USING NANOPORE SEQUENCING TO INTERROGATE THE GENOME AND EPIGENOME

by

Roham Razaghi

A dissertation submitted to Johns Hopkins University

in conformity with the requirements for the degree of

Doctor of Philosophy

Baltimore, Maryland

October, 2022

Abstract

This work involves using native RNA nanopore sequencing to directly characterize the transcriptome of a human cell line, GM12878. We demonstrated several new methods, and findings, including newly discovered isoforms, allele-specific isoforms, measurement of polyadenylation length, and even measurement of RNA modifications. We also describe an application of nanopore RNA sequencing and chemical labeling to measure the secondary structure of RNA. Lastly, we demonstrate an analysis framework for looking at a new file format for single-molecule/long-read modification data.

Thesis Committee

Winston Timp (Primary Advisor, Primary Reader)
Associate Professor
Department of Biomedical Engineering
Johns Hopkins School of Engineering

Michael C. Schatz (Chair of Thesis Committee, Secondary Reader)
Professor
Department of Computer Science
Johns Hopkins School of Engineering

Kasper Daniel Hansen
Associate Professor
Department of Biostatistics
Johns Hopkins Bloomberg School of Public Health

Table of Contents

List of Figures

Chapter 1

Introduction

1.1 History of Nanopore Sequencing

25 years of development after the initial patent in 1995, nanopore sequencing has emerged as a viable commercial platform for nucleic acid sequencing and contributed to massive strides in genomics and transcriptomics. Nanopore sequencing operates in a similar fashion to a Coulter counter - allowing for characterization of a polymer (DNA, RNA, or cDNA) based on its interaction with ionic current flowing through the pore. In contrast to most other sequencing methods which operate through sequencing by synthesis, nanopore is characterizing the molecule directly which enables longer reads and evaluation of nucleotide modifications on DNA or RNA.

It can be argued that the genomics era began with the resolution of the three-dimensional structure of DNA in the 1950s (Franklin and Gosling, 1953; Watson and Crick, 1953). This was followed by the first sequencing of nucleic acids using chromatography based methods in the 1960s (Holley et al.,

1965; Wu, 1972; Sanger, Brownlee, and Barrell, 1965). Chromatographic methods like these were then optimized (Sanger, Nicklen, and Coulson, 1977) throughout the 70s and 80s, resulting in automated capillary electrophoresis sequencing available by 1990 (Luckey et al., 1990). This technology development, a coupling of molecular biology advances to advances in engineering and computational analysis, led in part to the completion of the first human genome in 2001, an achievement which cost >$3 billion and took ∼13 years (Venter et al., 2001; Lander et al., 2001) But technology development has continued, reducing the costs and improving the speed of sequencing (Schloss et al., 2020). A goal of a $1000 genome was set and strived for via different methods of technology development.

Technological advancements led to the introduction of the massively parallel high throughput Next Generation Sequencing (NGS). These technologies allow for rapid and cheaper sequencing than the previously used Sanger. Illumina sequencing in particular, can generate billions of sequencing reads enabling the goal of whole human genome sequencing for under $1000 (Buermans and Dunnen, 2014; Davies, 2015; Mardis, 2006). The technology involves local clonal amplification of DNA template molecules and identification of nucleotides through detection of fluorescent signals. While generating sequencing reads greater than 99.9% accurate (Q30), Illumina sequencing is subject to cycle dephasing brought on by increasing read length, inverted repeats and GC rich sequences which decreases signal to noise and precludes read accuracy (Nakamura et al., 2011). Due to this technological hurdle, GC rich regions are under-represented in the sequencing reads and read length

is limited. Alternatively, long read sequencing, sequencing, offers many advantages over NGS. While NGS can generate reads up to 600bp in length, single molecule sequencing can routinely sequence reads longer than 10kb. Long reads improve de novo assembly, mappability, transcript isoform identification, phasing of alleles, and detection of structural variants. Additionally, third generation sequencing can be performed on native nucleic acid, both DNA and RNA, therefore reducing PCR amplification bias and preserving epigenetic information in the form of base modifications.

There are currently two methods of single molecule sequencing, Oxford Nanopore Technologies (ONT) and Pacific Biosciences (PacBio). Both rely on distinct biophysical principles to evaluate the order of nucleotides on single molecules. PacBio sequencing captures sequence information during the replication process by detecting fluorophore incorporation at a tethered polymerase. This process occurs within a zero mode wave-guide (ZMW) which limits fluorophore detection primarily to the volume around the polymerase (Levene et al., 2003). Read length in PacBio sequencing is dependent on the processivity of the polymerase - the longer the polymerase stays on the molecule, the longer the sequencing read can be. Because PacBio sequencing uses a circularized library, there is an opportunity to sequence the same molecule multiple times (HiFi data), improving the single read accuracy to greater than 99.9% (Q30) at the cost of sequencing yield or molecular read length (25 kb) (Nurk et al., 2020). In contrast, CLR sequencing can achieve exceeding 50kb, with a decrease in read accuracy to below 85% (Nurk et al., 2020; Wei and Zhang, 2018).

In contrast, nanopore sequencing relies on fluctuations in electrical current to characterize a biopolymer as it passes through a biological nanopore. Nanopore sequencing is unique from other sequencing methodologies (i.e. Sanger, Illumina, PacBio) because it characterizes the molecule directly, rather than as a result of DNA synthesis. The methodology was first suggested that polymers could be characterized by measuring the altered current as polymers pass through protein pores in the late 1990s (Church et al., 1998; Deamer, Akeson, and Branton, 2016). This idea became a reality when Kasianowicz et al. characterized DNA and RNA in α-hemolysin (α-HL) nanopores using techniques developed in electrophysiology for ion-channel measurements (Kasianowicz et al., 1996).

As experimentation continued, initial enthusiasm was curbed because of significant roadblocks in application to actual sequencing, mainly that the speed of nucleic acid translocation was too fast (1-10μs) to resolve individual bases. Improving signal to noise required each base to be in the pore for > 100μs. This pointed to the necessity of racheting the strand through the pore so that each step allowed sufficient time to identify the next base in the strand's sequence (Meller, Nivon, and Branton, 2001; Deamer, Akeson, and Branton, 2016; Hornblower et al., 2007). It was not until 2012 that the two essential components of a functioning nanopore sequencer (ie. translocation control at single-nucleotide resolution and discrimination among bases) were in place by utilizing a mutant MspA nanopore and phi29 DNA polymerase (Manrao et al., 2012).

With all the pieces in place, commercial development began with the announcement at AGBT in 2012 and the first release of the R6 MinION nanopore sequencer by Oxford Nanopore Technologies in 2014. The MinION is a compact and portable device with 2,048 individually addressable protein nanopores of which 512 can sequence simultaneously. (Cherf et al., 2012; Yeh et al., 2012). Initial testing of the MinION showed it could yield 50-150Mb with reads up to 15kb long (Mikheyev and Tin 2014, W. Timp et al. 2014). With these early iterations, error rate in the MinIONs was a significant concern with per read accuracy at ∼67% (Mikheyev and Tin 2014; W. Timp et al. 2014). Rapid and significant improvements have occurred throughout the past few years leading to additional iterations of the protein pore, motor protein and membrane. The dominant pore version is currently the R9.4, which is derived from the *Escherichia coli* CsgG pore and also has a newer version of the motor protein (E8) to increase translocation speed ((Loose, 2017; Lu, Giordano, and Ning, 2016; Goyal et al., 2014). Current yield varies between the different flowcell options from ∼1Gb for the smallest (flongle) flowcell with 128 sequencing channels to ∼10-20Gb for the minION with 512 sequencing channels to ∼100Gb for the PromethION with ∼3000 sequencing channels.

To sequence DNA on a nanopore instrument, first high quality, long DNA molecules have to be extracted intact from samples. This is a specialized technique and a sample specific problem which has required the generation of specialized extraction methods such as the wrinkled silica of Circulomics Nanobind and modified spin columns of RevoluGen's Firemonkey (Zhang et al. 2016; Gong et al. 2019). Depletion of short molecules can also occur after

extraction using a combination of polyethylene glycol and NaCl, the Circulomics short read eliminator kit (SRE) or with the BluePippin automated gel cassette from SAGE sciences (Kovaka et al., 2021; Wiley and Miller, 2020; Law, Warren, and McCallion, 2020; Kovaka et al., 2021). Once purified, sequencing adapters with bound motor protein are ligated to double-stranded DNA. The bound motor protein ensures translocation control and a tether included in the adaptor places the molecule on the surface, reducing the pore capture problem to a 2D diffusion concern, limiting the amount of time a pore stays empty, though still only a small fraction of input molecules is sequenced. Single strands of the library molecule are then sequenced from 5'-3', generating a sequencing read (Figure 1.1 A).

But because this sequencing is not dependent on synthesis, we can also characterize RNA molecules directly by passing them through a pore, as the original experiments on nanopore sequencing did. By attaching a sequencing adaptor to the 3' end of RNA molecules with a bound motor protein compatible with RNA, direct RNA nanopore sequencing is possible (Figure 1.1 B).

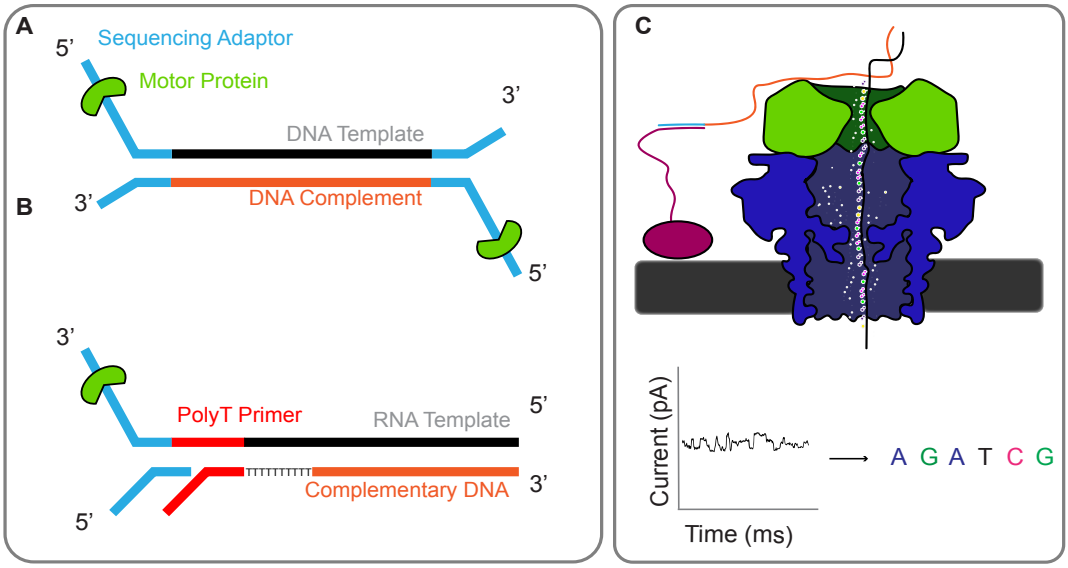

Figure 1.1: *Nanopore sequencing A) For DNA nanopore sequencing, adapters (blue) are ligated to end-prepped and dA-tailed native DNA . Each adapter has a motor protein (green) for aiding in translocation through the pore from the 5′ end to the 3′ end. B) In nanopore direct RNA libraries an adapter with a 10(dT) (red) overhang ligates to the polyadenylated tail of transcripts followed by a reverse transcription step to linearize the RNA template, can also be replaced with a custom adapter so that any RNA sequence can be targeted with a custom primer. Then, sequencing adapters with motor proteins (green) are ligated in the last step of library preparation. RNA is sequenced from the 3′ polyadenylated tail to the 5′ cap. C) DNA electrophoretically translocates the modified CsgG protein pore. Electric current is measured as a function of time and nucleic acid bases can be associated with signature electrical fluctuations.*

1.2 Accuracy

Despite the many improvements to nanopore sequencing over the years, problems with translocation control and signal to noise that riddled the technology in its early years are still pervasive. Nanopore sequencing in its current formulation does not interrogate a single base at a time - rather multiple bases have a significant impact on the current. The primary influence on the current is generally accepted to be a group of 5-6 nucleotides for the R9.4 pore. With the

newest iteration of the motor protein (named E8), DNA translocates through the pore at an average speed of ∼450b/s and the electrical current is sampled about 9 times per k-mer (4kHz) (Rang, Kloosterman, and de Ridder 2018; de Lannoy, de Ridder, and Risse 2017). However, the speed of translocation is not quite uniform, making the basecalling of homopolymeric regions a difficult task and results in systematic errors mostly in the form of insertions and deletions (indels) (Winston Timp, Comer, and Aksimentiev 2012; O'Donnell, Wang, and Dunbar 2013). This speed can further be influenced by the number of library molecules loaded in the pore - too many and the speed drops due to a lack of fuel for the motor protein, too few and the speed is faster. The rate of translocation for RNA is much lower (∼70b/s) and fluctuates significantly. This results in a lower signal to noise ratio making native RNA sequencing more erroneous (∼85%) (Viehweger et al., 2019).

To translate the electrical signal per k-mer into sequence, the first generation of ONT basecallers relied on Hidden Markov Models (HMMs), these are a class of probabilistic models that allow prediction of a sequence of unknown variables (nucleotide bases) from a set of observed variables (ionic current). It was shown that HMMs could successfully decode trinucleotides in 2012 (Winston Timp, Comer, and Aksimentiev 2012). ONT's first HMM basecaller, Albacore was released in 2017 and relied on an intermediary stage known as "event detection". The later versions of Albacore transitioned to 'raw basecalling' with a transducer-based model which calls bases directly from the signal data skipping the event detection step. This improved basecalling in homopolymeric regions with single read accuracy of Q9.2 and consensus

accuracy of Q21.9 (Wick, Judd, and Holt, 2019). ONT's next version of the basecaller, Guppy, was released in late 2017. Early versions of Guppy did not perform much better than Albacore, however its GPU compatibility made it substantially faster to run. The most recent version of Guppy contains two basecalling algorithms, the current baseline algorithm contained in Albacore and the addition of the "flip-flop" algorithm. Flip-flop uses a neural network implementation that substantially improves consensus accuracy and calling of long homopolymer regions. Guppy flip-flop generates reads with consensus accuracy of 99.5%, a significant improvement over the previous basecallers (Wick, Judd, and Holt, 2019). ONT also offers a neural network training toolkit, Tiayaki, which can be used to develop models for Guppy. While the model generation and training process can often be time-consuming and labor intensive, the performance of a neural net relies heavily on the quality of the training data. Therefore the most accurate base calls are obtained from models trained on native DNA of the same species (Wick, Judd, and Holt, 2019). Additionally, several independently developed basecallers from researchers have become available including Chiron, Nanocall, DeepNano, and basecRAWler (David et al., 2017; Boža, Brejová, and Vinař, 2017).

The resulting basecalling should be assessed at both the individual read level and at as a measure of the consensus accuracy. Errors at the read level complicate interpretation of heterogeneous samples or low-coverage data, while errors at the consensus level are harder to solve even with high coverage sequencing. Even low accuracy individual reads can lead to a high accuracy consensus as long as the error is random rather than systematic. Unfortunately,

nanopore is subject to specific errors in certain stretches and conformations of k-mers. In some cases this can be resolved by examining reads on the plus and minus strand - the reverse complement strand may not have the same systematic error as the original strand. This can even be employed to increase *per molecule* accuracy with the “1D^2" sequencing mode where both strands are sequenced, though this can reduce the yield. Low complexity sequences, e.g. homopolymers, are hard to resolve accurately. While systematic errors in low-complexity regions are still a pervasive problem with nanopore sequencing, random errors can be solved with increased sequencing coverage. Methods such as Intramolecular-ligated Nanopore Consensus Sequencing (INC-Seq) and Rolling Circle Amplification to Concatemeric Consensus (R2C2)capitalize on the strand displacement and processivity of phi29 DNA polymerase to perform rolling circle amplification on circularized template molecules (Li, Xiong, and Yi, 2016; Volden et al., 2018; Cole et al., 2020). The resulting library consists of long DNA or cDNA molecules made up of multiple repeating units. After sequencing, these repeating sequences are corrected by generating a consensus sequence improving accuracy to greater than 94% and 97%, respectively for R2C2 and INC-seq (Li, Xiong, and Yi, 2016; Volden et al., 2018).

Libraries generated from high quality HMW DNA can generate sequencing runs with N50s greater than 100kb, albeit with a significant cost to sequencing yield (Jain et al., 2018). Often reads can be up to megabases long, these reads have been colloquially referred to as “whales", and their appearance in sequencing runs is becoming more common as UHMW DNA extraction

methods are improving. Often these “whales" can also be incorrectly split by ONT’s MinKNOW software during sequencing, but can be computationally “reattached" with tools like BulkVis (Payne et al., 2019).

1.3 DNA Assembly

Genome assembly aims reconstruct the full genome sequence of the organism by first organizing the sequencing reads to *contigs*, which are then ordered and oriented into larger *scaffolds* with gaps between them contigs (Figure 1.2 A). Most plant and animal genomes have high levels of repeated and duplicated sequences that cause ambiguities in the ordering of genome segments (Simpson and Pop 2015). A great example of this is the human genome: after nearly two decades of improvements from its initial completion, the current human reference genome (GRCh38) is the most accurate and complete vertebrate genome ever produced. However, gaps represented by stretches of Ns still persist (Chaisson et al., 2015; Guo et al., 2017).

While the development of NGS has revolutionized the field of genomics by making whole genome sequencing rapid and affordable, short reads alone result in fragmented assemblies because most repetitive sequences longer than the read length can not be resolved (Nagarajan and Pop, 2009). Contigs assembled from long reads can be $\sim$30 to 300-fold longer than those assembled from short reads (Rhie et al., 2020). Long reads reduce the number of gaps, but the decrease in accuracy when compared to NGS adds additional challenges to long read assembly. The most common method for de novo assembly, the de Bruijn graph, is confounded by sequencing errors (Nagarajan and Pop 2009;

Simpson and Pop 2015). The overlap-layout-consensus (OLC) algorithm was revived for long read assembly because it can handle inconsistencies in read length and a relatively high number of sequencing errors making it excellent for assembly from nanopore or PacBio data (Koren et al., 2017).

The high error rate from long-read sequencing data can work to the detriment of assembly. The initial assembly algorithms can correct many errors simply looking at the consensus (Koren et al., 2017; Kolmogorov et al., 2019), which will eliminate many errors with sufficient coverage, but systematic errors will persist in the resulting assembly. One solution to these errors is subsequent polishing algorithms such as nanopolish or medaka. Nanopolish examines the assembly and assesses the likelihood of alternative sequences using the raw electrical data to find the most likely consensus (Loman, Quick, and Simpson, 2015). Nanopolish substantially increases the per base accuracy of the consensus sequence and also improves basecalling of homopolymer tracks (Wick, Judd, and Holt, 2019; Loman, Quick, and Simpson, 2015). Medaka instead uses a trained neural network against the aligned reads - by training on known sequences, they have established models that perform well to correct errors using only the basecalled reads.

Alternatively accuracy can be increased by simple rounds of consensus generation using either the original long reads (Racon) or highly accurate Illumina short reads (Racon, Pilon, FreeBayes, or POLCA) (Walker et al., 2014; Garrison and Marth, 2012; Zimin and Salzberg, 2019). Unfortunately these short-read consensus polishers are limited by the mappability of the short reads - so cannot polish inside highly repetitive areas or other areas of .

Other genome assembly methods involve using both the short and long reads for assembly in a process known as “hybrid assembly". These hybrid approaches can use long reads to construct the structure of the genome and fill in the bases with short accurate reads or combine short reads together into longer “super-reads" and scaffold these “super-reads" into “mega-reads" using the long read nanopore data (Koren et al., 2012; Zimin et al., 2013). Lastly, contigs can be scaffolded using a variety of data modalities that capture long range information (i.e. mate pair, HiC, optical mapping). Nanopore sequencing has been used to generate chromosome level reference genomes for a multitude of model and non-model organisms and entire microbial communities (Nicholls et al., 2019; Hamner et al., 2019).

1.4 Structural Variants

Since the development of high-throughput sequencing methods (Illumina), our understanding and study of mutations/alterations to the human genome have exploded. Characterization of small nucleotide variations inform Mendelian diseases, genetic predispositions, and different cancers have allowed us to form mechanistic insight. However, most of these studies focused on what is easiest to detect with short-read sequencing, small nucleotide variations. These variations are identified via alignment to the reference genome and identification via a suite of tools (e.g., freebayes or GATK).

In contrast, structural variations, defined as genomic alterations larger than 50bp which encompass deletions, duplications, insertions, inversions and translocations, describe major rearrangements in the genome. But these

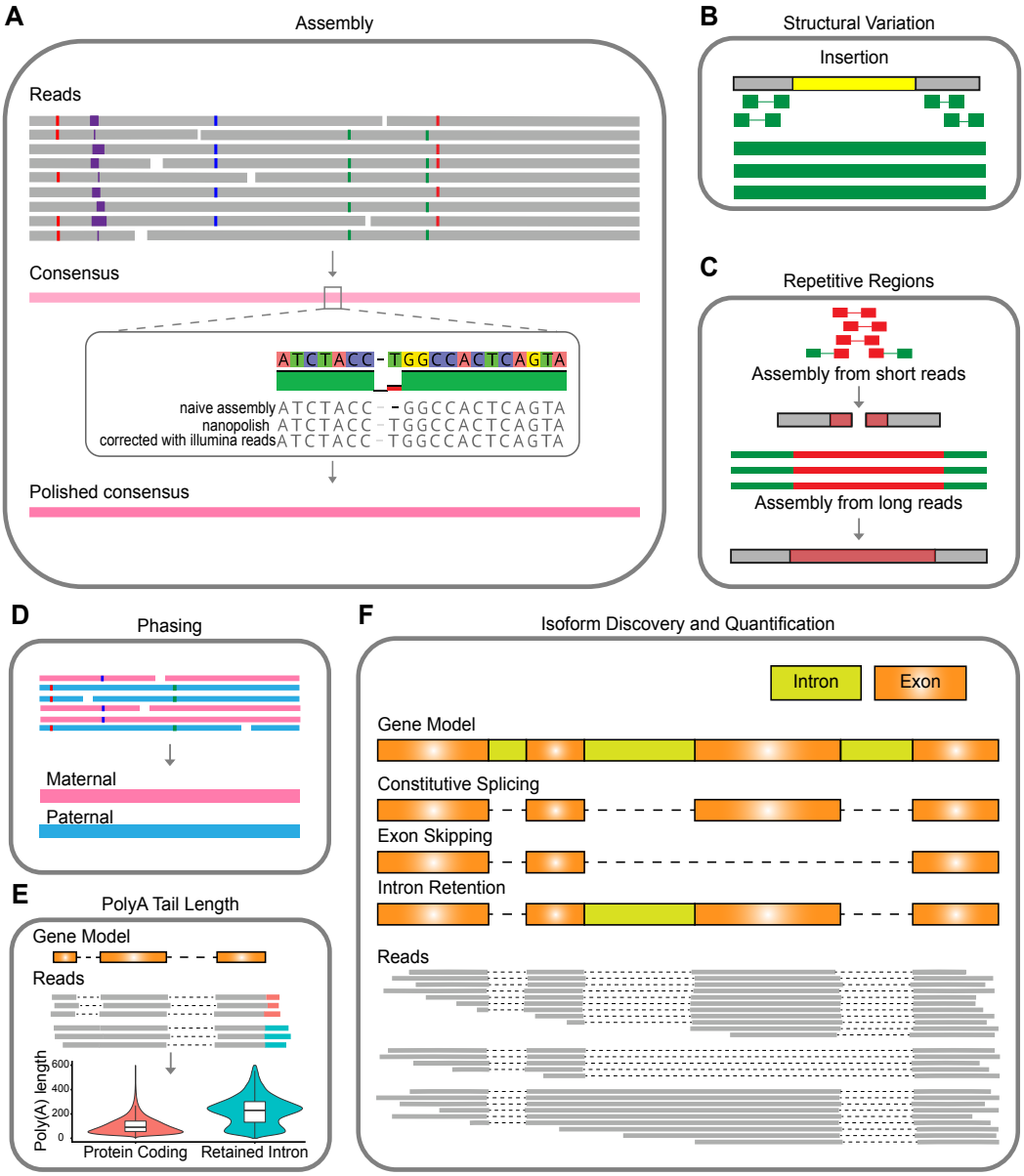

Figure 1.2: *Long read genome assembly A) Overview of whole genome de novo assembly from nanopore sequencing reads.* ***B)*** *Example of aligned long and short reads over an insertion.* ***C)*** *Example of aligned long and short reads across a large repetitive array.* ***D)*** *Longer reads allow for each read to contain multiple SNPs, this makes phasing of reads into maternal and paternal alleles possible.* ***E)*** *PolyA tail Analysis* ***F)*** *Isoform discovery*

are hard to identify with short-read sequencing - though in principle changes in coverage and split read alignments can identify SVs, in practice issues of mappability

Detecting and visualizing structural variations (SVs) is critical for understanding the relationship between SVs, human traits, and diseases. Deducing SVs from Illumina paired end data is well established and highly used, however these methods lack sensitivity (only 10-70% of variants detected), and have very high false positive rates (up to 89%) (Pang et al., 2010; English, Salerno, and Reid, 2014; Mahmoud et al., 2019). Conversely, long read sequencing considerably increases correct detection of SVs because of both higher quality genome assemblies and accurate mapping of complex regions (Mahmoud et al. 2019; Sedlazeck et al. 2018). Ultra long nanopore reads have been used to anchor long and complex variants allowing for accurate analysis of previously unexplored regions of the human genome such as complex repetitive arrays (ie. centromeres) and nested SVs (ie. INVDUPs and INVDELs) involved in human cancer (Jain et al., 2018; Miga et al., 2020) (Figure 1.2 B).

By exploiting nanopore long read data it is possible to reconstruct the diverse architecture of SVs responsible for normal human genetic variation and those implicated in human disease (Miao et al. 2018; Gong et al. 2018; Norris et al. 2016; Sakamoto et al. 2019). Currently nanopore technology is becoming more widely utilized in the medical community as a routine diagnostic tool for discovery of novel Mendelian diseases (Mantere, Kersten, and Hoischen 2019). The developments of software pipelines such as Sniffles, FreeBayes,

NanoVAR, SVIM, NanoSV, and Picky, increase accuracy in detecting and diagnosing novel SVs involved in human disease that gene panels and whole exome sequencing have failed to detect (Mantere, Kersten, and Hoischen 2019; Gong et al. 2018; Heller and Vingron 2019; Sedlazeck et al.; Garrison and Marth 2012; Miao et al. 2018; Tham et al. 2020; Cretu Stancu et al. 2017)

Expansion of variable number tandem repeats (VNTRs) causes more than 30 Mendelian human disorders. Short read sequencing is ineffective at sequencing through GC rich VNTRs due to PCR biases and has difficulty mapping longer VNTRs. For this reason, long read sequencing is useful in identifying novel VNTRs and quantifying the number of tandem repeats. Basecalling in VNTR regions has been reported to be more highly error prone due to systematic errors in tandem repeat calling (Mitsuhashi et al., 2019). While the basecalls in these regions are unreliable, signal level analysis algorithms (ie. STRique, NanoSatellite) have been developed in order to estimate tandem repeat numbers from raw nanopore signal data (Gießelmann et al 2018, De Roeck et al 2018). NanoSTRique uses a dynamic time warping (DTW) approach to identify VNTR spanning reads from the raw signal data and follows this with a hidden Markov model based count on the signal of interest (Gießelmann et al., 2018). This approach was successful in quantifying VNTR number for the repeat expansion loci associated with Frontotemporal Dementia (FTD) and Amyotrophic Lateral Sclerosis (ALS) (Gießelmann et al., 2018). NanoSatellite follows a similar DTW approach and has been used in clinical settings to characterize the VNTR alleles in Alzheimer's disease (Roeck et al.).

In addition to detecting novel SVs, advances in nanopore targeted sequencing technologies make it possible to sequence native DNA at specific regions at high depth to more deeply probe SVs at known locations (Gabrieli et al., 2018; Gießelmann et al., 2018; Kovaka et al., 2021). The 'nanopore Cas9 Targeted-Sequencing' (nCATS) method utilizes the ability of Cas9 to make cuts at specific locations then ligates nanopore adapters in order to enrich for specific loci without any amplification biases or loss of DNA modifications. Targeted long read sequencing can detect SVs ranging from large chromosomal deletions to SNPs with high accuracy and sensitivity even in long repetitive loci such as human tumor suppressor gene BRCA1, which is responsible for the onset of many breast and ovarian cancers (Gilpatrick et al., n.d.). Additionally, nanopore adaptive sequencing with the ONT ReadUntil API allows nanopore devices to selectively eject individual reads from the pore in real-time. This has inspired the development of open source software UNCALLED and ReadUntil to rapidly match streaming nanopore current signals to a reference sequence (Kovaka et al. 2020; Payne et al. 2020). UNCALLED enriched 148 human genes associated with hereditary cancers enabling accurate detection of SNPs, indels, structural variants (SVs), and methylation and detected twice as many SVs compared to 50x coverage Illumina sequencing (Kovaka et al. 2020; Payne et al. 2020). ReadUntil adaptive sequencing was used to enrich 25,600 target regions covering nearly 10,000 genes and 717 genes implicated in cancer (Kovaka et al. 2020; Payne et al. 2020).

1.5 RNA Sequencing

RNA sequencing has emerged as a crucial tool over recent years to investigate different characteristics of the transcriptome such as differential gene expression, splicing variation, gene annotations, ribosomal profiling, etc. Illumina is the current gold standard short-read RNA sequencing platform accounting for the majority of published RNA-seq data on SRA (Stark, Grzelak, and Hadfield 2019). Long-read cDNA sequencing has significantly improved the quality of transcriptome-wide analysis by identifying longer transcripts. These longer reads not only enhance the detection of splice-junctions but also result in capturing diverse isoforms (Stark, Grzelak, and Hadfield 2019). This has caused the emergence of new computational tools that integrate these long-read cDNA reads in genome annotation (Cook et al., 2019; Lagarde et al., 2017). It has also been shown that ONT cDNA sequencing is capable of generating full-length transcript reads even with low RNA input (for example, single-cell experiments) (Oikonomopoulos et al., 2016). This technology then can be used to investigate different characteristics of RNA such as splicing variation, kinetics, alternative polyadenylation, and post-translational modifications at the isoform level and their relevance in a variety of fields in biology and medicine.

1.6 Full-length isoform discovery

Perhaps one of the most obvious advantages is the possibility of novel isoform discovery even in very well-characterized samples. For example, studies

have identified a considerable number of novel isoforms in lymphoblastoid cell lines (Workman et al., 2018; Jong et al., 2017). Another interesting study recently revealed the potential of nanopore sequencing for investigation of full-length circRNAs in human and mouse brains. Rahimi et al. reported more than 200 novel exons used in circRNAs (Rahimi et al., 2019).

Despite its strength in full length discovery, long-read cDNA/RNA sequencing still faces significant challenges. First is the 5′ truncation common to these methods. This could happen due to a variety of reasons such as RNA degradation, and sample handling (Stark, Grzelak, and Hadfield 2019). Some studies have tried to address this issue for both long-read RNA and cDNA sequencing. Jiang et al. utilized a 5′-Cap capturing approach to look at the impact of Piwi on the exonization of TEs in loci. In this approach, 7-methylguanosine 5′-capped RNAs are enriched using a biotinylated RNA adapter. In the case of long-read cDNA sequencing, different reverse transcriptases can be used that convert only 5′-capped mRNAs to cDNA (Jiang et al., 2019). Sessegolo et al. also have investigated the usage of commercialized TeloPrime amplification kit (Lexogen) that is selective to both capped and polyadenylated RNA molecules (Sessegolo et al. 2019).

Polyadenylation is one of the vital RNA regulation mechanisms that impacts nuclear transport, RNA stability, and translation initiation. Nanopore sequencing has made the assessment of poly(A) tail length in a transcriptome-wide high-throughput manner possible. Recently, nanopolish-polya was introduced as a computational tool that accurately estimates the poly(A) tail length at the read-level in direct-RNA sequencing. Nanopolish achieves this

by segmenting the signal into four regions of start, leader, adapter, poly(A) tail, and transcript using a Hidden Markov Model (HMM). After correcting for the different translocation speeds across the reads, the length of poly(A) tail is estimated. This was then used to show the correlation of poly(A) tail length with different characteristics of RNA at both transcriptome and isoform levels (Workman et al., 2018). Another tool that implements a similar approach is tailfindr, an alignment-free poly(A) length estimator that works for not only direct-RNA sequencing but also cDNA sequencing (Krause et al., 2019).

1.7 Base Modifications

Another major advantage of nanopore sequencing is the ability to obtain information about non-canonical nucleic acid bases from raw signal data. Base modifications to DNA and RNA play major roles in integral cellular processes such as aging, gene regulation, imprinting, gene expression, transcript localization and disease (Field et al., 2018; Gibney and Nolan, 2010; Kumar, Chinnusamy, and Mohapatra, 2018; Macdonald, 2012; Liyanage et al., 2014). The current 'gold standard' method for profiling 5mC in DNA is bisulfite sequencing (Patterson et al., 2011). Sodium bisulfite treatment converts unmodified cytosines to uracil, while leaving methylated cytosines unchanged (Clark et al., 1994; Frommer et al., 1992). Research has shown the presence of other DNA base modifications in prokaryotes such as N6-Methyladenosine and N4-methylcytosine. Although a variety of methods have been developed to detect these modifications, they are not nearly as accurate and high throughput as bisulfite sequencing. Research has shown the presence of other

DNA base modifications in prokaryotes such as N6-Methyladenosine and N4-methylcytosine. Although a variety of methods have been developed to detect these modifications, they are not nearly as accurate and high throughput as bisulfite sequencing. Despite being the current gold standard, bisulfite treatment is harsh and damaging to the DNA, resulting in DNA degradation and significant sample loss (Kint et al., 2018). In a newer method called Enzymatic Methyl-seq (EM-seq), 5mC and 5hmC modified bases are detected more accurately. EM-seq is specifically superior to bisulfite sequencing due to the fact that it has less sequencing bias and requires minimal DNA input (Vaisvila et al. 2019). In addition, the short reads gained from bisulfite sequencing reveal short-range patterns, however, long range methylation information can reveal allele specific patterns, particularly those involved in imprinting (Gigante et al., 2019).

Base modifications from non-canonical nucleotides introduce unique deviations in the signal data making them detectable. Calling base modifications typically involves traditional basecalling, mapping the raw signal to a genomic reference and then computing if a base is modified based on evidence from the signal (Gouil and Keniry, 2019). Nanopolish methylation caller is a pretrained package that detects 5-methylcytosine in a CpG context by employing a HMM. The HMM uses a table of event level distributions characteristic to every k-mer, termed a pore model, to decipher the methylation state of k-mers (David et al., 2017). Other software such as signalAlign, mcaller, DeepSignal, and DeepMod use either HMMs or neural networks to detect both 5mC and 6-mA modifications (McIntyre et al., 2019; Rand et al., 2017). ONT also offers

software packages Taiyaki, Megaladon and Tombo for training and calling nucleic acid base modifications.

There are over 100 known post-translational modifications to RNA, but the most frequently studied include N6-methyladenosine (m6A), 5-methylcytosine, inosine, pseudouridine, 7-methylguanosine, and N1-methyladenosine (Zhao, Roundtree, and He, 2017). 5'cap and poly(A) tail play arguably the most crucial roles in RNA regulation processes such as transcript stability, splicing, nuclear export, and translation initiation (Roundtree et al., 2017). Profiling RNA modifications have been proven to be crucial in order to better understand their role in RNA regulation and human disease. Current methods for profiling modifications in RNA are often complex, inefficient, and do not offer combinatorial measurement of multiple modifications simultaneously, i.e., detection of 6-mA and 5-mC combinations (Li, Xiong, and Yi, 2016). Since the introduction of direct RNA sequencing by ONT, a few studies have tried to investigate the possibility of RNA modification detection using this platform.

