



Error rates for nanopore discrimination among cytosine, methylcytosine, and hydroxymethylcytosine along individual DNA strands

Jacob Schreiber, Zachary L. Wescoe, Robin Abu-Shumays, John T. Vivian, Baldandorj Baatar, Kevin Karplus, and Mark Akeson¹

Nanopore Group, Department of Biomolecular Engineering, University of California, Santa Cruz, CA 95064

Edited* by Daniel Branton, Harvard University, Cambridge, MA, and approved September 26, 2013 (received for review June 4, 2013)

Cytosine, 5-methylcytosine, and 5-hydroxymethylcytosine were identified during translocation of single DNA template strands through a modified *Mycobacterium smegmatis* porin A (M2MspA) nanopore under control of phi29 DNA polymerase. This identification was based on three consecutive ionic current states that correspond to passage of modified or unmodified CG dinucleotides and their immediate neighbors through the nanopore limiting aperture. To establish quality scores for these calls, we examined ~3,300 translocation events for 48 distinct DNA constructs. Each experiment analyzed a mixture of cytosine-, 5-methylcytosine-, and 5-hydroxymethylcytosine-bearing DNA strands that contained a marker that independently established the correct cytosine methylation status at the target CG of each molecule tested. To calculate error rates for these calls, we established decision boundaries using a variety of machine-learning methods. These error rates depended upon the identity of the bases immediately 5' and 3' of the targeted CG dinucleotide, and ranged from 1.7% to 12.2% for a single-pass read. We estimate that Q40 values (0.01% error rates) for methylation status calls could be achieved by reading single molecules 5–19 times depending upon sequence context.

MspA | epigenetics

Epigenetic modifications of DNA help regulate gene transcription in biological cells. In mammals, 5-methylcytosine (mC) modification of CG dinucleotides is known to influence development (1, 2) and contribute to human diseases including cancer (3). Other modifications have been detected at carbon 5 of cytosine including 5-hydroxymethylcytosine (hmC) (4), and more recently 5-formylcytosine, and 5-carboxycytosine (5). Physiological roles for hmC in carcinogenesis and embryonic stem cell differentiation have been proposed (6).

High-throughput techniques for mC detection are based on bisulfite treatment of genomic DNA (7). In the conventional assay, cytosine (but not mC nor hmC) is converted to uracil (8). Thus, positions not converted to uracil identify cytosines that were modified in the original genomic sequence. In a landmark paper, Lister et al. (9) used this technique to map genome-wide cytosine methylation in human embryonic stem cells and fetal lung fibroblasts at single-nucleotide precision. Recently, bisulfite strategies for discriminating between mC and hmC using the Tet1 enzyme (10) or by chemical modification of hmC (11) have been described.

Single-molecule techniques have emerged as possible alternatives to bisulfite treatment for detecting epigenetic modifications of DNA (12). These single-molecule approaches share several useful features including few processing steps before sequence analysis, long reads that routinely exceed several thousand nucleotides, and the ability to read native DNA strands in heterogeneous mixtures. The most advanced of these single-molecule techniques, from Pacific Biosciences, uses fluorescence to detect labeled nucleotide triphosphates during daughter-strand elongation. This elongation is catalyzed by a DNA polymerase adsorbed at the base of a zero-mode waveguide (13, 14). Using the Pacific Biosciences device, cytosine can be distinguished from

mC on the DNA template strand based on differences in light pulse duration and time between light pulses as G nucleotides are incorporated during elongation.

Nanopores have been used to characterize nucleic acids (15). Nanopore-based DNA strand sequencing is a newer single-molecule technique that is anticipated for commercial release in the near term. Enzymatic control at single-nucleotide precision (16, 17) coupled to a protein pore that can resolve 3- to 4-nt “words” (18) were key developments that facilitated this technology. Using synthetic DNA strands suspended as static complexes, cytosine, mC, and hmC in the pore sensor could be distinguished from one another using a modified α -hemolysin nanopore (19).

In this report, we used DNA synthesis by a wild-type phi29 DNA polymerase (phi29 DNAP) to processively move DNA templates through the *Mycobacterium smegmatis* porin A (M2MspA) pore at single-nucleotide precision. This allowed us to read individual cytosine, mC, and hmC nucleobases at a target CG dinucleotide within strands of known base composition. Using independent markers on each template strand, we calculated single-pass error rates for calling methylation status at the target CG for all 16 permutations of nucleobases immediately flanking the CG dinucleotide.

Results

Comparison of Ionic Current Patterns for Cytosine, mC, hmC. The nanopore device is shown schematically in Fig. 1. A single M2MspA pore was embedded in a ~30- μ m-diameter lipid bilayer that separated two wells each containing ~100 μ L of a 0.3 M KCl solution. A representative synthetic DNA substrate is

Significance

Modification of cytosine bases in DNA can determine when genes are turned on in biological cells. These modifications are important during cell differentiation, embryogenesis, and aberrant cell growth in cancer. Here, we present a nanopore technique that permits direct detection of cytosine, 5-hydroxymethylcytosine, and 5-methylcytosine on individual synthetic DNA strands of known sequence. This technique focuses on three ionic current amplitudes that occur as an enzyme motor pulls the cytosine on a captured DNA strand through the nanopore. For genomic DNA, we predict that a given strand must be read 5–19 times to achieve cytosine methylation calls that are sufficiently accurate for epigenetic studies.

Author contributions: K.K. and M.A. designed research; J.S., Z.L.W., R.A.-S., and B.B. performed research; J.S. and K.K. contributed new reagents/analytic tools; J.S., Z.L.W., R.A.-S., and J.T.V. analyzed data; and J.S., Z.L.W., R.A.-S., and M.A. wrote the paper.

Conflict of interest statement: M.A. is a consultant to Oxford Nanopore Technologies (Oxford, UK).

*This Direct Submission article had a prearranged editor.

