

# Nanopore sequencing for N1-methylpseudouridine in RNA reveals sequence-dependent discrimination of the modified nucleotide triphosphate during transcription

Aaron M. Fleming\* and Cynthia J. Burrows\*

Dept. of Chemistry, University of Utah, Salt Lake City, UT 84112-0850

\*To whom correspondence should be addressed. E-mail: [burrows@chem.utah.edu](mailto:burrows@chem.utah.edu) or [aflenm@chem.utah.edu](mailto:aflenm@chem.utah.edu)

## Abstract

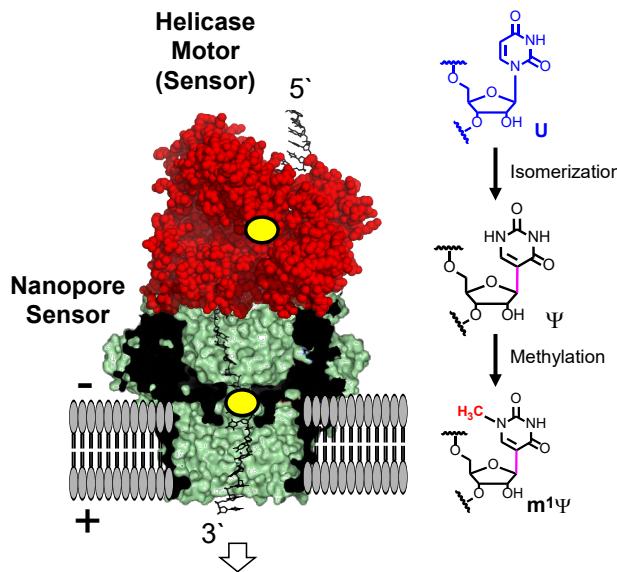
Direct RNA sequencing with a commercial nanopore platform was used to sequence RNA containing uridine (U), pseudouridine ( $\Psi$ ), or N1-methylpseudouridine ( $m^1\Psi$ ) generated by in vitro transcription (IVT). The base calling data as well as the ionic currents and dwell times for U,  $\Psi$ , or  $m^1\Psi$  as they translocated through the helicase and nanopore proteins identified diagnostic signatures for  $\Psi$  and  $m^1\Psi$ ; however, the two modifications yielded similar patterns although both were different from U. Understanding the nanopore signatures for  $\Psi$  and  $m^1\Psi$  enabled a running start T7 RNA polymerase assay to study how competing mixtures of UTP with  $\Psi$ TP or  $m^1\Psi$ TP lead to nucleotide selection in all possible adjacent sequence contexts. For UTP vs.  $\Psi$ TP,  $\Psi$ TP was favorably incorporated in singly-modified contexts, while doubly-modified contexts found high yields of  $\Psi$ TP insertion on the 5' side and lower yields on the 3' side. For UTP vs.  $m^1\Psi$ TP, UTP was favorably selected except in 5'-XA (X = U or  $m^1\Psi$ ) where the ratio was determined by their relative NTP concentrations. Experiments with chemically-modified triphosphates and DNA templates designed based on the structure of T7 RNA polymerase provide a model to explain the observations. These results may aid in future efforts that employ IVT to make therapeutic mRNAs with sub-stoichiometric amounts of  $m^1\Psi$ .

## Introduction

Inspection of native RNA across all phyla of life has identified >150 chemical modifications that include the addition of alkyl groups on the base and/or sugar, isomerization, sulfurization, oxidation, or reduction of the nucleobases (1-3). These chemical changes are found in tRNA, rRNA, mRNA, and small and large non-coding RNAs. The epitranscriptome refers to these modifications essential for the transcriptome's functional relevancy for cellular processes. Additionally, chemical modification of RNA has found its way into clinical applications where the successes of therapeutic siRNAs and mRNA vaccines were largely achieved due to the site-specific chemical decorations of the polymer (4,5). Identification and quantification of RNA modifications have been pursued by nuclease and phosphatase digestion of target strands to nucleosides for LC-MS/MS analysis resulting in the complete loss of all sequence information (1). The development of nanopores as a third-generation sequencing platform has brought about many enabling advancements, one of which is the ability to directly sequence RNA. Out of the many benefits to direct RNA sequencing is the opportunity to sequence chemical modifications.

Nanopore sequencing on a commercial platform (Oxford Nanopore Technologies) is achieved by the utility of a helicase as a motor protein to deliver the RNA into a protein nanopore at a velocity that is ATP-dependent (Figure 1) (6). An electrophoretic force serves to guide the direction of the RNA 3' to 5' into the nanopore protein where a small central constriction zone exists. As the strand passes the constriction zone with a length of 5-nt of RNA (i.e., k-mer), the ionic current changes as a function of sequence (7). In recent iterations of base-calling software, the current vs. time traces are deconvoluted via a recurrent neural network (8). Exciting developments using nanopores have showcased direct RNA sequencing for modifications in tRNA, rRNA, mRNA, small/large RNAs, and viral RNAs from biological sources (7,9-15). The predominant native RNA modifications inspected include pseudouridine ( $\Psi$ ),  $N^6$ -methyladenosine, and 2'-O-methyl nucleotides. Our work with  $\Psi$  demonstrated base calling

error analysis combined with current and dwell time analysis minimizes the false discovery rate for modification detection in noisy nanopore data (12). Some challenges remain before this approach becomes a routine technique in the RNA researcher's toolbox.



**Figure 1.** The Oxford Nanopore Technologies platform sequences U, Ψ, and  $m^1\Psi$  directly in RNA.

In the present work, the commercial nanopore sequencer was employed to directly sequence N1-methylpseudouridine ( $m^1\Psi$ ) in RNA (Figure 1). This modification results from isomerization and methylation of the uridine nucleotide, and was first discovered in archaea tRNA by the McCloskey laboratory (16). Since then, this modification was found to be essential for the successful use of mRNA vaccines to defend against SARS-CoV-2 infections (5). These synthetic mRNA vaccines are produced by *in vitro* transcription (IVT), in which all U sites are converted to  $m^1\Psi$  by feeding T7 RNA polymerase  $m^1\Psi$ -nucleotide triphosphate ( $m^1\Psi$ TP) instead of UTP (5). Using the same approach for RNA production as the mRNA vaccines, synthetic RNAs with  $m^1\Psi$  or Ψ were prepared and nanopore sequenced herein. The obtained data was inspected for changes in the ionic current, dwell time, and base calling profile between the modified RNAs and the U-containing RNA control strand of the same sequence. This analysis led to the design of

new RNAs using the N1-ethyl or N1-propyl derivatives of  $\Psi$  to provide trending data regarding the nanopore signatures. These foundational results for the signals associated with  $\Psi$  and  $m^1\Psi$  enabled a study of T7 RNA polymerase selection of the modified nucleotide triphosphate when the canonical UTP was present. All adjacent sequence contexts were studied. The nanopore sequencing found many unexpected sequence-context effects for canonical vs. non-canonical nucleotide triphosphate selection by the RNA polymerase. The present direct RNA sequencing via nanopores provided new insight into the clinically relevant modification  $m^1\Psi$ , and how T7 RNA polymerase selects mixed competing NTPs during mRNA synthesis. These final results will be of critical importance for future manufacturing of therapeutic RNAs made by in vitro transcription that aim to have sub-stoichiometric incorporation of modified nucleotides such as  $m^1\Psi$ .

## Materials and Methods

***RNA synthesis by in vitro transcription.*** In vitro transcription was performed using the MEGAscript T7 transcription kit according to the manufacturer's instructions. The duplex DNA templates for the IVT reactions were synthesized via commercial sources to have a T7 promoter for initiation of transcription and ended with a poly-A tail for sequencing library preparation (Figure S1). The IVT reactions were incubated overnight at 37 °C in a PCR thermocycler. After the overnight incubation, turbo DNase I treatment was performed on all samples at 37 °C, followed by purification using Quick Spin Columns for radiolabeled RNA purification. To install  $\Psi$  or  $m^1\Psi$ , IVT was conducted in the presence of pseudouridine-5'-triphosphate ( $\Psi$ TP) or  $m^1\Psi$ TP instead of UTP. Success in the synthesis of the RNA transcripts was verified by agarose gel electrophoresis by comparison to a ladder of known lengths. All experiments were conducted in duplicate.

