Secondary Structure Detection Through Direct Nanopore RNA Sequencing

Alan Shaw[1], Jonathan M. Craig[2], Hossein Amiri[1,4], Jeonghoon Kim[1,3,5], Heather E. Upton[4,6], Sydney C. Pimentel[4,7], Jesse R. Huang[2], Susan Marqusee[1,3,4,8], Jens H. Gundlach[2], Kathleen Collins[1,4,6], and Carlos J. Bustamante[1,3,4,5,8,10,11,12]

1. Institute for Quantitative Biosciences, University of California, Berkeley, CA, 94720
2. Department of Physics, University of Washington, Seattle, WA, 98195
3. Biophysics Graduate Group, University of California, Berkeley, CA, 94720
4. Department of Molecular and Cell Biology, University of California, Berkeley, CA, 94720
5. Jason L. Choy Laboratory of Single-Molecule Biophysics, University of California, Berkeley, CA 94720
6. Bakar Fellows Program, University of California, Berkeley, CA, 94720
7. Present address: NYU Grossman School of Medicine 550 First Avenue New York, NY 10016
8. Department of Chemistry, University of California, Berkeley, CA 94720
9. Chan Zuckerberg Biohub, San Francisco, CA, USA
10. Department of Physics, University of California, Berkeley, CA 94720
11. Kavli Energy Nanoscience Institute, University of California, Berkeley, CA 94720
12. Howard Hughes Medical Institute, University of California, Berkeley, CA 94720

*Correspondence to: Carlos J. Bustamante, carlosb@berkeley.edu; Susan Marqusee, marqusee@berkeley.edu; Kathleen Collins, kcollins@berkeley.edu; Jens H. Gundlach, gundlach@uw.edu.

Abstract

Techniques that can directly sequence RNAs and determine secondary structures are essential to establish how an RNA molecule folds and how folding affects its function. We describe the development of a direct RNA nanopore sequencing technique using an engineered *Bombyx mori* retroelement reverse transcriptase (RT) to thread RNA through the *Mycobacterium smegmatis* porin A (MspA) nanopore in single-nucleotide steps. First, we establish the map that correlates RNA sequence and MspA ion current. We find that during sequencing the RT can sense the strength of RNA secondary structures 11-12 nt downstream of its front boundary, modifying its stepping dynamics accordingly. Using this feature, we achieve simultaneous RNA nanopore sequencing and structure detection, without the need of prior conversion to complementary DNA (cDNA) or RNA modifications. The sequence-dependent ion currents open the way to utilize the MspA nanopore to investigate the single-molecule activity of other processive RNA translocases such as the ribosome.

Introduction

RNA plays critical roles as a messenger for protein synthesis as well as in a wide diversity of biological processes including, among others, regulation of transcription and translation, protein synthesis, and nuclear organization[1,2]. The structure of RNA is vital to its biological function[3], and therefore high-throughput methods to assign RNA structures have been developed, which are largely based on Illumina sequencing technology[4,5]. However, indirect sequencing of RNA through its conversion to cDNA via reverse transcription can only infer the existence of RNA structures from altered sequencing fidelity or termination events of DNA polymerization. Another approach, the Oxford Nanopore (ONT) platform, typically performs RNA sequencing by sequencing cDNA. Advances in the ONT platform have enabled RNA strand sequencing in combination with SHAPE-seq reagents, and the presence of RNA structures was revealed through the detection of modified bases and altered helicase translocation kinetics[6,7].

However, no technique to-date is capable of directly sequencing RNA and detecting RNA secondary structures simultaneously without the use of chemical modifications or digestion.

In this study, we present a nanopore-based technique that directly sequences RNA *and* detects RNA structure from the sequencing translocation kinetics, without the need of prior cDNA conversion or RNA modifications. This advancement was achieved using a processive eukaryotic cellular RT to thread RNA into a MspA nanopore one base at a time, as it synthesizes cDNA (**Figure 1A**). The MspA nanopore instrument is based on the setup developed previously[8]. This setup consists of two wells filled with electrolyte solutions that are separated by an insulating lipid bilayer in which a single MspA nanopore has been inserted (**Supplementary Figure 1**). When voltage is applied across the membrane, ion current flows through the pore. When biological polymers such as DNA, RNA, or peptides enter the pore, the measured current drops to an extent determined by the sequence of the biopolymer spanning the constriction of the MspA nanopore[8–10].

In the first application strategy, target RNAs are hybridized to a short DNA primer that is tagged with cholesterol to direct the RNA/DNA complex to the nanopore membrane (**Figure 1A**). Addition of the RT results in binding of the enzyme to the RNA/DNA junction and consequent initiation of cDNA synthesis. Upon application of a voltage bias across the membrane, the RNA 3' end of an elongation complex is drawn through a backwards inserted MspA nanopore (**Figure 1A , Supplemental Figure 1**) and the RT comes to rest on top of the pore, preventing additional RNA transit through it. The rate of cDNA synthesis by RT dictates the 3'-to-5' passage of the RNA template chain through the pore in discrete steps until synthesis is complete. Using this setup, we established the first MspA nanopore *RNA quadromer map* that connects each detected ion current with a unique 4-nucleotide sequence spanning the pore, and enables RNA sequencing with the MspA nanopore.

The MspA nanopore technique is also a powerful tool to dissect the biophysical function of molecular motors[8] as the exact position of the motor protein on its template can be determined, and its single-molecule biophysical parameters (such as dwell time, pauses, backtracking activity) at every single step on the template can be detected and analyzed. In this study, we challenged the RT with different RNA structural barriers and quantified its translocation kinetics, which revealed RT's ability to sense secondary structures ahead of its front boundary. We discuss possible mechanisms for this downstream sensing. Combining RNA sequencing and monitoring of the single molecule kinetics of the RT, we have demonstrated simultaneous RNA sequencing and structure detection without the need of prior conversion to cDNA or chemical modifications of RNA.

