ACS NANO
Cite This: ACS Nano 2018, 12, 7067−7075
www.acsnano.org

ARTICLE

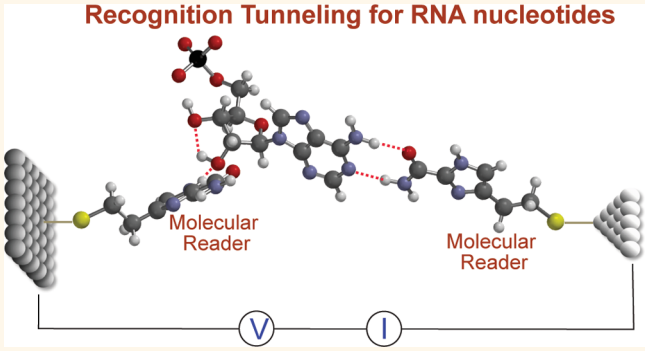
Recognition Tunneling of Canonical and Modified RNA Nucleotides for Their Identification with the Aid of Machine Learning

JongOne Im,[†,∥] Suman Sen,[†,‡,∥] Stuart Lindsay,[*,†,‡,§] and Peiming Zhang[*,†]

[†]Biodesign Institute, [‡]School of Molecular Sciences, and [§]Department of Physics, Arizona State University, Tempe, Arizona 85287, United States

Supporting Information

ABSTRACT: In the present study, we demonstrate a tunneling nanogap technique to identify individual RNA nucleotides, which can be used as a mechanism to read the nucleobases for direct sequencing of RNA in a solid-state nanopore. The tunneling nanogap is composed of two electrodes separated by a distance of <3 nm and functionalized with a recognition molecule. When a chemical entity is captured in the gap, it generates electron tunneling currents, a process we call recognition tunneling (RT). Using RT nanogaps created in a scanning tunneling microscope (STM), we acquired the electron tunneling signals for the canonical and two modified RNA nucleotides. To call the individual RNA nucleotides from the RT data, we adopted a machine learning algorithm, support vector machine (SVM), for the data analysis. Through the SVM, we were able to identify the individual RNA nucleotides and distinguish them from their DNA counterparts with reasonably high accuracy. Since each RNA nucleoside contains a hydroxyl group at the 2′-position of its sugar ring in an RNA strand, it allows for the formation of a tunneling junction at a larger nanogap compared to the DNA nucleoside in a DNA strand, which lacks the 2′ hydroxyl group. It also proves advantageous for the manufacture of RT devices. This study is a proof-of-principle demonstration for the development of an RT nanopore device for directly sequencing single RNA molecules, including those bearing modifications.



KEYWORDS: *recognition tunneling, RNA sequencing, RNA nucleotides, machine learning, SVM*

Next generation sequencing (NGS) has become a workhorse for genomics, being extended to transcriptomics (cDNA-based RNA-sequencing, aka RNA-seq).[1] To sequence RNA, NGS requires a reverse transcription synthesis and PCR amplification of complementary DNA (cDNA)[2]—a process that not only results in loss of information on modifications of RNA but also introduces artifacts and bias to the samples.[3,4] Ozsolak et al. first reported a single molecule method for direct sequencing of RNA by enzymatic synthesis,[5] but it cannot avoid the loss of information about modifications on RNA. In fact, there are more than 100 distinct types of modified RNA nucleosides (http://mods.rna.albany.edu/mods/). There is an urgent need for a technology that can directly read both canonical and modified nucleosides in an RNA strand.

With a vision of a low-cost and hand-held device for sequencing, nanopores—orifices of nanometer diameters—have widely been explored as an electronic sequencing device since the beginning of the NIH $1000 genome program.[6] Oxford Nanopore Technologies is the first company that launched a nanopore sequencing product, MinIon, on the market, based on ionic current measurements (https://nanoporetech.com). The MinIon sequencer has recently finished sequencing a human genome.[7] Furthermore, it has also sequenced RNA transcripts directly but only shown marginal effectiveness in identifying methylated bases.[8] The protein−nanopore sequencing reads at least three nucleotides at a time,[9] a situation unlikely to be improved in even atomically thin pores[10] because the access resistance dominates the signal.[11]

Received: April 15, 2018
Accepted: June 22, 2018
Published: June 22, 2018

DOI: 10.1021/acsnano.8b02819
ACS Nano 2018, 12, 7067−7075

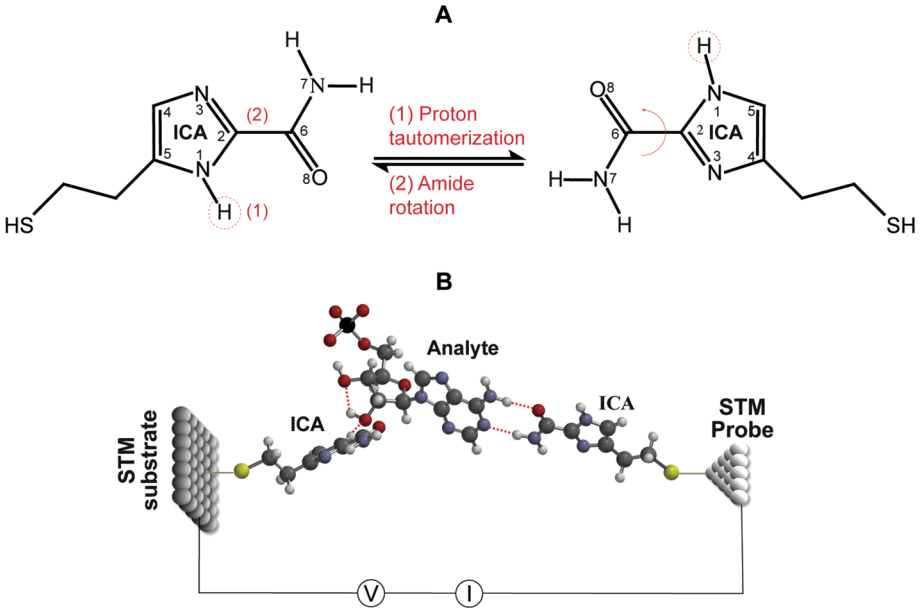


Figure 1. (A) Structure of ICA and its tautomeric equilibrium. (B) Illustration of a scanning tunneling microscope (STM) setup for the RT measurement.

To address the resolution issue of nanopores, we have been developing a recognition tunneling (RT) method to read individual nucleotides in a nucleic acid sequence.[12] According to Zwolak and Di Ventra's theoretical study, the electron tunneling transverse to the long axis of a DNA strand can be localized enough to sense just one base at a time.[13] Thus, we devised an RT device consisting of a two-electrode nanogap functionalized with a recognition molecule. It can capture an analyte to form a tunneling junction through noncovalent chemical bonds. Such a structure allows for the electron tunneling through a nanogap of ~2.5−3.0 nm and avoids the nonspecific adsorption that could result in the RT device losing its functions rapidly.[14] The RT device can also be incorporated into solid-state nanopores using semiconductor fabrication techniques for massively parallel production. Previously, we demonstrated that the DNA bases, including methylated cytosine, could be identified by RT.[15] That led us to explore the potential of RT for use in the identification of RNA nucleotides. In this paper, we show that RT can read not only the canonical but also the modified RNA nucleotides, which may allow us to use our DNA sequencing chip (currently under development) for direct sequencing of RNA with some alterations. Meanwhile, we extended a machine learning algorithm, support vector machine (SVM) that we have developed for analysis of RT data,[16−18] for the present study. The SVM is a supervised learning model to separate different classes in a hyperdimensional space,[19] used in genomics[20,21] and single molecule detection.[22,23] Here, we report a proof-of-principle study using a scanning tunneling microscope (STM) and free nucleotides to demonstrate the RT measurement. A real sequencing device would require tunnel junctions embedded in the solid-state nanopores, a means for translocation control, and the recognition molecules that can effectively discriminate among the RNA nucleotides.

