academy of Sciences*, 107(37):16060–16065, 9 2010.

[40] Hongxu Ding, Ioannis Anastopoulos, Andrew D. Bailey, Joshua Stuart, and Benedict Paten. Towards inferring nanopore sequencing ionic currents from nucleotide chemical structures. *Nature Communications*, 12(1):6545, 12 2021.

[41] Hongxu Ding, Andrew D Bailey, Miten Jain, Hugh Olsen, and Benedict Paten. Gaussian mixture model-based unsupervised nucleotide modification number detection using nanopore-sequencing readouts. *Bioinformatics*, 36(19):4928–4934, 12 2020.

[42] Dan Dominissini, Sharon Moshitch-Moshkovitz, Schraga Schwartz, Mali Salmon-Divon, Lior Ungar, Sivan Osenberg, Karen Cesarkas, Jasmine Jacob-Hirsch, Ninette Amariglio, Martin Kupiec, Rotem Sorek, and Gideon Rechavi. Topology of the human and mouse m6A RNA methylomes revealed by m6A-seq. *Nature*, 485(7397):201–206, 2012.

[43] Dan Dominissini, Sigrid Nachtergaele, Sharon Moshitch-Moshkovitz, Eyal Peer, Nitzan Kol, Moshe Shay Ben-Haim, Qing Dai, Ayelet Di Segni, Mali Salmon-Divon, Wesley C Clark, Guanqun Zheng, Tao Pan, Oz Solomon, Eran Eyal, Vera Hershkovitz, Dali Han, Louis C Doré, Ninette Amariglio, Gideon Rechavi,

and Chuan He. The dynamic N1 -methyladenosine methylome in eukaryotic messenger RNA. *Nature*, 530(7591):441–446, 2 2016.

[44] Ian Dunham, Anshul Kundaje, Shelley F. Aldred, Patrick J. Collins, Carrie A. Davis, Francis Doyle, Charles B. Epstein, Seth Frietze, Jennifer Harrow, Rajinder Kaul, Jainab Khatun, Bryan R. Lajoie, Stephen G. Landt, Bum-Kyu Kyu Lee, Florencia Pauli, Kate R. Rosenbloom, Peter Sabo, Alexias Safi, Amartya Sanyal, Noam Shoresh, Jeremy M. Simon, Lingyun Song, Nathan D. Trinklein, Robert C. Altshuler, Ewan Birney, James B. Brown, Chao Cheng, Sarah Djebali, Xianjun Dong, Jason Ernst, Terrence S. Furey, Mark Gerstein, Belinda Giardine, Melissa Greven, Ross C. Hardison, Robert S. Harris, Javier Herrero, Michael M. Hoffman, Sowmya Iyer, Manolis Kellis, Pouya Kheradpour, Timo Lassmann, Qunhua Li, Xinying Lin, Georgi K. Marinov, Angelika Merkel, Ali Mortazavi, Stephanie L. Stephen C.J. J Parker, Timothy E. Reddy, Joel Rozowsky, Felix Schlesinger, Robert E. Thurman, Jie Wang, Lucas D. Ward, Troy W. Whitfield, Steven P. Wilder, Weisheng Wu, Hualin S. Xi, Kevin Y. Yip, Jiali Zhuang, Bradley E. Bernstein, Eric D. Green, Chris Gunter, Michael Snyder, Michael J. Pazin, Rebecca F. Lowdon, Laura A.L. L Dillon, Leslie B. Adams, Caroline J. Kelly, Julia Zhang, Judith R. Wexler, Peter J. Good, Elise A. Feingold, Gregory E. Crawford, Job Dekker, Laura Elnitski, Peggy J. Farnham, Morgan C. Giddings, Thomas R. Gingeras, Roderic Guigó, Timothy J. Hubbard, W. James Kent, Jason D. Lieb, Elliott H. Margulies, Richard M. Myers, John A. Stamatoyannopoulos, Scott A. Tenenbaum, Zhip-

ing Weng, Kevin P. White, Barbara Wold, Yanbao Yu, John Wrobel, Brian A. Risk, Harsha P. Gunawardena, Heather C. Kuiper, Christopher W. Maier, Ling Xie, Xian Chen, Tarjei S. Mikkelsen, Shawn Gillespie, Alon Goren, Oren Ram, Xiaolan Zhang, Li Wang, Robbyn Issner, Michael J. Coyne, Timothy Durham, Manching Ku, Thanh Truong, Matthew L. Eaton, Alex Dobin, Andrea Tanzer, Julien Lagarde, Wei Lin, Chenghai Xue, Brian A. Williams, Chris Zaleski, Maik Röder, Felix Kokocinski, Rehab F. Abdelhamid, Tyler Alioto, Igor Antoshechkin, Michael T. Baer, Philippe Batut, Ian Bell, Kimberly Bell, Sudipto Chakrabortty, Jacqueline Chrast, Joao Curado, Thomas Derrien, Jorg Drenkow, Erica Dumais, Jackie Dumais, Radha Duttagupta, Megan Fastuca, Kata Fejes-Toth, Pedro Ferreira, Sylvain Foissac, Melissa J. Fullwood, Hui Gao, David Gonzalez, Assaf Gordon, Cédric Howald, Sonali Jha, Rory Johnson, Philipp Kapranov, Brandon King, Colin Kingswood, Guoliang Li, Oscar J. Luo, Eddie Park, Jonathan B. Preall, Kimberly Presaud, Paolo Ribeca, Daniel Robyr, Xiaoan Ruan, Michael Sammeth, Kuljeet Singh Sandhu, Lorain Schaeffer, Lei-Hoon Hoon See, Atif Shahab, Jorgen Skancke, Ana Maria Suzuki, Hazuki Takahashi, Hagen Tilgner, Diane Trout, Nathalie Walters, Hao Huaien Wang, Yoshihide Hayashizaki, Alexandre Reymond, Stylianos E. Antonarakis, Gregory J. Hannon, Yijun Ruan, Piero Carninci, Cricket A. Sloan, Katrina Learned, Venkat S. Malladi, Matthew C. Wong, Galt P. Barber, Melissa S. Cline, Timothy R. Dreszer, Steven G. Heitner, Donna Karolchik, Vanessa M. Kirkup, Laurence R. Meyer, Jeffrey C. Long, Morgan Maddren, Brian J. Raney, Linda L.

Grasfeder, Paul G. Giresi, Anna Battenhouse, Nathan C. Sheffield, Kimberly A. Showers, Darin London, Akshay A. Bhinge, Christopher Shestak, Matthew R. Schaner, Seul Ki Kim, Zhuzhu Zhengdong Zhancheng Zhuzhu Z. Zhang, Piotr A. Mieczkowski, Joanna O. Mieczkowska, Zheng Liu, Ryan M. McDaniell, Yunyun Ni, Naim U. Rashid, Min Jae Kim, Sheera Adar, Zhuzhu Zhengdong Zhancheng Zhuzhu Z. Zhang, Tianyuan Wang, Deborah Winter, Damian Keefe, Vishwanath R. Iyer, Meizhen Zheng, Ping Wang, Jason Gertz, Jost Vielmetter, E. Christopher Partridge, Katherine E. Varley, Clarke Gasper, Anita Bansal, Shirley Pepke, Preti Jain, Henry Amrhein, Kevin M. Bowling, Michael Anaya, Marie K. Cross, Michael A. Muratet, Kimberly M. Newberry, Kenneth McCue, Amy S. Nesmith, Katherine I. Fisher-Aylor, Barbara Pusey, Gilberto DeSalvo, Stephanie L. Stephen C.J. J Parker, Sreeram Suganthi Balasubramanian, Nicholas S. Davis, Sarah K. Meadows, Tracy Eggleston, J. Scott Newberry, Shawn E. Levy, Devin M. Absher, Wing H. Wong, Matthew J. Blow, Axel Visel, Len A. Pennachio, Hanna M. Petrykowska, Alexej Abyzov, Bronwen Aken, Daniel Barrell, Gemma Barson, Andrew Berry, Alexandra Bignell, Veronika Boychenko, Giovanni Bussotti, Claire Davidson, Gloria Despacio-Reyes, Mark Diekhans, Iakes Ezkurdia, Adam Frankish, James Gilbert, Jose Manuel Gonzalez, Ed Griffiths, Rachel Harte, David A. Hendrix, Toby Hunt, Irwin Jungreis, Mike Kay, Ekta Khurana, Jing Leng, Michael F. Lin, Jane Loveland, Zhi Lu, Deepa Manthravadi, Marco Mariotti, Jonathan Mudge, Gaurab Mukherjee, Cedric Notredame, Baikang Pei, Jose Manuel Rodriguez, Gary Saun-