***T7 RNA polymerase studies with mixed NTPs.*** Studies on NTP discrimination by T7 RNA polymerase were conducted using the MEGAscript T7 transcription kit with some changes to the manufacturer's protocol as described. The NTP concentrations were 2 mM for ATP, GTP, and CTP, while the UTP and  $\Psi$ TP or  $m^1\Psi$ TP (Trilink Biotechnologies) were 1 mM each to achieve

a total concentration of U and its derivatives of 2 mM. To ensure the UTP and  $\Psi$ TP or  $m^1\Psi$ TP were at a 1:1 ratio, the stock solution concentrations were determined using UV-vis spectroscopy with established extinction coefficients (UTP:  $\lambda_{262\text{ nm}} = 10,000\text{ L mol}^{-1}\text{cm}^{-1}$ ;  $\Psi$ TP:  $\lambda_{262\text{ nm}} = 7,550\text{ L mol}^{-1}\text{cm}^{-1}$ ;  $m^1\Psi$ TP:  $\lambda_{271\text{ nm}} = 8,870\text{ L mol}^{-1}\text{cm}^{-1}$ ). The IVT reactions were allowed to progress for 2 h at 37 °C before termination. Addition of turbo DNase I following the manufacturer's protocol terminated the reaction. Further studies on T7 RNA polymerase NTP selection were conducted with N1-ethylpseudouridine triphosphate ( $e^1\Psi$ TP:  $\lambda_{271\text{ nm}} = 7,800\text{ L mol}^{-1}\text{cm}^{-1}$ ) or N1-propylpseudouridine triphosphate ( $p^1\Psi$ TP:  $\lambda_{271\text{ nm}} = 8,900\text{ L mol}^{-1}\text{cm}^{-1}$ ; Trilink Biotechnologies). All experiments were conducted in duplicate.

**Nanopore library preparation and sequencing.** The poly-A tail containing RNAs generated by IVT were the input strands in the direct RNA sequencing kit (SQK-RNA002) from Oxford Nanopore Technologies (ONT). The protocol was followed without changes and the library prepared samples were directly used for sequencing. The samples were applied to the ONT Flongle™ flowcell running the R9.4.1 chemistry following the manufacturer's protocol with an applied voltage of 180 mV. All sequencing experiments were conducted until all nanopores were no longer functional that generally yielded >15,000 reads.

**Data analysis.** The ionic current vs. time traces in fast5 file format were base called using guppy v.6.0.7 to obtain the fastq sequencing read files used in the subsequent data analyses. The fastq files were aligned to the reference sequences using minimap2 ‘-ax map-ont -L’ to generate bam files (17). The bam file alignment statistics were determined with the flagstat function in Samtools and then the files were indexed with Samtools for visualization in Integrative Genomics Viewer (IGV) to obtain the base call information at the modification sites (18,19). The Nanopore-Psu, Nanocompore, and Tombo tools were used as described on their Github sites (7,11,20). The current and dwell time data were extracted, resquiggled to the reference, and

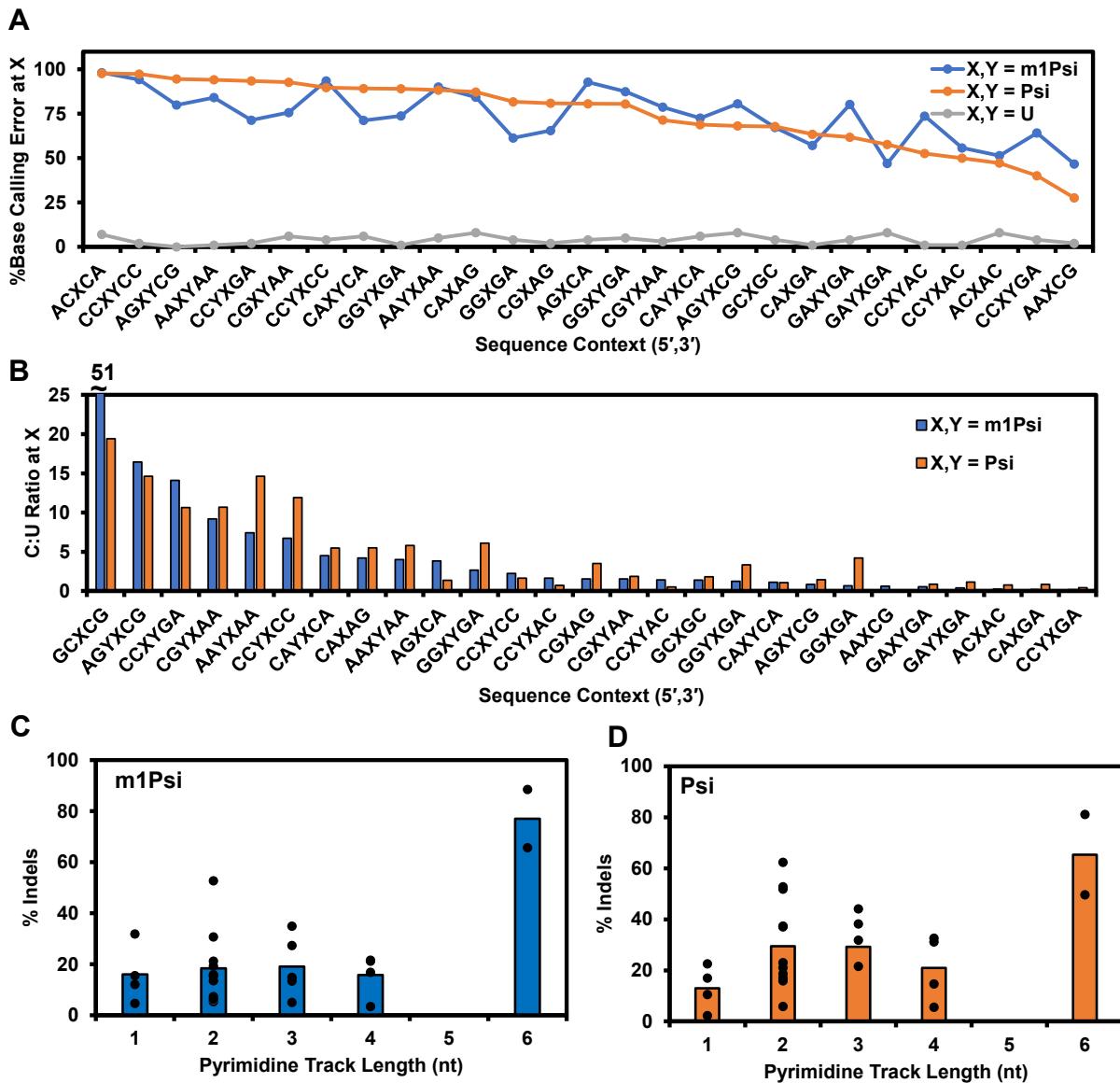
analyzed using Nanopolish as described in the user manual (21). The data were plotted and analyzed in either python, Origin, or Excel for visualization.

## Results and Discussion

**Experimental setup.** The RNA to address the performance of the nanopore sequencer on  $m^1\Psi$  with comparisons to U and  $\Psi$  were synthesized by IVT, like the present approach for mRNA vaccine development (Figure S1) (5). The use of commercially available  $m^1\Psi$ TP or  $\Psi$ TP allowed the generation of RNAs with 100% incorporation of the modified nucleotides. The RNA strands sequenced had the U,  $\Psi$ , or  $m^1\Psi$  spaced >18 nts apart such that only one site of analysis interacted with the nanopore-helicase setup at a time. A limitation to this approach is sequence contexts with the parent nucleotide U cannot be generated. The goal is to study T7 RNA polymerase activity in the presence of modified and canonical nucleotides to measure selectivity imposed by the enzyme; therefore, mixed sequence contexts with U and  $\Psi$  or  $m^1\Psi$  is not an issue. All possible 5' and 3' adjacent sequence contexts flanking a single site and sites for double incorporation of the modification were designed into the duplex DNA template used for transcription of the RNA. The nanopore sequencer using the CsgG protein nanopore (v 9.4.1 flow cells) has a sensing window for RNA (i.e., k-mer) of ~5 nt (7). The RNA sequences studied partially covered the sequence space for the nanopore k-mer; however, not all sequence combinations were studied.

**Base-calling error analysis.** Nanopore sequencing was conducted on the RNA strands following standard protocols to generate current vs. time traces that were base called with Guppy (v 6.0.7), followed by alignment to the reference with minimap2, and visualization of the base calls was achieved with IGV. First, the base-calling error identified by the presence of calls for C, A, G, or insertion and deletions (indels) for the U,  $\Psi$ , and  $m^1\Psi$  RNAs were quantified. In the data for sequences with two U,  $\Psi$ , or  $m^1\Psi$  adjacent to one another, each of the two in the pair gave unique base call errors; thus, in the plots shown in Figure 2, the modified sites in the sequences

are shown with the letters X and Y, where the data above the sequence are for the position labeled X. The U-containing RNA strands gave percent errors <10% for the sequence contexts studied (Figures 2A gray line). The percent error for each of the sequence contexts possessing  $\Psi$  was quantified and ordered from high to low (Figure 2A orange line) to find that the error ranged from nearly 100% down to ~20% with a strong dependency on the sequence context. The percent base calling error for each  $m^1\Psi$ -containing context was plotted (Figure 2A blue line). The data were not identical to those observed for  $\Psi$ , but they did show a similar trend. The base call analysis identifies modification of U to either  $m^1\Psi$  or  $\Psi$  generates base called data with much higher error than U.