RESULTS AND DISCUSSION

To read individual RNA nucleotides, we used 4(5)-(2-mercaptoethyl)-1*H*-imidazole-2-carboxamide (ICA), which was originally developed as a universal reader to read individual DNA nucleobases through hydrogen bonding in a tunneling nanogap.[24] As shown in Figure 1A, ICA has a structure with a rotatable amide connected to an imidazole ring for the molecular recognition, a flexible two carbon chain bearing a thiol group at its end for attachment to the metal surface, and a proton that is exchangeable between two tautomers. For the RT measurement, we employed a scanning tunneling microscope (STM) to generate nanogaps by controlling its conductance. Its substrate and tip were functionalized with ICA (Figure 1B). Four canonical RNA-nucleosides (designated rA, rC, rG, and rU) and two modified RNA-nucleosides (N^6-methyladenosine and inosine, designated rm^6A and rI) were used as analytes (see Figure 2*i* for their structures). By convention, their monophosphates are designated rNMP (N represents different nucleosides), generally called nucleotides. These two modified RNA nucleotides have been found in epitranscriptomes,[25] which may be related to the evolution of cancer.[26]

RT of RNA Nucleoside Monophosphates. RT measurements were carried out in a 1.0 mM phosphate buffered aqueous solution (pH 7.4), collecting useful signals up to 25 kHz in frequency. Palladium (Pd) metal was used as an electrode material for both substrates and probes. The probes were partially insulated with polyethylene,[27] and both probes and substrates were functionalized with ICA immediately before use.[17] We found that a tunneling set point of 4 pA at a bias of 0.5 V, which corresponds to a gap distance around 2.4 nm,[28] produced RT signals from the RNA nucleotides. At this set point, a phosphate buffer control was almost free of spikes (Figure S1, Supporting Information (SI)). For the RT measurement, the tunnel gap was first stabilized under the predefined set point of current and bias in STM until the baseline current was noise free, and then an analyte injected into the liquid cell with a final concentration of ~0.1 mM, followed by recording electrical signals for 20 min. Each analyte was measured three times using freshly prepared probes, substrates, and samples. As a result, we collected three sets of RT data for each of the RNA nucleotides. A typical RT

DOI: 10.1021/acsnano.8b02819
ACS Nano 2018, 12, 7067−7075

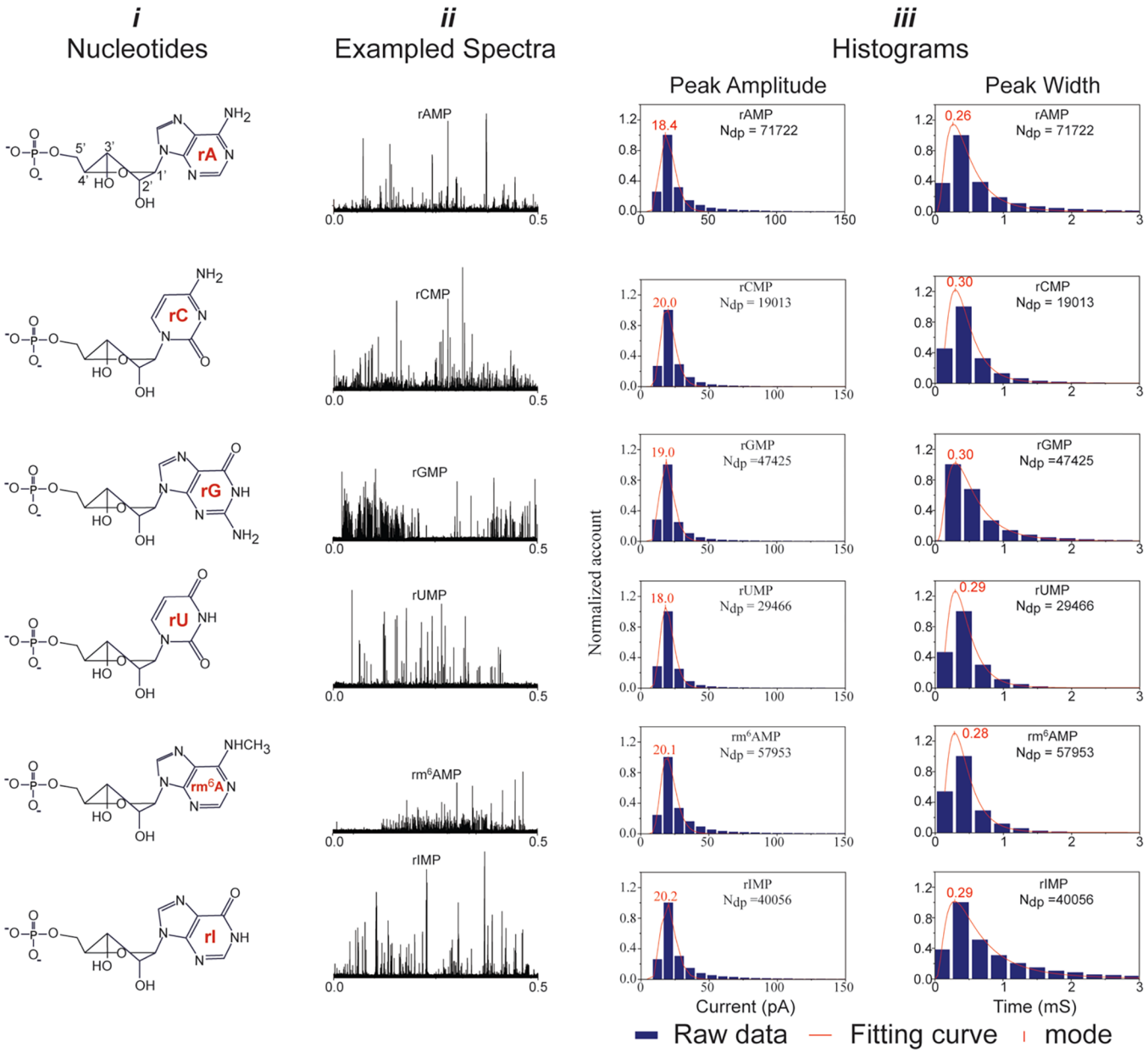


Figure 2. (i) Chemical structures of RNA nucleoside monophosphates used in this study; (ii) trace examples of RT current vs time signals generated with ICA at a set point of 4 pA and 0.5 V; (iii) histograms of RT current amplitude and. peak widths with their fitting curves (shown in red).

