Supplementary Information

Discrimination of methylcytosine from hydroxymethylcytosine in DNA molecules

Meni Wanunu*[1], Devora Cohen-Karni*[2,3], Robert R. Johnson*[4], Lauren Fields[2], Jack Benner[2],Neil Peterman[1], Yu Zheng[2], Michael L. Klein[4], and Marija Drndic[1]

Table of contents:

Complete Author list for references with >16 authors:

4. Lister, R.; Pelizzola, M.; Dowen, R. H.; Hawkins, R. D.; Hon, G.; Tonti-Filippini, J.; Nery, J. R.; Lee, L.; Ye, Z.; Ngo, Q. M.; Edsall, L.; Antosiewicz-Bourget, J.; Stewart, R.; Ruotti, V.; Millar, A. H.; Thomson, J. A.; Ren, B.; Ecker, J. R. *Nature* 2009, 462, (7271), 315-322.

25. Branton, D.; Deamer, D. W.; Marziali, A.; Bayley, H.; Benner, S. A.; Butler, T.; Di Ventra, M.; Garaj, S.; Hibbs, A.; Huang, X. H.; Jovanovich, S. B.; Krstic, P. S.; Lindsay, S.; Ling, X. S. S.; Mastrangelo, C. H.; Meller, A.; Oliver, J. S.; Pershin, Y. V.; Ramsey, J. M.; Riehn, R.; Soni, G. V.; Tabard-Cossa, V.; Wanunu, M.; Wiggin, M.; Schloss, J. A. *Nat. Biotechnol.* 2008, 26, (10), 1146-1153.

28. MacKerell, A. D.; Bashford, D.; Bellott; Dunbrack, R. L.; Evanseck, J. D.; Field, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S.; Joseph-McCarthy, D.; Kuchnir, L.; Kuczera, K.; Lau, F. T. K.; Mattos, C.; Michnick, S.; Ngo, T.; Nguyen, D. T.; Prodhom, B.; Reiher, W. E.; Roux, B.; Schlenkrich, M.; Smith, J. C.; Stote, R.; Straub, J.; Watanabe, M.; Wiorkiewicz-Kuczera, J.; Yin, D.; Karplus, M. *The Journal of Physical Chemistry B* 1998, 102, (18), 3586-3616.

SI-1. PCR amplification of DNA with different cytosine compositions (C, mC, and hmC).

All DNA molecules in this paper were prepared by PCR using Phusion DNA polymerase (Finnzymes/NEB). The 3kb sequence was amplified from T4 genomic DNA. The 400 bp and 1100 bp samples were amplified from pBR322 plasmid (NEB). To verify that modified cytosines do not introduce mismatches, we sequenced all types of products following PCR amplification. To make DNA samples with mixed cytosine proportions, we have added different cytosine mononucleotide ratios in the PCR mix. Following PCR amplification, the percentage of hmC was qualitatively determined by digestion with a methylation dependent restriction enzyme (MspJI) (see Figure S1). The 3 kbp DNA products were subjected to a 1% agarose gel electrophoresis as shown in Figure S1A below. The PCR products were then incubated with MspJI modification-dependent restriction endonuclease.[1] As demonstrated in Figure S1B, DNA with modified cytosines was digested, and the extent of digestion qualitatively correlates with the fraction of C with respect to hmC in the PCR mix.

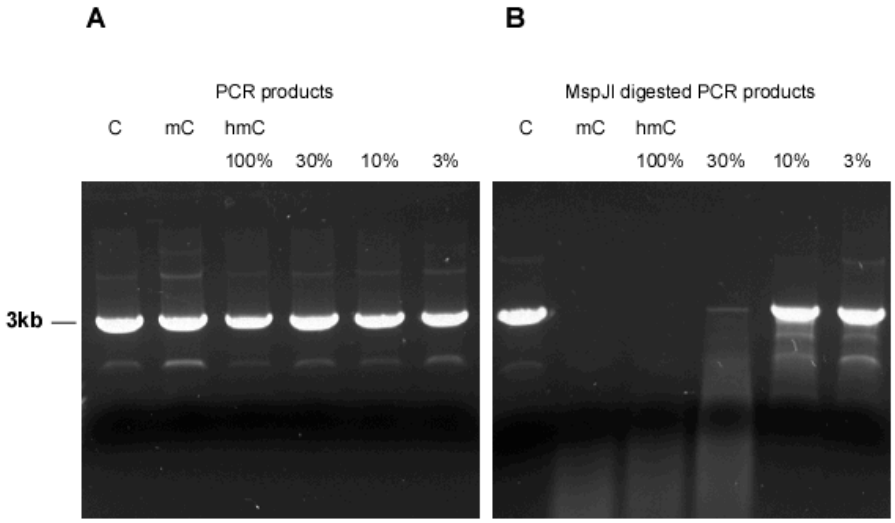


Figure S1. The mC and hmC incorporated 3 kbp DNA PCR products. (A) Purified 3 kb PCR product on 1% agarose gel. (B) Digested by a cytosine-modification dependent restriction enzyme (MspJI, NEB). Sites containing mC or hmC are digested (lane 2, 3), while unmodified C remains intact (lane 1). Qualitatively, samples that contain more hmC than C are more extensively digested (lanes 4, 5, 6).

[1]Zheng, Y., D. Cohen-Karni, et al. (2010). "A unique family of Mrr-like modification-dependent restriction endonucleases." Nucleic Acids Res. Advance Access published on May 5, 2010.

SI-2. Current amplitude histograms for 410 bp C-DNA, mC-DNA, and hmC-DNA.