Results and Discussion

Establishment of MspA nanopore sequencing of RNA. Nanopore sequencing requires two main components: 1) a processive motor to either pull or feed single-stranded (ss) DNA or RNA through the nanopore in discrete steps, and 2) an *a priori* knowledge of the ion currents corresponding to all the possible sequences that partially block the nanopore. In this study, we used the MspA nanopore, which has been previously exploited for DNA[9] and peptide[10] sequencing. A major challenge to our goal was to identify a motor protein that could translocate the RNA template through the MspA nanopore in a processive and controlled manner. We initially tested two classes of enzymes: RNA helicase and RT. No readily available enzymes

tested in either category, including the NS3 helicase from Hepatitis C virus (HCV)[11] and retroviral RT[12], could retain and processively thread ssRNA through the nanopore under the necessary cross-membrane voltage bias. With the DNA primer and ssRNA capture strategy of our experimental design, we determined that a bacterial self-splicing intron RT from *Eubacterium rectale*[13] was able to thread ssRNA through the nanopore, but the processivity of the enzyme under nanopore translocation conditions was not optimal and resulted in short RNA translocation events (**Supplementary Figure 2A**, left panel). In comparison, we found that a truncated and modified form of a retroelement RT from *Bombyx mori*[14] had the necessary processivity under nanopore experimental conditions and consistently generated long enough RNA translocation events (**Supplementary Figure 2B**, left panel). Upon further analysis, as described in the following section, we determined that the *B. mori* RT was able to generate read lengths up to 400 nt on a 550 nt RNA template (**Supplementary Figure 2B**, right panel) while the RT from *E. rectale* was only able to generate reads up to 60 nt on the same RNA template (**Supplementary Figure 2A**, right panel). Therefore, all further experiments were conducted with the RT from *B. mori*, referred to as bmRT below. RNA sequences used in this study are summarized in **Supplementary Table 1**.

Construction and Validation of the MspA RNA Quadromer Map. A nanopore sequencing approach requires a library of currents corresponding to all the possible sequences that can be found inside the nanopore. For the MspA nanopore, this correlation is referred to as the "quadromer map"[8] since the ion current is determined by the 4 nt that span the constriction site of the pore. Because RNA sequencing with the MspA nanopore had not yet been achieved, we first proceeded to obtain the RNA quadromer map for the MspA nanopore. Comparing RNA translocation events we collected with the first RNA template (RNA1) (**Supplementary Table 1**), we noticed that RNA sequencing traces consistently ended with a very similar signature followed by RNA signal stillness in the pore (**Supplementary Figure 3**), which coincides with bulk biochemistry observations that bmRT does not readily dissociate from its RNA template upon completion of cDNA synthesis[14]. Therefore, we could assume that the ion current signals obtained close to the end of a translocation event originate from sequences close to the 5' end of the RNA. Based on this observation, we performed bmRT-directed nanopore translocation of RNAs with and without two eight-nucleotide polyadenosine (polyA) tract insertions that are about 80 nt apart near RNA1's 5' end (RNA1 and RNA1_PolyA, **Figure 1B**, **Supplementary Table 1**). Because polyA generates a signature of high ion current[9], we could roughly assign ion currents to the 80 nt RNA region flanked by the two polyA regions (**Figure 1B**, top and middle panels), and after removing instrument noise or erratic enzyme behavior, we constructed the corresponding sequence of consensus nanopore ion currents. **Figure 1B**, bottom panel, shows the consensus ion currents of part of this region. Next, having the precedent that for DNA sequencing using the MspA nanopore the sequence "TT" often correlates with a local ion current minimum[9], we were able to match the consensus ion currents to the known sequence of RNA1 with single nucleotide accuracy (**Figure 1B** bottom panel, **Supplementary Note 1**). This analysis allowed us to generate the RNA quadromer map (**Supplementary Table 2**), whose information content is comparable to the published DNA quadromer map data[8] (**Supplementary Note 1**). Comparison between ion currents predicted by the resulting RNA quadromer map with existing DNA quadromer maps (**Supplementary Figure 4A**), revealed significant differences between the two, highlighting the importance of newly derived RNA quadromer map for reliable

RNA sequencing. A representative segment of consensus ion current related to sequence is shown in **Figure 1B** (bottom panel). To further verify the quadromer map, we used it to predict the ion current pattern for a different RNA sequence (RNA2, **Supplementary Table 1**). The predicted ion currents matched well with the experimentally determined ones (**Supplementary Figure 5**). The quality of the match is similar to that of previously reported MspA nanopore sequencing of ssDNA[9] (**Supplementary Note 1**), demonstrating the technique's suitability for RNA sequencing. Importantly, the observation that each nucleotide in the RNA template can be assigned to a single step in the consensus ion current series confirms that the bmRT takes single-nucleotide steps on its RNA template and sequentially releases a single nt of RNA at a time to enter the nanopore.

We note that the nanopore reports on the RNA sequence partially blocking the current through the constriction site of the pore. In order to relate the dwell times of the RT with the presence of RNA structures in front of the enzyme (next sections), we need to know the exact location of the enzyme on the RNA template when a particular sequence is in the pore. In other words, we need to establish the offset between the constriction site of the pore and the catalytic site of bmRT. To this end, we exploited a particular feature of this enzyme when it reaches the 5' end of RNA template: it extends its cDNA product via non-templated addition generating up to five nt of 3' overhang[14,15]. Based on the range of positions at which the enzyme stops threading RNA into the nanopore, we estimated that the distance between the enzyme's catalytic site and the constriction site of the nanopore is 17 nt (**Supplementary Figure 6**). This offset allowed us to define the position of the bmRT catalytic site in nanopore sequencing ion-current traces.

Detection of RNA structure via nanopore sequencing kinetics. Stable RNA secondary structures have been shown to affect the kinetic rates of molecular motors such as the ribosome[16], RNA helicases[17], and retroviral RTs[18,19]. We aimed to challenge the bmRT with RNA structures during nanopore sequencing and characterize the changes in kinetic behavior of bmRT as it encounters these barriers. To extract kinetic information corresponding to the RNA sequence, we determined the average dwell time before each RT step on the RNA template. This procedure involved pooling data obtained for that step from multiple sequencing traces of the same sequence and fitting them to a single exponential function (**Figure 2A**). As shown in **Figure 2A**, the dwell time distribution of bmRT can be described by a single exponential function, which suggests that bmRT has a single dominant rate limiting step between each translocation steps.