Table 1. Primary Parameters of RT Signals for the RNA Nucleotides Derived from Curve Fitting of the Raw Data with the Lognormal Function[a,b]

	current amplitude (pA)			peak width (ms)		
RNA nucleotide	median	mean	σ	median	mean	σ
rAMP	20.3	21.1	5.9	0.40	0.50	0.36
rCMP	19.9	20.7	5.6	0.40	0.46	0.26
rGMP	19.5	20.2	5.3	0.45	0.56	0.40
rUMP	19.5	20.1	5.3	0.38	0.44	0.25
avg	19.8 ± 0.8	20.5 ± 0.5	5.5 ± 0.3	0.41 ± 0.03	0.49 ± 0.05	0.32 ± 0.07
modified RNA nucleotide	median	mean	σ	median	mean	σ
rIMP	20.2	20.9	5.8	0.53	0.72	0.65
rm^{6}AMP	20.5	21.4	6.1	0.38	0.43	0.25

[a]Fitting errors are smaller than 0.1 for all parameters and all fittings have adjusted $R^2 > 0.98$. [b]Each standard deviation (σ) is related to the mean value that it is close to in the same row.

DOI: 10.1021/acsnano.8b02819
ACS Nano 2018, 12, 7067−7075

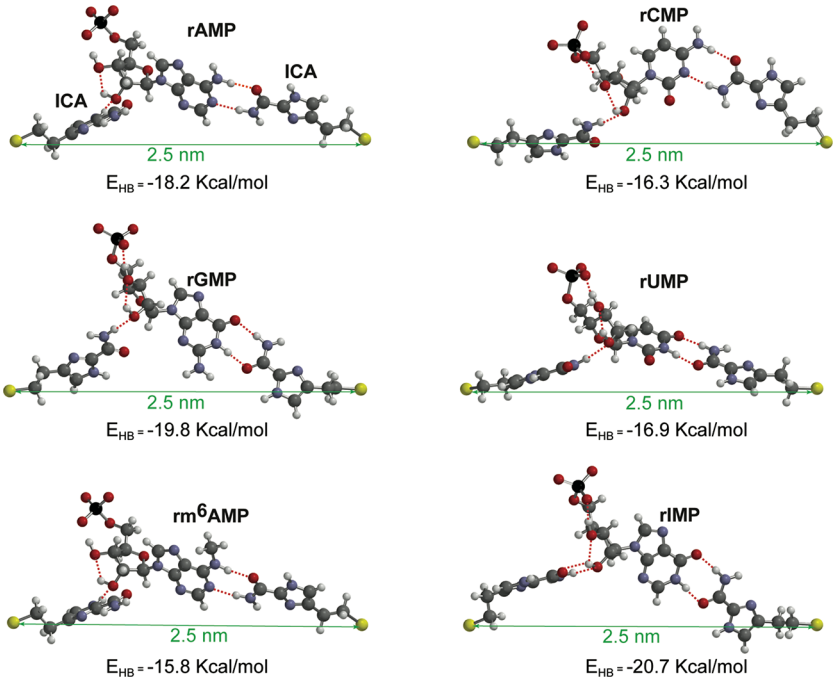


Figure 3. Computational models of canonical and modified RNA nucleotides interacting with ICA through hydrogen bonding in a tunneling gap, calculated by density functional theory (B3LYP/6-31G*) in a water environment with two sulfur atoms at a fixed distance of 2.5 nm.

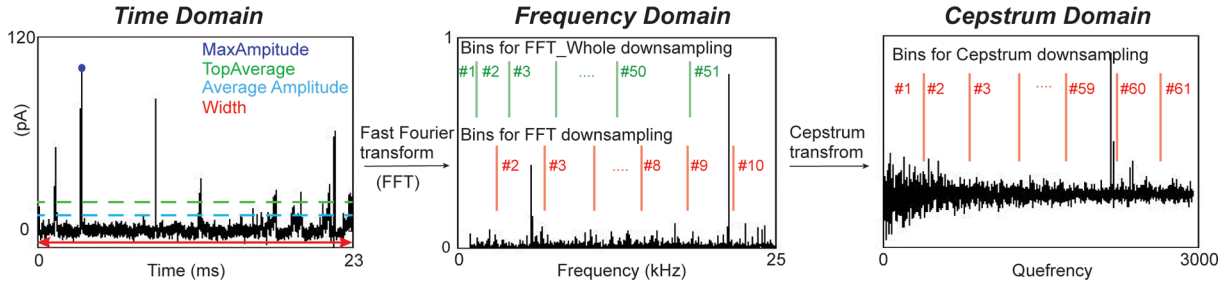


Figure 4. Process of extracting features from an RT spectrum for signal characterization (see the Experimental Methods for details).

current trace vs time recording for each measured nucleotide is displayed in Figure 2*ii*, appearing as a train of stochastic current spikes. To analyze the RT data, we first determined the current amplitude and peak width of each spike, which are defined as the primary parameters, and then plotted them in histograms (Figure 2*iii*). Since they are not symmetrically distributed and skewed to the right, we fit each histogram to the Lognormal function (red curves in Figure 2*iii*). The fitting results for each nucleotide are summarized in Table 1. First, we found that each canonical RNA nucleotide has a marginally higher conductance on average, but a similar distribution, compared to its DNA counterpart (Figure S2 and Table S1, SI). This result may indicate that these nucleotides formed a very similar complex with ICA in the tunneling gap. In contrast, the peak widths for the canonical RNA nucleotides are significantly smaller than those for their DNA counterparts on average. They appear widely distributed with relatively large standard derivations (σ) for both of the DNA and RNA nucleotides. The peak width may indicate the residence time of a nucleotide in the tunneling nanogap, reflecting the stability of a nucleotide binding to the ICA molecules to a certain degree. In this regard, the DNA nucleotide may form a complex with ICA that is more stable than its RNA counterpart does. Interestingly, the modified RNA nucleotide rIMP formed the most stable complex and rm^{6}AMP the least stable complex with ICA in the tunneling nanogap, but their conductivities are very similar.

To better understand the RT data, we carried out a computer modeling study. We first built the hydrogen-bonded complexes of RNA nucleotides with two ICA molecules fixed at a distance of 2.5 nm between their sulfur atoms using the molecular mechanics, followed by a DFT calculation with B3LYP/6-31G* in a water environment (see Experimental Methods for details). As shown in Figure 3, each RNA nucleotide interacts with one ICA molecule through its nucleobase by two hydrogen bonds and the other one through its 2′ hydroxyl group by one hydrogen bond. The structural similarity among these complexes may be the reason why we measure small differences in their conductance. We calculated the intermolecular hydrogen bonding energies (E_{HB}) by subtracting the DFT energies of two ICA molecules and one RNA nucleotide in their free states from the total energy of their complexes. The results show stabilities of the RNA nucleotide-ICA triplets in order of rIMP > rGMP > rAMP > rUMP ~ rCMP > rm^{6}AMP, well correlated to the mean value order of their RT peak widths.

