Single-Molecule RNA Sequencing for Simultaneous Detection of m6A and 5mC

Takahito Ohshiro[1†], Masamitsu Konno[2†], Ayumu Asai[2,3†], Yuki Komoto[2], Akira Yamagata[2,4], Yuichiro Doki[5], Hidetoshi Eguchi[5], Ken Ofusa[2,4], Masateru Taniguchi[1*], Hideshi Ishii[2*]

Affiliations

1. *The Institute of Science and Industrial Research, Osaka University, 8-1 Mihogaoka, Ibaraki, Osaka, 567-0047, Japan*
2. *Center of Medical Innovation and Translational Research, Graduate School of Medicine, Osaka University, 2-2 Yamadaoka, Suita, Osaka, 565-0871, Japan*
3. *Artificial Intelligence Research Center, The Institute of Scientific and Industrial Research, Osaka University, 8-1 Mihogaoka, Ibaraki, Osaka, 567-0047, Japan*
4. *Prophoenix Division, Food and Life-Science Laboratory, Idea Consultants, Inc., 1-24-22 Nanko-kita, Suminoe-ku, Osaka-city, Osaka, 559-8519, Japan*
5. *Gastroenterological Surgery, Department of Surgery, Graduate School of Medicine, Osaka University, 2-2 Yamadaoka, Suita, Osaka, 565-0871, Japan*

6.

Table of Contents

S1. Picking and analysis

The signal rise time is considered to be the start of the signal from the estimated baseline until it reaches a value exceeding six times the noise level as the threshold value. The start of the signal is when it exceeds one times the noise level. Next, the signal is determined to have ended when a data point below one times the noise level appears from the estimated baseline after the signal starts. The baseline is defined as the mode of the histogram for the last 2000 data points at each time point. Noise is defined as the mode of the standard deviation of 200 data points obtained by dividing the latest 10,000 data points at each time point into 500 equal parts.