To examine the effect of RNA secondary structures on bmRT dwell times, we designed an RNA template that contains two repeats of the same sequence (RNA3 in **Supplementary Table 1**) in which one repeat is partially base paired to form the stem of a stable RNA hairpin and the other is not (**Figure 2B** top panel). The RT kinetic profiles derived from the average dwell times of the individual steps of the enzyme obtained for the repeats in the presence and absence of the hairpin structural barrier were compared (**Figure 2B** bottom panel). This analysis revealed a major pause when the catalytic site of the enzyme is 2 nt away from the start of the RNA hairpin. This pausing indicates that the hairpin duplex represents a barrier that slows bmRT translocation along the RNA template, and makes it possible to use this enzyme to simultaneously detect RNA structures concomitant with sequencing. Interestingly, the two kinetic profiles were indistinguishable in the remainder of the hairpin sequence, suggesting that the invasion by the enzyme of the hairpin is sufficient to greatly destabilize it.

As a second test, we hybridized an RNA oligonucleotide to a region of the same RNA3 template sequence to create a double-stranded (ds) barrier for the enzyme (**Figure 2C** top panel). RT kinetic profiles obtained in the presence and absence of the hybridized oligonucleotide showed pauses in translocation at distinct positions within the dsRNA region (**Figure 2C** bottom panel, and **Supplementary Figure 7**). As in the case of the hairpin, dwell times with and without the dsRNA barrier remained similar for most translocation steps. To rule out the possibility that direct contact with the 5' phosphate of the RNA may have caused the pause right before the terminal RNA hairpin, we designed a pair of RNA oligonucleotides that hybridize to RNA3, one with a 5' ssRNA polyA overhang and another without it. We found that the kinetic profile of bmRT was similar for both oligos (**Supplementary Figure 8**), suggesting that the pause we observed before the terminal RNA hairpin (**Figure 2B**) is most likely due to the presence of a stable RNA secondary structure and not the presence of a 5' phosphate.

Modeling of the impact of RNA structures on bmRT translocation kinetics. Surprisingly, bmRT dwell times at most positions do not appear to be changed by the presence of the secondary structures in the RNA either in the form of a hairpin or in the form of a duplex. Rather, the enzyme seems to pause at certain particular positions in this region and be unaffected in the regions of secondary structure that surround them. This behavior suggests that bmRT functions as an active helicase capable of destabilizing RNA structures[20]. To explain why the enzyme slows down at certain specific locations within the secondary structures, we constructed an active helicase model to quantitatively describe the kinetic profile of bmRT as a function of barrier stability.

As a general description, the translocation cycle of bmRT consists of a residence phase during which events such as dNTP binding and catalysis occur, followed by a stepping phase in which the motor attempts to move along its track. The overall observed dwell time at each position would equal $k_{resid}^{-1} + k_{step}^{-1}$, where k_{resid} and k_{step} are the rates of completing a residence and the rate of stepping of the enzyme, respectively. As we show in **Supplementary Note 2**, the observed dwell times in the presence of barriers can be well explained if the stepping rate depends not only on the base pairing stability of the nucleotide that is stepped over (at the helicase site of the enzyme), but also on the stability of several downstream nucleotides:

$$k_{step} = P_u P_{mu} k_{ss} \qquad (1)$$

where P_u is the probability that the stepped-over nucleotide is in its unpaired state, P_{mu} is the probability that the following downstream segment of length m is in its unpaired state, and k_{ss} is the stepping rate over single-stranded RNA (in the absence of barrier). P_u is a function of the Gibbs free energy difference between the unpaired and paired states of the nucleotide:

$$P_u = (1 + \exp(-\beta\,(\Delta G_{bp} - \Delta G_d)))^{-1} \qquad (2)$$

where β is $(k_B T)^{-1}$, ΔG_{bp} is the free energy of base pairing for the nucleotide, and ΔG_d is the destabilization energy due to the helicase. A large negative value of ΔG_d would represent a more "active" helicase[20]. Similarly,

$$P_{mu} = (1 + \exp(-\beta\,(\textstyle\sum_{i=1}^{m} \Delta G_{bp,i} - m\Delta G_d)))^{-1} \qquad (3)$$

where m is the length of the downstream segment following the stepped-over nucleotide, $\sum_{i=1}^{m} \Delta G_{bp,i}$ is the total base-pairing free energy of the downstream segment, and ΔG_d is the same as above (per nucleotide).

Knowing the sequence of the dsRNA barrier, the value of ΔG_{bp} at each nucleotide position can be estimated precisely using the nearest neighbor rules[22] (as the difference in the

ΔG of the barrier before and after opening of the given nucleotide). Additionally, k_{resid} can be determined from the observed translocation rates in the absence of the barrier. This leaves k_{ss}, ΔG_d, and m as the only free parameters in the model. After fitting these parameters, dwell times predicted using Eqs. 1 to 3 are in excellent agreement with the kinetic profiles obtained in the presence of different barriers, with the major and minor points of slowdown properly reproduced (**Figure 3A** and **Supplementary Figure 9D**). We fit the model independently to five data sets but the parameters converged to similar values in all cases: k_{ss} ~200 s^{-1}, ΔG_d ~–2.6 kcal/mol, m ~10-11 (**Supplementary Figure 9D**). Furthermore, the best fit is obtained if the bmRT catalytic site nucleotide (–1 position) and the next nucleotide (–2) are assumed to be always unpaired, indicating that the helicase site of bmRT is at position –3 (**Figure 3B**). Indeed, structure prediction based on homology modeling of bmRT suggests that position –2 cannot accommodate dsRNA[15], in agreement with this assumption.

Our model suggests that bmRT interacts with 11-12 nt of the downstream template (including the stepped-over nucleotide itself, **Figure 3B**). Although no direct structural evidence currently exists for this interaction, the formation of a stable complex between bmRT and its RNA template prior to target-priming and cDNA synthesis in the cell[14,21] suggests the existence of an extensive bmRT-RNA binding interface. Significantly, previous single molecule optical tweezers studies on the kinetics of two other RNA motors, the NS3 helicase from HCV[17], and the RT from the murine leukemia virus[19] have revealed that these motors can also sense and slow down in response to RNA secondary structures 6 to 8 nt downstream of the enzymes, suggesting that downstream sensing of structured regions in RNA is not uncommon in RNA helicases. A downstream sensing range of at least 3 nt was similarly inferred from the helicase kinetics and structural analysis of the bacterial ribosome[16,22,23]. Downstream sensing could be mediated by direct interaction of the helicase with the RNA to destabilize its folding [17,19,23], or by a mechanism in which the *kinetic* stability of the junction arises from a long-range allosteric coupling through the double helix[24].