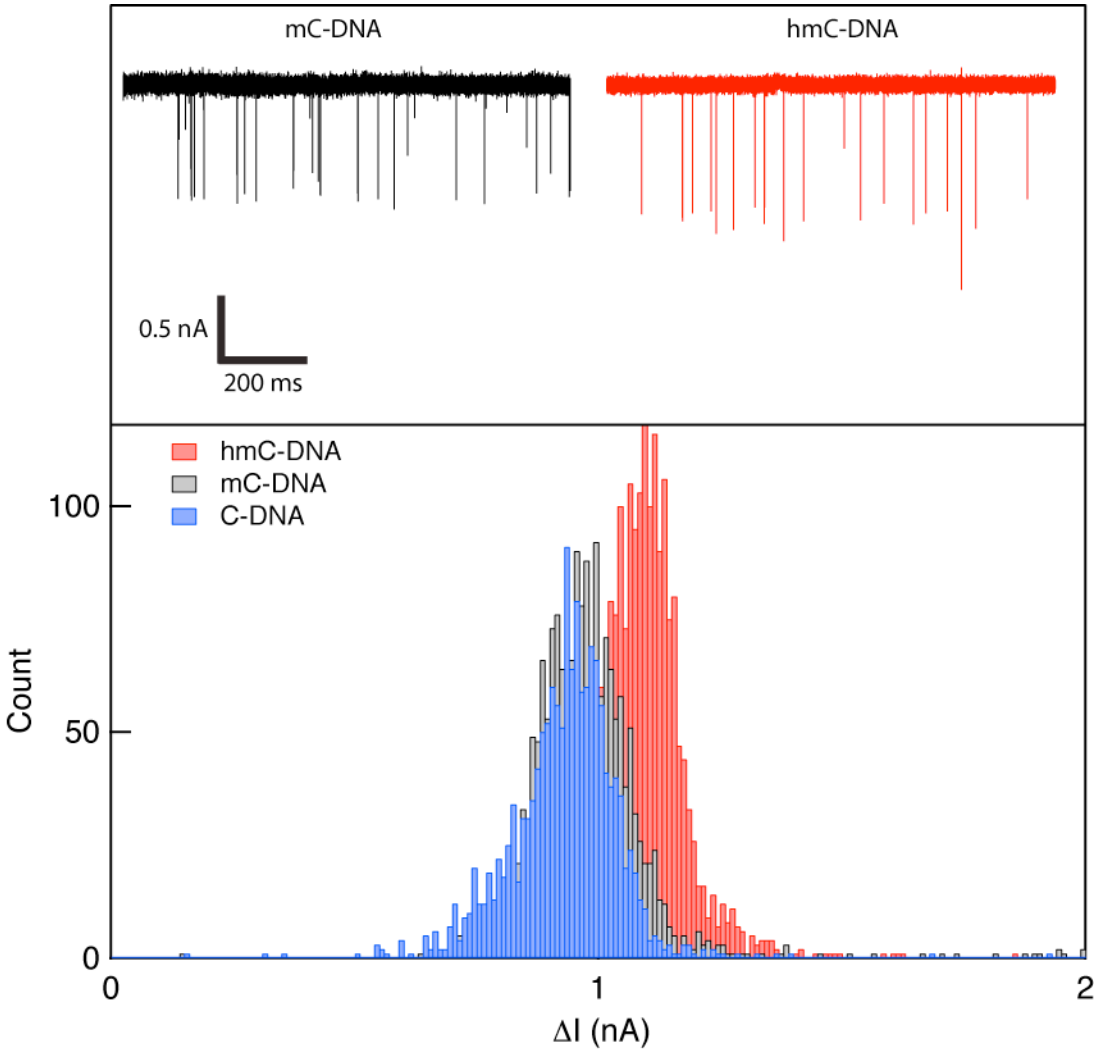


Figure S2. Current amplitude histograms for 410 bp DNA fragments with different cytosine modifications. (Top) Continuous 1-second current traces for 410 bp mC-DNA (black) and hmC-DNA (red), as obtained using a 4 nm pore at 21°C and 300 mV applied voltage. (Bottom) Histograms of the mean current amplitudes (ΔI) for >1,000 events for each of the three modified cytosine variants. The histograms clearly show that hmC-DNA produces deeper amplitudes than C-DNA and mC-DNA.

SI-3. Mass spectrometry analysis of DNA products with mixed cytosines

In order to quantify the different cytosine contents in DNA produced using PCR, we performed liquid chromatography/mass spectrometry (LC/MS) experiments on the digested PCR product samples. Purified PCR fragments were digested using Antarctic Phosphatase (NEB) and Phosphodiesterase I from Crotalus Adamanteus venom (Sigma). Digested nucleoside samples were analyzed by reverse phase liquid chromatography (LC) and electrospray ionization/time-of-flight mass spectrometry (ESI-TOF MS). A reverse phase column, 1 x 150 mm, Develosil RP-Aqueous C30, 3 μm particles, 140 Å pore size (Phenomenex), was developed at a flow rate of 20 μl/min at 30^{o}C using an Agilent 1100 capillary LC connected directly to an Agilent 6210 series ESI-TOF MS. The column was equilibrated with 50mM ammonium acetate, pH 6 in water. A nucleoside sample of up to 8 μl volume was injected, initial conditions were maintained for two-minutes and then the column was developed with a 15-minute linear gradient from 0% to 22.5% acetonitrile, and was held at 22.5% acetonitrile for an additional five minutes. Nucleosides eluted approximately 16-23 minutes following injection. Mass spectra were acquired over a range of 100 to 400 m/z at one cycle/sec and 10,000 transients per scan. The following ionization conditions were used: a VCap of 4000 V, 300 C with a drying gas 7.0 l/min; a nebulizer pressure of 15 psi and a fragmentor voltage of 215 V. The acquired spectra were extracted with Agilent MassHunter Qualitative Analysis Software (with Bioconfirm B 2.0.2) software using mass ranges of 136.0-136.1, 112.0-112.1, 152.0-152.1, 127.0-127.1, 126.0-126.2, 142.0-142.2, 305.1-305.2 for the liberated base fragments of dA, dC, dG, dT, d(5m)C, d(5hm)C and d(5Glucose,hm)C, respectively.