Calling of Individual Nucleotides by SVM. As shown in Figure 2, the current amplitudes of the RNA nucleotides, as

DOI: 10.1021/acsnano.8b02819
ACS Nano 2018, 12, 7067−7075

well as their peak widths, are overlapped with one another so these parameters cannot simply be used for the identification of the RNA nucleotides. We have been successful in calling the DNA nucleotides using the SVM.[17] In the present study, we adopted it to call the RNA nucleotides from their RT signals. First, the RT data in the time domain were subjected to Fourier transform (FFT), followed by cepstrum transform (Figure 4, and also see the Experimental Methods for details). From the three domains, a series of features were extracted (Table S2, SI) and used to index each of the RT spikes. Meanwhile, we identified and removed those spikes common to all analytes owing to contamination, capture events that were insensitive to chemical variation, noise spikes generated by STM electronics and servo control.[29] Since SVM is a supervised learning process, it requires training to build a model for the classification of newly generated data. In the present study, we randomly took 10% of the RT data of the four canonical RNA nucleotides from their first two sets of collections to train the SVM with the extracted features until each of the spikes was successfully assigned to its corresponding nucleotides (Table 2, 100% classification). It is essential that the training sets contain a significant amount of data, but not all of it; otherwise, the testing becomes meaningless, which is why we chose the10% for training. Also, we envisage that a real sequencing device would be calibrated before the data acquisition, and 10% data allocation for calibration is reasonable. Then the trained SVM was engaged to call the RNA nucleotides from the remaining RT data, and the calling accuracy is designated as optimistic accuracy (because this is an ideal case where the training and test data come from the same pool). As a result, the SVM called rAMP and rGMP with an accuracy of 99.1% and 92.0% and rUMP and rCMP with an accuracy of 90.7% and 87.6%, respectively. On average, the optimistic accuracy of calling the canonical RNA nucleotides was ~92.4% (Table 2). Furthermore, we engaged the same SVM to call the RNA nucleotides from the third set of their RT data. The resultant calling accuracy is designated as predictive accuracy, representative of a situation where the existing RT data are used as a reference for SVM to analyze the data generated by a newly fabricated RT device. The predictive accuracy of calling the canonical RNA nucleotides was 83.8% on average, ~8% lower than the optimistic accuracy. By comparison, the optimistic accuracy of calling the canonical RNA nucleotides is about 5% lower than the one for the DNA counterparts on average (Table S3, SI). Unexpectedly, the predictive accuracy of calling the canonical RNA nucleotides is 20% higher than the one for DNA nucleotides. We also notice that the accuracies for calling the purine nucleotides are higher than those for the pyrimidine nucleotides, which correlates with the stability of their hydrogen bonding complexes.

To test the SVM for a possible bias to individual nucleotides, we trained it with different combinations of the features. Three of the SVMs achieved a 100% training accuracy. As shown in Table 3, these SVMs called each of the canonical RNA nucleotides with the variations less than 2% for their optimistic accuracies and 5% for their predictive accuracies (highlighted in red). We have noticed that the SVM trained with more features would not always give a higher calling accuracy. By comparison, it would allow us to select the most appropriate SVM for the calling. In Table 3, we also include the error rates of each nucleotide miscalled by the SVMs (black numbers), to which it is worth paying attention. For example, rCMP was miscalled as rUMP with an error rate of 10.7%, much higher than as rAMP and rGMP (0.6% and 2.3%, respectively). This result indicates that the SVM miscalling is not random, having a bias in favor of a particular nucleotide. In the extreme, the modified RNA nucleotides were completely miscalled because they were not involved in the SVM training. The nucleotide rm^{6}AMP was miscalled rAMP with an error rate of as high as 77% (bold number in the rm^{6}AMP column, Table 3). This miscalling could be attributed to the structural similarity between rm^{6}AMP and rAMP. Both of them share the same Watson Crick base pairing edges so they may form similar

Table 2. Highest Accuracy (%) Achieved by SVM Calling of the Individual Canonical RNA Nucleotides

RNA nucleotide[a]						
SVM training		SVM calling accuracy (%)				
running Mode	separation (%)	rAMP	rCMP	rGMP	rUMP	mean ± σ
optimistic	100	99.1	87.6	92.0	90.7	92.4 ± 4.9
predictive	N/A	97.2	82.2	88.2	67.7	83.8 ± 12.4

[a]Taken from Table 3 (*vide infra*).

Table 3. Accuracy of RNA Nucleotides Called by the SVMs Trained with Different Numbers of Features[a,b]

SVM training			Calling Rate (%)				Calling Rate (%)	
			Nucleotides in training				Nucleotides Not in training	
Feature Number	Running Mode	Separation (%)	rAMP	rCMP	rGMP	rUMP	rm^{6}AMP	rIMP
103	Optimistic	100	99.2|0.4|0.3|0.1	0.6|86.4|2.3|10.7	0.3|1.7|92.3|5.6	0.4|8.2|2.7|88.7	N/A	N/A
	Predictive	N/A	97.9|0.2|1.8|0.1	1.5|85.8|1.8|10.9	0.5|1.9|89.5|8.1	0.8|23.8|10.7|64.7	**77.3**|6.4|4.2|12.1	1.4|35.2|24.8|**38.6**
125	Optimistic	100	99.1|0.5|0|0.4	0.2|87.6|2.6|9.6	0.3|2.6|92.0|5.1	0.1|7.3|1.9|90.7	N/A	N/A
	Predictive	N/A	97.2|0.4|2.0|0.4	4.0|82.2|2.4|11.4	0.4|2.7|88.2|8.7	0.3|22.6|9.4|67.7	**75.0**|7.1|3.5|14.4	2.3|**30.9**|**27.1**|39.7
142	Optimistic	100	99.1|0.5|0.3|0.1	0.4|86.6|2.3|10. 7	0.3|2.8|90.9|5.9	0.3|7.9|2.0|89.8	N/A	N/A
	Predictive	N/A	97.5|0.4|2.0|0.1	2.4|81.3|4.2|12.1	0.2|3.1|85.2|11.5	0.6|21.7|9.2|68.5	**76.7**|5.3|3.6|14.4	1.6|**28.8**|**28.0**|41.6

[a]See Table S4-1 (SI) for a list of features involved in each training and Table S4-2 (SI) for the event numbers generated by each nucleotide used in the SVM analysis. [b]The SVM calling results are listed in the order of A|C|G|U for each nucleotide. The correct calls are highlighted in red, and incorrect ones are shown black.

DOI: 10.1021/acsnano.8b02819
ACS Nano 2018, 12, 7067−7075

Table 4. Accuracy of the SVM Calling rm6AMP[a−c]

	Calling Rate (%)					Calling Rate (%)
	Nucleotides in training					Nucleotide not in training
Running Mode	rAMP	rCMP	rGMP	rUMP	rm^{6}AMP	rIMP
Optimistic	**99.5**|0.4|0 |0|0.1	0.7|**79.1**|3.6 |6.4|10.2	0.5|2.1|**89.0** |1.5|6.9	0.3|8.5|3.3| **77.9** |10.0	0|6.8|3.2| 4.9|**85.1**	N/A
	Average: 86.1±8.7					
Predictive	**98.1**|0.1|1.6 |0.1|0.1	0.9|**16.7**|1.8 |**74.7**|6.0	0.1|1.4|**86.3** |2.9|9.3	0.7|15.2|13.3 |**58.9**|11.9	0.5|13.7|8.9| **28.5**|**48.4**	**78.0**|4.6|2.2|6.3|8.9
	Average: 61.7±32.2					

[a]See Table S5-1 (SI) for a list of features involved in the SVM training and Table S5-2 (SI) for the event numbers generated by each nucleotide used in the SVM analysis. [b]The results were generated using the SVM with 100% training accuracy. [c]The SVM calling results are listed in order of A|C|G|U|m^{6}A for each nucleotide. The correct calls are highlighted in red, and incorrect ones are shown black.