**Figure 2.** The presence of  $\Psi$  and  $m^1\Psi$  yield sequence-dependent base-calling errors during direct RNA nanopore sequencing. (A) Plots of sequence context vs. percent total error, which include calls to A, C, G, and indels, for U (grey),  $\Psi$  (orange), and  $m^1\Psi$  (blue). The values were generated from >1000 aligned reads. (B) A plot of sequence context vs. the C:U base calling ratio from the data in panel A for  $\Psi$  (orange) and  $m^1\Psi$  (blue). The percent indels vs. pyrimidine track length from the panel A data for (C)  $m^1\Psi$  and (D)  $\Psi$ . Some of the C:U base calling ratio data were previously reported by our laboratory (12) and have been reanalyzed for the present comparisons.

The base calling data were then inspected more closely to identify the neural network base caller error modes at the modified sites. Previous studies found  $\Psi$  is miscalled as a C (10-

12,14). The C to U ratio for  $\Psi$  and  $m^1\Psi$  in each context were compared to find they produced similar sequence dependency in the C relative to U calls (Figure 2B). Sequence-dependent trends in the base calling percent error or base call ratio were not observed. Next, the indel profiles for each context bearing  $\Psi$  or  $m^1\Psi$  were quantified. The literature identified indels increase on the v 9.4.1 flow cell chemistry of the ONT platform in homopolymer sequence runs (22); with this in mind and knowing that the modifications can code like a C, the indel percentage was plotted against the length of the pyrimidine track flanking the site. These data do not enable the study of all possible sequence contexts, although when the pyrimidine track length was 6 nucleotides long, indel signatures for  $\Psi$  and  $m^1\Psi$  increased. The increased frequency of indels observed in the reported data for 2-nt long tracks results from the over-representation of sequences with two modifications adjacent to one another in the sequence.

The key finding in the base calling analysis was that  $\Psi$  and  $m^1\Psi$  cause changes to the raw data that the neural network base caller interprets to yield similar base-calling errors. Direct RNA nanopore sequencing for  $\Psi$  has been reported and tools developed to quantify the data to determine the extent of occupancy at modified sites (10-15). Many of these approaches rely on base calling features. The dominant and unique base calling signature for  $\Psi$  is to be called as a C (Figure 2B), which was found herein and previously (10,12,14,15). This approach in the analysis is subject to false positives when natural U→C variations exist in the RNA that can masquerade as a  $\Psi$  when inspecting base calling features exclusively; a solution to this challenge is described below (12).

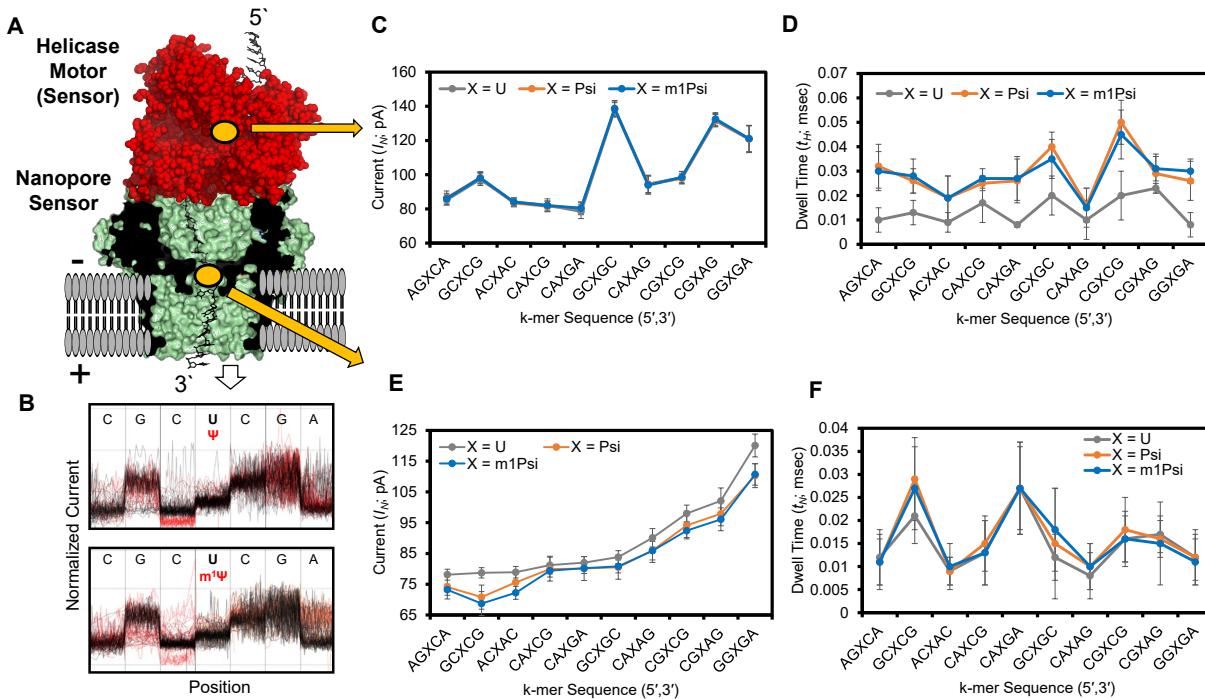
The present work with  $m^1\Psi$  identified its signature in the nanopore base calling data is like  $\Psi$ . This observation points to another challenge for direct RNA sequencing for modifications with nanopores, and that is different modifications can yield very similar signatures. After exploring the nanopore signatures for the >150 known modifications, the community might find, for example, that a particular U is modified in an unknown sample but the question of what the modification is

will remain. Parallel biological and sequencing studies on native cells and those with suspected writer protein knockouts or knockdowns can aid in the identification of the modification (7,10). Alternatively, the use of low-throughput and targeted assays such as SCARLT or mass spectrometry sequencing will be needed to identify the modification, and even these will be challenged by modification isomers (e.g.,  $m^5C$ ,  $m^4C$ , and  $C_m$ ) (23,24).

***Ionic current level and dwell time analyses.*** Because the base calling data are derived from the current changes measured as the modification moves through the sensor, a question is whether  $\Psi$  and  $m^1\Psi$  have similar current profiles. Another finding from our prior work was that  $\Psi$  impacted the helicase motor protein translocation kinetics to provide a second signal to follow the presence of the modification (12), which was confirmed in another study (25). We then inspected the current vs. time traces for U,  $\Psi$ , and  $m^1\Psi$  to measure how they differ regarding the residual current levels and dwell times when the subject sites were in the nanopore, and the residual current levels and dwell times when the sites were in the helicase active site.

In Figure 3B, a stacked plot of current levels for either  $\Psi$  or  $m^1\Psi$  relative to U obtained when they are in the nanopore sensor was generated using the Tombo tool (20). All other Tombo-generated plots are shown in Figure S2. This current-level analysis illustrates a very important point regarding nanopore sequencing for modifications. The nanopore constriction houses ~5 nts of RNA called a k-mer and the change in current as a result of the modification does not have to occur when it is positioned in the center of the k-mer; as these examples show, the largest deviation in current occurred when the C 5' to the modification was centered in the k-mer. As a consequence, each sequence context containing U or either of the two modifications was first analyzed visually with Tombo and statistically with Nanocompare (7) to locate the sites of the largest current deviation in the nanopore and dwell time deviation in the helicase for the plots provided. To avoid confusion, the k-mer sequences in all plots have the modification site in the center, while the actual k-mers analyzed are provided in the figure legend. This analysis was

restricted to sequence contexts that had one modification because those with two generally induced current and dwell time changes that occurred at multiple locations (Figure S2). Single  $\Psi$  and  $m^1\Psi$  modifications were analyzed in all possible adjacent sequence contexts that do not include the parent nucleotide U (9 contexts), and the 5'-GXC context was studied in two different 5-nt k-mers (5'-AGXCA and 5'-CGXCG where X = U,  $\Psi$ , or  $m^1\Psi$ ).



**Figure 3.** Analysis of currents and dwell times as sites of single U,  $\Psi$ , or  $m^1\Psi$  resided in the nanopore or helicase sensors. (A) A structural view of the nanopore sequencer set up. (B) Comparative current level analysis for U vs.  $\Psi$  (top) or  $m^1\Psi$  (bottom) generated by Tombo. These plots illustrate that the maximal current level change does not always occur when the modification is centered in the central constriction zone of the CsgG protein. Additionally, Tombo plots compress the time domain on the x-axis for visualization; therefore, the x-axis or position does not reflect the true residence times. (C) Comparative plots of the helicase current levels ( $I_H$ ) and (D) dwell times ( $t_H$ ) for the sequence context studied. (E) Plots for comparison of the nanopore current levels ( $I_N$ ) and dwell times ( $t_N$ ) for the sequence context studied. The position of the maximal differences between U and the modifications was derived from Tombo and Nanocompare analysis. The actual k-mers analyzed for the nanopore sensor  $I_N$  and  $t_N$  values are provided in parentheses written 5' to 3' where X = U,  $\Psi$ , or  $m^1\Psi$ : AGXCA (GXCAC), GCXCG (GCXCG), ACXAC (XACCA), CAXCG (ACAXC), CAXGA (GCAXG), GCXGC (GCXGC), CAXAG (CAXAG), CGXCG (CGXCG), CGXAG (ACGXA), and GGXGA (XGACA).