ders, Andrea Sboner, Stephen Searle, Cristina Sisu, Catherine Snow, Charlie Steward, Electra Tapanari, Michael L. Tress, Marijke J. Van Baren, Stefan Washietl, Laurens Wilming, Amonida Zadissa, Zhuzhu Zhengdong Zhancheng Zhuzhu Z. Zhang, Michael Brent, David Haussler, Alfonso Valencia, Nick Addleman, Roger P. Alexander, Raymond K. Auerbach, Sreeram Suganthi Balasubramanian, Keith Bettinger, Nitin Bhardwaj, Alan P. Boyle, Alina R. Cao, Philip Cayting, Alexandra Charos, Yong Cheng, Catharine Eastman, Ghia Euskirchen, Joseph D. Fleming, Fabian Grubert, Lukas Habegger, Manoj Hariharan, Arif Harmanci, Sushma Iyengar, Victor X. Jin, Konrad J. Karczewski, Maya Kasowski, Phil Lacroute, Hugo Lam, Nathan Lamarre-Vincent, Jin Lian, Marianne Lindahl-Allen, Renqiang Min, Benoit Miotto, Hannah Monahan, Zarmik Moqtaderi, Xinmeng J. Mu, Henriette O'Geen, Zhengqing Ouyang, Dorrelyn Patacsil, Debasish Raha, Lucia Ramirez, Brian Reed, Minyi Shi, Teri Slifer, Heather Witt, Linfeng Wu, Xiaoqin Xu, Koon Kiu Yan, Xinqiong Yang, Kevin Struhl, Sherman M. Weissman, Luiz O. Penalva, Subhradip Karmakar, Raj R. Bhanvadia, Alina Choudhury, Marc Domanus, Lijia Ma, Jennifer Moran, Alec Victorsen, Thomas Auer, Lazaro Centanin, Michael Eichenlaub, Franziska Gruhl, Stephan Heermann, Burkhard Hoeckendorf, Daigo Inoue, Tanja Kellner, Stephan Kirchmaier, Claudia Mueller, Robert Reinhardt, Lea Schertel, Stephanie Schneider, Rebecca Sinn, Beate Wittbrodt, Jochen Wittbrodt, Gaurav Jain, Gayathri Balasundaram, Daniel L. Bates, Rachel Byron, Theresa K. Canfield, Morgan J. Diegel, Douglas Dunn, Abigail K. Ebersol, Tristan Frum,

Kavita Garg, Erica Gist, R. Scott Hansen, Lisa Boatman, Eric Haugen, Richard Humbert, Audra K. Johnson, Ericka M. Johnson, Tattyana V. Kutyavin, Kristen Lee, Dimitra Lotakis, Matthew T. Maurano, Shane J. Neph, Fiedencio V. Neri, Eric D. Nguyen, Hongzhu Qu, Alex P. Reynolds, Vaughn Roach, Eric Rynes, Minerva E. Sanchez, Richard S. Sandstrom, Anthony O. Shafer, Andrew B. Stergachis, Sean Thomas, Benjamin Vernot, Jeff Vierstra, Shinny Vong, Hao Huaien Wang, Molly A. Weaver, Yongqi Yan, Miaohua Zhang, Joshua M. Akey, Michael Bender, Michael O. Dorschner, Mark Groudine, Michael J. MacCoss, Patrick Navas, George Stamatoyannopoulos, Kathryn Beal, Alvis Brazma, Paul Flicek, Nathan Johnson, Margus Lukk, Nicholas M. Luscombe, Daniel Sobral, Juan M. Vaquerizas, Serafim Batzoglou, Arend Sidow, Nadine Hussami, Sofia Kyriazopoulou-Panagiotopoulou, Max W. Libbrecht, Marc A. Schaub, Webb Miller, Peter J. Bickel, Balazs Banfai, Nathan P. Boley, Haiyan Huang, Jingyi Jessica Li, William Stafford Noble, Jeffrey A. Bilmes, Orion J. Buske, Avinash D. Sahu, Peter V. Kharchenko, Peter J. Park, Dannon Baker, James Taylor, Lucas Lochovsky, Ian Dunham, Jason Ernst, Terrence S. Furey, Mark Gerstein, Belinda Giardine, Melissa Greven, Ross C. Hardison, Robert S. Harris, Javier Herrero, Michael M. Hoffman, Sowmya Iyer, Manolis Kellis, Jainab Khatun, Pouya Kheradpour, Anshul Kundaje, Timo Lassmann, Qunhua Li, Xinying Lin, Georgi K. Marinov, Angelika Merkel, Ali Mortazavi, Stephanie L. Stephen C.J. J Parker, Timothy E. Reddy, Joel Rozowsky, Felix Schlesinger, Robert E. Thurman, Jie Wang, Lucas D. Ward, Troy W. Whitfield, Steven P.

Wilder, Weisheng Wu, Hualin S. Xi, Kevin Y. Yip, Jiali Zhuang, Bradley E. Bernstein, Ewan Birney, Ian Dunham, Eric D. Green, Chris Gunter, Michael Snyder, Michael J. Pazin, Rebecca F. Lowdon, Laura A.L. L Dillon, Leslie B. Adams, Caroline J. Kelly, Julia Zhang, Judith R. Wexler, Eric D. Green, Peter J. Good, Elise A. Feingold, Bradley E. Bernstein, Ewan Birney, Gregory E. Crawford, Job Dekker, Laura Elnitski, Peggy J. Farnham, Mark Gerstein, Morgan C. Giddings, Thomas R. Gingeras, Eric D. Green, Roderic Guigó, Ross C. Hardison, Timothy J. Hubbard, Manolis Kellis, W. James Kent, Jason D. Lieb, Elliott H. Margulies, Richard M. Myers, Michael Snyder, John A. Stamatoyannopoulos, Scott A. Tenenbaum, Zhiping Weng, Kevin P. White, Barbara Wold, Jainab Khatun, Yanbao Yu, John Wrobel, Brian A. Risk, Harsha P. Gunawardena, Heather C. Kuiper, Christopher W. Maier, Ling Xie, Xian Chen, Morgan C. Giddings, Bradley E. Bernstein, Charles B. Epstein, Noam Shoresh, Jason Ernst, Pouya Kheradpour, Tarjei S. Mikkelsen, Shawn Gillespie, Alon Goren, Oren Ram, Xiaolan Zhang, Li Wang, Robbyn Issner, Michael J. Coyne, Timothy Durham, Manching Ku, Thanh Truong, Lucas D. Ward, Robert C. Altshuler, Matthew L. Eaton, Manolis Kellis, Sarah Djebali, Carrie A. Davis, Angelika Merkel, Alex Dobin, Timo Lassmann, Ali Mortazavi, Andrea Tanzer, Julien Lagarde, Wei Lin, Felix Schlesinger, Chenghai Xue, Georgi K. Marinov, Jainab Khatun, Brian A. Williams, Chris Zaleski, Joel Rozowsky, Maik Röder, Felix Kokocinski, Rehab F. Abdelhamid, Tyler Alioto, Igor Antoshechkin, Michael T. Baer, Philippe Batut, Ian Bell, Kimberly Bell, Sudipto

