Supplementary Information

Fast DNA sequencing with a graphene-based nanochannel device

Seung Kyu Min, Woo Youn Kim, Yeonchoo Cho & Kwang S. Kim

S1. Transmission curves for tilted geometries from the π-stacked structures

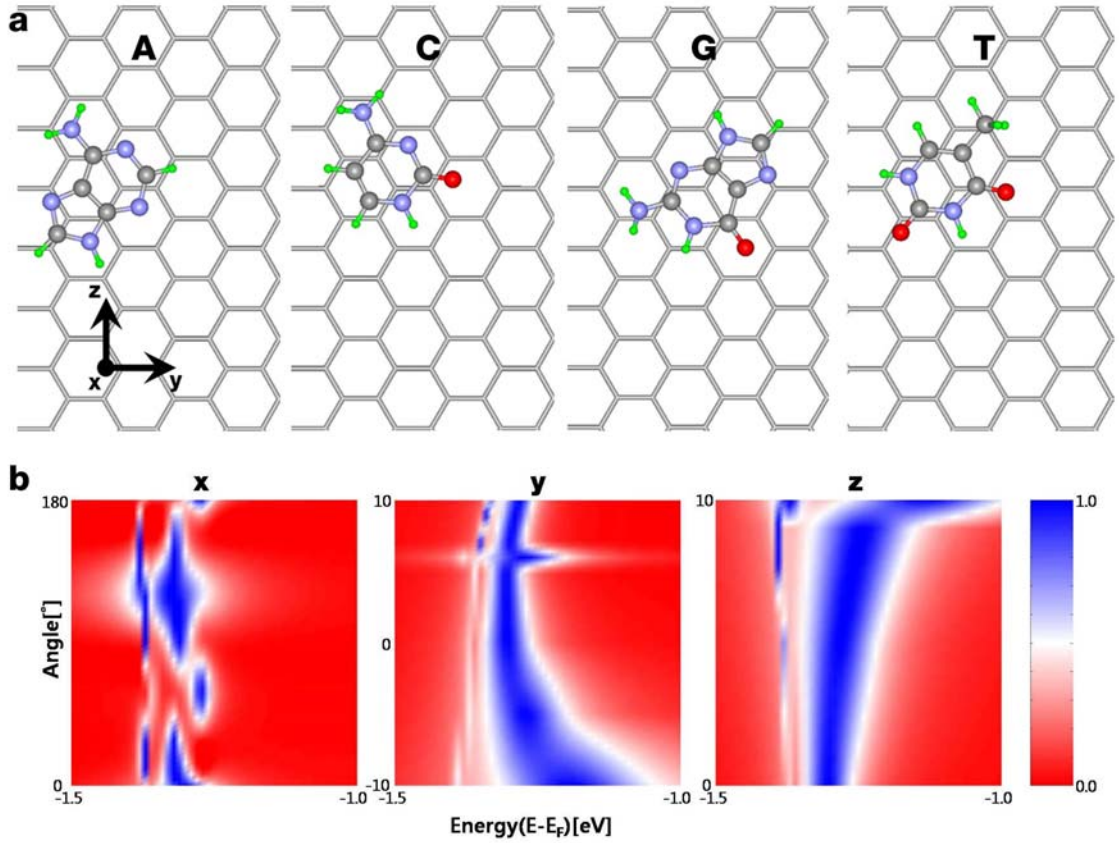

Figure S1| a, A/C/G/T-AGNR most populated structure from the NAMD simulations. **b,** transmission spectra for the rotation of Ade along each axis.

In practical situations or simulations, DNA bases are slightly off the minimum energy geometries due to the backbone strains and fluctuations arising from the temperature. Thus, we calculate the transmission spectra for slightly tilted geometries of Ade from a graphene nanoribbon (GNR) with respect to the minimum energy geometry (Figure S1). Rotation along the stacked axis, x-axis, does not show significant changes in the position of sharp drops in the transmission curves. For tilted geometries, rotated along the z-axis, the positions of sharp drops remain almost the same within the change of 10˚ in tilting angle. The 10˚ tilted geometry is a few fractions of 1 kcal/mol unstable compared to the equilibrium geometry. Since the geometries deviate from the equilibrium geometry within a few degrees at room temperature, statistical measurements give a certain distribution of currents or transmission curves.