Using the parameters obtained from bmRT kinetics, we deduced the motor's characteristic curve for overall translocation speed as a function of average ΔG_{bp} (red curve in **Figure 3C**). The sigmoidal shape of the curve becomes sharper for larger values of m, i.e. for motors displaying longer ranges of downstream RNA structure sensing (**Figure 3C**). Due to this shape, translocation is barely affected by isolated base pairs, and slows down significantly only if the average stability of the entire downstream segment exceeds ΔG_d in magnitude (–2.6 kcal/mol per nucleotide in the case of bmRT).

Since RNA is known to spontaneously form secondary structures of short polymer lengths[25], we also quantified the average dwell time at each nt on the RNA3 template alone (without the hybridizing RNA oligonucleotide) and analyzed the correlation between dwell times and the presence of dsRNA predicted by mfold [26] (**Supplementary Note 3**). As expected, longer dwell times are only observed in front of downstream RNA regions that have high probability of being double-stranded, most of which have high GC content (**Supplementary Figure 10**). Indeed, by incorporating the predicted base pairing probabilities into our active helicase model, we can qualitatively reproduce the observed pattern of dwell times with an overall correlation coefficient of ~0.6 (**Supplementary Figure 11**).

Detection of RNA aptamer-ligand complex formation. Finally, we explored the possibility to utilize our experimental assay to detect the binding of an RNA aptamer to its ligand

that stabilizes its tertiary structure. We designed an RNA template that contains a single Broccoli RNA aptamer[27] at its 5' end (RNA3_Broccoli in **Supplementary Table 1**). This aptamer has two G-quadruplexes (GQ) and can bind the fluorescent ligand BI, which stabilizes the folding of the aptamer RNA[27,28]. The Broccoli GQs are preceded by a short RNA duplex (**Figure 4A**) which was shown to be important in folding of the aptamer based on sequence truncation experiments[27].

We first showed that the aptamer binds to BI under our experimental conditions (**Supplementary Figure 12**). Using our assay, we then compared the single-molecule kinetic profiles of bmRT on Broccoli RNA with and without the presence of BI (**Figure 4B**). Results indicate that binding to BI and stabilization of the Broccoli RNA structure led to a significant pause of the bmRT when the helicase site of the enzyme is still 1 nt away from the start of the Broccoli RNA duplex. Based on our active helicase model, at this position the downstream sensing range (**Figure 3B**) covers the Broccoli duplex and its continuous stack of nucleotides up to the first GQ, and a slowdown is indeed expected (**Supplementary Figure 9,** middle panel). Broccoli mutation experiments show that replacement of either of the Gs in the GQs for another base results in a significant loss of BI fluorescence, and that GQ formation is critical to the formation of stable Broccoli RNA structure[28]. This observation, in combination with our single-molecule kinetics data, indicate that both the short RNA duplex that precedes the GQ and the GQ in BI-bound Broccoli RNA represent stable barriers that slow down bmRT, and that the nanopore-based RNA sequencing approach described here can be used to identify stabilized secondary structures such as those of ligand-bound RNA aptamers.

Conclusion

Characterization of RNA sequence and structure is critical to understanding the complex roles that RNA molecules play in normal physiology and diseases. In this study, we present for the first time an RNA nanopore sequencing method that provides both sequence and structure information simultaneously. To this end, we established an assay to follow the translocation of RNA through the MspA nanopore with single nucleotide resolution, gated by an engineered eukaryotic RT, and we generated the RNA quadromer map that allows us to reliably assign ion current signals to RNA sequences. We also showed that kinetics of individual bmRT enzymes can reveal, with single-nucleotide resolution, how translocation rate varies with the sequence-dependent stability of encountered structural barriers. We have found that the pausing of the bmRT indicates when a particularly extensive, stable secondary structure is encountered by the enzyme. Our results suggest that bmRT functions as a processive helicase that actively unfolds incoming dsRNA and senses secondary structures 11-12 bp downstream of its front boundary. In addition, we showed that slow-down in kinetic rates can be used to detect the presence of RNA aptamer-ligand complexes.

Our method has high promise for interrogating RNA sequence and structure directly without the need of prior RNA modifications. Furthermore, our technique directly provides biophysical information on how RNA structure barriers impact the biophysical behavior of a eukaryotic RNA molecular motor protein. Finally, the quadromer map for RNA in the MspA nanopore presented here opens up possibilities to utilize the MspA nanopore tweezers to investigate the activity and dynamics of other processive RNA translocases such as the

ribosome or synthetases such as RNA polymerases, and do so with single nucleotide resolution and in a sequence dependent manner.

Materials and Methods

RNA template preparation

RNA template sequences were ordered as dsDNA gBlocks that contain the T7 promoter from Integrated DNA Technologies (IDT) and inserted into a linearized pRZ plasmid using the infusion cloning kit (Thermo Fisher) and transformed into Sure2 cells (Agilent) following manufacturer's instructions. Positive colonies were screened with Sanger sequencing. PCR was used to amplify templates for *in vitro* transcription using the MEGAscript kit following manufacturer's instructions (Thermo Fisher). The RNA templates were purified with the MEGAclear kit (Thermo Fisher), and concentration determined with a Nanodrop spectrophotometer (Thermo Fisher). RNA oligonucleotide and DNA primer with 5' cholesterol modification were ordered from IDT. The RNA template was mixed with DNA primer (and when relevant a 10-fold excess of RNA oligonucleotide for dsRNA barrier experiments) to a final concentration of 0.8 µM and 2 µM respectively, in buffer containing 20 mM Tris pH 8.0 and 20 mM NaCl and heated to 75°C for 90 seconds and immediately placed on ice until further use.

