Nucleotide and Structural Label Identification in Single RNA Molecules with Quantum Tunneling Spectroscopy

Electronic Supplementary Information

Gary R. Abel, Jr.,[1,2] Lee E. Korshoj,[1,2] Peter B. Otoupal,[1] Sajida Khan,[1,2] Anushree Chatterjee,[1] Prashant Nagpal[1,2,3*]

[1]*Department of Chemical and Biological Engineering, University of Colorado Boulder*
[2]*Renewable and Sustainable Energy Institute (RASEI), University of Colorado Boulder*
[3]*Materials Science and Engineering, University of Colorado Boulder*

Table of Contents:

Supporting Figures:

Generating kernel density estimates from parameter distributions

Using the mean and standard deviation of all measurements $x_1, x_2, x_3, \ldots x_n$ in a data set (rA, rG, rC, rU, and rN + NMIA), the parameters can be summarized as kernel density estimation curves, with probability density, f, found from the following expression:

$$f(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)$$

Here, n is the number of data points, h is the bandwidth parameter calculated from the common estimation in terms of n and standard deviation σ as $h \approx 1.06\sigma n^{-1/5}$, and K is the Gaussian kernel defined as $K(u) = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}u^2\right)$.

Base-calling algorithms used in RNA sequence identification and structural label mapping

Supervised machine learning was used for base calling in both RNA sequence identification (rA, rG, rC, rU discrimination) and structural label mapping (rN ± NMIA label detection). Specifically, the naïve Bayes classifier formed the basis of our base-calling algorithm (and all modified versions of the algorithm). As described in the text, this system uses Bayes' theorem to take in values for each of the parameters and assign a classification based on the probability of each parameter corresponding to the various classes. In the case of RNA sequence identification, there are four classes: rA, rG, rC, and rU. For structural mapping, there are two classes: rN − NMIA and rN + NMIA, (where N is A, G, C, or U). To ensure that both RNA sequence identification and structural mapping results were as robust as possible, the testing and training data was drawn from a library of parameter values for each ribonucleotide and NMIA-labeled ribonucleotide formed from hundreds of STS measurements, as described in the main text. These libraries were randomly split into fourths for base calling analyses with 4-fold cross-validation of results. Additionally, at each coverage level, *n*X (or *n* repeated measurements/reads), 200 combinations of *n* measurements were tested and averaged for each class (i.e., 800 total different combinations of *n* measurements for each 4-fold cross-validation trial for RNA sequence identification results). By using separate testing and training data sets with 4-fold cross-validation and extensive combining of testing measurements, the results presented here rigorously demonstrate the ability of our algorithm to accurately base call signals from unknown measurements.

The output from our base calling algorithms (detailed in Figures 3-6 of the main text) includes recall of each class, overall accuracy, and confidence in the base call. From a confusion matrix analysis, recall of a specific class is True Positives/(True positives + False Negatives), and the overall accuracy is the average recall of all classes multiplied by 100%. Confidence is another important metric for assessing base calling. The confidence in calling a particular base can be calculated using the probability values from the base calling algorithm in the form $C_i = (P_i - P_j)/P_i$. Here, C_i is the confidence for calling base i, P_i is the probability value associated with the called base, and P_j is the second largest probability (for the second most probable base). This confidence value is also indicative of the signal-to-noise level.

The algorithm employed for the initial RNA sequence identification analysis and for all of the structural label mapping results was a typical naïve Bayes classifier. This algorithm was applicable since a single set of parameters was used for classification. For RNA sequence identification, this includes the results using only HOMO and LUMO as parameters as well as the results with all 12 biophysical parameters (results in Figure 3a,b). For the structural label mapping, this includes the label detection for each modified ribonucleotide using a single subset of the 12 parameters (results in Figure 6e,f).

A more advanced, modified naïve Bayes classification algorithm was used for the optimal RNA sequence identification results (Figure 3c and Figure 4a, as well as the conductance screening results in Figure 4b,c). The modified algorithm uses multiple subsets of the 12 biophysical parameters in different combinations along with probability weighting coefficients. This algorithm is based on the idea that different parameter subsets can lead to maximal recall of different nucleobases, and can therefore be used together for fine-tuning nucleobase recognition. Specifically, the algorithm uses four different subsets of parameters, which were put together around a pair of two "fundamental" parameters that demonstrated maximum recall. Additional parameters are stacked alongside the fundamental parameters to provide small perturbations that can increase or decrease recall for individual ribonucleotides. These additional parameters are chosen from base calling tests in which parameters are successively added to the fundamental pair to search for specific combinations where recall is enhanced, in analogy to the 'parameter tuning' described previously.[1] When implemented, this process leads to four separate base calls (one call for each of the four parameter sets run through a standard naïve Bayes classifier). The four calls are resolved into a single call with weight coefficients that have been optimized (through numerical

convergence) to provide a weight for each ribonucleotide called by each parameter set. In the end, the resolved base calls are at a higher accuracy than any one of the single parameter sets alone. A schematic and description of the modified naïve Bayes algorithm, along with the specific parameter subsets and weight coefficients for results both with and without conductance screening, is provided in Figure S5. Detailed output from the algorithm at 35X coverage (including probability values, confidence of base calling, and accuracy) for no conductance screening, low-conductance screening, and high-conductance screening are shown in Figures S6, S7, and S8, respectively. These plots show all 800 base calls from which a representative section of 50 base calls was selected and shown in Figure 4.

References

1 L. E. Korshoj, S. Afsari, S. Khan, A. Chatterjee and P. Nagpal, *Small,* 2017, **13**, 1603033.

2 M. Petri, D. M. Kolb, U. Memmert and H. Meyer, *Electrochim. Acta,* 2003, **49**, 175-182.

3 M. W. Schmidt, K. K. Baldridge, J. A. Boatz, S. T. Elbert, M. S. Gordon, J. H. Jensen, S. Koseki, N. Matsunaga, K. A. Nguyen and S. Su, *Journal of computational chemistry,* 1993, **14**, 1347-1363.

