Electronic Supplementary Material (ESI) for ChemComm.

Supplementary information

for

Detection of alcohol-derived cancer marker by single-molecule quantum sequencing

Y. Komoto, T. Ohshiro, & M. Taniguchi*

Correspondence to: taniguti@sanken.osaka-u.ac.jp

This PDF file includes:

SI1. Methods

Sample preparation

2′-Deoxyguanosine monohydrate 99-100% (Sigma Aldrich) and N^2-Ethyl-2′-deoxyguanosine ≥98% (Sigma Aldrich) were dissolved in milliQ water to prepare a 1 μM solution. The concentration of the mixed solution was dG, N2-Et-dG, 1 μM: 3 μM and 3 μM: 1 μM, respectively.

Device fabrication

Schematic strructure of SMQS device is shown in Figure 1c. Polyimide film was coated as an insulating layer onto thin-silicon substrate with spin coating. A gold nanowire was drawn on the silicon substrate by electron-beam lithography. Then, SiO_2 film was coated on the gold nanowire by chemical vapor deposition. Narrowest part of nanowire is several ten nm. Polyimide layer under gold nanowire were removed by dry-etching to form free-standing gold-wire. SEM image of the substrate is shown in Figure 1d. For the electrical measurement of the aqueous solution, a solution holding cover made of poly dimethyl siloxane (PDMS) was attached on the substrate. PDMS was bonded onto the substrate after the hydrophilization by vacuum plasma. The PDMS cover has a microchannel that connects the hole for introducing the sample solution and the Au narrow wire.

Electrical measurements

Single-molecule conductance measurements were performed at the optimal gap distance of the nanogap electrodes as described previously in reference 18-20 in main text. The solutions were injected into the microchannel drawn on the PDMS. The Au nanowire sensing part was

immersed with the solutions. The diffused nucleobases are detected as a pulse signal as they pass through the nanogap. A lithographically fabricated gold nanowire on a thin silicone substrate was broken by mechanically bending the substrate under ambient condition with application of a bias voltage of 100 mV, and the single detection part of the nanogap electrodes was formed. Throughout the junction breaking process, the junction conductance (G) was monitored using a picoammeter (Keithley 6487). A series of conductance jumps of the order of $G_0 = 2e^2/h$ (where e and h are the elementary charge and Planck's constant, respectively) was observed, and the final conductance was 1 G_0. Several seconds after reaching the 1 G_0 state, a gold atomic junction naturally ruptured in the nanowire, creating a nanogap. After the nanogap formed, current amplifier is switched from picoammeter (Keithley 6487) to homemade current amplifier to detect picoamperes-scale current signals. The measurement time for each gap distance is 5 min. The current profile was recorded with homemade current amplifier (10^9 V/A) and Digital Multimeter (National Instruments, PXIe-4081). Sampling rate was 10 kHz. The bias voltage was applied from battery source. Schematic view of the setup is drawn in Figure S1. The gap size was controlled using the piezo bias voltage every 1 sec as the current adjust as target value in feedback loop during the measurement. The target gap width was 0.58, 0.6 and 0.62 nm. The gap distance was estimated from the baseline tunnelling current as following section.

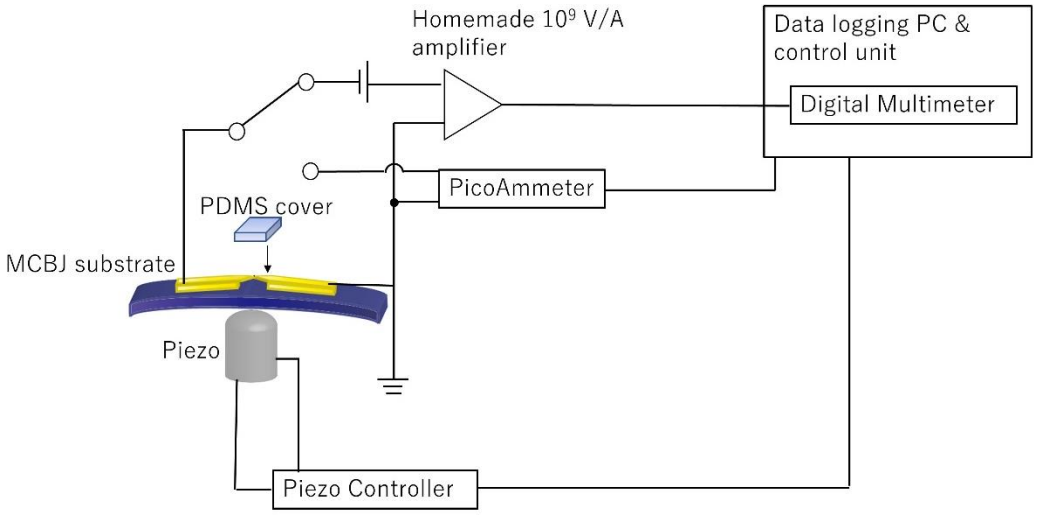

Figure S1. Schematic illustration of the setup

Estimation of gap distance

The gap distance is estimated using by following current equation of direct tunneling current

$$I = const \exp\left(-\frac{4\pi}{h}\sqrt{2mwl}\right).$$

Here, h, m,w, and l represents plank constant, electron mass, work function of gold electrode, gap distance. We used electron mass of 9.1×10^{-31} kg as m, and work function of Au(111) as w. Effective mass and work function of gold nanogap not (111) surface should be used for accurate estimation. Furthermore, the inelastic gold gap broadening just after breaking atomic junction is not under consideration. Hence, the experimental gap length is larger than the target width of 0.58, 0.6 and 0.62 nm .