S2. Mononucleotide signal detection

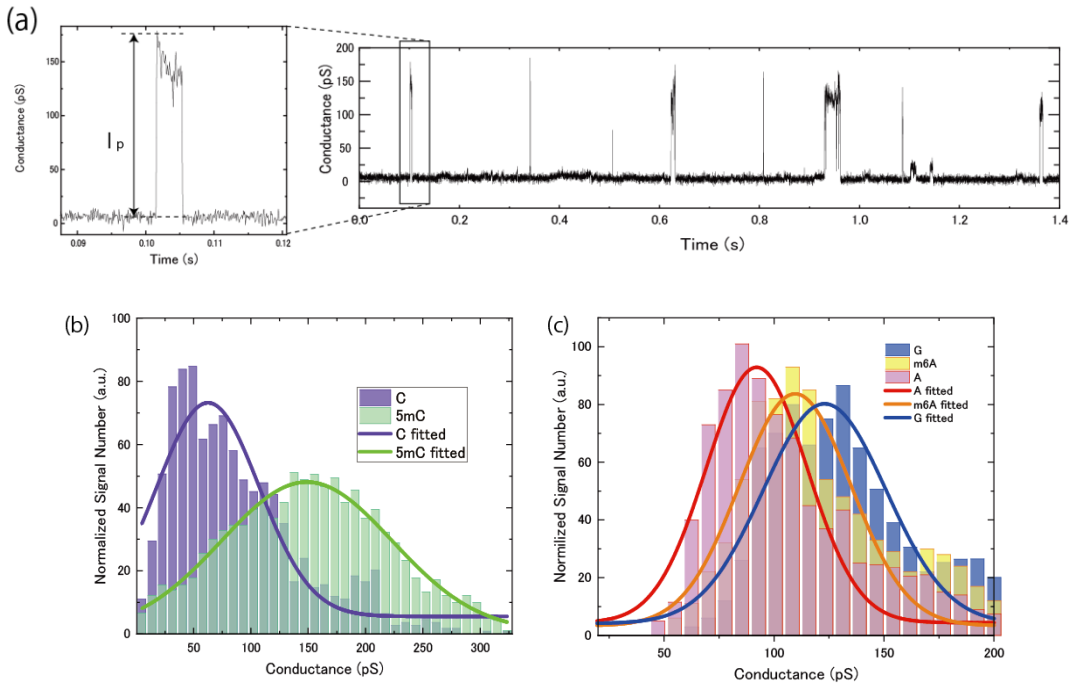

Figure S1. Mononucleotide signal detection. Single-molecule electrical signals for m6A and 5mC are measured by using a 0.64-nm gap electrode, which was tuned by a nano-fabricated mechanically controllable break junction (nano-MCBJ). Figure S1a shows typical *I*-*t* profiles for 5mC in aqueous solution. The observed electrical signals were characterized by I_p, which were defined as the maximum current and the duration of the current, respectively. Single-molecule conductance (I_p/V) histograms were then constructed from the I_p data using approximately 1000 signals for each molecule and were analyzed using Gaussian fit curves. We defined the peak values of the Gaussian fit curves as the single-nucleotide conductance. The single-nucleotide conductances were found to be 111 pS for m6A compared with 92 pS for adenosine (Figure S1c). Similarly, the conductance of 5mC was found to be 149 pS compared with 64 pS for cytidine (Figure S1b). The conductance values determined by this single-molecule electrical detection are closely related to characteristic energy levels, particularly the HOMO energy level, which is calculated by density functional theory[20,23]. The conductance differences caused by methylation for the pyrimidine, i.e., C and 5mC, and purine, i.e., A and m6A, are due differences in the π-conjugation induced by the electron-donating character of methyl substitutions. Together with these data, the order of conductance values order was found to be the following: 5mC (149 pS)>G (123 pS)>m6A (111 pS)>A (92 pS)>C (64 pS)>rUMP (50 pS). The values relative to G, which are the values normalized to the G

conductance value (123 pS) were ordered as follows: 5mC (1.21)>rGMP (= 1)>m6A (0.90)>rAMP (0.77)>rCMP (0.52)>rUMP (0.41) (Table 1).

S3. Base calling

The base sequences from the conductance profiles were determined by a Phred base-calling method, which is widely used for conventional genome sequencer analysis[S1,S2]. In the case of the conventional sequencer, base calling is based on the optical intensity for each wavelength of the base-attached optical probes. In our case, base calling was performed based on the tunnel-current intensity; the detailed procedure is described in previous reports[22] and is only briefly described here.

In the first step, the data histograms were derived from the conductance–time profiles, and each peak value of the conductance (μ) and standard deviation (σ) for each base is determined by the Gaussian fitting of the datapoint histogram. In each conductance-profile histogram, there are several peaks. Among them, the minimum and maximum values of the peaks correspond to the relative conductance values of baseline and G, respectively (Table 1). In the second step, the peak conductance and the standard deviation values were used to calculate each base probability and each base element. In the third step, base species were sequentially assigned in the time profiles based on the maximum probability values of each base type or the baseline. For this assignment, the integral value of the base probability for each length of time was calculated to reduce the abrupt signal change, which was mainly due to electrical noise. In this study, we set the time length for integration to 0.3 msec, which is comparable to the experimentally determined minimum retention time around the electrodes. The maximum integral value of each base-probability value was sequentially assigned to the base type or baseline for each of the time regions. For each of the assigned bases, the accuracy of the base assignment was quantified by the *Q* score[S2]. Signals with low *Q* values were due to the large conductance distributions and their overlap. To ensure the accuracy of read-sequence signals, only read-sequence signals with over six *Q* scores (P>75%) were used for subsequent resequencing.

S4. Signal assembly and conductance plotting/mapping

Signal assembly was performed based on tunnel-current intensity, and the detailed procedure has been described previously[22]; the procedure is only briefly described here. A sequence position number was assigned for each read base in the "long-read signals" and compared to the miRNA-200c-5p reference sequence; over six right-read base signals were described as "long-right read signals" and used for assembly and plotting. The read direction was sometimes found to be changed in the terminal position, and therefore a "duplicated" read was observed. In the assembly process, these duplicated-read regions were automatically deleted after sequence-position assignment, and the straight-right read regions were used. For the assembly, we used the miRNA-200c-5p reference sequence and all the methylated derivatives from all the signals, and the resulting conductance profiles of all the long-read signals are shown in Figs. 2 a–c.