Chakrabortty, Xian Chen, Jacqueline Chrast, Joao Curado, Thomas Derrien, Jorg Drenkow, Erica Dumais, Jackie Dumais, Radha Duttagupta, Megan Fastuca, Kata Fejes-Toth, Pedro Ferreira, Sylvain Foissac, Melissa J. Fullwood, Hui Gao, David Gonzalez, Assaf Gordon, Harsha P. Gunawardena, Cédric Howald, Sonali Jha, Rory Johnson, Philipp Kapranov, Brandon King, Colin Kingswood, Guoliang Li, Oscar J. Luo, Eddie Park, Jonathan B. Preall, Kimberly Presaud, Paolo Ribeca, Brian A. Risk, Daniel Robyr, Xiaoan Ruan, Michael Sammeth, Kuljeet Singh Sandhu, Lorain Schaeffer, Lei-Hoon Hoon See, Atif Shahab, Jorgen Skancke, Ana Maria Suzuki, Hazuki Takahashi, Hagen Tilgner, Diane Trout, Nathalie Walters, Hao Huaien Wang, John Wrobel, Yanbao Yu, Yoshihide Hayashizaki, Jennifer Harrow, Mark Gerstein, Timothy J. Hubbard, Alexandre Reymond, Stylianos E. Antonarakis, Gregory J. Hannon, Morgan C. Giddings, Yijun Ruan, Barbara Wold, Piero Carninci, Roderic Guigó, Thomas R. Gingeras, Kate R. Rosenbloom, Cricket A. Sloan, Katrina Learned, Venkat S. Malladi, Matthew C. Wong, Galt P. Barber, Melissa S. Cline, Timothy R. Dreszer, Steven G. Heitner, Donna Karolchik, W. James Kent, Vanessa M. Kirkup, Laurence R. Meyer, Jeffrey C. Long, Morgan Maddren, Brian J. Raney, Terrence S. Furey, Lingyun Song, Linda L. Grasfeder, Paul G. Giresi, Bum-Kyu Kyu Lee, Anna Battenhouse, Nathan C. Sheffield, Jeremy M. Simon, Kimberly A. Showers, Alexias Safi, Darin London, Akshay A. Bhinge, Christopher Shestak, Matthew R. Schaner, Seul Ki Kim, Zhuzhu Zhengdong Zhancheng Zhuzhu Z. Zhang, Piotr A. Mieczkowski, Joanna O. Mieczkowska, Zheng Liu, Ryan M.

McDaniell, Yunyun Ni, Naim U. Rashid, Min Jae Kim, Sheera Adar, Zhuzhu Zhengdong Zhancheng Zhuzhu Z. Zhang, Tianyuan Wang, Deborah Winter, Damian Keefe, Ewan Birney, Vishwanath R. Iyer, Jason D. Lieb, Gregory E. Crawford, Guoliang Li, Kuljeet Singh Sandhu, Meizhen Zheng, Ping Wang, Oscar J. Luo, Atif Shahab, Melissa J. Fullwood, Xiaoan Ruan, Yijun Ruan, Richard M. Myers, Florencia Pauli, Brian A. Williams, Jason Gertz, Georgi K. Marinov, Timothy E. Reddy, Jost Vielmetter, E. Christopher Partridge, Diane Trout, Katherine E. Varley, Clarke Gasper, The ENCODE Project Consortium, Overall coordination (data analysis Coordination), Data production leads (data Production), Lead analysts (data Analysis), Writing Group, NHGRI project management (scientific Management), Principal investigators (steering Committee), Boise State University Analysis), University of North Carolina at Chapel Hill Proteomics groups (data production, Broad Institute Group (data production, , Boise State University Analysis), University of North Carolina at Chapel Hill Proteomics groups (data production, Broad Institute Group (data production, , Cold Spring Harbor Center for Genomic Regulation Barcelona RIKEN Sanger Institute University of Lausanne Genome Institute of Singapore group (data production, University of Geneva analysis), Data coordination center at U C Santa Cruz (production data Coordination), Duke University University of Texas Austin University of North Carolina-Chapel Hill group (data production, E B I analysis), Genome Institute of Singapore group (data production Analysis), , HudsonAlpha Institute UC Irvine Stanford group (data

production, and Caltech analysis). An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414):57–74, 9 2012.

[45] Richard Durbin, Sean R. Eddy, Anders Krogh, and Graeme Mitchison. *Biological Sequence Analysis*. Cambridge University Press, 4 1998.

[46] David Duvenaud, Dougal Maclaurin, Jorge Aguilera-Iparraguirre, Rafael Gómez-Bombarelli, Timothy Hirzel, Alán Aspuru-Guzik, and Ryan P. Adams. Convolutional Networks on Graphs for Learning Molecular Fingerprints. *Advances in Neural Information Processing Systems*, 2015-Janua:2224–2232, 9 2015.

[47] Kerstin A Effenberger, Veronica K Urabe, and Melissa S. Jurica. Modulating splicing with small molecular inhibitors of the spliceosome. *Wiley Interdisciplinary Reviews: RNA*, 8(2):e1381, 3 2017.

[48] E. C. Fieller, H. O. Hartley, and E. S. Pearson. Tests for Rank Correlation Coefficients. I. *Biometrika*, 44(3/4):470, 12 1957.

[49] E. C. FIELLER and E. S. PEARSON. Tests for rank correlation coefficients. II. *Biometrika*, 48(1-2):29–40, 1961.

[50] R A Fisher. Frequency Distribution of the Values of the Correlation Coefficient in Samples from an Indefinitely Large Population. *Biometrika*, 10(4):507, 1915.

[51] Benjamin A Flusberg, Dale R Webster, Jessica H Lee, Kevin J Travers, Eric C Olivares, Tyson A Clark, Jonas Korlach, and Stephen W Turner. Direct detec-

tion of DNA methylation during single-molecule, real-time sequencing. *Nature Methods*, 7(6):461–465, 6 2010.

[52] Jean Baptiste Fourmann, Olexandr Dybkov, Dmitry E. Agafonov, Marcel J Tauchert, Henning Urlaub, Ralf Ficner, Patrizia Fabrizio, and Reinhard Lührmann. The target of the DEAH-box NTP triphosphatase Prp43 in saccharomyces cerevisiae spliceosomes is the U2 snRNP-intron interaction. *eLife*, 5(APRIL2016), 4 2016.

[53] M. Frommer, L. E. McDonald, D. S. Millar, C. M. Collis, F. Watt, G. W. Grigg, P. L. Molloy, and C. L. Paul. A genomic sequencing protocol that yields a positive display of 5- methylcytosine residues in individual DNA strands. *Proceedings of the National Academy of Sciences of the United States of America*, 89(5):1827–1831, 1992.

[54] Mattia Furlan, Anna Delgado-Tejedor, Logan Mulroney, Mattia Pelizzola, Eva Maria Novoa, and Tommaso Leonardi. Computational methods for RNA modification detection from nanopore direct RNA sequencing data. *RNA Biology*, 00(00):1–10, 2021.

[55] Hasindu Gamaarachchi, Chun Wai Lam, Gihan Jayatilaka, Hiruna Samarakoon, and Martin A Smith. GPU Accelerated Adaptive Banded Event Alignment for Rapid Comparative Nanopore Signal Analysis. *bioRxiv Bioinformatics*, pages 1–37, 2019.

[56] Yubang Gao, Xuqing Liu, Bizhi Wu, Huihui Wang, Feihu Xi, Markus V.

Kohnen, Anireddy S. N. Reddy, and Lianfeng Gu. Quantitative profiling of N6-methyladenosine at single-base resolution in stem-differentiating xylem of Populus trichocarpa using Nanopore direct RNA sequencing. *Genome Biology*, 22(1):22, 12 2021.