S2. Molecular dynamics simulations

We used the CHARMM parameter set for nucleobases (Foloppe, N. & MacKerell, Jr., A.D. All-Atom Empirical Force Field for Nucleic Acids: I. Parameter Optimization Based on Small Molecule and Condensed Phase Macromolecular Target Data. J. Comput. Chem. 21, 86-104 (2000)). For the GNR device, we used the CHARMM parameters of a benzene molecule, while the charges for non-edge carbons are set to zero. The GNR device is fixed during the simulation. For silicon nitride, the parameters are set to describe the experimental dielectric constants. (This is well described in the NAMD tutorial: http://www.ks.uiuc.edu/Training/SumSchool/materials/sources/tutorials/02-namd-tutorial/namd-tutorial-html/)

To expedite the highly computationally-intensive simulations, the applied force of 30 kcal/mol/Å (~0.25 nN/base) is given much stronger than that to be imposed in the real system merely for the sake of computational efficiency, which nevertheless, provides a favourable outcome to mimic a realistic applied electric field. In the simulations, the vertical distances for planar π-stacking (face-to-face conformation) between bases and graphene are around 3.6 Å. Because of such fast translocation, in the period of transit for which each base interacts well with the graphene (corresponding to ~50% total transit time), each base shows some changes from the optimal geometry except for the special Cyt case near the terminal of the ssDNA strand (for which the tilt angle is up to 20˚). The realistic dynamics with much weaker force should give highly stable π-stacked structures for nearly the whole transit time.

S3. 2D transient autocorrelation function (TACF)

From the TACF equation in the text, $C(t,t_0;\tau=1)$ can be interpreted as an autocorrelation function with a time domain from t_0 to $t_0+\tau$. In the usual autocorrelation function, τ is infinite and the function is not dependent on t_0 and so is often set to 0 or the function is simply denoted as $C(t)$. In DNA sequencing, the transmission spectra varies depending on the time as the ssDNA passes through the GNR. Thus, we are less interested in the stationary spectra over the whole time but the transient spectra for a small duration whose range is small enough to extract the information of each single base interacting with the GNR.

Let us consider an exemplary time-dependent transmission function $J(E,t)$:

$$J(E,t) = z(E-E_1)\left[\theta(t-t_{1i})-\theta(t-t_{1f})\right]+z(E-E_2)\left[\theta(t-t_{2i})-\theta(t-t_{2f})\right]$$

where

$$z(x) = \frac{1}{\sqrt{2\pi}}\exp(ix).$$

Here, $t_{1i} < t_{2i} < t_{1f} < t_{2f}$ and $\theta(t)$ is a Heaviside step function (Figs. S2a and S2b). t_i and t_f describe the time when the interaction between a base and GNR begins and ends, respectively. For simplicity, we assume that each base has its characteristic peak at E_i which does not depend on its orientations in the transmission function. Then, we can divide $J(E,t)$ into three time domains:

$$J(E,t) = \begin{cases} z(E-E_1) & DOMAIN\ 1: t_{1i} < t < t_{2i} \\ z(E-E_1)+z(E-E_2) & DOMAIN\ 2: t_{2i} < t < t_{1f} \\ z(E-E_2) & DOMAIN\ 3: t_{1f} < t < t_{2f} \end{cases}$$

Then, the inner product between $J(E,t_1)$ and $J(E,t_2)$, $\int dE\ J(E,t_1)J(E,t_2)$, depends on the location of t_1 and t_2 (see the following table).

t_1 \ t_2	DOMAIN 1	DOMAIN 2	DOMAIN 3
DOMAIN 1	1	1	0
DOMAIN 2	1	2	1
DOMAIN 3	0	1	1

Fig. S2a shows a case that TACF is equal to 1. For the whole time integration in the TACF equation, both t_1 and t_2 are located in DOMAIN 1. However, in Fig. S2b, the TACF is less than 1 since the time integration contains zero-correlation region. At a given t_0 with a very small t, the non-unity region begins after $t = t_f - t_0$, which makes triangular patterns. In Figs. S2c and S2d,

we calculated 2-dimensional TACF from a model case of the 5'-GACT-3' sequence where [t_i, t_f] for G, A, C, and T are [0.0, 3.0], [2.5, 5.2], [4.5, 8.0], and [6.8, 10.0], respectively. It shows distinguishable triangular patterns when two bases are interacting with GNR.