Table 5. Accuracy of the SVM Calling rIMP[a−c]

	Calling Rate (%)					Calling Rate (%)
	Nucleotides in training					Nucleotide not in training
Running Mode	rAMP	rCMP	rGMP	rUMP	rIMP	rm^{6}AMP
Optimistic	89.4|0.2|0.2 |0|10.2	0.1|83.4|3.1 |9.1|4.3	0.1|2.5|91.2 |5.1|1.1	0.2|9.2|3.6 |83.7|3.3	3.3|6.7|1.2 |10.5|78.2	N/A
	Average: 85.2±5.2					
Predictive	66.8|0.1|1.3 |0.2|31.6	3.6|85.1|1.5 |6.9|2.9	0.2|1.9|88.2 |8.7|1.0	0.3|21|10 |60.7|8.0	63.9|0.2|2.1 |0.1|33.7	1.4|35.3|23.8|34.5|5.0
	Average: 66.9±21.9					

[a]See Table S6-1 (SI) for a list of features involved in the SVM training and Table S6-2 (SI) for the event numbers generated by each nucleotide used in the SVM analysis. [b]The results were generated using the SVM with 100% training accuracy. [c]The SVM calling results are listed in order of A|C|G|U|I for each nucleotide. The correct calls are highlighted in red, and incorrect ones are shown black.

hydrogen-bonding complexes with ICA in the tunneling gap (see Figure 2), resulting in the miscalling bias of the SVM. On the other hand, rIMP was miscalled rUMP, rCMP, and rGMP with an error rate of ∼39%, 35%, and 25%, respectively. This multinucleotide miscalling may be explained by the universal base characteristic of inosine, which can base-pair with canonical nucleobase A, C, G, and U(T).[30] Without supervising, the SVM would misidentify those unexpected incomers in a sample.

To call the modified RNA nucleotide, we pooled its RT data with those of the canonical RNA nucleotides in the SVM training. Table 4 shows the accuracy results of SVM calling rm^{6}AMP as well as the four canonical RNA nucleotides. The SVM called rm^{6}AMP with an optimistic accuracy of 85.1% and miscalled it as rAMP, rCMP, rGMP, and rUMP with an error rate of 0%, 6.8%, 3.2%, and 4.9%, respectively. In reverse, the SVM miscalled rAMP, rCMP, rGMP, and rUMP as rm^{6}AMP with an error rate of 0.1%, 10.2%, 6.9% and 10.0%, respectively. These results show that the SVM can effectively distinguish between rAMP and rm^{6}AMP because there is a negligible miscall between them. Since its RT data was not involved in the training, rIMP was miscalled mainly as rAMP with an error rate as high as 78%, but as rm^{6}AMP only with an 8.9% error rate. Thus, the SVM can be used to distinguish rm^{6}AMP from rIMP.

In the same manner, we trained the SVM with a pool of the RT data of rIMP and four canonical RNA nucleotides. As a result, the SVM called rIMP with an optimistic accuracy of 78.2%, while miscalling it as rAMP, rCMP, rGMP, and rUMP with an error rate of 3.3%, 6.7%, 1.2% and 10.5%, respectively (Table 5). Conversely, the SVM miscalled rAMP, rCMP, rGMP, and rUMP as rIMP with an error rate of 10.2%, 4.3%, 1.1%, and 3.3%, respectively. All of these miscalling rates are much lower than the random pick (the probability of which is 20%). Without including rm^{6}AMP in training, the SVM miscalled it majorly as rCMP, rGMP, or rUMP with a rate of 35.3%, 23.8%, and 34.5% respectively. Again, it is confirmed that the SVM can distinguish these two modified RNA nucleotides from each other as rm^{6}AMP was miscalled as rIMP only with an error rate of 8.9% and rIMP as rm^{6}AMP with an error rate of 5.0%.

The calling accuracy of the SVM falls with the increase in the complexity of the data pool. When the SVM was trained to call all the six RNA nucleotides, its optimistic accuracy was reduced to ∼71.4% on average (Table S7, SI), compared to ∼92.4% for calling the four (Table 2) and ∼86.0% for calling the five (Tables 4 and 5). We notice that the predictive accuracy is even more susceptible to the data complexity. On average, the predictive accuracy is ∼46.4% for calling the six nucleotides (Table S7, SI), much lower than that for calling the four (∼83.8%, Table 2). Simultaneously, the miscalling rate increases with the decrease of the predictive accuracy. For example, rCMP was miscalled as rUMP with an error rate of as high as 74.7% (Table 4), and rIMP was miscalled as rAMP

DOI: 10.1021/acsnano.8b02819
ACS Nano 2018, 12, 7067−7075

Table 6. Accuracy of the SVM Trained with Different Number of Features to Call the Eight Canonical Nucleotides from a RT Data Pool[a]

SVM training (4 dNMPs + 4 rNMPs)		optimistic accuracy (%)								
		DNA nucleotide				RNA nucleotide				
no. of features	separation (%)	dAMP	dCMP	dGMP	dTMP	rAMP	rCMP	rGMP	rUMP	avg
41	100	77.4	75.5	78.1	82.7	93.9	82.3	90.7	80.5	82.6 ± 6.5
			subaverage: 78.4 ± 3.1				subaverage: 86.9 ± 6.5			
54	100	73.6	74.7	81.5	83.2	96.4	80.7	91.3	81.4	82.9 ± 7.7
			subaverage: 78.3 ± 4.8				subaverage: 87.5 ± 7.7			
70	100	75.6	77.5	77.2	83.5	97.0	79.8	90.6	83.9	83.1 ± 7.4
			subaverage: 78.5 ± 3.5				subaverage: 87.8 ± 7.7			
92	100	78.5	79.5	78.2	79.1	96.2	79.6	88.9	83.3	82.9 ± 6.5
			subaverage: 78.8 ± 0.6				subaverage: 87.0 ± 7.2			
93	100	72.0	79.4	77.4	81.2	97.0	82.1	87.5	85.2	82.7 ± 7.5
			subaverage: 77.5 ± 4.0				subaverage: 88.0 ± 6.4			

[a]See Table S8-1 (SI) for a list of features involved in each training and Table S8-2 (SI) for the event numbers generated by each nucleotide used in the SVM analysis.

with an error rate of 63.9% (Table 5). There is only one exception in which the predictive accuracy of calling rCMP is higher than the optimistic one (Table 5). In terms of the optimistic accuracy, the miscalling rate is only up to ∼10% in the worst-case scenario. In general, the tips are handcrafted for the STM measurement, which inevitably results in a variation from one tip to another even though we follow the same protocol for the preparation. That may be one of the sources which would reduce the predictive accuracy. We expect that the massively parallel production of the RT chips in a silicon wafer should reduce the variation from device to device to improve the predictive accuracy. Nonetheless, a reference run may be needed for each of the RT devices to have its SVM with a high calling accuracy.