A standard calibration curve of solutions of A, G, T, C, and hmC at different proportions of hmC and C were prepared, maintaining the condition [A] = [G] = [T] = [C] + [hmC]. Extracted ion chromatograms for different bases in different samples and standards are shown in Figure S3. The 0.3 and 0.6 hmC/C ratios are presented for both the 3 kb product and the standard dNTP mix samples. The C (red) and hmC (black) peaks differ in their ratios between the two sample sets.

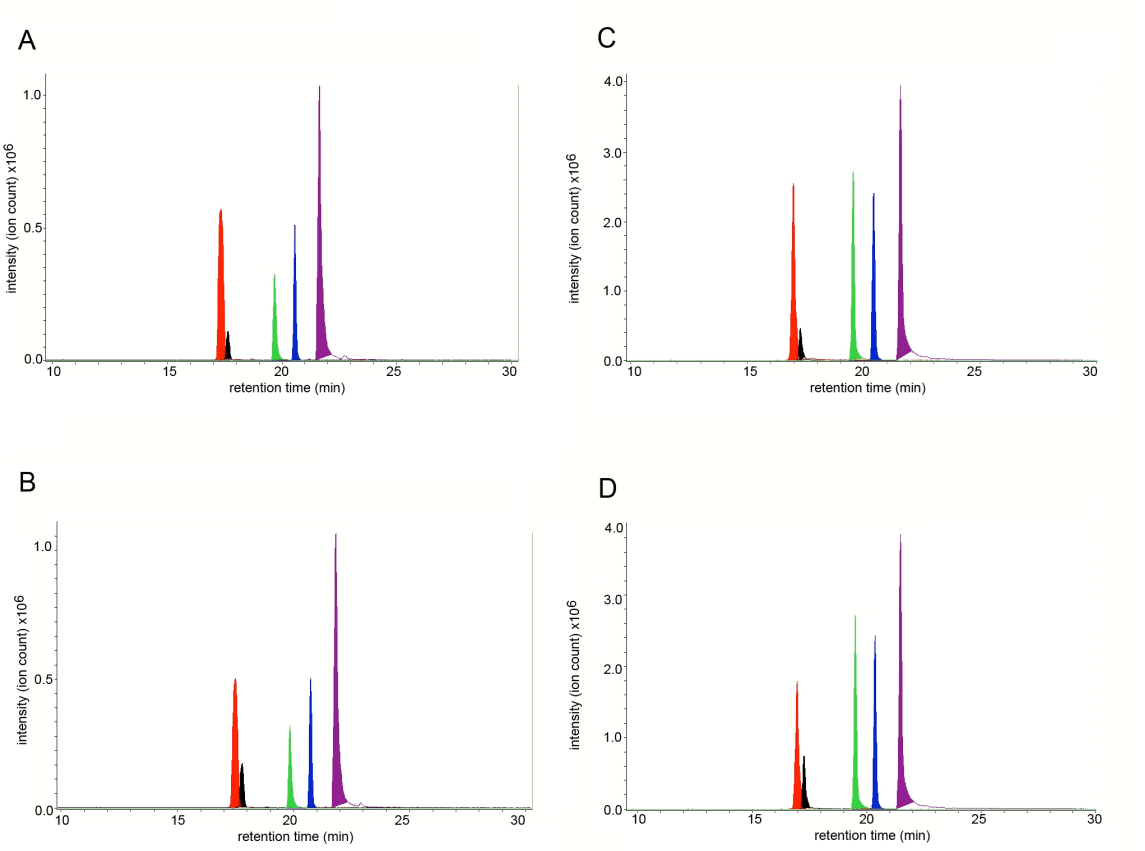


Figure S3. Extracted ion chromatograms for different mononucleosides. The different nucleosides have different retention times. Colors: C (red), mC (black), G (green), T (blue), and A (purple). A) Analysis of the digested 3kb PCR product that used 30% hmC in the PCR mix (30% with respect to cytosines). B) Analysis of the 3kb PCR product with 60% hmC. C) Analysis of a standard dNTP mix that contains 30% hmC. D) Analysis of a standard dNTP mix with 60% hmC.

The intensity of the peaks corresponds to the amount, as detected by MS. In Figure S4a, we show the relative area for each type of mononucleoside as a function of the hmC/C ratio in the nucleotide mix. The relative areas of A, G, and T are not identical, despite their being of equal concentrations due to variation in the ionization from one chemical species to another. To account for this, we normalized the peak areas based on the standards. As expected, increasing the relative concentration of hmC relative to C results in a decrease in the relative area of C (red), and simultaneously, an increase in the relative area of hmC (purple). Finally, based on the normalized curves, in Figure S4b we display the ratios of hmC with respect to C in DNA samples generated with different relative concentrations of hmC to C. Linearity in the curves suggests that the incorporation of different cytosines into growing DNA strands in PCR is non-selective.