¹To whom correspondence should be addressed. E-mail: makeson@soe.ucsc.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1310615110/-DCSupplemental.

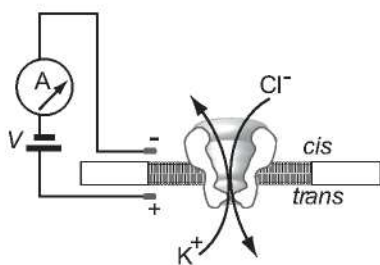
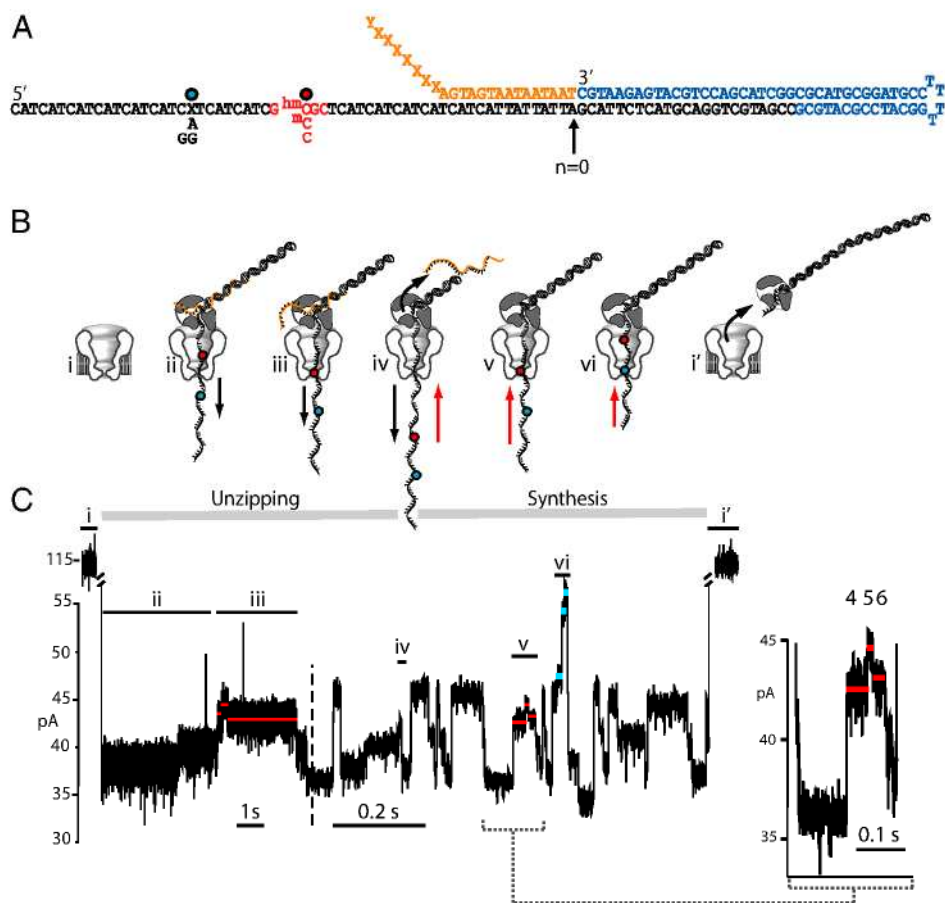


Fig. 1. Schematic of the nanopore device. A single M2MspA channel was inserted into a 30-μm-diameter horizontal lipid bilayer that separated two 100-μL wells, termed cis and trans. The wells were filled with buffer composed of 0.3 M KCl, 10 mM HEPES/KOH (pH 8.00 ± 0.05) at 23 ± 1 °C. An applied voltage between Ag/AgCl electrodes (180-mV trans side positive, unless otherwise noted) was used to capture and analyze individual DNA strands moving through the nanopore controlled by phi29 DNAP.

shown in Fig. 2A, along with an annealed blocking oligomer that protected the 3'-OH terminus from phi29 DNAP-catalyzed strand elongation in bulk phase (16, 20). The single-stranded 59-nt

Fig. 2. Strategy for reading epigenetic modifications on a target CG dinucleotide. (A) Example DNA substrate. An 82-nt DNA template strand (black letters) was annealed to a 55-nt DNA strand (blue letters) bearing a four-base loop and a short duplex stem. This hybrid formed a 37-bp double-stranded segment and a 59-nt 5'-overhang. A blocking oligomer (gold letters) was annealed to this DNA substrate so that the 5' end of the block was stacked against the 3' end of the DNA hairpin. The blocking oligomer included a seven-residue abasic tail (Xs) capped by a single C3-spacer (Y). This tail served to protect the blocking oligomer from 3' exonuclease digestion and as a lever to promote its removal at the nanopore orifice under voltage load. In bulk phase, the blocking oligomer prevented phi29 DNAP access to the 3'-OH of the DNA substrate. The target CG dinucleotide (red letters) is 28–29 nt from the first base copied during synthesis (A at $n = 0$, arrow). The nearest neighbors to the CG dinucleotide are also highlighted in red. In this particular DNA strand, the neighboring bases are G (5') and C (3') and the modified cytosine residue (red circle) is hmC. It was coupled to an abasic residue 10 positions toward the 5' end (X, blue circle) from the cytosine of interest. This abasic residue served as an independent marker that identified an hmC-bearing DNA strand. Cytosine and mC at the CG dinucleotide in otherwise identical DNA strands were identified by a "GG" or "A" marker, respectively (blue circle). DNA substrates bearing all other 15 permutations of nearest neighbors and the three variants of cytosine with the corresponding markers (Fig. S1), were assembled and analyzed in the same manner. (B) Capture and translocation of DNA bound to phi29 DNAP. Colors and symbols correspond to those shown in A. (i) Open channel. (ii) Capture of a DNA-phi29 DNAP complex. (iii) Unzipping of the blocking oligomer as the polymerase is driven 5' to 3' along the template strand by the applied voltage. The target cytosine nucleotide (red circle) is shown at the center of the M2MspA pore constriction that dictates the measured ionic current. (iv) The blocking oligomer is completely unzipped from the captured DNA template, placing the 3'-OH of the primer strand at the catalytic center of phi29 DNAP. Synthesis of the primer strand now drives DNA template movement against the applied voltage (red arrow). (v) The target cytosine (red circle) reenters the pore constriction from the opposite direction and is reread. (vi) The marker (blue circle) enters the pore constriction as synthesis proceeds, identifying the strand as bearing cytosine, mC, or hmC in the target CG dinucleotide. (i') Primer strand synthesis continues, eventually pulling the template strand out of the pore into the cis compartment. (C) Representative ionic current trace. Letters i–vi and i' correspond to cartoons in B. The vertical dashed line demarks a change in the timescale so that unzipping and synthesis (which is ~10-fold faster) can be viewed in a continuous trace. The three horizontal red lines correspond to ionic current states 4, 5, and 6, which afford the most discrimination between cytosine, mC, and hmC during translocation of a target CG dinucleotide through the M2MspA pore constriction. These key states are expanded in the *Inset at Right*.