[57] Daniel R. Garalde, Elizabeth A. Snell, Daniel Jachimowicz, Botond Sipos, Joseph H. Lloyd, Mark Bruce, Nadia Pantic, Tigist Admassu, Phillip James, Anthony Warland, Michael Jordan, Jonah Ciccone, Sabrina Serra, Jemma Keenan, Samuel Martin, Luke McNeill, E. Jayne Wallace, Lakmal Jayasinghe, Chris Wright, Javier Blasco, Stephen Young, Denise Brocklebank, Sissel Juul, James Clarke, Andrew J. Heron, and Daniel J. Turner. Highly parallel direct RNA sequencing on an array of nanopores. *Nature Methods*, 15(3):201–206, 3 2018.

[58] Audrey P. Gasch, Paul T. Spellman, Camilla M. Kao, Orna Carmel-Harel, Michael B. Eisen, Gisela Storz, David Botstein, and Patrick O. Brown. Genomic Expression Programs in the Response of Yeast Cells to Environmental Changes. *Molecular Biology of the Cell*, 11(12):4241–4257, 12 2000.

[59] Daniela Georgieva, Qian Liu, Kai Wang, and Dieter Egli. Detection of base analogs incorporated during DNA replication by nanopore sequencing. *Nucleic Acids Research*, 48(15):e88–e88, 9 2020.

[60] Parveen Goyal, Petya V. Krasteva, Nani Van Gerven, Francesca Gubellini, Imke Van Den Broeck, Anastassia Troupiotis-Tsaïlaki, Wim Jonckheere, Gérard

Péhau-Arnaudet, Jerome S. Pinkner, Matthew R. Chapman, Scott J. Hultgren, Stefan Howorka, Rémi Fronzes, and Han Remaut. Structural and mechanistic insights into the bacterial amyloid secretion channel CsgG. *Nature*, 516(7530):250–253, 2014.

[61] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. Connectionist temporal classification. In *Proceedings of the 23rd international conference on Machine learning - ICML '06*, pages 369–376, New York, New York, USA, 2006. ACM Press.

[62] Pawel Grzechnik, Sylwia A. Szczepaniak, Somdutta Dhir, Anna Pastucha, Hannah Parslow, Zaneta Matuszek, Hannah E. Mischo, Joanna Kufel, and Nicholas J. Proudfoot. Nuclear fate of yeast snoRNA is determined by co-transcriptional Rnt1 cleavage. *Nature Communications*, 9(1):1783, 5 2018.

[63] Benjamin Guglielmi and Michel Werner. The yeast homolog of human PinX1 is involved in rRNA and small nucleolar RNA maturation, not in telomere elongation inhibition. *Journal of Biological Chemistry*, 277(38):35712–35719, 2002.

[64] J. S. Hardwick, F. G. Kuruvilla, J. K. Tong, A. F. Shamji, and S. L. Schreiber. Rapamycin-modulated transcription defines the subset of nutrient-sensitive signaling pathways directly controlled by the Tor proteins. *Proceedings of the National Academy of Sciences*, 96(26):14866–14870, 12 1999.

[65] Ralf Hauenschild, Lyudmil Tserovski, Katharina Schmid, Kathrin Thüring,

Marie Luise Winz, Sunny Sharma, Karl Dieter Entian, Ludivine Wacheul, Denis L.J. Lafontaine, James Anderson, Juan Alfonzo, Andreas Hildebrandt, Andres Jäschke, Yuri Motorin, and Mark Helm. The reverse transcription signature of N-1-methyladenosine in RNA-Seq is sequence dependent. *Nucleic Acids Research*, 43(20):9950–9964, 11 2015.

[66] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition, 12 2015.

[67] Annika U Heininger, Philipp Hackert, Alexandra Z. Andreou, Kum Loong Boon, Indira Memet, Mira Prior, Anne Clancy, Bernhard Schmidt, Henning Urlaub, Enrico Schleiff, Katherine E Sloan, Markus Deckers, Reinhard Lührmann, Jörg Enderlein, Dagmar Klostermeier, Peter Rehling, and Markus T Bohnsack. Protein cofactor competition regulates the action of a multifunctional RNA helicase in different pathways. *RNA Biology*, 13(3):320–330, 1 2016.

[68] Joseph Heitman, N. Rao Movva, and Michael N. Hall. Targets for Cell Cycle Arrest by the Immunosuppressant Rapamycin in Yeast. *Science*, 253(5022):905–909, 8 1991.

[69] Martin E Hess, Simon Hess, Kate D. Meyer, Linda A W Verhagen, Linda Koch, Hella S. Brönneke, Marcelo O. Dietrich, Sabine D. Jordan, Yogesh Saletore, Olivier Elemento, Bengt F. Belgardt, Thomas Franz, Tamas L. Horvath, Ulrich Rüther, Samie R. Jaffrey, Peter Kloppenburg, and Jens C. Brüning. The fat

mass and obesity associated gene (Fto) regulates activity of the dopaminergic midbrain circuitry. *Nature Neuroscience*, 16(8):1042–1048, 8 2013.

[70] Takuto Hideyama, Takenari Yamashita, Hitoshi Aizawa, Shoji Tsuji, Akiyoshi Kakita, Hitoshi Takahashi, and Shin Kwak. Profound downregulation of the RNA editing enzyme ADAR2 in ALS spinal motor neurons. *Neurobiology of Disease*, 45(3):1121–1128, 3 2012.

[71] Sepp Hochreiter and Jurgen Jurgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1–32, 1997.

[72] John D Hunter. Matplotlib: A 2D Graphics Environment. *Computing in Science & Engineering*, 9(3):90–95, 5 2007.

[73] Miten Jain, Sergey Koren, Karen H Miga, Josh Quick, Arthur C Rand, Thomas A Sasani, John R Tyson, Andrew D Beggs, Alexander T Dilthey, Ian T Fiddes, Sunir Malla, Hannah Marriott, Tom Nieto, Justin O'Grady, Hugh E Olsen, Brent S Pedersen, Arang Rhie, Hollian Richardson, Aaron R Quinlan, Terrance P Snutch, Louise Tee, Benedict Paten, Adam M. Phillippy, Jared T Simpson, Nicholas James Loman, and Matthew Loose. Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nature Biotechnology*, 36(4):338–345, 4 2018.

[74] Nicky Jonkhout, Julia Tran, Martin A. Smith, Nicole Schonrock, John S. Mattick, and Eva Maria Novoa. The RNA modification landscape in human disease. *Rna*, 23(12):1754–1769, 2017.

[75] Qida Ju and Jonathan R. Warner. Ribosome synthesis during the growth cycle ofSaccharomyces cerevisiae. *Yeast*, 10(2):151–157, 2 1994.

[76] John Karijolich, Athena Kantartzis, and Yi-Tao Yu. RNA Modifications: A Mechanism that Modulates Gene Expression. In *Methods in molecular biology (Clifton, N.J.)*, volume 629, pages 1–19. Springer US, 2010.

[77] J. J. Kasianowicz, E. Brandin, D. Branton, and D. W. Deamer. Characterization of individual polynucleotide molecules using a membrane channel. *Proceedings of the National Academy of Sciences*, 93(24):13770–13773, 11 1996.

[78] Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, pages 1–15, 12 2014.

[79] Thomas N. Kipf and Max Welling. Semi-Supervised Classification with Graph Convolutional Networks. *5th International Conference on Learning Representations, ICLR 2017 - Conference Track Proceedings*, pages 1–14, 9 2016.

[80] Jonas Korlach and Stephen W. Turner. Going beyond five bases in DNA sequencing. *Current Opinion in Structural Biology*, 22(3):251–261, 2012.

[81] Witold Kot, Nikoline S Olsen, Tue K Nielsen, Geoffrey Hutinet, Valérie de Crécy-Lagard, Liang Cui, Peter C Dedon, Alexander B Carstens, Sylvain Moineau, Manal A Swairjo, and Lars H Hansen. Detection of preQ0 deazaguanine modifications in bacteriophage CAjan DNA using Nanopore sequencing

reveals same hypermodification at two distinct DNA motifs. *Nucleic Acids Research*, 48(18):10383–10396, 10 2020.