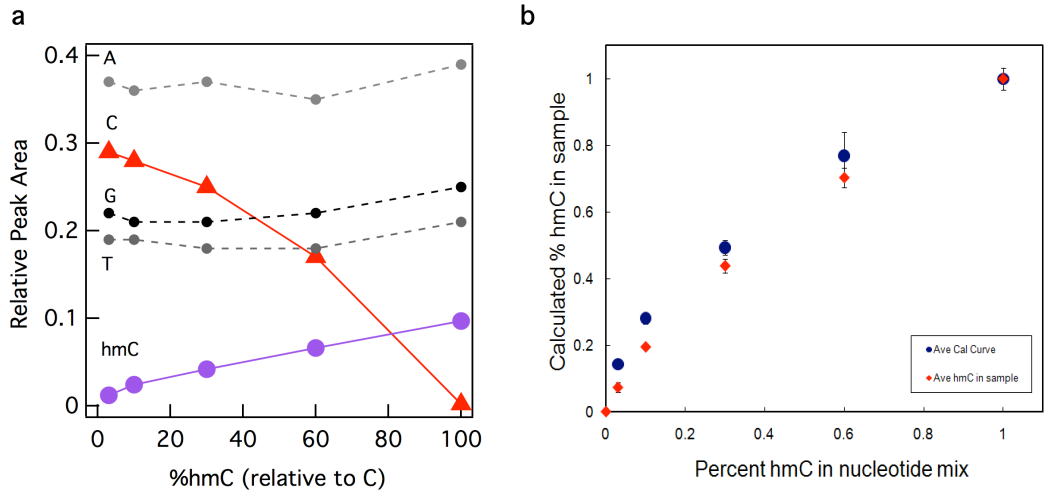


Figure S4. LC/MS analysis of PCR-amplified DNA with modified cytosines. (a) Normalized relative peak areas for the indicated nucleoside bases at different proportions of hmC and C in a nucleotide mix (see text, each point represents the average of three experiments). (b) Comparison of %hmC over C in the PCR samples as compared to the standard calibration curve.

SI-4. Raw fluorescence annealing curves for 3 kbp C-DNA, mC-DNA, and hmC-DNA.

The figure below shows the raw fluorescence of SYBR Green I® in the presence of 3kbp DNA with different cytosine types. To collect the data, samples were prepared in a 96-well plates in triplicates, sealed using a plastic adhesive sticker, and inserted into a RT-PCR instrument programmed to heat to 95°C and hold for one minute, heat to 98°C and hold for one minute, then repeat a cycle in which the temperature is decreased by 0.2°C for 30 sec, followed by exciting the samples at 488 nm and measuring the fluorescence emission at 530 nm. SYBR Green I is a dye that intercalates between the bases of dsDNA, and when doing so, its quantum yield increases dramatically. The presence of dsDNA in the sample is associated with as increase in fluorescence of SYBR Green I in the solution. Raw fluorescence intensities are shown as the temperature of the solution is cooled from 98°C to 40°C in decrements of 0.2°C. After complete annealing, the linear increase in fluorescence is typical for SYBR Green I. Therefore, the data shown in Figure 4b in the paper are the inverse of the differential of the curves shown in Figure SI-4 here, i.e., -df/dT.

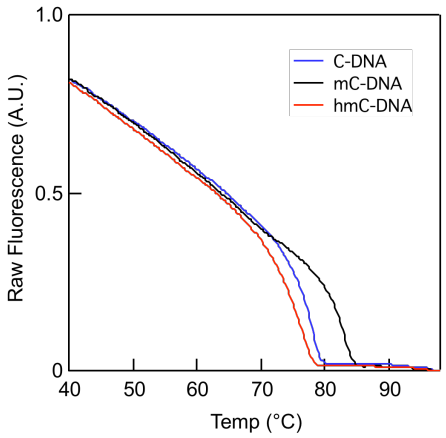


Figure S5. Annealing curves for DNA with different cytosine modifications.

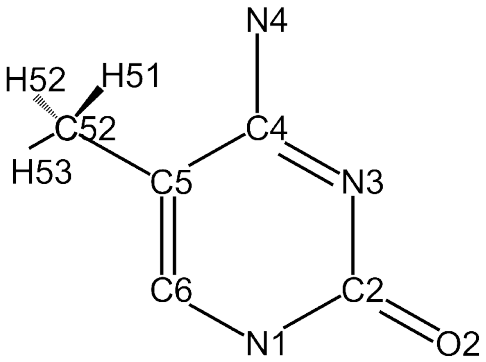
SI-5. Computational details.

Electronic structure calculations were performed with Gaussian 03[2] using the B3LYP functional and the 6-311G* basis set. Methyl and hydroxymethyl modifications can rotate about the C-C bond that connects these groups to the pyrimidine ring. Thus, to ensure that the optimized structures represented the global energy minimum, our procedure included a scan of the potential energy surface associated with rotations about this bond.

Molecular dynamics simulations were performed on $d(A*CT)_9 \cdot (AGT)_9$ duplexes using NAMD[3]. Here, *C represents either C, mC or hmC. The duplexes were simulated in aqueous solution containing 1 M KCl at neutral pH. Each system was simulated for approximately 0.12 μs under ambient temperature and pressure[4] using a 1.5 fs time step. The DNA and counter ions were modeled with the CHARMM force field.[5] The force field for the cytosine modifications was developed following the standard CHARMM protocol and is given in SI. The TIP3P[6] water model was employed for the solvent. Electrostatic interactions were computed using the particle mesh Ewald method and a grid point density of about 1/Å. Analyses of the trajectories was performed in VMD.[7] X3DNA was used to calculate the local helical parameters of the duplexes.[1]

Table S1. Parameters for modified cytosines. Only the parameters for the modified atoms are listed. All other values are taken from the CHARMM cytosine parameters.

