# Cytosine Methylation Variant Calling with MinION Nanopore Sequencing

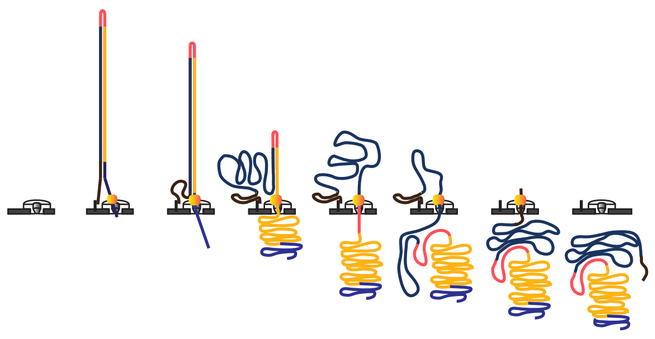
Arthur C. Rand, Miten Jain, Jordan Eizenga, Audrey Musselman-Brown, Hugh E. Olsen, Mark Akeson and Benedict Paten

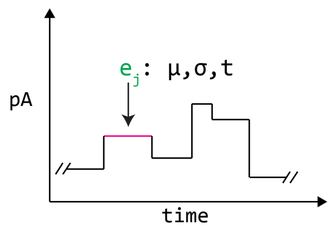
Department of Biomolecular Engineering, University of California, Santa Cruz

## Abstract

Chemical modifications to DNA regulate cellular state and function. The Oxford Nanopore MinION is a portable single-molecule DNA sequencer that can sequence long fragments of genomic DNA. Here we show that the MinION can be used to detect and map three cytosine variants: cytosine, 5-methylcytosine, and 5-hydroxymethylcytosine. We present a probabilistic method that enables expansion of the nucleotide alphabet to include bases containing chemical modifications. Our results on synthetic DNA show that individual cytosine base modifications can be classified with accuracy up to 95% in a three-way comparison and 98% in a two-way comparison. We also demonstrate that 5-methylcytosine can be accurately mapped in E. coli genomic DNA