[82] Yoshihiko Kotake, Koji Sagane, Takashi Owa, Yuko Mimori-Kiyosue, Hajime Shimizu, Mai Uesugi, Yasushi Ishihama, Masao Iwata, and Yoshiharu Mizui. Splicing factor SF3b as a target of the antitumor natural product pladienolide. *Nature Chemical Biology*, 3(9):570–575, 9 2007.

[83] Skirmantas Kriaucionis and Nathaniel Heintz. The Nuclear DNA Base 5-Hydroxymethylcytosine Is Present in Purkinje Neurons and the Brain. *Science*, 324(5929):929–930, 5 2009.

[84] Denis L.J. Lafontaine, Cécile Bousquet-Antonelli, Yves Henry, Michèle Caizergues-Ferrer, and David Tollervey. The box H + ACA snoRNAs carry Cbf5p, the putative rRNA pseudouridine synthase. *Genes and Development*, 12(4):527–537, 1998.

[85] DENIS L.J. LAFONTAINE and DAVID TOLLERVEY. Nop58p is a common component of the box C+D snoRNPs that is required for snoRNA stability. *RNA*, 5(3):S135583829998192X, 3 1999.

[86] Bruno Lapeyre and Suresh K. Purushothaman. Spb1p-Directed Formation of Gm2922 in the Ribosome Catalytic Center Occurs at a Late Processing Stage. *Molecular Cell*, 16(4):663–669, 11 2004.

[87] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard,

and L. D. Jackel. Backpropagation Applied to Handwritten Zip Code Recognition. *Neural Computation*, 1(4):541–551, 12 1989.

[88] Nina B Leeds, Eliza C Small, Shawna L Hiley, Timothy R Hughes, and Jonathan P Staley. The Splicing Factor Prp43p, a DEAH Box ATPase, Functions in Ribosome Biogenesis. *Molecular and Cellular Biology*, 26(2):513–522, 1 2006.

[89] Adrien Leger, Paulo P Amaral, Luca Pandolfini, Charlotte Capitanchik, Federica Capraro, Isaia Barbieri, Valentina Migliori, Nicholas M Luscombe, Anton J Enright, Kostantinos Konstantinos Tzelepis, Jernej Ule, Tomas Fitzgerald, Ewan Birney, Tommaso Leonardi, and Tony Kouzarides. RNA modifications detection by comparative Nanopore direct RNA sequencing. *bioRxiv*, page 843136, 2019.

[90] Adrien Leger and Tommaso Leonardi. pycoQC, interactive quality control for Oxford Nanopore Sequencing. *Journal of Open Source Software*, 4(34):1236, 2 2019.

[91] H. J. Levene, J. Korlach, S. W. Turner, M. Foquet, H. G. Craighead, and W. W. Webb. Zero-mode waveguides for single-molecule analysis at high concentrations. *Science*, 299(5607):682–686, 2003.

[92] Heng Li. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, 34(18):3094–3100, 9 2018.

[93] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, and Richard Durbin. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–2079, 8 2009.

[94] Sheng Li and Christopher E. Mason. The Pivotal Regulatory Landscape of RNA Modifications. *Annual Review of Genomics and Human Genetics*, 15(1):127–150, 8 2014.

[95] W.-Q. Liang and M J Fournier. Synthesis of functional eukaryotic ribosomal RNAs in trans: Development of a novel in vivo rDNA system for dissecting ribosome biogenesis. *Proceedings of the National Academy of Sciences*, 94(7):2864–2868, 4 1997.

[96] Huanle Liu, Oguzhan Begik, Morghan C. Lucas, Jose Miguel Ramirez, Christopher E. Mason, David Wiener, Schraga Schwartz, John S. Mattick, Martin A. Smith, and Eva Maria Novoa. Accurate detection of m6A RNA modifications in native RNA sequences. *Nature Communications*, 10(1):1–9, 9 2019.

[97] Jianzhao Liu, Yuanxiang Zhu, Guan-Zheng Luo, Xinxia Wang, Yanan Yue, Xiaona Wang, Xin Zong, Kai Chen, Hang Yin, Ye Fu, Dali Han, Yizhen Wang, Dahua Chen, and Chuan He. Abundant DNA 6mA methylation during early embryogenesis of zebrafish and pig. *Nature Communications*, 7(1):13052, 12 2016.

[98] Kuanqing Liu, Daniel A Santos, Jeffrey A Hussmann, Yun Wang, Benjamin M

Sutter, Jonathan S Weissman, and Benjamin P Tu. Regulation of translation by methylation multiplicity of 18S rRNA. *Cell Reports*, 34(10):108825, 3 2021.

[99] Qian Liu, Li Fang, Guoliang Yu, Depeng Wang, Chuan-le Le Xiao, and Kai Wang. Detection of DNA base modifications by a deep recurrent neural network on Oxford Nanopore sequencing data. *Nature Communications*, 10(1), 2019.

[100] Qian Liu, Daniela C. Georgieva, Dieter Egli, and Kai Wang. NanoMod: a computational tool to detect DNA modifications using Nanopore long-read sequencing data. *bioRxiv*, page 277178, 2018.

[101] Nicholas J. Loman, Joshua Quick, and Jared T. Simpson. A complete bacterial genome assembled de novo using only nanopore sequencing data. *Nature Methods*, 12(8):733–735, 2015.

[102] Daniel A. Lorenz, Shashank Sathe, Jaclyn M. Einstein, and Gene W. Yeo. Direct RNA sequencing enables m6A detection in endogenous transcript isoforms at base-specific resolution. *Rna*, 26(1):19–28, 2020.

[103] Frank Lyko. The DNA methyltransferase family: A versatile toolkit for epigenetic regulation, 2018.

[104] A. MacDonald, C.J. Scarrott, D. Lee, B. Darlow, M. Reale, and G. Russell. A flexible extreme value mixture model. *Computational Statistics & Data Analysis*, 55(6):2137–2157, 6 2011.

[105] H D Madhani and C Guthrie. Genetic interactions between the yeast RNA

helicase homolog Prp16 and spliceosomal snRNAs identify candidate ligands for the Prp16 RNA-dependent ATPase, 7 1994.

[106] Kerstin C. Maier, Saskia Gressel, Patrick Cramer, and Björn Schwalb. Native molecule sequencing by nano-ID reveals synthesis and stability of RNA isoforms. *Genome Research*, 30(9):1332–1344, 9 2020.

[107] Virginie Marchand, Florence Blanloeil-Oillo, Mark Helm, and Yuri Motorin. Illumina-based RiboMethSeq approach for mapping of 2'-O-Me residues in RNA. *Nucleic Acids Research*, 44(16):e135–e135, 9 2016.

[108] Marcus H Stoiber, Joshua Quick, Rob Egan, Ji Eun Lee, Susan E Celniker, Robert Neely, Nicholas Loman, Len Pennacchio, and James B Brown. De novo Identification of DNA Modifications Enabled by Genome-Guided Nanopore Signal Processing. *biorXiv*, 2016.

[109] Arnold Martin, Susanne Schneider, and Beate Schwer. Prp43 Is an Essential RNA-dependent ATPase Required for Release of Lariat-Intron from the Spliceosome. *Journal of Biological Chemistry*, 277(20):17743–17750, 5 2002.

[110] Leland McInnes, John Healy, and James Melville. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv*, 2 2018.

[111] Alexa B. R. McIntyre, Noah Alexander, Kirill Grigorev, Daniela Bezdan, Heike Sichtig, Charles Y Chiu, and Christopher E Mason. Single-molecule sequencing

detection of N6-methyladenine in microbial reference materials. *Nature Communications*, 10(1):579, 12 2019.