In one of the most recent studies, Liu et al. have combined using the raw signal with basecalling errors to accurately call m6A modifications in a new tool called EpiNano. EpiNano takes advantage of support vector machines (SVMs) to train a model on in-vitro transcribed reads with all possible 5-mers that incorporate m6A. They report that using 5-mer current intensity, read quality, per-base quality, per-base mismatch frequency, per-base deletion frequency, and per-base insertion frequency as features of the proposed SVM improve the accuracy of calling m6A sites up to 97%-99% (H. Liu et al. 2019). Although these studies have contributed significantly to this area using this

new technology, there are still crucial challenges that require more thorough investigations. While it is relatively easy to identify modification sites when investigating reads in aggregate, confidently calling modified bases at a single-read level is difficult. ONT's Taiyaki enables users to train their own models for basecalling of nanopore dRNA reads. While this can be a major enhancement, Constructing long RNA molecules with modifications as training sets is quite expensive and challenging.

1.8 Dynamics and labeling

Exogenous labels can be added to native nucleic acid and later detected upon sequencing. This methodology has been utilized to study structure and dynamics of DNA and RNA. In the case of RNA, studying nascent RNA can provide a general understanding of how enhancer-mediated gene regulation works (J. (Wang et al., 2018). Often multiple exons in a pre-mRNA go through differential splicing. In order to understand the dynamics and regulation of these splicing events across nascent transcripts, comprehensive methods to assess RNA processing phenomena in vivo. In a recent study Drexler et al. have developed a technique named nano-COP in which nascent RNAs are directly sequenced using nanopore sequencing. nano-COP uses 4sU labeling of RNA to capture nascent RNA. This provides a valuable tool to study the dynamics and patterns of RNA splicing omitting amplification bias. Using this method, it was shown that co-transcriptional splicing often happens after transcription of several kilobases of pre-mRNA by RNA polymerase II, suggesting splicing machinery starts working as transcription takes place

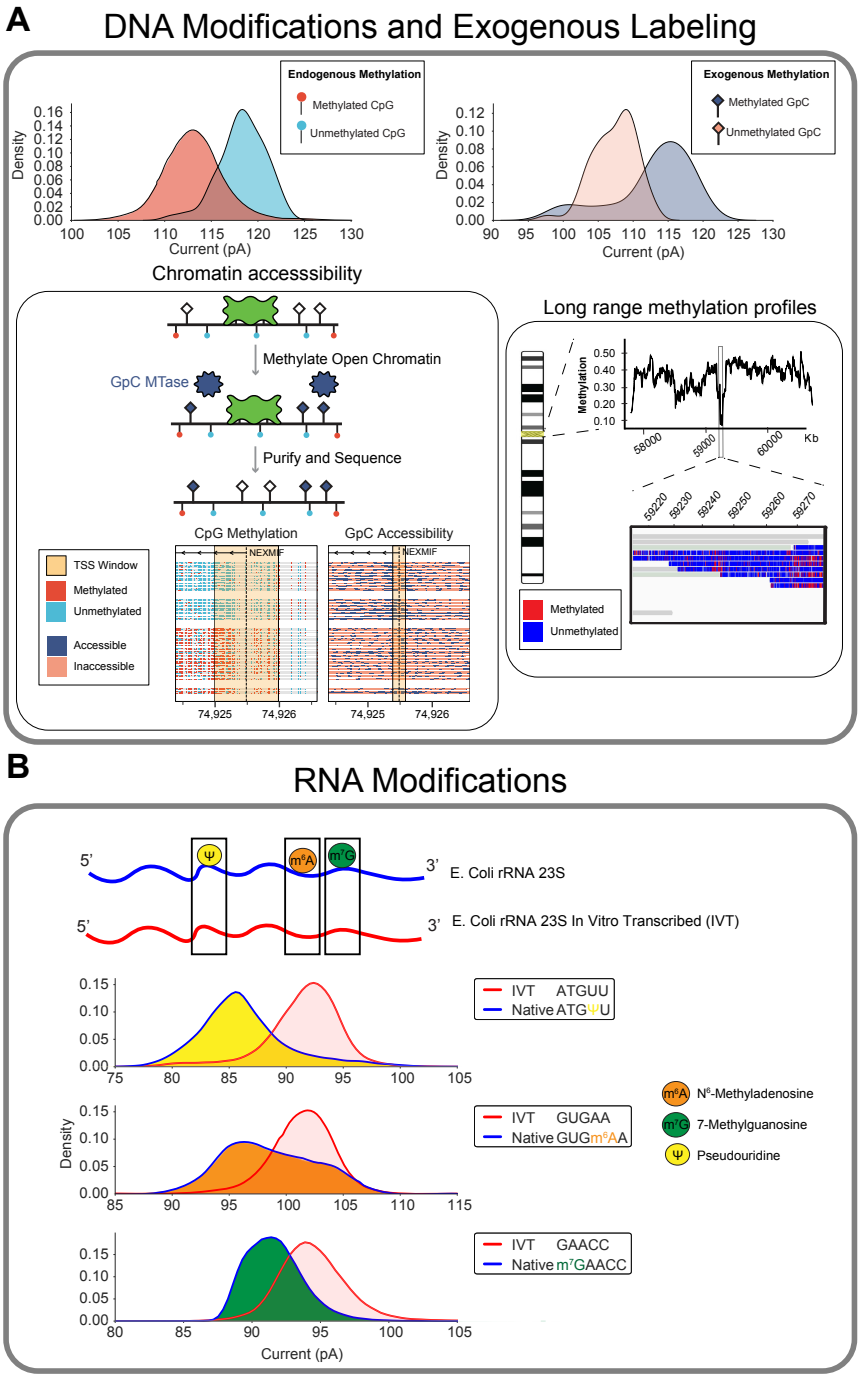

Figure 1.3: *A) DNA modifications and Exogenous labeling. B) RNA modifications*

(Drexler, Choquet, and Stirling Churchman, 2019).

Another study introduces a somewhat similar method called nano-ID to investigate RNA metabolism at single RNA molecules and isoforms level in different cell states and conditions. nano-ID utilizes 4sU labeling without enriching for the labeled RNA. By incorporating some of the RNA modification methods described above, the labeled (nascent transcripts) and unlabeled (existing pre-mRNA) sequencing reads can be distinguished computationally. Maier et al. reported the change in synthesis rate, stability, and splicing pattern at the isoform level in human cells treated by heat shock by this method (Maier et al., 2019).

Signal level analysis has also been used to capture dynamics of DNA during genome replication using pulsed in BrdU with either D-NAscent or RepNano (Hennion et al., 2018). These methods detect differences in BrdU incorporation frequency across individual molecules. This information can be used to reveal the location of active replication origins, fork direction, termination sites, and fork pausing/stalling events. Nanopore sequencing can also detect non-endogenous GpC methylation for profiling chromatin accessibility (Lee et al. 2018, Shipony, Marinov, and Swaffer 2018). The SMAC-seq method treats DNA with m6A and CpG and GpC 5mC methyltransferases which preferentially methylate DNA in open regions of chromatin. When the DNA is subsequently sequenced, highly methylated regions are indicative of open chromatin regions (Shipony, Marinov, and Swaffer 2018). The MeSMLR-seq and NanoNOMe methods treat DNA with GpC 5mC methyltransferases to profile chromatin accessibility and nucleosome occupancy (Lee et al., 2018).

Detection of methylated DNA has also proven useful in the case of binning metagenomic contigs, associating mobile genetic elements with their host genomes, and identifying misassembled metagenomic contigs (Tourancheau et al. 2020).

1.9 References

Boža, Vladimír, Broňa Brejová, and Tomáš Vinař (2017). "DeepNano: Deep recurrent neural networks for base calling in MinION nanopore reads". en. In: *PLoS One* 12.6, e0178751.

Buermans, H P J and J T den Dunnen (2014). "Next generation sequencing technology: Advances and applications". en. In: *Biochim. Biophys. Acta* 1842.10, pp. 1932–1941.

Chaisson, Mark J P, John Huddleston, Megan Y Dennis, Peter H Sudmant, Maika Malig, Fereydoun Hormozdiari, Francesca Antonacci, Urvashi Surti, Richard Sandstrom, Matthew Boitano, Jane M Landolin, John A Stamatoyannopoulos, Michael W Hunkapiller, Jonas Korlach, and Evan E Eichler (2015). "Resolving the complexity of the human genome using single-molecule sequencing". In: *Nature* 517.7536, pp. 608–611.

Cherf, Gerald M, Kate R Lieberman, Hytham Rashid, Christopher E Lam, Kevin Karplus, and Mark Akeson (2012). "Automated forward and reverse ratcheting of DNA in a nanopore at 5-Å precision". en. In: *Nat. Biotechnol.* 30.4, pp. 344–348.

Church, George, David W Deamer, Daniel Branton, Richard Baldarelli, and John Kasianowicz (1998). "Characterization of individual polymer molecules based on monomer-interface interactions". Pat. 5795782.

Clark, S J, J Harrison, C L Paul, and M Frommer (1994). "High sensitivity mapping of methylated cytosines". en. In: *Nucleic Acids Res.* 22.15, pp. 2990–2997.

Cole, Charles, Ashley Byrne, Matthew Adams, Roger Volden, and Christopher Vollmers (2020). "Complete characterization of the human immune cell transcriptome using accurate full-length cDNA sequencing". en. In: *Genome Res.* 30.4, pp. 589–601.

Cook, David E, Jose Espejo Valle-Inclan, Alice Pajoro, Hanna Rovenich, Bart P H J Thomma, and Luigi Faino (2019). "Long-Read Annotation: Automated Eukaryotic Genome Annotation Based on Long-Read cDNA Sequencing". en. In: *Plant Physiol.* 179.1, pp. 38–54.

David, Matei, L J Dursi, Delia Yao, Paul C Boutros, and Jared T Simpson (2017). *Nanocall: an open source basecaller for Oxford Nanopore sequencing data*.

Davies, Kevin (2015). *The $1,000 Genome: The Revolution in DNA Sequencing and the New Era of Personalized Medicine*. en. Simon and Schuster.

Deamer, David, Mark Akeson, and Daniel Branton (2016). "Three decades of nanopore sequencing". en. In: *Nat. Biotechnol.* 34.5, pp. 518–524.

Drexler, Heather L, Karine Choquet, and L Stirling Churchman (2019). "Human co-transcriptional splicing kinetics and coordination revealed by direct nascent RNA sequencing". en.

English, Adam C, William J Salerno, and Jeffrey G Reid (2014). "PBHoney: identifying genomic variants via long-read discordance and interrupted mapping". en. In: *BMC Bioinformatics* 15, p. 180.

Field, Adam E, Neil A Robertson, Tina Wang, Aaron Havas, Trey Ideker, and Peter D Adams (2018). "DNA Methylation Clocks in Aging: Categories, Causes, and Consequences". en. In: *Mol. Cell* 71.6, pp. 882–895.

Franklin, R E and R G Gosling (1953). "Molecular configuration in sodium thymonucleate". en. In: *Nature* 171.4356, pp. 740–741.

Frommer, M, L E McDonald, D S Millar, C M Collis, F Watt, G W Grigg, P L Molloy, and C L Paul (1992). "A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands". en. In: *Proc. Natl. Acad. Sci. U. S. A.* 89.5, pp. 1827–1831.

Gabrieli, Tslil, Hila Sharim, Dena Fridman, Nissim Arbib, Yael Michaeli, and Yuval Ebenstein (2018). "Selective nanopore sequencing of human BRCA1 by Cas9-assisted targeting of chromosome segments (CATCH)". en. In: *Nucleic Acids Res.* 46.14, e87.

Garrison, Erik and Gabor Marth (2012). "Haplotype-based variant detection from short-read sequencing". In: arXiv: `1207.3907 [q-bio.GN]`.

Gibney, E R and C M Nolan (2010). "Epigenetics and gene expression". en. In: *Heredity* 105.1, pp. 4–13.

Gießelmann, P, B Brändl, E Raimondeau, R Bowen, and others (2018). "Repeat expansion and methylation state analysis with nanopore sequencing". In.

Gigante, Scott, Quentin Gouil, Alexis Lucattini, Andrew Keniry, Tamara Beck, Matthew Tinning, Lavinia Gordon, Chris Woodruff, Terence P Speed, Marnie E Blewitt, and Matthew E Ritchie (2019). "Using long-read sequencing to detect imprinted DNA methylation". en. In: *Nucleic Acids Res.* 47.8, e46.

Gilpatrick, Timothy, Isac Lee, James E Graham, Etienne Raimondeau, Rebecca Bowen, Andrew Heron, Fritz J Sedlazeck, and Winston Timp (n.d.). *Targeted Nanopore Sequencing with Cas9 for studies of methylation, structural variants, and mutations*.

Gouil, Quentin and Andrew Keniry (2019). "Latest techniques to study DNA methylation". en. In: *Essays Biochem.* 63.6, pp. 639–648.

Goyal, Parveen, 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 (2014). "Structural and mechanistic insights into the bacterial amyloid secretion channel CsgG". en. In: *Nature* 516.7530, pp. 250–253.

Guo, Yan, Yulin Dai, Hui Yu, Shilin Zhao, David C Samuels, and Yu Shyr (2017). "Improvements and impacts of GRCh38 human reference on high throughput sequencing data analysis". en. In: *Genomics* 109.2, pp. 83–90.

Hamner, Steve, Bonnie L Brown, Nur A Hasan, Michael J Franklin, John Doyle, Margaret J Eggers, Rita R Colwell, and Timothy E Ford (2019). "Metagenomic Profiling of Microbial Pathogens in the Little Bighorn River, Montana". en. In: *Int. J. Environ. Res. Public Health* 16.7.

Hennion, M, J M Arbona, C Cruaud, F Proux, B Le Tallec, and others (2018). "Mapping DNA replication with nanopore sequencing". In: *BioRxiv*.

Holley, R W, G A Everett, J T Madison, and A Zamir (1965). "NUCLEOTIDE SEQUENCES IN THE YEAST ALANINE TRANSFER RIBONUCLEIC ACID". en. In: *J. Biol. Chem.* 240, pp. 2122–2128.

Hornblower, Breton, Amy Coombs, Richard D Whitaker, Anatoly Kolomeisky, Stephen J Picone, Amit Meller, and Mark Akeson (2007). "Single-molecule analysis of DNA-protein complexes using nanopores". en. In: *Nat. Methods* 4.4, pp. 315–317.

Jain, Miten, Hugh E Olsen, Daniel J Turner, David Stoddart, Kira V Bulazel, Benedict Paten, David Haussler, Huntington F Willard, Mark Akeson, and Karen H Miga (2018). "Linear assembly of a human centromere on the Y chromosome". en. In: *Nat. Biotechnol.* 36.4, pp. 321–323.

Jiang, Feng, Jie Zhang, Qing Liu, Xiang Liu, Huimin Wang, Jing He, and Le Kang (2019). "Long-read direct RNA sequencing by 5'-Cap capturing reveals the impact of Piwi on the widespread exonization of transposable elements in locusts". en. In: *RNA Biol.* 16.7, pp. 950–959.

Jong, Lucy C de, Simone Cree, Vanessa Lattimore, George A R Wiggins, Amanda B Spurdle, kConFab Investigators, Allison Miller, Martin A Kennedy, and Logan C Walker (2017). "Nanopore sequencing of full-length BRCA1 mRNA transcripts reveals co-occurrence of known exon skipping events". en. In: *Breast Cancer Res.* 19.1, p. 127.

Kasianowicz, J J, E Brandin, D Branton, and D W Deamer (1996). "Characterization of individual polynucleotide molecules using a membrane channel". en. In: *Proc. Natl. Acad. Sci. U. S. A.* 93.24, pp. 13770–13773.

Kint, Sam, Ward De Spiegelaere, Jonas De Kesel, Linos Vandekerckhove, and Wim Van Criekinge (2018). "Evaluation of bisulfite kits for DNA

methylation profiling in terms of DNA fragmentation and DNA recovery using digital PCR". en. In: *PLoS One* 13.6, e0199091.

Kolmogorov, Mikhail, Jeffrey Yuan, Yu Lin, and Pavel A Pevzner (2019). "Assembly of long, error-prone reads using repeat graphs". en. In: *Nat. Biotechnol.* 37.5, pp. 540–546.

Koren, Sergey, Michael C Schatz, Brian P Walenz, Jeffrey Martin, Jason T Howard, Ganeshkumar Ganapathy, Zhong Wang, David A Rasko, W Richard McCombie, Erich D Jarvis, and Adam M Phillippy (2012). "Hybrid error correction and de novo assembly of single-molecule sequencing reads". en. In: *Nat. Biotechnol.* 30.7, pp. 693–700.

Koren, Sergey, Brian P Walenz, Konstantin Berlin, Jason R Miller, Nicholas H Bergman, and Adam M Phillippy (2017). "Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation". en. In: *Genome Res.* 27.5, pp. 722–736.

Kovaka, Sam, Yunfan Fan, Bohan Ni, Winston Timp, and Michael C Schatz (2021). "Targeted nanopore sequencing by real-time mapping of raw electrical signal with UNCALLED". In: *Nature biotechnology* 39.4, pp. 431–441.

Krause, Maximilian, Adnan M Niazi, Kornel Labun, Yamila Nicole Torres Cleuren, Florian Sebastian Müller, and Eivind Valen (2019). "tailfindr: Alignment-free poly(A) length measurement for Oxford Nanopore RNA and DNA sequencing". en. In: *RNA*.

Kumar, Suresh, Viswanathan Chinnusamy, and Trilochan Mohapatra (2018). "Epigenetics of Modified DNA Bases: 5-Methylcytosine and Beyond". en. In: *Front. Genet.* 9, p. 640.

Lagarde, Julien, Barbara Uszczynska-Ratajczak, Silvia Carbonell, Sílvia Pérez-Lluch, Amaya Abad, Carrie Davis, Thomas R Gingeras, Adam Frankish, Jennifer Harrow, Roderic Guigo, and Rory Johnson (2017). "High-throughput annotation of full-length long noncoding RNAs with capture long-read sequencing". en. In: *Nat. Genet.* 49.12, pp. 1731–1740.

Lander, E S et al. (2001). "Initial sequencing and analysis of the human genome". en. In: *Nature* 409.6822, pp. 860–921.

Law, William D, René L Warren, and Andrew S McCallion (2020). "Establishment of an eHAP1 human haploid cell line hybrid reference genome assembled from short and long reads". en. In: *Genomics*.

Lee, I, R Razaghi, T Gilpatrick, N Sadowski, and others (2018). "Simultaneous profiling of chromatin accessibility and methylation on human cell lines with nanopore sequencing". In: *BioRxiv*.

Levene, M J, J Korlach, S W Turner, M Foquet, H G Craighead, and W W Webb (2003). "Zero-mode waveguides for single-molecule analysis at high concentrations". en. In: *Science* 299.5607, pp. 682–686.

Li, Xiaoyu, Xushen Xiong, and Chengqi Yi (2016). "Epitranscriptome sequencing technologies: decoding RNA modifications". en. In: *Nat. Methods* 14.1, pp. 23–31.

Liyanage, Vichithra R B, Jessica S Jarmasz, Nanditha Murugeshan, Marc R Del Bigio, Mojgan Rastegar, and James R Davie (2014). "DNA modifications: function and applications in normal and disease States". en. In: *Biology* 3.4, pp. 670–723.

Loman, Nicholas J, Joshua Quick, and Jared T Simpson (2015). "A complete bacterial genome assembled de novo using only nanopore sequencing data". en. In: *Nat. Methods* 12.8, pp. 733–735.

Loose, Matthew W (2017). "The potential impact of nanopore sequencing on human genetics". en. In: *Hum. Mol. Genet.* 26.R2, R202–R207.

Lu, Hengyun, Francesca Giordano, and Zemin Ning (2016). "Oxford Nanopore MinION Sequencing and Genome Assembly". en. In: *Genomics Proteomics Bioinformatics* 14.5, pp. 265–279.

Luckey, J A, H Drossman, A J Kostichka, D A Mead, J D'Cunha, T B Norris, and L M Smith (1990). "High speed DNA sequencing by capillary electrophoresis". en. In: *Nucleic Acids Res.* 18.15, pp. 4417–4421.

Macdonald, William A (2012). "Epigenetic mechanisms of genomic imprinting: common themes in the regulation of imprinted regions in mammals, plants, and insects". en. In: *Genet. Res. Int.* 2012, p. 585024.

Mahmoud, Medhat, Nastassia Gobet, Diana Ivette Cruz-Dávalos, Ninon Mounier, Christophe Dessimoz, and Fritz J Sedlazeck (2019). "Structural variant calling: the long and the short of it". en. In: *Genome Biol.* 20.1, p. 246.

Maier, Kerstin C, Saskia Gressel, Patrick Cramer, and Björn Schwalb (2019). "Native molecule sequencing by nano-ID reveals synthesis and stability of RNA isoforms". en.

Manrao, Elizabeth A, Ian M Derrington, Andrew H Laszlo, Kyle W Langford, Matthew K Hopper, Nathaniel Gillgren, Mikhail Pavlenok, Michael Niederweis, and Jens H Gundlach (2012). "Reading DNA at single-nucleotide resolution with a mutant MspA nanopore and phi29 DNA polymerase". en. In: *Nat. Biotechnol.* 30.4, pp. 349–353.

Mardis, Elaine R (2006). "Anticipating the 1,000 dollar genome". en. In: *Genome Biol.* 7.7, p. 112.

McIntyre, Alexa B R, Noah Alexander, Kirill Grigorev, Daniela Bezdan, Heike Sichtig, Charles Y Chiu, and Christopher E Mason (2019). *Single-molecule sequencing detection of N6-methyladenine in microbial reference materials*.

Meller, A, L Nivon, and D Branton (2001). "Voltage-driven DNA translocations through a nanopore". en. In: *Phys. Rev. Lett.* 86.15, pp. 3435–3438.

Miga, Karen H, Sergey Koren, Arang Rhie, Mitchell R Vollger, Ariel Gershman, Andrey Bzikadze, Shelise Brooks, Edmund Howe, David Porubsky, Glennis A Logsdon, Valerie A Schneider, Tamara Potapova, Jonathan Wood, William Chow, Joel Armstrong, Jeanne Fredrickson, Evgenia Pak, Kristof Tigyi, Milinn Kremitzki, Christopher Markovic, Valerie Maduro, Amalia Dutra, Gerard G Bouffard, Alexander M Chang, Nancy F Hansen, Françoisen Thibaud-Nissen, Anthony D Schmitt, Jon-Matthew Belton, Siddarth Selvaraj, Megan Y Dennis, Daniela C Soto, Ruta Sahasrabudhe, Gulhan Kaya, Josh Quick, Nicholas J Loman, Nadine Holmes, Matthew Loose, Urvashi Surti, Rosa Ana Risques, Tina A Graves Lindsay, Robert Fulton, Ira Hall, Benedict Paten, Kerstin Howe, Winston Timp, Alice Young, James C Mullikin, Pavel A Pevzner, Jennifer L Gerton, Beth A Sullivan, Evan E Eichler, and Adam M Phillippy (2020). "Telomere-to-telomere assembly of a complete human X chromosome". en. In: *Nature*, p. 735928.

Mitsuhashi, Satomi, Martin C Frith, Takeshi Mizuguchi, Satoko Miyatake, Tomoko Toyota, Hiroaki Adachi, Yoko Oma, Yoshihiro Kino, Hiroaki Mitsuhashi, and Naomichi Matsumoto (2019). "Tandem-genotypes: robust detection of tandem repeat expansions from long DNA reads". en. In: *Genome Biol.* 20.1, p. 58.

Nagarajan, Niranjan and Mihai Pop (2009). "Parametric complexity of sequence assembly: theory and applications to next generation sequencing". en. In: *J. Comput. Biol.* 16.7, pp. 897–908.

Nakamura, Kensuke, Taku Oshima, Takuya Morimoto, Shun Ikeda, Hirofumi Yoshikawa, Yuh Shiwa, Shu Ishikawa, Margaret C Linak, Aki Hirai, Hiroki Takahashi, Md Altaf-Ul-Amin, Naotake Ogasawara, and Shigehiko Kanaya (2011). "Sequence-specific error profile of Illumina sequencers". en. In: *Nucleic Acids Res.* 39.13, e90.

Nicholls, Samuel M, Joshua C Quick, Shuiquan Tang, and Nicholas J Loman (2019). "Ultra-deep, long-read nanopore sequencing of mock microbial community standards". en. In: *Gigascience* 8.5.

Nurk, Sergey, Brian P Walenz, Arang Rhie, Mitchell R Vollger, Glennis A Logsdon, Robert Grothe, Karen H Miga, Evan E Eichler, Adam M Phillippy,

and Sergey Koren (2020). “HiCanu: accurate assembly of segmental duplications, satellites, and allelic variants from high-fidelity long reads”. en.

Oikonomopoulos, Spyros, Yu Chang Wang, Haig Djambazian, Dunarel Badescu, and Jiannis Ragoussis (2016). “Benchmarking of the Oxford Nanopore MinION sequencing for quantitative and qualitative assessment of cDNA populations”. en. In: *Sci. Rep.* 6, p. 31602.

Pang, Andy W, Jeffrey R MacDonald, Dalila Pinto, John Wei, Muhammad A Rafiq, Donald F Conrad, Hansoo Park, Matthew E Hurles, Charles Lee, J Craig Venter, Ewen F Kirkness, Samuel Levy, Lars Feuk, and Stephen W Scherer (2010). “Towards a comprehensive structural variation map of an individual human genome”. en. In: *Genome Biol.* 11.5, R52.

Patterson, Kate, Laura Molloy, Wenjia Qu, and Susan Clark (2011). “DNA methylation: bisulphite modification and analysis”. en. In: *J. Vis. Exp.* 56.

Payne, Alexander, Nadine Holmes, Vardhman Rakyan, and Matthew Loose (2019). “BulkVis: a graphical viewer for Oxford nanopore bulk FAST5 files”. en. In: *Bioinformatics* 35.13, pp. 2193–2198.

Rahimi, Karim, Morten T Venø, Daniel M Dupont, and Jørgen Kjems (2019). “Nanopore sequencing of full-length circRNAs in human and mouse brains reveals circRNA-specific exon usage and intron retention”. en.

Rand, Arthur C, Miten Jain, Jordan M Eizenga, Audrey Musselman-Brown, Hugh E Olsen, Mark Akeson, and Benedict Paten (2017). “Mapping DNA methylation with high-throughput nanopore sequencing”. en. In: *Nat. Methods* 14.4, pp. 411–413.

Rhie, A, S A McCarthy, O Fedrigo, J Damas, G Formenti, and others (2020). “Towards complete and error-free genome assemblies of all vertebrate species”. In: *bioRxiv*.

Roundtree, Ian A, Molly E Evans, Tao Pan, and Chuan He (2017). “Dynamic RNA Modifications in Gene Expression Regulation”. In: *Cell* 169.7, pp. 1187–1200.

Sanger, F, G G Brownlee, and B G Barrell (1965). “A two-dimensional fractionation procedure for radioactive nucleotides”. en. In: *J. Mol. Biol.* 13.2, pp. 373–398.

Sanger, F, S Nicklen, and A R Coulson (1977). “DNA sequencing with chain-terminating inhibitors”. en. In: *Proc. Natl. Acad. Sci. U. S. A.* 74.12, pp. 5463–5467.

Schloss, Jeffery A, Richard A Gibbs, Vinod B Makhijani, and Andre Marziali (2020). "Cultivating DNA Sequencing Technology After the Human Genome Project". en. In: *Annu. Rev. Genomics Hum. Genet.* 21, pp. 117–138.

Venter, J C et al. (2001). "The sequence of the human genome". en. In: *Science* 291.5507, pp. 1304–1351.

Viehweger, Adrian, Sebastian Krautwurst, Kevin Lamkiewicz, Ramakanth Madhugiri, John Ziebuhr, Martin Hölzer, and Manja Marz (2019). "Direct RNA nanopore sequencing of full-length coronavirus genomes provides novel insights into structural variants and enables modification analysis". en. In: *Genome Res.*

Volden, Roger, Theron Palmer, Ashley Byrne, Charles Cole, Robert J Schmitz, Richard E Green, and Christopher Vollmers (2018). "Improving nanopore read accuracy with the R2C2 method enables the sequencing of highly multiplexed full-length single-cell cDNA". en. In: *Proc. Natl. Acad. Sci. U. S. A.* 115.39, pp. 9726–9731.

Walker, Bruce J, Thomas Abeel, Terrance Shea, Margaret Priest, Amr Abouelliel, Sharadha Sakthikumar, Christina A Cuomo, Qiandong Zeng, Jennifer Wortman, Sarah K Young, and Ashlee M Earl (2014). "Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement". en. In: *PLoS One* 9.11, e112963.

Wang, Jing, Yue Zhao, Xiaofan Zhou, Scott W Hiebert, Qi Liu, and Yu Shyr (2018). "Nascent RNA sequencing analysis provides insights into enhancer-mediated gene regulation". en. In: *BMC Genomics* 19.1, pp. 1–18.

Watson, J D and F H Crick (1953). "Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid". en. In: *Nature* 171.4356, pp. 737–738.

Wei, Ze-Gang and Shao-Wu Zhang (2018). "NPBSS: a new PacBio sequencing simulator for generating the continuous long reads with an empirical model". en. In: *BMC Bioinformatics* 19.1, p. 177.

Wick, Ryan R, Louise M Judd, and Kathryn E Holt (2019). "Performance of neural network basecalling tools for Oxford Nanopore sequencing". en. In: *Genome Biol.* 20.1, p. 129.

Wiley, G and M J Miller (2020). "A highly contiguous genome for the Golden-fronted Woodpecker (Melanerpes aurifrons) via a hybrid Oxford Nanopore and short read assembly". In: *bioRxiv*.

Workman, R E, A Tang, P S Tang, M Jain, J R Tyson, and others (2018). "Nanopore native RNA sequencing of a human poly (A) transcriptome". In: *BioRxiv*.

Wu, R (1972). “Nucleotide sequence analysis of DNA”. en. In: *Nat. New Biol.* 236.68, pp. 198–200.

Yeh, Li-Hsien, Mingkan Zhang, Sang W Joo, and Shizhi Qian (2012). *Slowing down DNA translocation through a nanopore by lowering fluid temperature*.

Zhao, Boxuan Simen, Ian A Roundtree, and Chuan He (2017). “Post-transcriptional gene regulation by mRNA modifications”. en. In: *Nat. Rev. Mol. Cell Biol.* 18.1, pp. 31–42.

Zimin, Aleksey V, Guillaume Marçais, Daniela Puiu, Michael Roberts, Steven L Salzberg, and James A Yorke (2013). “The MaSuRCA genome assembler”. en. In: *Bioinformatics* 29.21, pp. 2669–2677.

Zimin, Aleksey V and Steven L Salzberg (2019). “The genome polishing tool POLCA makes fast and accurate corrections in genome assemblies”. en.

Chapter 2

Nanopore native RNA sequencing of a human poly(A) transcriptome

This chapter is a published manuscript at the journal Nature Methods reprinted in compliance with the journal policies. I am a co-first author on this study responsible for the majority of downstream analysis. More specifically, analysis of allele-specific expression, poly(A) tail length, and base modification are the result of my work.

Rachael E Workman et al. (2019). "Nanopore native RNA sequencing of a human poly (A) transcriptome". In: *Nature methods* 16.12, pp. 1297–1305

2.1 Introduction

The roles of RNA in cell function are numerous and complex. Beyond the fundamental importance of mRNA, tRNA, and ribosomal RNA in translation, several classes of non-coding RNA (ncRNA) regulate cellular processes including division, differentiation, and programmed cell death (Cech and Steitz, 2014; Su et al., 2016).

Sequencing by synthesis (SBS) strategies have dominated RNA sequencing since the early 1990s. Typically this involves generation of cDNA templates by reverse transcription (RT) (Temin and Mizutani, 1970; Baltimore, 1970) coupled with PCR amplification (Saiki et al., 1988). Sequential base identification along template strands is generated by DNA polymerase-dependent incorporation of complementary nucleotides into daughter strands. A high throughput version of this basic technique (RNA-seq (Nagalakshmi et al., 2008; Wilhelm et al., 2008; Mortazavi et al., 2008; Lister et al., 2008; Cloonan et al., 2008; Marioni et al., 2008; Morin et al., 2008)) can be implemented to determine both reference-based and *de novo* transcriptomes at high coverage (Wang, Gerstein, and Snyder, 2009). Single molecule SBS strategies have been employed to sequence RNA without a PCR step using Pacific Biosciences and Helicos platforms. Read lengths were <25 nt (Vilfan et al., 2013) and <34 nt (Ozsolak et al., 2009) respectively.

Nanopore RNA strand sequencing has emerged as an alternative single molecule strategy (Garalde et al., 2018; Jenjaroenpun et al., 2018; Smith et al., 2019). It differs from SBS-based platforms in that native RNA nucleotides, rather than copied DNA nucleotides, are identified as they thread through and touch a nanoscale sensor. Nanopore RNA strand sequencing shares the core features of nanopore DNA sequencing, *i.e.* a processive helicase motor regulates movement of a bound polynucleotide driven through a protein pore by an applied voltage. As the polynucleotide advances through the pore in single nucleotide steps, ionic current impedance reports on the segment of bases that occupy a narrow reading head as a function of time. This series of

ionic current segments is converted into nucleotide sequence using an ONT algorithm trained with known RNA molecules.

Here we describe sequencing and analysis of a human poly(A) transcriptome from the GM12878 cell line using the Oxford Nanopore (ONT) platform. We demonstrate that long native RNA reads allow for discovery and characterization of RNA isoforms that are difficult to observe using short read cDNA methods (Steijger et al., 2013; Venturini et al., 2018). Because native RNA strands are directly read by nanopores, nucleotide modifications and 3′ poly(A) tail lengths can be determined from the ionic current signal without additional processing steps. Data and resources are posted online at:

(`https://github.com/nanopore-wgs-consortium/NA12878/blob/master/RNA.md`).

2.2 RNA preparation, nanopore sequencing, and computational pipeline

The protocol we used to isolate and sequence native poly(A) RNA from a human B-lymphocyte cell line (GM12878) is summarized in Figure 2.1a and detailed in methods. Briefly, 750 ng of poly(A) RNA was adapted for nanopore sequencing using ONT protocols and library reagents. Adapted poly(A) RNA (160-400 ng) was then added to the MinION flow cell and sequenced for ∼24-48 hours. A typical ionic current trace during *TP53* mRNA translocation through a nanopore is shown in Figure 2.1b. The ionic current readout for each poly(A) RNA strand was basecalled using Albacore version 2.1.0 (ONT).

We also performed nanopore cDNA sequencing using the identical GM12878

RNA sample and analysis pipeline, but with modified parameters appropriate for cDNA sequencing (methods). Both the RNA and cDNA data were archived and used for downstream analyses (Figure 2.1c). They are available on GitHub at:

(`https://github.com/nanopore-wgs-consortium/NA12878/blob/master/RNA.md`).