The current and dwell values were obtained from Nanopolish resquiggled data (i.e., aligning base calls with current levels). Previously it was found that resquiggling with all available software (Tombo or Nanopolish) can filter out data and lead to biases in ways that are not well understood (10), which was confirmed herein (Figure S3); however, no other tools are available for this type of analysis. With this known issue in mind, we proceeded with caution and avoid overarching claims regarding the data. The  $I_H$ ,  $\Psi$ , or  $m^1\Psi$  first interact with the sequencer in the helicase where the residual current ( $I_H$ ) and/or average dwell time ( $t_H$ ) can be impacted. The  $I_H$  values for U,  $\Psi$ , and  $m^1\Psi$  for each sequence context were nearly identical (Figure 3C). This is consistent with the fact that the source of the current deviations is in the nanopore, not the helicase. When the  $t_H$  values were inspected for U,  $\Psi$ , or  $m^1\Psi$  in the helicase, the times were longer for  $\Psi$  and  $m^1\Psi$  relative to U (Figure 3D). This is consistent with our previous finding regarding the helicase stalling on  $\Psi$  (12). A comparison of the average  $t_H$  for  $\Psi$  and  $m^1\Psi$  for the contexts studied found they are similar (Figure 3D).

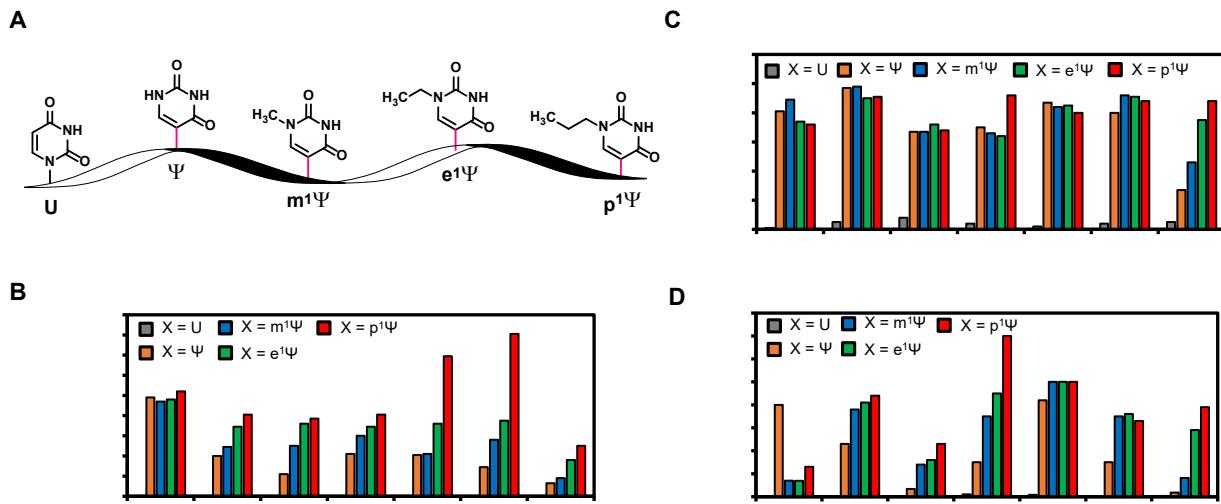
Next, the interrogated site interacts with the nanopore sensor in which the residual currents ( $I_N$ ) and average dwell times ( $t_N$ ) were measured. The  $I_N$  values for the k-mers found changes between U vs.  $\Psi$  or  $m^1\Psi$  (Figure 3E). All U-containing k-mers had greater  $I_N$  values relative to  $\Psi$  or  $m^1\Psi$  in the same context, which shows the modifications are more blocking to the current. The  $I_N$  values between  $\Psi$  and  $m^1\Psi$  were similar. The average  $t_N$  values for the three nucleotides in each sequence context gave small differences (Figure 3F), but they were less than those observed for the modifications in the helicase. The key regions  $\Psi$  and  $m^1\Psi$  impact the recorded data are in the dwell time while situated in the helicase active site and in current level while located in the nanopore central constriction zone. The changes in  $I_H$ ,  $t_H$ ,  $I_N$ , and  $t_N$  for each of the singly-modified k-mers were inspected for correlations with the base calling error, C:U ratio, and indel frequency using a Spearman's rank-order correlation test. The analysis found for  $\Psi$  the

base calling error and C:U ratio correlated with  $I_N$  ( $\rho = 0.88$  and  $\rho = 0.84$ ) but not with the other physical values measured; in contrast, indel frequency for  $\Psi$  did not correlate with any parameter measured (Figure S4). The correlation analysis for  $m^1\Psi$ , found  $I_N$  correlated with base calling error ( $\rho = 0.61$ ), while  $I_N$  and  $t_N$  correlated with the C:U ratio ( $\rho = 0.58$  and  $\rho = -0.61$ ; Figure S4). These correlations are not surprising; however, they led to a testable hypothesis as described below.

The nanopore sequencer is unique because proteins are used in a setup to generate ionic current vs. time traces for DNA or RNA as they pass through the platform under an electrophoretic force. The nanopore protein and helicase motor can function differently on RNA with chemical modifications compared to the all-canonical strands to provide signatures for the presence of these changes. The U modifications  $\Psi$  and  $m^1\Psi$  impacted the dwell time in the helicase sensor and the current level in the nanopore sensor to provide more data for confident calls of the modifications. Inspection of the current vs. time data can be beneficial when conducting de novo sequencing of RNA for modifications. Using  $\Psi$  and  $m^1\Psi$  as an example, they are base called like C nucleotides; therefore, natural U→C variations would exist as false positives when looking for these modifications in base calling analysis. An approach to differentiate modifications from sequence variations is to look at the helicase dwell time  $t_H$ , as we previously proposed (12).

**Study of N1-methyl-, N1-ethyl-, and N1-propyl- $\Psi$  derivatives.** Two main motivations led us to study the N1-methyl, N1-ethyl, and N1-propyl derivatives of  $\Psi$  (i.e.,  $m^1\Psi$ ,  $e^1\Psi$ , and  $p^1\Psi$ ; Figure 4A) with the nanopore sequencer. First, the previous observation that the nanopore current values  $I_N$  showed a correlation with base calling features for  $\Psi$  and  $m^1\Psi$ . This led us to wonder whether systematically increasing the alkyl size would impact the  $I_N$  value leading to a trending change in the base calling features. Said another way, can systematic chemical changes to the base result in expected changes in how the neural network base caller identifies bases. Second, a patent filed for use of nucleotide modifications for mRNA vaccines has included claims

with methyl, ethyl, and propyl derivatives (26), and therefore, nanopore sequencing of these modifications may have relevance in the future. Moreover, nucleotide triphosphates for N1-ethyl and N1-propylpseudouridine are commercially available.



**Figure 4.** Nanopore sequence analysis of the N1-alkyl derivatives of  $\Psi$ . (A) The structures of U,  $\Psi$ ,  $m^1\Psi$ ,  $e^1\Psi$ , and  $p^1\Psi$ . (B) The difference between the U  $I_N$  value and the N1-alkyl  $\Psi$   $I_N$  values ( $\Delta I_N = U I_N - X I_N$ ; where  $X = \Psi$ ,  $m^1\Psi$ ,  $e^1\Psi$ , or  $p^1\Psi$ ). (C) The percent base calling error, and (D) the C:U base call ratio for each of the nucleotides studied. These studies were conducted on seven different sequence contexts where  $> 500$  single-molecule measurements from two replicates provided the average values reported.