The size of τ can be chosen as an optimal value (Fig. S3). If τ is larger than the transit time of a single base, the resulting information is extracted for multiple bases, giving indistinguishable data. For most cases, the optimal τ should not be larger than the minimal value of the single-base transit time. Since the minimal value of the single-base transit time is ~0.1 ns in our simulations (Fig. 3e in manuscript), τ=0.05 ns gives the best result (Fig. S3a).

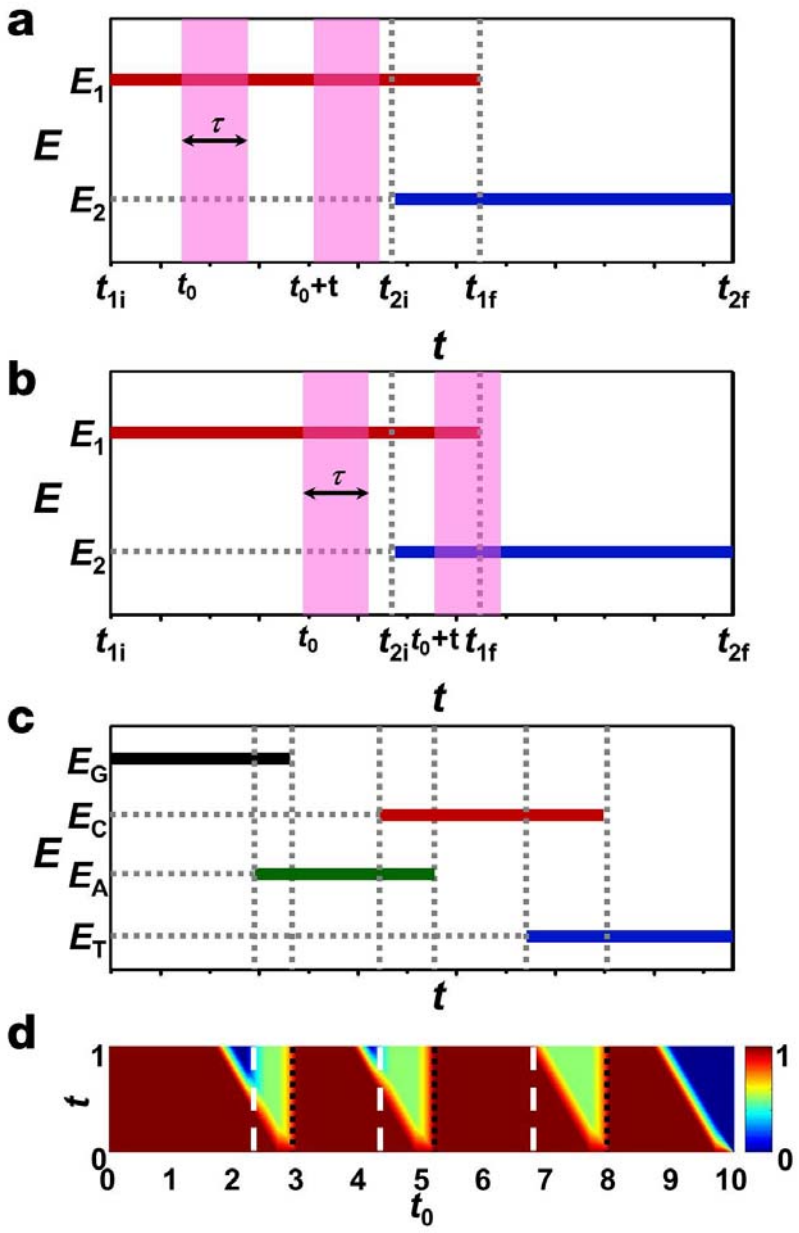

Figure S2| a, b, time-dependent transmission function J(E,t) for two time-dependent signals at E1 (red) and E2 (blue) with time integration range (pink) in correlation calculation. A case that TACF is 1 **(a)** and less than 1 **(b)**. **c.** A time-dependent transmission function J(E,t) for a model sequence, 5'-GACT-3'. For Gua, [t_i, t_f] = [0.0, 3.0]. Likewise,for Ade, Cyt, and Thy, [t_i, t_f] = [2.5, 5.2], [4.5, 8.0], and [6.0, 10.0], respectively. **d,** 2-dimensional TACF $C(t,t_0;\tau=0.3)$ for this model case. A black dotted line indicates the time that a DNA base passes out, while a white dashed line indicates the time that the following DNA base comes in.