2.3 Native poly(A) RNA sequencing statistics

Each of six laboratories performed five nanopore sequencing runs. Together, these thirty runs produced 13.0 million poly(A) RNA strand reads, of which 10.3 million qualified as pass reads (Albacore generated Q-value > 7). Throughput varied between 50K and 831K pass poly(A) reads per flow cell. The 10.3 million pass RNA nanopore reads had an N50 length of 1,334 bases, and a median length of 771 bases. Of these, 9.9 million aligned to the GRCh38 human genome reference sequence using minimap2 version 2.1 with a splice-aware setting (-ax splice -uf -k14) (Li, 2018). This algorithm was chosen because it aligns nanopore reads to exons in the human genome while spanning across introns (Tang et al., 2018). The 360,000 unaligned pass reads had a median read length of 211 bases, suggesting that shorter nanopore reads were more difficult to align.

We next aligned the RNA pass reads to the GENCODE v27 transcriptome reference using minimap2. This resulted in aligned reads ranging in length from 85 nt (a fragment of an mRNA encoding Ribosomal Protein RPL39), to 21kb (an mRNA encoding spectrin repeat containing nuclear envelope protein

2 (*SYNE2*)). A comprehensive list of the genes and isoforms represented among the aligned native RNA reads can be found on GitHub.

MarginStats (version 0.1) (Jain et al., 2015a) was employed to calculate percent identity and the number of matches, mismatches, and indels per aligned read in this population. The median identity across flow cells and laboratories averaged 86+/-0.86% (Figure 2.2a), with mismatch, insertion, and deletion errors of 2.4%, 4.3%, and 4.4% respectively. The basecaller seldom confused G-for-C or C-for-G (0.38% and 0.47% errors respectively); C-to-T and T-to-C errors were substantially higher (3.62% and 2.23% respectively) (Figure 2.2b). We compared the observed read length *vs* expected transcript length as defined by GENCODE v27, and found general agreement (Figure 2.2c). The discrete clusters below the diagonal represent incorrect assignments to GENCODE isoforms, and the diffuse shading represents fragmented RNA (see text below concerning RNA truncation).

For nanopore cDNA data, we observed a median identity of 85% (Figure 2.2d) which is comparable to recent published nanopore DNA results (Jain et al., 2018). The substitution error patterns for cDNA data were similar to those for native RNA data (Figure 2.2e). There was a weaker correspondence between observed vs expected read lengths for cDNA (Figure 2.2f).

2.4 Kmer coverage

Previous Kmer analyses indicated that some nucleotide sequences are over- or under-represented in nanopore-based DNA sequence reads (Jain et al., 2015a; Jain et al., 2018). In this study, we assessed nanopore RNA and cDNA 5-mer

coverage using reads aligned to GENCODE v27 isoforms. Only reads that covered 90% or more of a given reference sequence were chosen. Of the 10.3 million RNA pass reads, 2.9 million RNA reads were selected based on this criterion. Of the 15.1 million pass cDNA reads, 3.9 million pass cDNA reads were selected. These reads included all 1024 possible 5-mers.

The largest deviation from expectation often occurred for homopolymer-rich 5-mers that were under-represented in native RNA and over-represented in cDNA. This is similar to previous 5-mer analysis for ONT MinION DNA sequence data (Jain et al., 2015a; Jain et al., 2018).

2.5 Nanopore sequencing performance assessed using mitochondrially-encoded RNA

We reasoned that mitochondrial poly(A) transcripts could be used to benchmark nanopore sequencing performance because they are abundant in all human cells, they are single exon, and they vary substantially in length (349-2,379 nt). Of the 9.9 million aligned nanopore poly(A) RNA strand reads, approximately 10% (950,879) aligned to the mitochondrial genome (Figure 2.3a and public UCSC track: `https://goo.gl/erWFyu`). As expected, most of these poly(A) transcripts corresponded to mitochondrial ribosomal RNA or to mitochondrial mRNA. Overall, the nanopore RNA reads recapitulated known features of the human MT-transcriptome.

MT-RNA read length analysis was revealing. Figure 3b shows 5,000 reads that aligned to *MT-CO2* or to *MT-ND4L/ND4* genes. In each panel, a dominant band corresponded closely to the expected transcript length (732 nt and

1,673 nt for *MT-CO2* and *MT-ND4L/ND4* respectively). However, for each of these, a population of truncated reads was randomly distributed between the dominant band and about 300 nt in length. When we quantified the fraction of truncated reads as a function of nominal transcript length for ten MT-mRNA of the heavy strand (Online Methods), we found a strong linear anti-correlation in most cases (Figure 2.3c). The single outlier was *MT-ND5* which is the mitochondrial transcript with a 568 nt 3′ UTR.

These MT-poly(A) RNA truncations could occur at any of several non-biological steps during the sequencing process, or they could arise from regulated enzymatic degradation in the mitochondrion (Szczesny et al., 2012). Here we considered three possible non-biological causes that were specific to the nanopore platform.

One systematic cause of read truncations occurred because the enzyme that controls translocation through the pore is 10-15 nt from the nanopore sensor. Thus, when the enzyme releases the last base at the 5′ end, the strand is rapidly driven through the pore at microseconds per nucleotide which prevents reading the terminal 10-15 nt. This phenomenon was evident by close inspection of read coverage at the 5′ end of mitochondrial mRNA transcripts (`https://goo.gl/erWFyu`), and is expected for all direct RNA reads in the present ONT protocol.

Another possible cause was ionic current signal artifacts associated with enzyme stalls during RNA translocation, or with extraneous voltage spikes. Similar artifacts have been shown to disrupt strand reads during MinION sequencing of DNA (Payne et al., 2018). Systematic analysis of 2,729 *MT-CO1*

reads within bulk FAST5 files from Lab 1 identified 527 reads which started or ended abnormally (methods). By including ionic current segments that were identified before or after many of these truncations, we reconstructed 300 reads with longer alignments to *MT-CO1* (Figure 2.3d). As anticipated, this phenomenon was length dependent (Figure 2.3e), ranging from 4.2% of reads with rescued segments for ND3 (346 nt nominal length) to 17.6% for ND5 (2379 nt nominal length).

A third possible cause was strand breaks during nanopore sequencing runs. As a test, we analyzed *MT-CO1* read-length distribution for each of the six laboratories as a function of time on ONT flow cells. As expected, we found that the read frequency at all lengths declined steadily over 36 hours, however the full-length fraction declined by only 5% (Figure 2.3f). This analysis also revealed that initial input RNA quality differed substantially between laboratories. For Lab 1 (representative of Labs 1-5) the full length *MT-CO1* fraction was 49.7% for 0-6 hours on the sequencer (Figure 2.3f, left); by comparison, for Lab 6 the full length *MT-CO1* fraction was 31.1% (Figure 2.3f, right). A possible cause is that Lab 6 used a separate GM12878 cell pellet. For this reason, isoform-level analyses in the following sections omitted Lab 6 data, and focused on 8.17 million aligned poly(A) RNA reads from Labs 1-5.

2.6 Isoform detection and analysis

Long nanopore reads could improve resolution of RNA exon-exon connectivity, allowing for discovery of unannotated RNA isoforms. However, these

reads averaged 14% per-read base call errors, confounding precise determination of splice sites. Also, biological RNA processing and *in vitro* 5′-end truncations (see above) can make it difficult to define transcription start sites (TSS).

To overcome these limitations we employed FLAIR (Full-Length Alternative Isoform Analysis of RNA). In this strategy, we first replaced any nanopore-based splice sites bearing apparent sequencing errors with splice sites supported by GENCODE v27 annotations or by Illumina GM12878 cDNA data (Tilgner et al., 2014; Cho et al., 2014). Second, to overcome TSS uncertainty caused by truncated RNA reads, we considered only reads with 5′ ends proximal to promoter regions (defined by ENCODE promoter chromatin states for the GM12878 cell line (Bernstein et al., 2005; Ernst and Kellis, 2010; Ernst et al., 2011)). And third, we used FLAIR to group reads into isoforms according to chains of splice junctions.

We then compiled two FLAIR isoform sets using different supporting read criteria (see methods):

i) A FLAIR-sensitive set that included isoforms with three or more uniquely mapped reads (see GitHub link). This large set could be useful for isoform discovery, at the risk of false positives.

ii) A FLAIR-stringent set that was compiled by filtering set (i) for isoforms having three or more supporting reads that spanned $\geq$ 80% of the isoform with $\geq$ 25 nt coverage into the first and last exon.

We used the FLAIR-stringent dataset to screen for unannotated isoforms because it had the most strict criteria for isoform assignment. This set was

composed of 33,984 isoforms from 10,793 genes. Over half (52.6%) of the isoforms had a splice junction chain unannotated in GENCODE v27 (13.0% of total assigned reads). Of the unannotated isoforms, 46.5% (7,961) had new combinations of annotated splice junctions, 13.3% (2,281) had retained introns, and 6.9% (1,180) had an unannotated exon. Figure 4a shows an example set of lncRNA isoforms arising from an unannotated transcription start site with multiple splice variants. We performed the same analysis using the FLAIR-sensitive set.

To better characterize long non-coding RNAs (lncRNAs), we then segregated the FLAIR-stringent isoforms into three categories (methods): i) lncRNAs that lacked an annotated start codon; ii) isoforms from protein-coding genes with premature termination codons upstream of the last splice junction; and iii) known protein-coding isoforms. Non-coding genes had more complex splicing patterns per gene than did coding genes, as measured by Shannon entropy (Figure 2.4b). This is consistent with prior studies that demonstrated increased alternative splicing in non-coding exons (Deveson et al., 2018; Gonzàlez-Porta et al., 2013).

As a conservative alternative to FLAIR, we compiled two GENCODE-based isoform sets:

iii) A GENCODE-sensitive set that included isoforms with one or more reads that mapped uniquely to GENCODE v27. We implemented a lower coverage threshold than we did for FLAIR because GENCODE is carefully curated.

iv) A GENCODE-stringent set that was compiled by filtering set (iii) for

isoforms having one or more supporting reads that spanned $\geq$ 80% of the isoform with $\geq$ 25 nt coverage into the first and last exon.

To estimate the sequencing depth required to completely characterize the GM12878 transcriptome, we plotted the number of isoforms detected in the GENCODE-sensitive and FLAIR-stringent isoform sets versus the number of subsampled reads in 10% increments. We then fitted a hyperbolic function to the data (Figure 2.4c). It is evident that the curves did not saturate and that additional reads would be required to capture a complete GM12878 transcriptome.

2.7 Assignment of transcripts to parental alleles

Allele-specific expression (ASE) is the preferential transcription of RNA from the paternal or maternal copy of a gene.. Although the importance of this phenomenon has been characterized (Baralle and Giudice, 2017), the consequences are not fully understood. This is partly due to technical limitations of haplotype identification using short read sequencing technologies.

We reasoned that the long nanopore RNA reads would be easier to assign to the parental allele of origin due to the greater chance of encountering a heterozygous SNP. Reads with at least two heterozygous SNPs were assigned to the parental allele of origin using HapCUT2 (Edge, Bafna, and Bansal, 2017). To discover the most possible genes, we used the FLAIR-sensitive dataset. In it, we found 3,751 genes with at least 10 haplotype informative reads. 3,707 of these genes were from autosomal chromosomes and 44 were from the X-chromosome. Among autosomal genes, 228 (6.1%) showed significant

ASE (binomial test, p<0.001), and among X-chromosome genes, 23 (95.7%) showed significant ASE (binomial test, p<0.001). X-chromosome expression was biased, with 22/23 allele-specific X-linked genes originating from the maternal allele, consistent with previous results for this cell line (Rozowsky et al., 2011). The sole paternally expressed X-linked locus encoded the lncRNA XIST, which is transcribed from the inactive X-chromosome and recruits epigenetic silencing machinery for X-inactivation in females (Brown et al., 1991). The remaining genes were expressed equally from both parental alleles.

We combined these allele-specific reads with isoforms from the FLAIR-sensitive set to mine for allele-specificity (methods). We identified 5 genes with one isoform expressed from one allele and another isoform expressed from the other allele (binomial test, P<0.001). One of these genes, *IFIH1*, had a paternal isoform with exon 8 retained, while the maternal isoform did not retain exon 8 (Figure 2.4d). We note that the closest SNV used in allele-assignment was 886 nt away from the alternative splicing event in this transcript. This would be undetectable using short read sequencing.

2.8 poly(A) analysis

Transcript poly(A) tails are thought to play a role in post-transcriptional regulation, including mRNA stability and translational efficiency (Eckmann, Rammelt, and Wahle, 2011; Preiss, 2013). However, these homopolymers can be several hundred nucleotides long making them difficult to measure using short-read SBS data (Subtelny et al., 2014; Chang et al., 2014).

In this study, we measured poly(A) tail lengths directly using a low variance the ionic current signal associated with the 3′ end of each poly(A) strand (Figure 1b, iii). Briefly, we developed a computational method ('nanopolish-polya', https://github.com/jts/nanopolish) to segment this signal and estimate how many ionic current samples were drawn from the poly(A) tail region. Then, by correcting for the rate at which the RNA molecule passes through the pore, nanopolish-polya estimates the length of the poly(A) tail.

To test this method we obtained six MinION-derived poly(A) RNA control datasets generated by ONT (ENA accession PRJEB28423). These datasets consisted of ionic current traces for synthetic *S. cerevisiae* enolase transcripts appended with 3′ poly(A) tails of 10, 15, 30, 60, 80 or 100 nucleotides. A second version of the 60nt poly(A) tailed construct (60nt-kN) contained a 10nt randomer between the enolase sequence and the 3′ poly(A).

Poly(A) tail length estimates for these synthetic controls are shown in Figure 2.5a. Median estimates fell within 4 nucleotides of the expected tail length for the 10-to-80 poly(A) datasets; for the 100nt dataset, the median estimate was 109nt. We observed that 66%-80% of the estimated lengths fell within 2 median absolute deviations of the expected tail length. The predicted tail length distribution for the 60nt-kN dataset (bearing the 10nt random sequence insert) contained a higher proportion of short poly(A) tails than expected, which may indicate amplification errors specific to this sample due to the 10nt sequence insert.

A limitation of our approach is the inability to detect when the poly(A) region stalls in the nanopore sensor during translocation, causing over-estimation

of the tail length. From the control data, we estimate this occurs for 1-3% of the sequenced molecules. Also, as the length of the poly(A) tail increases, the variance of our estimator does as well. This is expected because the number of ionic current samples in the tail region for a fixed expected tail length is not deterministic and has substantial variation due to the kinetics of translocation. We found that we were able to offset some of this inherent variance by using the overall transcript translocation rate as an estimator of the poly(A) rate.

We applied this poly(A) length estimator to the complete GM12878 native poly(A) RNA sequence dataset. Overall, the poly(A) length distribution centered at $\sim$50nt, with a broad dispersion of longer poly(A) tails for some transcripts. When we segregated mitochondrial-encoded transcripts from nuclear-encoded transcripts, we found that the mitochondrial transcripts had poly(A) lengths which peaked at 52nt, with a mean of 59nt and almost no poly(A) tail lengths greater than 100nt (Figure 2.5b). This is consistent with results for mitochondrial poly(A) RNA from other human cell lines (Temperley et al., 2010). Conversely, nuclear transcripts showed a broader length distribution, with a peak at 58nt, a mean of 112nt, and a large number of poly (A) tails greater than 200nt.

Next, we measured poly(A) tail length differences between genes with at least 500 reads. Figure 2.5c shows the distribution of poly(A) size for genes found to have the 2 longest poly(A) tails, genes with the two shortest poly(A) tails, and the gene with the median poly(A) length.

For some genes, e.g. the RNA-binding protein DDX5, multiple poly(A) length peaks were observed (Figure 2.5c), suggesting the presence of poly(A)

tail-length sub-populations that are isoform specific. To explore this, we analyzed genes in the Gencode-stringent dataset, and found 215 genes that had isoforms with significantly different poly(A) lengths.

When we compared two Gencode isoforms of *DDX5*, we noted that an intron-retaining isoform (ENST00000581230, '230') had a median poly(A) tail length of 327nt, compared with the protein-coding isoform (ENST00000225792,'792'), which had a median poly(A) tail length of 125nt. (Figure 2.5d). This difference motivated us to explore the relationship between poly(A) length and RNA intron-retention. For this analysis, we classified each isoform in Gencode-sensitive as either protein-coding or intron-retaining. We observed that a subset of transcripts with retained introns tended to have longer poly(A) tails (median 232nt) than did transcripts without introns (median 91nt) (t-test p-value < 2.2e-16, Figure 5e). This result is consistent with a previous observation that nuclear transcripts with retained introns tend to have longer poly(A) tails, priming them for degradation through recognition by the nuclear poly(A) binding protein (PABPN1) (Bresson et al., 2015).

2.9 Modification detection

Nanopore sequencing has been used to identify base modifications in DNA (Simpson et al., 2017; Rand et al., 2017) and RNA (Garalde et al., 2018; Smith et al., 2019). N6-methyladenine (m6A) is the most common internal modification on mRNA (Liu and Pan, 2016), and has been implicated in many facets of RNA metabolism (Dai et al., 2018). m6A dysregulation has been linked to human diseases, including obesity and cancer (Sibbritt, Patel, and Preiss,

2013). Because m6A modifications are enriched in 3' UTRs, with two-thirds of these containing miRNA sites (Meyer et al., 2012), the impact of this modification appears to be largely regulatory, as opposed to altering protein coding sequence.

Prior work has documented that base modifications result in changes in ionic current distributions for a given kmer (Rand et al., 2017; Simpson et al., 2017). We focused our studies on the GGACU binding motif of METTL3, a subunit of the m6A methyltransferase complex (Roost et al., 2015). As an example, we compared the raw current signal at a putative m6A site (chr19:3976327) for eukaryotic elongation factor 2 (*EEF2)* RNA versus the signal for an *in vitro* transcribed copy produced from GM12878 mRNA (methods). This comparison revealed an ionic current change attributable to m6A (Figure 2.6a). To validate this result, we used synthetic oligomers that were identical except for the presence or absence of m6A within the GGACU motif (Figure 2.6b). After sequencing, we used nanopolish eventalign to extract current levels for the 5-mers which contained the modified base and the unmodified base. This revealed a clear current difference (Figure 2.6c) consistent with the *EEF2* result.

To determine if m6A modifications differed between isoforms of the same gene, we screened Gencode-sensitive isoforms for ionic current changes at the GGACU motif. We found 86 genes (198 isoforms) where the median current levels at a single GGACU were significantly different between gene isoforms (Kruskal-Wallis, Student's t-test, and Kolmogorov-Smirnov statistical testing with Bonferroni multiple testing correction). An example is illustrated for the

SNHG8 gene (Figure 2.6d).

Another post-transcriptional modification, A-to-I RNA editing (Licht et al., 2016), commonly occurs in introns, UTRs, and Alu elements. It plays a role in splicing and regulating innate immunity (Nishikura, 2010; Tajaddod, Jantsch, and Licht, 2016) and is associated with numerous diseases (Gallo et al., 2017). NGS detects A-to-I editing as a nucleotide variant in cDNA sequences (A-to-G).

Previous nanopore experiments documented the presence of systematic base miscalls in regions of *E. coli* 16S rRNA bearing modified RNA bases (Smith et al., 2019). Consistent with this, we found systematic base miscalls at putative inosine bearing positions in the GM12878 aryl hydrocarbon receptor (*AHR*) data (chr7:17,345,148-17,345,157). To cross-validate this result, we compared our cDNA sequence data relative to the GM12878 reference and found that putative inosines were detected as a base change (A-to-G) (single inosine for the CUACU 5-mer, and multiple inosines for the AAAAA 5-mer).

The ionic current distribution for the putative single inosine 5-mer (CUACU) was modestly different from the canonical 5-mer (Figure 6e). The ionic current distribution for the inosine containing AAAAA 5-mer was more complex, possibly reflecting the presence of multiple inosines (Figure 2.6f).

2.10 Discussion

Nanopore RNA sequencing has two useful features: 1) The sequence composition of each strand is read as it existed in the cell. This permits direct detection of post-transcriptional modifications including nucleotide alterations and

polyadenylation; 2) reads can be continuous over many thousands of nucleotides providing splice-variant and haplotype phasing. Although each of these features is useful in itself, the combination is unique and likely to provide new insights into RNA biology. The two principal drawbacks of the present ONT nanopore RNA sequencing platform is the relatively high error rate (compared to Illumina cDNA sequencing), and uncertainty about the 5' end of the transcript.

We were concerned that read fragmentation in the nanopore data was caused by RNA degradation on the nanopore flow cells during sequencing. However, we found minimal ($\sim$5%) reduction in the full-length fraction of a 1.6 kb mRNA (*MT-CO1*) over 36 hours. Preliminary analysis indicated that read truncations were more often caused by electronic signal noise due to current spikes of unknown origin. We showed that meaningful biological signals can be recovered from bulk Fast5 files around these truncations, suggesting that future improvements to the MinKNOW read segmentation pipeline are needed.

When combined with more accurate short Illumina reads, long nanopore reads allowed for end-to-end documentation of RNA transcripts bearing numerous splice junctions, which would not be possible using either platform alone. We documented a high proportion (52.6%) of unannotated isoforms, similar to other long-read transcriptome sequencing studies (e.g., 35.6% and 49%) (Tardaguila et al., 2018; Anvar et al., 2018). While many of these unannotated isoforms are low abundance and their protein coding potential unknown, it is important to catalog them because subtle splicing changes can impact

function (Wang et al., 2016; Bradley et al., 2012). We also note that the number of detected isoforms did not saturate using the nanopore poly(A) RNA dataset, indicating that greater sequence depth will be necessary to give a comprehensive picture of the GM12878 poly(A) transcriptome.

A variety of techniques have been used to examine allele-specific expression (ASE) (Rozowsky et al., 2011; Turro et al., 2011; Pandey et al., 2013; Mayba et al., 2014; Skelly et al., 2011; Tilgner et al., 2014; Deonovic et al., 2017). However, identification of ASE is limited using short read platforms because heterozygous variants are rare within any given window of a few hundred nucleotides. Our nanopore approach for ASE discovery has the advantage of long reads, but the disadvantage of high base call errors. We attempted to mitigate the effects of these errors by requiring multiple heterozygous variants and a stringent false-discovery rate (FDR) during ASE analysis. Therefore, the number of genes that we report as demonstrating ASE (167) is likely an underestimation. We report nearly exclusive use of the maternal X-chromosome, with the only paternal transcripts originating from the XIST locus, consistent with previous findings (Rozowsky et al., 2011). Importantly, we have shown that nanopore sequencing enables allele-specific isoform studies, especially in cases where the splicing variation does not have a heterozygous variant within range of conventional short-read sequencing.

Polyadenylation of RNA 3' ends regulates RNA stability and translation efficiency by modulating RNA-protein binding and RNA structure (Eckmann, Rammelt, and Wahle, 2011). However, transcriptome-wide poly(A) analysis has been difficult due to basecalling and dephasing errors (Chang et al., 2014).

Recently implemented modifications to the Illumina strategy address these limitations (Chang et al., 2014; Subtelny et al., 2014; Woo et al., 2018); but can not resolve distal relationships, such as between splicing and poly(A) length. Nanopore poly(A) tail length estimation using nanopolish-polya offers the advantages of both direct length assessment and maintenance of information about isoform and modification status per transcript. Our preliminary studies revealed differences in poly(A) length distribution between mitochondrial and nuclear genes, between different nuclear genes, and between different isoforms of the same gene. We note in particular an increase in poly(A) tail length for some intron-retaining isoforms. This is consistent with previous work showing that hyper-adenylation targets intron-retaining nuclear transcripts for degradation through recognition by a poly(A)-binding protein (PABPN1) (Bresson et al., 2015).Additionally, deadenylation of cytoplasmic transcripts is a core part of the RNA degradation pathway (Yi et al., 2018), suggesting that time course experiments investigating RNA decay kinetics (Parker and Song, 2004) could be possible with this technology.

We have demonstrated detection of N6-methyladenosine and inosine modifications in human poly(A) RNA. This validates prior work which showed modification-dependent ionic current shifts associated with m6A (*S. cerevisiae*) (Garalde et al., 2018), pseudouridine (Smith et al., 2019), and m7G (*E. coli*) (Smith et al., 2019). Differences in m6A modification level proved to be discernible at the isoform level for human *SNGH8* mRNA (Figure 2.6d), thus documenting splicing variation and modification changes simultaneously.

Although other methods exist for high throughput analysis of RNA modifications (Li, Xiong, and Yi, 2016), they often require enrichment which limits quantification, and they are usually short-read based. The latter precludes analysis of long-distance interactions between modifications, and between modifications and other RNA features such as splicing and poly(A) tail length. The capacity to detect these long-range interactions is likely to be important given recent work suggesting links between RNA modifications, splicing regulation, and RNA transport and lifetime (Roundtree et al., 2017; Lee, Kim, and Kim, 2014). We argue that nanopore native RNA sequencing could deliver this long-range information for entire transcriptomes. However, this will require algorithms trained on large, cross-validated datasets as has been accomplished for cytosine and adenine methylation in genomic DNA (Simpson et al., 2017; Rand et al., 2017).

2.11 Conclusions

Oxford Nanopore devices sequence long native RNA strands directly. In this study, we showed that these long reads improved human poly(A) RNA isoform characterization, including allele specificity. Because native RNA strands were read directly, m6A and inosine nucleotide modifications could be detected without intermediate preparative steps. We introduced a new tool (nanopolish-polya) that estimates 3′ poly(A) tails on individual RNA strands based on nanopore ionic current signals. Applied to the GM12878 transcriptome, it revealed differences in RNA poly(A) tail lengths between nuclear and mitochondrially encoded genes, and between splice variants of

genes.

2.12 Data Availability

Sequence data including raw signal files (FAST5), event-level data (FAST5), base-calls (FASTQ) and alignments (BAM) are available as an Amazon Web Services Open Data set for download from `https://github.com/nanopore-wgs-consortium/NA12878`. The scripts used for various analyses are also available from the same GitHub under nanopore-human-transcriptome/scripts.

2.13 Figures

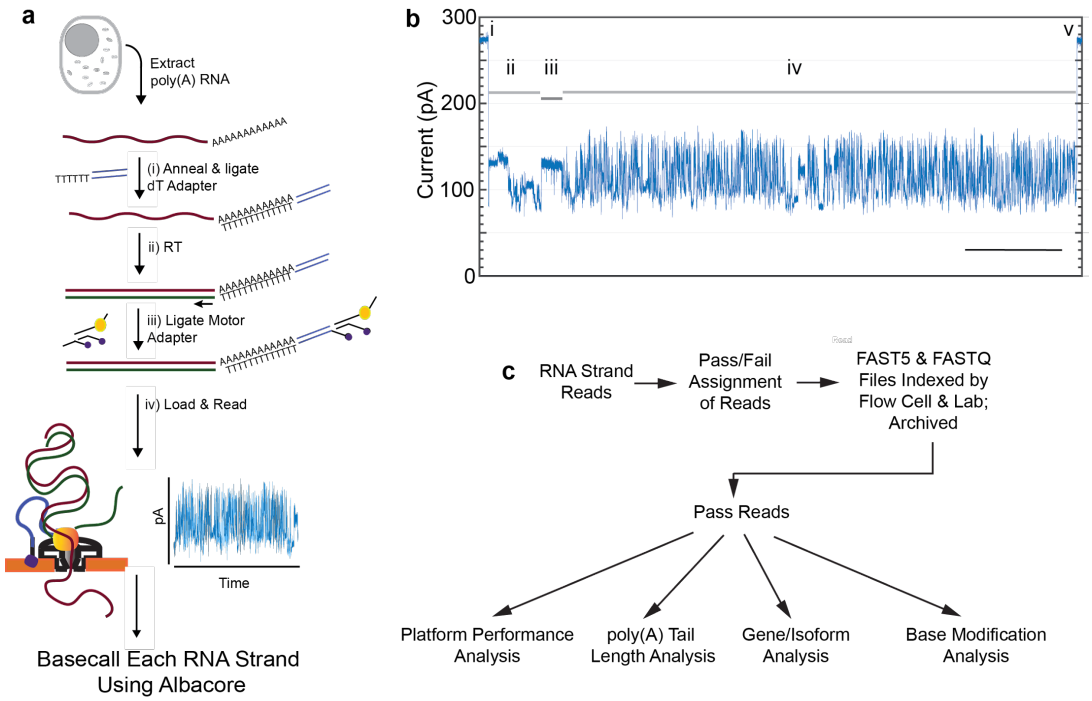

Figure 2.1: *(a) RNA is isolated from cells followed by poly(A) selection using poly(dT) beads. Poly(A) RNA is then prepared for nanopore sequencing using the following steps: (i) A duplex adapter bearing a poly(dT) overhang is annealed to the RNA poly(A) tail, followed by ligation of the strand abutting the poly(A) tail; ii) the poly(dT) complement is extended by reverse transcription; iii) a proprietary ONT adapter bearing a motor enzyme is ligated to the first adapter; and (iv) the product is loaded onto the ONT flow cell for reading by ionic current impedance. The ionic current trace for each poly(A) RNA strand is base called using a proprietary ONT algorithm (Albacore). (b) A representative ionic current trace for a 2.3 kb TP3 transcript ionic current components: (i) Strand capture; ii) ONT adapter translocation; iii) poly(A) RNA tail translocation; iv) mRNA translocation; and (v) exit of the strand into the trans compartment. Bar is 5 seconds. (c) Processing of the RNA strand reads in silico, followed by data analysis.*

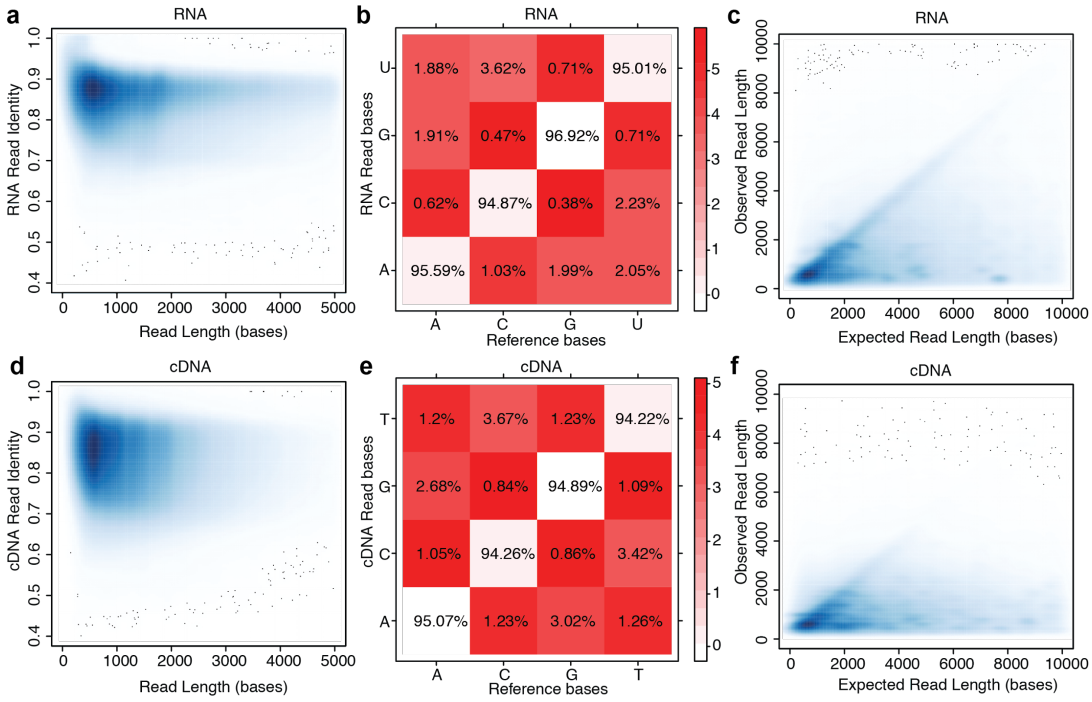

Figure 2.2: *(a) Alignment identity vs. read length for native RNA reads. (b) Substitution matrix for native RNA reads. (c) Observed v. expected read length for ∼9.7 million native RNA reads. (d) Alignment identity vs. read length for cDNA reads. (e) Substitution matrix for cDNA reads. (f) Observed vs. expected read length for ∼14.1 million cDNA reads.*

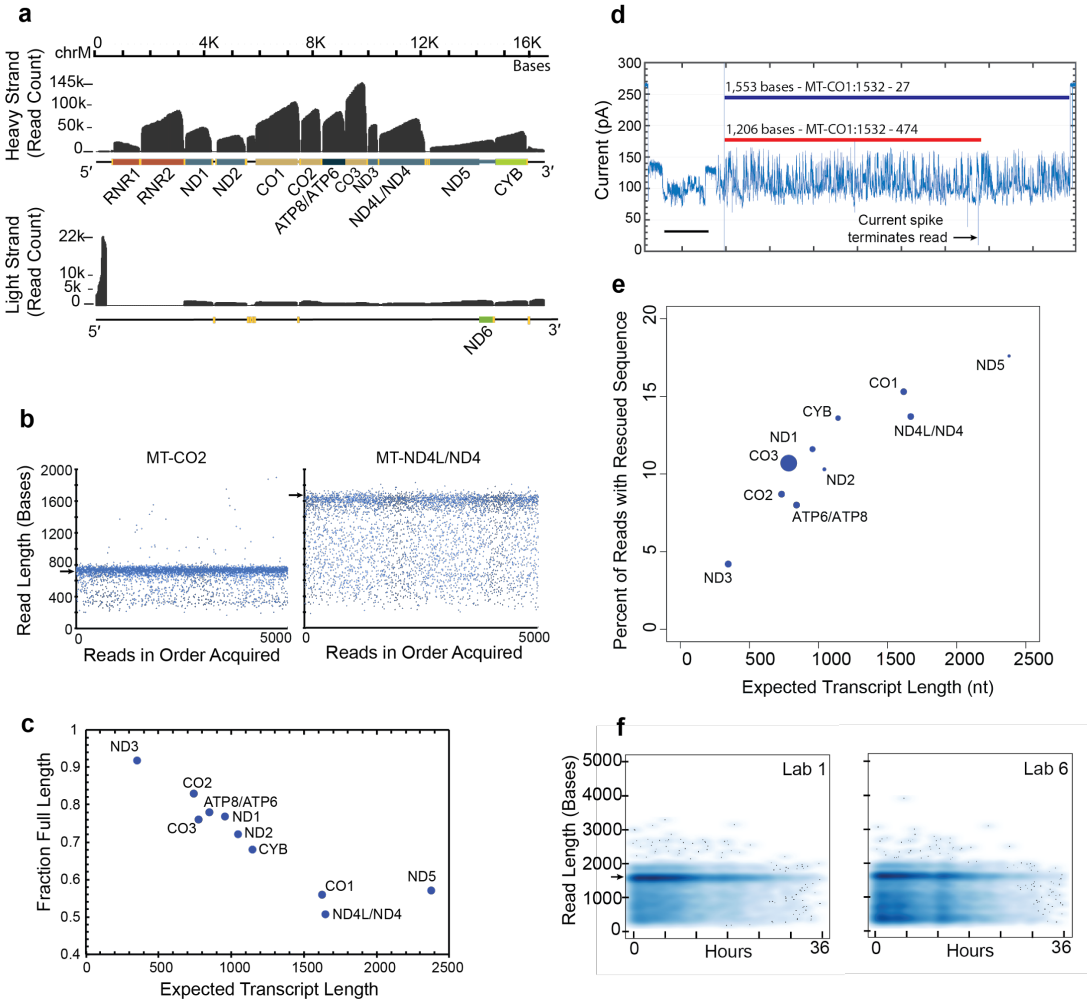

Figure 2.3: *(a) Read coverage of the H strand (top) and the L strand (bottom). (b) Distribution of nanopore read lengths for MT-CO2 and MT-ND4L/ND4 transcripts. Each point represents one of approximately 5000 reads in the order acquired from a single Lab 1 MinION experiment. Horizontal arrows are expected transcript read lengths. (c) Relationship between expected transcript read length and fraction of nanopore poly(A) RNA reads that were full length. Each point is for a protein coding transcript on the H strand. Labels are for mitochondrial genes without the MT prefix. See Online Methods for definition of 'Full Length'. (d) Ionic current trace for translocation of a MT-CO1 transcript. It is representative of traces where the read was artificially truncated by a signal anomaly. The red line represents the MinKNOW segmented read (positions 474-1532 of the MT-CO1 gene), and the blue line represents the manually segmented and rescued read (positions 27-1532 of the MT-CO1 gene). The time bar is two seconds. (e) Percent of artificially truncated strand reads where sequence was recovered from the ionic current signal. Points are for protein coding transcripts as in panel c. (f) MT-CO1 poly(A) transcript read length vs MinION run time.*