Supplementary Figures

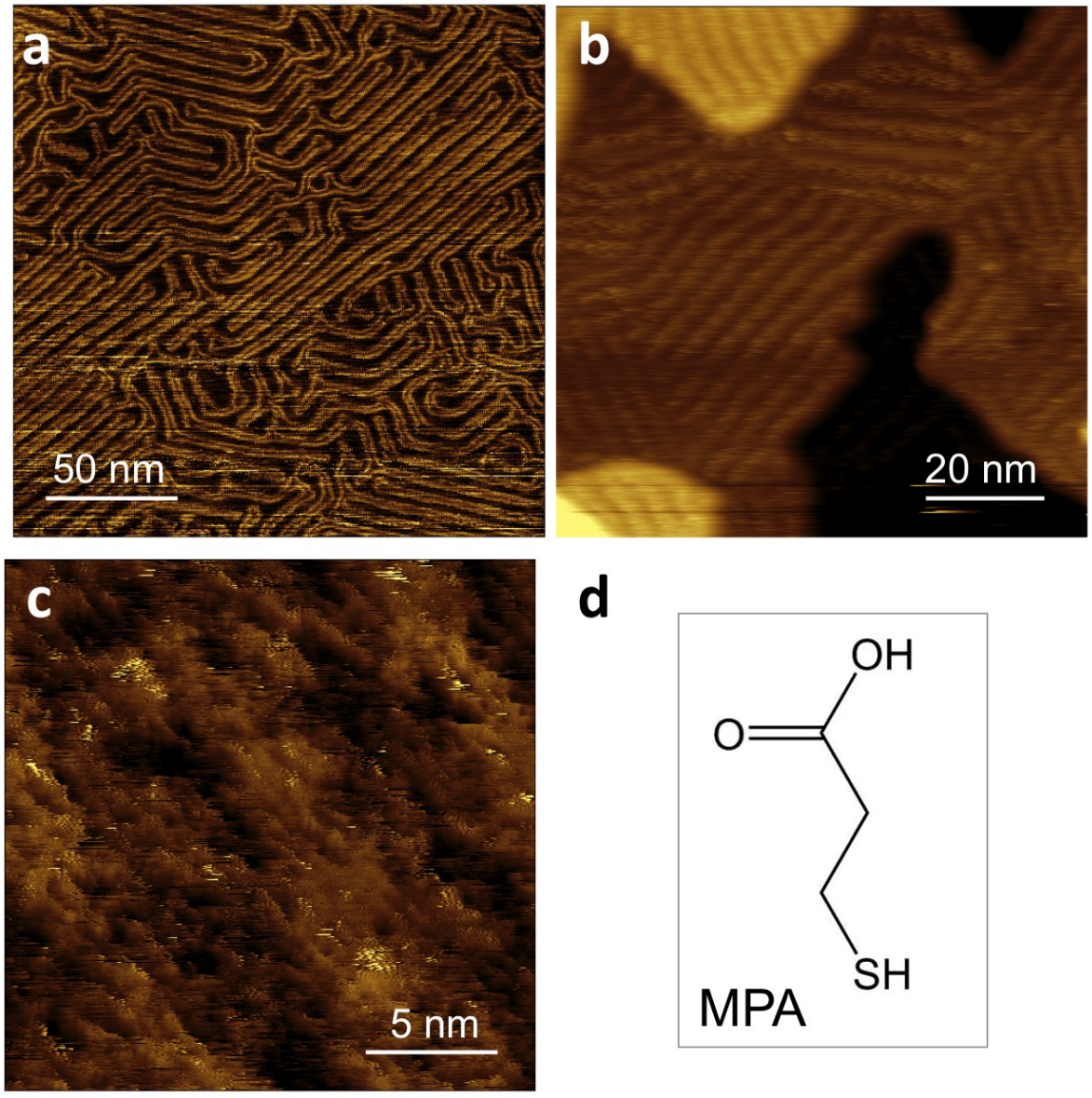

Figure S1— Representative STM images of Au, MPA/Au, and RNA on MPA/Au. (a) STM image of a clean Au (111) terrace, with the characteristic herringbone reconstruction visible. (b) STM image of MPA monolayer on Au (111). Several atomic steps and terraces are visible, as well as the previously observed threefold-symmetric striped domains. [2] (c) STM image of a densely-packed layer of poly-$(rC)_7$ RNA on MPA, with clusters of individual RNA strands visible. Images were acquired while operating in constant current mode, with an imaging bias of −0.5 V and a setpoint of −100-200 pA. (d) Molecular structure of 3-mercaptopropionic acid (MPA). STM images were leveled by mean plane subtraction and flattened line-by-line using Gwyddion image analysis software (http://gwyddion.net/).

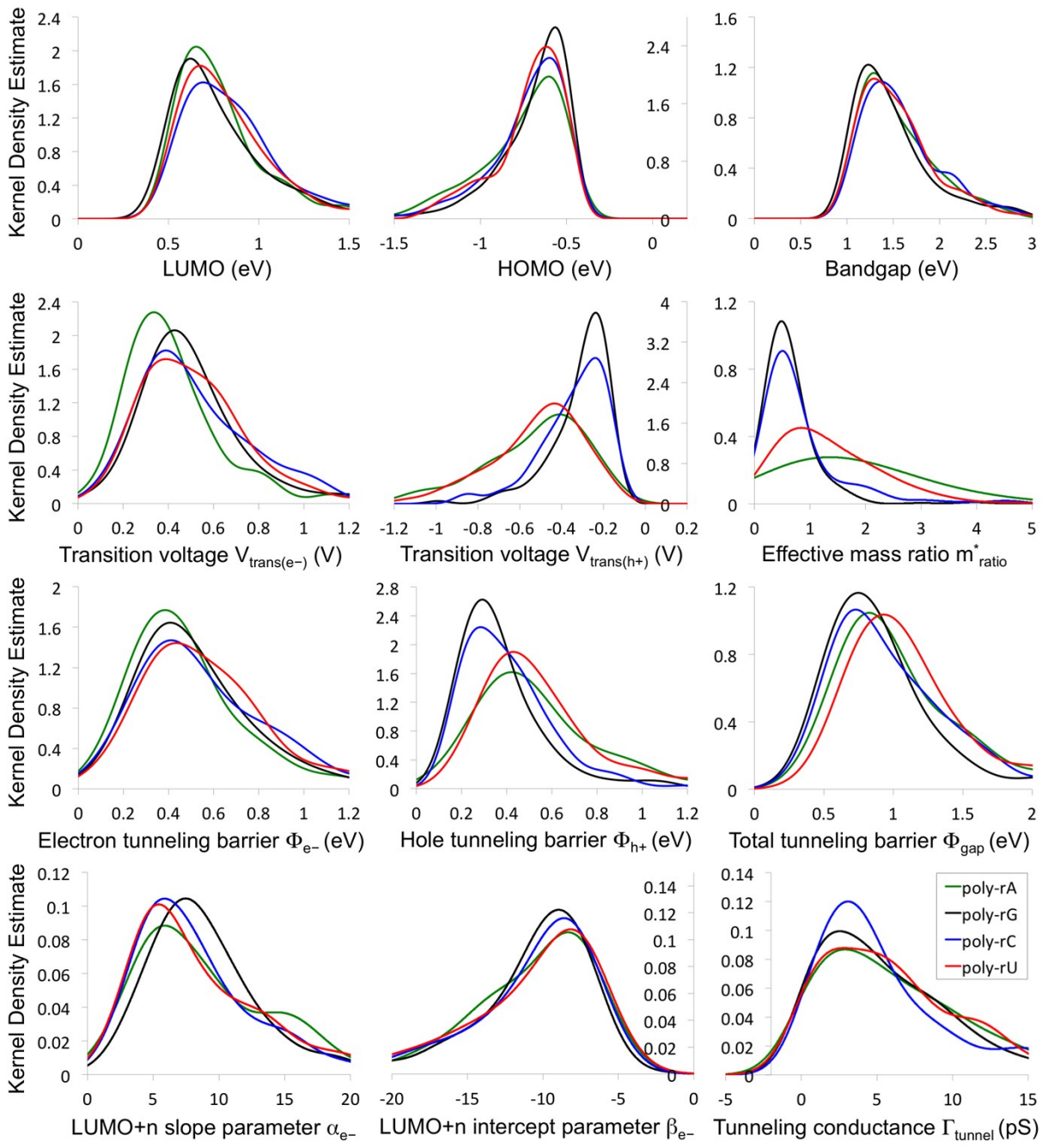

Figure S2— Kernel density estimates of all 12 biophysical parameters. Data was collected on unmodified poly-$(rN)_7$ RNA at high surface-coverage on MPA. Each curve represents ~200 STS measurements.