The signal count of methylated and non-methylated bases for each base position were performed as follows. In the case of position #4 in miRNA-200c-5p (Fig. 3a, first column), we used the miRNA-200c-5p reference sequence and all the methylated derivatives determined from the obtained signals; of these, signals containing cytidine (C) and methylated cytidine (5mC) in the #4 position of miRNA-200c-5p were determined from the captured miRNA sample solution. From the signals for the captured miRNA-200c-5p, we extracted all the over six right-read base signals containing cytidine (C) and methylated cytidine (5mC) in the #4 position of miRNA 200c-5p, and the total signal number was found to be 1606. The number of signals for 5mC and C were 10 and 1596, respectively, so that the methylation rate was 0.6% (10/1606). We used a similar method for determining the methylated signal number and its methylation rate for all the potential methylated positions (Figs. 3a and 3b and Figs. 4a and 4b).

S5. Methylation detection of MALDI–TOF mass spectrum

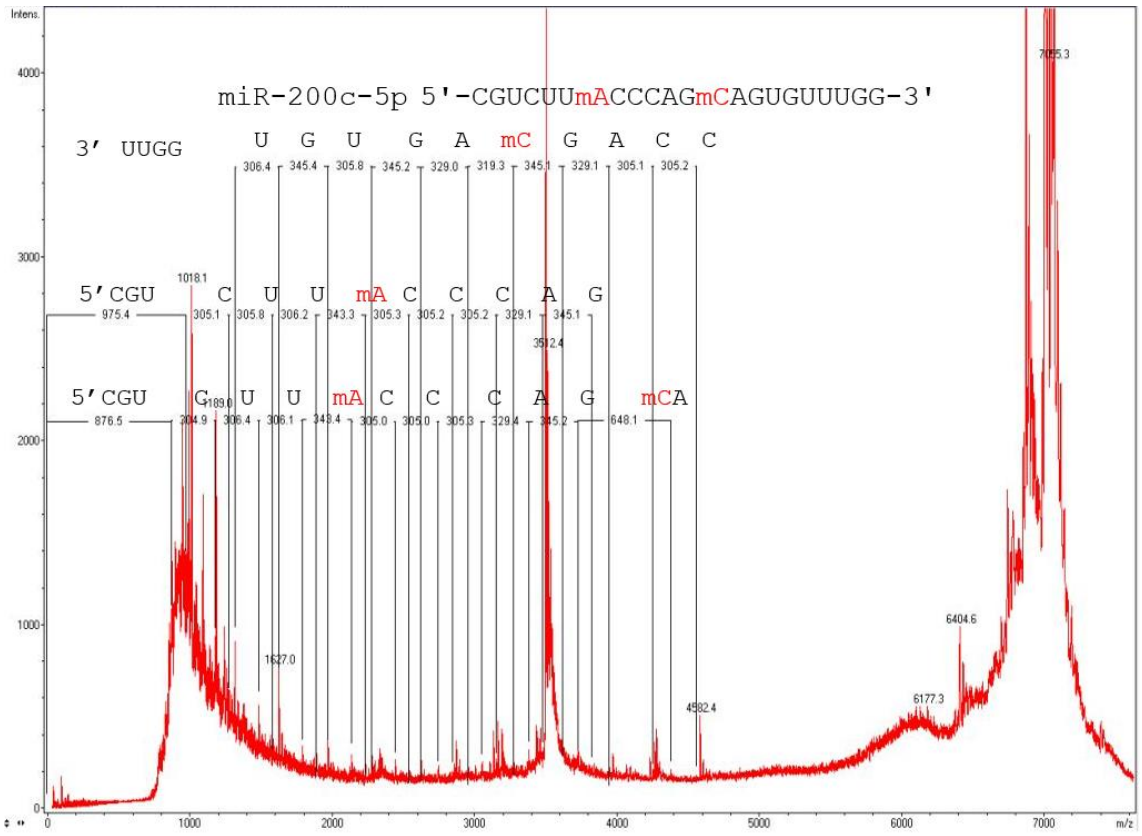

Figure S2. Methylation detection of MALDI–TOF–MS/MS spectrum of miR-200c-5p. Fragment ions from base 4-9, 4-10, and 18-9 are shown. The adenosine at position 7 and the cytidine at position 13 show an additional peak (+14 m/z), indicating methylation (m6A and 5mC).

Supporting information references

[S1]. Ewing, B. & Green, P. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res.* **8**, 186–194 (1998).

[S2]. Ewing, B., Hillier, L., Wendl, M. C. & Green, P. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res.* **8**, 175–185 (1998).