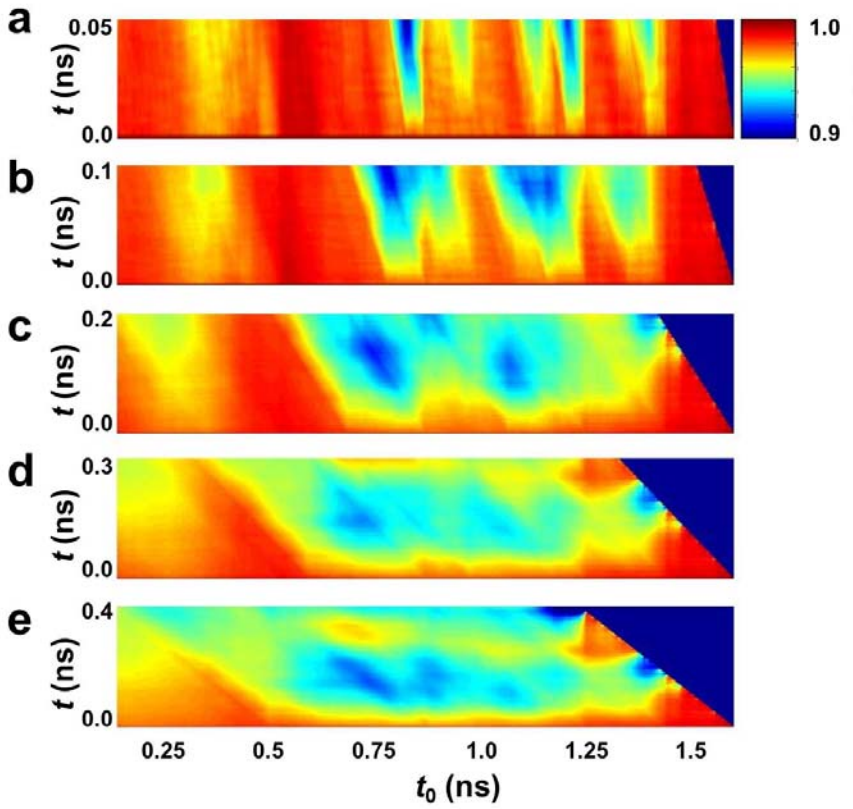

Figure S3| Dependence of 2-dimensional TACF on τ from DNA sequencing data in the text, **a,** τ=0.05 ns **b,** τ=0.1 ns **c,** τ=0.2 ns **d,** τ=0.3 ns and **e,** τ=0.4 ns.

S4. Interpretation of Fig. 3 (text)

It is not easy to identify each base from Fig. 3a in the text. As the test DNA strand 5'-GCATCGCT-3' enters the GNR device that holds only one base at once, the forefront base Gua begins to stack on the GNR, resulting in appearance of the peaks around E-E_F=-0.65 and -1.65 eV. Subsequently, the second base Cyt approaches to the GNR, while the Gua escapes. Thus, a new pattern appears at E-E_F = 1.8 and -1.2 eV, while the original pattern for the Gua fades away. Likewise, other bases have their own specific patterns (Ade: -1.3, and Thy: -1.6 eV). Though the patterns of Ade and Cyt (Gua and Thy) overlap at E-E_F = -1.2 (-1.6) eV, the Cyt (Gua) has another unique pattern at 1.8 (-0.65) eV. Thus, in principle, each DNA base is distinguishable by our in-house developed data mining technique and 2-dimensional transient time correlation analysis.