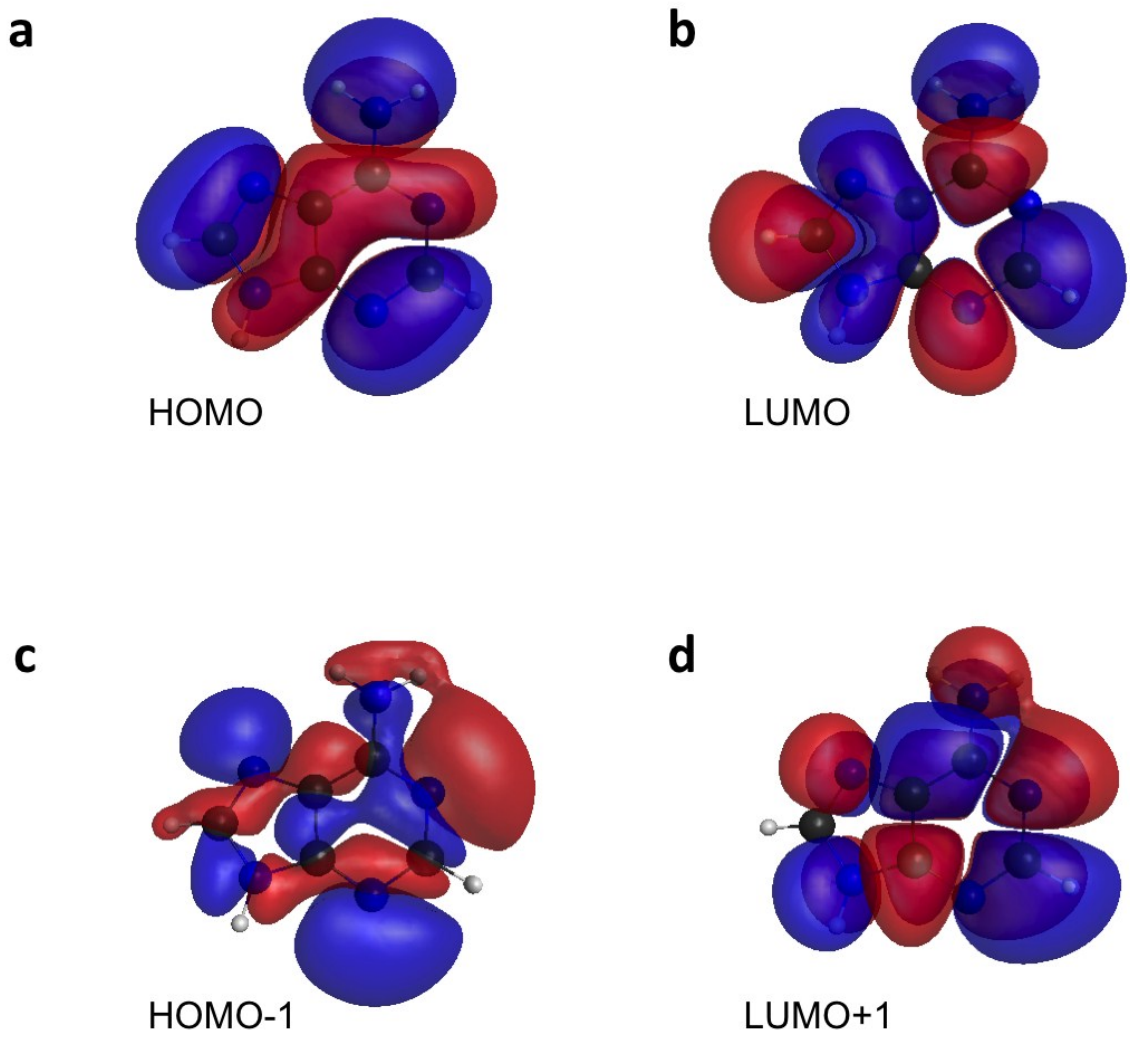

Figure S3— Results of density functional theory (DFT) calculations of the molecular orbitals of adenine. Shown are the calculated molecular orbital wavefunction isosurfaces for the (a) HOMO, (b) LUMO, (c) HOMO−1, and (d) LUMO+1 energy levels of the adenine nucleobase. While the electron wavefunctions of the frontier orbitals (LUMO, HOMO) mainly extend over the conjugated nucleobase, modifications of the adjacent sugar are potentially better characterized by probing the higher-energy orbital wavefunctions due to their better overlap with the sugar backbone. DFT calculations were performed with the GAMESS software package, [3] using the restricted Hartree-Fock method with a 6-311++G(2d,2p) basis set and the Becke 3-parameter hybrid density functional (B3YLP).