Preparation of RNA motor enzymes

E. rectale RT and N-terminally truncated *B. mori* RT were expressed and purified as described previously[16]. In short: The open reading frame of the enzymes was codon optimized and ordered from GenScript, and inserted with an N-terminal maltose binding protein tag into the MacroLab vector 2bct that contains a C-terminal 6xHis tag (https://qb3.berkeley.edu/facility/qb3-macrolab/). The enzymes were expressed in Rosetta2(DE3)pLysS cells in 2xYT medium and induced with isopropylthio-β-galactoside. Cells were lysed by sonication on ice and a three-step purification process (nickel-agarose column, heparin-Sepharose column, HiPrep 16/60 Sephacryl S-200HR size exclusion column) was used to purify the enzymes. The purified enzymes were stored in 25 mM HEPES pH 7.4, 800 mM KCl, 10% glycerol, and 1 mM DTT and stored at -80°C. Working stocks were stored at -20°C after RT dilution to a final concentration of 20 µM in 25 mM HEPES pH 7.5, 800 mM KCl, and 50% glycerol.

MspA nanopore instrumentation

The MspA nanopore instrument is a custom-built instrument based on the design from the Gundlach lab[8]. In more detail, 2 wells of about 120 µl in volume were drilled into a Teflon block and the two wells were connected with Teflon tubing. One end of the tube was heat-shrunk and a small hole (about 20 um in diameter) was created using a fine surgical needle. Electrodes were prepared by inserting an Ag/AgCl pellet in heat shrink tubing. The Teflon block was mounted onto a custom-made aluminum block. Under the aluminum block is a Peltier that is connected to a temperature control unit (TED200C, Thorlabs). An Axopatch 200b (Molecular Devices) was connected to the electrodes and used to apply voltage and measure ion current. The Axopatch 200b is connected to a PC using National Instrument's data acquisition card (DAQ) and controlled with a custom LabVIEW code. The well that contains the 20 um hole is

referred to as the *cis* well, and is where all the biochemical components are introduced during sequencing data acquisition. The other well is referred to at the *trans* well.

Nanopore Experiments

The two wells and tubing were first filled with standard experiment buffer (40 mM HEPES pH 7.5, 400 mM KCl). 180 mV was applied to the system. Dry Lipid (4ME 16:0 DIETHER PC 10MG, Avanti polar lipids) was mixed with hexadecane (Sigma-Aldrich) until the consistency resembled that glue, followed by application of the lipid-hexadecane mixture to the tip of the Teflon tubing in the *cis* well. Lipid bilayer was generated by introducing an air bubble via a pipette to the surface of the tubing. Afterwards, MspA protein (the M2-NNN MspA mutant[8]) was added to the well to a final concentration of about 0.02 µg/ml. After successful insertion of a single backwards pore, we reduced the system's voltage to 140 mV and buffer-exchanged the *cis* well to RT experiment buffer (40 mM HEPES pH 7.5, 320 mM KCl, 3 mM $MgCl_2$, 5 mM DTT, 24 µM dNTP), heated up the system to 36°C, and added the RNA/DNA primer complex to the well to a final concentration of about 15 nM RNA. Afterwards, we added the RT to a final concentration of about 150 nM and started data acquisition.

Nanopore Broccoli ligand binding experiment

Broccoli RNA template sequences were ordered as dsDNA gBlock as above and inserted into a linearized pRZ plasmid using infusion cloning kit (Takara Bio) and transformed into Stellar cells following manufacturer's instructions. Positive colonies were screened with Sanger sequencing. PCR was used to amplify templates for in vitro transcription with T7 RNA polymerase (NEB). The RNA product was extracted with phenol and concentration was measured by Nanodrop spectrophotometer (Thermo Fisher). Ligand for Broccoli RNA aptamer BI (LuceRNA) was prepared in 50mM DMSO and further diluted in water. Binding of the ligand to the RNA template was tested by varying the ratio of ligand to RNA in buffer containing 20 mM Tris pH 8.0 and 20 mM NaCl and heated to 75°C for 90 seconds and immediately placed on ice. The fluorescence intensity was quantified using ImageJ. In the nanopore experiment using BI ligand, 1:15 ratio of RNA to ligand was used.

Data Processing

The data processing pipeline is based on methods described previously[8]. In short: raw data (collected at 50 kHz) was down sampled to 2 kHz, and RNA translocation events were identified by using a custom GUI written in MATLAB. A point of change algorithm[8] was used to identify steps within a continuous series of RNA translocation events. The steps identified and their corresponding dwell times were then used for additional data processing as described in the main article.

Acknowledgements

AS was supported by the k99/r00 award from the National Human Genome Research Institute grant number 5K99HG011492. HEU, SCP and KC were supported by funding from the University of California, Berkeley Bakar Fellows Program and N.I.H. DP1HL156819 (KC). JMC, JRH, and JHG were supported by National Human Genome Research Institute grant R01HG005115. AS, JK were supported by N.I.H R01-GM0325543 (CJB) and NSF

MCB1616591 (SM). CJB was supported by Director, Office of Science, Office of Basic Energy Sciences of the U.S. Department of Energy (DOE), contract no. DE-AC02-05CH11231 Nanomachine program and Molecular Foundry. SM is a Chan Zuckerberg Biohub Investigator. The authors thank all members of the Marqusee, Gundlach, Collins, and Bustamante lab for helpful discussions and support.

Conflict of Interest

Engineered *B. mori* RT and sequence variants with improved properties are included in patent applications filed by University of California, Berkeley with HEU, SCP, and KC as named inventors. HEU and KC are founders of Karnateq Inc., which licensed the RT technology. An additional patent application describing nanopore sequencing applications was filed by University of California, Berkeley with AS, HEU, SCP, JMC, JHG, SM, CJB, and KC as named inventors.

Author Contributions

AS, JHG, and CJB conceived the study. AS, SM, KC, JMC, and CJB designed the experiments. SM, JHG, KC, and CJB supervised the work. AS, JK performed the nanopore experiments and collected data. AS, HA, and JMC designed, cloned, and prepped the RNA sequences. AS, HA, and JMC analyzed the data and prepared figures. HEU, SCP, and KC conceived and engineered the bmRT enzyme and optimized its biochemistry. AS built the nanopore instrument with support from JMC, JRH, and JHG. AS, JMC, HA, JK, KC, and CJB wrote the manuscript. All authors contributed to editing the manuscript.