overhang of the DNA substrate contained the CG dinucleotide that is the focus of this study. The nucleobases immediately adjacent to the CG are highlighted. This 4-mer was varied to include all 48 possible permutations (16 permutations each for cytosine, mC, and hmC at the target CG; Fig. S1). These DNA strands also included identifiers 10 nt 5' of the cytosine target that allowed independent determination of the correct methylation status of each captured strand (e.g., an abasic marker "X" in Fig. 2A indicated hmC at the CG dinucleotide). To simplify analysis, the remainder of the template strand was composed of CAT trinucleotide repeats. This trinucleotide results in a discernible three-step ionic current series (18).

The sequence of events during polymerase-regulated DNA movement through the nanopore is diagrammed in Fig. 2B. A corresponding ionic current trace is also shown (Fig. 2C). The sequence is as follows: (i) open channel (~115 pA); (ii) capture and voltage-dependent ratcheting of the template through the nanopore toward the trans compartment, advancing in single-nucleotide steps, as the blocking oligomer is unzipped from the template strand absent catalysis; (iii) unzipping continues as the CG dinucleotide passes through the nanopore limiting aperture, resulting in a characteristic series of ionic current steps; (iv) release of the blocking oligomer and positioning of the phi29

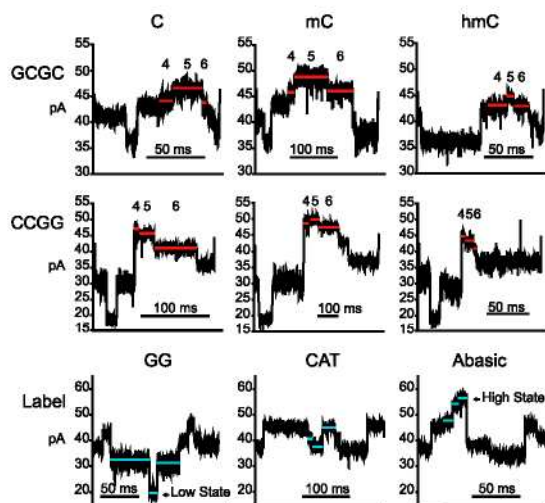


Fig. 3. Ionic current traces comparing individual DNA strands bearing cytosine, mC, and hmC. The first row shows traces for the GCGC context, and the second row shows traces for the CCGG context. Ionic current states 4, 5, and 6 appeared to give the best discrimination for methylation status at the target CG dinucleotide and are highlighted with red lines (Movie S1). The third row shows traces for downstream markers used to identify cytosine-, mC-, and hmC-bearing DNA strands, respectively, with blue lines highlighting the characteristic states for each marker. A total of 639 CCGG-bearing DNA strands was analyzed, of which 175 contained cytosine, 287 contained mC, and 177 contained hmC at the CG dinucleotide. A total of 448 GCGC-bearing DNA strands was analyzed, of which 205 contained cytosine, 134 contained mC, and 109 contained hmC at the CG dinucleotide. Traces shown were filtered using a 2-kHz low-pass digital filter.

DNAP at the primer–template junction allowing catalytic addition of nucleotides to the daughter strand, reversing the direction of DNA template translocation toward the *cis* compartment; (v) polymerase-driven elongation of the primer strand continues to move the DNA strand against the nanopore electric field in single-nucleotide increments, pulling the CG dinucleotide and its immediate neighbors through the M2MspA sensor, resulting in a characteristic ionic current series that mirrors (iii); (vi) daughter-strand elongation continues and the marker moves into the pore-limiting aperture, resulting in current steps that unambiguously identify the strand; and (i') return to open channel following release of the DNA molecule into the *cis* compartment as the DNA template strand became too short to be retained in the pore electric field. Control experiments confirmed that the marker at (vi) did not influence the ionic current measured during translocation of the CG dinucleotide and its neighboring bases (v).

To illustrate the effect of cytosine methylation state on ionic current, we chose two DNA 4-mer contexts (CCGG and GCGC) that are frequently methylated in the human genome (9). Example traces are shown in Fig. 3 along with labels used for independent identification of the methylation status of each molecule. Within these traces, eight consecutive ionic current states were quantified. These states were predicted to include those influenced by the target CG dinucleotide along with flanking states characteristic of the background CAT-dependent ionic current pattern. The positions of states 1–8 were chosen based on the following: (i) the number of nucleotide additions (and associated discrete current states) that must occur from the start of primer extension to CG entry into the M2MspA-limiting aperture (16–17 nt); and (ii) disruption in the repeating CAT-dependent current pattern.