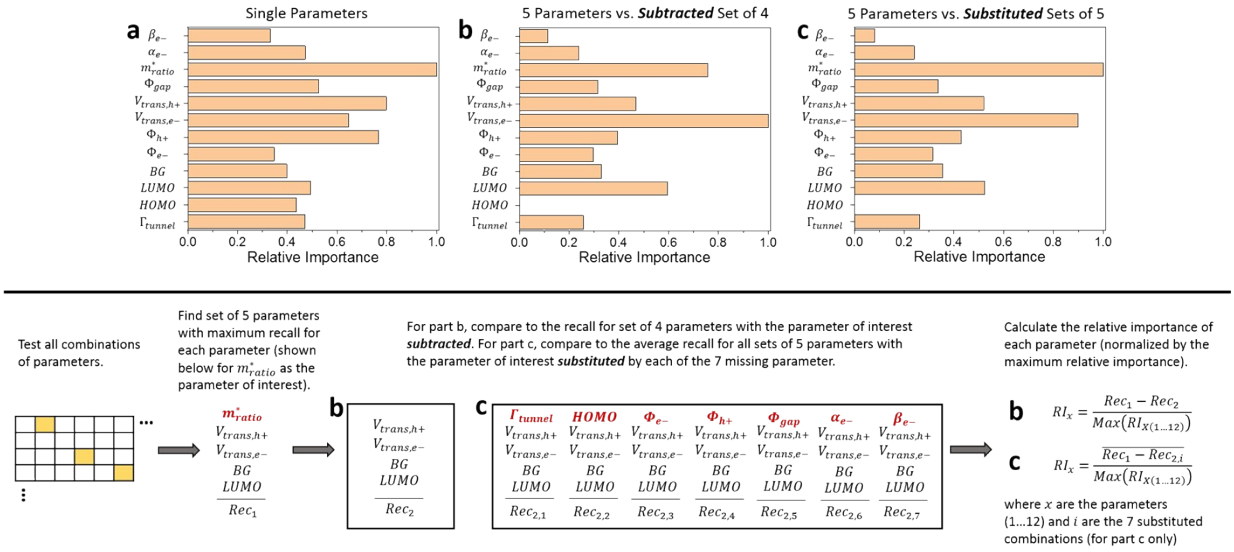

Figure S4— Determining relative importance of the parameters. (a) Relative importance of the 12 parameters calculated from recall using each individual parameter, normalized to the maximum. (b) Relative importance of the 12 parameters calculated from comparing the set of 5 parameters with best recall to the set of 4 parameters where the parameter of interest is removed, normalized by the maximum. (c) Relative importance of the 12 parameters calculated from comparing the set of 5 parameters with best recall to the average recall of the 7 additional sets of 5 parameters where the parameter of interest is substituted by one of the missing parameters, normalized by the maximum. Shown at the bottom is a schematic description of the method used to generate plots b and c. All recall results are for 25X coverage and no conductance screening. All three cases (a, c, c) show a consistent trend in the relative importance of the different parameters.

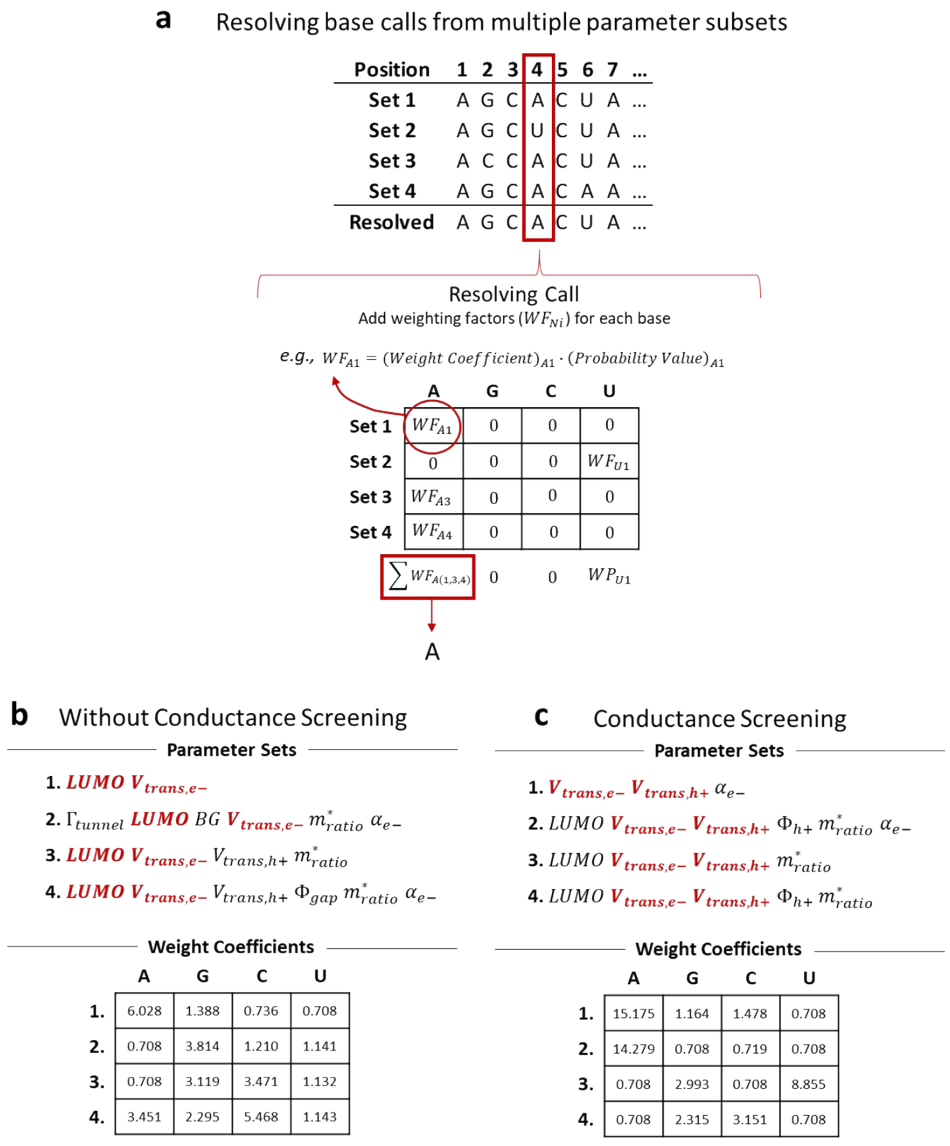

Figure S5— Base calling example and optimal parameter sets. (a) Resolving a single base call from each of the four base calls of the parameter subsets. The four subsets of parameters (labeled Set 1-4) each make a base call for every position in an unknown RNA sequence. To resolve the calls into a single call (shown as an example for Position 4), the call made by each parameter subset is given a weighting factor calculated as the product of the weight coefficient for ribonucleotide rN and the probability value output from the standard naïve Bayes classifier for ribonucleotide rN (where N is A, G, C, or U). The weighting factors for each ribonucleotide are summed, and the largest determines the resolved base call. Specific parameter subsets (with fundamental parameters highlighted in red) and weight coefficients are shown for the modified naïve Bayes (b) without conductance screening and (c) with conductance screening.