5-Methylcytosine



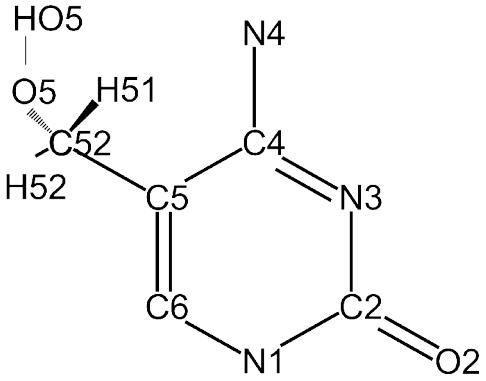
Atom Name	Type	Charge
C5	CN3	-0.06
C52	CT3	-0.27
H51	HA	0.09
H52	HA	0.09
H53	HA	0.09

Bond	Force Constant	Equilibrium Distance
CN3 – CT3	230	1.478

Angle	Force Constant	Equilibrium Angle
CN3 – CT3 – HA	33.43	110.1
CN3 – CN3 – CT3	38.0	124.6
CN2 – CN3 – CT3	38.0	118.7

Dihedral	Force Constant	Multiplicity	Phase
CN2 – CN3 – CT3 – HA	0.46	3	0
CN3 – CN3 – CT3 – HA	0.46	3	0

5-Hydroxymethylcytosine



Atom Name	Type	Charge
C5	CN3	-0.06
C52	CT2	0.05
H51	HA	0.09
H52	HA	0.09
O5	OH1	-0.66
HO5	H	0.43

Bond	Force Constant	Equilibrium Distance
CN3 – CT3	230	1.49

Angle	Force Constant	Equilibrium Angle
CN3 – CT2 – HA	33.43	110.1
CN3 – CN3 – CT2	38.0	124.6
CN2 – CN3 – CT2	38.0	118.7
CN3 – CT2 – OH1	75.7	110.1

Dihedral	Force Constant	Multiplicity	Phase
CN3 – CN3 – CT2 – HA	0.46	3	0
CN2 – CN3 – CT2 – HA	0.46	3	0
CN3 – CN3 – CT2 – OH1	0.46	3	0
CN2 – CN3 – CT2 – OH1	0.46	3	0

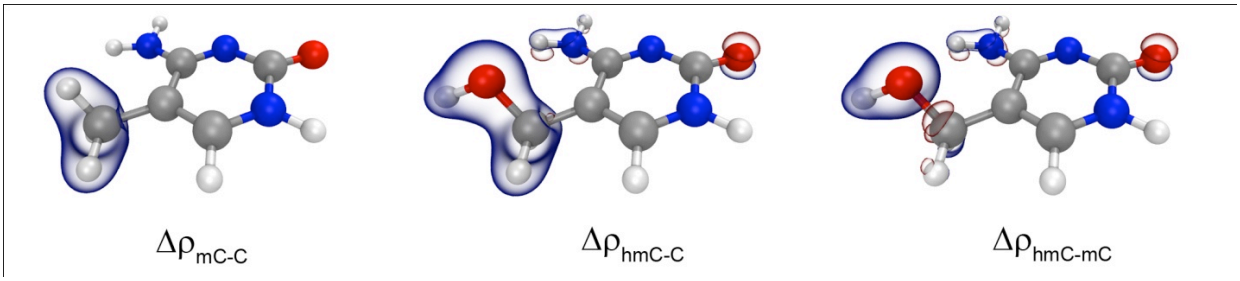


Figure S6. Electron density of mC relative to C (left), hmC relative to C (middle) and hmC relative to mC (right). Blue and red isosurfaces indicate regions of increased and diminished electron density, respectively. The change in electronic structure is localized around the modification itself. The hydroxyl group in hmC polarizes the position 4 amine and position 2 keto groups causing minor geometric changes. However, a Mulliken population analysis reveals no appreciable differences in charge around these atoms. Additionally, counterpoise corrected calculations of G-*C base pair binding energy show that these small structural differences have negligible effect on binding. Thus, using the existing CHARMM charge parameters from C in mC and hmC is justified.