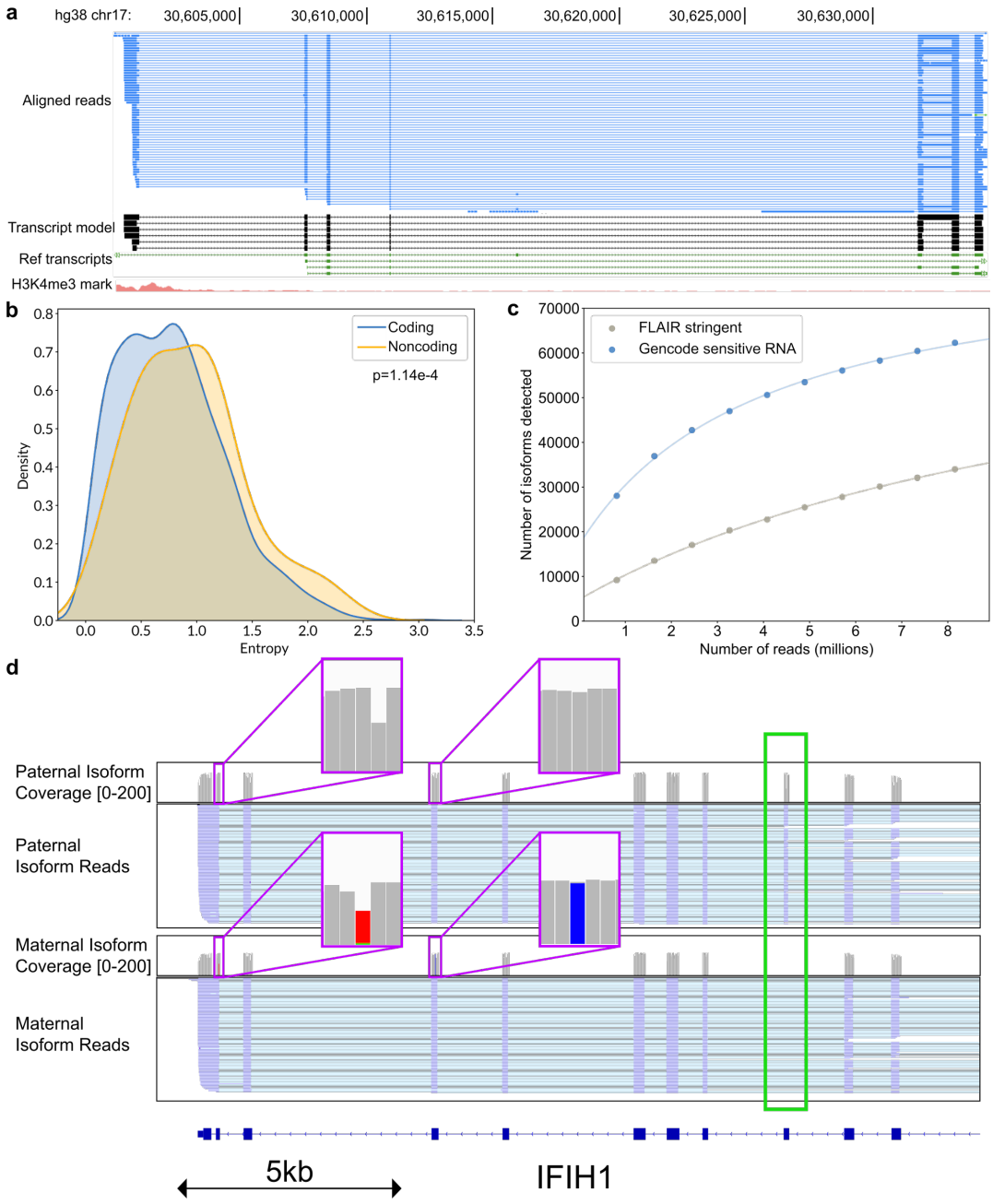

Figure 2.4: *(a) Genome browser view of unannotated isoforms found in the native RNA data. (b) Distributions of the Shannon entropy of isoform expression for coding versus noncoding genes detected by FLAIR. (c) Saturation plot showing the number of isoforms discovered (y-axis) in relation to the different numbers of reads (x-axis) of total native RNA and cDNA data used. (d) IGV view of the allele-specific isoforms of the gene IFIH1. Purple boxes (inset) indicate location of SNPs used to assign allele specificity, alternatively spliced exon is indicated with a green box.*

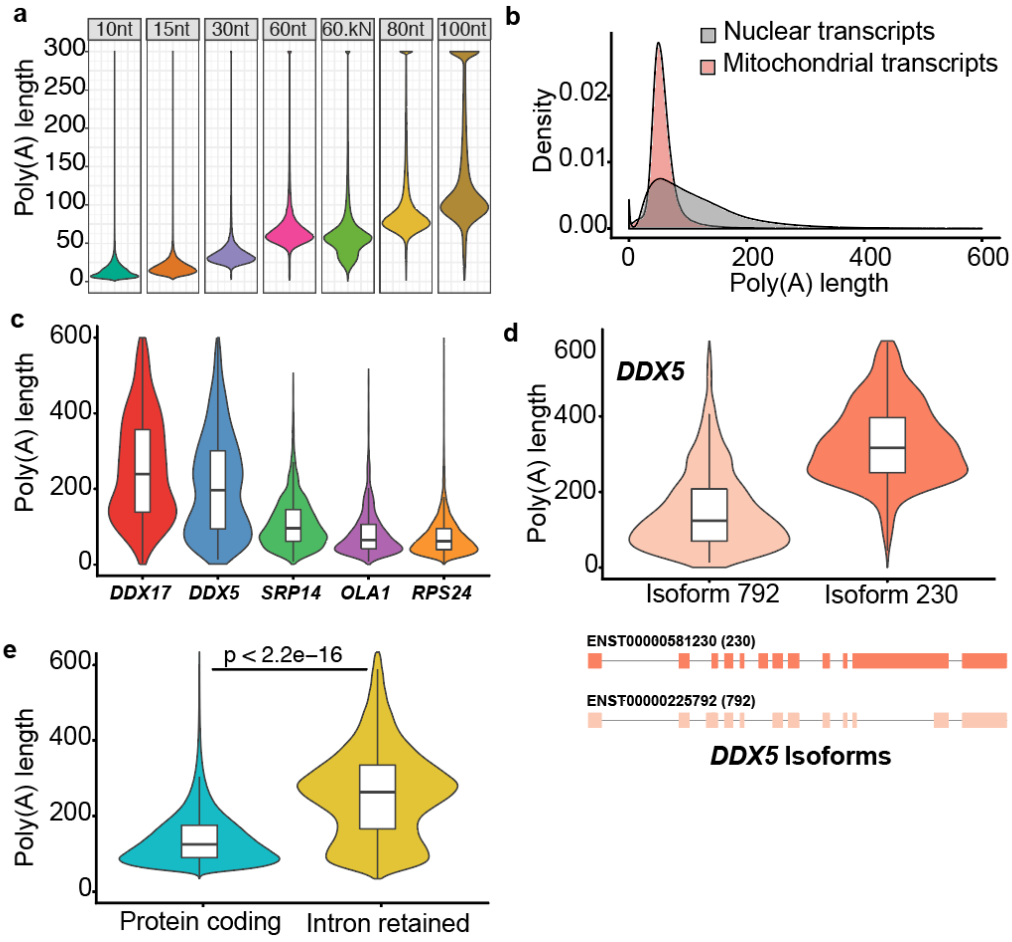

Figure 2.5: *(a) Estimate of poly(A) lengths for a synthetic enolase control transcript bearing 3' poly(A) tails of 10, 15, 30, 60, 80 or 100 nucleotides. '60kN' contained all possible combinations of a 10nt random sequence inserted between the enolase sequence and the 3' poly(A) 60mer. (b) Poly(A) length distributions for transcripts encoded in the mitochondrial genome versus nuclear-encoded genes. (c) Violin plots showing the range of poly(A) tail lengths sequenced, with the longest (DDX5, DDX17), shortest (RPS24, OLA1), and average (SRP14) poly(A) distributions plotted. (d) Distribution of poly(A) tail lengths and gene models for two isoforms of DDX5 plotted. (e) Distribution of poly(A) tail lengths for intron-retaining and intron-free transcripts identified using Gencode-Sensitive isoform set, Kruskal-Wallis p-value denoted.*

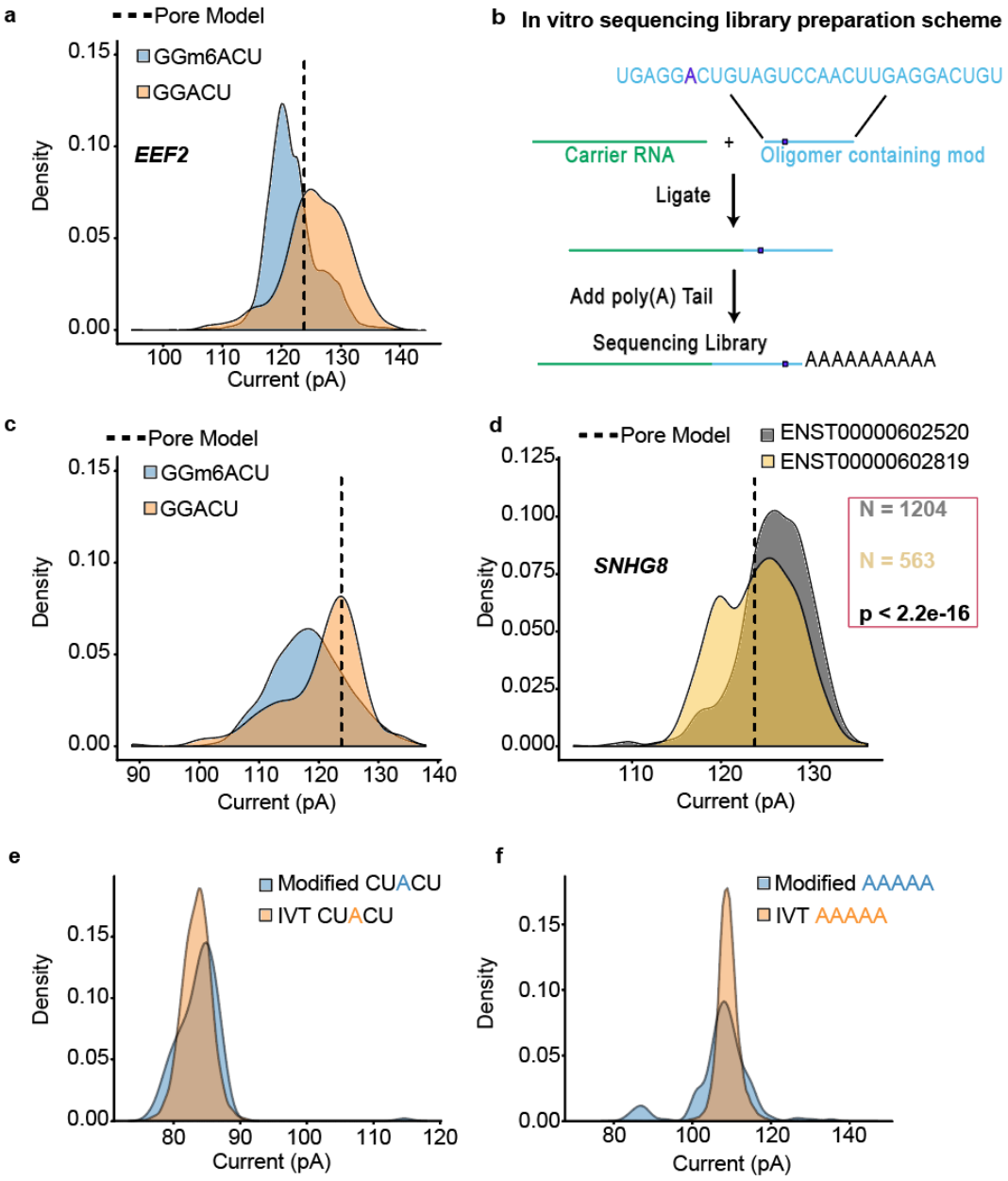

Figure 2.6: *(a) Comparing current signal from m6A-modified and unmodified GGACU motifs. (b) Schematic for the oligomer-ligation preparation. (c) Comparing current signal from m6A-modified and unmodified GGACU motifs. (d) Current distributions for GGACU motifs within SNHG8 gene isoforms. (e) Ionic current distributions for putative inosine-bearing CUACU 5-mer in the 3′-UTR region of the AHR gene. (f) Ionic current distributions for putative inosine-bearing AAAAA 5-mer in the 3′-UTR region of the AHR gene.*

2.14 Methods

Unless otherwise noted, kit based protocols described below followed the manufacturer's instructions.

2.14.1 GM12878 cell tissue culture

GM12878 cells (passage 4) were received from the Coriell Institute and cultured in RPMI media (Invitrogen cat# 21870076) supplemented with 15% non heat-inactivated FBS (Lifetech cat# 12483020) and 2mM L-Glutamax (Lifetech cat# 35050061). Cells were grown to a density of 1 x 10^6 / ml before subsequent dilution of every $\sim$3 days and expanded to 9 x T75 flasks (45 ml of media in each). Cells were centrifuged for 10 min at 100 x g (4^{o}C), washed in 1/10th volume of PBS (pH 7.4) and combined for homogeneity. The cells were then evenly split between 8 x 15ml tubes and pelleted at 100g for 10 mins at 4^{o}C. The cell pellets were then snap frozen in liquid nitrogen and immediately stored at -80^{o}C before shipping on dry ice. Two tubes of 5 x 10^7 frozen GM12878 cell pellets from passage 10 from a single passage, cultured at UBC, were distributed and used at UBC, OICR, JHU, and UCSC. Two tubes of cells from passage 11 were distributed to UoN from UBC, and an independently cultured passage of GM12878 was used at UoB. (University of British Columbia (UBC), University of Birmingham (UoB), Ontario Institute of Cancer Research (OICR), Johns Hopkins University (JHU), University of Nottingham (UoN), and University of California Santa Cruz (UCSC))

2.14.2 Total RNA Isolation

The following protocol was used by each of the six institutions. Four ml of TRI-Reagent (Invitrogen AM9738) was added to a frozen pellet of 5 x 10^7 GM12878 cells and vortexed immediately. This sample was incubated at room temperature for 5 minutes. Four hundred μl BCP (1-Bromo-3-chloro-propane) or 200 μl CHCl3 (Chloroform) was added per ml of sample, vortexed, incubated at room temperature for 5 minutes, vortexed again, and centrifuged for 10 minutes at 12,000g (4^{o}C). The aqueous phase was pooled in a LoBind Eppendorf tube and combined with an equal volume of isopropanol. The tube was mixed, incubated at room temperature for 15 minutes, and centrifuged for 15 minutes at 12,000g (4^{o}C). The supernatant was removed, the RNA pellet was washed with 750 μl 80% ethanol and then centrifuged for 5 minutes at 12,000g (4^{o}C). The supernatant was removed. The pellet was air-dried for 10 minutes, resuspended in nuclease free water (100 μl final volume), quantified, and either stored at -80^{o}C or processed further by poly(A) purification.

2.14.3 Poly(A) RNA isolation

One hundred μg aliquots of total RNA were diluted in 100 μl of nuclease free water and poly(A) selected using NEXTflex Poly(A) Beads (BIOO Scientific Cat#NOVA-512980). Resulting poly(A) RNA was eluted in nuclease free water and stored at -80^{o}C.

2.14.4 MinION native RNA sequencing of GM12878 poly(A) RNA

Biological poly(A) RNA (500-775 ng) and a synthetic control (Lexogen SIRV Set 3, 5 ng) were prepared for nanopore direct RNA sequencing generally following the ONT SQK-RNA001 kit protocol, including the optional reverse transcription step recommended by ONT. One difference from the standard ONT protocol was in the use of Superscript IV (Thermo Fisher) for reverse transcription. RNA sequencing on the MinION and GridION platforms was performed using ONT R9.4 flow cells and the standard MinKNOW (version 1.7.14) protocol script recommended by ONT, with one exception, i.e. we restarted the sequencing runs at several time points to improve active pore counts and throughput during the first 24hrs.

2.14.5 cDNA synthesis

First strand cDNA synthesis was performed using Superscript IV (Thermo Fisher) and 100 ng of poly(A) purified RNA. Reverse transcription and strand-switching primers were provided by ONT in the SQK-PCS108 kit. After reverse transcription, PCR was performed using LongAmp Taq Master Mix (NEB) under the following conditions: 95^{o}C for 30 seconds, 11-15 cycles (95^{o}C for 15 seconds, 62^{o}C for 15 seconds, 65^{o}C for 15 minutes), 65^{o}C for 15 minutes, hold at 4^{o}C. The 15 cycle PCR was performed when using the SQK-PCS108 kit and 11 cycle PCR was performed when using the SQK-LSK308 kit. PCR products were purified using 0.8X AMPure XP beads.

2.14.6 MinION sequencing of GM12878 cDNA

cDNA sequencing libraries were prepared using 1 μg of cDNA following the standard ONT protocol for SQK-PCS108 (1D sequencing) or SQK-LSK308 (1D^2 sequencing) with one exception. That is, we used 0.8X aAMPure XP beads for cleanup. We used standard ONT MinKNOW scripts for MinION sequencing with one exception. That is, we restarted the sequencing runs at several time points to improve active pore counts and throughput during the first 24 hours.

2.14.7 Acquiring continuous data for nanopore sequencing runs and resegmenting reads

For a subset of runs, “bulk FAST5 files" containing continuous raw current traces and read decisions made by MinKNOW were recorded for more detailed analysis. This can be enabled in MinKNOW by looking at “Additional options" under “Output" when configuring a run to start in MinKNOW. Options were set to capture raw signal data and the read table. Events were not captured to reduce file size (Payne et al., 2018). Bulk FAST5 files were investigated using BulkVis (Payne et al., 2018)and scripts available on GitHub (`https://github.com/nanopore-wgs-consortium/NA12878/tree/master/nanopore-human-transcriptome/scripts/bulk_signal_read_correction`). To identify reads with abnormal start or ends the read classifications made by MinKNOW in the 2 seconds before and after each read start or end respectively. Read starts should include ‘pore’, ‘good_single’, ‘inrange’ or ‘unblocking’ classifications (Payne et al., 2018). Read ends should also end with these categories.

Reads which did not start or end with these classifications were considered as potentially abnormal. Additional signal before and after the read was extracted from the bulk FAST5 file and a new synthetic read created for base calling (using Albacore version 2.1.3). For abnormal read starts, signal up to the start of the previous read was prepended. For abnormal read ends, signal up to the start of the following read was appended. Base calling is disrupted by signal incorrectly classified as open pore. Therefore these incorrect signal chunks were replaced with signal matching the mean for each read to generate a corrected read. These reads were recalled and mapped against the candidate targets using minimap2 with standard ONT parameters. This method can result in incorrectly concatenated reads and so mapping to the target was used to filter out such sequences. The difference in target coverage for each read was used to indicate recovery of sequence data. All corrected read files, basecalls, mapping files and scripts used to generate them are available on GitHub (link cited above).

2.14.8 Length analysis of mitochondrial protein-coding transcripts

In this analysis, we limited the test population for each gene to reads that aligned to a 50 nt sequence at the 3′ prime end of its ORF, except for *MT-ND5* where alignment was to a 50 nt sequence at the end of its 568 nt 3′ UTR. Full length was defined as extending to at least within 25 nt of the genes expected 5′ terminus. This limit was chosen because the processive enzyme that regulates RNA translocation is distal from the CsgG nanopore limiting aperture and necessarily falls off before the 5′ end is read. The sharpest coverage drop-off is

typically at 10 nt from the 5′ transcript end; we chose the 25 nt limit to ensure that all likely full length reads were captured in the count.

2.14.9 In vitro transcription

cDNA synthesis was performed according to ONT instructions (SQK-PCS108 kit) by combining Superscript IV (Thermo Fisher), RT and ONT strand switching primers, and 100 ng of poly(A) purified RNA. Next, an 11 cycle PCR reaction was performed using the ONT SQK-LSK308 kit but with a modified version of the primer that included a T7 promoter as recommended by NEB (Catalog number E2040S). The PCR reaction was run under the following conditions: 95°C for 30 seconds, 11 cycles (95°C for 15 seconds, 62°C for 15 seconds, 65°C for 15 minutes), 65°C for 15 minutes, hold at 4°C.

PCR products were purified using 0.8X AMPure XP beads. Next, *in vitro* transcription was performed using the NEB HiScribe T7 High Yield RNA Synthesis Kit following NEB instructions. The IVT product was poly(A) tailed using the same kit. The resulting IVT RNA was purified using LiCl precipitation and then adapted for RNA sequencing on the MinION the using SQK-RNA001 kit.

2.14.10 Oligomer Ligation

The oligomer containing the N6-methyladenosine modification was obtained as a lyophilized pellet from Trilink BioTechnologies and resuspended to 20 μM using TE buffer (Quality Biological Cat#351-011-721). The firefly luciferase (FLuc) transcript used as the carrier molecule was produced by *in*

vitro transcription using the HiScribe™ ARCA mRNA Kit (with tailing) (NEB Cat#E2060) and supplied protocol with the following exception: after DNase treatment, the reaction was terminated and the RNA purified using 1X Agencourt RNAClean XP beads (Beckman Coulter A63987). The oligomer was then treated with T4 polynucleotide kinase (PNK) (NEB Cat#M0201) to phosphorylate the 5'end for ligation. After phosphorylation, the oligomer was purified using the Oligo Clean & Concentrator kit (Zymo Research Cat#D4060). The phosphorylated oligomer and FLuc transcript were quantified, combined in equimolar amounts, and ligated using T4 RNA Ligase 1 (NEB Cat#M0204). The reaction mixture was incubated at 16^{o}C overnight. After incubation, the RNA was purified using RNAClean XP beads. The ligated product was poly(A) tailed using *E. coli* Poly(A) Polymerase (NEB HiScribe™ ARCA mRNA Kit) according to the supplier's instructions. After A-tailing, the RNA was purified using RNAClean XP beads. The isolated RNA was poly(A) selected using NEXTflex Poly(A) Beads. The resulting poly(A) RNA was eluted in nuclease free water and immediately prepared for sequencing using Oxford Nanopore's direct RNA sequencing kit (SQK-RNA001) and protocol.

2.14.11 Basecalling, alignments, and percent identity calculations

We used the ONT Albacore workflow (version 2.1.0) for basecalling direct RNA and cDNA data. A strand read with an average sequence quality of 7 or higher (Q7) was classified as pass (default setting for Albacore (version 2.1.0). We used minimap2 version 2.1 (recommended parameters i.e. *-ax splice -uf -k14* for alignments to the human genome and *-ax map-ont* for alignments to

the human transcriptome) to align the nanopore RNA and cDNA reads to the GRCh38 human genome reference and to the GENCODE v27 transcriptome reference. We used marginStats (version 0.1) (Jain et al., 2015b) to calculate alignment identities and errors for pass RNA strand reads and pass 1D cDNA strand reads. Substitutions were calculated using custom scripts available within marginAlign (version 0.1) (Jain et al., 2015a).

2.14.12 Kmer analysis

We assessed nanopore RNA and cDNA 5-mer coverage using GENCODE isoforms. The read sequences were filtered by length and only reads covering 90% or more of the respective reference sequence were chosen. We calculated expected 5-mer counts from the set of reference sequences and observed 5-mer counts from the set of read sequences. For plotting purposes, we normalized the read and reference counts to coverage per megabase. The scripts are available within marginAlign (Jain et al., 2015a).

2.14.13 Isoform detection and characterization

To define isoforms from the sets of native RNA and cDNA reads, we used FLAIR v1.4, a version of FLAIR (Tang et al., 2018) with additional considerations for native RNA nanopore data. For our analysis, we first removed reads generated by lab 6, because a disproportionate number of those molecules appeared to be truncated prior to addition to the nanopore flow cell. We also removed 71,276 aligned reads with deletions greater than 100 bases caused

by minimap2 version 2.1. We then selected reads that had TSSs within promoter regions that were computationally derived from ENCODE ChIP-Seq data (Ernst and Kellis, 2010; Ernst et al., 2011). Using FLAIR-correct, we corrected primary genomic alignments for pass reads based on splice junction evidence from GENCODE v27 annotations and Illumina short-read sequencing of GM12878. This step also removes reads containing non-canonical splice junctions not present in the annotation or short-read data. The filtered and corrected reads were then processed by FLAIR-collapse which generates a first-pass isoform set by grouping reads on their splice junctions chains. Next, pass reads were realigned to the first-pass isoform set, retaining alignments with MAPQ>0. Isoforms with fewer than 3 supporting reads or those which were subsets of a longer isoform were filtered out to compile the FLAIR-sensitive isoform set. A FLAIR-stringent isoform set was also compiled by filtering the FLAIR-sensitive set for isoforms which had 3 supporting reads that spanned $\geq$ 80% of the isoform and a minimum of 25nt into the first and last exons. Unannotated isoforms were defined as those with a unique splice junction chain not found in GENCODE v27. Isoforms were considered intron-retaining if they contained an exon which completely spanned another isoform's splice junction. Isoforms with unannotated exons were defined as those with at least one exon that did not overlap any existing annotated exons in GENCODE v27. Isoforms at unannotated loci were defined as isoforms that only contain unannotated exons. Genes that did not contain an annotated start codon were considered non-coding genes.

2.14.14 Defining promoter regions in GM12878 for isoform filtering

Promoter chromatin states for GM12878 were downloaded from the UCSC Genome Browser in BED format from the hg18 genome reference. Chromatin states were derived from an HMM based on ENCODE ChIP-Seq data of nine factors (Ernst and Kellis, 2010; Ernst et al., 2011). The liftover tool (Hinrichs et al., 2006)was used to convert hg18 coordinates to hg38. The active, weak, and poised promoter states were used.

2.14.15 Haplotype Assignment and Allele-Specific Analysis

We obtained genotype information for GM12878 from existing phased Illumina platinum genome data generated by deep sequencing of the cell donors' familial trio (Eberle et al., 2016). The bcftools package was used to filter for only variants that are heterozygous in GM12878. Starting with aligned reads, we used the extractHAIRS utility of the haplotype-sensitive assembler HapCUT2 (Edge, Bafna, and Bansal, 2017) to identify reads with allele-informative variants. For allelic assignment, we required a read to contain at least two variants, and required that greater than 75% of identified variants agreed on the parental allele of origin – this stringent threshold was selected to reduce the chances of incorrect assignment from nanopore sequencing errors. Through this approach, each read was annotated as maternal, paternal or unassigned. To identify genes that demonstrated a very strong bias for a single allele, we performed a binomial test of all reads assigned to a parental allele, with an FDR of 0.001. We also visually inspected numerous genes displaying

genes demonstrating allele-specificity using IGV, to increase our confidence in proper mapping of the reads and evaluate the presence of variants.

We further integrated this haplotype-specific analysis with our isoform pipeline to explore for the presence of allele-specific isoforms. If reads for a specific isoform originated from a single parental allele (binomial test, FDR 0.001), the isoform was assigned as allele specific. We then filtered for any genes which contained both maternal and paternal allele-specific isoforms, and visually inspected these isoforms using IGV to compare location of variants and splicing events.

2.14.16 Poly(A) tail length analysis

The Supplemental Note describes use of nanopolish-polya version 0.10.2 (`https://github.com/jts/nanopolish`) to estimate polyadenylated tail lengths of nanopore native RNA sequence reads. We used the Kurskal-Wallis test as implemented in Python to determine statistically significant changes between isoforms; code is available at [`https://github.com/nanopore-wgs-consortium/NA12878/tree/master/nanopore-human-transcriptome/scripts`]

2.14.17 Modification detection and analysis

We focused our initial efforts on m6A modification in genes previously identified as enriched in modifications from m6A immunoprecipitation sequencing data on human cell lines (Roost et al., 2015; Molinie et al., 2016). We aligned native RNA reads and IVT RNA reads to candidate genes and then extracted ionic current information (mean current and standard deviation

in pA) for specific 5-mers using nanopolish eventalign (version 0.10.2). We compared ionic current kernel density estimates (KDE) for GGACU within the 3′ UTR of the *EEF2* gene in native RNA with the KDE for its canonical IVT RNA counterpart. The extent and directionality of current shifts observed by m6A modification within the GGACU motif were orthogonally investigated using an in-vitro oligomer ligation assay, as described above. We compared KDEs for the modified and unmodified GGACU motifs within the synthetic oligomer. Statistical testing (Kruskal-Wallis, Student's t-test, Kolmogorov-Smirnov and Bonferroni correction) was implemented in Python with code available at [https://github.com/nanopore-wgs-consortium/NA12878/tree/master/nanopore-human-transcriptome/scripts].

For detecting A-to-I editing, we focused on the 3′-UTR region of the human aryl hydrocarbon receptor (*AHR*) gene. Using the UCSC Genome Browser, we identified systematic G base variant calls in *AHR* cDNA data (probable inosine substitutions in RNA). We then tested for systematic base miscalls at the corresponding positions in native RNA data. Next, we used nanopolish eventalign (version 0.10.2) to extract ionic current information for two putative inosine-containing 5-mers (CUACU and AAAAA), and for their respective IVT-derived canonical 5-mers from chromosome 7. Ionic current distributions for CUACU and AAAAA 5-mers between the biological and IVT data were compared using kernel density estimates.

2.15 References

Anvar, Seyed Yahya, Guy Allard, Elizabeth Tseng, Gloria M Sheynkman, Eleonora de Klerk, Martijn Vermaat, Raymund H Yin, Hans E Johansson, Yavuz Ariyurek, Johan T den Dunnen, Stephen W Turner, and Peter A C 't Hoen (2018). "Full-length mRNA sequencing uncovers a widespread coupling between transcription initiation and mRNA processing". en. In: *Genome Biol.* 19.1, p. 46.

Baltimore, David (1970). "Viral RNA-dependent DNA Polymerase: RNA-dependent DNA Polymerase in Virions of RNA Tumour Viruses". In: *Nature* 226, p. 1209.

Baralle, Francisco E and Jimena Giudice (2017). "Alternative splicing as a regulator of development and tissue identity". en. In: *Nat. Rev. Mol. Cell Biol.* 18.7, pp. 437–451.

Bernstein, Bradley E, Michael Kamal, Kerstin Lindblad-Toh, Stefan Bekiranov, Dione K Bailey, Dana J Huebert, Scott McMahon, Elinor K Karlsson, Edward J Kulbokas 3rd, Thomas R Gingeras, Stuart L Schreiber, and Eric S Lander (2005). "Genomic maps and comparative analysis of histone modifications in human and mouse". en. In: *Cell* 120.2, pp. 169–181.

Bradley, Robert K, Jason Merkin, Nicole J Lambert, and Christopher B Burge (2012). "Alternative splicing of RNA triplets is often regulated and accelerates proteome evolution". en. In: *PLoS Biol.* 10.1, e1001229.

Bresson, Stefan M, Olga V Hunter, Allyson C Hunter, and Nicholas K Conrad (2015). "Canonical Poly(A) Polymerase Activity Promotes the Decay of a Wide Variety of Mammalian Nuclear RNAs". en. In: *PLoS Genet.* 11.10, e1005610.

Brown, Carolyn J, Andrea Ballabio, James L Rupert, Ronald G Lafreniere, Markus Grompe, Rossana Tonlorenzi, and Huntington F Willard (1991). "A gene from the region of the human X inactivation centre is expressed exclusively from the inactive X chromosome". In: *Nature* 349, p. 38.

Cech, Thomas R and Joan A Steitz (2014). "The noncoding RNA revolution-trashing old rules to forge new ones". en. In: *Cell* 157.1, pp. 77–94.

Chang, Hyeshik, Jaechul Lim, Minju Ha, and V Narry Kim (2014). "TAIL-seq: genome-wide determination of poly(A) tail length and 3′ end modifications". en. In: *Mol. Cell* 53.6, pp. 1044–1052.

Cho, Hyunghoon, Joe Davis, Xin Li, Kevin S Smith, Alexis Battle, and Stephen B Montgomery (2014). "High-resolution transcriptome analysis with long-read RNA sequencing". en. In: *PLoS One* 9.9, e108095.

Cloonan, Nicole, Alistair R R Forrest, Gabriel Kolle, Brooke B A Gardiner, Geoffrey J Faulkner, Mellissa K Brown, Darrin F Taylor, Anita L Steptoe, Shivangi Wani, Graeme Bethel, Alan J Robertson, Andrew C Perkins, Stephen J Bruce, Clarence C Lee, Swati S Ranade, Heather E Peckham, Jonathan M Manning, Kevin J McKernan, and Sean M Grimmond (2008). "Stem cell transcriptome profiling via massive-scale mRNA sequencing". en. In: *Nat. Methods* 5.7, pp. 613–619.

Dai, Dongjun, Hanying Wang, Liyuan Zhu, Hongchuan Jin, and Xian Wang (2018). "N6-methyladenosine links RNA metabolism to cancer progression". en. In: *Cell Death Dis.* 9.2, p. 124.

Deonovic, Benjamin, Yunhao Wang, Jason Weirather, Xiu-Jie Wang, and Kin Fai Au (2017). "IDP-ASE: haplotyping and quantifying allele-specific expression at the gene and gene isoform level by hybrid sequencing". en. In: *Nucleic Acids Res.* 45.5, e32.

Deveson, Ira W, Marion E Brunck, James Blackburn, Elizabeth Tseng, Ting Hon, Tyson A Clark, Michael B Clark, Joanna Crawford, Marcel E Dinger, Lars K Nielsen, John S Mattick, and Tim R Mercer (2018). "Universal Alternative Splicing of Noncoding Exons". en. In: *Cell Syst* 6.2, 245–255.e5.

Eberle, Michael A, Epameinondas Fritzilas, Peter Krusche, Morten Källberg, Benjamin L Moore, Mitchell A Bekritsky, Zamin Iqbal, Han-Yu Chuang, Sean J Humphray, Aaron L Halpern, Semyon Kruglyak, Elliott H Margulies, Gil McVean, and David R Bentley (2016). "A reference data set of 5.4 million phased human variants validated by genetic inheritance from sequencing a three-generation 17-member pedigree". In: *Genome Res.*

Eckmann, Christian R, Christiane Rammelt, and Elmar Wahle (2011). "Control of poly(A) tail length". en. In: *Wiley Interdiscip. Rev. RNA* 2.3, pp. 348–361.

Edge, Peter, Vineet Bafna, and Vikas Bansal (2017). "HapCUT2: robust and accurate haplotype assembly for diverse sequencing technologies". en. In: *Genome Res.* 27.5, pp. 801–812.

Ernst, Jason and Manolis Kellis (2010). "Discovery and characterization of chromatin states for systematic annotation of the human genome". en. In: *Nat. Biotechnol.* 28.8, pp. 817–825.