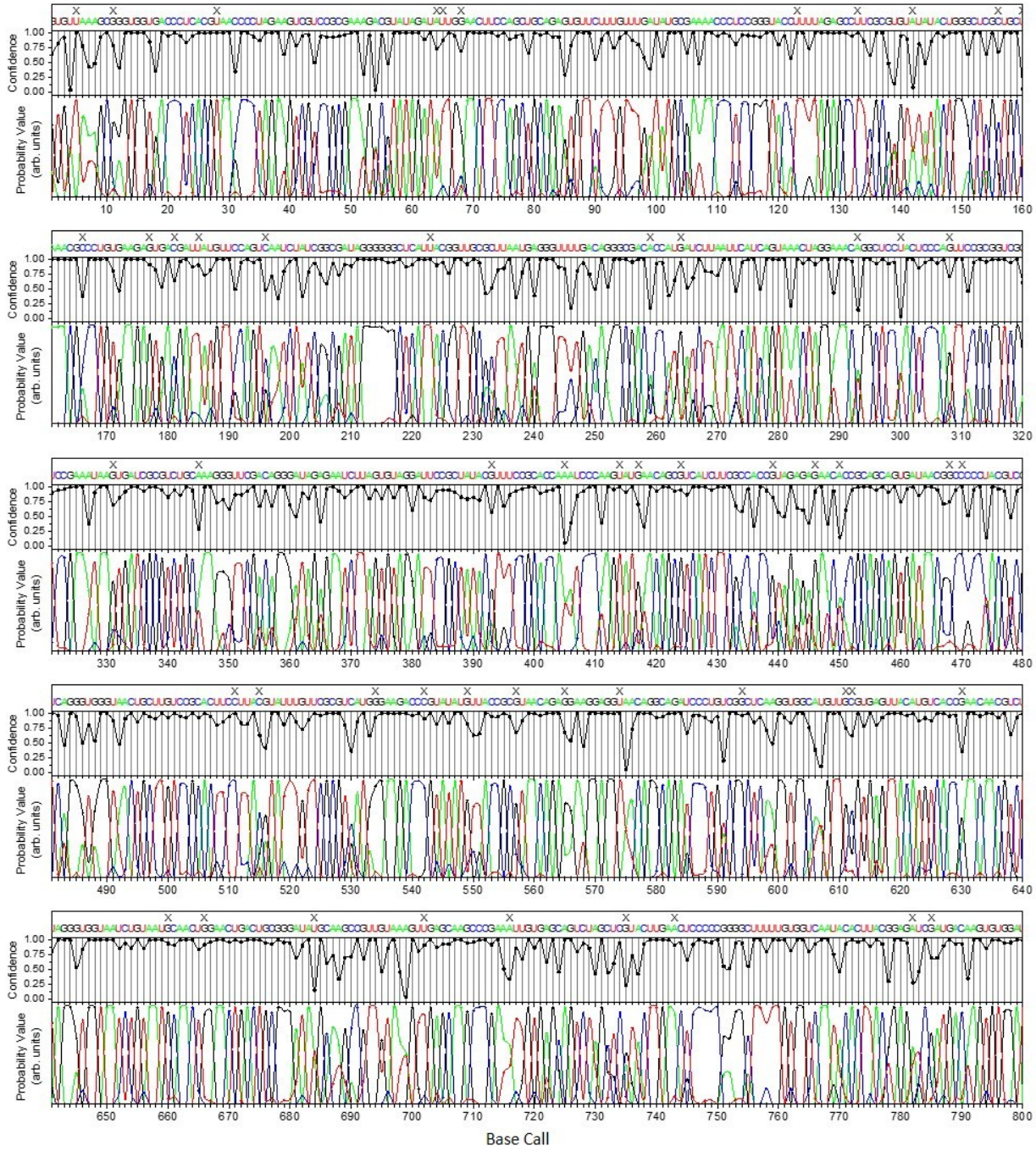

Figure S6— Detailed base calling output at 35X coverage for no conductance screening. Probability values (obtained from the base calling algorithm), confidence of base calling, and accuracy (X indicates incorrect calls) for the complete set of 800 base calls at 35X coverage (a small set of 50 base calls was shown in Figure 4a, main text).

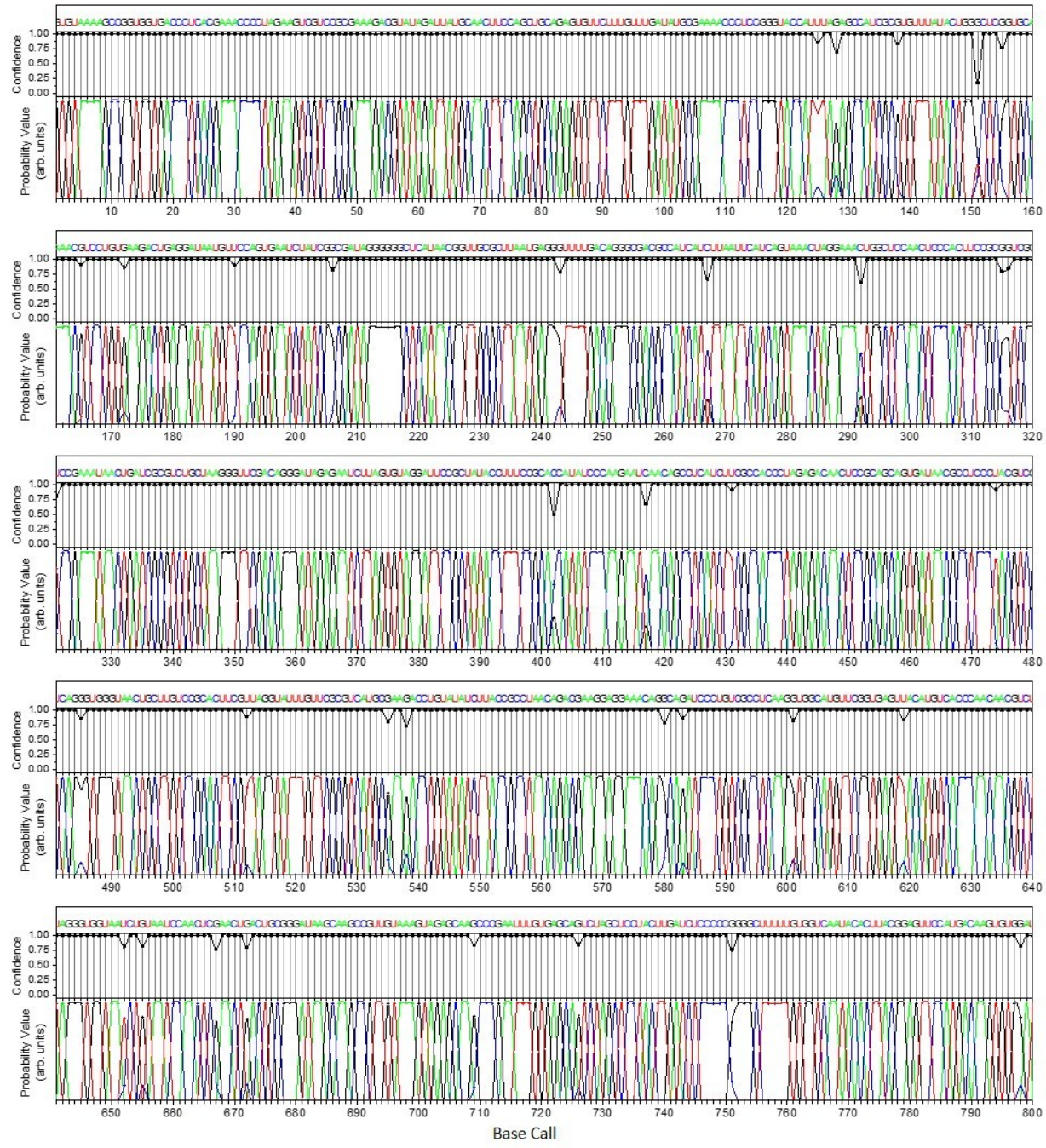

Figure S7— Detailed base calling output at 35X coverage for low-conductance screening. Probability values (obtained from the base calling algorithm), confidence of base calling, and accuracy (X indicates incorrect calls) for the complete set of 800 base calls at 35X coverage (a small set of 50 base calls was shown in Figure 4b, main text).