[112] Carolin A. Müller, Michael A. Boemo, Paolo Spingardi, Benedikt M. Kessler, Skirmantas Kriaucionis, Jared T. Simpson, and Conrad A. Nieduszynski. Capturing the dynamics of genome replication on individual ultra-long nanopore sequence reads. *Nature Methods*, 16(5):429–436, 2019.

[113] Hitoshi Nakatogawa and Koreaki Ito. The Ribosomal Exit Tunnel Functions as a Discriminating Gate. *Cell*, 108(5):629–636, 3 2002.

[114] S. Kundhavai Natchiar, Alexander G. Myasnikov, Hanna Kratzat, Isabelle Hazemann, and Bruno P. Klaholz. Visualization of chemical modifications in the human 80S ribosome structure. *Nature*, 551(7681):472–477, 2017.

[115] Peng Ni, Neng Huang, Zhi Zhang, De-Peng Peng Wang, Fan Liang, Yu Miao, Chuan-Le Le Xiao, Feng Luo, and Jianxin Wang. DeepSignal: detecting DNA methylation state from Nanopore sequencing reads using deep-learning. *Bioinformatics*, 35(22):4586–4595, 11 2019.

[116] Intawat Nookaew, Piroon Jenjaroenpun, Hua Du, Pengcheng Wang, Jun Wu, Thidathip Wongsurawat, Sun Hee Moon, En Huang, Yinsheng Wang, and Gunnar Boysen. Detection and Discrimination of DNA Adducts Differing in Size, Regiochemistry, and Functional Group by Nanopore Sequencing. *Chemical Research in Toxicology*, 33(12):2944–2952, 12 2020.

[117] T G Obrig, W J Culp, W L McKeehan, and B Hardesty. The mechanism by which cycloheximide and related glutarimide antibiotics inhibit peptide synthesis on reticulocyte ribosomes. *The Journal of biological chemistry*, 246(1):174–81, 1 1971.

[118] Siew Loon Ooi, Dmitry A. Samarsky, Maurille J. Fournier, and Jef D. Boeke. Intronic snoRNA biosynthesis in saccharomyces cerevisiae depends on the lariat-debranching enzyme: Intron length effects and activity of a precursor snoRNA. *Rna*, 4(9):1096–1110, 9 1998.

[119] Shatakshi Pandit, Bert Lynn, and Brian C Rymond. Inhibition of a spliceosome turnover pathway suppresses splicing defects. *Proceedings of the National Academy of Sciences of the United States of America*, 103(37):13700–13705, 9 2006.

[120] Matthew T Parker, Katarzyna Knop, Anna V Sherwood, Nicholas J Schurch, Katarzyna Mackinnon, Peter D Gould, Anthony JW Hall, Geoffrey J Barton, and Gordon G Simpson. Nanopore direct RNA sequencing maps the complexity of Arabidopsis mRNA processing and m6A modification. *eLife*, 9, 1 2020.

[121] Steven Parker, Marcin G Fraczek, Jian Wu, Sara Shamsah, Alkisti Manousaki, Kobchai Dungrattanalert, Rogerio Alves de Almeida, Edith Invernizzi, Tim Burgis, Walid Omara, Sam Griffiths-Jones, Daniela Delneri, and Raymond T. O'Keefe. Large-scale profiling of noncoding RNA function in yeast. *PLoS Genetics*, 14(3):e1007253, 3 2018.

[122] Karl Pearson. X. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 50(302):157–175, 7 1900.

[123] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12(85):2825–2830, 2011.

[124] Brigitte Pertschy, Claudia Schneider, Marén Gnädig, Thorsten Schäfer, David Tollervey, and Ed Hurt. RNA Helicase Prp43 and Its Co-factor Pfa1 Promote 20 to 18 S rRNA Processing Catalyzed by the Endonuclease Nob1. *Journal of Biological Chemistry*, 284(50):35079–35091, 12 2009.

[125] Elisabeth Petfalski, Thomas Dandekar, Yves Henry, and David Tollervey. Processing of the Precursors to Small Nucleolar RNAs and rRNAs Requires Common Components. *Molecular and Cellular Biology*, 18(3):1181–1189, 3 1998.

[126] Anton S. Petrov, Chad R. Bernier, Burak Gulen, Chris C. Waterbury, Eli Hershkovits, Chiaolong Hsiao, Stephen C. Harvey, Nicholas V. Hud, George E. Fox,

Roger M. Wartell, and Loren Dean Williams. Secondary Structures of rRNAs from All Three Domains of Life. *PLoS ONE*, 9(2):e88222, 2 2014.

[127] Dorota Piekna-Przybylska, Wayne A Decatur, and Maurille J Fournier. New bioinformatic tools for analysis of nucleotide modifications in eukaryotic rRNA. *RNA*, 13(3):305–312, 3 2007.

[128] Yury S Polikanov, Sergey V Melnikov, Dieter Söll, and Thomas A Steitz. Structural insights into the role of rRNA modifications in protein synthesis and ribosome assembly. *Nature Structural & Molecular Biology*, 22(4):342–344, 4 2015.

[129] William Poole, David L. Gibbs, Ilya Shmulevich, Brady Bernard, and Theo A. Knijnenburg. Combining dependent ¡i¿P-¡/i¿ values with an empirical adaptation of Brown's method. *Bioinformatics*, 32(17):i430–i436, 9 2016.

[130] Anna Portela and Manel Esteller. Epigenetic modifications and human disease. *Nature Biotechnology*, 28(10):1057–1068, 2010.

[131] Ted Powers and Peter Walter. Regulation of Ribosome Biogenesis by the Rapamycin-sensitive TOR-signaling Pathway in ¡i¿Saccharomyces cerevisiae¡/i¿. *Molecular Biology of the Cell*, 10(4):987–1000, 4 1999.

[132] Andrea Dal Pozzolo, Olivier Caelen, Reid A. Johnson, and Gianluca Bontempi. Calibrating Probability with Undersampling for Unbalanced Classification. In *2015 IEEE Symposium Series on Computational Intelligence*, pages 159–166. IEEE, 12 2015.

[133] Eun-Ang Raiber, Robyn Hardisty, Pieter van Delft, and Shankar Balasubramanian. Mapping and elucidating the function of modified bases in DNA. *Nature Reviews Chemistry*, 1(9), 2017.

[134] Gokul Ramaswami, Rui Zhang, Robert Piskol, Liam P Keegan, Patricia Deng, Mary A O'Connell, and Jin Billy Li. Identifying RNA editing sites using RNA sequencing data alone. *Nature Methods*, 10(2):128–132, 2 2013.

[135] Arthur C Rand, Miten Jain, Jordan M Eizenga, Audrey Musselman-Brown, Hugh E Olsen, Mark Akeson, and Benedict Paten. Mapping DNA methylation with high-throughput nanopore sequencing. *Nature Methods*, 14(4):411–413, 4 2017.

[136] Jean Paul Renaud, Ashwin Chari, Claudio Ciferri, Wen Ti Liu, Hervé William Rémigy, Holger Stark, and Christian Wiesmann. Cryo-EM in drug discovery: Achievements, limitations and prospects. *Nature Reviews Drug Discovery*, 17(7):471–492, 2018.

[137] M. Ronaghi. Pyrosequencing Sheds Light on DNA Sequencing. *Genome Research*, 11(1):3–11, 1 2001.

[138] Jonathan M. Rothberg, Wolfgang Hinz, Todd M. Rearick, Jonathan Schultz, William Mileski, Mel Davey, John H. Leamon, Kim Johnson, Mark J. Milgrew, Matthew Edwards, Jeremy Hoon, Jan F. Simons, David Marran, Jason W. Myers, John F. Davidson, Annika Branting, John R. Nobile, Bernard P. Puc, David Light, Travis A. Clark, Martin Huber, Jeffrey T. Branciforte, Isaac B. Stoner,

Simon E. Cawley, Michael Lyons, Yutao Fu, Nils Homer, Marina Sedova, Xin Miao, Brian Reed, Jeffrey Sabina, Erika Feierstein, Michelle Schorn, Mohammad Alanjary, Eileen Dimalanta, Devin Dressman, Rachel Kasinskas, Tanya Sokolsky, Jacqueline A. Fidanza, Eugeni Namsaraev, Kevin J. McKernan, Alan Williams, G. Thomas Roth, and James Bustillo. An integrated semiconductor device enabling non-optical genome sequencing. *Nature*, 475(7356):348–352, 7 2011.