Ernst, Jason, Pouya Kheradpour, Tarjei S Mikkelsen, Noam Shoresh, Lucas D Ward, Charles B Epstein, Xiaolan Zhang, Li Wang, Robbyn Issner, Michael Coyne, Manching Ku, Timothy Durham, Manolis Kellis, and Bradley E Bernstein (2011). "Mapping and analysis of chromatin state dynamics in nine human cell types". en. In: *Nature* 473.7345, pp. 43–49.

Gallo, Angela, Dragana Vukic, David Michalík, Mary A O'Connell, and Liam P Keegan (2017). "ADAR RNA editing in human disease; more to it than meets the I". en. In: *Hum. Genet.* 136.9, pp. 1265–1278.

Garalde, Daniel R, 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 (2018). "Highly parallel direct RNA sequencing on an array of nanopores". en. In: *Nat. Methods*.

Gonzàlez-Porta, Mar, Adam Frankish, Johan Rung, Jennifer Harrow, and Alvis Brazma (2013). "Transcriptome analysis of human tissues and cell lines reveals one dominant transcript per gene". en. In: *Genome Biol.* 14.7, R70.

Hinrichs, A S, D Karolchik, R Baertsch, G P Barber, G Bejerano, H Clawson, M Diekhans, T S Furey, R A Harte, F Hsu, J Hillman-Jackson, R M Kuhn, J S Pedersen, A Pohl, B J Raney, K R Rosenbloom, A Siepel, K E Smith, C W Sugnet, A Sultan-Qurraie, D J Thomas, H Trumbower, R J Weber, M Weirauch, A S Zweig, D Haussler, and W J Kent (2006). "The UCSC Genome Browser Database: update 2006". en. In: *Nucleic Acids Res.* 34.Database issue, pp. D590–8.

Jain, Miten, Ian T Fiddes, Karen H Miga, Hugh E Olsen, Benedict Paten, and Mark Akeson (2015a). "Improved data analysis for the MinION nanopore sequencer". en. In: *Nat. Methods* 12.4, pp. 351–356.

Jain, Miten, Ian T Fiddes, Karen H Miga, Hugh E Olsen, Benedict Paten, and Mark Akeson (2015b). "Improved data analysis for the MinION nanopore sequencer". In: *Nat. Methods* 12.4, pp. 351–356.

Jain, Miten, 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 J Loman, and Matthew Loose (2018). "Nanopore sequencing and assembly of a human genome with ultra-long reads". In: *Nat. Biotechnol.* 36, p. 338.

Jenjaroenpun, Piroon, Thidathip Wongsurawat, Rui Pereira, Preecha Patumcharoenpol, David W Ussery, Jens Nielsen, and Intawat Nookaew (2018).

"Complete genomic and transcriptional landscape analysis using third-generation sequencing: a case study of Saccharomyces cerevisiae CEN.PK113-7D". en. In: *Nucleic Acids Res.*

Lee, Mihye, Boseon Kim, and V Narry Kim (2014). "Emerging roles of RNA modification: m(6)A and U-tail". en. In: *Cell* 158.5, pp. 980–987.

Li, Heng (2018). "Minimap2: pairwise alignment for nucleotide sequences". en. In: *Bioinformatics*.

Li, Xiaoyu, Xushen Xiong, and Chengqi Yi (2016). "Epitranscriptome sequencing technologies: decoding RNA modifications". en. In: *Nat. Methods* 14.1, pp. 23–31.

Licht, Konstantin, Utkarsh Kapoor, Elisa Mayrhofer, and Michael F Jantsch (2016). "Adenosine to Inosine editing frequency controlled by splicing efficiency". en. In: *Nucleic Acids Res.* 44.13, pp. 6398–6408.

Lister, Ryan, Ronan C O'Malley, Julian Tonti-Filippini, Brian D Gregory, Charles C Berry, A Harvey Millar, and Joseph R Ecker (2008). "Highly integrated single-base resolution maps of the epigenome in Arabidopsis". en. In: *Cell* 133.3, pp. 523–536.

Liu, Nian and Tao Pan (2016). "N6-methyladenosine–encoded epitranscriptomics". en. In: *Nat. Struct. Mol. Biol.* 23.2, pp. 98–102.

Marioni, John C, Christopher E Mason, Shrikant M Mane, Matthew Stephens, and Yoav Gilad (2008). "RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays". en. In: *Genome Res.* 18.9, pp. 1509–1517.

Mayba, Oleg, Houston N Gilbert, Jinfeng Liu, Peter M Haverty, Suchit Jhunjhunwala, Zhaoshi Jiang, Colin Watanabe, and Zemin Zhang (2014). "MBASED: allele-specific expression detection in cancer tissues and cell lines". en. In: *Genome Biol.* 15.8, p. 405.

Meyer, Kate D, Yogesh Saletore, Paul Zumbo, Olivier Elemento, Christopher E Mason, and Samie R Jaffrey (2012). "Comprehensive analysis of mRNA methylation reveals enrichment in 3' UTRs and near stop codons". en. In: *Cell* 149.7, pp. 1635–1646.

Molinie, Benoit, Jinkai Wang, Kok Seong Lim, Roman Hillebrand, Zhi-Xiang Lu, Nicholas Van Wittenberghe, Benjamin D Howard, Kaveh Daneshvar, Alan C Mullen, Peter Dedon, Yi Xing, and Cosmas C Giallourakis (2016). "m6A-LAIC-seq reveals the census and complexity of the m6A epitranscriptome". In: *Nat. Methods* 13, p. 692.

Morin, Ryan, Matthew Bainbridge, Anthony Fejes, Martin Hirst, Martin Krzywinski, Trevor Pugh, Helen McDonald, Richard Varhol, Steven Jones,

and Marco Marra (2008). “Profiling the HeLa S3 transcriptome using randomly primed cDNA and massively parallel short-read sequencing”. en. In: *Biotechniques* 45.1, pp. 81–94.

Mortazavi, Ali, Brian A Williams, Kenneth McCue, Lorian Schaeffer, and Barbara Wold (2008). “Mapping and quantifying mammalian transcriptomes by RNA-Seq”. en. In: *Nat. Methods* 5.7, pp. 621–628.

Nagalakshmi, Ugrappa, Zhong Wang, Karl Waern, Chong Shou, Debasish Raha, Mark Gerstein, and Michael Snyder (2008). “The transcriptional landscape of the yeast genome defined by RNA sequencing”. en. In: *Science* 320.5881, pp. 1344–1349.

Nishikura, Kazuko (2010). “Functions and regulation of RNA editing by ADAR deaminases”. en. In: *Annu. Rev. Biochem.* 79, pp. 321–349.

Ozsolak, Fatih, Adam R Platt, Dan R Jones, Jeffrey G Reifenberger, Lauryn E Sass, Peter McInerney, John F Thompson, Jayson Bowers, Mirna Jarosz, and Patrice M Milos (2009). “Direct RNA sequencing”. en. In: *Nature* 461.7265, pp. 814–818.

Pandey, Ram Vinay, Susanne U Franssen, Andreas Futschik, and Christian Schlötterer (2013). “Allelic imbalance metre (Allim), a new tool for measuring allele-specific gene expression with RNA-seq data”. en. In: *Mol. Ecol. Resour.* 13.4, pp. 740–745.

Parker, Roy and Haiwei Song (2004). “The enzymes and control of eukaryotic mRNA turnover”. en. In: *Nat. Struct. Mol. Biol.* 11.2, pp. 121–127.

Payne, Alexander, Nadine Holmes, Vardhman Rakyan, and Matthew Loose (2018). “BulkVis: a graphical viewer for Oxford nanopore bulk FAST5 files”. en. In: *Bioinformatics*.

Preiss, Thomas (2013). *The End in Sight: Poly(A), Translation and mRNA Stability in Eukaryotes*. en. Landes Bioscience.

Rand, Arthur C, Miten Jain, Jordan M Eizenga, Audrey Musselman-Brown, Hugh E Olsen, Mark Akeson, and Benedict Paten (2017). “Mapping DNA methylation with high-throughput nanopore sequencing”. en. In: *Nat. Methods* 14.4, pp. 411–413.

Roost, Caroline, Stephen R Lynch, Pedro J Batista, Kun Qu, Howard Y Chang, and Eric T Kool (2015). “Structure and thermodynamics of N6-methyladenosine in RNA: a spring-loaded base modification”. en. In: *J. Am. Chem. Soc.* 137.5, pp. 2107–2115.

Roundtree, Ian A, Molly E Evans, Tao Pan, and Chuan He (2017). “Dynamic RNA Modifications in Gene Expression Regulation”. en. In: *Cell* 169.7, pp. 1187–1200.

Rozowsky, Joel, Alexej Abyzov, Jing Wang, Pedro Alves, Debasish Raha, Arif Harmanci, Jing Leng, Robert Bjornson, Yong Kong, Naoki Kitabayashi, Nitin Bhardwaj, Mark Rubin, Michael Snyder, and Mark Gerstein (2011). "AlleleSeq: analysis of allele-specific expression and binding in a network framework". en. In: *Mol. Syst. Biol.* 7, p. 522.

Saiki, R K, D H Gelfand, S Stoffel, S J Scharf, R Higuchi, G T Horn, K B Mullis, and H A Erlich (1988). "Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase". en. In: *Science* 239.4839, pp. 487–491.

Sibbritt, Tennille, Hardip R Patel, and Thomas Preiss (2013). "Mapping and significance of the mRNA methylome". en. In: *Wiley Interdiscip. Rev. RNA* 4.4, pp. 397–422.

Simpson, Jared T, Rachael E Workman, P C Zuzarte, Matei David, L J Dursi, and Winston Timp (2017). "Detecting DNA cytosine methylation using nanopore sequencing". en. In: *Nat. Methods* 14.4, pp. 407–410.

Skelly, Daniel A, Marnie Johansson, Jennifer Madeoy, Jon Wakefield, and Joshua M Akey (2011). "A powerful and flexible statistical framework for testing hypotheses of allele-specific gene expression from RNA-seq data". en. In: *Genome Res.* 21.10, pp. 1728–1737.

Smith, Andrew M, Miten Jain, Logan Mulroney, Daniel R Garalde, and Mark Akeson (2019). *Reading canonical and modified nucleobases in 16S ribosomal RNA using nanopore native RNA sequencing*.

Steijger, Tamara, Josep F Abril, Pär G Engström, Felix Kokocinski, RGASP Consortium, Tim J Hubbard, Roderic Guigó, Jennifer Harrow, and Paul Bertone (2013). "Assessment of transcript reconstruction methods for RNA-seq". en. In: *Nat. Methods* 10.12, pp. 1177–1184.

Su, Ye, Haijiang Wu, Alexander Pavlosky, Ling-Lin Zou, Xinna Deng, Zhu-Xu Zhang, and Anthony M Jevnikar (2016). "Regulatory non-coding RNA: new instruments in the orchestration of cell death". en. In: *Cell Death Dis.* 7.8, e2333.

Subtelny, Alexander O, Stephen W Eichhorn, Grace R Chen, Hazel Sive, and David P Bartel (2014). "Poly(A)-tail profiling reveals an embryonic switch in translational control". en. In: *Nature* 508.7494, pp. 66–71.

Szczesny, Roman J, Lukasz S Borowski, Michal Malecki, Magdalena A Wojcik, Piotr P Stepien, and Pawel Golik (2012). "RNA degradation in yeast and human mitochondria". en. In: *Biochim. Biophys. Acta* 1819.9-10, pp. 1027–1034.

Tajaddod, Mansoureh, Michael F Jantsch, and Konstantin Licht (2016). "The dynamic epitranscriptome: A to I editing modulates genetic information". en. In: *Chromosoma* 125.1, pp. 51–63.

Tang, Alison D, Cameron M Soulette, Marijke J van Baren, Kevyn Hart, Eva Hrabeta-Robinson, Catherine J Wu, and Angela N Brooks (2018). "Full-length transcript characterization of SF3B1 mutation in chronic lymphocytic leukemia reveals downregulation of retained introns". en.

Tardaguila, Manuel, Lorena de la Fuente, Cristina Marti, Cécile Pereira, Francisco Jose Pardo-Palacios, Hector Del Risco, Marc Ferrell, Maravillas Mellado, Marissa Macchietto, Kenneth Verheggen, Mariola Edelmann, Iakes Ezkurdia, Jesus Vazquez, Michael Tress, Ali Mortazavi, Lennart Martens, Susana Rodriguez-Navarro, Victoria Moreno-Manzano, and Ana Conesa (2018). "SQANTI: extensive characterization of long-read transcript sequences for quality control in full-length transcriptome identification and quantification". en. In: *Genome Res.*

Temin, H M and S Mizutani (1970). "RNA-dependent DNA polymerase in virions of Rous sarcoma virus". en. In: *Nature* 226.5252, pp. 1211–1213.

Temperley, Richard J, Mateusz Wydro, Robert N Lightowlers, and Zofia M Chrzanowska-Lightowlers (2010). "Human mitochondrial mRNAs—like members of all families, similar but different". In: *Biochimica et Biophysica Acta (BBA) - Bioenergetics* 1797.6, pp. 1081–1085.

Tilgner, Hagen, Fabian Grubert, Donald Sharon, and Michael P Snyder (2014). "Defining a personal, allele-specific, and single-molecule long-read transcriptome". en. In: *Proc. Natl. Acad. Sci. U. S. A.* 111.27, pp. 9869–9874.

Turro, Ernest, Shu-Yi Su, Ângela Gonçalves, Lachlan J M Coin, Sylvia Richardson, and Alex Lewin (2011). "Haplotype and isoform specific expression estimation using multi-mapping RNA-seq reads". en. In: *Genome Biol.* 12.2, R13.

Venturini, Luca, Shabhonam Caim, Gemy George Kaithakottil, Daniel Lee Mapleson, and David Swarbreck (2018). "Leveraging multiple transcriptome assembly methods for improved gene structure annotation". en. In: *Gigascience* 7.8.

Vilfan, Igor D, Yu-Chih Tsai, Tyson A Clark, Jeffrey Wegener, Qing Dai, Chengqi Yi, Tao Pan, Stephen W Turner, and Jonas Korlach (2013). "Analysis of RNA base modification and structural rearrangement by single-molecule real-time detection of reverse transcription". en. In: *J. Nanobiotechnology* 11, p. 8.

Wang, Lili, Angela N Brooks, Jean Fan, Youzhong Wan, Rutendo Gambe, Shuqiang Li, Sarah Hergert, Shanye Yin, Samuel S Freeman, Joshua Z Levin, Lin Fan, Michael Seiler, Silvia Buonamici, Peter G Smith, Kevin F Chau, Carrie L Cibulskis, Wandi Zhang, Laura Z Rassenti, Emanuela M Ghia, Thomas J Kipps, Stacey Fernandes, Donald B Bloch, Dylan Kotliar, Dan A Landau, Sachet A Shukla, Jon C Aster, Robin Reed, David S DeLuca, Jennifer R Brown, Donna Neuberg, Gad Getz, Kenneth J Livak, Matthew M Meyerson, Peter V Kharchenko, and Catherine J Wu (2016). "Transcriptomic Characterization of SF3B1 Mutation Reveals Its Pleiotropic Effects in Chronic Lymphocytic Leukemia". en. In: *Cancer Cell* 30.5, pp. 750–763.

Wang, Zhong, Mark Gerstein, and Michael Snyder (2009). "RNA-Seq: a revolutionary tool for transcriptomics". In: *Nat. Rev. Genet.* 10, p. 57.

Wilhelm, Brian T, Samuel Marguerat, Stephen Watt, Falk Schubert, Valerie Wood, Ian Goodhead, Christopher J Penkett, Jane Rogers, and Jürg Bähler (2008). "Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution". en. In: *Nature* 453.7199, pp. 1239–1243.

Woo, Yu Mi, Yeonui Kwak, Sim Namkoong, Katla Kristjánsdóttir, Seung Ha Lee, Jun Hee Lee, and Hojoong Kwak (2018). "TED-Seq Identifies the Dynamics of Poly(A) Length during ER Stress". en. In: *Cell Rep.* 24.13, 3630–3641.e7.

Workman, Rachael E, Alison D Tang, Paul S Tang, Miten Jain, John R Tyson, Roham Razaghi, Philip C Zuzarte, Timothy Gilpatrick, Alexander Payne, Joshua Quick, et al. (2019). "Nanopore native RNA sequencing of a human poly (A) transcriptome". In: *Nature methods* 16.12, pp. 1297–1305.

Yi, Hyerim, Joha Park, Minju Ha, Jaechul Lim, Hyeshik Chang, and V Narry Kim (2018). "PABP Cooperates with the CCR4-NOT Complex to Promote mRNA Deadenylation and Block Precocious Decay". en. In: *Mol. Cell* 70.6, 1081–1088.e5.

Chapter 3

Direct detection of RNA modifications and structure using single molecule nanopore sequencing

This chapter is a published manuscript at the journal Cell Genomics reprinted in compliance with the journal policies. I am a co-author on this study responsible for the analysis of electrical nanopore signals to infer RNA modification sites and secondary structure.

William Stephenson et al. (2022). "Direct detection of RNA modifications and structure using single-molecule nanopore sequencing". In: *Cell genomics* 2.2, p. 100097

3.1 Introduction

Modifications are present on many classes of RNA including tRNA, rRNA and mRNA. These modifications modulate diverse biological processes such as genetic recoding, mRNA export and folding. In addition, modifications

can be introduced to RNA molecules using chemical probing strategies that reveal RNA structure and dynamics. Many methods exist to detect RNA modifications by short-read sequencing; however, limitations on read length inherent to short-read-based methods dissociate modifications from their native context, preventing single-molecule modification analysis. Here we demonstrate direct RNA nanopore sequencing to detect endogenous and exogenous RNA modifications on long RNAs at the single-molecule level. We detect endogenous 2′-*O*-methyl and base modifications across *E. coli* and *S. cerevisiae* ribosomal RNAs as shifts in current signal and dwell times distally through interactions with the helicase motor protein. We further use the 2′-hydroxyl reactive SHAPE reagent, acetylimidazole, to probe RNA structure at the single-molecule level with readout by direct nanopore sequencing.

Over 100 distinct modifications of RNA have been identified, occurring on either the nucleobase or the ribose sugar. These modifications exhibit diverse effects on RNA structure and function, including modulation of stability, translation efficiency, structural dynamics, nuclear export and translational recoding, (Wendy, Tristan, and Cassandra, 2016; Jun, Hyosuk, and Christine, 2016; Daniel et al., 2019) and in some cases, are installed, read or removed by modification “writers", “readers" and “erasers", suggesting a dynamic model of post-transcriptional gene regulation (Deepak et al., 2018; Zaccara et al., 2019). The 2′-*O*-methyl (Nm) modification occurs in the 5′ cap of eucaryotic mRNAs (m^7GpppNmNm) and extensively in ribosomal RNAs (rRNA). Nm modifications have also been detected within coding regions of mRNA (Qing et al., 2017) and appear to tune cognate tRNA selection during translation,

thereby adjusting protein synthesis dynamics (**choi20182'**). Pseudouridine (Ψ), often referred to as the fifth base due to its widespread inclusion in diverse classes of RNA, is generated by isomerization of uracil, to create a nucleoside with distinct hydrogen bonding and base pairing properties and increased base stacking propensity relative to uridine (Emily, Anna, and Eric, 2017; Schwartz et al., 2014). The absence or reduction of RNA modifications have been implicated in multiple diseases including cancer, heart disease, and genetic diseases (Nicky et al., 2017; Bianca et al., 2005; Yi, Thomas, and Meier, 2014). Comprehensive detection and localization of RNA modifications within their native context will improve our understanding of RNA modification function and regulation, and their role in disease.

Current methods to detect post-transcriptional RNA modifications fall into three broad classes. Immunoprecipitation methods use antibodies specific to individual modifications to enrich short fragments of RNA with the modifications, which are then converted into cDNA and sequenced with short reads (Mark and Yuri, 2017; Xiaoyu, Xushen, and Chengqi, 2016). These approaches can be applied genome-wide, but do not always provide nucleotide resolution and are limited by the availability and specificity of the pulldown reagents. An alternative family of approaches takes advantage of the propensity of reverse transcriptase (RT) enzymes to either terminate cDNA synthesis or incorporate noncomplementary nucleotides when a modified RNA base is encountered (Qing et al., 2017; Thomas et al., 2014; Kate, 2019; Matthias et al., 2009). Finally, modified nucleosides can be directly detected by mass spectrometry (MALDI

or LC-MSMS); unique mass to charge ratios (m/z) and peaks can provide information relating to precise chemical identities and abundances of modified nucleosides and their immediate sequence context (Rebecca et al., 2015; Ning et al., 2019). Each of these methods has proven useful in specific contexts, but all have limited resolution and typically only probe one modification at a time.

Exogenous modifications of RNA have been used as temporal tags to assay RNA dynamics, including stability, turnover, and splicing timing (Veronika et al., 2017; Heather, Karine, and L, 2020; Kentaro et al., 2020; Liying et al., 2020). In addition, chemical probing is widely used to monitor RNA structure. Chemical probes modify either the nucleobase (for example, dimethyl sulfate DMS) or the ribose 2′-hydroxyl group, the site of modification in Selective 2′-Hydroxyl Acylation analyzed by Primer Extension (SHAPE) strategies. The location of SHAPE reagent-induced modification is detected by exploiting RT termination at the modified residue or through detection of a mutation opposite the modified site during RT readthrough. This latter approach, termed Selective 2′-Hydroxyl Acylation analyzed by Primer Extension and Mutational Profiling (SHAPE-MaP) (Nathan et al., 2014), yields single nucleotide resolution reactivity patterns and can be employed transcriptome wide (Yue et al., 2014; Anthony et al., 2018). However, SHAPE-MaP technology is limited by current constraints of short-read sequencing. An ideal method for investigating both endogenous and exogenous RNA modifications would detect multiple classes of modifications and would allow observation of multiple modifications on the same RNA molecule.

Direct RNA nanopore sequencing has emerged as a promising technology for full-length sequencing and analysis of both cell-derived and synthetic RNA molecules. In the commercial platform developed by Oxford Nanopore Technologies (ONT), RNA is translocated via a motor protein through a biological nanopore suspended in a membrane (Figure 3.1A). As the RNA transits through the pore under voltage bias, the observed changes in picoampere ionic current are characteristic of the chemical identity and sequence of the 5 nucleotides (“kmer") positioned at the pore constriction (Garalde et al. 2018; Workman et al. 2019).

Here we adapt direct RNA sequencing on the ONT platform to detect both endogenous and exogenous RNA modifications. Importantly, we show that the raw current signal from nanopore sequencing detects RNA modifications, independent of the chemical nature of the modification. In addition, we describe modification-dependent signals in the time domain, relating to the translocation rate of single molecules through the nanopore. This time domain signal provides a complementary dimension of information that may be incorporated with current signal for *de novo* identification of nucleotide modification classes. Building on these insights, we developed nanoSHAPE that combines long-read, direct RNA sequencing with a new SHAPE reagent that, by virtue of its high reactivity and small adduct size, enables full-length probing of structure in long RNAs.

3.2 Results

3.2.1 Identification of specific modifications at defined locations within 16S rRNA

We first applied direct ONT RNA sequencing to rRNAs from *Escherichia coli* and *Saccharomyces cerevisiae*; these rRNA are highly abundant and harbor well-characterized modifications. We generated in vitro transcripts, which are devoid of modifications, of the small and large subunit RNAs, as controls. These in vitro transcribed (IVT) controls, were sequenced independently to provide a modification-free baseline against which native, cell-derived RNAs could be compared to identify putative sites of modification (Andrew et al., 2019; Felix et al., 2019). IVT controls and native RNAs exhibited reads at the expected lengths for *E. coli* (16S: 1.5 kb, 23S: 2.9 kb) (Figure 3.1B) and *S. cerevisiae* (18S: 1.8 kb, 25S: 3.4 kb). Notably, median quality scores for full-length molecules from *E. coli* were lower than for the native samples as compared to IVT controls (16S: -0.53, 23S: -0.31, t-test $p < 0.05$); results were similar for *S. cerevisiae*. This reduction likely reflects the effect that high modification levels have on read quality using the current iteration of modification un-aware base calling software. Coverage was lower towards the 5′ ends for all samples, including IVT controls as expected from the configuration of direct RNA nanopore sequencing, which translocates RNA in the 3′ to 5′ direction.

Reads were processed using both Tombo (Marcus et al., 2016) and Nanopolish (Jared et al., 2017), which perform raw current signal level to sequence alignment. As an example, we highlight a region of current signal mapped

with Tombo on 16S rRNA in *E. coli* containing two modifications in proximity: N4,2′-*O*-dimethylcytosine (m^4Cm) at position 1402 and 5-methylcytosine (m^5C) at position 1407 (Figure 3.1C). We observed a clear deviation in current signal distribution of the native sample at position 1402, whereas current signal deviation due to the chemically distinct m^5C modification at position 1407 was less pronounced and spread over 3 positions, 1406 - 1408 highlighting the often-distributed nature of complex nanopore signals in the kmer context.

3.2.2 Comprehensive rRNA modification detection

We next examined whether nanopore sequencing could detect all known modifications in the small and large subunit rRNAs of both *E. coli* (Triinu and Jaanus, 2010) and *S. cerevisiae* (Masato et al., 2016). Normalized current and dwell time differences between native and IVT samples were observed across the small and large subunit rRNAs consistent with the locations of known modifications. To more rigorously assess modification signals, which are highly dependent on kmer sequence context, we performed non-parametric Kolmogorov-Smirnov testing (KS) across all positions for current and dwell time from raw signal aligned data using both Tombo and Nanopolish (Methods, Figure 3.1D). Peaks in the KS statistic profile indicated distributional differences between the IVT unmodified control sample and the native samples. The KS statistic for the current signal was strongly correlated between Tombo and Nanopolish (Pearson correlation, $r = 0.75$ - 0.79) however the dwell time KS statistic profiles were only moderately correlated (Pearson correlation, $r = 0.45$ - 0.55). Generally, KS statistic peaks for current were

observed within $\pm$2 nt of the known modifications.

To assess modification class features we collated all RNA modifications across the small and large ribosomal subunits from both *E. coli* and *S. cerevisiae*, aligning them by their known modification position and considered the aggregate median KS statistic profile across modification class: Nm, Ψ, and base (e.g., base modifications excluding Ψ). The median KS statistic for current of all modification classes had an appreciable signal at the site of modification (pore constriction) as expected (Figure 3.1E). With respect to the dwell time, the median KS statistic profile exhibited two peaks for Nm and pseudouridine modifications but not for base modifications (Figure 3.1E). The primary peak occurred at the nanopore constriction (relative modification position 0); however, secondary peaks were observed approximately 10 nucleotides in the 3′ direction with Nm modifications exhibiting a larger median KS statistic than Ψ. This 10-nt distance is the 'registration distance' (Xr) from the pore constriction where kmer currents are measured, to the motor protein that sits atop the nanopore. The registration distance in DNA nanopore sequencing experiments using a more terminally located pore constriction in MspA and a different motor protein suggested an Xr of $\sim$ 20 nt (Elizabeth et al., 2012). In the R9.4.1 version nanopore system used here, a CsgG pore is used, which has a centrally located pore constriction, likely explaining the smaller Xr (Daniel et al., 2018). Interestingly, these observations suggest that at least in some sequence contexts, the motor protein kinetics are more sensitive to Nm than Ψ modifications. The 2′-*O*-methylation confers approximately -0.2 kcal/mol of stacking free energy to single stranded RNA and favors the

C3′-endo conformation of RNA (P, N, and M, 1974). Thus, changes in RNA conformation or steric and chemical interactions with particular amino acid residues of the motor protein may perturb translocation times. Interestingly, all types of modifications exhibit a median KS statistic peak for dwell time above background at relative modification position 0, indicating that transit times through the nanopore constriction itself are also perturbed by many of the modifications studied here. However, we did not observe significant changes in deletion or insertion error frequencies as might be expected from these dwell time differences. The majority of errors observed due to Ψ or Nm modifications were mismatches centered at the modification site, with Ψ having the highest average error of 48% mismatch.

3.2.3 RNA translocation rate is sensitive to nucleotide modifications and sequence composition

We next investigated the dwell time as a function of position to ascertain whether modifications and/or sequence content perturb dwell times. Comparison of dwell times for both native and IVT samples at a distance of Xr in the 3′ direction from selected sites within *E. coli* 16S (1402 m^4Cm), *E. coli* 23S (2552 Um), *S. cerevisiae* 18S (1428 Gm), and *S. cerevisiae* 25S (2220 Am) rRNAs revealed a significant increase (Mann-Whitney U test) in native samples relative to IVT controls (Figure 3.2A) To assess the extent of motor protein pausing due to sequence rather than the presence of modifications, we explored sequence similarities for the top 1% of ranked dwell times from all IVT samples in the region spanning the entire sequencing complex from upstream of the motor protein to the exit of the nanopore. We observed a strong guanosine

enrichment approximately 9-11 nt upstream of the pore constriction (Figure 3.2B), consistent with the registration distance of 10 nt, that corresponds to the location of the active site of the motor protein. We quantified the representation of each nucleotide in the motor protein active site, defined as the three nucleotides at positions 9-11 3′ of the nanopore constriction, across dwell time percentiles for all IVT samples and observed over-representation of guanosine in the highest percentiles (Figure 3.2C). Collectively these data indicate that even absent modifications, the motor protein used in direct RNA nanopore sequencing experiments has a tendency to pause on guanosine-rich sequences. These observations are consistent with pausing over guanosine-rich sequences by single-molecule picometer resolution nanopore tweezers (SPRNT) experiments with DNA (using a Hel308 based translocation mechanism) (Jonathan et al., 2019). Pausing over guanosine-rich regions may be a general feature of enzyme-based translocation; these regions may sterically hinder the enzyme or higher-than-average single-stranded stacking energies might decelerate translocation (Reid et al., 2015).

3.2.4 The 1-acetylimidazole reagent generates a compact SHAPE adduct

As nanopore sequencing was able to detect endogenous Nm modifications in rRNA, we hypothesized that we could detect 2′-*O*-adducts resulting from exposure of folded RNAs to electrophilic SHAPE reagents, which would then enable us to interrogate RNA structure using nanopore sequencing. A long-term advantage to exploiting nanopore sequencing for structural probing is the possibility of detection of multiple modifications per molecule,

enabling analysis of phased and correlated structural information over long sequence distances (Homan et al. 2014; Sengupta, Rice, and Weeks 2019). We initially attempted nanopore sequencing using RNA that had been modified with established SHAPE reagents 2-methylnicotinic acid imidazolide (NAI), 1-methyl-7-nitroisatoic anhydride (1M7), and *N*-methylisatoic anhydride (NMIA). In our hands, at the concentrations tested, readout of these experiments with nanopore sequencing resulted in few full-length reads and poor alignment accuracy as compared to an unmodified control sample. A similar poor alignment accuracy was reported recently when a short RNA modified with high concentrations of NAI (100 mM) was analyzed using nanopore sequencing.(Aw et al. 2020) We observed that the alignment percentage and fraction of full-length reads were poor at both high (200mM) and low (25mM) concentrations of NAI, precluding the detection of multiple modifications on single long RNA molecules. We hypothesized that the observed inefficient and incomplete translocation was due to the presence of multiple bulky 2′-*O*-aryl adducts that result from reaction of RNA with these SHAPE reagents.

We therefore searched for a reagent that would produce a smaller adduct, more chemically similar to native Nm modifications. We examined five carbonyl-imidazolide candidates, for structure-selective 2′-*O*-acylation. We detected covalent adduct formation for NAI, as expected, and for 1-acetylimidazole (AcIm). AcIm was previously identified as a 2′-hydroxyl-reactive reagent (Maryam et al., 2019). The proposed reaction of AcIm with the 2′- hydroxyl of RNA (Figure 3.3A) results in the most compact possible acetyl adduct using

an electrophilic carbonyl reagent.

The relative rate of reaction for a SHAPE reagent with the 2'-hydroxyl of RNA is mirrored by its rate of reaction with water (Edward et al., 2005). We investigated the timescale of AcIm reactivity by monitoring the change in absorbance of AcIm in reaction buffer; imidazole was monitored as a control. The AcIm signal, centered at 250 nm, decayed via a single exponential, with a half-life of 3 minutes at 37 °C (Figure 3.3B), consistent with prior measurements of N-acetylimidazole hydrolysis (Maryam et al., 2019; B, 1982). The AcIm spectrum decays to that observed for imidazole, supporting hydrolysis of AcIm into imidazole and, non-absorbing, acetate.

To assess AcIm reactivity with RNA, we extracted RNA from *E. coli* and treated total RNA with 100 mM NAI, 13 mM NMIA, 100 mM AcIm, or DMSO (as a vehicle control). We then obtained per-nucleotide reactivity profiles using mutational profiling (SHAPE-MaP) (Nathan et al., 2014) and aligned the resulting cDNAs to the 16S and 23S rRNAs (Figure 3.3C). Reactivity profiles for the three reagents were highly correlated, indicating that AcIm is a robust SHAPE reagent. AcIm reacted with each of the four canonical RNA nucleotides and preferentially reacted with conformationally flexible (unpaired) nucleotides (Figure 3.3D). Receiver operator characteristic curve analysis demonstrated that there was virtually no difference in discrimination between paired and unpaired nucleotides for NAI, NMIA, and AcIm (area under curve ~ 0.8 for all probes) (Figure 3.3E). In sum, AcIm generates small adducts, reacts broadly with all four ribonucleotides, and generates SHAPE-MaP reactivity profiles consistent with known reagents.

3.2.5 nanoSHAPE: Direct RNA nanopore sequencing of AcIm modified RNA

We next assessed whether AcIm chemical probing could be used to guide RNA secondary structure modeling based on single-molecule direct RNA nanopore sequencing, a method we call nanoSHAPE. We focused on an in vitro transcribed pri-miRNA transcript of the miR-17∼92 cluster, which spans 951 nucleotides and folds to form a series of well-defined hairpin structures (Steven et al., 2011; Saikat et al., 2012). The p ri-miR-17∼92 is predicted to form a moderately structured complex (predicted $\Delta G_{centroid} = -298.80$ kcal/mol) with a low number of isoenergetic suboptimal conformations (ensemble diversity, ED = 152.9), making it a suitable substrate for folding studies. We first assessed the structure of pri-miR-17∼92 by SHAPE-MaP using both AcIm and NAI. Reactivities for the two reagents were highly correlated (Spearman's rho = 0.78). Furthermore, secondary structure modeling informed by the AcIm SHAPE-MaP reactivity profile produced a centroid structure consistent with the current pri-miR transcript structure models (Steven et al., 2011). These experiments indicate that both chemical probes are suitable for investigating this structure and provide a control for comparison with nanoSHAPE. We next assessed the compatibility of AcIm with nanopore sequencing by performing a series of direct RNA nanopore sequencing experiments using either unmodified pri-miR-17∼92, or RNA modified with 5, 20, 50, 75, 100, 150 or 200 mM final concentrations of AcIm. Prior to AcIm modification, the terminal RNA nucleoside was modified through oxidation and beta-elimination to remove the 3'-nucleoside and leave a 3'-phosphate. After AcIm modification, the

3′-phosphate was removed by phosphatase treatment to leave a 3′-hydroxyl, allowing ligation with RNA sequencing adapters (Figure 3.4A). At higher AcIm modification rates, we observed noticeable decreases in read quality and fraction of full-length reads obtained. This signal degradation reduced the percentage of reads successfully aligned by Tombo. The coverage was higher at the 3′ end at all concentrations, consistent with the 3′ to 5′ read direction of the RNA through the nanopore. The coverage, fraction of full-length reads, and aligned read percentage became inadequate at the highest AcIm concentration (200 mM) so this condition was excluded from further analysis. The lower fraction of full-length reads observed after modification with AcIm suggested that reads were truncating at sites of AcIm modification. We mapped direct RNA nanopore sequencing read termini from unmodified RNA and RNA modified with 25 mM NAI or 150 mM AcIm to the pri-miR-17∼92 RNA to the parent sequence and compared to the SHAPE-MaP reactivity profile. The control and AcIm modified data had similar read termini profiles (Spearman's rho = 0.76), which in turn were similar to the SHAPE-MaP reactivity profile albeit shifted by approximately 10-15 nt in the 3′ direction, which implicates motor protein involvement in read truncation events. These observations suggest, first, that in the absence of SHAPE reagent modification, intrinsic RNA structures cause a small degree of truncation and, second, that modification with AcIm slightly accentuates read truncations in these same regions. Read termini in NAI-modified RNA had a different profile (Spearman's rho = 0.12 and 0.19 for control and AcIm respectively) with most termini mapping to the 3′ end of the pri-miR-17∼92 sequence, consistent with the very small

fraction of full-length reads obtained from NAI-modified sequencing experiments. Thus, NAI is a poor nanoSHAPE probe as these adducts pose serious problems for motor protein processing and processivity.