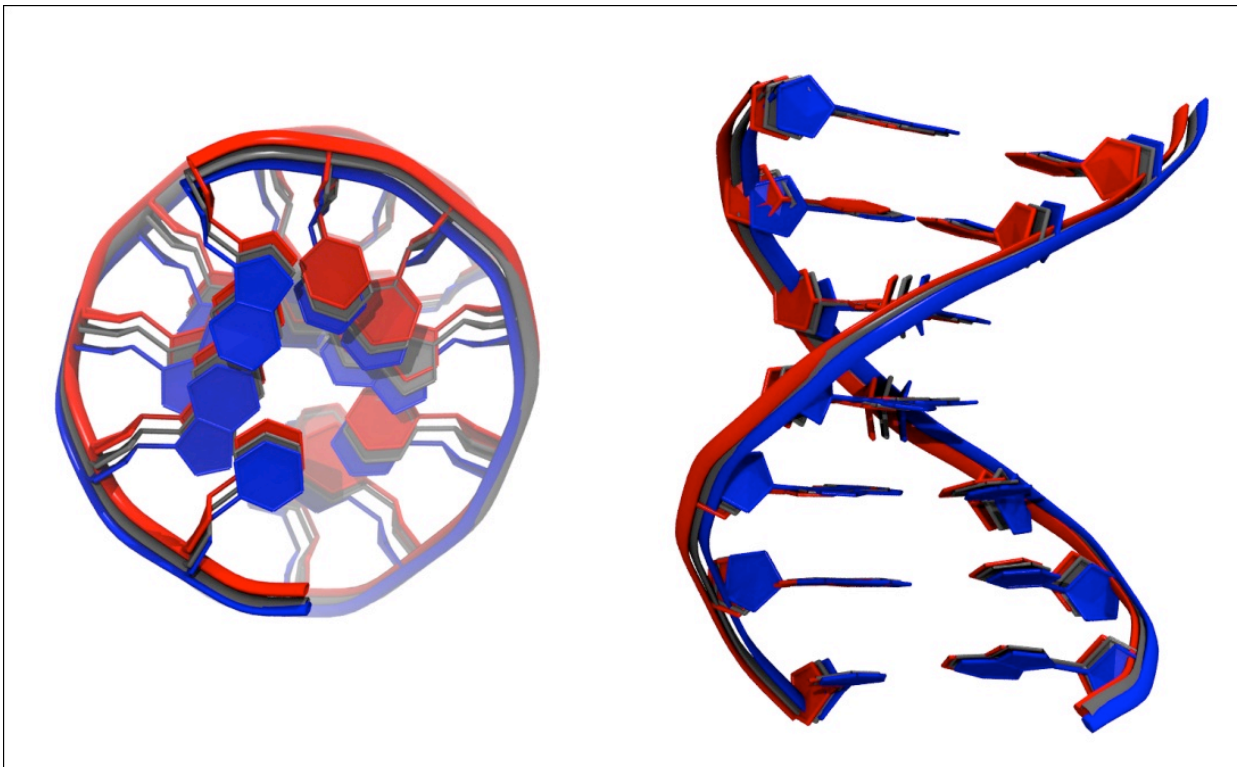


Figure S7. Superimposed average structures of C-DNA (blue), mC-DNA (gray), hmC-DNA (red). The structure changes systematically from C-DNA to hmC-DNA with mC-DNA intermediate. These changes are due to steric effects and, thus, follow the size of the chemical modification.

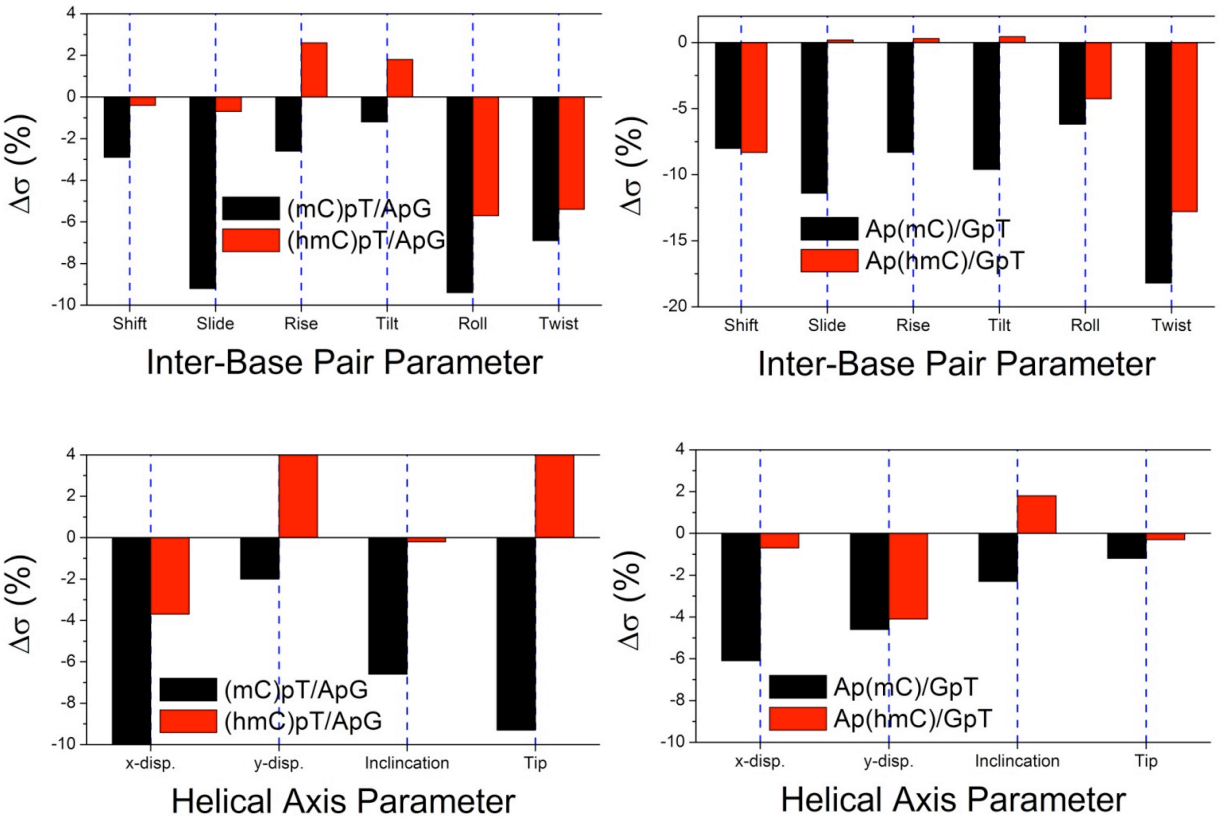


Figure S8. Differences in standard deviations (fluctuations) of inter-base pair and helical axis parameters for Ap(*C)/GpT and (*C)pT/ApG steps relative to unmodified ApC/GpT and CpT/ApG steps. Fluctuations in these parameters depend on both the steric effects and polarity of the modifications. These plots show that these two effects oppose one another: increasing the size of the modification increases the local rigidity of the duplex while increasing the modification's polarity decreases rigidity. Because of these opposite effects, these parameters follow the trend G-C >≈ G-hmC > G-mC. The intra-base pair fluctuations (see manuscript text), on the other hand, are affected less by steric effects and tend to follow the modifications polarity. Thus, the fluctuations in G-*C base pairs follows G-hmC > G-C > G-mC.