[139] Paul Ryvkin, Y. Y. Leung, I. M. Silverman, Micah Childress, Otto Valladares, Isabelle Dragomir, Brian D Gregory, and L.-S. Wang. HAMR: high-throughput annotation of modified ribonucleotides. *RNA*, 19(12):1684–1692, 12 2013.

[140] Yogesh Saletore, Kate Meyer, Jonas Korlach, Igor D Vilfan, Samie Jaffrey, and Christopher E Mason. The birth of the Epitranscriptome: deciphering the function of RNA modifications. *Genome Biology*, 13(10):175, 2012.

[141] F. Sanger, S. Nicklen, and A. R. Coulson. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences*, 74(12):5463–5467, 12 1977.

[142] Peter Schattner, Wayne A Decatur, Carrie A Davis, Manuel Ares Jr, Maurille J Fournier, and Todd M Lowe. Genome-wide searching for pseudouridylation guide snoRNAs: analysis of the Saccharomyces cerevisiae genome. *Nucleic Acids Research*, 32(14):4281–4296, 8 2004.

[143] Markus Schosserer, Nadege Minois, Tina B. Angerer, Manuela Amring, Hanna

Dellago, Eva Harreither, Alfonso Calle-Perez, Andreas Pircher, Matthias Peter Gerstl, Sigrid Pfeifenberger, Clemens Brandl, Markus Sonntagbauer, Albert Kriegner, Angela Linder, Andreas Weinhäusel, Thomas Mohr, Matthias Steiger, Diethard Mattanovich, Mark Rinnerthaler, Thomas Karl, Sunny Sharma, Karl-Dieter Entian, Martin Kos, Michael Breitenbach, Iain B.H. Wilson, Norbert Polacek, Regina Grillari-Voglauer, Lore Breitenbach-Koller, and Johannes Grillari. Methylation of ribosomal RNA by NSUN5 is a conserved mechanism modulating organismal lifespan. *Nature Communications*, 6(1):6158, 5 2015.

[144] M. Schuster and K. K Paliwal. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681, 1997.

[145] Schraga Schwartz, Sudeep D. Agarwala, Maxwell R. Mumbach, Marko Jovanovic, Philipp Mertins, Alexander Shishkin, Yuval Tabach, Tarjei S. Mikkelsen, Rahul Satija, Gary Ruvkun, Steven A. Carr, Eric S. Lander, Gerald R. Fink, and Aviv Regev. High-Resolution Mapping Reveals a Conserved, Widespread, Dynamic mRNA Methylation Program in Yeast Meiosis. *Cell*, 155(6):1409–1421, 12 2013.

[146] Schraga Schwartz, Douglas A. Bernstein, Maxwell R. Mumbach, Marko Jovanovic, Rebecca H. Herbst, Brian X. León-Ricardo, Jesse M. Engreitz, Mitchell Guttman, Rahul Satija, Eric S. Lander, Gerald Fink, and Aviv Regev. Transcriptome-wide mapping reveals widespread dynamic-regulated pseudouridylation of ncRNA and mRNA. *Cell*, 159(1):148–162, 9 2014.

[147] Andrew Shafik, Ulrike Schumann, Maurits Evers, Tennille Sibbritt, and Thomas Preiss. The emerging epitranscriptomics of long noncoding RNAs. *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms*, 1859(1):59–70, 1 2016.

[148] Sunny Sharma, Johannes David Hartmann, Peter Watzinger, Arvid Klepper, Christian Peifer, Peter Kötter, Denis L.J. Lafontaine, and Karl Dieter Entian. A single N1-methyladenosine on the large ribosomal subunit rRNA impacts locally its structure and the translation of key metabolic enzymes. *Scientific Reports*, 8(1):11904, 8 2018.

[149] Sunny Sharma, Jun Yang, Rob van Nues, Peter Watzinger, Peter Kötter, Denis L.J. Lafontaine, Sander Granneman, and Karl Dieter Entian. Specialized box C/D snoRNPs act as antisense guides to target RNA base acetylation. *PLoS Genetics*, 13(5):1–23, 2017.

[150] Jay Shendure, Shankar Balasubramanian, George M. Church, Walter Gilbert, Jane Rogers, Jeffery A. Schloss, and Robert H. Waterston. DNA sequencing at 40: past, present and future. *Nature*, 550(7676):345–353, 10 2017.

[151] Carrie L. Simms, Liewei L. Yan, and Hani S. Zaher. Ribosome Collision Is Critical for Quality Control during No-Go Decay. *Molecular Cell*, 68(2):361–373, 10 2017.

[152] Jared T Simpson, Rachael E Workman, P C Zuzarte, Matei David, L J Dursi,

and Winston Timp. Detecting DNA cytosine methylation using nanopore sequencing. *Nature Methods*, 14(4):407–410, 4 2017.

[153] Katherine E. Sloan, Ahmed S. Warda, Sunny Sharma, Karl-Dieter Entian, Denis L. J. Lafontaine, and Markus T. Bohnsack. Tuning the ribosome: The influence of rRNA modification on eukaryotic ribosome biogenesis and function. *RNA Biology*, 14(9):1138–1152, 9 2017.

[154] Andrew M Smith, Miten Jain, Logan Mulroney, Daniel R Garalde, and Mark Akeson. Reading canonical and modified nucleobases in 16S ribosomal RNA using nanopore native RNA sequencing. *PLOS ONE*, 14(5):e0216709, 5 2019.

[155] Ankur Jai Sood, Coby Viner, and Michael M. Hoffman. DNAmod: the DNA modification database. *Journal of Cheminformatics*, 11(1):30, 12 2019.

[156] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A Simple Way to Prevent Neural Networks from Overfitting Nitish. *Journal of Machine Learning Research*, 15(1):1929–1958, 6 2014.

[157] William Stephenson, Roham Razaghi, Steven Busan, Kevin M. Weeks, Winston Timp, and Peter Smibert. Direct detection of RNA modifications and structure using single molecule nanopore sequencing. *bioRxiv*, pages 1–31, 2020.

[158] Marcus Stoiber and James Brown. BasecRAWller: Streaming Nanopore Basecalling Directly from Raw Signal. *bioRxiv*, 2017.

[159] Student. The Probable Error of a Mean. *Biometrika*, 6(1):1, 3 1908.

[160] Hajime Suzuki and Masahiro Kasahara. Acceleration of Nucleotide Semi-Global Alignment with Adaptive Banded Dynamic Programming. *bioRxiv*, 25(18):1–1, 2017.

[161] Naoko Tanaka, Anna Aronova, and Beate Schwer. Ntr1 activates the Prp43 helicase to trigger release of lariat-intron from the spliceosome. *Genes & Development*, 21(18):2312–2325, 9 2007.

[162] Masato Taoka, Yuko Nobe, Yuka Yamaki, Ko Sato, Hideaki Ishikawa, Keiichi Izumikawa, Yoshio Yamauchi, Kouji Hirota, Hiroshi Nakayama, Nobuhiro Takahashi, and Toshiaki Isobe. Landscape of the complete RNA chemical modifications in the human 80S ribosome. *Nucleic Acids Research*, 46(18):9289–9298, 10 2018.

[163] Masato Taoka, Yuko Nobe, Yuka Yamaki, Yoshio Yamauchi, Hideaki Ishikawa, Nobuhiro Takahashi, Hiroshi Nakayama, and Toshiaki Isobe. The complete chemical structure of Saccharomyces cerevisiae rRNA: partial pseudouridylation of U2345 in 25S rRNA by snoRNA snR9. *Nucleic Acids Research*, 44(18):8951–8961, 10 2016.