For both the CCGG and GCGC contexts, initial inspection suggested that ionic current states 4, 5, and 6 were highest for mC, intermediate for cytosine, and lowest for hmC. By comparison, ionic current states 1–3 and 7–8 appeared similar regardless of methylation status. To quantitatively test the importance of states 4, 5, and 6 for all 16 4-mer contexts, we implemented

a forest of extremely randomized trees (21). A forest of extremely randomized trees is an ensemble of decision tree classifiers where each classifier only uses a few features to make predictions. Features that are important to making a correct classification will be present in the well-performing classifiers. Using this strategy applied to DNA substrates bearing all 4-mer contexts (examples traces, Fig. S2), we confirmed that states 4, 5, and 6 were key for correct cytosine methylation state classification (Fig. 4). Graphical representations based on these three values for each CCGG-bearing strand (639 events; Fig. 5A) and for each GCGC-bearing strand (448 events; Fig. 5B) revealed discrete population clusters that were representative of graphs for all sequence contexts (Fig. S3). Outliers are distinguishable when looking at their corresponding ionic current trace (Fig. S4).

Error Estimates Using Decision Boundaries Based on Three Machine-Learning Methods.

A practical device for calling cytosine methylation status will require automated and statistically robust partitioning of the ionic current space between three populations (cytosine, mC, and hmC). Thus, for each of the 16 sequence contexts examined in this study, we implemented Naive Bayes, Support Vector Machine (SVM), and Random Forest algorithms to learn decision boundaries between the three cytosine populations. These algorithms were built and tested using ionic current states 4, 5, and 6 for each of 3,300 translocation events partitioned between the 16 4-mer contexts. In brief, we used 80% of the event data for a given context to train each classifier. This resulted in decision boundaries that classified cytosine vs. mC vs. hmC for that sequence context. These decision boundaries were then used to predict methylation status of the remaining 20% of the data. The error rate was then calculated by comparing those predictions against the known methylation status reported by the marker on each strand. We repeated this process four times by holding out a different 20% fraction of the data for testing. The number of incorrect calls in each of these iterations divided by the number of total events in the dataset yielded the error estimate for that dataset. Finally, we repeated the entire process 20 times to get a distribution of error estimates.

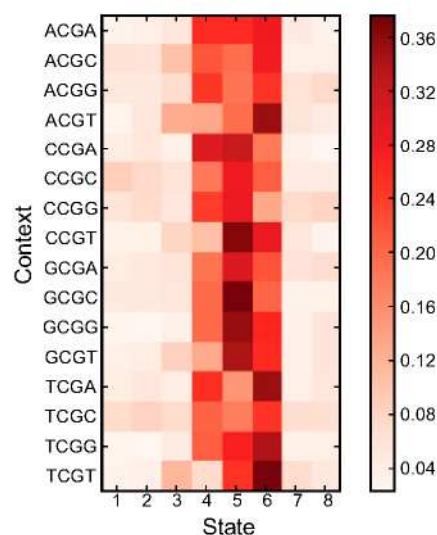


Fig. 4. Quantitative identification of ionic current states that discriminate between cytosine, mC, and hmC. Each row of the heat map corresponds to one 4-mer sequence context. Each column indicates one of the eight measured ionic current states. The “heat” of a cell is the importance of the state (Methods), derived from fitting a forest of extremely randomized trees to that 4-mer context, normalized to sum to 1 for each row. The legend at Right shows the color coding of the heat map. States 4, 5, and 6 were almost always the most important states for a correct classification across all sequence contexts.

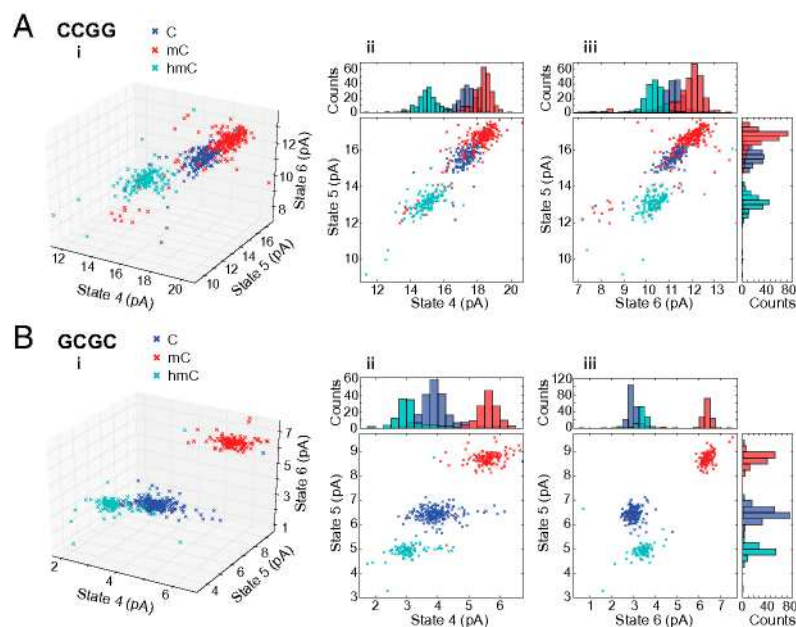


Fig. 5. Comparison of key ionic current states for DNA strands bearing cytosine, mC, and hmC at the target CG dinucleotide. Data are from experiments where a mixture of three cytosine variants in a given DNA sequence context was added to the *cis* compartment bathing a single M2MspA nanopore. In all panels, Xs represent normalized ionic current values for an individual DNA strand read one time. Blue Xs indicate cytosine at position 29 of the template strand and a “GG” label, red Xs indicate a mC at position 29 and an “A” label, and teal indicates an hmC at position 29 and an abasic label. (A) CCGG. In this context, discrimination between cytosine, mC, and hmC was the weakest for the 16 4-mer contexts analyzed. (B) GCGC. In this context, discrimination between cytosine, mC, and hmC was the strongest for the 16 4-mer contexts analyzed. Horizontal panels for both A and B represent the following: (i) 3D plots for ionic current states 4, 5, and 6; (ii) 2D plots for ionic current states 5 vs. 4; (iii) 2D plots for ionic current states 5 vs. 6. For the CCGG and GCGC contexts, 639 and 448 total DNA strands were read, respectively. Six replicate traces for CCGG and GCGC templates, including an example outlier for each, are shown in Fig. S4.