We next performed KS statistical testing for current and dwell time distributions across all AcIm concentrations as compared to the unmodified control. KS peaks in current were observed in all profiles primarily in single-stranded regions of pri-miR-17∼92. We determined Spearman's rank order correlations for both KS of current and KS of dwell time (shifted by Xr) against the AcIm SHAPE-MaP profile. The correlation was greatest for current, and maximized at 150 mM (current rho $= 0.51$, dwell time rho $= 0.28$). To assess the capability of single-molecule-based reconstruction of reactivity profiles, we performed per-read statistical testing and normalization for every nucleotide position within 1000 individual full-length molecules of pri-miR-17∼92 across all AcIm concentrations (Figure 3.4B,C) The median mutation rates from SHAPE-MaP libraries derived from pri-miR-17∼92 modified with 25 mM and 200 nM AcIm were 0.03% and 0.1%, respectively, which corresponds to 0.285 and 0.951 detected adducts per full-length read, respectively. These low values suggest that the current MaP approach does not detect the AcIm adduct efficiently. Based on the single molecule nanopore data, and after statistical testing and detection, we obtained approximately 105 called modification sites per full-length read at 150 mM AcIm (Figure 3.4D). This value is likely notably inflated due to the distributed nature of modification detection and due to the high level of noise intrinsic to current generation nanopores. Nonetheless, these comparisons suggests that nanopore detection is more efficient at detecting

the compact 2′-ribose AcIm adduct than is MaP-RT.

We next examined the number of single-molecule reads which are required to obtain an optimal correlation with SHAPE-MaP. We sub-sampled full-length reads from n = 1 to n = 1000 and calculated the normalized reactivity profiles from the per-read current statistical testing as a function of number of reads. The Spearman's rank correlation of the normalized reactivity against the SHAPE-MaP reactivity profile reached 95% of the maximum correlation at around 200 reads for each AcIm concentration, with higher correlations with MaP data obtained at higher AcIm concentrations (Figure 3.4E). In the normalized reactivity profile derived from nanoSHAPE, we observe less distinctive reactivity features closer to the 5′ end of the pri-miR-17∼92 transcript. To explore this phenomenon, we calculated the Spearman's rank correlation on a progressively shortened normalized reactivity profile, trimming from the 5′ end. This procedure revealed a maximum correlation at about 300 nucleotides from the 5′-end of the transcript (rho = 0.53, 150mM AcIm) (Figure 3.4F), indicating that the 5′ end of RNAs may not be well resolved by this approach. Poor structural resolution at the 5′ end may be due to reduced coverage in this region resulting from incomplete reverse transcription due to the presence of AcIm adducts that can cause cDNA truncation. Reverse transcription is not strictly required for direct RNA sequencing using the nanopore platform; however, the presence of a cDNA is known to stabilize the sequenced RNA strand increasing overall yield and throughput, potentially favoring the recovery of RNA molecules with a full-length cDNA annealed.

3.2.6 nanoSHAPE facilitates RNA structure modeling

We performed secondary structure modeling using nanoSHAPE and SHAPE-MaP reactivities as pseudo-free energy constraints introduced into a nearest-neighbor RNA folding algorithm (Katherine et al., 2009; Kevin and 2021, n.d.). The two structural models differ in that nanoSHAPE centroid structure includes fewer long-range base pairs and has larger loop sizes (specifically for hairpins 17 and 19a) than does the SHAPE-MaP-based structure (Figure 3.5A). Hairpin 17 and 19a are both predicted to have internal bulges (U151 and G435-U437) at the bases of their loops towards the 3′ side, based on both SHAPE-MaP and unconstrained modeling (Saikat et al., 2012). Reactivity at these positions, observed by nanoSHAPE, which features a 3′ to 5′ read direction, may over-detect reactivity at loop closing base pairs, leading to prediction of larger loop sizes**.** Importantly, centroid structures for both SHAPE-MaP- and nanoSHAPE-constrained predictions contain the six miRNA hairpins expected to occur in the 17∼92 cluster.

We next performed partition function calculations for RNA structures arising from SHAPE-MaP-constrained and nanoSHAPE-constrained pri-miR-17∼92 sequences. Partition function base pair probabilities between all possible nucleotides (*i*,*j*) exhibited generally concordant connectivity patterns indicating broad agreement between the collective predicted structural ensembles (Figure 3.5B). We then benchmarked minimum free energy and centroid secondary structures predicted from nanoSHAPE versus predictions from SHAPE-MaP constrained modeling relative to models obtained with no probing data (NPD). In general, MaP-constrained models were the most distinct,

consistent with extensive prior work showing SHAPE-MaP data generally substantially change RNA structure models relative to no-probing-data models, in the direction of the correct structure (Nathan et al., 2014; Kevin and 2021, n.d.). nanoSHAPE data also clearly perturbed the structural ensemble, relative to the no-probing-data ensemble, to become more similar to the SHAPE-MaP-informed model (Figure 3.5C). We conclude that nanoSHAPE produces reactivity patterns and secondary structure predictions for the pri-miR-17∼92 sequence broadly consistent with high-throughput sequencing based RNA chemical probing and structural profiling approaches.

3.3 Discussion

The long-read direct RNA nanopore sequencing on the ONT platform is a promising tool for characterizing RNA at the single-molecule level. RNA molecules exhibit diverse chemical and structural states that serve as effectors and modulators of RNA function, interaction, and dynamics. We sought to use the direct measurement of RNA, rather than a cDNA copy, to examine RNA chemical modifications and secondary structure. Our direct RNA sequencing approach was able to detect native modifications in ribosomal RNA from *E. coli* and *S. cerevisiae* at both the nucleobase (Ψ) and backbone (Nm). The majority of these positions are modified stoichiometrically, making them good systems for benchmarking endogenous modification detection. In our dataset comparing endogenous rRNAs to in vitro transcribed controls, we performed raw signal to sequence alignment with both Tombo (Marcus et al., 2016) and Nanopolish (Jared et al., 2017) and identified rRNA modification positions to

within $\pm$ 2 nt of known modified sites.

Direct RNA nanopore sequencing is uniquely positioned to answer questions about the dynamics and ordering of modification installation on rRNA and, in principle, has the potential to address both quantification and long-range phasing of modifications. Signal discrimination remains an outstanding challenge for direct RNA nanopore sequencing for modification detection, which is a function of the kmer sequence context and ability to align raw current signal to sequence. Development of training sets consisting of known modifications in all possible kmer sequence contexts will be required for RNA modification identification without resorting to comparison with an IVT control. The ability to call modifications without an IVT control, coupled with increasing yield of direct RNA sequencing should allow investigation of other, less abundant cellular mRNAs and long non-coding RNAs.

In addition to current signal levels, we characterized changes in the current level dwell times in direct RNA sequencing. Ribose modifications, both the endogenous Nm, and exogenous SHAPE reagent modifications, notably extend dwell times. Dwell time changes due to RNA modifications at the pore constriction have been recently reported (Leger et al. 2019). Here we demonstrated that dwell time is dependent on motor protein translocation kinetics mediated at a registration distance (Xr $\sim$10 nt in the 5′ direction) from the pore constriction. Additionally, we observed that motor protein dwell times are influenced by primary sequence. Translocation rates have been shown to be sequence dependent for DNA using the Hel308 motor protein and a MspA nanopore (Jonathan et al., 2019). We found that average translocation

rates varied across the nanopore array complicating direct comparison of dwell times across array channels. However, we suggest that large dwell times observed in direct RNA nanopore sequencing experiments may be used to infer sites (at Xr distance) of Nm or Ψ modifications in the absence of IVT controls, provided that the sequence context around the putative modification is not G-rich. Full characterization and incorporation of this extra dimension of information will require channel- or even read-specific normalization to faithfully compare translocation rates across the nanopore array.

The detection of naturally occurring Nm modifications in rRNA suggested the possibility of applying this approach to detect experimentally introduced 2'-*O* adducts, as used in RNA structure probing methods. Key to our success was identifying a SHAPE reagent specifically tailored for nanopore sequencing. AcIm had favorable properties including a short (but still experimentally manageable) half-life, small adduct size, detection by mutational profiling, and commercial availability. The small adduct (2'-*O*-acetyl) created by modifying RNA with AcIm is detectable in direct RNA nanopore sequencing experiments. High rates of AcIm modification do lower the number of full-length reads, overall yield, and alignment rates; however, the resulting yield and data quality are vastly superior to that obtained with reagents that yield bulkier adducts. Consistent with our analysis, an NAI analog (NAI-N3) was independently (Jong et al., 2020) found to induce drastic decreases in yield. In that work, NAI-N3 yielded hit-rates of 1-2% over the readable fraction of RNA, whereas AcIm achieved median hit-rates up to 11% on full-length single molecules. Application of nanoSHAPE with AcIm to the analysis

of the structure of the pri-miR-17∼92 transcript revealed that nanoSHAPE-data-constrained modeling yielded RNA structures broadly similar to those obtained with SHAPE-MaP data.

3.4 Limitations of the study

Direct RNA nanopore sequencing, and by extension nanoSHAPE, have limitations. The method requires high concentrations of target RNA and a free 3′-hydroxyl for ligation to the 5′ phosphate of the first adapter required for nanopore sequencing. Electrophilic SHAPE reagents, AcIm included, covalently modify the 3′-hydroxyl, preventing ligation. We ameliorated this challenge by chemical treatment to create a terminal phosphate; after modification, the RNA was then treated with a phosphatase to enable poly(A) tailing and ligation to the sequencing adapter. If nanoSHAPE is to be properly extended transcriptome-wide and to in-cell structural probing experiments with high single molecule modification rates, novel methods for selection and enrichment of target RNAs and protection of the 3′-hydroxyl (or enrichment of molecules with ligate-able 3′-hydroxyl ends) in cells may be required to ensure sufficient yield with the current direct RNA nanopore sequencing method.

nanoSHAPE is also limited by the poor resolution of reactivity profiles at the 5′ ends of longer RNA molecules. This loss of resolution is likely due to multiple factors including coverage bias inherent to the 3′-to-5′ direction of RNA nanopore sequencing and the difficulty of translocation through a highly structured RNA, like pri-miR-17∼92. It is also possible that cDNA synthesis on highly AcIm-modified RNA, which is an optional step that facilitates

translocation, was incomplete in our experiments.

Finally, nanoSHAPE is limited by the method of signal analysis used to identify intrinsic posttranscriptional modifications and SHAPE adduct sites. In this work, we used a comparative approach, comparing current signals of modified RNAs to those of unmodified RNAs of the same sequence to identify sites of difference. Modification detection may be improved by using methods that employ trained models for signal classification. However, de novo methods are reliant on a training or ground truth set containing the modification in all kmer sequence contexts. A second challenge in adduct detection is distinguishing authentic chemical modifications from current signal and dwell time changes induced by the underlying RNA structure. Extensive benchmarking with native RNAs of known structure would inform deconvolution of adduct versus structure effects.

Despite these challenges, nanoSHAPE demonstrates significant promise. Long-read single-molecule sequencing will permit investigation of RNA structural ensembles for long RNAs directly. Adduct detection and sequencing throughput are poised to improve as direct RNA nanopore sequencing technology and analyses mature. Direct sequencing of AcIm-modified RNA will be crucial to deciphering RNA energy landscapes, alternative folding pathways, and phasing of distal RNA structural elements.

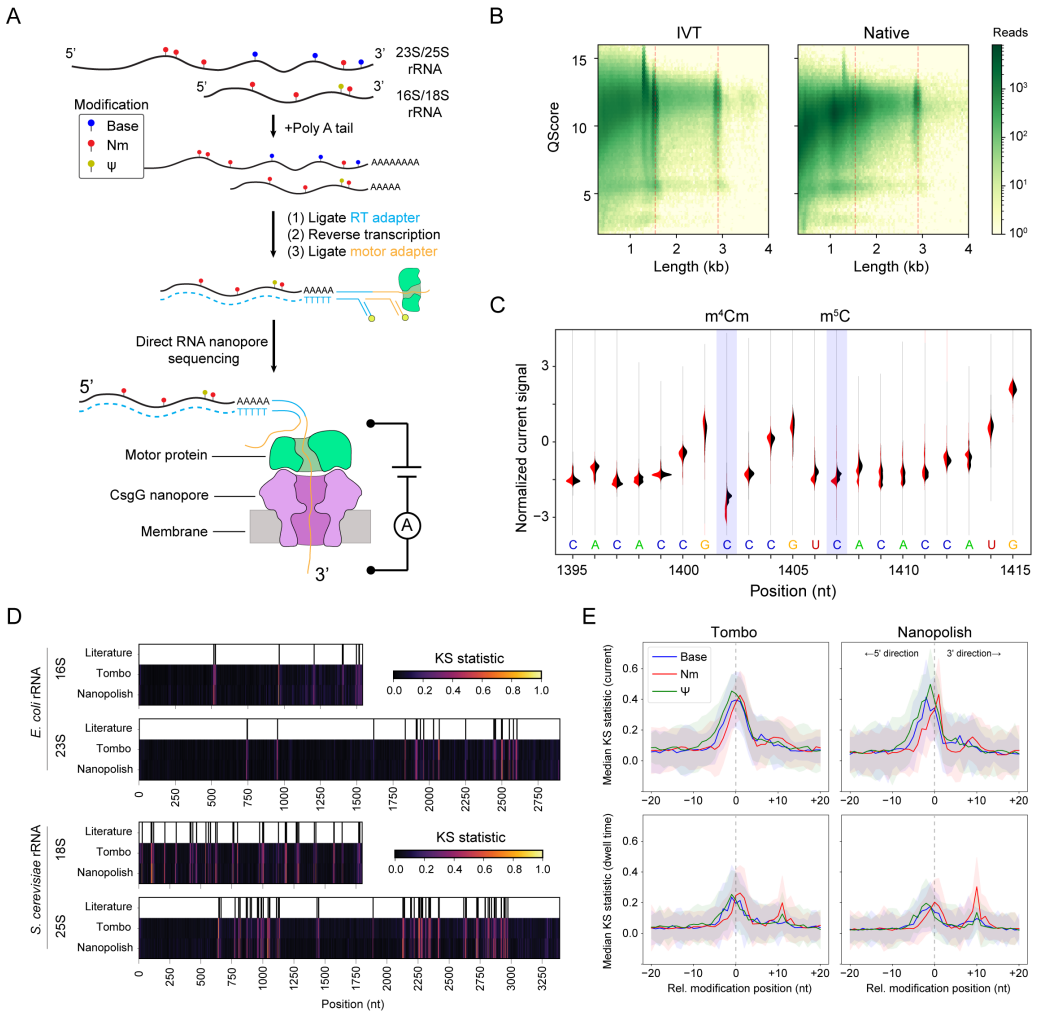

Figure 3.1: *Direct RNA nanopore sequencing and modification detection. A) Scheme for direct RNA nanopore sequencing. rRNAs containing native modifications are poly(A) tailed before ligation of adapters and RT. Ionic current blockage events are characteristic of the kmer sequence of RNA transiting through the pore constriction.* ***B)*** *Read quality heatmap for IVT rRNA (left) and native rRNA (right) from E. coli. Dashed red lines indicated the expected lengths for 16S rRNA (1.5 kb) and 23S rRNA (2.9 kb)* ***C)*** *Normalized native (red) and IVT (black) current signal alignment for 16S rRNA from E. coli spanning positions 1395 - 1415 performed using Tombo. Sites of known modifications within this window are highlighted in blue.* ***D)*** *Positional Kolmogorov-Smirnov (KS) statistical testing of current signals across rRNA from E. coli and S. cerevisiae using both Tombo and Nanopolish. Modification positions described in the literature are indicated as black lines.* ***E)*** *Median current and dwell KS statistic profiles separated by modification type (base (excluding Ψ), blue; 2'-O-methyl, red; and Ψ, green) and aligned by modification position from both Tombo and Nanopolish. Colored shaded regions represent the standard deviation of the KS statistic.*

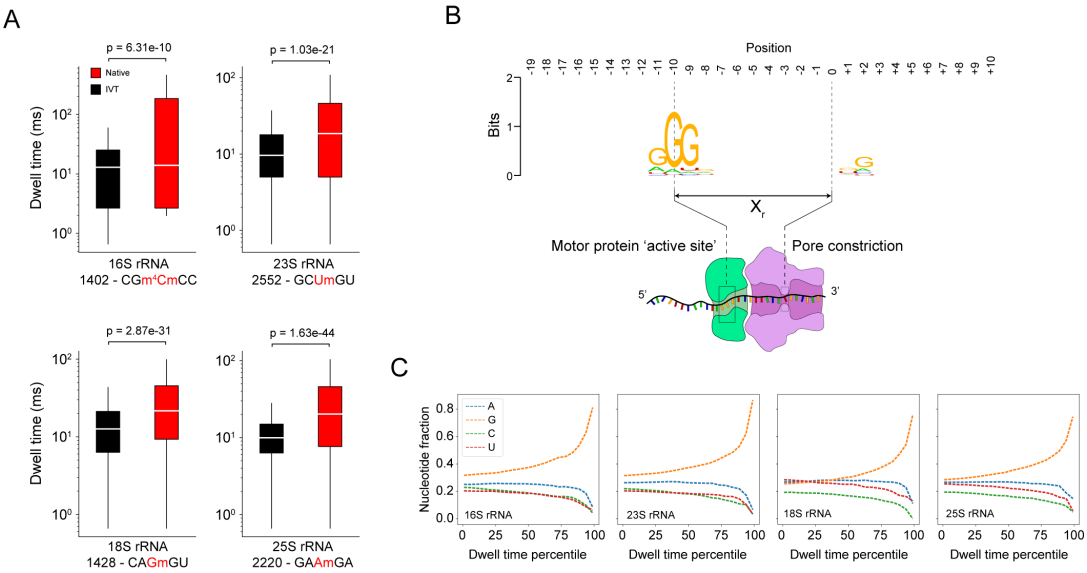

Figure 3.2: *Dependence of dwell time on modifications and sequence. A) Dwell time comparison of native (red) and IVT (black) samples (Mann-Whitney U test, $n = 1000$ reads) at selected Nm modification sites in 16S and 23S from E. coli and 18S and 25S from S. cerevisiae. The kmer is indicated below the x-axis along with the position of the modification (red). Dwell times are from +Xr from the centered modification site kmer.* ***B)*** *Sequence motif from the top 1% of dwell times (IVT samples only, 16S, 23S, 18S, and 25S) spanning a 30-nucleotide window encompassing the entire biomolecular sequencing complex.* ***C)*** *Nucleotide representation (fraction) within the trimer (positions -9, -10, and -11) from IVT samples as a function of dwell time percentile. The highest dwell time percentiles are enriched for guanosine within the trimer.*

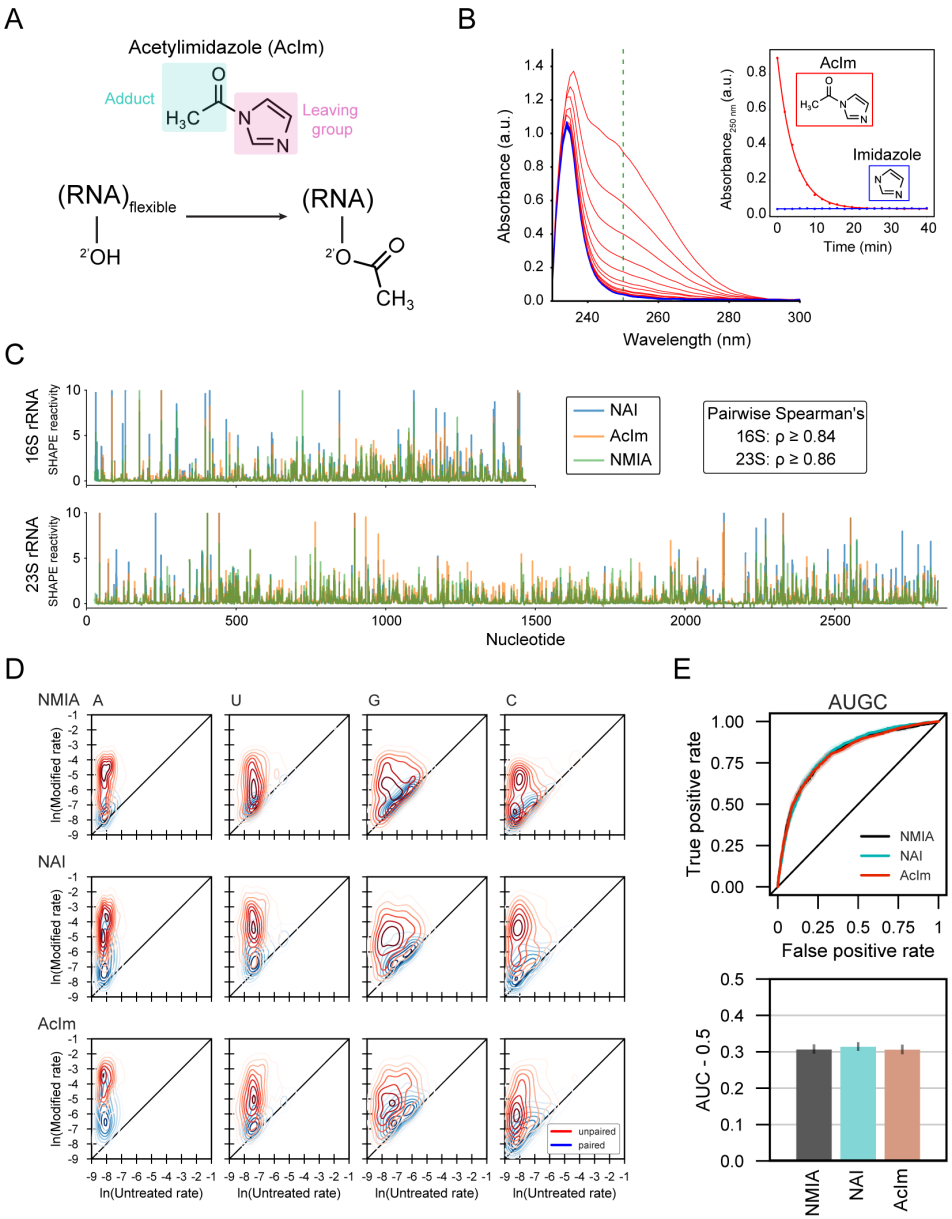

Figure 3.3: *Acetylimidazole generates small adducts detectable by SHAPE-MaP. A) Schematic of acylation of RNA with acetylimidazole (AcIm).* ***B)*** *Time dependence of hydrolysis for AcIm analyzed by changes in UV absorbance. Experiment performed at 37 °C.* ***C)*** *SHAPE-MaP reactivity profiles using NMIA, NAI, and AcIm for E. coli 16S and 23S rRNAs.* ***D)*** *Two-dimensional kernel density estimates for NMIA, NAI, and AcIm adduct-induced mutation rates for E. coli 16S and 23S rRNAs with unmodified control and SHAPE-modified rates on the x- and y-axes, respectively.* ***E)*** *Receiver operator characteristic curve and associated area under the curve (AUC) for MaP reactivities as a function of pooled nucleotide base pairing status for SHAPE reagents.*

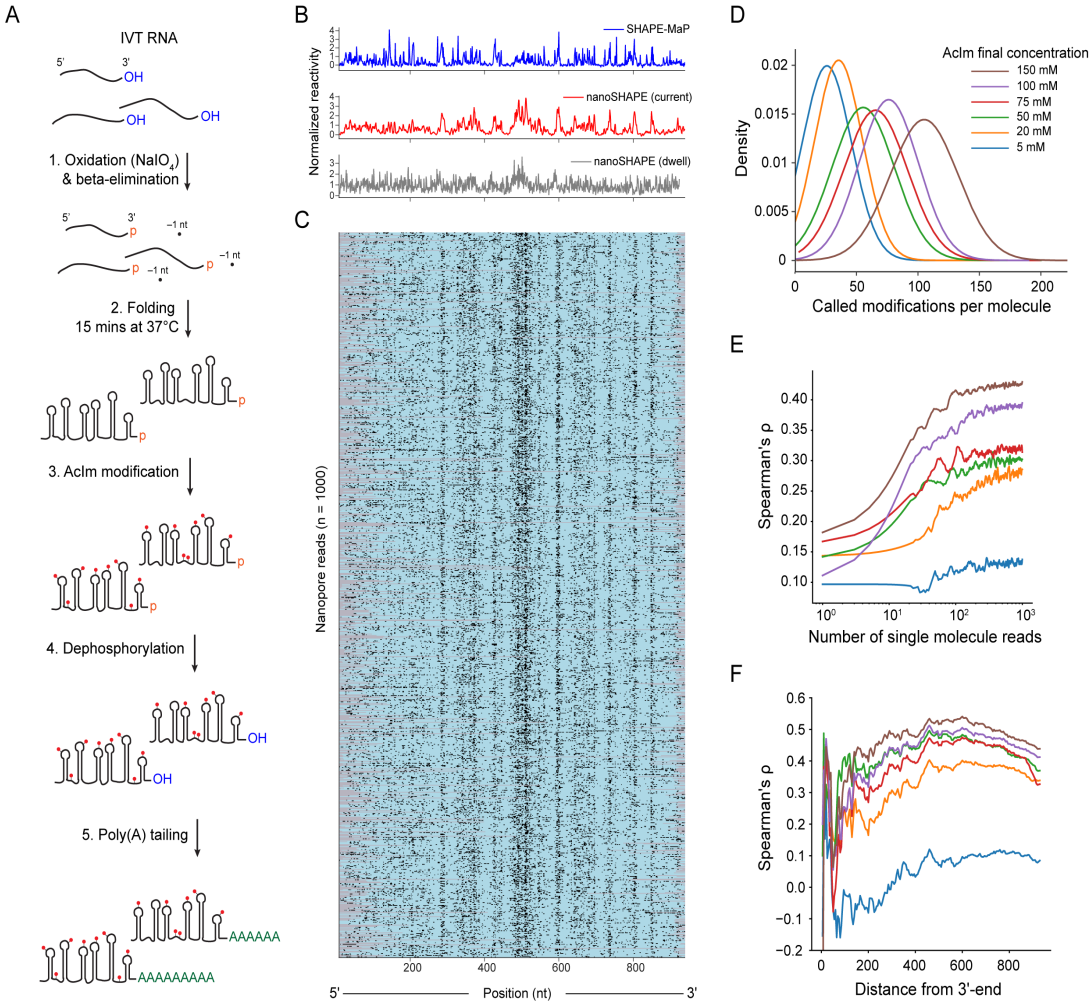

Figure 3.4: *Direct structural probing of a pri-miR-17∼92 transcript RNA using AcIm and nanopore sequencing. A) Scheme for 3′-end modification, SHAPE probing, and 3′ end tailing.* ***B)*** *Normalized SHAPE-MaP reactivity (blue), and nanoSHAPE reactivity detected by changes in current (red) and dwell time (grey) for the pri-miR-17∼92 RNA.* ***C)*** *Heatmap of 1000 nanopore reads of pri-miR-17∼92 modified with 150 mM AcIm. Modifications were determined by per-nucleotide Student's t-test using Fisher's method context of ± 1 of the current signal. Per-nucleotide p-values were corrected using the Benjamini-Hochberg procedure and binarized. Nucleotides scored as modified and unmodified are shown in black and teal, respectively; unmapped regions are grey.* ***D)*** *Kernel density estimate of the number of called modifications per pri-miR-17∼92 molecule as a function of AcIm concentration. Called modifications correspond to an upper limit.* ***E)*** *Spearman's rank order correlation (rho) between nanoSHAPE and SHAPE-MaP as a function of the number of contributing pri-miR-17∼92 molecules across the AcIm concentrations tested.* ***F)*** *Spearman's rho as a function of the distance from the 3′-end of the pri-miR-17∼92 RNA.*

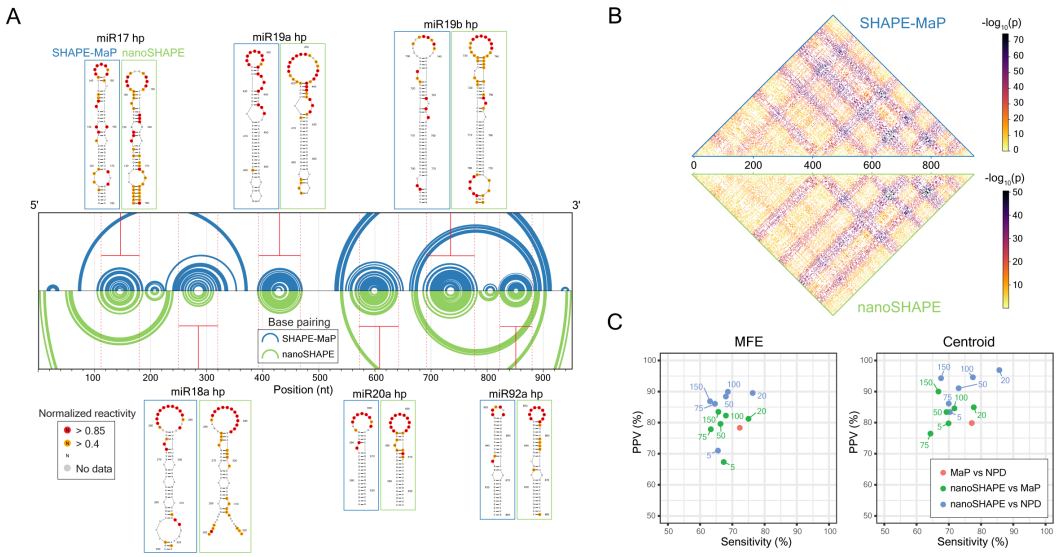

Figure 3.5: *Comparison of RNA structure modeling based on SHAPE-MaP and nanoSHAPE reactivities. A) Secondary structure models, visualized as arc diagrams, for the pri-miR-17∼92 RNA. Centroids for structures modeled using SHAPE-MaP (blue) or nanoSHAPE (green) constraints are shown. Secondary structure models for the constituent miRNA hairpins are shown with overlaid SHAPE-MaP and nanoSHAPE reactivities. SHAPE data correspond to the 25 mM AcIm concentration.* ***B)*** *Probabilities of all possible base pairs in pri-miR-17∼92 based on SHAPE-MaP (top) or nanoSHAPE (bottom) reactivity constraints, shown as a partition function dot plot. Probabilities are displayed as -log$_{10}$(probability base pair (i,j)).* ***C)*** *Similarity in structure models, reported as relative positive predictive value (PPV) and sensitivity for minimum free energy (MFE) and centroid structures. Pairwise comparison was performed between SHAPE-MaP constrained (MaP), nanoSHAPE constrained, and no probing data (NPD) secondary structure models.*

3.5 Methods

3.5.1 E.coli and S.Cerevisiae

E. coli (K-12 MG1655) cells were grown in an overnight culture at 37 °C in 5% CO_2 in 10 mL of freshly prepared Luria Bertani broth (LB) or 10 mL M9 minimal salts. 75 mL of pre-warmed media was inoculated with 1 mL of overnight culture. *E. coli* cells were grown to $OD_{600} = 0.5$, typically over 3 - 4 hours at 37 °C. *S. cerevisiae* (S288C), colonies were picked from an agar plate and incubated at 30 °C in ∼7 mL of YPD broth for two days.

3.5.2 rRNA extraction (E. coli and S. cerevisiae)

25 mL of *E. coli* cells were pelleted at 3280 g at 4 °C for 12 minutes. Cells were lysed in 16.5 mL lysis buffer (15mM Tris pH 8, 450 mM sucrose, 8 mM EDTA, 0.4 mg/ml lysozyme) for 5 minutes at room temperature then 10 minutes at 4 °C. The pellet was collected at 3280 g for 5 minutes and then resuspended in 2 mL proteinase K buffer (50 mM HEPES, 200 mM NaCl, 5 mM $MgCl_2$, 1.5% SDS, 0.2 mg/mL Proteinase K). The solution was vortexed for 10 seconds and incubated at room temperature for 5 minutes, and then 4 °C for 10 minutes. Nucleic acids were extracted twice with 1 volume of phenol:chloroform:isoamyl alcohol (25:24:1) followed by two subsequent chloroform extractions prior to ethanol precipitation and resuspension in 88 μl RNAse-free water. Purified nucleic acids were treated with Turbo DNAse (10 μl Turbo DNAse buffer [10X] and 2 μl Turbo DNAse) at 37 °C for 1 hour. Finally, RNA was purified with 0.8X vol AmpureXP beads. For *S. cerevisiae*,

RNA extraction was carried out using the YeaStar RNA kit (Zymo Research) according to instructions.

3.5.3 Generation of rRNA IVT controls

gDNA was extracted from *E. coli* using the method described for RNA above up to and including the ethanol precipitation step. After resuspension in 88 μl RNAse-free water, purified nucleic acids were treated with 1-5 μl 1 mg/ml RNaseA (Qiagen) for 45 minutes at 37 °C to degrade RNA. gDNA was then purified with 0.5X SPRI or subsequent ethanol precipitation. gDNA was purified from *S. cerevisiae* using the YeaStar Genomic DNA kit (Zymo Research) according to instructions. Primers for amplifying rDNA amplicons, which include a T7 transcription promoter for subsequent in vitro transcription (IVT) are detailed in Table S2. Amplicons were generated by PCR using Kapa HiFi DNA polymerase and purified by SPRI. T7 transcription templates were transcribed using HiScribe T7 Quick High Yield RNA Synthesis Kit (New England Biolabs). Reactions were cleaned up using the MEGAClear Transcription Clean-up Kit (Invitrogen) before nanopore sequencing library preparation.

3.5.4 Poly(A) tailing of RNA

Oxford Nanopore Technologies direct RNA sequencing requires a poly(A) tail for first adapter ligation. Both rRNA and pri-miR-17∼92 (951 nt) samples were poly(A) tailed. Briefly, 0.5 – 1.0 ug of RNA was poly(A) tailed with 2 ul ATP [10mM], 1 ul [5U] of *E. coli* Poly(A) polymerase (EPAP) (New England Biolabs)

and 2 ul EPAP reaction buffer [2X] in a final volume of 20 ul. Reactions were carried out for 15 minutes and quenched with 0.5 ul of 0.5 M EDTA before purification with 1X SPRI.

3.5.5 Nanopore library preparation

Direct RNA sequencing was performed using the Oxford Nanopore Technologies kit (SQK-RNA002) as directed with the RCS control RNA. Sequencing was performed on the MinION device using either standard flowcells (FLO-MIN106D) for rRNA experiments or a mixture of standard and flongle flowcells (FLO-FLG001) for pri-miR-17∼92 experiments. Sequencing was carried out until the number of active nanopores dropped below 5% of the initial total number of pores, typically 12-36 hours.