S5. Transmission signal of the terminal Cyt of a ssDNA

The detection of the third Cyt in the sequence is not clear as depicted in Fig. 3c, because it does not form a well stacked structure in our fast molecular dynamics. There are two reasons: (1) the terminal region is not properly stacked due to the terminal effect in the fast simulations, and (2) only Cyt can have the edge-to-face conformation ("T"-shaped structure, i.e., the Cyt plane is perpendicular to the graphene surface), though less stable than the stacked structure (face-to-face conformation). Thus, if there is no other significant peak representing a certain base, it needs to be assigned as Cyt as long as the Cyt signal is significant. This is particularly true at the terminal of the sequence. Repetitive (or equivalently long time) measurements should eventually clarify the Cyt signal, since the face-to-face conformation is statistically dominant. If the force is reduced realistically by orders of magnitude, the Cyt will form the stacked structure dominantly because the stacked structure is ~2 kcal/mol more stable than the edge-to-face structure in the narrow GNR.

S6. Comparison between graphene nanoribbons and carbon nanotube electrodes

By exploiting highly stable π-π stacking interactions in DNA sequencing, the orientations of DNA bases are clearly determined so that stable I-V characteristics can be obtained. This approach can be exploited by both GNR and CNT, well known π-conjugated systems, and hence both can be candidates for our DNA sequencing scheme. However, their geometrical differences make GNR advantageous over CNT. When a ssDNA passes by a CNT, a few DNA bases can stack on the device by π-π stacking interactions at once. In the case of GNR, on the other hand, only a single base forms a proper π-π stacking interaction, while neighboring bases have the edge-to-face conformation (Fig. S4). It should be noted that the π-π stacking interaction significantly influences the I-V characteristics, while the edge-to-face interaction on the edges of the GNR hardly affects the I-V characteristics (Fig. S5). Thus, the overlap of signals between neighboring bases is avoided by the flat structure of the GNR. Although it is possible to distinguish different π-stacked multi-bases in both cases, a difficulty arises when the same kind of bases appear in a series.

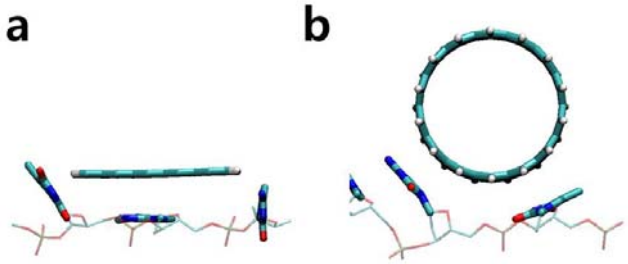

Figure S4| Snapshots of ssDNA-AGNR MD simulation, **a,** and ssDNA-CNT MD simulation**, b**. Double to multiple bases simultaneously form the stacked structures with the CNT for most of the time in MD simulations.