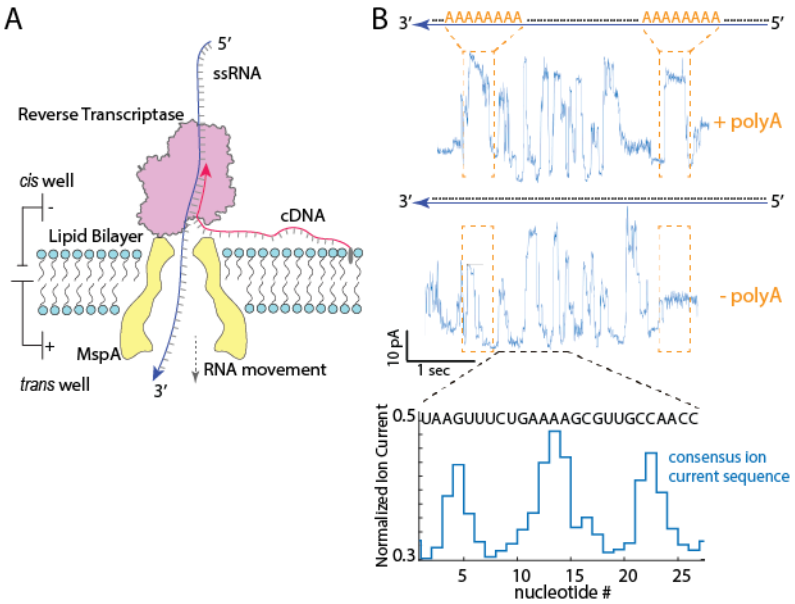

Figure 1. Sequencing RNA with a eukaryotic RT and MspA nanopore. **A**. The instrument setup of the MspA nanopore sequencer. A lipid bilayer was generated to separate two wells containing buffer solutions, a single MspA nanopore was inserted into the lipid membrane, and a bias of 140 mV was applied to the system during data acquisition. Template RNA (blue line) is hybridized with a DNA primer (red line) that is tagged with cholesterol (grey oval) and anchored to the bilayer. After introduction of the RT, it forms an elongation complex that can get captured by the nanopore. The RT will come to rest on top of the nanopore and by cDNA synthesis will continuously thread RNA into the pore in discrete steps. The positions of the polyA sequences are not to scale due to illustration purposes. **B**. Top and middle panel: Ion current signal from the translocation of RNA1 and RNA1_polyA. RNA1_polyA has two polyA sequences inserted close to the 5' end of the RNA about 80 nt apart (top panel, highlighted in orange) while RNA1 does not have polyA inserted (middle panel). The orange rectangles highlight the regions where polyA is inserted, and as seen in the top panel, polyA insertion gave rise to high ion current signal (top panel) that does not exist in the data without polyA insertion (middle panel, orange rectangles). Bottom panel: The blue line represents a segment of the normalized ion current levels from the middle panel, along with RNA sequence aligned to the ion current signals.

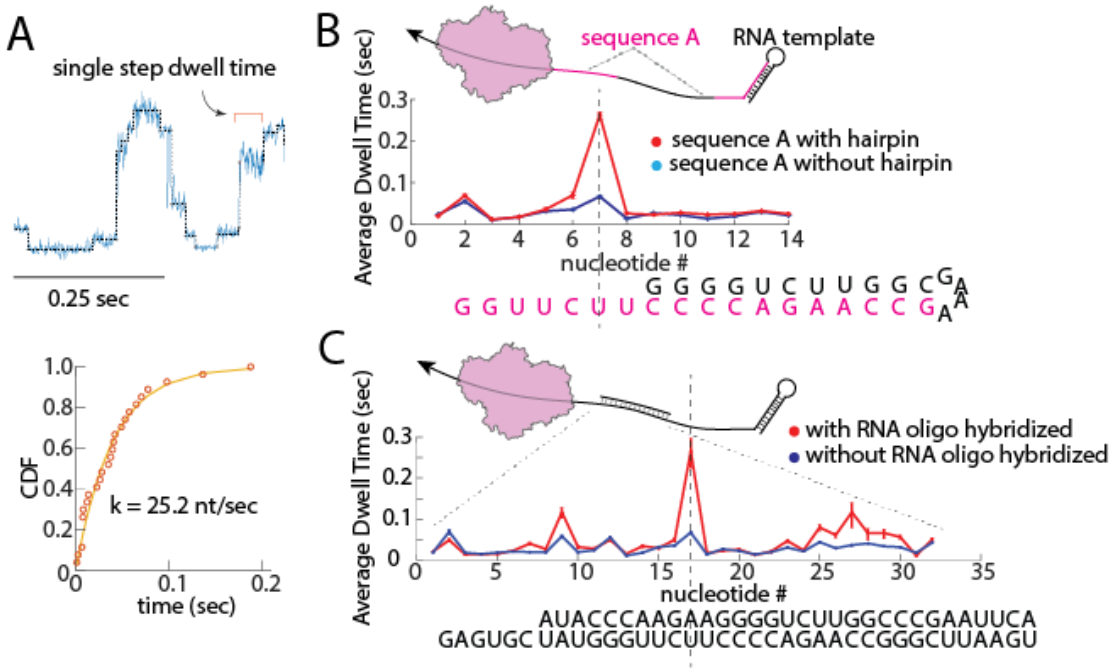

Figure 2. Detecting RNA secondary structures by analyzing the single molecule kinetics of bmRT. **A**. Top panel: Overlay of a segment of raw nanopore ion current from RNA translocation (blue) and the steps found via a point of change algorithm[8] (grey). Single steps can be detected, and their individual dwell times can be quantified by fitting the cumulative distribution function (CDF) of the dwell time of the same step obtained from difference RNA translocation traces to a single exponential (bottom panel). **B**. Top panel: RNA template that contains two repeats of sequence A (highlighted in magenta). The second repeat base pairs to form a stable 5' terminal hairpin. Bottom panel: Dwell time distribution of the first sequence A repeat and second sequence A repeat overlayed, the sequence underneath represent the sequence in the enzyme's catalytic site at every step. Sequence A is highlighted in magenta and the remainder of the terminal hairpin is in black. **C**. Top panel: a 32 nt RNA oligonucleotide (short black line) was hybridized to the RNA template. Bottom panel: Dwell time distribution comparison between the same RNA sequence with (red line) and without (blue line) hybridization of the RNA oligonucleotide. Error bars are 95% confidence interval.