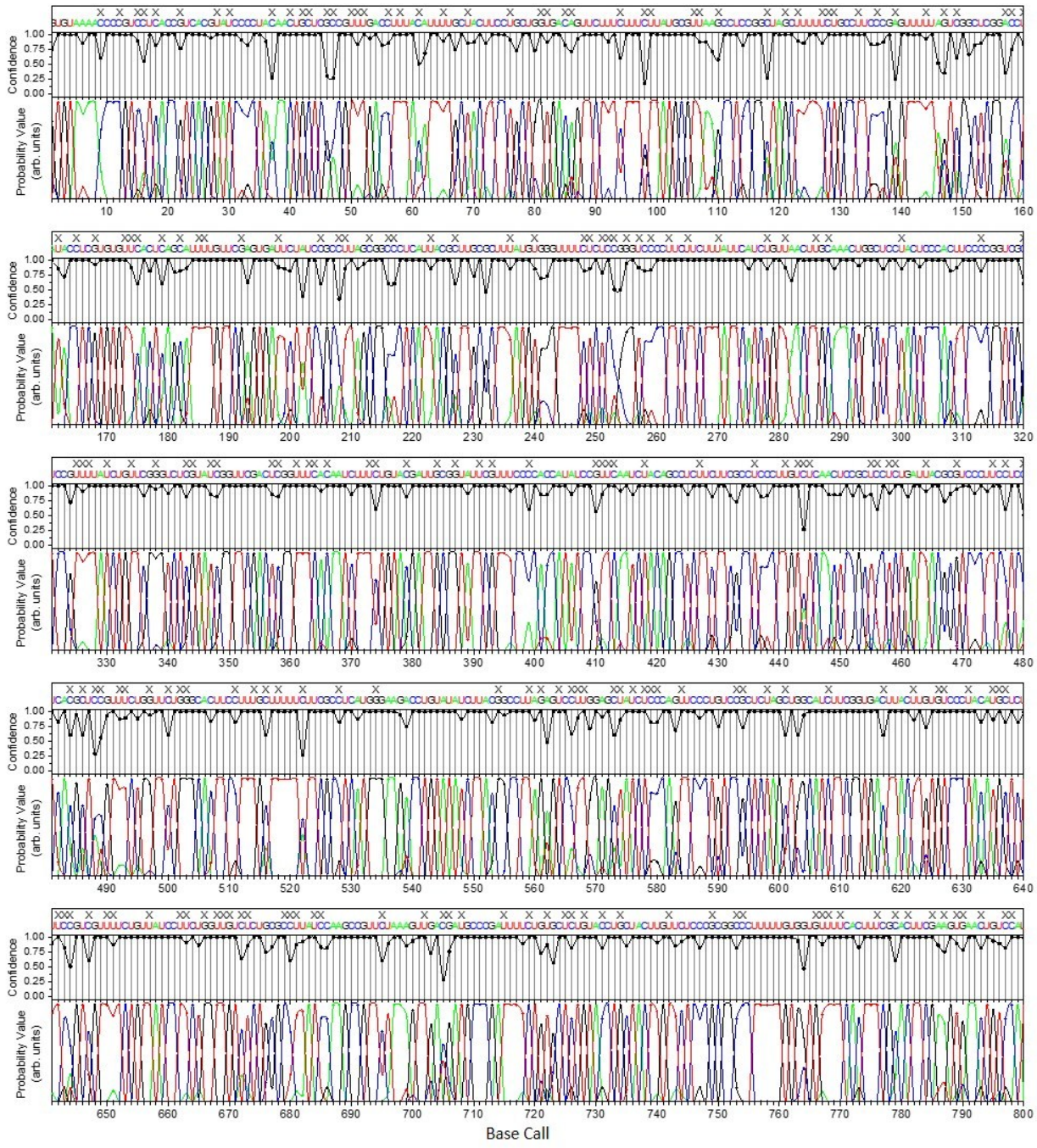

Figure S8— Detailed base calling output at 35X coverage for high-conductance screening. Probability values (obtained from the base calling algorithm), confidence of base calling, and accuracy (X indicates incorrect calls) for the complete set of 800 base calls at 35X coverage (a small set of 50 base calls was shown in Figure 4c, main text).