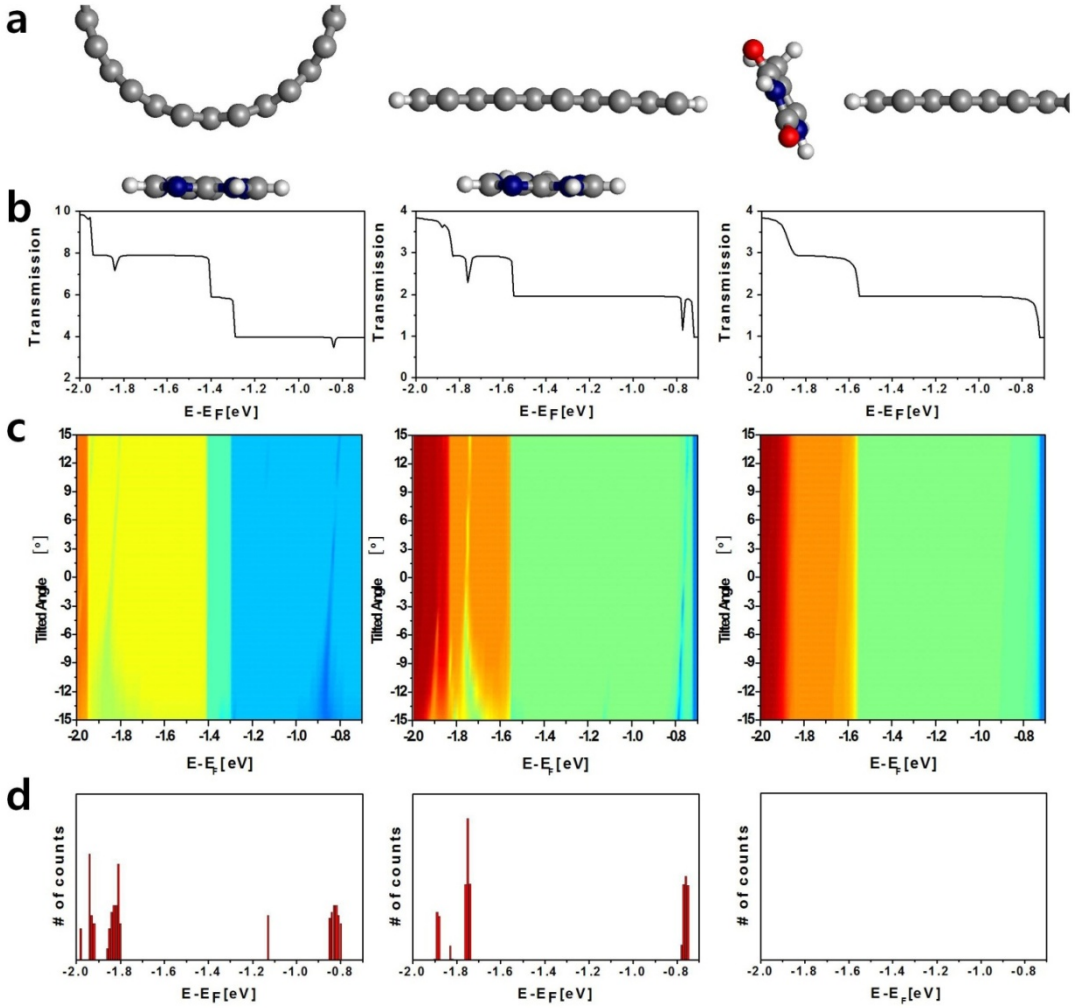

Figure S5| Comparison between GNRs and CNTs for a single base binding. **a,** Structures of equilibrium geometries; **b,** Transmission functions at the equilibrium geometries; **c,** Change of transmission functions as bases rotate; **d,** Distribution functions for kink positions. Left, middle and right columns are cases for A-CNT, A-GNR, and T-GNR, respectively.