A subset of singly-modified sequence contexts in RNA was synthesized by IVT, libraries were prepared and then sequenced by the nanopore. The data were resquiggled with Nanopolish to then find the values  $I_N$ ,  $t_N$ ,  $I_H$ , and  $t_H$  for each context and N1-alkyl  $\Psi$  derivative. Noteworthy, as the alkyl group size increased the percentage of reads lost during resquiggling with Nanopolish increased (Figure S3). Figure 4B provides a plot of the difference in  $I_N$  values between the U-containing RNA with each N1-modified  $\Psi$  studied (i.e.,  $\Delta I_N = [U I_N - \text{modified } \Psi I_N]$ ). All current and dwell time average values with the associated errors are provided and described in Figure S5. Out of the seven sequence contexts studied, six showed a trend in the  $\Delta I_N$  values with the size of

the alkyl group. The exception was a modification in the context 5'-GGXGA that produced a large deviation from U to  $\Psi$ , for which the alkyl group size did not change this already large difference. Analysis of the percent base calling error for the modifications in the different sequence contexts did not identify a visible trend, except the k-mer 5'-CAXAG. This standout context is the one that showed the least base calling error for  $\Psi$ , and therefore, there was room for an increased change to be observed for the increasing alkyl group size on N1. Lastly, the C:U base calling ratio was inspected, and in all but one case, the value increased going from  $\Psi$  to the N1-alkyl derivatives, and the change varied with sequence context. The context 5'-GGXGA was the one that gave a decrease in the C:U ratio for  $\Psi$  and its derivatives, and the reason for this is not known.

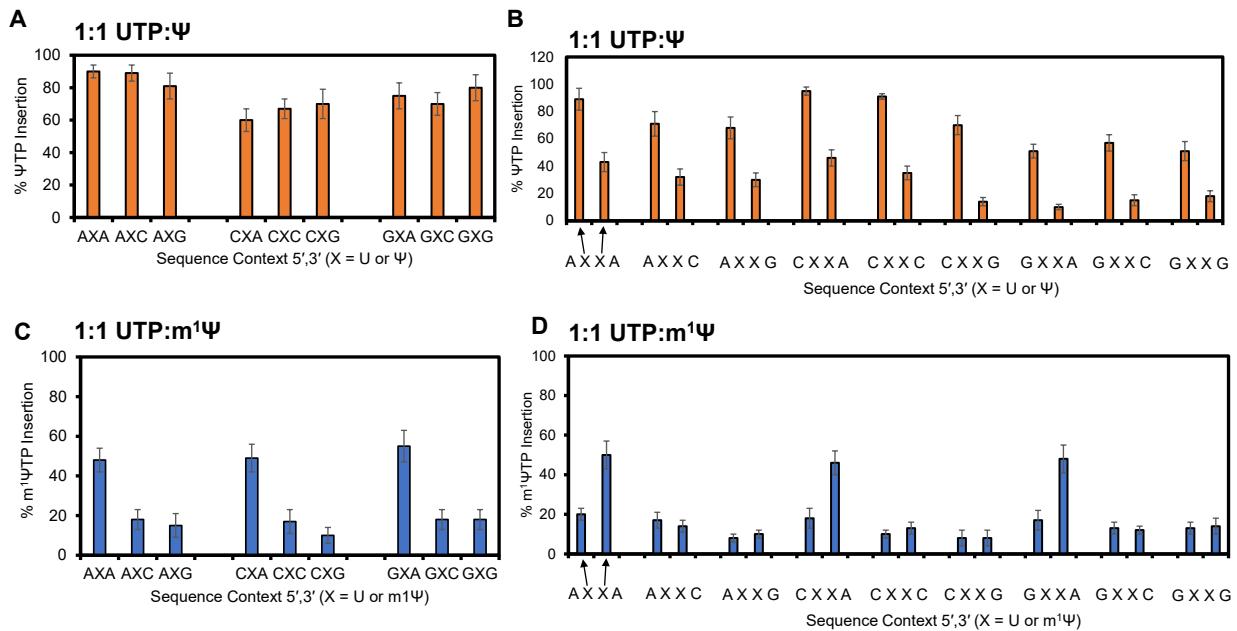
These data identify the series  $\Psi$ ,  $m^1\Psi$ ,  $e^1\Psi$ , and  $p^1\Psi$  generate nanopore-derived base calling error differences from U that can be followed to identify their presence in an RNA strand. Additionally, the currents and dwell times for the N1-alkyl  $\Psi$  derivatives do change (Figure S5), but the data are subject to loss during the resquiggling process resulting in a low confidence in the values obtained (Figure S3). Hopefully, future efforts on the computational side of nanopore sequencing analysis will find a solution to this limitation. These data may assist future efforts that use selective chemistry to add chemical reporter groups to target modifications for nanopore sequencing (27). This pilot study provides promise for this approach but reminds us that sequence context will influence the results in unexpected ways.

**T7 RNA polymerase NTP selection for mixtures of UTP with  $\Psi$ TP or  $m^1\Psi$ TP.** In the final set of studies, we used our knowledge of nanopore sequencing signatures for  $\Psi$  or  $m^1\Psi$  to explore NTP selection by T7 RNA polymerase during IVT. Currently, mRNA vaccines are produced by T7 RNA polymerase mediated transcription, in which all the U nucleotides are completely replaced with  $m^1\Psi$  (5). A prior study found the partial replacement of the canonical nucleotides with modified forms ( $m^5C$  and  $s^2U$ ) in therapeutic mRNAs generated by IVT can be effective (28); however, the studies did not know whether there existed sequences that T7 RNA

polymerase favored or disfavored insertion of the non-canonical nucleotide triphosphates. Therefore, we addressed this question for UTP vs.  $\Psi$ TP or  $m^1\Psi$ TP mixtures during T7 RNA polymerase synthesis of an mRNA. The duplex DNA template provided coding potential to interrogate all possible immediate sequences contexts in singly-modified and doubly-modified contexts.

The RNA polymerase evaluation for NTP selection was conducted with a one-to-one ratio of UTP to  $\Psi$ TP or UTP to  $m^1\Psi$ TP. To ensure the competing NTPs were mixed in a one-to-one ratio, the stock solution concentrations were determined by UV-vis spectroscopy using extinction coefficients provided by the manufacturer of the nucleotides. The RNA strands were synthesized and then sequenced with the nanopore. The synthesized RNAs were analyzed by agarose gel analysis to confirm the full-length extension of the template (Figure S6). Quantification of the  $\Psi$  or  $m^1\Psi$  insertion was achieved using a recently published tool, Nanopore-Psu (11), which was trained on base calling data to use the associated errors for prediction of the modification status at each site. To ensure the best accuracy in quantification, and because the tool was designed for  $\Psi$ , not  $m^1\Psi$ , we used the aligned nanopore reads to mix them in known ratios of U with  $\Psi$  or  $m^1\Psi$  to develop calibration curves (Figure S7). This approach allowed the repurposing of Nanopore-Psu for occupancy prediction of another U modification (i.e.,  $m^1\Psi$ ) in the sequence contexts studied.

The U vs.  $\Psi$  competition for the T7 polymerase active site in singly-modified contexts found  $\Psi$  was favorably installed with a 60-90% yield (Figure 5A). In the same competition but for sites in which the template DNA strand codes for insertion of two adjacent U/ $\Psi$  residues, something very interesting was found. In each context, the 5' site favored  $\Psi$ TP insertion (>50%) while the 3' site gave a >2-fold reduction in  $\Psi$ TP insertion (Figure 5B). This suggests insertion of the first modified nucleotide influences the insertion of the second.



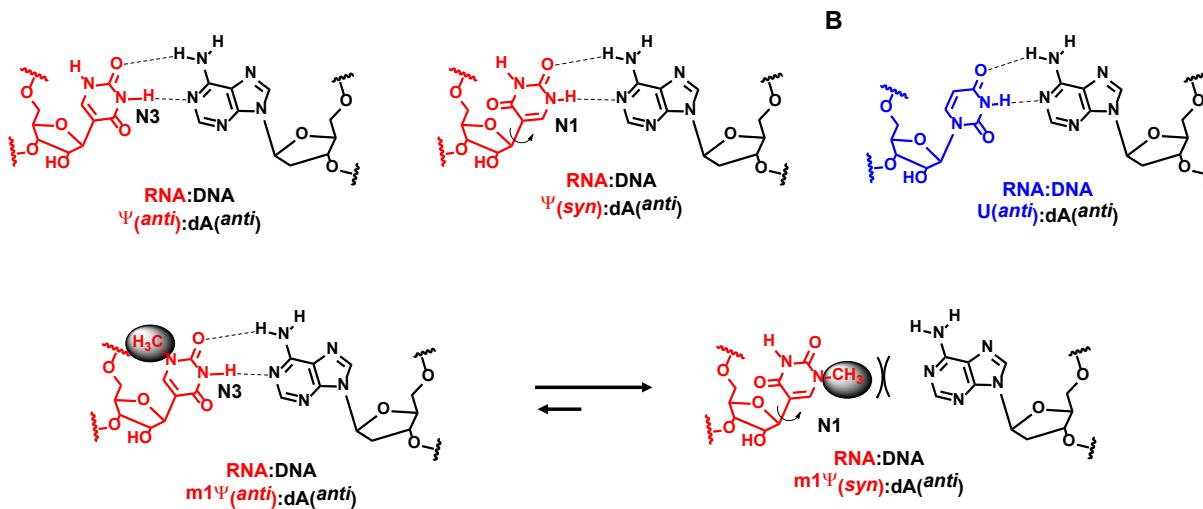
**Figure 5.** Yields of  $\Psi$ TP or  $m^1\Psi$ TP incorporation when competed with UTP for insertion and elongation by T7 RNA polymerase. Percent insertion yields for  $\Psi$ TP in (A) singly-modified or (B) doubly modified sequence contexts. Percent insertion yields for  $m^1\Psi$ TP in (C) singly-modified or (D) doubly-modified sequence contexts. The yields were determined via direct RNA sequencing with the commercial nanopore platform and the base calling data were analyzed with the published tool Nanopore-Psu (11).