3.5.6 Reagents

All standard laboratory reagents, including AcIm, were purchased from Millipore-Sigma, with the exception of NMIA purchased from Invitrogen/Thermo Fisher Scientific. NAI was synthesized from 2-methylnicotinic acid and 1,1'-carbonyldiimidazole, as described. (Robert et al., 2012) Briefly, 137 mg (1mmol) 2-methylnicotinic acid was dissolved in 0.5 mL anhydrous DMSO. A solution of 162 mg (1 mmol) 1,1'-carbonyldiimidazole in 0.5 mL anhydrous DMSO was added dropwise over 5 min. The resulting solution was stirred at room temperature using a PTFE coated micromagnet until gas evolution was complete and then stirred at room temperature for 1 h further. The resulting solution was used as a 1.0 M stock solution (assuming complete conversion) containing

a 1:1 mixture of the desired compound and imidazole. The NAI stock solution was aliquoted and frozen at -80 °C when not in use. The reagent is stable for several months if stored in anhydrous DMSO at -80 °C. The stock solution should be warmed to room temperature prior to opening.

3.5.7 AcIm hydrolysis

AcIm hydrolysis was tracked at 37 ºC by time resolved UV absorbance using a Nanodrop 2000 spectrophotometer in [1x] modification buffer (100 mM HEPES pH 8.0, 100 mM NaCl, 10mM $MgCl_2$) every 2 minutes for 40 minutes. Imidazole spectra were collected every 2 minutes in [1x] modification buffer for 40 minutes.

3.5.8 SHAPE-MaP on rRNA

Extracted rRNA was treated with NAI [100 mM final], AcIm [100 mM final], NMIA [13 mM final] or DMSO (unmodified control). All SHAPE-MaP experiments were performed with 10% volume fraction of DMSO. Modification was carried out at 37 °C for 3 half-lives of the chemical probe used. For mutational profiling RT, 1 μl of nonamer primer [200 ng/μl or 2μM] was added to 1-3 μg of rRNA in 10 μl nuclease free water. The samples were incubated at 65 °C for 5 minutes then cooled on ice. 8 μl of [2.5x] MaP buffer (125 mM Tris pH 8.0, 187.5 mM KCl, 15 mM $MnCl_2$, 25 mM DTT, and 1.25 mM dNTPs) was added and incubated at 42 °C for 2 minutes. 1 μl of SuperScript II reverse transcriptase was added and mixed well before incubating the reaction at 42 °C for 2-3 hours, and then at 70 °C to inactivate the polymerase. cDNA

was exchanged into water using G-50 columns (GE Life Sciences) the volume increased to 68 μl using nuclease free water. Second strand synthesis was carried out (Second Strand Synthesis Enzyme mix; New England Biolabs) and the dsDNA was used to generate a Nextera library for sequencing on an Illumina MiSeq, as described.(Smola et al. 2015)

3.5.9 RNA modification (pri-miR-17 92)

In order to protect the 3'-OH of pri-miR-17∼92 RNA from modification with acylating reagents, the terminal 3' nucleotide was oxidized followed by a beta-elimination reaction to remove the terminal nucleotide leaving a terminal phosphate. Then RNA modification was carried out prior to dephosphorylation and nanopore library preparation. Briefly, pri-miR-17∼92 was incubated at 37 ºC for 30 minutes with shaking in oxidation buffer ($NaIO_4$ [20mM], Lysine-HCl [200mM] pH 8.5, final volume: 40 μl). The reaction was quenched with 2 μl of ethylene glycol then purified using 1x SPRI, eluting into beta-elimination buffer (Sodium borate [33.75mM], boric acid [50mM], pH 9.5) incubating at 45 ºC for 45 minutes. RNA was again purified by 1x SPRI. 1-2.5 μg of IVT pri-miR-17∼92 RNA was diluted into 7 μl water and heated to 95 ºC for 2 min and immediately placed on ice (2 min). 6μl of folding buffer [3.3x] (333 mM Tris-HCl pH 8.0, 333 mM NaCl and 33 mM MgCl2) and 5 μl HEPES pH 8.0 [200 mM] were added and the RNA was allowed to fold for 20 min at 37 ºC. 2 μl of DMSO (control) or SHAPE reagent (NAI or AcIm) were added to a new tube, then folded RNA was added and mixed by pipetting. Modification was carried out for at least 3 half-lives at 37 ºC. RNA was then

dephosphorylated by adding 22 μl RNAse-free water, 5 μl Antarctic Phosphatase reaction buffer [10x], 2 μl Antarctic Phosphatase [5k U/mL] (NEB) and 1 μl RNase inhibitor and incubating at 37 ºC for 30 minutes with shaking. The phosphatase was inactivated by incubating the reaction at 65 ºC for 5 minutes. Finally, the RNA was purified using 1x SPRI prior to poly(A) tailing.

3.5.10 SHAPE-MaP on pri-miR-17 92

After pri-miR-17∼92 RNA was dephosphorylated, mutational profiling reverse transcription (RT) was performed. 1 μl of nonamer primer [200 ng/μl or 2μM] was added to 1-3 μg of RNA in 10 μl nuclease free water. The samples were incubated at 65 °C for 5 minutes then cooled on ice. 8 μl of [2.5x] MaP buffer (125 mM Tris pH 8.0, 187.5 mM KCl, 15 mM $MnCl_2$, 25 mM DTT, and 1.25 mM dNTPs) was added and incubated at 42 °C for 2 minutes. 1 μl of SuperScript II reverse transcriptase was added and mixed well before incubating the reaction at 42 °C for 2-3 hours, and then at 70 °C to inactivate the polymerase. cDNA was exchanged into water using G-50 columns (GE Life Sciences) the volume increased to 68 μl using nuclease free water. Second strand synthesis was carried out (Second Strand Synthesis Enzyme mix; New England Biolabs) and the dsDNA was used to generate a Nextera library for sequencing on an Illumina MiSeq, as described.(Smola et al. 2015)

3.5.11 Nanopore data processing

Multi-fast5 reads were basecalled using guppy (v3.1.5). Base called multi-fast5 reads were then converted to single read fast5s using the Oxford Nanopore

Technologies API, ont_fast5 (v1.0.1). Fastqs were mapped to their respective transcriptomes for *E. coli* (NC_000913.3.fa) and *S. cerevisiae* (R1-1-1_19960731.fsa) using minimap2 (v2.11).

3.5.12 Nanopolish and Tombo analysis of data

Tombo (v1.5.1) and Nanopolish (v0.11.1) were both used to detect native modifications in rRNA datasets as well as detect modifications deposited from SHAPE reagents. Comparisons were performed between native and IVT samples for the rRNA datasets and between modified (at indicated concentrations) and unmodified samples for pri-miR-17$\sim$92. Nanopolish *eventalign* module was used to align current intensities and dwell times to reference sequences. Kolmogorov–Smirnov (KS) statistical testing was performed in order to detect modified nucleotides. Using Tombo, raw signal squiggles were assigned to reference sequences using *resquiggle*. Next, modified base detection was carried out using the *detect_modifications model_sample_compare* method. Per-read statistical testing (AcIm modified RNA) was performed with a $\pm$ 1 nucleotide Fisher's method context adjustment. The requisite text output was obtained using *text_output browser_files* method. Reactivity profiles from Tombo per-read statistical testing were further adjusted using Benjamini-Hochberg procedure for multiple testing. Adjusted per-read reactivity profiles were used to calculate percentage modification per genomic position. This percentage profile was then normalized using the normalization procedure described in SHAPE-MaP method.(Smola et al. 2015) Single molecule positional current, standard deviation of current, and dwell time data were extracted

as numpy arrays directly from single read fast5 data using custom written python scripts.

3.5.13 RNA structure modeling

Centroid structures and free energies were obtained using the RNAfold (v2.4.13) (Vienna) web server. Options were to avoid isolated base pairs and temperature = 37 ºC. R-chie (Lai et al. 2012) was used for displaying base pairing (arc) of centroid structures. The RNAstructure (v6.2) software suite (Reuter and Mathews 2010) was used for partition function calculation and associated dot plot visualization. The following options were used for partition function calculation: maximum percent energy difference = 10%, maximum number of structures = 50, window size = 3, temperature = 37 ºC. PPV and sensitivity were determined by performing pairwise comparison between the respective minimum free energy (MFE) and centroid structures of miR-17∼92 for sequence alone prediction, SHAPE-MaP and nanoSHAPE constrained experiments. CT files from constrained and unconstrained RNAfold predictions were used as input to the RNAstructure scorer function to determine PPV and sensitivity. PPV corresponds to the percentage of predicted base pairs that are in the “accepted" structure and sensitivity corresponds to the percentage of known base pairs correctly predicted in the “accepted" structure.

3.6 References

Andrew, Smith, Jain Miten, Mulroney Logan, Garalde Daniel, and Akeson Mark (2019). "Reading Canonical and Modified Nucleobases in 16S Ribosomal RNA Using Nanopore Native RNA Sequencing". In: *PloS One* 14 (5), e0216709. DOI: 10.1371/journal.pone.0216709.

Anthony, Mustoe, Steven Busan Greggory, Rice Christine, Hajdin Brant, Peterson Vera, Ruda Neil, Kubica Razvan, Nutiu Jeremy, Baryza Kevin, and Weeks (2018). "Pervasive Regulatory Functions of MRNA Structure Revealed by High-Resolution SHAPE Probing". In: *Cell* 173 (1), pp. 181–195. DOI: 10.1016/j.cell.2018.02.034.

B, Gour-Salin (1982). "Hydrolysis Rates of Some Acetylimidazole Derivatives". In: *Can. J. Chem* 61, p. 2059.

Bianca, Gonzales, Henning Dale, So Rolando, Dixon Jill, Dixon Michael, and Valdez Benigno (2005). "The Treacher Collins Syndrome (TCOF1) Gene Product Is Involved in Pre-RRNA Methylation". In: *Human Molecular Genetics* 14 (14), pp. 2035–2043. DOI: 10.1093/hmg/ddi208.

Daniel, Eyler, Monika K, Franco Zahra, Batool Monica, Wu Michelle, Dubuke Malgorzata, Dobosz-Bartoszek Joshua, Jones Yury, Polikanov Bijoyita, Roy Kristin, and Koutmou (2019). "Pseudouridinylation of MRNA Coding Sequences Alters Translation". In: *Proceedings of the National Academy of Sciences of the United States of America* 116, pp. 23068–23074. DOI: 10.1073/pnas.1821754116.

Daniel, Garalde, Elizabeth A, Snell Daniel, Jachimowicz Botond, Sipos Joseph, Lloyd Mark, Bruce Nadia, and Pantic (2018). "Highly Parallel Direct RNA Sequencing on an Array of Nanopores". In: *Nature Methods* 15 (3), pp. 201–206. DOI: 10.1038/nmeth.4577.

Deepak, Patil, Brian F, Pickering Samie, and Jaffrey (2018). "Reading M6A in the Transcriptome: M6A-Binding Proteins". In: *Trends in Cell Biology* 28 (2), pp. 113–127. DOI: 10.1016/j.tcb.2017.10.001.

Edward, Merino, Kevin A, Wilkinson Jennifer, Coughlan Kevin, and Weeks (2005). "RNA Structure Analysis at Single Nucleotide Resolution by Selective 2-Hydroxyl Acylation and Primer Extension (SHAPE)". In: *Journal of the American Chemical Society* 127 (12), pp. 4223–4231. DOI: 10.1021/ja043822v.

Elizabeth, Manrao, Ian M, Derrington Andrew, Laszlo Kyle, Langford Matthew, Hopper Nathaniel, Gillgren Mikhail, Pavlenok Michael, Niederweis Jens, and Gundlach (2012). "Reading DNA at Single-Nucleotide Resolution

with a Mutant MspA Nanopore and Phi29 DNA Polymerase". In: *Nature Biotechnology* 30 (4), pp. 349–353. DOI: 10.1038/nbt.2171.

Emily, Harcourt, Kietrys Anna, and Kool Eric (2017). "Chemical and Structural Effects of Base Modifications in Messenger RNA". In: *Nature* 541 (7637), pp. 339–346. DOI: 10.1038/nature21351.

Felix, Grünberger, Knüppel Robert, Jüttner Michael, Fenk Martin, Borst Andreas, Reichelt Robert, Hausner Winfried, Soppa Jörg, Ferreira-Cerca Sebastien, and Grohmann Dina (2019). "Nanopore-Based Native RNA Sequencing Provides Insights into Prokaryotic Transcription, Operon Structures, RRNA Maturation and Modifications". In: *BioRxiv*. DOI: 10.1101/2019.12.18.880849.

Heather, Drexler, Choquet Karine, and Stirling Churchman L (2020). "Splicing Kinetics and Coordination Revealed by Direct Nascent RNA Sequencing through Nanopores". In: *Molecular Cell* 77 (5), pp. 985–998. DOI: 10.1016/j.molcel.2019.11.017.

Jared, Simpson, Workman Rachael, Zuzarte P, David L, Dursi Winston, and Timp (2017). "Detecting DNA Cytosine Methylation Using Nanopore Sequencing". In: *Nature Methods* 14 (4), pp. 407–410. DOI: 10.1038/nmeth.4184.

Jonathan, Craig, Andrew H, Laszlo Ian, Nova Henry, Brinkerhoff Matthew, Noakes Katherine, Baker Jasmine, Bowman Hugh, Higinbotham Jonathan, Mount Jens, and Gundlach (2019). "Determining the Effects of DNA Sequence on Hel308 Helicase Translocation along Single-Stranded DNA Using Nanopore Tweezers". In: *Nucleic Acids Research* 47 (5), pp. 2506–2513. DOI: 10.1093/nar/gkz004.

Jong, Aw, Ashley Shaun, Lim Jia, Xu Wang, Finnlay R, Lambert Wen, Tan Yang, Shen Yu, and Zhang (2020). "Determination of Isoform-Specific RNA Structure with Nanopore Long Reads". In: *Nature Biotechnology*. DOI: 10.1038/s41587-020-0712-z.

Jun, Jiang, Seo Hyosuk, and Chow Christine (2016). "Post-Transcriptional Modifications Modulate RRNA Structure and Ligand Interactions". In: *Accounts of Chemical Research* 49 (5), pp. 893–901. DOI: 10.1021/acs.accounts.6b00014.

Kate, Meyer (2019). "DART-Seq: An Antibody-Free Method for Global M6A Detection". In: *Nature Methods* 16 (12), pp. 1275–1280. DOI: 10.1038/s41592-019-0570-0.

Katherine, Deigan, Tian W, Li David, Mathews Kevin, and Weeks (2009). "Accurate SHAPE-Directed RNA Structure Determination". In: *Proceedings of the National Academy of Sciences* 106 (1), pp. 97–102.

Kentaro, Kawata, Wakida Hiroyasu, Yamada Toshimichi, Taniue Kenzui, Han Han, Seki Masahide, Suzuki Yutaka, and Akimitsu Nobuyoshi (2020). "Metabolic Labeling of RNA Using Multiple Ribonucleoside Analogs Enables the Simultaneous Evaluation of RNA Synthesis and Degradation Rates". In: *Genome Research* 30 (10), pp. 1481–1491. DOI: 10.1101/GR.264408.120.

Kevin, Weeks and 2021 (n.d.). "SHAPE Directed Discovery of New Functions in Large RNAs". In: *Accounts of Chemical Research* 54 (10), pp. 2502–2517.

Liying, Meng, Guo Yilan, Tang Qi, Huang Rongbing, Xie Yuchen, and Chen Xing (2020). "Metabolic RNA Labeling for Probing RNA Dynamics in Bacteria". In: *Nucleic Acids Research* 48 (22), pp. 12566–12576. DOI: 10.1093/nar/gkaa1111.

Marcus, Stoiber, Quick Joshua, Egan Rob, Lee Ji, Celniker Susan, Neely Robert, Loman Nicholas, Pennacchio Len, and Brown James (2016). "De Novo Identification of DNA Modifications Enabled by Genome-Guided Nanopore Signal Processing". In: *BioRxiv*, p. 94672. DOI: 10.1101/094672.

Mark, Helm and Motorin Yuri (2017). "Detecting RNA Modifications in the Epitranscriptome: Predict and Validate". In: *Nature Reviews Genetics* 18 (5), pp. 275–291.

Maryam, Habibian, Velema Willem, Kietrys Anna, Onishi Yoshiyuki, and Kool Eric (2019). "Polyacetate and Polycarbonate RNA: Acylating Reagents and Properties". In: *Organic Letters* 21 (14), pp. 5413–5416. DOI: 10.1021/acs.orglett.9b01526.

Masato, Taoka, Nobe Yuko, Yamaki Yuka, Yamauchi Yoshio, Ishikawa Hideaki, Takahashi Nobuhiro, Nakayama Hiroshi, and Isobe Toshiaki (2016). "The Complete Chemical Structure of Saccharomyces Cerevisiae RRNA: Partial Pseudouridylation of U2345 in 25S RRNA by SnoRNA SnR9". In: *Nucleic Acids Research* 44 (18), pp. 8951–8961. DOI: 10.1093/nar/gkw564.

Matthias, Schaefer, Pollex Tim, Hanna Katharina, and Lyko Frank (2009). "RNA Cytosine Methylation Analysis by Bisulfite Sequencing". In: *Nucleic Acids Research* 37 (2). DOI: 10.1093/nar/gkn954.

Nathan, Siegfried, Steven Busan Greggory, Rice Julie, Nelson Kevin, and Weeks (2014). "RNA Motif Discovery by SHAPE and Mutational Profiling (SHAPE-MaP)". In: *Nature Methods* 11 (9), pp. 959–965.

Nicky, Jonkhout, Tran Julia, Smith Martin, Schonrock Nicole, Mattick John, and Novoa Eva (2017). "The RNA Modification Landscape in Human Disease". In: *Rna* 23 (12), pp. 1754–1769. DOI: 10.1261/rna.063503.117.

Ning, Zhang, Shi Shundi, Jia Tony, Ziegler Ashley, Yoo Barney, Yuan Xiaohong, Li Wenjia, and Zhang Shenglong (2019). "A General LC-MS-Based RNA Sequencing Method for Direct Analysis of Multiple-Base Modifications in RNA Mixtures". In: *Nucleic Acids Research* 47 (20), e125. DOI: 10.1093/nar/gkz731.

P, Prusiner, Yathindra N, and Sundaralingam M (1974). "Effect of Ribose O(2)-Methylation on the Conformation of Nucleosides and Nucleotides". In: *BBA Section Nucleic Acids And Protein Synthesis* 366 (2), pp. 115–123. DOI: 10.1016/0005-2787(74)90325-6.

Qing, Dai, Moshitch-Moshkovitz Sharon, Han Dali, Kol Nitzan, Amariglio Ninette, Rechavi Gideon, Dominissini Dan, and He Chuan (2017). "Nm-Seq Maps 2-O-Methylation Sites in Human MRNA with Base Precision". In: *Nature Methods* 14 (7), pp. 695–698. DOI: 10.1038/nmeth.4294.

Rebecca, Rose, Quinn Ryan, Sayre Jackie, and Fabris Daniele (2015). "Profiling Ribonucleotide Modifications at Full-Transcriptome Level: A Step toward MS-Based Epitranscriptomics (RNA (2015) 21 (2143))". In: *Rna* 21 (12), p. 2143. DOI: 10.1261/rna.054908.115.

Reid, Brown, Casey T, Andrews Adrian, and Elcock (2015). "Stacking Free Energies of All DNA and RNA Nucleoside Pairs and Dinucleoside-Monophosphates Computed Using Recently Revised AMBER Parameters and Compared with Experiment". In: *Correction. Journal of Chemical Theory and Computation* 11 (5), pp. 2315–2328.

Robert, Spitale, Crisalli Pete, Flynn Ryan, Torre Eduardo, Kool Eric, and Chang Howard (2012). "RNA SHAPE Analysis in Living Cells". In: *Nature Chemical Biology* 9 (1), pp. 18–20. DOI: 10.1038/nchembio.1131.

Saikat, Chakraborty, Mehtab Shabana, Patwardhan Anand, and Krishnan Yamuna (2012). In: *Pri-MiR-17-92a Transcript Folds into a Tertiary Structure and Autoregulates Its Processing* 18, pp. 1014–1028. DOI: 10.1261/rna.031039.111.

Schwartz, Schraga Douglas, Bernstein Maxwell, Mumbach Marko, Jovanovic Rebecca, Herbst Brian, León-Ricardo Jesse, and Engreitz (2014). "Transcriptome-Wide Mapping Reveals Widespread Dynamic-Regulated Pseudouridylation of NcRNA and MRNA". In: *Cell* 159 (1), pp. 148–162. DOI: 10.1016/j.cell.2014.08.028.

Stephenson, William, Roham Razaghi, Steven Busan, Kevin M Weeks, Winston Timp, and Peter Smibert (2022). "Direct detection of RNA modifications and structure using single-molecule nanopore sequencing". In: *Cell genomics* 2.2, p. 100097.

Steven, Chaulk, Gina L, Thede Oliver, Kent Zhizhong, Xu Emily, Gesner Richard, Veldhoen, Suneil K, and Khanna (2011). "Role of Pri-MiRNA Tertiary Structure in MiR-17∼92 MiRNA Biogenesis". In: *RNA Biology* 8 (6). DOI: 10.4161/rna.8.6.17410.

Thomas, Carlile, Maria F, Rojas-Duran Boris, Zinshteyn Hakyung, Shin Kristen, Bartoli Wendy, and Gilbert (2014). "Pseudouridine Profiling Reveals Regulated MRNA Pseudouridylation in Yeast and Human Cells". In: *Nature* 515 (7525), pp. 143–146. DOI: 10.1038/nature13802.

Triinu, Siibak and Remme Jaanus (2010). "Subribosomal Particle Analysis Reveals the Stages of Bacterial Ribosome Assembly at Which RRNA Nucleotides Are Modified". In: *Rna* 16 (10), pp. 2023–2032. DOI: 10.1261/rna.2160010.

Veronika, Herzog, Reichholf Brian, Neumann Tobias, Rescheneder Philipp, Bhat Pooja, Burkard Thomas, Wlotzka Wiebke, Arndt Von, Haeseler Johannes, Zuber Stefan, and Ameres (2017). "Thiol-Linked Alkylation of RNA to Assess Expression Dynamics". In: *Nature Methods* 14 (12), pp. 1198–1204. DOI: 10.1038/nmeth.4435.

Wendy, Gilbert, Bell Tristan, and Schaening Cassandra (2016). "Messenger RNA Modifications: Form, Distribution, and Function". In: *Science* 352 (6292), pp. 1408–1412. DOI: 10.1126/science.aad8711.

Xiaoyu, Li, Xiong Xushen, and Yi Chengqi (2016). "Epitranscriptome Sequencing Technologies: Decoding RNA Modifications". In: *Nature Methods* 14 (1), pp. 23–31.

Yi, Yu, Tao U. Thomas, and Meier (2014). "RNA-Guided Isomerization of Uridine to Pseudouridine -Pseudouridylation". In: *RNA Biology* 11 (12), pp. 1483–1494.

Yue, Wan, Qu Kun, Zhang Qiangfeng, Flynn Ryan, Manor Ohad, Ouyang Zhengqing, and Zhang Jiajing (2014). "Landscape and Variation of RNA Secondary Structure across the Human Transcriptome". In: *Nature* 505 (7485), pp. 706–709.

Zaccara, Sara Ryan, Ries Samie, and Jaffrey (2019). "Reading, Writing and Erasing MRNA Methylation". In: *Nature Reviews. Molecular Cell Biology* 20.

Chapter 4

Modbamtools: Analysis of single-molecule epigenetic data for long-range profiling, heterogeneity, and clustering

This chapter is currently a preprint deposited on bioRxiv. I am the first author leading this work and responsible for the full implementation of modbamtools.

Roham Razaghi et al. (2022). “Modbamtools: Analysis of single-molecule epigenetic data for long-range profiling, heterogeneity, and clustering”. In: *bioRxiv*

4.1 Abstract

The advent of long-read sequencing methods provides new opportunities for profiling the epigenome - especially as the methylation signature comes for “free" when native DNA is sequenced on either Oxford Nanopore or Pacific Biosciences instruments. However, we lack tools to visualize and analyze

data generated from these new sources. Recent efforts from the GA4GH consortium have standardized methods to encode modification location and probabilities in the BAM format. Leveraging this standard format, we developed a technology-agnostic tool, modbamtools to visualize, manipulate and compare base modification/methylation data in a fast and robust way. modbamtools can produce high quality, interactive, and publication-ready visualizations as well as provide modules for downstream analysis of base modifications. Modbamtools comprehensive manual and tutorial can be found at `https://rrazaghi.github.io/modbamtools/`.

4.2 Introduction

Direct single-molecule sequencing methods, e.g. Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT), have recently greatly expanded in throughput and yield. In addition to the canonical base sequencing data that these platforms generate, modifications on the nucleic acids can be measured directly, either via delays in the incorporation of bases (IPD, PacBio (Flusberg et al., 2010)) or perturbations in the electrical current (ONT (Simpson et al., 2017)). These have been accompanied by development of software tools to measure and call modifications within this data, but the output formats of these calls were not standardized precluding easy downstream development. Modification data files have typically been stored as enormous (terabyte scale) tsv/csvs and early efforts to incorporate 5-methylcytosine information from ONT into a “bisulfite-like" BAM file required complex manipulations(Lee et al., 2020).

More recently, the Global Alliance for Genomics and Health (GA4GH) (Rehm et al., 2021) standards group proposed an addition to the BAM file spec, incorporating two new tags (MM and ML) for SAM/BAM alignment files. The MM tag is used to locate the strand and position the modification was observed on, and the ML tag is the probability of each modification being present (`http://samtools.github.io/hts-specs`). Although these tags were introduced as an adaptation to long-read base modification data, it is anticipated that all technologies will eventually incorporate this file format.

Single molecule base modification callers have rapidly adapted to the new standard format. Currently, for nanopore data, most modification calling tools can output BAM files with tags, including guppy, bonito, Megalodon, and nanopolish (Simpson et al., 2017). Similarly, Primrose, and ccsmeth (Ni et al., 2022) can be used for PacBio reads. An updated list of compatible tools generating these alignment files can be found at `https://rrazaghi.github.io/modbamtools/`.

Here we introduce modbamtools, a suite of tools to explore modifications in single-molecule data using this new format. With this tool we generate interactive and batch visualization and analysis for methylation frequency and single-molecule methylation. Profiling methylation across individual molecules, we can look at coordination of long-range methylation effects, e.g. enhancer-promoter interactions, and the degree of variation of methylation "noise" within regions. We have also generated modules to phase reads by using genetic variation or through methylation alone via a read clustering approach, to enable exploration of allele-specific methylation and epigenetic

heterogeneity.

4.3 Results

4.4 Usage and Examples

We developed modbamtools, a software package that provides analysis and interactive visualization of single-read base modification data along with other highly used formats for genomic tracks (GTF, bigwig, bedgraph, etc). Modbamtools utilizes core python modules including numpy (Walt, Colbert, and Varoquaux, 2011), pandas (McKinney and Others, 2011), scikit-learn (Pedregosa et al., 2011), pysam (Heger et al., 2014), click, plotly (Plotly Technologies Inc., 2015), modbampy, pybigwig (Ryan, Gruning, and Ramirez, 2016), pypdf2, pillow, and hdbscan (McInnes, Healy, and Astels, 2017). We have made modbamtools easily accessible through PyPI ('pip install modbamtools').

The tool has three main elements ('calcMeth', 'calcHet', 'cluster') and a plotting function that allows for interactive plotting of single-read base modification data. This generates a multi-panel plot (Figure 4.1) consisting of an annotation track, methylation frequency track, and single-read plots. The annotation track can display other sets of genomics data including gene models, other epigenetic data (e.g. ENCODE ChIP-seq), and genetic variation. Methylation frequencies along with a smoothed average frequency is plotted on top of the reads similar to a conventional genome browser. The methylation frequency plot shows the per locus frequency of modified to total called bases.

Finally, the single-read plots represent each individual single molecule with base modifications indicated as blue for unmodified and red for modified. These figures can be output as HTML, PDF, PNG, or SVG. The HTML provided is generated with plotly and is interactive, allowing magnification. Multiple plots can be output in batch mode by providing a BED file of regions of interest resulting in a multiple page HTML or PDF report.

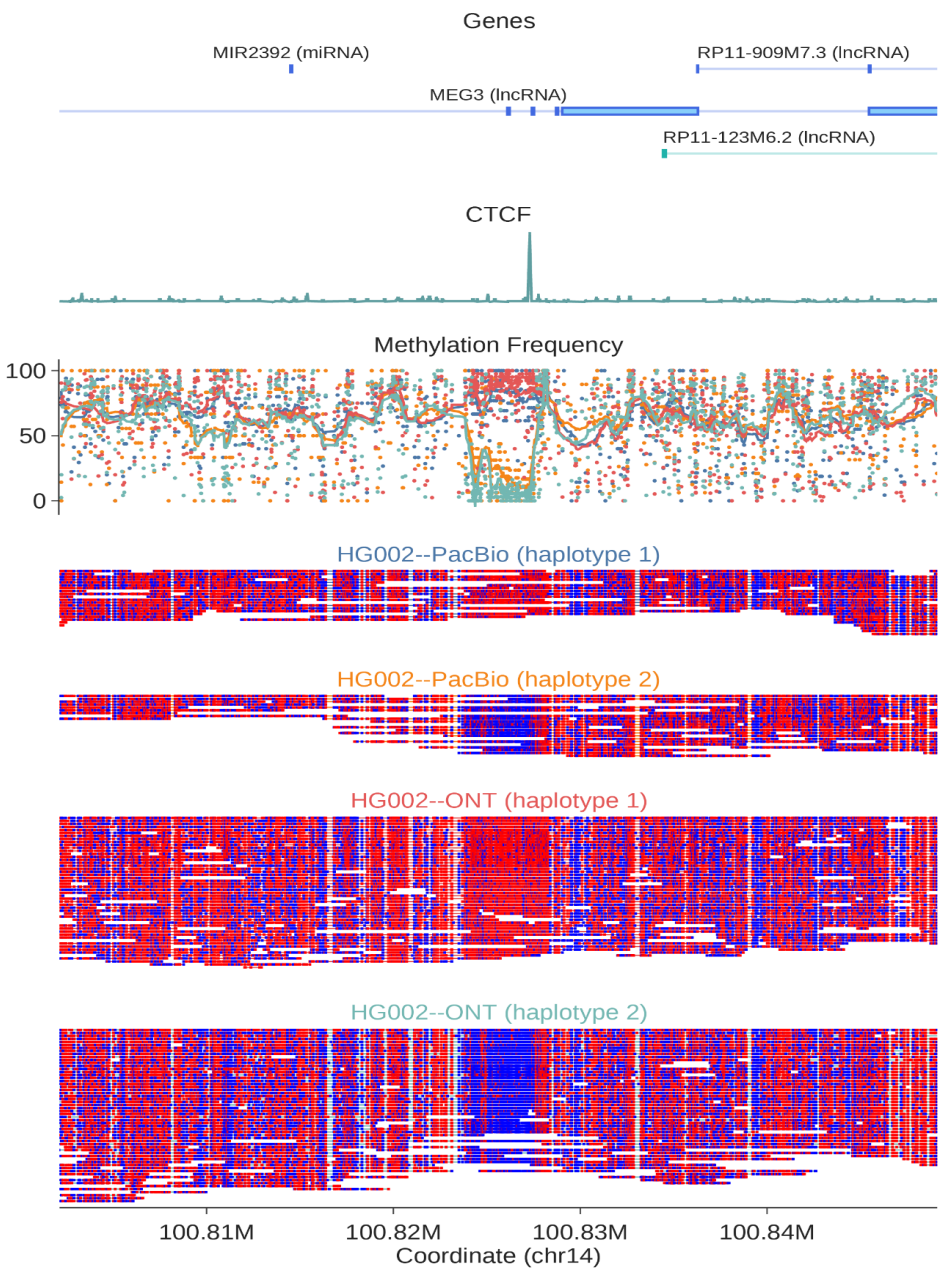

Figure 4.1: *Example of modbamtools output onMEG3 (chr14 : 100,802,849,111) locus using both PacBio and ONT single-molecule data from the HG002 Genome in a Bottle cell line. "Genes" track shows GENCODE (Release 38, GRCH38) gene models and the "CTCF" track shows CTCF CHiP-seq ENCODE track from GM12878. Methylation frequency track is colored according to platform and haplotype, with colors indicated by the title of the single-molecule*

Using appropriate tools, e.g. clair (Zheng et al., 2021) or whatshap (Martin et al., 2016), BAM files can have the haplotypes of reads encoded with the commonly used “HP" tag. Our tool has the ability to group the alignments based on phase tag (HP) in BAM files. Using this HP tag, we can separate reads according to haplotype, plotting each haplotype’s methylation frequency as different colored lines and the single reads as separate plot elements. We show an example of this module on methylation calls from the HG002 cell line at the *MEG3* long noncoding RNA (lncRNA), using public single-molecule methylation data from both ONT and PacBio platforms (Figure 4.1). *MEG3* has known monoallelic expression in many tissues and loss of this regulation has been implicated in development of type 2 diabetes mellitus (Rosa et al., 2005; Kameswaran et al., 2014). From this data, we observe clear examples of allele-specific methylation at a CTCF binding site and *MEG3* promoter region.

Beyond clustering according to genomic haplotype, we have implemented a method to cluster single-molecule reads based on methylation status alone using Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) (McInnes, Healy, and Astels, 2017). This is a useful feature for regions without many SNPs for phasing reads into haplotypes (Gershman et al., 2022). Clustering can also be used to quantify different cell types or to profile early cancer detection from a heterogeneous sample (Wang et al., 2021; Houseman et al., 2008; Gkountela et al., 2019; Tian et al., 2020). Clustering can be performed either as a part of the plotting command or separately (‘–cluster‘ command) with the input of a batch file for locations used for the clustering. As shown in Figure 4.2, we can cluster the *SNURF* gene promoter based purely

on methylation signal at this locus. This paternally imprinted locus can also be phased based on genotyping information, demonstrating the agreement of our clustering approach with classical methods.