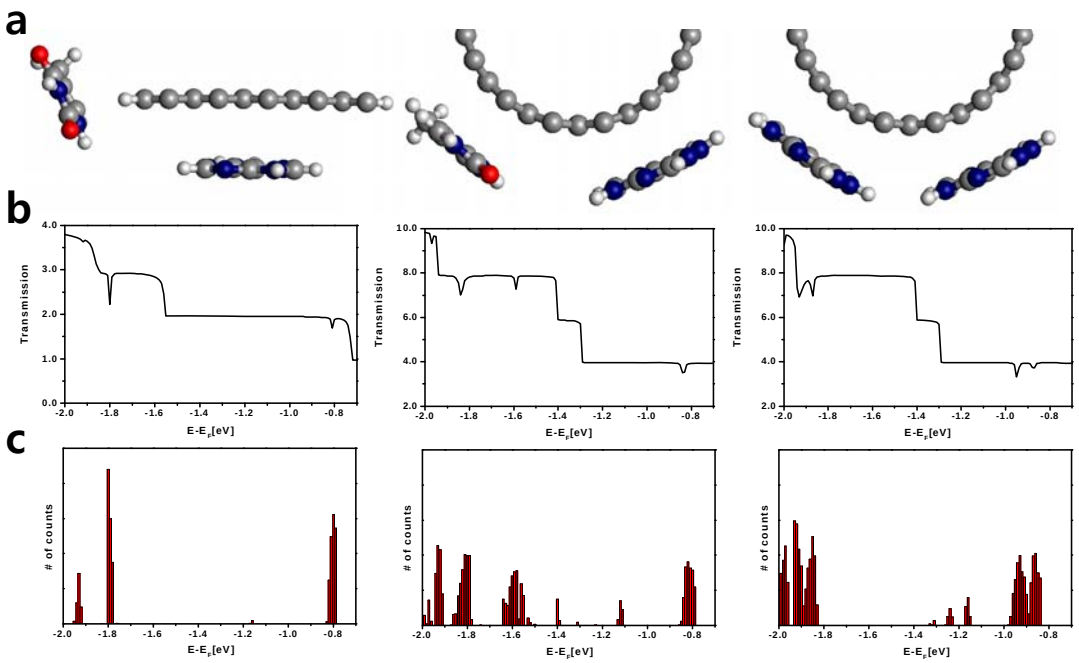

Figure S6| Comparison between GNRs and CNTs for double to multiple bases binding. **a,** Structures of equilibrium geometries; **b,** Transmission functions at the equilibrium geometries; **C,** Distribution functions for kink positions. Left, middle and right columns are cases for AT-GNR, AT-CNT, and AA-CNT, respectively.

We investigated the effects of geometrical changes for a single base-CNT/GNR. In Fig. S5, a stacked Ade (A) is shown in the case of CNT, while in the cae of GNR Ade (A) and Thy (T) are shown in stacked and edge-to-face conformations, respectively. We study the transmission functions as a function of the tilted angle (θ) with respect to the equilibrium geometry (Fig. S5a-c). The distribution function for the kink positions in transmission functions is analyzed by taking into account the Boltzmann factor (~exp[-ΔE/k_BT]) (Fig. S5d). As a result, the transmission functions for A-CNT/GNR show two main kinks at around -0.8 and -1.8 eV, and the kink positions slightly change according to the tilted angle for the stacked conformations. The width of the distribution for A-CNT (~0.05 eV) is larger than that of A-GNR (~0.02 eV). For the edge-to-face conformation, no peaks appear because the edge-to-face interactions hardly affect transmission functions.

The analysis of transmission functions for two bases interacting with GNR or CNT is more helpful to understand the differences (Fig. S6). The structure and transmission functions from the reference geometry are shown in Figs. S6a and b, respectively. The distribution

function for the kink positions is shown in Fig. S6c. Here, we investigated Ade-Thy (AT) interacting with AT-GNR/CNT and Ade-Ade (AA) interacting with CNT (AA-CNT). For the case of AT-GNR, the transmission functions only from Ade is reflected, while both Ade and Thy contribute to the transmission functions for the AT-CNT case. Since Ade and Thy show distant kink positions (Ade: -0.8 -1.8, Thy: -1.6 eV) in transmission, independent information of the two bases can be obtained. However, as an extreme case, AA-CNT shows overlaps of the kink positions, and furthermore the interaction between two bases causes a shift of the positions. Consequently, GNR is superior to CNT for accurate DNA sequencing analysis.

S7. Advantages of narrow GNRs

Thanks to the recent progress of synthesis of narrow GNRs, our sequencing scheme would be feasible in the near future. Our scheme is to search the chage of conductance in a certain energy range by means of gate voltage sweeping. Here, the extent of change, which is a relative quantity, depends on the number of conduction channels which is determined by the width of GNR or CNT. In the case of ~1 nm armchair GNR used in this paper, the conductance drops from 2 to 1 G_0, i.e., by 50 %, in quantum conductance unit ($G_0 = 2e^2/h$) at E-E_F=[-2.0,-1.0] eV. In the case of CNT(14,0) which has ~1 nm diameter, the conductance drops from 8 to 7 G_0, i.e., by 12.5% which is almost at the noise level. Wider GNRs as well as CNTs have more conduction channels, so that the conductance changes by DNA base stacking is less noticeble. Thus, narrower GNRS promise more reliable resolution in the measurement of conductance change.