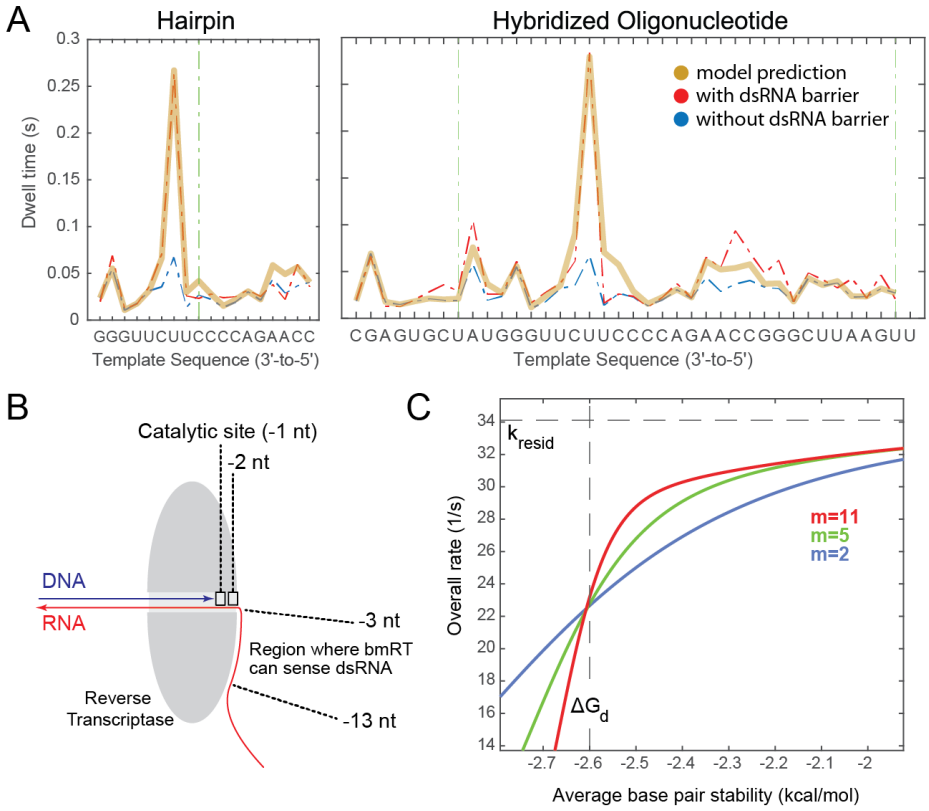

Figure 3. An active helicase model to describe the helicase activity of bmRT. **A.** Agreement of model predictions with experimental kinetic profiles of bmRT over RNA segments that can form a hairpin (left, data from **Figure 2B**) or hybridize to an oligonucleotide (right, data from **Figure 2C**). The major pauses in the presence of barriers (**Figure 2B-C**) are reproduced by the model. See Supplementary Figure 9 for details. **B.** Schematic drawing of the bmRT elongation complex showing the expected relative positions of the polymerase catalytic site (–1) and the closest helicase site (–3) during the dwell time of the enzyme. With this arrangement, the –1 and –2 RNA nucleotides are both unpaired. After incorporation of an incoming dNTP at position –1, translocation by one step would require that the –3 nucleotide becomes unpaired. In this model, the helicase can sense RNA structures both at the –3 position and at further downstream nucleotides up to position –13 or –14 (total length of 11-12 nt), possibly due to preferential binding of the helicase to ssRNA. **C.** Model prediction for the dependence of overall translocation rate as a function of the average base pair stability in downstream RNA. The sigmoidal becomes sharper with increased length (m) of the downstream sensing region following position -3. This plot uses helicase destabilization energy (ΔG_d) of –2.6kcal/mol, single-strand stepping rate of 200 nt/s, and mean residence rate of 34 nt/s, as obtained from the fit to the measurements (Supplementary Figure 9D).

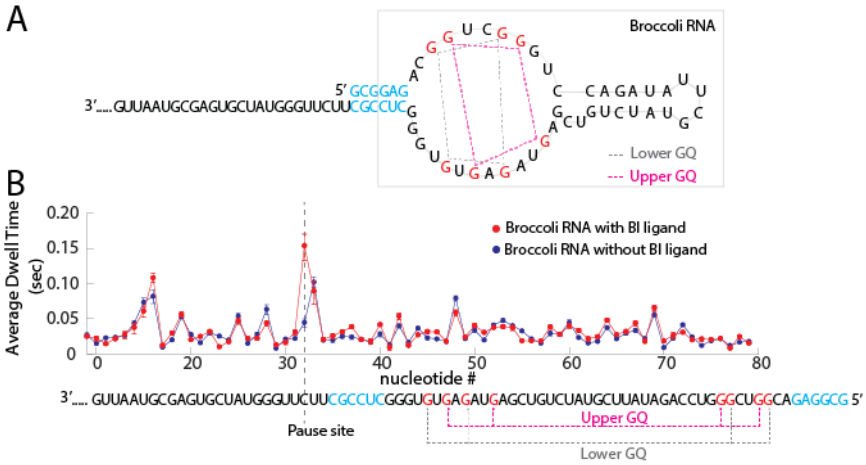

Figure 4. Detecting Broccoli RNA-BI ligand binding using direct RNA nanopore sequencing. **A.** sequence design of RNA3_Broccoli, the G bases involved in GQ formation are highlighted in red. A 6nt RNA duplex that precedes GQs is highlighted in blue. Information about which G bases are involved in GQ formation is obtained from ref. 25. **B.** Average dwell time of bmRT along the Broccoli RNA sequence. The Broccoli RNA sequence is highlighted in orange and its upper and lower GQ is marked in magenta and grey respectively. The BI binding site is located on top of the upper GQ. A strong pause was observed when the bmRT's catalytic site is at nt # 32, which is 12 nt from the start of the lower GQ and 2 nt from the start of the dsRNA duplex. Also see Supplementary Figure 9. Error bars are 95% confidence interval.