Nanopore sequencing results from the competition between UTP and  $m^1\Psi$ TP for T7 RNA polymerase incorporation were even more different. In sites where the template codes for a single modification, UTP was installed with >80% yield with one exception; when the template DNA coded for ATP insertion 3' to the modified nucleotide, T7 RNA polymerase inserted UTP and  $m^1\Psi$ TP with the frequency at which they were mixed in solution (i.e., 1:1; Figure 5C). Last, sites that could incorporate two UTP or  $m^1\Psi$ TP adjacent to one another found UTP was the major triphosphate incorporated except when there was a 3' ATP inserted in the strand, which gave a nearly 1:1 ratio of the two competing pyrimidines being inserted (Figure 5D). The ability to directly sequence RNA with nanopores and quantitatively call modifications led to the discovery of the sequence-dependent bias that T7 RNA polymerase has for the selection of UTP vs.  $\Psi$ TP or  $m^1\Psi$ TP.

Using the Nanopore-Psu tool to analyze direct RNA nanopore sequencing data, we studied T7 RNA polymerase selection of competing NTPs in many sequence contexts in a running start assay (Figure 6A). To address these results a better understanding of T7 RNA polymerase is needed. The phage RNA polymerase T7 is a single subunit DNA-dependent RNA polymerase, it does not have an exonuclease domain for proofreading, and with these limitations, the polymerase maintains high fidelity transcription (1 error in  $10^4$  NTPs polymerized) (29). How T7 RNA polymerase maintains high fidelity RNA synthesis has been addressed by x-ray crystallography and computational studies (30-34). Experimental work identified that T7 RNA polymerase easily accepts modified NTPs (35), NTP elongation kinetics are influenced by steric factors (36), and NTP selection is impacted in solutions where the dielectric constant and water activity have been altered with PEG-200 (37). The key points in discrimination of the incoming NTP are to ensure the exclusion of dNTPs by checking the presence of the 2'-OH group, and that the NTP and templating DNA nucleotide form a viable base pair to maintain the fidelity of the mRNA code. Why does T7 RNA polymerase discriminate between UTP vs.  $\Psi$ TP or  $m^1\Psi$ TP, of which can both form viable base pairs with the dA nucleotide in the template DNA strand?

**T7 RNA polymerase selectivity for  $\Psi$ TP when competing with UTP.** When UTP competed with  $\Psi$ TP for incorporation in singly modified sites, the yield of  $\Psi$ TP incorporation was >60% (Figure 5A). We hypothesize the reason for the favorability of  $\Psi$ TP incorporation is the greater freedom for *syn* and *anti*-glycosidic bond angles of the base relative to the sugar in this C-nucleoside that was found to favor the *syn* conformation (38). Both conformations of  $\Psi$ TP display a face to the dA nucleotide that can form two hydrogen bonds with similar base-pair shapes (Figure 6A). Thus,  $\Psi$ TP is favorably inserted because both base conformations *syn* or *anti* relative to the sugar yield viable dA base pairs, whereas UTP can only base pair with dA in the *anti* conformation (Figure 6B). This claim is supported by modeling work on T7 RNA

polymerase that found phosphodiester bond formation for polymerization can only occur when the base pair has the right size and shape (34). When the template DNA strand had two adjacent dA nucleotides to direct insertion of the competing NTPs,  $\Psi$ TP was favored for incorporation on the 5' site but was disfavored by >2-fold for the 3' site in all sequence contexts. A possible explanation for this observation comes back to the *syn*- and *anti*-glycosidic bond conformations for  $\Psi$ . Installation of a *syn*  $\Psi$  at the 5' site based on these data disfavors a second non-canonical nucleotide from being incorporated at the 3' site. The 5' *syn*  $\Psi$  lacks a carbonyl group in the minor groove that is known to be important in DNA polymerases for progression (39). Detailed structural studies will need to be conducted to address this structural hypothesis for two adjacent  $\Psi$  incorporations by T7 RNA polymerase.



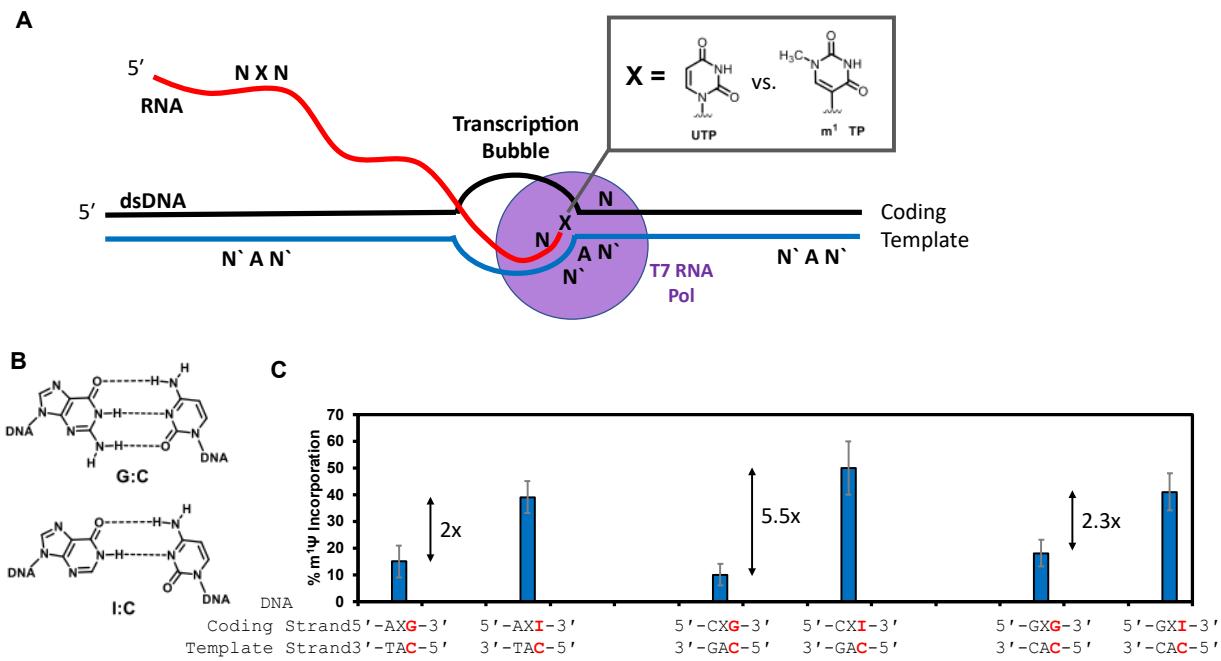
**Figure 6.** Base pairs formed by (A)  $\Psi$ , (B) U, and (C)  $m^1\Psi$  in RNA with a dA nucleotide in a template DNA strand.

**T7 RNA polymerase selectivity for UTP competed with  $m^1\Psi$ TP.** Considering  $m^1\Psi$ TP vs. UTP for incorporation by T7 RNA polymerase, the findings differed considerably compared to competitions with  $\Psi$ TP. In the sequence contexts that coded for a single U/ $m^1\Psi$ , UTP was favorably selected by >4-fold (Figure 5B). The sequence context exceptions were those with a 3' A in the RNA that gave U and  $m^1\Psi$  incorporation at a 1:1 ratio reflected by their concentrations in

solution. There are two mysteries regarding these observations. (1) Why does UTP outcompete  $m^1\Psi$ TP in all sequence contexts except one? (2) Why does a 3' A in the RNA result in higher yields for the modified nucleotide triphosphate?

The Chow laboratory used NOE NMR measurements to report on the *syn* vs. *anti* conformational equilibrium for  $m^1\Psi$  in the nucleoside context (38). They found  $m^1\Psi$  had a greater preference for the *syn* conformation than  $\Psi$ . Unlike  $\Psi$ ,  $m^1\Psi$  *syn* cannot base pair with dA in the DNA template strand (Figure 6C); therefore, the dominant glycosidic bond conformation for this modification is not compatible with T7 RNA polymerase to catalyze phosphodiester bond formation. This results in UTP outcompeting  $m^1\Psi$  for incorporation and elongation by this DNA-dependent RNA polymerase.