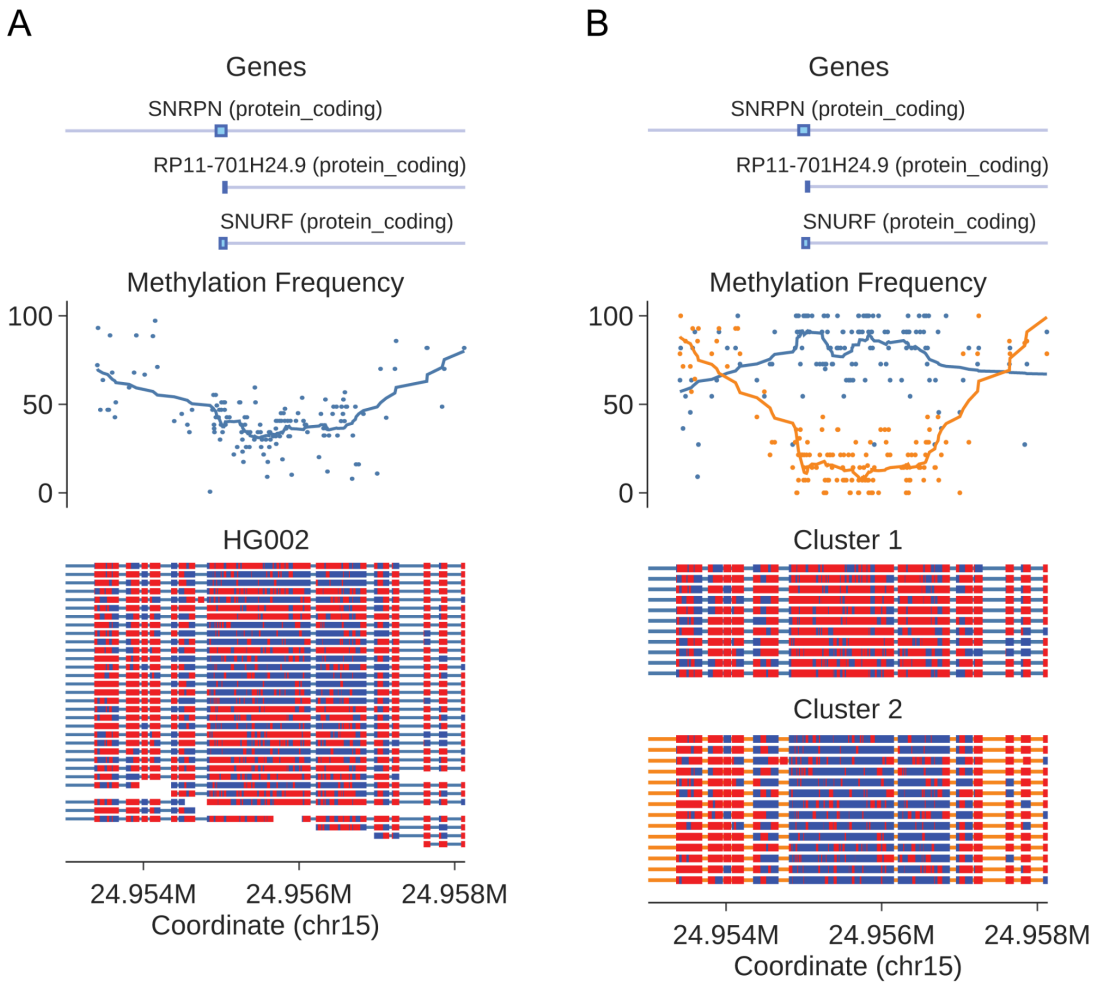

Figure 4.2: *A) single molecule methylation profile on geneSNRPN (chr15 : 24,953,958,133) from HG002 data as in Figure1.B) Single molecule methylation profile on gene SNRPN separated into clusters with 'modbamtools plot -cluster'*

Finally, using a BED file of genomic loci, we can profile the average methylation in each location, including methylation on each haplotype. The "calcMeth" module calculates methylation average across each single molecule

first then aggregates over all molecules which map to that region, rather than averaging CpG methylation per CpG then averaging across the region. This is especially useful with long reads to capture methylation variability more efficiently (Figure 4.3).

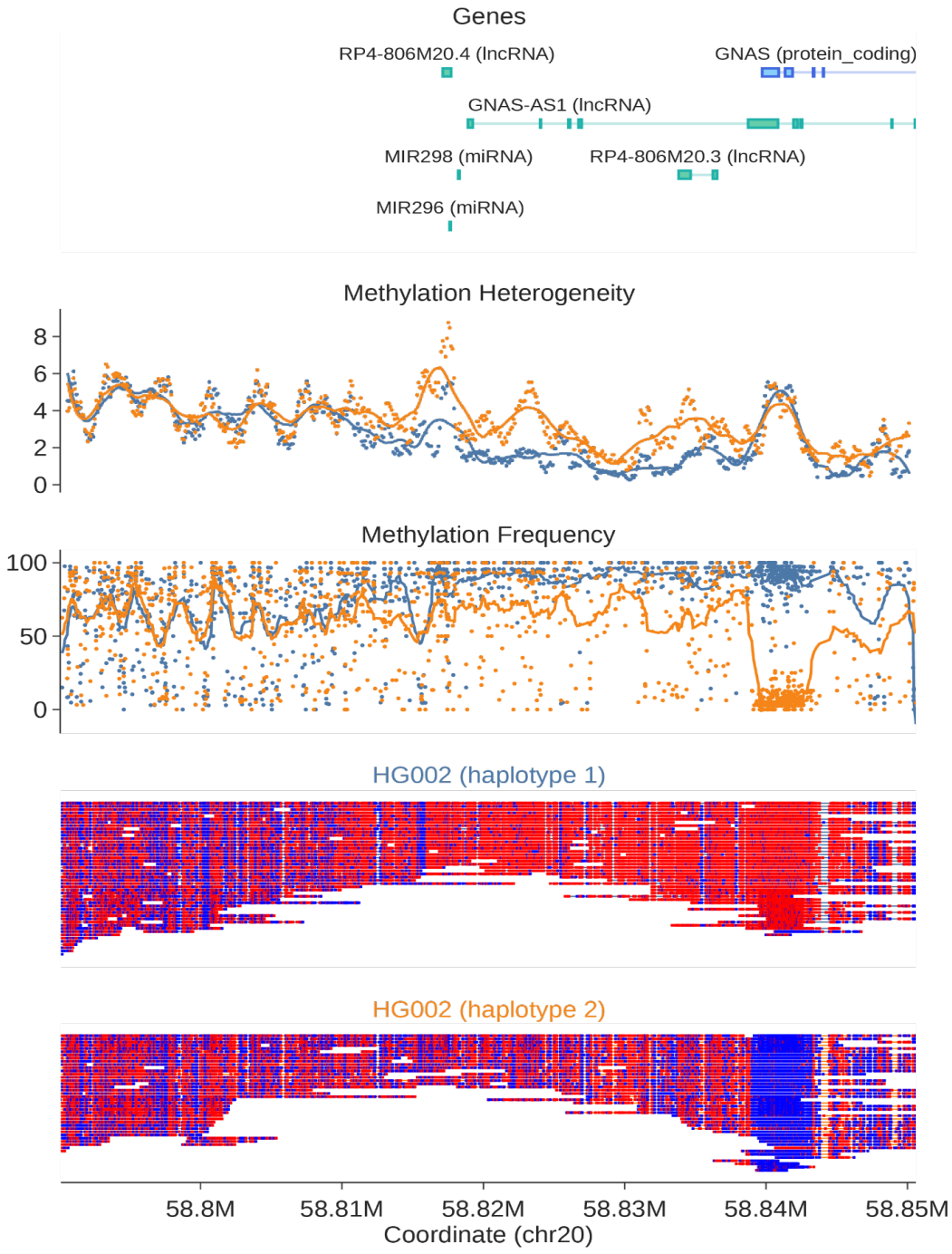

Figure 4.3: *Example of modbamtools plot with options for haplotype separation and calculating heterogeneity at theGNAS locus (chr20 : 58,790,850,596).*

With single-molecule methylation data, can quantify not only methylation frequency averaged across all reads, but also variability of methylation across individual molecules. A few studies have attempted to address this by proposing different algorithms to quantify this feature (Scherer et al., 2020; Landau et al., 2014; Guo et al., 2017; Landan et al., 2012; Xie et al., 2011). Here, we implemented a module to calculate methylation heterogeneity ("calcHet") that calculates this on genomic regions provided by the user. Similar to the clustering function, "–heterogeneity" option can be used with plotting command to visualize this; we have plotted it for the *GNAS* locus in Figure 4.3. There we observe areas of clear difference in methylation heterogeneity across the region, suggesting not only a change in methylation but a less ordered epigenetic state on one allele when compared to the other.

4.5 Conclusion

Advances in single-molecule sequencing throughput suggest we are at an inflection point where large scale data sets are on the horizon. These data types offer the unique advantage of providing DNA methylation data *as well as* primary sequence - but without tools to take advantage of it, these data will be "left on the table" and not used to their potential. Here we have described a toolset to take advantage of these data, using the newly described modification tags present in the SAM/BAM file specifications. This toolset is compatible with all modern modification callers. Modbamtools provides fast, robust, interactive visualization and analysis for alignment files containing

base modification tags.

4.6 Acknowledgments

We would like to thank Jared Simpson and Chris Wright for their helpful comments and contributions to the development of modified base alignment files. W.T. has two patents (8,748,091 and 8,394,584) licensed to ONT.

4.7 Funding

This study was supported by National Human Genome Research Institute (project no. 5R01HG009190) and National Cancer Institute (project no. 1U01CA253481-01A1)

4.8 Data Availability

Publicly available data on cell line HG002 was downloaded from s3://ont-open-data/gm24385_mod_2021.09/extra_analysis/bonito_remora (ONT HG002 WGS) and https://downloads.pacbcloud.com/public/dataset/HG002-CpG-methylation-202202/ (PacBio HG002 WGS). CTCF track was downloaded from https://www.encodeproject.org/experiments/ENCSR000DZN/. GENCODE Release 38 for GRCH38 was used for gene model tracks.

4.9 Code Availability and implementation

modbamtools source code is available at https://github.com/rrazaghi/modbamtools. A manual and tutorial are available at https://rrazaghi.github.io/modbamtools/.

4.10 References

Flusberg, Benjamin A, Dale R Webster, Jessica H Lee, Kevin J Travers, Eric C Olivares, Tyson A Clark, Jonas Korlach, and Stephen W Turner (2010). "Direct detection of DNA methylation during single-molecule, real-time sequencing". en. In: *Nat. Methods* 7.6, pp. 461–465.

Gershman, Ariel, Michael E G Sauria, Xavi Guitart, Mitchell R Vollger, Paul W Hook, Savannah J Hoyt, Miten Jain, Alaina Shumate, Roham Razaghi, Sergey Koren, Nicolas Altemose, Gina V Caldas, Glennis A Logsdon, Arang Rhie, Evan E Eichler, Michael C Schatz, Rachel J O'Neill, Adam M Phillippy, Karen H Miga, and Winston Timp (2022). "Epigenetic patterns in a complete human genome". en. In: *Science* 376.6588, eabj5089.

Gkountela, Sofia, Francesc Castro-Giner, Barbara Maria Szczerba, Marcus Vetter, Julia Landin, Ramona Scherrer, Ilona Krol, Manuel C Scheidmann, Christian Beisel, Christian U Stirnimann, Christian Kurzeder, Viola Heinzelmann-Schwarz, Christoph Rochlitz, Walter Paul Weber, and Nicola Aceto (2019). "Circulating Tumor Cell Clustering Shapes DNA Methylation to Enable Metastasis Seeding". en. In: *Cell* 176.1-2, 98–112.e14.

Guo, Shicheng, Dinh Diep, Nongluk Plongthongkum, Ho-Lim Fung, Kang Zhang, and Kun Zhang (2017). "Identification of methylation haplotype blocks aids in deconvolution of heterogeneous tissue samples and tumor tissue-of-origin mapping from plasma DNA". en. In: *Nat. Genet.* 49.4, pp. 635–642.

Heger, A, T G Belgrad, M Goodson, and K Jacobs (2014). *pysam: Python interface for the SAM/BAM sequence alignment and mapping format*.

Houseman, E Andres, Brock C Christensen, Ru-Fang Yeh, Carmen J Marsit, Margaret R Karagas, Margaret Wrensch, Heather H Nelson, Joseph Wiemels, Shichun Zheng, John K Wiencke, and Karl T Kelsey (2008). "Model-based clustering of DNA methylation array data: a recursive-partitioning algorithm for high-dimensional data arising as a mixture of beta distributions". en. In: *BMC Bioinformatics* 9, p. 365.

Kameswaran, Vasumathi, Nuria C Bramswig, Lindsay B McKenna, Melinda Penn, Jonathan Schug, Nicholas J Hand, Ying Chen, Inchan Choi, Anastassios Vourekas, Kyoung-Jae Won, Chengyang Liu, Kumar Vivek, Ali Naji, Joshua R Friedman, and Klaus H Kaestner (2014). "Epigenetic regulation of the DLK1-MEG3 microRNA cluster in human type 2 diabetic islets". en. In: *Cell Metab.* 19.1, pp. 135–145.

Landan, Gilad, Netta Mendelson Cohen, Zohar Mukamel, Amir Bar, Alina Molchadsky, Ran Brosh, Shirley Horn-Saban, Daniela Amann Zalcenstein,

Naomi Goldfinger, Adi Zundelevich, Einav Nili Gal-Yam, Varda Rotter, and Amos Tanay (2012). "Epigenetic polymorphism and the stochastic formation of differentially methylated regions in normal and cancerous tissues". en. In: *Nat. Genet.* 44.11, pp. 1207–1214.

Landau, Dan A, Kendell Clement, Michael J Ziller, Patrick Boyle, Jean Fan, Hongcang Gu, Kristen Stevenson, Carrie Sougnez, Lili Wang, Shuqiang Li, Dylan Kotliar, Wandi Zhang, Mahmoud Ghandi, Levi Garraway, Stacey M Fernandes, Kenneth J Livak, Stacey Gabriel, Andreas Gnirke, Eric S Lander, Jennifer R Brown, Donna Neuberg, Peter V Kharchenko, Nir Hacohen, Gad Getz, Alexander Meissner, and Catherine J Wu (2014). "Locally disordered methylation forms the basis of intratumor methylome variation in chronic lymphocytic leukemia". en. In: *Cancer Cell* 26.6, pp. 813–825.

Lee, Isac, Roham Razaghi, Timothy Gilpatrick, Michael Molnar, Ariel Gershman, Norah Sadowski, Fritz J Sedlazeck, Kasper D Hansen, Jared T Simpson, and Winston Timp (2020). "Simultaneous profiling of chromatin accessibility and methylation on human cell lines with nanopore sequencing". en. In: *Nat. Methods* 17.12, pp. 1191–1199.

Martin, Marcel, Murray Patterson, Shilpa Garg, Sarah O Fischer, Nadia Pisanti, Gunnar W Klau, Alexander Schöenhuth, and Tobias Marschall (2016). "WhatsHap: fast and accurate read-based phasing". en.

McInnes, Leland, John Healy, and Steve Astels (2017). "hdbscan: Hierarchical density based clustering". In: *J. Open Source Softw.* 2.11, p. 205.

McKinney, Wes and Others (2011). "pandas: a foundational Python library for data analysis and statistics". In: *Python for high performance and scientific computing* 14.9, pp. 1–9.

Ni, Peng, Jinrui Xu, Zeyu Zhong, Jun Zhang, Neng Huang, Fan Nie, Feng Luo, and Jianxin Wang (2022). "DNA 5-methylcytosine detection and methylation phasing using PacBio circular consensus sequencing". en.

Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, and Others (2011). "Scikit-learn: Machine learning in Python". In: *the Journal of machine Learning research* 12, pp. 2825–2830.

Razaghi, Roham, Paul W Hook, Shujun Ou, Michael Schatz, Kasper D Hansen, Miten Jain, and Winston Timp (2022). "Modbamtools: Analysis of single-molecule epigenetic data for long-range profiling, heterogeneity, and clustering". In: *bioRxiv*.

Rehm, Heidi L et al. (2021). "GA4GH: International policies and standards for data sharing across genomic research and healthcare". en. In: *Cell Genom* 1.2.

Rosa, Alberto L, Yuan-Qing Wu, Bernard Kwabi-Addo, Karen J Coveler, V Reid Sutton, and Lisa G Shaffer (2005). "Allele-specific methylation of a functional CTCF binding site upstream of MEG3 in the human imprinted domain of 14q32". en. In: *Chromosome Res.* 13.8, pp. 809–818.

Ryan, D, B Gruning, and F Ramirez (2016). "pyBigWig 0.2. 4". In: *Cited on*, p. 3.

Scherer, Michael, Almut Nebel, Andre Franke, Jörn Walter, Thomas Lengauer, Christoph Bock, Fabian Müller, and Markus List (2020). "Quantitative comparison of within-sample heterogeneity scores for DNA methylation data". en. In: *Nucleic Acids Res.* 48.8, e46.

Simpson, Jared T, Rachael E Workman, P C Zuzarte, Matei David, L J Dursi, and Winston Timp (2017). "Detecting DNA cytosine methylation using nanopore sequencing". en. In: *Nat. Methods* 14.4, pp. 407–410.

Tian, Zijian, Lingfeng Meng, Xingbo Long, Tongxiang Diao, Maolin Hu, Miao Wang, Ming Liu, and Jianye Wang (2020). "DNA methylation-based classification and identification of bladder cancer prognosis-associated subgroups". en. In: *Cancer Cell Int.* 20, p. 255.

Walt, Stefan van der, S Chris Colbert, and Gael Varoquaux (2011). "The NumPy Array: A Structure for Efficient Numerical Computation". In: *Computing in Science Engineering* 13.2, pp. 22–30.

Wang, Jieyu, Jun Li, Ruifang Chen, Huiran Yue, Wenzhi Li, Beibei Wu, Yang Bai, Guohua Zhu, and Xin Lu (2021). "DNA methylation-based profiling reveals distinct clusters with survival heterogeneity in high-grade serous ovarian cancer". en. In: *Clin. Epigenetics* 13.1, p. 190.

Xie, Hehuang, Min Wang, Alexandre de Andrade, Maria de F Bonaldo, Vasil Galat, Kelly Arndt, Veena Rajaram, Stewart Goldman, Tadanori Tomita, and Marcelo B Soares (2011). "Genome-wide quantitative assessment of variation in DNA methylation patterns". en. In: *Nucleic Acids Res.* 39.10, pp. 4099–4108.

Zheng, Zhenxian, Shumin Li, Junhao Su, Amy Wing-Sze Leung, Tak-Wah Lam, and Ruibang Luo (2021). "Symphonizing pileup and full-alignment for deep learning-based long-read variant calling". en.

Chapter 5

Conclusion and Future direction

This thesis summarizes the work I have performed in the Timp lab at Johns Hopkins University during my PhD. My dissertation has resulted in 9 manuscripts (one as a co-first author, and one as a first author) (Razaghi et al., 2022; Kovaka et al., 2019; Lee et al., 2020; Workman et al., 2019; Stephenson et al., 2022; Tiek et al., 2022; Gershman et al., 2021; Gershman et al., 2022; Kandathil et al., 2021). The focus of my work has revolved around sequencing technology methods development using third-generation sequencing.

First in my dissertation work, we generated one of the largest to date direct-RNA nanopore sequencing datasets. This data being one of the first of its kind, showcased the potential of dRNA sequencing in the fields of transcriptomics and epitranscriptomics. We demonstrated how dRNA can be utilized to advance our knowledge of allele-specific expression, poly(A) tail, and RNA base modifications (Workman et al., 2019).

We then focused on a follow up study with collaboration with New York Genome Center to investigate the potential of RNA exogenous labeling. This labeling on long RNA molecules proved to be effective in inferring RNA

secondary structure. This coupled with analysis enabled by the first study provided a way to cluster alternative structures in an isoform-specific manner while measuring endogenous RNA modification (Stephenson et al., 2022).

For the second half of my dissertation, I focused on DNA and base modifications. I was involved in a study that adapts the sequencing technique NOMe-seq to third-generation sequencing (NanoNOMe) (Lee et al., 2020; Lay, Kelly, and Jones, 2018). This enables us to have a phased epigenome that includes the simultaneous measurement of endogenous (DNA CpG methylation) and exogenous (chromatin accessibility) methylation profiles. Inspired by this study, we have been trying to develop a "multi-color" assay where different DNA-protein interactions can be profiled in addition to DNA methylation and chromatin accessibility. This can be achieved by fusing protein A/G to a DNA modifying enzyme. This construct can be pulled to a protein of interest with proper antibody treatments. The enzyme then modifies the DNA in close proximity of the protein of interest. Although, we gathered some preliminary data on this proposed assay, we need more time to refine and optimize further in the near future.

Lastly, I focused on one of the recent sequencing library preparation kits called ultra-long nanopore sequencing (SQK-ULK001). This kit has enabled the genomics community to produce reads up to 2-3 Mb in length. In our hands, we routinely generate 80-100 Gb of sequencing data with N50s of above 100 kb using a Promethion flowcell. These ultra-long reads provide a crucial tool to close any gaps in the human genome mostly due to repetitive elements (Nurk et al., 2022). With this better mappability, we can also study

the epigenome more comprehensively and discover novel biology (Gershman et al., 2022). One of the challenges of studying methylation using nanopore sequencing (especially ultra-long sequencing) is navigating rather large text files. Recently, htslib has provided a new way to store base modification information into alignment files. I developed modbamtools to provide the community a set of tools to navigate, visualize, and manipulate this new file format efficiently. One of the future directions for this study is to expand the functions further to adapt to the growing cell-free DNA (cfDNA) methylation analysis (Razaghi et al., 2022). Another potential future project would be to utilize ultra-long reads to investigate combinatorial methylation states at different regulatory elements (enhancers, promoters, gene bodies, etc) on a phased single molecule level. This can provide better insight on how methylation regulates gene expression levels.

5.1 References

Gershman, Ariel, Tatiana G Romer, Yunfan Fan, Roham Razaghi, Wendy A Smith, and Winston Timp (2021). "De novo genome assembly of the tobacco hornworm moth (Manduca sexta)". In: *G3* 11.1, jkaa047.

Gershman, Ariel, Michael EG Sauria, Xavi Guitart, Mitchell R Vollger, Paul W Hook, Savannah J Hoyt, Miten Jain, Alaina Shumate, Roham Razaghi, Sergey Koren, et al. (2022). "Epigenetic patterns in a complete human genome". In: *Science* 376.6588, eabj5089.

Kandathil, Abraham J, Andrea L Cox, Kimberly Page, David Mohr, Roham Razaghi, Khalil G Ghanem, Susan A Tuddenham, Yu-Hsiang Hsieh, Jennifer L Evans, Kelly E Coller, et al. (2021). "Plasma virome and the risk of blood-borne infection in persons with substance use disorder". In: *Nature communications* 12.1, pp. 1–7.

Kovaka, Sam, Aleksey V Zimin, Geo M Pertea, Roham Razaghi, Steven L Salzberg, and Mihaela Pertea (2019). "Transcriptome assembly from long-read RNA-seq alignments with StringTie2". In: *Genome biology* 20.1, pp. 1–13.

Lay, Fides D, Theresa K Kelly, and Peter A Jones (2018). "Nucleosome occupancy and methylome sequencing (NOMe-seq)". In: *DNA Methylation Protocols*. Springer, pp. 267–284.

Lee, Isac, Roham Razaghi, Timothy Gilpatrick, Michael Molnar, Ariel Gershman, Norah Sadowski, Fritz J Sedlazeck, Kasper D Hansen, Jared T Simpson, and Winston Timp (2020). "Simultaneous profiling of chromatin accessibility and methylation on human cell lines with nanopore sequencing". In: *Nature Methods* 17.12, pp. 1191–1199.

Nurk, Sergey, Sergey Koren, Arang Rhie, Mikko Rautiainen, Andrey V Bzikadze, Alla Mikheenko, Mitchell R Vollger, Nicolas Altemose, Lev Uralsky, Ariel Gershman, et al. (2022). "The complete sequence of a human genome". In: *Science* 376.6588, pp. 44–53.

Razaghi, Roham, Paul W Hook, Shujun Ou, Michael Schatz, Kasper D Hansen, Miten Jain, and Winston Timp (2022). "Modbamtools: Analysis of single-molecule epigenetic data for long-range profiling, heterogeneity, and clustering". In: *bioRxiv*.

Stephenson, William, Roham Razaghi, Steven Busan, Kevin M Weeks, Winston Timp, and Peter Smibert (2022). "Direct detection of RNA modifications and structure using single-molecule nanopore sequencing". In: *Cell genomics* 2.2, p. 100097.

Tiek, Deanna M, Beril Erdogdu, Roham Razaghi, Lu Jin, Norah Sadowski, Carla Alamillo-Ferrer, J Robert Hogg, Bassem R Haddad, David H Drewry, Carrow I Wells, et al. (2022). "Temozolomide-induced guanine mutations create exploitable vulnerabilities of guanine-rich DNA and RNA regions in drug-resistant gliomas". In: *Science Advances* 8.25, eabn3471.

Workman, Rachael E, Alison D Tang, Paul S Tang, Miten Jain, John R Tyson, Roham Razaghi, Philip C Zuzarte, Timothy Gilpatrick, Alexander Payne, Joshua Quick, et al. (2019). "Nanopore native RNA sequencing of a human poly (A) transcriptome". In: *Nature methods* 16.12, pp. 1297–1305.

Roham Razaghi

858-717-6027 | rrazagh1@jhu.edu | 414 Light St Apt 3804, Baltimore, MD 21202

Executive Summary

- Collaborative bioinformatician/computational biologist and molecular biologist with 5+ years of experience in utilizing statistical methods to analyze genomic, epigenomic, and transcriptomic datasets
- Adept in library preparation and data analysis for both next-generation sequencing (Illumina) and third-generation sequencing (Nanopore, PacBio) platforms. Vast experience in calling methylation, base modification, and tandem repeats using long-read sequencing
- Experienced in working with large datasets and applying exploratory analysis, statistical models (linear regression, time series modelling), multivariate analysis (dimensionality reduction, clustering, classification), probability sampling methods (random, stratified, cluster), and optimization (gradient descent)
- Expert in using python and R libraries for data interpretation and visualization. Basic familiarity with SQL (sqlite) and Apache Spark (pyspark) to analyze large datasets efficiently
- Experienced in leadership and teamwork: initiated and led collaborations and met deadlines within academia and with industry including an international consortium of six laboratories. Working with two companies (New England BioLabs, Immagina Biotechnology) to develop new products for next/third-generation sequencing
- Excellent written and oral communicator: invited speaker at an international conference, three poster presentations, taught computational biology courses to PhD students, published seven manuscripts in high-impact journals (one as a co-first author), two preprints (currently under review at journals), and two manuscripts currently in preparation

Education

Johns Hopkins School of Medicine — Baltimore, MD
Ph.D. Biomedical Engineering — *June 2017 – May 2022 (expected)*

University of California, San Diego — La Jolla, CA
B.S. Bioengineering — *September 2014 – May 2017*

Fresno City College — Fresno, CA
Engineering — *September 2012 – June 2014*

Technical Skills

Computational: Python, R, Bash, Linux, Nanopore sequencing, PacBio Sequencing, Single-Cell RNA-seq, ATAC-seq, ChIP-seq, CUT&RUN, Bisulfite sequencing, Jupyter, Google Colab, Snakemake, SQL (sqlite3), AWS (S3, EC2, Spark, Elastic MapReduce), Statistical Methodologies, Data Visualization, HTML/CSS, Adobe Illustrator, ImageJ
Molecular Biology: Third/Next Generation Sequencing Sample Preparation, HMW DNA extraction, Viral Transduction/Transfection, DNA/RNA Exogenous Labeling, Cloning, Protein Expression, PCR

Experience

Graduate Student Researcher — June 2017 – Present
Johns Hopkins School of Medicine — *Baltimore, MD*

Nanopore Sequencing for Measurement of Endogenous and Exogenous Modifications in Nucleic Acids

Thesis Advisor, PI: Winston Timp

- Developed modbamtools (https://rrazaghi.github.io/modbamtools/), software suite for the visualization, manipulation, and comparison of nucleic acid modifications encoded in modified base bam files
- Developed computational methods to investigate allele-specific alternative splicing, polyadenylation, and modifications using long-read direct-RNA sequencing data
- Efficiently lead and participate in collaborations within JHU and outside, including international consortia
- Developed computational methods to study RNA structure using third generation sequencing platforms
- Developing tools to explore genomic and transcriptomic data including single-cell and long-read data. Expert in haplotype phasing of long-read data based on variants and methylation
- Utilizing extensive suite of wet lab and computational techniques to develop an assay for simultaneous profiling of methylation, chromatin accessibility, and lamina associated domains using third-generation sequencing

Bioinformatics, Data Scientist June 2021 – August 2021
Gilead Sciences *Foster City, CA*
Biomarker Department Intern

- Collaborated with clinical bioinformaticians, biomarker scientists, and data scientists to conduct multi-omics analyses of biomarker data collected from clinical studies to help uncover modes of action of Gilead's cure therapeutics in HBV patients
- Developed an analysis pipeline to analyze single-cell RNA sequencing data from multiple clinical trial cohorts
- Implemented a workflow to detect and visualize viral integrations from targeted long read Pacbio sequencing data

Senior Design Engineer and Undergraduate Researcher Nov 2015 – May 2017
University of California, San Diego *La Jolla, CA*
Neural Interaction Laboratory
Advisor, PI: Todd Coleman

- Developed a light-controllable genetic switch through the binding interaction of hybrid molecules in order to promote controlled and precise gene, protein, and hormone production endogenously

Teaching and Leadership

JHU BME Equity, Diversity, and Inclusion Committee | *Co-Founder and Co-Chair* June 2020 – June 2021
We are a group of Ph.D. students in the biomedical engineering department at Johns Hopkins University working to make our department and communities more equitable, diverse, and inclusive (EDI). I co-founded this group in 2020 after the murder of George Floyd and the subsequent discussions centered around race, equity, diversity, and inclusion. We are working with faculty and staff in the BME department to advocate for Ph.D. students and EDI issues

- BME EDI launched the BME PhD Mentorship Program, a program to pair first-year students with upper-year students, who can be a resource for navigating the BME PhD experience
- Successfully advocated for and placed student representatives on both faculty and admission committee meetings
- Developed an online mini-course series including guidelines to courses online that are aligned with each track in the biomedical engineering department
- Launched REU/Internship Application Assistance Program to expand our efforts to help URM and first-generation students reach graduate school
- Implemented mental health and wellness education resources for students

Teaching Jan 2019 – May 2020
The best way to convey material is to first give students a grasp of the basic fundamentals underlying more advanced processes. To that end, I have taught different foundational lab classes, both benchwork and computational labs. Having the students perform experiments or work problems, either coding or in wet labs, gives students new insights. When paired to the didactic coursework already present in our curriculum this gives the students a stronger foundation to move forward in their careers.

Methods in Nucleic Acid Sequencing (Head TA and Instructor):

Sequencing technology is a rapidly progressing field that requires experience in both wet (molecular biology) and dry (computational analysis) techniques. This laboratory course consists of three experimental modules that will provide students with valuable hands-on experience in DNA sequencing and analysis. Students learn basic sequencing library preparation, perform sequencing experiments and analyze the resulting data. Experiments include human targeted sequencing, metagenomic sequencing and genome assembly.

- Designed genomics, metagenomics, and transcriptomics experiments
- Taught DNA/RNA extraction methods, NGS library preparation, Nanopore sequencing
- Deployed a local GALAXY server to host and analyze the generated sequencing data
- Developed lectures and held coding sessions

Computational Biology and Bioinformatics (TA and Instructor):

BCMB core course, taught November-January. This course challenges students to explore publicly available tools and resources that they can use in their own research, and to understand the basic molecular biology and computational concepts that the tools are built on.

- Designed lectures and held office hours
- Created problem sets (Unix, Bash, genomic data manipulation, BLAST, sequence alignment, etc)

Computational Biology Bootcamp (Head TA and Instructor):

This intensive one week class is meant to immerse students in computation, and to provide them with the foundational tools to be able to apply modern computational techniques and appropriate statistics to their data. Students learn how to work in a command line shell and different "notebook" style computing environments including Jupyter and Rmarkdown. Throughout the course, students apply these skills to different practical analysis problems for exploratory data analysis, visualization, and interpretation. The presented problems run the gamut from biophysics to cellular and systems biology to genomics.

- Designed and created the course content using Jupyter Notebooks
- Taught Python, Unix, Bash scripting, data visualization, and exploratory analysis
- Deployed the course on Microsoft Azure (JupyterHub)
- Held frequent coding sessions and office hours

Publications

* co-first author

Peer Reviewed

W. Stephenson, **R. Razaghi**, S. Busan, K. M. Weeks, W. Timp, and P. Smibert, "Direct detection of RNA modifications and structure using single molecule nanopore sequencing," *Cell genomics*, vol. 2, no. 2, p. 100 097, 2022.

A. Gershman, M. E. Sauria, P. W. Hook, S. J. Hoyt, **R. Razaghi**, S. Koren, N. Altemose, G. V. Caldas, M. R. Vollger, G. A. Logsdon, A. Rhie, E. E. Eichler, M. C. Schatz, R. J. O'Neill, A. M. Phillippy, K. H. Miga, and W. Timp, "Epigenetic patterns in a complete human genome," *Science*, 2022. DOI: 10.1126/science.abj5089.

A. Gershman, T. G. Romer, Y. Fan, **R. Razaghi**, W. A. Smith, and W. Timp, "De novo genome assembly of the tobacco hornworm moth (manduca sexta)," *G3*, vol. 11, no. 1, pp. 1–9, 2021.

A. J. Kandathil, A. L. Cox, K. Page, D. Mohr, **R. Razaghi**, K. G. Ghanem, S. A. Tuddenham, Y.-H. Hsieh, J. L. Evans, K. E. Coller, *et al.*, "Plasma virome and the risk of blood-borne infection in persons with substance use disorder," *Nature communications*, vol. 12, no. 1, pp. 1–7, 2021.

I. Lee, **R. Razaghi**, T. Gilpatrick, M. Molnar, A. Gershman, N. Sadowski, F. J. Sedlazeck, K. D. Hansen, J. T. Simpson, and W. Timp, "Simultaneous profiling of chromatin accessibility and methylation on human cell lines with nanopore sequencing," *Nature Methods*, pp. 1–9, 2020.

S. Kovaka, A. V. Zimin, G. M. Pertea, **R. Razaghi**, S. L. Salzberg, and M. Pertea, "Transcriptome assembly from long-read RNA-seq alignments with StringTie2," *Genome biology*, vol. 20, no. 1, pp. 1–13, 2019.

R. E. Workman*, A. D. Tang*, P. S. Tang*, M. Jain*, J. R. Tyson*, **R. Razaghi***, P. C. Zuzarte, T. Gilpatrick, A. Payne, J. Quick, *et al.*, "Nanopore native RNA sequencing of a human poly (A) transcriptome," *Nature methods*, vol. 16, no. 12, pp. 1297–1305, 2019.

Preprints

D. M. Tiek, **R. Razaghi**, L. Jin, N. Sadowski, C. Alamillo-Ferrer, J. R. Hogg, B. R. Haddad, D. H. Drewry, C. I. Wells, J. E. Pickett, *et al.*, "Targeting destabilized DNA G-quadruplexes and aberrant splicing in drug-resistant glioblastoma," *bioRxiv*, p. 661 660, 2019.

Conference Talks and Posters

Posters

R. Razaghi and W. Timp, "Measurement of DNA methylation and nuclear organization with nanopore sequencing," *Advances in Genome Biology and Technology (AGBT)*, 2020.

R. Razaghi, T. Gilpatrick, N. Sadowski, P. Tang, R. Workman, J. Simpson, and W. Timp, "Signal analysis of nanopore RNA sequencing to interrogate poly (a) tails and post-transcriptional modifications," *Biophysical Journal*, vol. 116, no. 3, 356a, 2019.

R. Razaghi, T. Gilpatrick, N. Sadowski, P. Tang, R. Workman, J. Simpson, and W. Timp, "Signal analysis of nanopore RNA sequencing to interrogate poly (a) tails and post-transcriptional modifications," *NHGRI Advanced Genomic Technology Development Meeting*, 2019.

Talks

R. Razaghi, P. Hook, M. Jain, K. Hansen, and W. Timp, "Shedding light on the long-range interaction of the human epigenome using ultra-long nanopore sequencing," Talk, Oxford Nanopore Technologies London Calling, London, UK, May 21, 2021.

R. Razaghi and W. Timp, "Measurement of dna methylation and nuclear organization with nanopore sequencing," Talk, Genetics Virtual Week, US, Apr. 21, 2021.

R. Razaghi and W. Timp, "Measurement of DNA methylation and lamina-associated domains with nanopore sequencing," Talk, 4th International Conference on Epigenetics and Bioengineering (EpiBio), 2020.

R. Razaghi and W. Timp, "Signal analysis of nanopore RNA sequencing to interrogate poly(A) tails and post-transcriptional modifications," Talk, ABRF 2019 Annual Meeting, 2019.