Furthermore, we tested the SVM for the identification of both canonical DNA and RNA nucleotides in an RT data pool. The significant difference between these two type of nucleotides lies in their sugar structures. The DNA nucleotide has only one hydroxyl group at the 3′-position and its sugar mainly adopts a 2′-endo pucker (Figure S2*i*, SI). By comparison, the RNA nucleotide has two hydroxyl groups at the 2′ and 3′ position respectively, and its sugar mainly adopts a 3′-endo pucker (Figure 2*i*). In the same way, as described above, the SVM was trained with 10% of the RT data randomly taken from the pool, and then the trained SVMs were used to assign all the spikes remaining in the data pool. Again, a 100% training accuracy could be achieved using different combinations of the features, but a larger number of features would not always give a higher calling accuracy on average (Table 6). The highest optimistic accuracy of calling individual DNA and RNA nucleotides in the RT data pool was ∼83% on average, which was achieved with 70 of the features. Interestingly, the SVM called the RNA nucleotides with an accuracy of 88%, about ∼9% higher than that for calling the DNA nucleotides. Moreover, it called the purine RNA nucleotides with an accuracy of >90%, higher than those for the pyrimidine RNA nucleotides. This study demonstrates that RT can potentially be used to identify those RNA nucleotides misincorporated into a DNA sequence,[31] an error not addressed by NGS. It also suggests that more effective molecular readers may be needed specifically for the identification of different RNA nucleosides.

CONCLUSION

The goal of this study was to demonstrate the feasibility of identifying individual RNA nucleotides using a tunneling nanogap, which will be incorporated into a solid-state nanopore for reading the RNA sequences. As a proof of concept, we used STM, an instrument that can rapidly generate nanogaps, for the tunneling measurement. Given that chemical modifications in RNA play an important role in epitranscriptome,[32−34] we included two well-known modified RNA nucleotides, as well as the four canonical RNA nucleotides, in the present study. Under a voltage bias, the negatively charged nucleotide was injected to the nanogap and captured by ICA, where it would experience a series of thermal fluctuations—association with both of the substrate and tip, partial dissociation, reassociation, or full dissociation to be replaced by another molecule. An RT signal may reveal the information on these processes. Our DFT models show that two ICA molecules can interact with an RNA nucleotide via hydrogen bonding to the 2′-hydroxyl groups and Watson−Crick base pairing edges of the nucleotide respectively in a 2.5 nm wide nanogap. We found that the intermolecular hydrogen-bonding energies correlated well with the most probable peak widths, which suggests that hydrogen-bonding interactions may be responsible for the capture events in the tunneling nanogap. Previously, we confirmed that a non-hydrogen bonding reader could not generate RT signals with abasic deoxynucleotide and glucose in the tunneling nanogap, but ICA did,[17] a piece of evidence that ICA can interact with the ribose of an RNA nucleotide through hydrogen bonding. Similarly, the DNA nucleotides interact with ICA through their 3′-hydroxyl groups and nucleobases (Figure S3, SI). In this way, their tunneling pathways would be longer than their RNA counterparts. We found that the amplitude of tunneling current with the DNA nucleotides is marginally smaller than with RNA nucleotides on average. However, the canonical DNA nucleosides have larger signal peak widths than their RNA counterparts. As far as the modified RNA nucleotides are concerned, we also noticed that rIMP has the largest peak width and rm6AMP the smallest one, but their amplitudes are close to those of canonical RNA nucleotides.

The RNA nucleotides cannot be identified merely by the current amplitudes and peak widths of their RT data as having discussed above. In general, the RT signals are susceptible to

DOI: 10.1021/acsnano.8b02819
ACS Nano 2018, 12, 7067−7075

bonding geometry, which varies with the thermal fluctuations (and competition with water molecules),[35] containing rich information on the structures and physical properties of the tunneling junctions. We transformed the RT data collected in the time domain to the frequency and cepstrum data, creating an excessive number of features for the identification of the RNA nucleotides by machine learning. With the aid of SVM, we were able to identify the four canonical RNA nucleotides with the optimistic accuracy of >90% and to call up to five nucleotides with an optimistic accuracy of ∼85% on average. The calling accuracy may further be improved by deep learning,[36] which will be explored in the future. The SVM is a supervised machine learning method so that the training data have to closely represent those to be analyzed in order to achieve a high calling accuracy. Based on the present study, one would expect the loss of accuracy when the SVM is trained with the data acquired in one junction and applied to the data from another ("predictive" accuracy) because of a junction to junction variation. Also, one would expect to obtain the higher "optimistic" accuracy by running a reference sample through the junction for training an SVM or choosing one from those accumulated in a data bank with its training set closely matched to the RT data of the reference.

The present study provides a key parameter necessary for the design of an RT device. It shows that a gap distance of 2.5 nm should be sufficient for RNA nucleotides to elicit the RT signals. Since RNA contains a hydroxyl group in each of its sugar rings, a tunneling junction can be formed through the RNA nucleoside hydrogen bonding to the reader molecules such as ICA. Because DNA lacks a hydroxyl group in its sugar rings, most likely, it forms the tunneling junctions through its nucleobases, which requires the RT device to have a smaller nanogap for the interactions. A larger nanogap is always advantageous for the manufacturing production. Currently, we are working on a device with a tunneling nanogap embedded in a solid-state nanopore for sequencing.[37]

EXPERIMENTAL METHODS

RT Measurement. RNA nucleoside monophosphates were purchased as sodium salts from Sigma-Aldrich. Water was purified from a Milli-Q system with a specific resistance of 18 MΩ-cm and total organic carbon below 5 ppb. The stock solution of each analyte was prepared in a 1.0 mM phosphate buffer with pH 7.4. The Pd tips and substrates were functionalized with ICA following a procedure described previously.[18]

In a typical RT experiment, the measurement followed a process of mounting the functionalized Pd-STM probe and Pd-substrate into a PicoSPM STM, stabilizing the tunneling nanogap in a phosphate buffer (1.0 mM, 7.4 pH) until a clean baseline was generated (∼2 h), injecting an analyte solution to the liquid cell with a final concentration of ∼100 μM, and then collecting current recordings under a predefined substrate bias. For each analyte, three separate experiments were carried out each with freshly made probes, substrates, and samples. The RT data are stored in https://www.dropbox.com/sh/djmejsmwsd0gtsr/AADSuDhIgVAkmM9Es_lBQYrna?dl=0 for open access.

Computer Modeling. DFT calculations were performed using the program Spartan'16 for Windows, Wavefunction Inc. 2D chemical structures of the RNA nucleotides and ICA were drawn in ChemDraw Professional 17 and exported to Spartan'16 to generate the corresponding 3D structures, where hydrogen bonded triplexes were formed by energy minimization with the distance between two sulfur atoms constrained at a distance of 2.5 nm using the built-in MMFF94s molecular mechanics. The complexed structures were further optimized using B3LYP/6-31G* in water with the two sulfur atoms constrained at the distance of 2.5 nm.