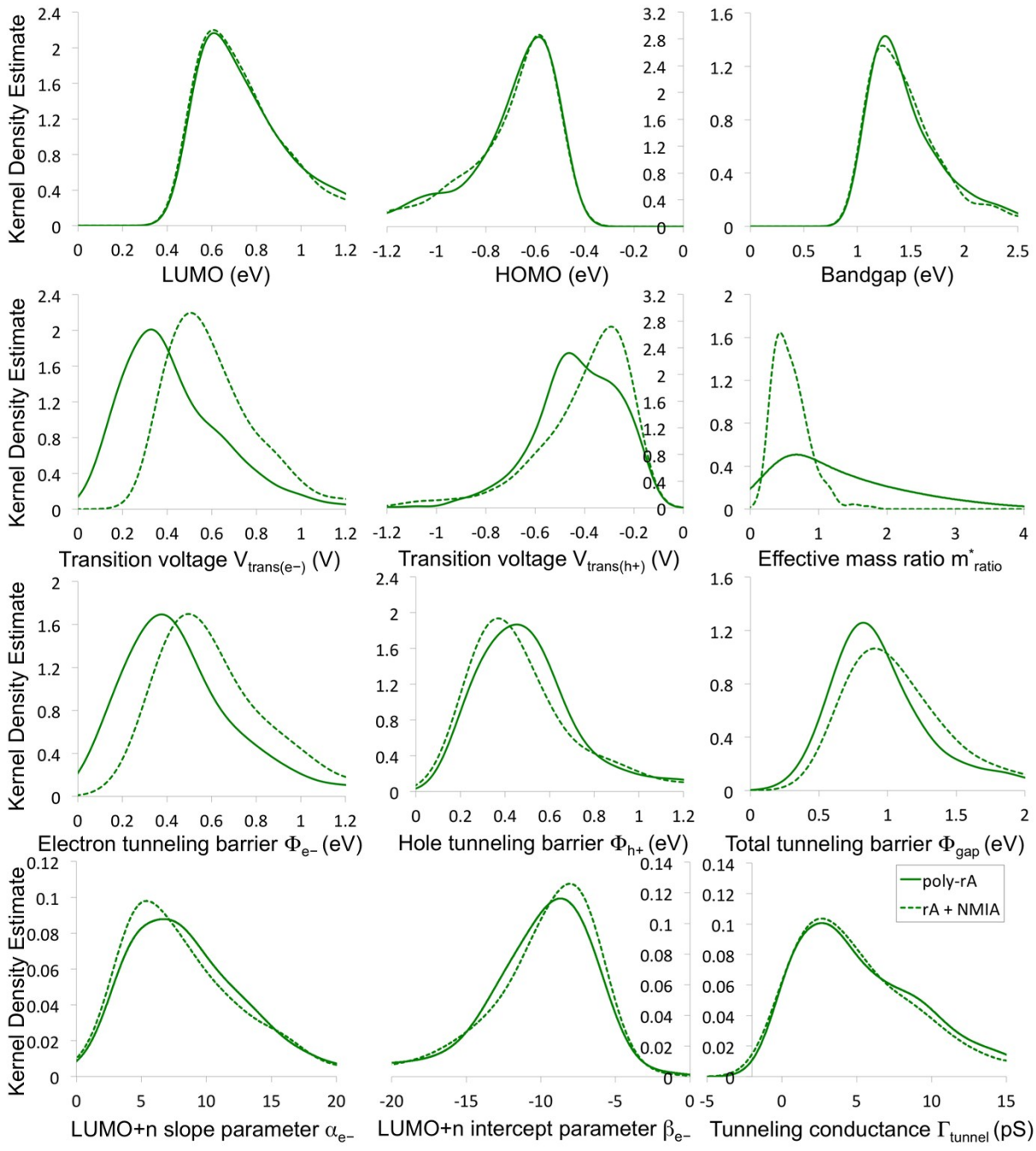

Figure S9— Kernel density estimates of all 12 parameters for rA ± NMIA. Data was collected on both unmodified and NMIA-modified poly-$(rA)_7$ RNA at low surface-coverage on MPA. Each curve represents the kernel density estimate for ~500 STS measurements.

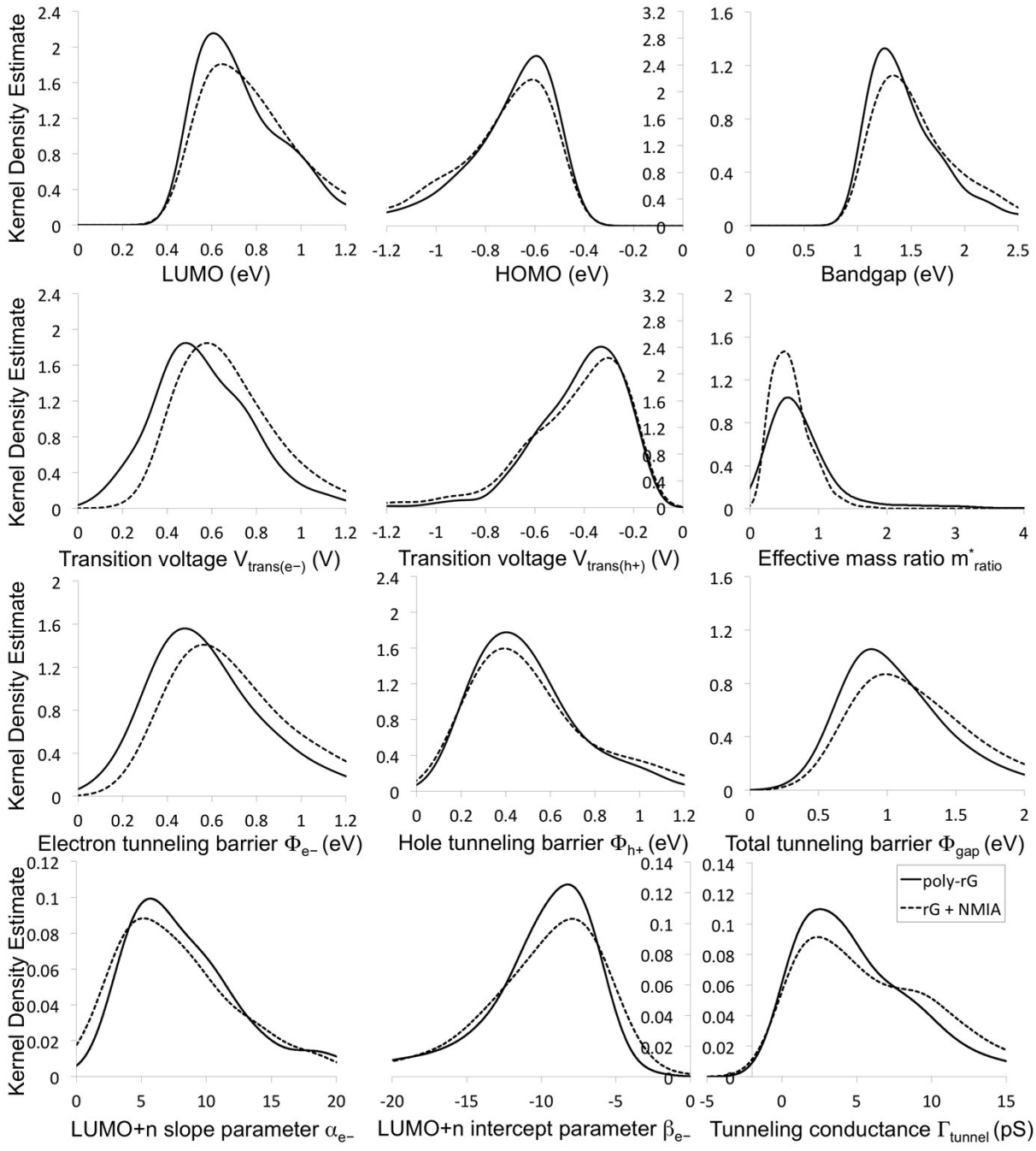

Figure S10— Kernel density estimates of all 12 parameters for rG ± NMIA. Data was collected on both unmodified and NMIA-modified poly-$(rG)_7$ RNA at low surface-coverage on MPA. Each curve represents the kernel density estimate for ~500 STS measurements.