Regarding the 3' A effect where  $m^1\Psi$ TP and UTP were selected based on their solution concentrations, the structures for the steps of NTP selection and polymerization by T7 RNA polymerase offer a clue that can be tested (30-34). The active site of T7 RNA polymerase is maintained in an open conformation by the O-, O'-helix during NTP selection that closes to generate the contacts needed for NTP discrimination and phosphodiester bond formation (30,33). Computational studies suggest closing of the O-, O'-helix is thermally regulated and that the open conformation provides space for NTPs to sample glycosidic bond conformations to yield viable base pairs for polymerization in the growing mRNA strand (32). In the open and closed conformations, the DNA base pair 3' to the templating nucleotide is still formed and a pi stack between F644 of the O, O'-helix, and the template DNA nucleotide exists (Figure 7A). In this case, the pi stack would be with a dT on the template DNA strand to form the weakest interactions possible, a pi stack with a pyrimidine in a two-hydrogen bond base pair. This may allow greater sampling of the open and closed states of the O-, O'-helix providing space and time for  $m^1\Psi$ TP to find the *anti* conformation to base pair with the template dA (Figure 6C).



**Figure 7.** The T7 RNA polymerase structure during transcription provides a testable hypothesis why UTP and  $m^1\text{TP}$  are selected equally when an A is 3' to the site in RNA. (A) A cartoon model of transcription by T7 RNA polymerase focusing on the base pairs. (B). Structures for the base pairs dG:dC and dI:dC. (C) Yields for  $m^1\text{TP}$  incorporation when a dI:dC base-pair follows the site of competition to give yields similar to the situation with a dA:dT base pair at that position.

To study this polymerase-substrate active site structural proposal, a duplex DNA was designed, synthesized, and used for IVT that had a DNA templating dT nucleotide for the UTP vs.  $m^1\text{TP}$  competition followed by a dC. The change was in the coding DNA strand where the native dG that forms three hydrogen bonds with dC was replaced with 2'-deoxyinosine (dI) for pairing with two-hydrogen bonds and pi stacking (dC) interaction with F644 but direct the insertion of a G 3' to the U/ $m^1\text{U}$  competition site (Figure 7B). The nanopore sequence analysis found in the weakened 3' base pair system in comparison to the dG:dC base pair the incorporation of  $m^1\text{U}$  increased by nearly >3-fold (Figure 7C). These results support the weaker protein-DNA interaction and DNA base pair on the 3' side of a site for insertion of U or  $m^1\text{U}$  impact the NTP insertion and elongation yields. The stronger interactions (base pairing and pi stacking) favor canonical NTP insertion,

while the weaker interactions allow non-canonical NTPs that must sample glycosidic bond angles for proper base pairing to compete. Controls were conducted with the dI:dC base pair on the 5' side of the templating dA in the duplex DNA to find no effect on the U/m<sup>1</sup>Ψ insertion and elongation ratio (Figure S8). When the dA:dT is swapped to have the dA in the template DNA strand, the pi stack is stronger and UTP outcompetes m<sup>1</sup>ΨTP as was studied in the doubly-modified sequence contexts (Figure 5D).

As a final study regarding the T7 RNA polymerase competition assay, UTP was mixed in equal ratios with either e<sup>1</sup>ΨTP or p<sup>1</sup>ΨTP during IVT. The RNAs made were then directly sequenced and the percentage of modification insertion was quantified via the base calling analysis using Nanopore-Psu (11). Only singly-modified sequence contexts were analyzed. The analysis found as the alkyl group length increased at N1 of pseudouridine, the incorporation yield ratio for the modified NTP increased (Figure S9). The yields ratios were close to 1:1 for UTP vs. p<sup>1</sup>ΨTP. Prior work with long alkyl groups attached to the uridine nucleoside found them to adopt equal populations of *syn* and *anti* conformations (38). These results are of interest in situations where RNA is to be synthesized by IVT that would like a balanced distribution of U nucleotides with N1-alkylpseudouridine nucleotides.

## Conclusion

Direct RNA sequencing with nanopores can yield characteristic signatures for modifications that have been demonstrated for many modifications such as Ψ (10-13,15). In the present work, RNA containing either U, Ψ, or m<sup>1</sup>Ψ in all adjacent sequence contexts was generated by IVT and then sequenced with the commercial nanopore platform. The base calling data obtained identified unique signatures for the two U modifications that can be used to predict their presence in RNA. Using available computational tools, the occupancy of Ψ or m<sup>1</sup>Ψ can be predicted, as well. The present studies used the tool Nanopore-Psu first developed to predict Ψ occupancy in RNA (11) in the base called data for the analysis of m<sup>1</sup>Ψ. Further, we found the Ψ

and  $m^1\Psi$  relative to U yielded similar changes to the ionic currents when located in the nanopore sensor and slowed the dwell times when the modifications pass through the active site of the helicase motor protein. In the big picture, this analysis demonstrated different RNA modifications on the same nucleotide can yield similar signatures. This illustration should bring caution to RNA researchers using nanopore sequencing to conduct *de novo* analysis for chemical modifications. A site of modification could be found but the identity could remain questionable.

Understanding the nanopore sequencing data for  $\Psi$  or  $m^1\Psi$  enabled a running start T7 RNA polymerase assay to be conducted to compete them ( $\Psi$ TP or  $m^1\Psi$ TP) against UTP during transcription. The nanopore sequencing data for the RNAs found  $\Psi$ TP outcompeted UTP in all singly-modified contexts that we hypothesize results from *syn* and *anti*  $\Psi$  pairing equally well with dA in the template DNA strand. In doubly-modified contexts,  $\Psi$ TP was favored on the 5' side and disfavored on the 3' side that we propose results from the unfavourability of incorporating two *syn* pyrimidines in a growing mRNA chain. In contrast, UTP outcompeted  $m^1\Psi$ TP in all contexts except when a 3' A occurred after the modification site in the RNA, where the yield ratios were the same as the NTP concentration ratio in the solution. A model for the results is proposed based on the prior finding that  $m^1\Psi$  strongly favors the *syn* conformation and cannot pair with dA in the *anti* conformation allowing UTP to outcompete the modification (38). The exception is when there is an A coded for in the RNA sequence 3' to the modification, which based on the solved structures for the polymerase (30,33) provides a more flexible active site to allow  $m^1\Psi$ TP to adopt the *anti* conformation to pair with the templating dA. The present results identify that when mRNAs are synthesized with a mixture of U and  $m^1\Psi$ , their NTP concentrations will not be a good predictor of the extent of modification in the RNA because T7 RNA polymerase will function as a gatekeeper for the insertion yield of the NTPs in different sequence contexts. This information will be needed in future situations where therapeutic mRNAs are produced with sub-stoichiometric levels of  $m^1\Psi$ , which has been successfully conducted with other RNA modifications (28).

## Data Availability

The data are available upon request.

## Supplementary Data

Supplementary data are available at XYZ.

## Acknowledgments

The National Institutes of Health provided financial support for this project (R01 GM093099). Oligonucleotide synthesis was provided by the University of Utah Health Sciences Core facilities that are supported in part by a National Cancer Institute Cancer Center Support grant (P30 CA042014).

## Conflict of Interest

AMF and CJB have a licensed patent for nanopore sequencing to Electronic BioSciences, and AMF is a paid consultant at Electronic BioSciences advising on the chemistry of nucleic acids.

## References

1. Jones, J.D., Monroe, J. and Koutmou, K.S. (2020) A molecular-level perspective on the frequency, distribution, and consequences of messenger RNA modifications. *Wiley Interdiscip. Rev. RNA*, e1586.
2. Roundtree, I.A., Evans, M.E., Pan, T. and He, C. (2017) Dynamic RNA modifications in gene expression regulation. *Cell*, **169**, 1187-1200.
3. Linder, B. and Jaffrey, S.R. (2019) Discovering and mapping the modified nucleotides that comprise the epitranscriptome of mRNA. *Cold Spring Harb Perspect. Biol.*, **11**, a032201.
4. Hu, B., Zhong, L., Weng, Y., Peng, L., Huang, Y., Zhao, Y. and Liang, X.-J. (2020) Therapeutic siRNA: state of the art. *Sig. Transduct. Target Ther.*, **5**, 101.
5. Nance, K.D. and Meier, J.L. (2021) Modifications in an emergency: the role of N1-methylpseudouridine in COVID-19 vaccines. *ACS Cent. Sci.*, 10.1021/acscentsci.1021c00197.
6. Branton, D. and Deamer, D. (2019) *Nanopore Sequencing An Introduction*. World Scientific Publishing Co. Pte. Ltd.