Extraction of Features from RT Data. To characterize the RT signals, we categorized current spikes as peaks and clusters. First, the baseline of tunneling current corresponding to the set point current of 4 pA was moved to be zero, and we defined all the individual spikes larger than 15 pA as peaks. A cluster was determined by convoluting the Gaussian window of a 4096 data-point full width and a unit height to the center of each peak. The convoluted Gaussian traces were summed, and a range of the cluster was determined by applying a threshold of 0.1 to the summed Gaussian trace. All the current spikes within the range were assigned into the same cluster including those smaller than 15 pA.

Signal features were extracted from three different domains of time, frequency, and cepstrum (Figure 4). Each peak or cluster detected in time-domain was Fourier transformed into 25 kHz frequency spectrum, which is the Nyquist frequency of the instrument. The resultant Fourier transformed spectrum was down-sampled into two types of windows: (a) uniform window, 10 of which were defined with a 2.5 kHz bandwidth for peaks and 51 of which with a 490 Hz bandwidth for clusters (red vertical lines, Figure 4) and (b) variable windows, a total of 51, the size of which was varied by a power 2 square law from 0 to the Nyquist frequency (green vertical lines, Figure 4). Furthermore, the frequency spectra were transferred to the cepstrum domain by inverse Fourier transform of a logarithm of the squared magnitude of the frequency spectrum. The cepstrum was down-sampled into 61 of the uniform windows with a 410 Hz bandwidth. All of the signal features are listed in Table S2. To avoid a problem with a large numeric number becoming dominant, we rescaled all the features by shifting its mean to zero and having its standard deviation (s.d.) be one.

Data Analysis. Data recorded in the time domain were characterized by averaged amplitude and peak width, which were determined using MATLAB. First, the baseline tunneling current was shifted to zero, and a threshold was set at 15 pA for the tunneling spikes. These data were exported to OriginPro 2018 for the analysis and fit with Lognormal using the built-in Levenberg−Marquardt algorithm to have an adjusted $R^2 > 0.98$.

SVM Analysis. We began with rescaled 264 signal features and filtered out "bad" features, which were determined by correlation coefficient between feature pairs as well as statistical variation score of each feature calculated by a ratio between repeated experiment and different analytes variations. Those features above 0.7 correlation coefficient were replaced with a representative feature. Also, we ranked the features based on their statistical variation scores and removed 15 features of the lowest scores. The subsets of features for SVM analysis were selected by randomly or by their statistical variation scores. The optimized feature set was selected to provide the maximum true positive accuracy. We adopted the kernel-mode SVM from https://github.com/vjethava/svm-theta. The SVM running parameters C and γ were optimized through cross-validation of the randomly selected data pool. Full details of the SVM (written in MATLAB) can be found in the Web site: https://github.com/ochensati/SVM_DNA_TunnelVision.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acsnano.8b02819.

Supporting tables and figures (PDF)

AUTHOR INFORMATION

Corresponding Authors

*E-mail: stuart.lindsay@asu.edu.

*E-mail: peiming.zhang@asu.edu.

ORCID

Peiming Zhang: 0000-0003-2831-2308

DOI: 10.1021/acsnano.8b02819
ACS Nano 2018, 12, 7067−7075

Author Contributions

[‖]J.I. and S.S. contributed equally to this work.

Notes

The authors declare the following competing financial interest(s): S.B., S.S., S.L., and P.Z. are named as inventors for patent applications.

ACKNOWLEDGMENTS

This work was supported by Grant No. 5R01HG009180-02 from the National Human Genome Research Institute.

REFERENCES

(1) Shendure, J.; Balasubramanian, S.; Church, G. M.; Gilbert, W.; Rogers, J.; Schloss, J. A.; Waterston, R. H. DNA Sequencing at 40: Past, Present and Future. *Nature* **2017**, *550*, 345−353.

(2) van Dijk, E. L.; Jaszczyszyn, Y.; Thermes, C. Library Preparation Methods for Next-Generation Sequencing: Tone down the Bias. *Exp. Cell Res.* **2014**, *322*, 12−20.

(3) Ozsolak, F.; Milos, P. M. RNA Sequencing: Advances, Challenges and Opportunities. *Nat. Rev. Genet.* **2011**, *12*, 87−98.

(4) Potapov, V.; Ong, J. L. Examining Sources of Error in PCR by Single-Molecule Sequencing. *PLoS One* **2017**, *12*, e0169774.

(5) Ozsolak, F.; Platt, A. R.; Jones, D. R.; Reifenberger, J. G.; Sass, L. E.; McInerney, P.; Thompson, J. F.; Bowers, J.; Jarosz, M.; Milos, P. M. Direct RNA Sequencing. *Nature* **2009**, *461*, 814−818.

(6) Branton, D.; Deamer, D. W.; Marziali, A.; Bayley, H.; Benner, S. A.; Butler, T.; Di Ventra, M.; Garaj, S.; Hibbs, A.; Huang, X.; Jovanovich, S. B.; Krstic, P. S.; Lindsay, S.; Ling, X. S.; Mastrangelo, C. H.; Meller, A.; Oliver, J. S.; Pershin, Y. V.; Ramsey, J. M.; Riehn, R.; et al. The Potential and Challenges of Nanopore Sequencing. *Nat. Biotechnol.* **2008**, *26*, 1146−1153.

(7) Jain, M.; Koren, S.; Miga, K. H.; Quick, J.; Rand, A. C.; Sasani, T. A.; Tyson, J. R.; Beggs, A. D.; Dilthey, A. T.; Fiddes, I. T.; Malla, S.; Marriott, H.; Nieto, T.; O'Grady, J.; Olsen, H. E.; Pedersen, B. S.; Rhie, A.; Richardson, H.; Quinlan, A. R.; Snutch, T. P.; et al. Nanopore Sequencing and Assembly of a Human Genome With Ultra-Long Reads. *Nat. Biotechnol.* **2018**, *36*, 338−345.

(8) Garalde, D. R.; Snell, E. A.; Jachimowicz, D.; Sipos, B.; Lloyd, J. H.; Bruce, M.; Pantic, N.; Admassu, T.; James, P.; Warland, A.; Jordan, M.; Ciccone, J.; Serra, S.; Keenan, J.; Martin, S.; McNeill, L.; Wallace, E. J.; Jayasinghe, L.; Wright, C.; Blasco, J. Highly Parallel Direct RNA Sequencing on an Array of Nanopores. *Nat. Methods* **2018**, *15*, 201−206.

(9) Laszlo, A. H.; Derrington, I. M.; Gundlach, J. H. MspA Aanopore as a Single-Molecule Tool: From Sequencing to SPRNT. *Methods* **2016**, *105*, 75−89.

(10) Arjmandi-Tash, H.; Belyaeva, L. A.; Schneider, G. F. Single Molecule Detection with Graphene and Other Two-Dimensional Materials: Nanopores and Beyond. *Chem. Soc. Rev.* **2016**, *45*, 476−493.