Theoretical Calculation

DFT calculations of N^2-Et-dG and dG isolated molecules were calculated using Gaussian 09. The basis set was B3LYP/6-31G (d, p).

SI2. Detail of ML analysis

At First, we perform Positive and Unlabelled data Classification (PUC) to remove noise signals observed in blank solution. PUC is appropriate algorithm for noise removal. The PUC algorithm is Elkan and Noto's method[5]. Inner classifer is gaussian naïve Bayes in scikit-learn library version 0.21.1.

After the noise removal, 534 signals were trained and classified with supervised ML of XGBoost classifier. A 10-fold cross-validation was performed and its average and standard deviation values provided the classification ratios and errors. The errors are standard deviation of 10-time classification.

In prediction of mixture solutions, The XGBoost supervised machine learning classifier were trained with dG and N^2-Et-dG signals from each pure solution after PUC noise removal. The signals from the mixtures were classified one by one with the trained classifier after PUC noise removal.

The analysis was performed using Python 3.6 with the XGBoost library.

SI3. Estimation of classification accuracy for accumulated signals

The classification performance index (*F*-measure) is 0.77. This accuracy is not the accuracy determined by using multiple signals during application but only that for a single pulse. The classification accuracy can be improved by statistical analysis. In the method reported in this manuscript, each signal is classified one by one; the molecule is classified with majority vote of all signal classification results.

Here, we consider the relation between the classification accuracy and the number of signals *n*. The prediction ratio for a single pulse of the true molecule *p* is set to 0.77. Then, the probability of accurate prediction by the majority vote *P* is described using the following equation:

$$P = \sum_{k>n/2} {}_nC_k p^k (1-p)^{n-k}. \quad (1)$$

where *k* denotes the number of true signals. The relation between *P* and *n* is shown in Figure S2. The accuracy determined by the majority vote is over 95% for 7 signals, 99% for 17 signals, and 99.9% for 29 signals.

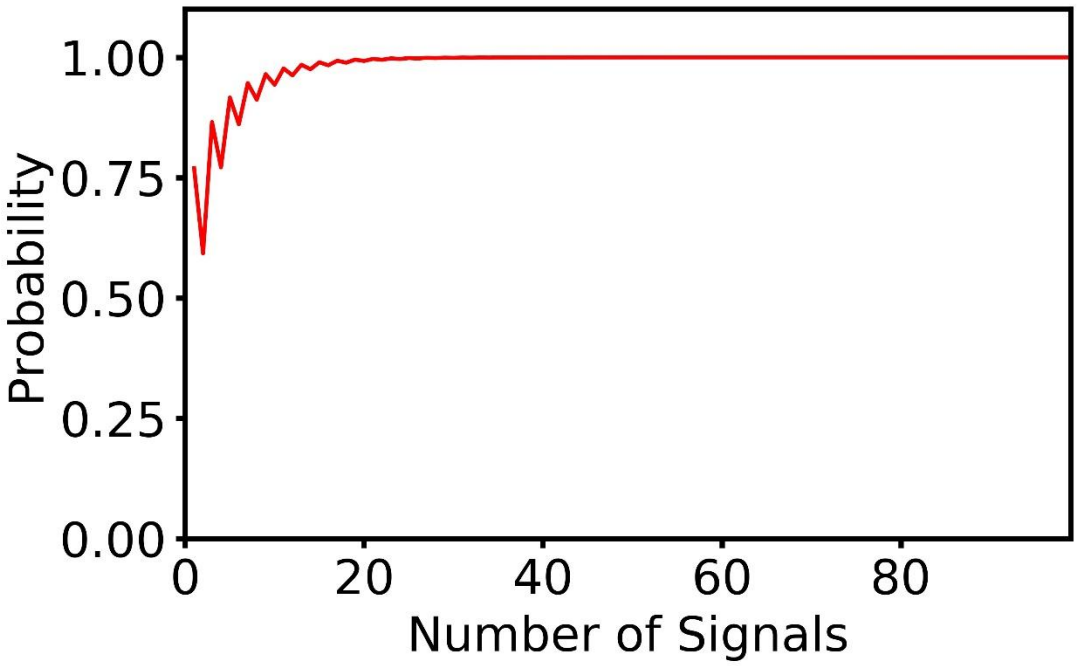

Figure S2. Relation between probability of accurate prediction by majority vote and number of signals.