## Nanopore Sequencing

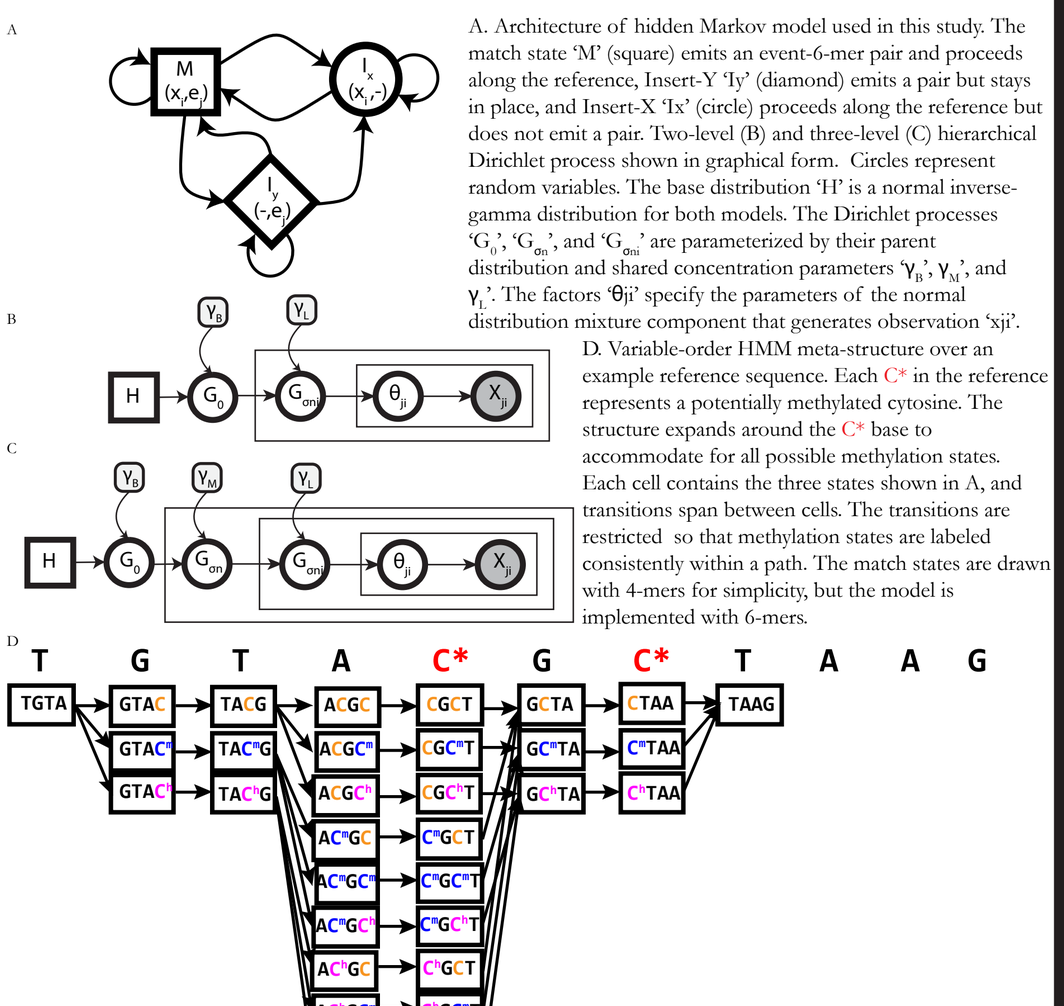





ATGCACTGAACA
ATGCAC
TGCACT
GCACTG ← $X_i$
CACTGA
ACTGAA
CTGAAC

A nanometer-sized protein pore embeded in a membrane.
The membrane seperates two chambers containing an ionic solution.
A voltage is applied, and the ionic current through the pore is recorded.
DNA is threaded through the pore, and partially blocks the ionic current.
The level of the ionic current ($e_j$) is due to six nucleotide words ($x_i$).

## Modeling Ionic Current with a hidden Markov model




A. Architecture of hidden Markov model used in this study. The match state 'M' (square) emits an event-6-mer pair and proceeds along the reference, Insert-Y 'Iy' (diamond) emits a pair but stays in place, and Insert-X 'Ix' (circle) proceeds along the reference but does not emit a pair. Two-level (B) and three-level (C) hierarchical Dirichlet process shown in graphical form. Circles represent random variables. The base distribution 'H' is a normal inverse-gamma distribution for both models. The Dirichlet processes '$G_0$', '$G_{\sigma n}$', and '$G_{\sigma ni}$' are parameterized by their parent distribution and shared concentration parameters '$\gamma_B$', $\gamma_M$', and $\gamma_L$'. The factors '$\theta$ji' specify the parameters of the normal distribution mixture component that generates observation 'xji'.