Table 1 shows the mean and SD of this distribution for every sequence context. None of the three classifiers consistently outperformed the others, and the difference between cytosine, mC, and hmC is large enough that any reasonable classifier would work. The error rates for methylation status calls ranged from 1.7% (GCGC) to 12.2% (CCGG) for a single read, with an average of ~6.5% error rate across all 16 sequence contexts. Using the lowest and highest error rates, we estimated the number of times an individual captured strand would have to be read for the majority of

calls to be in error less than 0.01% of the time (Q40; Fig. 6). This number ranged from 5 (GCGC) to 19 (CCGG) required reads.

Conclusion

We have shown that methylation status of cytosines can be determined on single DNA strands of known canonical base sequence using a biological nanopore coupled to a DNA polymerase motor. The error rates for calling cytosine, mC, and hmC averaged ~6.5% for a single read along individual captured strands. The size of this error depended upon the identity of bases immediately neighboring the target CG dinucleotide. We anticipate that the methylation status classifiers used in this study could be improved using larger training sets to delineate more precise decision boundaries. Also, alternate data-processing pipelines, e.g., principal component analysis, may include additional informative features beyond the three ionic current states used here and thus improve call accuracy. Multiple reads of individual genomic strands will be essential for most applications.

A nanopore device developed for epigenetics will likely focus first on well-documented reference genomes important in basic research (e.g., the mouse genome) and in health care (e.g., the human and dog genomes). This has important practical advantages. That is, a nanopore instrument capable of *de novo* DNA sequencing will require algorithms that extract single-base identity from ionic current states for all sequence contexts. This will necessarily include sequences where individual bases are difficult to discern (e.g., homopolymeric regions). In contrast, algorithms designed to detect methylation status of specific cytosines during resequencing would require high accuracy only for target cytosines and their nearest neighbors (as achieved at CG dinucleotides in this study). The remainder of each target genomic strand could be identified using lower precision ionic current patterns established from the reference sequence.

In its full embodiment, nanopore analysis of DNA methylation status could provide three significant advantages over conventional bisulfite-based methylation assays: (i) elimination of miscalls due to sample modification; the nanopore sensor touches and identifies each genomic DNA base directly, and therefore laboratory base modification is not needed; (ii) high accuracy along individual genomic DNA strands; upon capture, each strand can be retained in the nanopore allowing for as many reads as necessary to achieve the desired call accuracy; and (iii) linkage analysis of distant cytosine methylation sites. In principle, nanopore devices can analyze genomic DNA strands that are thousands of nucleotides in length.

Table 1. Accuracy of cytosine methylation calls using Naive Bayes, SVM, and Random Forest classifiers

Context	Count	Naive Bayes	SVM (RBF)	Random Forest
ACGA	169	96.1 (±0.4)	95.2 (±0.4)	95.6 (±0.5)
ACGC	163	93.0 (±0.8)	92.7 (±0.7)	93.0 (±0.8)
ACGG	165	94.0 (±0.8)	94.4 (±0.8)	95.1 (±0.9)
ACGT	165	94.9 (±0.4)	91.0 (±0.9)	93.9 (±0.6)
CCGA	124	92.8 (±0.9)	92.5 (±0.5)	94.3 (±0.6)
CCGC	171	93.6 (±0.5)	93.7 (±0.5)	93.1 (±1.1)
CCGG	639	87.8 (±0.3)	89.2 (±0.3)	90.3 (±0.4)
CCGT	171	94.4 (±0.3)	92.9 (±0.7)	93.5 (±0.7)
GCGA	154	89.7 (±0.7)	92.2 (±0.7)	90.6 (±1.6)
GCGC	448	98.2 (±0.2)	98.3 (±0.2)	98.1 (±0.2)
GCGG	176	95.3 (±0.8)	97.3 (±0.6)	96.4 (±0.8)
GCGT	141	95.4 (±0.4)	96.0 (±0.6)	95.2 (±0.9)
TCGA	206	94.4 (±0.6)	95.0 (±0.3)	95.8 (±0.6)
TCGC	152	89.2 (±0.8)	91.3 (±1.5)	90.2 (±1.2)
TCGG	142	90.5 (±0.9)	91.4 (±1.0)	90.9 (±1.3)
TCGT	158	90.9 (±0.9)	90.1 (±1.2)	91.4 (±1.1)

The first column at left indicates the 4-mer sequence context within the DNA template strand analyzed. The target CG dinucleotide is at the center of the 4-mer. The total number of DNA strands (cytosine-bearing strands plus mC-bearing strands plus hmC-bearing strands) used to establish decision boundaries and call accuracies for each context is shown in the second column. The mean and SD of the accuracy achieved using Naive Bayes, SVM, and Random Forest classifiers is given in columns 3–5, respectively. The worst accuracy was 87.8% (12.2% error rate) using the Naive Bayes classifier on the CCGG context, and the best accuracy was 98.3% (1.7% error rate) using the SVM classifier on the GCGC context. The accuracy of the highest performing classifier for each context is in boldface.