(11) Lindsay, S. The Promises and Challenges of Solid-State Sequencing. *Nat. Nanotechnol.* **2016**, *11*, 109−111.

(12) Lindsay, S.; He, J.; Sankey, O.; Hapala, P.; Jelinek, P.; Zhang, P.; Chang, S.; Huang, S. Recognition Tunneling. *Nanotechnology* **2010**, *21*, 262001.

(13) Zwolak, M.; Di Ventra, M. Electronic Signature of DNA Nucleotides *via* Transverse Transport. *Nano Lett.* **2005**, *5*, 421−424.

(14) Smith, T. The Hydrophilic Nature of a Clean Gold Surface. *J. Colloid Interface Sci.* **1980**, *75*, 51−53.

(15) Huang, S.; He, J.; Chang, S.; Zhang, P.; Liang, F.; Li, S.; Tuchband, M.; Fuhrmann, A.; Ros, R.; Lindsay, S. Identifying Single Bases in a DNA Oligomer with Electron Tunnelling. *Nat. Nanotechnol.* **2010**, *5*, 868−873.

(16) Zhao, Y.; Ashcroft, B.; Zhang, P.; Liu, H.; Sen, S.; Song, W.; Im, J.; Gyarfas, B.; Manna, S.; Biswas, S.; Borges, C.; Lindsay, S. Single-Molecule Spectroscopy of Amino Acids and Peptides by Recognition Tunnelling. *Nat. Nanotechnol.* **2014**, *9*, 466−473.

(17) Biswas, S.; Sen, S.; Im, J.; Biswas, S.; Krstic, P.; Ashcroft, B.; Borges, C.; Zhao, Y.; Lindsay, S.; Zhang, P. Universal Readers Based on Hydrogen Bonding or $\pi-\pi$ Stacking for Identification of DNA Nucleotides in Electron Tunnel Junctions. *ACS Nano* **2016**, *10*, 11304−11316.

(18) Im, J.; Biswas, S.; Liu, H.; Zhao, Y.; Sen, S.; Biswas, S.; Ashcroft, B.; Borges, C.; Wang, X.; Lindsay, S.; Zhang, P. Electronic Single-Molecule Identification of Carbohydrate Isomers by Recognition Tunnelling. *Nat. Commun.* **2016**, *7*, 13868.

(19) Mohammed, M.; Khan, M. B.; Bashier, E. B. M. *Machine Learning Algorithms and Applications*; CRC Press, 2017.

(20) Athamanolap, P.; Parekh, V.; Fraley, S. I.; Agarwal, V.; Shin, D. J.; Jacobs, M. A.; Wang, T. H.; Yang, S. Trainable High Resolution Melt Curve Machine Learning Classifier for Large-Scale Reliable Genotyping of Sequence Variants. *PLoS One* **2014**, *9*, e109094.

(21) Erlich, Y.; Mitra, P. P.; delaBastide, M.; McCombie, W. R.; Hannon, G. J. Alta-Cyclic: A Self-Optimizing Base Caller for Next-Generation Sequencing. *Nat. Methods* **2008**, *5*, 679−682.

(22) Winters-Hilt, S. Nanopore Detector Based Analysis of Single-Molecule Conformational Kinetics and Binding Interactions. *BMC Bioinf.* **2006**, *7* (Suppl 2), S21.

(23) Kolmogorov, M.; Kennedy, E.; Dong, Z.; Timp, G.; Pevzner, P. A. Single-Molecule Protein Identification by Sub-Nanopore Sensors. *PLoS Comput. Biol.* **2017**, *13*, e1005356.

(24) Liang, F.; Li, S.; Lindsay, S.; Zhang, P. Synthesis, Physicochemical Properties, and Hydrogen Bonding of 4(5)-Substituted 1-*H*-Imidazole-2-Carboxamide, A Potential Universal Reader for DNA Sequencing by Recognition Tunneling. *Chem. - Eur. J.* **2012**, *18*, 5998−6007.

(25) Li, X.; Xiong, X.; Yi, C. Epitranscriptome Sequencing Technologies: Decoding RNA Modifications. *Nat. Methods* **2017**, *14*, 23−31.

(26) Jiang, Q.; Crews, L. A.; Holm, F.; Jamieson, C. H. M. RNA Editing-Dependent Epitranscriptome Diversity in Cancer Stem Cells. *Nat. Rev. Cancer* **2017**, *17*, 381−392.

(27) Chang, S.; Sen, S.; Zhang, P.; Gyarfas, B.; Ashcroft, B.; Lefkowitz, S.; Peng, H.; Lindsay, S. Palladium Electrodes for Molecular Tunnel Junctions. *Nanotechnology* **2012**, *23*, 425202.

(28) Chang, S.; He, J.; Zhang, P.; Gyarfas, B.; Lindsay, S. Gap Distance and Interactions in a Molecular Tunnel Junction. *J. Am. Chem. Soc.* **2011**, *133*, 14267−14269.

(29) Chang, S.; Huang, S.; Liu, H.; Zhang, P.; Liang, F.; Akahori, R.; Li, S.; Gyarfas, B.; Shumway, J.; Ashcroft, B.; He, J.; Lindsay, S. Chemical Recognition and Binding kinetics in a Functionalized Tunnel Junction. *Nanotechnology* **2012**, *23*, 235101.

(30) Watkins, N. E., Jr.; SantaLucia, J., Jr. Nearest-Neighbor Thermodynamics of Deoxyinosine Pairs in DNA Duplexes. *Nucleic Acids Res.* **2005**, *33*, 6258−6267.

(31) Williams, J. S.; Kunkel, T. A. Ribonucleotides in DNA: Origins, Repair and Consequences. *DNA Repair* **2014**, *19*, 27−37.

(32) Liu, N.; Pan, T. N^6-Methyladenosine-Encoded Epitranscriptomics. *Nat. Struct. Mol. Biol.* **2016**, *23*, 98−102.

(33) Harcourt, E. M.; Kietrys, A. M.; Kool, E. T. Chemical and Structural Effects of Base Modifications in Messenger RNA. *Nature* **2017**, *541*, 339−346.

(34) Kleiner, R. E. Reading the RNA Code. *Biochemistry* **2018**, *57*, 11−12.

(35) Krstic, P.; Ashcroft, B.; Lindsay, S. Physical Model for Recognition Tunneling. *Nanotechnology* **2015**, *26*, 084001.

(36) Albrecht, T.; Slabaugh, G.; Alonso, E.; Al-Arif, S. Deep Learning for Single-Molecule Science. *Nanotechnology* **2017**, *28*, 423001.

(37) Pang, P.; Ashcroft, B. A.; Song, W.; Zhang, P.; Biswas, S.; Qing, Q.; Yang, J.; Nemanich, R. J.; Bai, J.; Smith, J. T.; Reuter, K.; Balagurusamy, V. S. K.; Astier, Y.; Stolovitzky, G.; Lindsay, S. Fixed-Gap Tunnel Junction for Reading DNA Nucleotides. *ACS Nano* **2014**, *8*, 11994−12003.

DOI: 10.1021/acsnano.8b02819
ACS Nano 2018, 12, 7067−7075