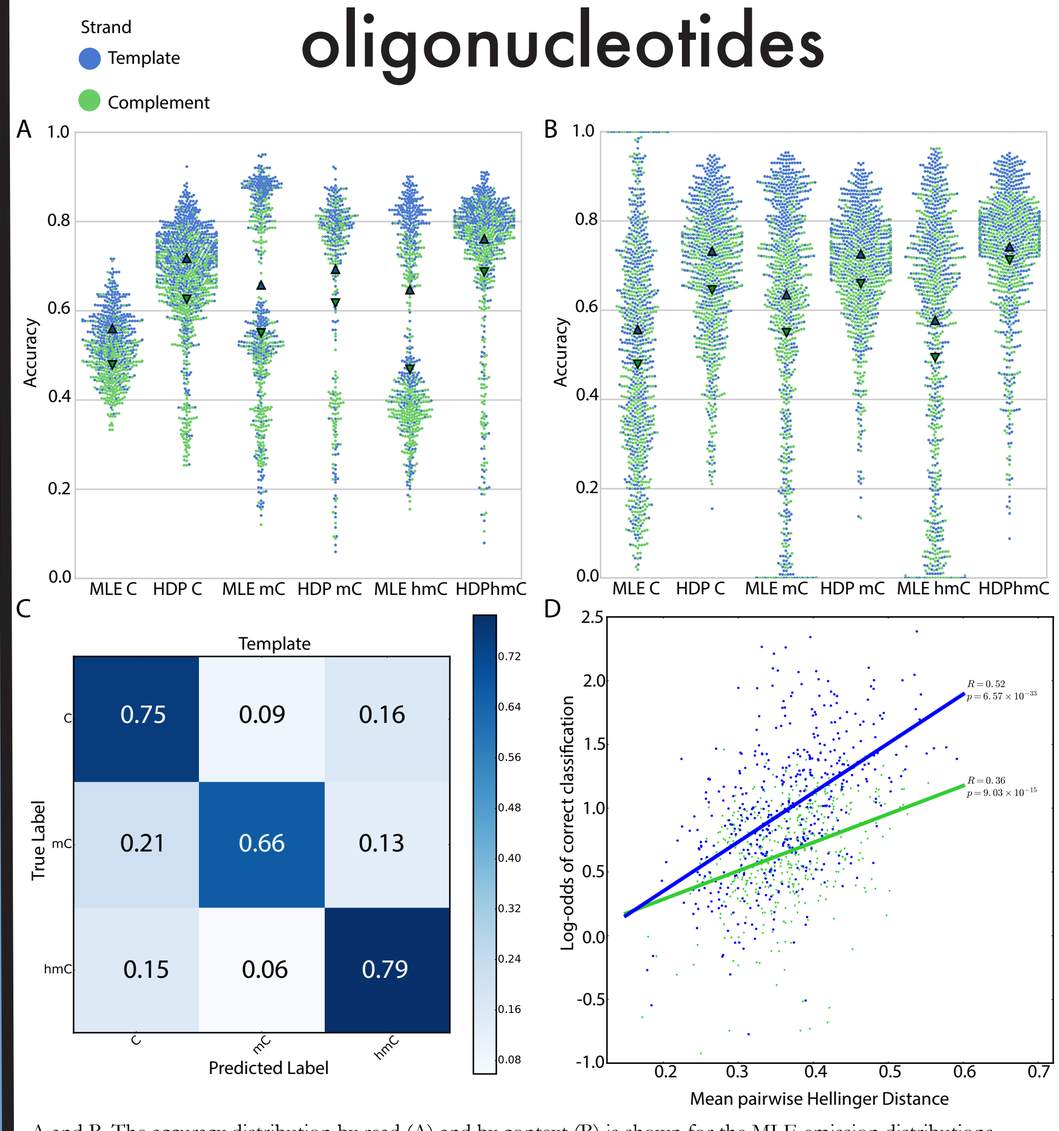
D. Variable-order HMM meta-structure over an example reference sequence. Each C* in the reference represents a potentially methylated cytosine. The structure expands around the C* base to accommodate for all possible methylation states. Each cell contains the three states shown in A, and transitions span between cells. The transitions are restricted so that methylation states are labeled consistently within a path. The match states are drawn with 4-mers for simplicity, but the model is implemented with 6-mers.

## Comparison of different HDP topologies

| | | Three-Way Accuracy | | |
|---|---|---|---|---|
| **Model** | Mean Accuracy (read) | Median Accuracy (read) | Mean Accuracy (site) | Median Accuracy (site) |
| **MLE** | 62% / 50% | 58% / 47% | 59% / 51% | 66% / 57% |
| **singlelevel** | 74% / 66 % | 79% / 72% | 73% / 63% | 76% / 69% |
| **multiset** | 74% / 67% | 80% / 76% | 73% / 67% | 76% / 70% |
| **composition** | 73% / 66% | 78% / 71% | 73% / 66% | 76% / 69% |
| **middleNts** | 72% / 64% | 76% / 69% | 72% / 64% | 75% / 67% |
| **group** | 73% / 65% | 78% / 71% | 72% / 66% | 75% / 69% |
| | | Two-Way Accuracy | | |
| **Model** | Mean Accuracy (read) | Median Accuracy (read) | Mean Accuracy (site) | Median Accuracy (site) |
| **singlelevel** | 83% / 78% | 86.5% / 84.5% | 82% / 77% | 83% / 78% |
| **multiset** | 83% / 78% | 86.5% / 84.5% | 82% / 77% | 83% / 78% |