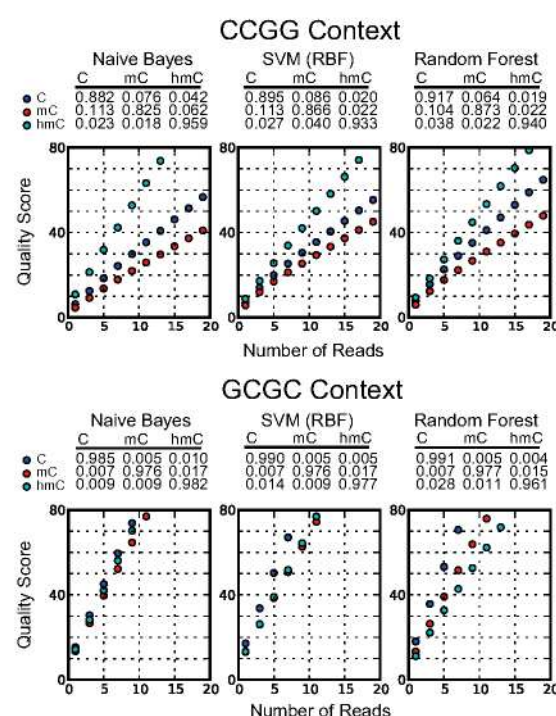


Fig. 6. Confusion matrices and quality score plots for each methylation status in the CCGG and GCGC contexts, using Naive Bayes, SVM, and Random Forest classifiers. Each confusion matrix shows the probability that a strand bearing a given cytosine variant (C, top row; mC, middle row; hmC, bottom row) will be classified as bearing a specific cytosine variant (C, left column; mC, center column; hmC, right column). The diagonal represents the probability of a correct call. The multinomial probability of the majority call being incorrect is then calculated using the values in the confusion matrix. The number of reads of a single strand needed to achieve a desired accuracy threshold based on the multinomial probability is depicted in the quality score plots below each confusion matrix. For the GCGC context (which has a higher accuracy along the diagonal of its confusion matrices), a Q40 quality score (0.01% error) for all cytosine variants was predicted to require only five strand rereads using the best classifier (Naive Bayes). By comparison, for the CCGG context (which has a lower accuracy along the diagonal of its confusion matrices), a Q40 quality score for all cytosine variants was predicted to require 19 strand rereads using the Naive Bayes classifier.

Methods

Proteins. The M2MspA protein (22) used to form single nanopores was prepared in our laboratory. Briefly, the ORF for M2MspA was constructed by overlapping PCR. This amplicon was sequenced and inserted into a pESUMO plasmid, then expressed and purified using the SUMOpro Expression System in *Escherichia coli* (LifeSensors). Wild-type bacteriophage phi29 DNA polymerase(exo+) was obtained from Enzymatics Corporation (833,000 U/mL; specific activity, 83,000 U/mg) and stored absent detergent.

DNA. Oligonucleotides were purchased from the Stanford Peptide and Nucleic Acid Facility and then purified by denaturing PAGE. Phosphoramidites used for the syntheses were from Glen Research. The phosphoramidite for the abasic residue was dideoxyribose-3'-[(2-cyanoethyl)-(N,N-diisopropyl)]-phosphoramidite. For mC, the phosphoramidite was 5'-dimethoxytrityl-N-benzoyl-5-methyl-2'-deoxycytidine-3'-[(2-cyanoethyl)-(N,N-diisopropyl)]-phosphoramidite. For hmC, the phosphoramidite was 5'-dimethoxytrityl-N-benzoyl-5-cyanoethoxy-methyl-2'-deoxycytidine-3'-[(2-cyanoethyl)-(N,N-diisopropyl)]-phosphoramidite. It was deprotected by 30% (wt/vol) ammonium hydroxide for 17 h at 75 °C. The 3'-spacer used in our blocking oligomers was made from (1-dimethoxytrityloxy-propanediol-3-succinoyl)-long-chain alkylamino-CPG (controlled pore glass) from Glen Research. DNA hybrids used for these studies (Fig. 2A) were composed of an 82-nt template, a 55-nt primer containing a 5'-hairpin, and a 15-nt blocking oligomer that protected the 3'-OH of the primer strand and thus limited DNA synthesis to hybrids captured and activated by the nanopore (20). Substrate hybrids were formed by combining the

template, hairpin primer, and blocking oligomer at a ratio of 1:1.2:1.2, incubating the mixture at 95 °C for 2.5 min, and snap chilling in an ice water bath. The presence of mC and hmC bases in template strands bearing GhmCGC, GmCGC, ChmCGG, and CmCGG 4-mers was verified by liquid chromatography/mass spectrometry (see below).

Verification of the Presence of Cytosine, mC, and hmC in Model DNA Strands Using Liquid Chromatography/Mass Spectrometry. 2'-Deoxycytidine-5'-monophosphate (dCMP), 5-methyl-2'-deoxycytidine-5'-monophosphate (mdCMP), and 5-hydroxymethyl-2'-deoxycytidine-5'-monophosphate (hmdCMP) standards were generated by digesting three reference 897-nt oligonucleotides (Zymo Research), one of which contained only unmodified cytosine, one of which contained only mC, and one of which contained only hmC. The 20-μL digestion reaction was performed at 37 °C for 3 h using 300 ng of DNA and 1 unit of DNA degradase (Zymo Research). Upon completion, the reactions were diluted with H₂O and run through NanoSep3K Omega spin columns (Pall Corporation) to separate the dNMPs from the enzyme and any undigested DNA. Two micrograms of each of the six templates (GCGC and CCGG contexts) were digested with 2 units of DNA degradase and processed in parallel with the DNA standards.

Samples containing 100 ng of digested standard DNA or 800 ng of sample DNA (GCGC and CCGG contexts) were analyzed by liquid chromatography–tandem mass spectrometry (LC-MS/MS) on a ThermoElectron Finnigan LTQ mass spectrometer (Thermo) coupled with a surveyor HPLC at the University of California Santa Cruz Mass Spectrometry Facility. Reverse-phase liquid chromatography was done with a Synergi Hydro 4-μm Fusion-RP 80A column (150 mm × 2.00 mm diameter; 4-μm particle size) (Phenomenex). Solvent A was 0.1% formic acid in water. Solvent B was 0.1% formic acid in methanol. The gradient was as follows: time (t) = 0–3 min, 100% solvent A; t = 3–5 min, 70% solvent A, 30% solvent B; t = 5–10 min, 10% solvent A, 90% solvent B; t = 10–20 min, 100% solvent B. The flow rate for chromatography was 200 μL/min.