[164] Masato Taoka, Yoshio Yamauchi, Yuko Nobe, Shunpei Masaki, Hiroshi Nakayama, Hideaki Ishikawa, Nobuhiro Takahashi, and Toshiaki Isobe. An analytical platform for mass spectrometry-based identification and chemical

analysis of RNA in ribonucleoprotein complexes. *Nucleic Acids Research*, 37(21):e140–e140, 11 2009.

[165] Yee Whye Teh, Michael I. Jordan, Matthew J. Beal, and David M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006.

[166] Haotian Teng, Minh Duc Cao, Michael B. Hall, Tania Duarte, Sheng Wang, and Lachlan J.M. M Coin. Chiron: translating nanopore raw signal directly into nucleotide sequence using deep learning. *GigaScience*, 7(5):1–10, 5 2018.

[167] David Tollervey and Tamás Kiss. Function and synthesis of small nucleolar RNAs. *Current Opinion in Cell Biology*, 9(3):337–342, 6 1997.

[168] R.-T. Tsai. Spliceosome disassembly catalyzed by Prp43 and its associated components Ntr1 and Ntr2. *Genes & Development*, 19(24):2991–3003, 12 2005.

[169] Chi Kang Tseng, Hsueh Lien Liu, and Soo Chen Cheng. DEAH-box ATPase Prp16 has dual roles in remodeling of the spliceosome in catalytic steps. *RNA*, 17(1):145–154, 1 2011.

[170] Erwin L. van Dijk, Hélène Auger, Yan Jaszczyszyn, and Claude Thermes. Ten years of next-generation sequencing technology. *Trends in Genetics*, 30(9):418–426, 9 2014.

[171] Robert Willem van Nues, Sander Granneman, Grzegorz Kudla, Katherine Elizabeth Sloan, Matthew Chicken, David Tollervey, and Nicholas James Watkins.

Box C/D snoRNP catalysed methylation is aided by additional pre-rRNA base-pairing. *The EMBO Journal*, 30(12):2420–2430, 6 2011.

[172] Adrian Viehweger, Sebastian Krautwurst, Kevin Lamkiewicz, Ramakanth Madhugiri, John Ziebuhr, Martin Hölzer, and Manja Marz. Direct RNA nanopore sequencing of full-length coronavirus genomes provides novel insights into structural variants and enables modification analysis. *Genome Research*, 29(9):1545–1554, 2019.

[173] Sara Vincenti, Valentina De Chiara, Irene Bozzoni, and Carlo Presutti. The position of yeast snoRNA-coding regions within host introns is essential for their biosynthesis and for efficient splicing of the host pre-mRNA. *RNA*, 13(1):138–150, 11 2006.

[174] Xiao Wang, Boxuan Simen Zhao, Ian A. Roundtree, Zhike Lu, Dali Han, Honghui Ma, Xiaocheng Weng, Kai Chen, Hailing Shi, and Chuan He. N6-methyladenosine modulates messenger RNA translation efficiency. *Cell*, 161(6):1388–1399, 2015.

[175] Joe H Ward. Hierarchical Grouping to Optimize an Objective Function. *Journal of the American Statistical Association*, 58(301):236–244, 3 1963.

[176] Jonathan R Warner. The economics of ribosome biosynthesis in yeast. *Trends in Biochemical Sciences*, 24(11):437–440, 11 1999.

[177] Michael Waskom. seaborn: statistical data visualization. *Journal of Open Source Software*, 6(60):3021, 4 2021.

[178] Nicholas J Watkins, Alexander Gottschalk, Gitte Neubauer, Berthold Kastner, Patrizia Fabrizio, Matthias Mann, and Reinhard Lührmann. Cbf5p, a potential pseudouridine synthase, and Nhp2p, a putative RNA- binding protein, are present together with Gar1p in all H BOX/ACA-motif snoRNPs and constitute a common bipartite structure. *RNA*, 4(12):1549–1568, 12 1998.

[179] P L Weaver, C Sun, and T H Chang. Dbp3p, a putative RNA helicase in Saccharomyces cerevisiae, is required for efficient pre-rRNA processing predominantly at site A3. *Molecular and cellular biology*, 17(3):1354–65, 3 1997.

[180] Ryan R. Wick, Louise M. Judd, Claire L. Gorrie, and Kathryn E. Holt. Completing bacterial genome assemblies with multiplex MinION sequencing. *Microbial Genomics*, 3(10):1–7, 10 2017.

[181] Ryan R. Wick, Louise M. Judd, and Kathryn E. Holt. Performance of neural network basecalling tools for Oxford Nanopore sequencing. *Genome Biology*, 20(1):129, 12 2019.

[182] Ronald J Williams and David Zipser. A Learning Algorithm for Continually Running Fully Recurrent Neural Networks. *Neural Computation*, pages 270–280, 1989.

[183] Daniel M. Wilson, Yu Li, Amber LaPeruta, Michael Gamalinda, Ning Gao, and

John L. Woolford. Structural insights into assembly of the ribosomal nascent polypeptide exit tunnel. *Nature Communications*, 11(1):5111, 12 2020.

[184] Rachael E. Workman, Alison D. Tang, Paul S. Tang, Miten Jain, John R. Tyson, Roham Razaghi, Philip C. Zuzarte, Timothy Gilpatrick, Alexander Payne, Joshua Quick, Norah Sadowski, Nadine Holmes, Jaqueline Goes de Jesus, Karen L. Jones, Cameron M. Soulette, Terrance P. Snutch, Nicholas Loman, Benedict Paten, Matthew Loose, Jared T. Simpson, Hugh E. Olsen, Angela N. Brooks, Mark Akeson, and Winston Timp. Nanopore native RNA sequencing of a human poly(A) transcriptome. *Nature Methods*, 16(12):1297–1305, 12 2019.

[185] Qingfa Wu, Erik Niebuhr, Huanming Yang, and Lars Hansen. Determination of the 'critical region' for cat-like cry of Cri-du-chat syndrome and analysis of candidate genes by quantitative PCR. *European Journal of Human Genetics*, 13(4):475–485, 4 2005.

[186] Qingfa Wu, Erik Niebuhr, Huanming Yang, and Lars Hansen. Determination of the 'critical region' for cat-like cry of Cri-du-chat syndrome and analysis of candidate genes by quantitative PCR. *European Journal of Human Genetics*, 13(4):475–485, 4 2005.

[187] Chuan Le Xiao, Song Zhu, Minghui He, De Chen, Qian Zhang, Ying Chen, Guoliang Yu, Jinbao Liu, Shang Qian Xie, Feng Luo, Zhe Liang, De Peng Wang, Xiao Chen Bo, Xiao Feng Gu, Kai Wang, and Guang Rong Yan. N

6 -Methyladenine DNA Modification in the Human Genome. *Molecular Cell*, 71(2):306–318, 2018.

[188] Jun Yang, Sunny Sharma, Peter Watzinger, Johannes David Hartmann, Peter Kötter, and Karl-Dieter Entian. Mapping of Complete Set of Ribose and Base Modifications of Yeast rRNA by RP-HPLC and Mung Bean Nuclease Assay. *PLOS ONE*, 11(12):e0168873, 12 2016.

[189] Yuan Yao, Lorenzo Rosasco, and Andrea Caponnetto. On Early Stopping in Gradient Descent Learning. *Constructive Approximation*, 26(2):289–315, 8 2007.

[190] Jerrold H. Zar. Spearman Rank Correlation: Overview. In *Wiley StatsRef: Statistics Reference Online*. Wiley, 9 2014.

[191] Ying Zhang, Tina Wölfle, and Sabine Rospert. Interaction of Nascent Chains with the Ribosomal Tunnel Proteins Rpl4, Rpl17, and Rpl39 of Saccharomyces cerevisiae. *Journal of Biological Chemistry*, 288(47):33697–33707, 11 2013.