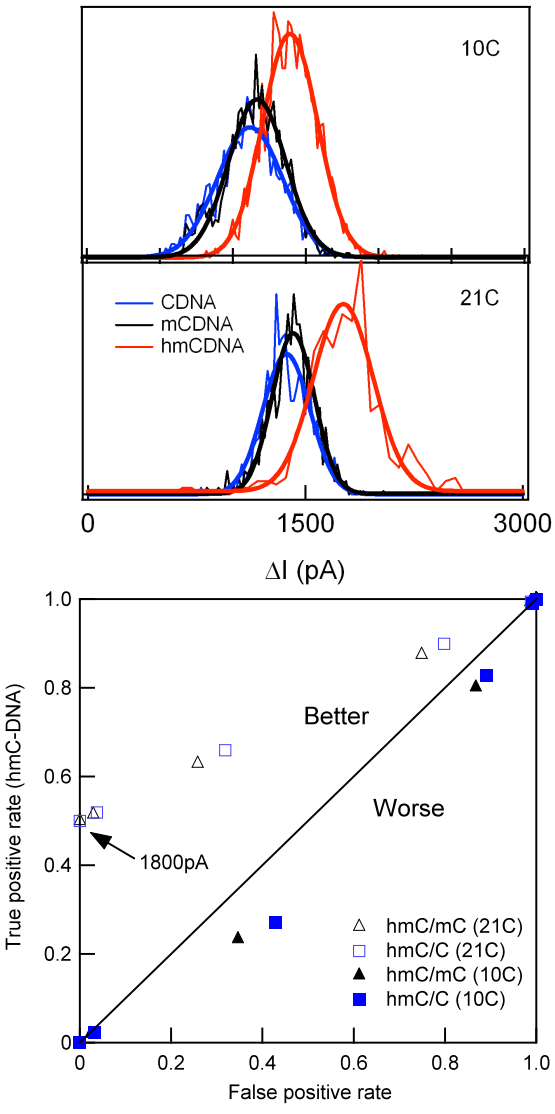
References:

1. Lu, X.J. and W.K. Olson, 3DNA: a versatile, integrated software system for the analysis, rebuilding and visualization of three-dimensional nucleic-acid structures. Nature Protocols, 2008. **3**(7): p. 1213-1227.
2. M. J. Frisch, G.W.T., H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, J. A. Montgomery, Jr., T. Vreven, K. N. Kudin, J. C. Burant, J. M. Millam, S. S. Iyengar, J. Tomasi, V. Barone, B. Mennucci, M. Cossi, G. Scalmani, N. Rega, G. A. Petersson, H. Nakatsuji, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, M. Klene, X. Li, J. E. Knox, H. P. Hratchian, J. B. Cross, V. Bakken, C. Adamo, J. Jaramillo, R. Gomperts, R. E. Stratmann, O. Yazyev, A. J. Austin, R. Cammi, C. Pomelli, J. W. Ochterski, P. Y. Ayala, K. Morokuma, G. A. Voth, P. Salvador, J. J. Dannenberg, V. G. Zakrzewski, S. Dapprich, A. D. Daniels, M. C. Strain, O. Farkas, D. K. Malick, A. D. Rabuck, K. Raghavachari, J. B. Foresman, J. V. Ortiz, Q. Cui, A. G. Baboul, S. Clifford, J. Cioslowski, B. B. Stefanov, G. Liu, A. Liashenko, P. Piskorz, I. Komaromi, R. L. Martin, D. J. Fox, T. Keith, M. A. Al-Laham, C. Y. Peng, A. Nanayakkara, M. Challacombe, P. M. W. Gill, B. Johnson, W. Chen, M. W. Wong, C. Gonzalez, and J. A. Pople, *Gaussian 03, Revision C.02*. 2004, Gaussian, Inc.: Wallingford CT.
3. Phillips, J.C., et al., *Scalable molecular dynamics with NAMD*. Journal of Computational Chemistry, 2005. **26**(16): p. 1781-1802.
4. Feller, S.E., et al., *Constant-Pressure Molecular-Dynamics Simulation - the Langevin Piston Method*. Journal of Chemical Physics, 1995. **103**(11): p. 4613-4621.
5. MacKerell, A.D., et al., *All-atom empirical potential for molecular modeling and dynamics studies of proteins*. Journal of Physical Chemistry B, 1998. **102**(18): p. 3586-3616.
6. Jorgensen, W.L., Quantum and Statistical Mechanical Studies of Liquids .10. Transferable Intermolecular Potential Functions for Water, Alcohols, and Ethers - Application to Liquid Water. Journal of the American Chemical Society, 1981. **103**(2): p. 335-340.
7. Humphrey, W., A. Dalke, and K. Schulten, *VMD: Visual molecular dynamics*. Journal of Molecular Graphics, 1996. **14**(1): p. 33-38.
8. Esteve, P.O., Chin, H.G., Benner, J., Feehery, G.R., Samaranayake, M., Horwitz, G.A., Jacobsen, S.E. and Pradhan, S. Regulation of DNMT1 Stability through SET7-Mediated Lysine Methylation in Mammalian Cells. Proc. Natl. Acad. Sci. 2009 106(13):5076-5081.

SI-6. Statistical analysis of the current amplitude data for C-DNA, mC-DNA, and hmC-DNA.

In this section we focus on statistical analysis of the current amplitude data, in order to establish that the discrimination among hmC-DNA and mC-DNA or C-DNA, as well as the detection of %hmC-DNA, is statistically significant.