The eluant from HPLC was processed by MS in negative mode over a full scan range of *m/z* 50–350, followed by MS/MS scans from the global time scheduled list of three compounds: *m/z* = 306.2 (dCMP), 320.2 (mdCMP), and 336.2 (hmdCMP). The electrospray voltage was set to 4.5 kV. The collision-induced dissociation at a normalized collision energy of 35% was used for MS/MS. Data were analyzed using the XCalibur software (Thermo). MS/MS for the modified dCMPs gave distinct breakdown products. The presence of dCMP in our template samples was confirmed in MS/MS by the presence of 263 and 245 *m/z* ions, mdCMP by the presence of 277 and 259 *m/z* ions, and hmdCMP by the presence of 292 and 275 *m/z* ions (Fig. S5).

Nanopore Experiments. Single M2MspA channels were inserted in lipid bilayers in 0.3 M KCl, 10 mM HEPES/KOH (pH 8.00 ± 0.05) at 23 °C as previously described (18, 22). The channels used for this study had currents between 110 and 120 pA at 180 mV (*trans*-side positive) upon insertion. All experiments were run at 180 mV. Current detection and voltage control for nanopore experiments were performed using an integrated patch-clamp amplifier (Axopatch 200B) set in whole-cell mode with a 5-kHz low-pass Bessel filtering. A Digidata 1440A (Molecular Devices) analog-to-digital converter sampled the data at 100 kHz.

Unless otherwise indicated, nanopore experiments were performed using 1 μM each of DNA substrates bearing XCGY, XmCGY, and XhmCGY hybrids added to the *cis* compartment with the same flanking nucleotides in the *X* and *Y* positions. The *cis* compartment was supplemented with 1 mM DTT, 2.25 μM phi29 DNA polymerase(exo+), 1 mM EDTA, 10 mM MgCl₂, and 1 mM each of the four dNTPs.

Data Collection. Extraction of “on pathway” events was semiautomated using pClampfit software (Axon Instruments) applied to ionic current data smoothed at 2,000 Hz with a low-pass digital Bessel filter. The criteria used to identify on pathway events were as follows: (i) an event must have begun with a drop from open-channel current to ~35–45 pA followed by procession through an unzipping regime wherein the ionic current states mirrored states observed during strand synthesis; (ii) all eight states of interest must have been present during strand synthesis; and (iii) the independent marker current state must have been identifiable one CAT ionic current series after the eighth quantified state. An event could contain an unexplained reversible drop in current as long as it did not occur during any of the eight states of interest or during the marker state. If an unexplained current drop below 0 pA occurred, the event was discarded. Using these criteria, ~3,300 events were extracted and analyzed constituting ~26% of all events 500 ms in duration or longer.

Once an event had been identified as on pathway, the eight ionic current states of interest were quantified. These states could be readily identified

because states 1 and 8 were always bounded by the highest current state produced by the CAT repeating pattern. In addition, the pattern produced by states 2–7 always differed significantly from the CAT-dependent three-state pattern into which they were inserted.

Feature Selection. For each translocation event, we compiled mean ionic current levels for eight states during passage of the target CG dinucleotide and its neighbors through the M2MspA pore. These values were accompanied by a label for the correct methylation state reported by the upstream marker on the same DNA strand. The Gini importance was calculated for each feature in a specific context using a forest of extremely randomized trees (21, 23). Feature selection was performed manually by taking features showing a Gini importance above a random model across all contexts. Briefly, data compiled for all translocation events for a given 4-mer context were fitted to a forest of 250 extremely randomized trees, with each tree using only three features. Among the 16 4-mer contexts, the number of compiled events ranged from $n = 124$ to $n = 639$. The open-source scikit-learn (version 0.13.1) (24) command we used to build the forest of extremely randomized trees and calculate the normalized Gini importance was as follows:

```
ExtraTreesClassifier (n_estimators = 250, compute_importances = True,
max_features = 3, max_depth = None, min_samples_split = 1, random_
state = 42).
```

Building Decision Boundaries. Using feature selection (see above), we identified ionic current states 4, 5, and 6 as important for calling cytosine methylation status. To remove 1–2 pA of baseline current drift due to evaporation, we normalized ionic current values for states 4, 5, and 6 of a given event by subtracting the mean of states 1, 2, 3, 7, and 8 (which were not used in classification) for the same event. Data shown in Fig. 5 are based on this normalization. The data were then converted to the standard scaler, $\mu = 0$, unit variance, for each state, which is a standard machine-learning preprocessing technique.

Three classifiers were used to estimate an error rate for the data: (i) SVM with a radial basis function (RBF), which is a nonprobabilistic method for calculating decision boundaries; (ii) Random Forest, which in this case was an ensemble of 30 trees; and (iii) Naive Bayes, which is a probabilistic method that uses Bayes' rule to classify each point. The scikit-learn commands for the classifiers were as follows:

```
SVC(gamma = 2, C = 1, kernel = 'rbf')
RandomForestClassifier(max_depth = 5, n_estimators = 30, max_features = 2)
GaussianNB().
```

Using each of these classifiers, we performed stratified five-fold cross-validation of data from each of the 16 4-mer contexts. This ensured that methylation status could be predicted for every event using decision

boundaries that were not derived using the event itself. Briefly, we trained decision boundaries from each classifier on 80% of the data, and used those decision boundaries to predict the methylation status of the remaining 20% of the dataset. Because each event included a label identifying a DNA strand's correct methylation status, we were able to determine whether the decision boundary had made a correct or incorrect call. This was then repeated four times, with a unique 20% subset of the data held out for testing each cycle. By summing the number of incorrect calls from each 20% of the dataset tested, we estimated the error by dividing the number of errors by the number of events. To ensure that this error estimate was not an outlier, the data were shuffled and the process repeated 20 times. This gave a distribution of error estimates for cytosine methylation status calls for the 4-mer context analyzed.