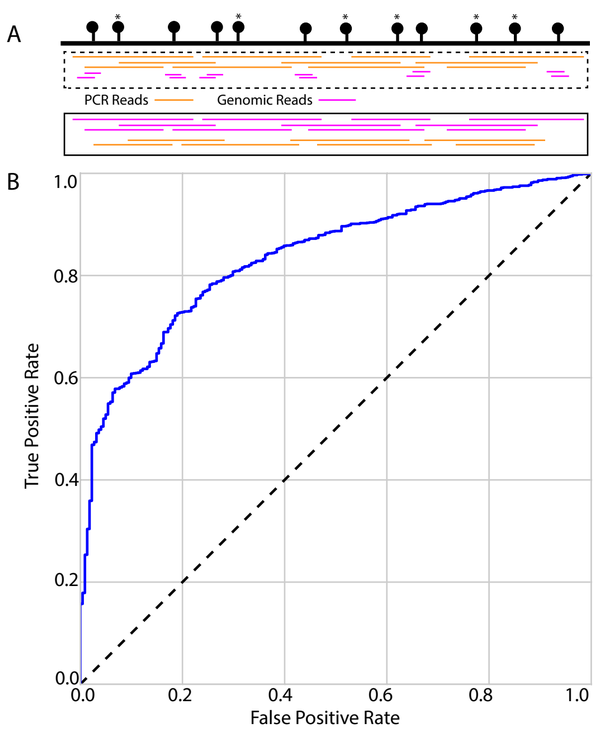
MLE is the maximum likelihood estimate of a normal distribution.
'Two-level' is an HDP model with no subgroupings of 6-mers, 'Multiset', 'Composition', 'MiddleNucleotides', and 'GroupMultiset' are three-level HDP models.
Three-way classification was performed between cytosine, 5-methylcytosine, and 5-hydroxymethylcytosine. Two-way classifications were between cytosine and 5-methylcytosine.

## Base modification calling accuracy results on synthetic oligonucleotides



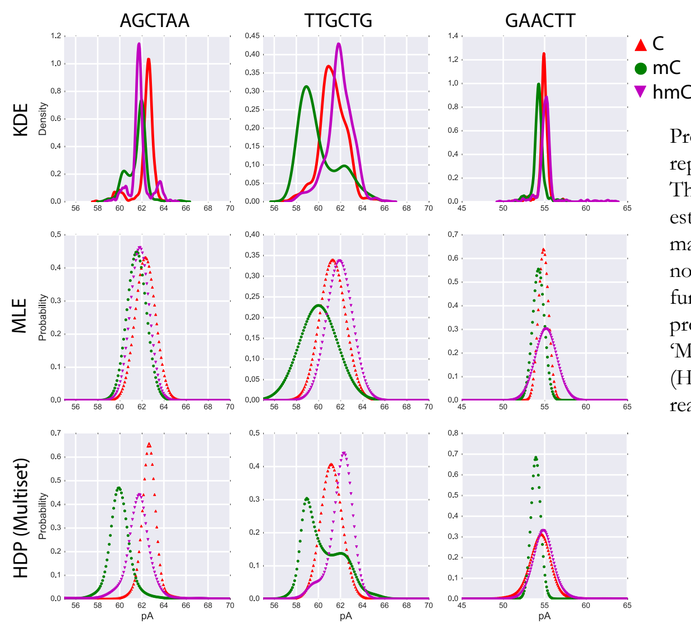

A and B. The accuracy distribution by read (A) and by context (B) is shown for the MLE emission distributions and the 'Multiset' HDP model on synthetic oligonucleotides. The triangles represent the mean of the distribution. C. Confusion matrix showing HMM-HDP three-way cytosine classification performance on template reads of synthetic oligonucleotides. D. Scatter plot shows the correlation between log-odds of correct classification and the mean pairwise Hellinger distance between the methylation statuses of the 6-mer distributions overlapping a cytosine.

## Mapping 5-methylcytosine in *E. coli* genomic DNA




A. Data partitioning for HDP training on E. coli. 1,709 high-confidence methylated CC*WGG sites (pins) were divided into training (unstarred) and test (starred). The HDP is trained on reads from PCR amplified DNA (orange lines) and events aligned to the training sites from genomic DNA reads (magenta lines). These combined data constitute the training dataset (dashed box). The trained model is then tested on genomic and PCR DNA reads aligned to the test sites from separate flow-cells. B. ROC plot shows HMM-HDP two-way classification performance on cytosines in test group (A, starred pins). Methylation calls are made by combining marginal probabilities from template and complement reads. Genomic reads were used to assess true positive rate, the PCR reads were used to assess the false positive rate.

## The HDP more realistically models ionic current distributions




Probability distributions for three representative 6-mers by multiple methods. The first row shows the kernel density estimate (KDE). The middle row shows maximum likelihood estimated (MLE) normal distribution probability density functions. The bottom row shows probability density functions from the 'Multiset' hierarchical Dirichlet process (HDP). All data shown are from template reads.