1) Top: Histograms of ΔI values for the three molecules at 21°C and 10°C (same data as inset of Figure 2). Bottom: Receiver operating characteristic (ROC) curves for the three different 3kbp cytosine variants at these two temperatures. Considering hmC-DNA as a positive result, and either C-DNA or mC-DNA is a negative result, the true positive rate was calculated as a function of the false positive rate for different threshold ΔI values. The ΔI range of data used in the calculation was 1000 – 2600 pA with 200 pA intervals for 21°C, whereas a range of 400 – 2200 pA with 200 pA intervals were used for the data at 10°C. This curve is useful to learn about sensitivity (ordinate) vs. specificity (abscissa) in our signals: At 21°C, we can observe *all* of the hmC-DNA signals by detecting all events with amplitudes above a threshold of 1000 pA (sensitivity of 1, or TPR = 1), although the rate of false positives (mC-DNA or C-DNA events) would also be 1 (i.e., our system would also detect all of mC-DNA events). In contrast, selecting a threshold of 1800 pA reduces our sensitivity to 0.52 (i.e., 50% of hmC-DNA events are undetected), but specificity increases to 99.7% because the false positive rate is 0.003. A similar trend is observed with C-DNA at 21°C (blue markers). In contrast, at 10°C it becomes quite difficult to distinguish among hmC-DNA and the other cytosines, and most of the points fall below the straight diagonal line, suggesting a bad trade-off between sensitivity and specificity. The conclusion of this ROC curve is that 21°C is more optimal for discrimination than 10°C, although in either case, a population of molecules is required in order to establish its identity.



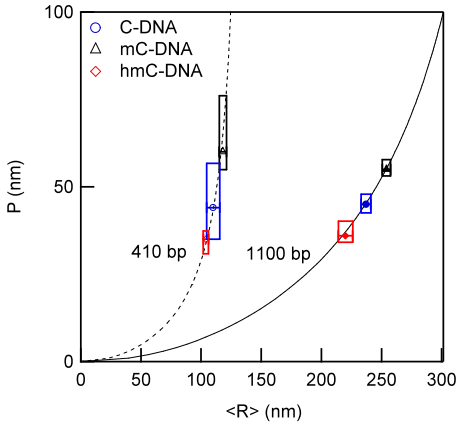
2) In the following table presents a statistical analysis of the current amplitude data for different % hmC-DNA samples. Specifically, to test the significance of the differences in the mean between the sample populations, we performed unpaired Student's T-tests for each neighboring population. That is, 3% hmC was compared to 0% hmC, 10% hmC compared to to 3% hmC, and 30% hmC was compared to 10% hmC. The null hypothesis of this t-test is that the two means are the same. Therefore, based on the t-value, the corresponding p-value represents the probability that our null hypothesis is correct. In order to analyze all samples, we formulated the hypotheses that every neighboring pair of samples is similar. This was done for all of the datasets that yielded Figure 3, for both the hmC-DNA with C-DNA background and the hmC-DNA with mC-DNA background. T-values and p-values for the 0% hmC samples are blank because the 3% hmC samples are being compared to it.

Sample	ΔI_{norm} **(pA)**	**St. Dev. (pA)**	***N*** **(events)**	***t***	***p*** **(two-tailed)**
30% hmC/C	249	129	1,419	32.9	<0.0001
10% hmC/C	82	100	1,392	9.13	<0.0001
3% hmC/C	27	99	1,427	1.69	0.09
0% hmC/C	0	111	1,419	-	-
0% hmC/mC	0	120	1,200	-	-
3% hmC/mC	56	106	1,502	3.90	0.001
10% hmC/mC	105	98	1,299	7.91	<0.0001
30% hmC/mC	225	130	1,401	16.0	<0.0001

The p-values on the right column of the above Table show that with the exception of 3% hmC/C sample, in which we can only be 91% certain that the 0% hmC/C mean is significantly different than the 3% hmC/C mean, our discrimination is quite robust. However, we stress that this discrimination is reliant upon obtaining sufficient statistics.

SI-7. Statistical analysis of persistence length from AFM data

The raw data collected from the AFM images are the mean end-to-end distance <*R*> (nm). Using Eqn. 1 in the manuscript and contour lengths of 0.34(nm/bp)*N(bp), the mean values of <*R*> for the two DNA lengths have been converted to the persistence length *P* (nm). The uncertainty of the mean μ_{err} was determined from the following general formula:



$$\mu_{err} = \frac{s}{\sqrt{n}},$$

where *s* is the standard deviation of the Gaussian fit and *n* is the number of measurements (indicated in Figure 6 of the manuscript). The figure to the right shows curves of *P* as a function of *R* based on Eqn 1. Superimposed on the curves are the experimental values of <*R*> determined from AFM, where error bars of $\pm\mu_{err}$ are shown. The boxes around each point are drawn in order to translate the error magnitude in <*R*> to the respective error in *P*. From these data we estimated the range of *P* values as quoted in the text, taking into account the largest error from each measurement.

Additionally, to test the significance of the differences in the mean between the sample populations, we performed unpaired Student's T-tests for each population with respect to C-DNA. For example, in the 1100 bp DNA case, assuming the null hypothesis that $\langle R\rangle_{mC\text{-}DNA} = \langle R\rangle_{C\text{-}DNA}$, we obtain $t = 2.90$ for the 428 measurements (190 for mC-DNA and 223 for C-DNA, degrees of freedom are $n\text{-}2 = 411$). This yields a p-value of 0.0040, which means that the probability of the null hypothesis to be correct is 0.4%. So we are 99.6% confident that $\langle R\rangle_{mC\text{-}DNA} \neq \langle R\rangle_{C\text{-}DNA}$. From the table below, we conclude that the <R> values are *significantly different for all three cytosine variants* with at least 99.6% certainty.

The following table presents the means, *t*-values, and *p*-values for our measurements (**NOTE**: C-DNA data do not have *t*-values and *p*-values because the other measurements are being compared to it).

Sample	**<*R*> (nm)**	**St. Dev. (nm)**	***n***	***t***	***p* (two-tailed)**
mC-DNA, 1100 bp	270	54	190	2.90	0.004
C-DNA, 1100 bp	255	47	223	-	-
hmC-DNA, 1100 bp	231	58	205	4.78	< 0.0001
mC-DNA, 410 bp	129	33	155	5.04	< 0.0001
C-DNA, 410 bp	112	19	127	-	-
hmC-DNA, 410 bp	102	24	117	3.45	0.001