Quality Score Estimates. To estimate the number of reads of a single DNA strand needed to achieve Q40 (0.01% incorrect calls) for a given sequence context, we first calculated confusion matrices for the Naive Bayes, Random Forest, and SVM classifiers using full precision accuracy calls. These confusion matrices quantified the specific types of miscalls rather than simply the total number of miscalls.

We then used a multinomial distribution based off this confusion matrix to estimate the number of reads needed to achieve Q40 for methylation status calls within a given sequence context. The multinomial equation was as follows:

$$\sum_{i=0}^N \sum_{j=0}^N \binom{N}{i} \binom{N-i}{j} p^i q^j r^{N-i-j} + \sum_{i=0}^N \sum_{j=0}^{\min(i-1, N-2i)} \binom{N}{i} \binom{N-i}{j} p^i q^j r^{N-i-j},$$

where N is the number of reads of a molecule, p is the probability of a correct call, q is the probability of an incorrect call in one manner, r is the probability of an incorrect call in the other manner, i is the number of events called correctly, j is the number of events called incorrectly in the manner corresponding to probability q , and $N - i - j$ is the number of events called incorrectly in the manner corresponding to probability r . When solved, this equation gives the probability of an incorrect call given N reads. Thus, to achieve Q40, we increased N until this probability dropped below 0.01%.

ACKNOWLEDGMENTS. We thank Jeff Nivala and Shawnie Miller for expression of the M2MspA protein, Max Cherf for conducting the initial experiments that led to this study, and Li Zhang for assistance with mass spectrometry analysis. We thank Peter Walker at the Stanford Peptide and Nucleic Acid facility for oligonucleotide synthesis. Funding support for the University of California Santa Cruz Mass Spectrometry Facility was provided by National Institutes of Health's National Center for Research Resources (Shared Instrumentation Grant S10-RR020939). The nanopore work was supported by National Human Genome Research Institute Grant HG006321-02 (to M.A.).

- Li E, Bestor TH, Jaenisch R (1992) Targeted mutation of the DNA methyltransferase gene results in embryonic lethality. *Cell* 69(6):915–926.
- Okano M, Bell DW, Haber DA, Li E (1999) DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development. *Cell* 99(3):247–257.
- Robertson KD (2005) DNA methylation and human disease. *Nat Rev Genet* 6(8):597–610.
- Tahiliani M, et al. (2009) Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. *Science* 324(5929):930–935.
- Ito S, et al. (2011) Tet proteins can convert 5-methylcytosine to 5-formylcytosine and 5-carboxylcytosine. *Science* 333(6047):1300–1303.
- Wu H, Zhang Y (2011) Mechanisms and functions of Tet protein-mediated 5-methylcytosine oxidation. *Genes Dev* 25(23):2436–2452.
- Frommer M, et al. (1992) A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. *Proc Natl Acad Sci USA* 89(5):1827–1831.
- Huang Y, et al. (2010) The behaviour of 5-hydroxymethylcytosine in bisulfite sequencing. *PLoS One* 5(1):e8888.
- Lister R, et al. (2009) Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* 462(7271):315–322.
- Yu M, et al. (2012) Base-resolution analysis of 5-hydroxymethylcytosine in the mammalian genome. *Cell* 149(6):1368–1380.
- Booth MJ, et al. (2012) Quantitative sequencing of 5-methylcytosine and 5-hydroxymethylcytosine at single-base resolution. *Science* 336(6083):934–937.
- Korlach J, Turner SW (2012) Going beyond five bases in DNA sequencing. *Curr Opin Struct Biol* 22(3):251–261.
- Eid J, et al. (2009) Real-time DNA sequencing from single polymerase molecules. *Science* 323(5910):133–138.
- Flusberg BA, et al. (2010) Direct detection of DNA methylation during single-molecule, real-time sequencing. *Nat Methods* 7(6):461–465.
- Deamer DW, Branton D (2002) Characterization of nucleic acids by nanopore analysis. *Acc Chem Res* 35(10):817–825.
- Cherf GM, et al. (2012) Automated forward and reverse ratcheting of DNA in a nanopore at 5-Å precision. *Nat Biotechnol* 30(4):344–348.
- Cockroft SL, Chu J, Amorin M, Ghadiri MR (2008) A single-molecule nanopore device detects DNA polymerase activity with single-nucleotide resolution. *J Am Chem Soc* 130(3):818–820.
- Manrao EA, et al. (2012) Reading DNA at single-nucleotide resolution with a mutant MspA nanopore and phi29 DNA polymerase. *Nat Biotechnol* 30(4):349–353.
- Wallace EV, et al. (2010) Identification of epigenetic DNA modifications with a protein nanopore. *Chem Commun (Camb)* 46(43):8195–8197.
- Olasagasti F, et al. (2010) Replication of individual DNA molecules under electronic control using a protein nanopore. *Nat Nanotechnol* 5(11):798–806.
- Geurts P, Ernst D, Wehenkel L (2006) Extremely randomized trees. *Mach Learn* 63(1):3–42.
- Butler TZ, Pavlenok M, Derrington IM, Niederweis M, Gundlach JH (2008) Single-molecule DNA detection with an engineered MspA protein nanopore. *Proc Natl Acad Sci USA* 105(52):20647–20652.
- Raileanu L, Stoffel K (2004) Theoretical comparison between the Gini index and information gain criteria. *Ann Math Artif Intell* 41(1):77–93.
- Pedregosa F, et al. (2011) Scikit-learn: Machine learning in Python. *J Mach Learn Res* 12:2825–2830.