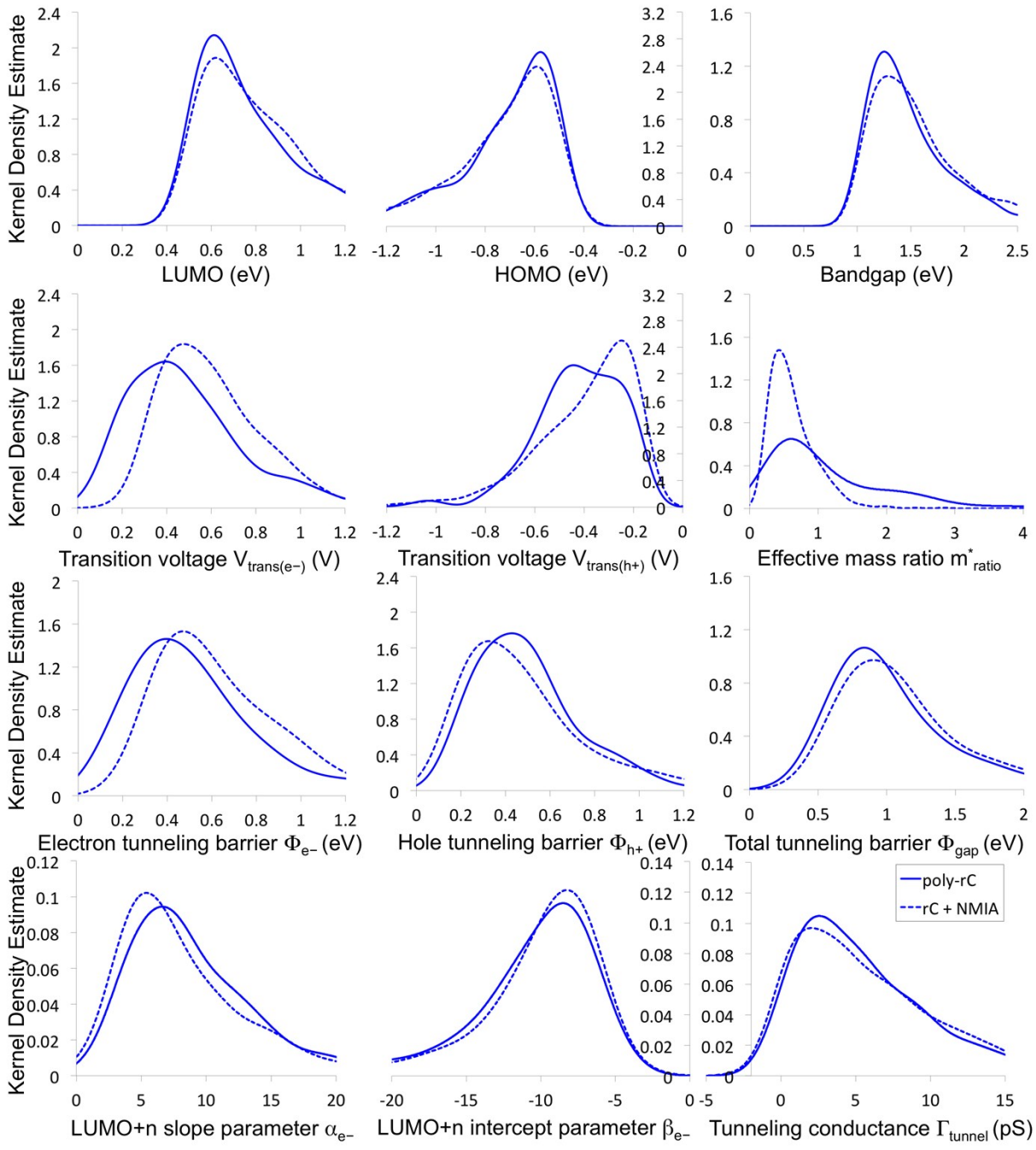

Figure S11— Kernel density estimates of all 12 parameters for rC ± NMIA. Data was collected on both unmodified and NMIA-modified poly-$(rC)_7$ RNA at low surface-coverage on MPA. Each curve represents the kernel density estimate for ~500 STS measurements.

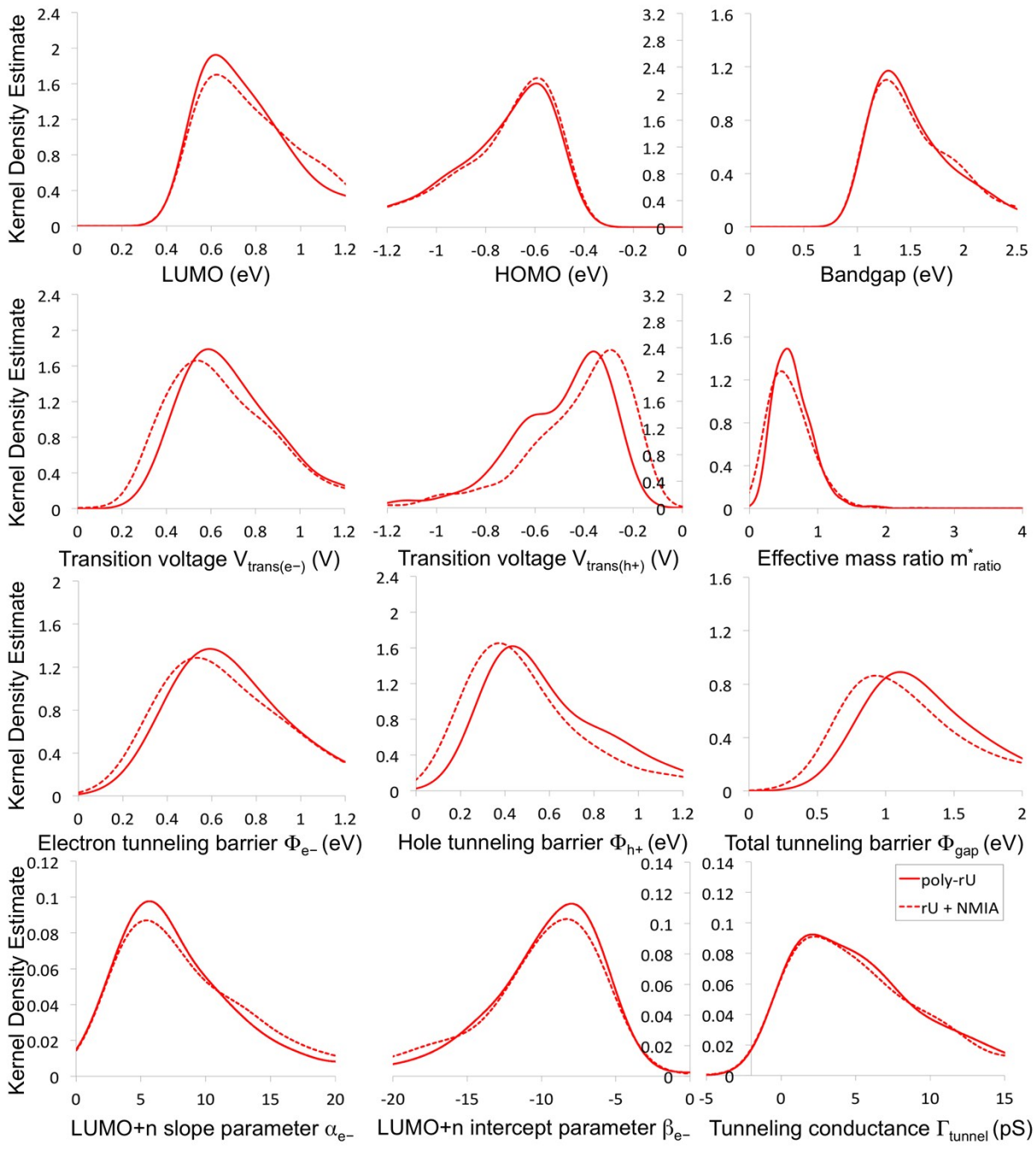

Figure S12— Kernel density estimates of all 12 parameters for rU ± NMIA. Data was collected on both unmodified and NMIA-modified poly-$(rU)_7$ RNA at low surface-coverage on MPA. Each curve represents the kernel density estimate for ~500 STS measurements.