7. Leger, A., Amaral, P.P., Pandolfini, L., Capitanchik, C., Capraro, F., Miano, V., Migliori, V., Toolan-Kerr, P., Sideri, T., Enright, A.J. *et al.* (2021) RNA modifications detection by comparative Nanopore direct RNA sequencing. *Nat. Commun.*, **12**, 7198.
8. Furlan, M., Delgado-Tejedor, A., Mulroney, L., Pelizzola, M., Novoa, E.M. and Leonardi, T. (2021) Computational methods for RNA modification detection from nanopore direct RNA sequencing data. *RNA Biol.*, **18**, 31-40.
9. Thomas, N.K., Poodari, V.C., Jain, M., Olsen, H.E., Akeson, M. and Abu-Shumays, R.L. (2021) Direct nanopore sequencing of individual full length tRNA strands. *ACS Nano*, **15**, 16642-16653.
10. Begik, O., Lucas, M.C., Pryszzc, L.P., Ramirez, J.M., Medina, R., Milenovic, I., Cruciani, S., Liu, H., Vieira, H.G.S., Sas-Chen, A. *et al.* (2021) Quantitative profiling of pseudouridylation dynamics in native RNAs with nanopore sequencing. *Nat. Biotechnol.*, 10.1038/s41587-41021-00915-41586.
11. Huang, S., Zhang, W., Katanski, C.D., Dersh, D., Dai, Q., Lolans, K., Yewdell, J., Eren, A.M. and Pan, T. (2021) Interferon inducible pseudouridine modification in human mRNA by quantitative nanopore profiling. *Genome Biol.*, **22**, 330.
12. Fleming, A.M., Mathewson, N.J., Howpay Manage, S.A. and Burrows, C.J. (2021) Nanopore dwell time analysis permits sequencing and conformational assignment of pseudouridine in SARS-CoV-2. *ACS Cent. Sci.*, **7**, 1707-1717.
13. Jenjaroenpun, P., Wongsurawat, T., Wadley, T.D., Wassenaar, Trudy M., Liu, J., Dai, Q., Wanchai, V., Akel, N.S., Jamshidi-Parsian, A., Franco, A.T. *et al.* (2020) Decoding the epitranscriptional landscape from native RNA sequences. *Nucleic Acids Res.*, **49**, e7.
14. Smith, A.M., Jain, M., Mulroney, L., Garalde, D.R. and Akeson, M. (2019) Reading canonical and modified nucleobases in 16S ribosomal RNA using nanopore native RNA sequencing. *PLoS One*, **14**, e0216709.
15. Makhamreh, A., Tavakoli, S., Gamper, H., Nabizadehmashhadtoroghi, M., Fallahi, A., Hou, Y.-M., Rouhanifard, S.H. and Wanunu, M. (2022) Messenger-RNA modification standards and machine learning models facilitate absolute site-specific pseudouridine quantification. *bioRxiv*, 2022.2005.2006.490948.
16. Pang, H., Ihara, M., Kuchino, Y., Nishimura, S., Gupta, R., Woese, C.R. and McCloskey, J.A. (1982) Structure of a modified nucleoside in archaeabacterial tRNA which replaces ribosylthymine. 1-Methylpseudouridine. *J. Biol. Chem.*, **257**, 3589-3592.
17. Li, H. (2018) Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, **34**, 3094-3100.
18. Danecek, P., Bonfield, J.K., Liddle, J., Marshall, J., Ohan, V., Pollard, M.O., Whitwham, A., Keane, T., McCarthy, S.A., Davies, R.M. *et al.* (2021) Twelve years of SAMtools and BCFtools. *GigaScience*, **10**, giab008.
19. Robinson, J.T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E.S., Getz, G. and Mesirov, J.P. (2011) Integrative genomics viewer. *Nat. Biotechnol.*, **29**, 24-26.

20. Stoiber, M., Quick, J., Egan, R., Eun Lee, J., Celtniker, S., Neely, R.K., Loman, N., Pennacchio, L.A. and Brown, J. (2017) De novo identification of DNA modifications enabled by genome-guided nanopore signal processing. *bioRxiv*, doi:10.1101/094672.
21. Loman, N.J., Quick, J. and Simpson, J.T. (2015) A complete bacterial genome assembled de novo using only nanopore sequencing data. *Nat. Meth.*, **12**, 733-735.
22. Van der Verren, S.E., Van Gerven, N., Jonckheere, W., Hambley, R., Singh, P., Kilgour, J., Jordan, M., Wallace, E.J., Jayasinghe, L. and Remaut, H. (2020) A dual-constriction biological nanopore resolves homonucleotide sequences with high fidelity. *Nat. Biotechnol.*, **38**, 1415-1420.
23. Liu, N., Parisien, M., Dai, Q., Zheng, G., He, C. and Pan, T. (2013) Probing N6-methyladenosine RNA modification status at single nucleotide resolution in mRNA and long noncoding RNA. *RNA*, **19**, 1848-1856.
24. Su, D., Chan, C.T., Gu, C., Lim, K.S., Chionh, Y.H., McBee, M.E., Russell, B.S., Babu, I.R., Begley, T.J. and Dedon, P.C. (2014) Quantitative analysis of ribonucleoside modifications in tRNA by HPLC-coupled mass spectrometry. *Nat. Protoc.*, **9**, 828-841.
25. Stephenson, W., Razaghi, R., Busan, S., Weeks, K.M., Timp, W. and Smibert, P. (2022) Direct detection of RNA modifications and structure using single-molecule nanopore sequencing. *Cell Genom.*, **2**, 100097.
26. Schrum, J.P., Siddiqui, S. and Ejube, K. (2010) USA.
27. Schibel, A.E., An, N., Jin, Q., Fleming, A.M., Burrows, C.J. and White, H.S. (2010) Nanopore detection of 8-oxo-7,8-dihydro-2'-deoxyguanosine in immobilized single-stranded DNA via adduct formation to the DNA damage site. *J. Am. Chem. Soc.*, **132**, 17992-17995.
28. Kormann, M.S., Hasenpusch, G., Aneja, M.K., Nica, G., Flemmer, A.W., Herber-Jonat, S., Huppmann, M., Mays, L.E., Illenyi, M., Schams, A. *et al.* (2011) Expression of therapeutic proteins after delivery of chemically modified mRNA in mice. *Nat. Biotechnol.*, **29**, 154-157.
29. Huang, J., Brieba, L.G. and Sousa, R. (2000) Misincorporation by wild-type and mutant T7 RNA polymerases: identification of interactions that reduce misincorporation rates by stabilizing the catalytically incompetent open conformation. *Biochemistry*, **39**, 11571-11580.
30. Temiakov, D., Patlan, V., Anikin, M., McAllister, W.T., Yokoyama, S. and Vassylyev, D.G. (2004) Structural basis for substrate selection by t7 RNA polymerase. *Cell*, **116**, 381-391.
31. Wang, B., Predeus, A.V., Burton, Z.F. and Feig, M. (2013) Energetic and structural details of the trigger-loop closing transition in RNA polymerase II. *Biophys. J.*, **105**, 767-775.
32. Wu, S., Li, L. and Li, Q. (2017) Mechanism of NTP binding to the active site of T7 RNA polymerase revealed by free-energy simulation. *Biophys. J.*, **112**, 2253-2260.

33. Yin, Y.W. and Steitz, T.A. (2004) The structural mechanism of translocation and helicase activity in T7 RNA polymerase. *Cell*, **116**, 393-404.
34. Wu, S., Wang, J., Pu, X., Li, L. and Li, Q. (2018) T7 RNA polymerase discriminates correct and incorrect nucleoside triphosphates by free energy. *Biophys. J.*, **114**, 1755-1761.
35. Milisavljević, N., Perlíková, P., Pohl, R. and Hocek, M. (2018) Enzymatic synthesis of base-modified RNA by T7 RNA polymerase. A systematic study and comparison of 5-substituted pyrimidine and 7-substituted 7-deazapurine nucleoside triphosphates as substrates. *Org. Biomol. Chem.*, **16**, 5800-5807.
36. Ulrich, S. and Kool, E.T. (2011) Importance of steric effects on the efficiency and fidelity of transcription by T7 RNA polymerase. *Biochemistry*, **50**, 10343-10349.
37. Takahashi, S., Matsumoto, S., Chilka, P., Ghosh, S., Okura, H. and Sugimoto, N. (2022) Dielectricity of a molecularly crowded solution accelerates NTP misincorporation during RNA-dependent RNA polymerization by T7 RNA polymerase. *Sci. Rep.*, **12**, 1149.
38. Chang, Y.C., Herath, J., Wang, T.H. and Chow, C.S. (2008) Synthesis and solution conformation studies of 3-substituted uridine and pseudouridine derivatives. *Bioorg. Med. Chem.*, **16**, 2676-2686.
39. Morales, J.C. and Kool, E.T. (1999) Minor groove interactions between polymerase and DNA: More essential to replication than Watson-Crick hydrogen bonds? *J. Am. Chem. Soc.*, **121**, 2323-2324.