References

1. Yao, R.-W., Wang, Y. & Chen, L.-L. Cellular functions of long noncoding RNAs. *Nat. Cell Biol.* **21**, 542–551 (2019).
2. Batista, P. J. & Chang, H. Y. Long noncoding RNAs: cellular address codes in development and disease. *Cell* **152**, 1298–1307 (2013).
3. Mortimer, S. A., Kidwell, M. A. & Doudna, J. A. Insights into RNA structure and function from genome-wide studies. *Nat Rev Genet* **15**, 469–479 (2014).
4. Loughrey, D., Watters, K. E., Settle, A. H. & Lucks, J. B. SHAPE-Seq 2.0: systematic optimization and extension of high-throughput chemical probing of RNA secondary structure with next generation sequencing. *Nucleic Acids Res* **42**, (2014).
5. Umeyama, T. & Ito, T. DMS-seq for In Vivo Genome-Wide Mapping of Protein-DNA Interactions and Nucleosome Centers. *Curr Protoc Mol Biol* **123**, e60 (2018).

6. Parker, M. T. *et al.* Nanopore direct RNA sequencing maps the complexity of Arabidopsis mRNA processing and m6A modification. *Elife* **9**, e49658 (2020).
7. Stephenson, W. *et al.* Direct detection of RNA modifications and structure using single-molecule nanopore sequencing. *Cell Genom* **2**, 100097 (2022).
8. Laszlo, A. H., Derrington, I. M. & Gundlach, J. H. MspA nanopore as a single-molecule tool: From sequencing to SPRNT. *Methods* **105**, 75–89 (2016).
9. Laszlo, A. H. *et al.* Decoding long nanopore sequencing reads of natural DNA. *Nat Biotechnol* **32**, 829–833 (2014).
10. Brinkerhoff, H., Kang, A. S. W., Liu, J., Aksimentiev, A. & Dekker, C. Multiple rereads of single proteins at single-amino acid resolution using nanopores. *Science* **374**, 1509–1513 (2021).
11. Cheng, W., Arunajadai, S. G., Moffitt, J. R., Tinoco, I. & Bustamante, C. Single-base pair unwinding and asynchronous RNA release by the hepatitis C virus NS3 helicase. *Science* **333**, 1746–1749 (2011).
12. Herschhorn, A. & Hizi, A. Retroviral reverse transcriptases. *Cell Mol Life Sci* **67**, 2717–2747 (2010).
13. Zhao, C., Liu, F. & Pyle, A. M. An ultraprocessive, accurate reverse transcriptase encoded by a metazoan group II intron. *RNA* **24**, 183–195 (2018).
14. Upton, H. E. *et al.* Low-bias ncRNA libraries using ordered two-template relay: Serial template jumping by a modified retroelement reverse transcriptase. *Proc Natl Acad Sci U S A* **118**, e2107900118 (2021).
15. Pimentel, S. C., Upton, H. E. & Collins, K. Separable structural requirements for cDNA synthesis, nontemplated extension, and template jumping by a non-LTR retroelement reverse transcriptase. *J Biol Chem* **298**, 101624 (2022).
16. Qu, X. *et al.* The Ribosome Uses Two Active Mechanisms to Unwind mRNA During Translation. *Nature* **475**, 118–121 (2011).

17. Cheng, W., Dumont, S., Tinoco, I. & Bustamante, C. NS3 helicase actively separates RNA strands and senses sequence barriers ahead of the opening fork. *Proc Natl Acad Sci U S A* **104**, 13954–13959 (2007).
18. Vilfan, I. D. *et al.* Analysis of RNA base modification and structural rearrangement by single-molecule real-time detection of reverse transcription. *J Nanobiotechnology* **11**, 8 (2013).
19. Malik, O., Khamis, H., Rudnizky, S., Marx, A. & Kaplan, A. Pausing kinetics dominates strand-displacement polymerization by reverse transcriptase. *Nucleic Acids Res.* **45**, 10190–10205 (2017).
20. Manosas, M., Xi, X. G., Bensimon, D. & Croquette, V. Active and passive mechanisms of helicases. *Nucleic Acids Res* **38**, 5518–5526 (2010).
21. Eickbush, T. H. & Eickbush, D. G. Integration, Regulation, and Long-Term Stability of R2 Retrotransposons. *Microbiol Spectr* **3**, MDNA3-0011–2014 (2015).
22. Amiri, H. & Noller, H. F. A tandem active site model for the ribosomal helicase. *FEBS Lett* **593**, 1009–1019 (2019).
23. Amiri, H. & Noller, H. F. Structural evidence for product stabilization by the ribosomal mRNA helicase. *RNA* **25**, 364–375 (2019).
24. Kim, S. *et al.* Probing allostery through DNA. *Science* **339**, 816–819 (2013).
25. Doty, P., Boedtker, H., Fresco, J. R., Haselkorn, R. & Litt, M. Secondary structure in ribonucleic acids*. *Proceedings of the National Academy of Sciences* **45**, 482–499 (1959).
26. Zuker, M. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res* **31**, 3406–3415 (2003).
27. Filonov, G. S., Moon, J. D., Svensen, N. & Jaffrey, S. R. Broccoli: rapid selection of an RNA mimic of green fluorescent protein by fluorescence-based selection and directed evolution. *J Am Chem Soc* **136**, 16299–16308 (2014).
28. Puchta, O. *et al.* Genotype-phenotype map of an RNA-ligand complex. 2020.12.17.423258 Preprint at https://doi.org/10.1101/2020.12.17.423258 (2020).

29. Vanegas, P. L., Horwitz, T. S. & Znosko, B. M. Effects of non-nearest neighbors on the thermodynamic stability of RNA GNRA hairpin tetraloops. *Biochemistry* **51**, 2192–2198 (2012).

30. Ross, B. C. Mutual Information between Discrete and Continuous Data Sets. *PLoS One* **9**, e87357 (2014).