Currently preprint, 2023, pp. 1–12

doi:

PAPET: a collection of performant algorithms to identify 5-methyl cytosine from PacBio SequelII data

Romain Groux,[1] Ioannis Xenarios[1] and Emanuel Schmid-Siegert[1,*]

[1]JSR Life Sciences, NGS-AI division, Rue de la Corniche 3A, 1066, Canton de vaud, Switzerland
[*]Emanuel Schmid-Siegert. emanuel.schmid-siegert@selexis.com

Abstract

CpGs methylation is an important feature for the regulation of gene expression in vertebreate genomes. In this paper, we present the PAcBio Predicting Epigenetics Toolkit (PAPET) algorithms. PAPET is a collection of general algorithms and tools to train predictive models and predict epigenetics from SequelII data. This set of tools is worth for the PacBio user community to keep up with the fast evolving pace of PacBio's sequencing technology. We apply this framework to predict CpG methylation from SequelII data and demonstrate that the classifiers obtained compare equally with their best in class counterparts. PAPET is implemented in C++ to ensure resource efficiency and an easy scalability to large datasets. Moreover, PAPET is fully multi-threaded. The source code is available at https://github.com/ngs-ai-org/papet.

Key words: Genomics, PacBio, long read, algorithm, machine-learning, epigenitics, vaudois cooking, 5mC, methylation, CpG

Introduction

Cytosine followed by guanine (CpG) dinucleotides are an important feature of vertebrate genomes. They are palyndromic sequences, meaning that CpGs always occur by pairs on the opposite DNA strands. CpGs can be divided into two groups : CpG islands (CGIs) and individual CpGs.

CGIs are genomic regions with more CpGs than expected by chance and account for ∼5% of all CpGs in vertebrate genomes. The remaining ∼95% of CpGs occur individually in the genome (Vinson and Chatterjee, 2012).

CpGs are a predominant feature of the regulatory sequences. CpGs are often found in transcription factor binding sites. As such, CpGs are enriched in RNA polymerase II gene (RNAPII) promoters in vertebrates, with more than your sign in the code is ignored ∼50% of individual CpGs falling within. This is also true for CGIs with current evidences suggesting that CGIs can favor the recruitment of the RNAPII complexe by creating an open chromatin environment making such genes active as their default state (Deaton and Bird, 2011). It is estimated that ∼70% of RNAPII gene promoters in vertebrates contain CGI sequences, with an enrichment for housekeeping and developmental genes (Deaton and Bird, 2011; Lenhard *et al.*, 2012). Currently, CGIs seem to form the most prominent regulatory sequence feature in vertebrate genomes (Saxonov *et al.*, 2006; Lenhard *et al.*, 2012).

A CpG can undergo covalent modification, namely methyl deposition on the 5th atom of the cytosine ring leading to the creation of 5-methylcytosine (5mC) (Hotchkiss, 1948). This phenomenon is termed CpG methylation. Such CpG methylation is catalyzed by a dedicate set of DNA-methyltransferases. In human, Dnmt1 and Dnmt3 are responsible for it. Dnmt1 shows a strong specificity for hemi-methylated CpGs, ensuring that both strands of the DNA share the same methylation status while Dnmt3 is responsible for de novo CpG methylation (Breiling and Lyko, 2015). Demethylation on the other hand is performed through DNA replication and oxydation of 5mC into 5-hydroxymethylcytosine (5hmC) and further by Tet factors (Wu and Zhang, 2017).

5mc and CpG methylation are an important feature of regulation of gene expression. As mentioned above, if individual CpG and CGI tend to be positive regulators of gene expression, CpG methylation acts as a repressor. CGI methylation is thought to be the last step of Polycomb-mediated gene regulation, leading to hetero-chromatin compaction and enforcing a stable gene silencing (Deaton and Bird, 2011). On the other hand, it has been demonstrated that CpG methylation modulates the affinity of TFs for their binding sites, here again pointing to a functional role of CpG methylation in gene silencing (Héberlé and Bardet, 2019).

For these reasons, CpG methylation was coined an epigenetic modification. The definition of epigenetics being an open debate (Deans and Maggert, 2015), for this paper, we will align ourselves on the definitions proposed by (Holliday, 1994).

Genome bi-sulfite sequencing (WGBS) is currently considered as the gold standard when it comes to genome-wide CpG methylation measurements. Indeed, it has been adopted by major international efforts such as the ENCODE Consortium (Dunham *et al.*, 2012), the NIH Roadmap Epigenomics Consortium (Bernstein *et al.*, 2010), the Blueprint Project (Adams *et al.*, 2012) or the IHEC (IHEC, nd). CpG methylation

© To be announced

detection by WGBS relies on the conversion of unmethylated cytosine into uracil, leaving 5mC as is in the library. Upon sequencing, any CpG in the reads can be considered methylated. However, WGBS cannot differentiate between 5mC and other cytosine methylation forms such as 5hmC.

The emergence of the long read sequencing technologies Oxford Nanopore Technologies (ONT) and Pacific Biosciences (PacBio) has revolutionized the field of sequencing in general and the one of epigenetics analysis in particular. Not only because of the long reads it produces but also because of the so-called kinetic data produced. Kinetic data are information measuring the timing of the DNA polymerase, in the case of PacBio (Eid *et al.*, 2009) or that of a pore in the case of ONT (Deamer *et al.*, 2016). These kinetic data were early on shown to be usable to identify chemical modification of the DNA bases and also, unlike WGBS, to differentiate between different epigenetic marks (Flusberg *et al.*, 2010; Schreiber *et al.*, 2013; Jain *et al.*, 2016). Additionally, this information is provided without the need of extra DNA treatment steps, such as required by WGBS.

This lead to the creation of different types of CpG methylation prediction methods from raw long read sequencing data for both ONT (Simpson *et al.*, 2016; Stoiber *et al.*, 2017; Nanopore, 2022; Ni *et al.*, 2019; Liu *et al.*, 2019; Yuen *et al.*, 2021) and PacBio sequencing data (Suzuki *et al.*, 2016; Tse *et al.*, 2021; Ni *et al.*, 2022; PacBio, 2022b).

Long read sequencing technologies, PacBio and ONT, are evolving at fast pace to provide the scientific community with increased the performance. Any modification on the sequencer or the sequencing chemistry has the potential to modify the sequencing kinetics. Thus predictive models using the kinetic signal as predictive values must be updated accordingly. This was the case when the PacBio RSII system was replaced by the newest SequelII system. RSII users could rely on AgIn for CpG methylation predictions (Suzuki *et al.*, 2016) but newer SequelII users had no option, until dedicated softwares were released recently (Ni *et al.*, 2022; PacBio, 2022b).

In this paper, we describe the "PacBio Predicting Epigenetics Toolkit" (PAPET) algorithms. We present a full framework to model PacBio kinetic signal, train models and use them as predictors to detect CpG methylation in PacBio sequencing data and equip the PacBio community with the ability of creating tailored models and predictors.

Methods

The kinetic signal measured by PacBio sequencer around a CpG can be modelled as a sampling from a mixture of two kinetic signal classes: a class of methylated CpG signal and a class of unmethylated CpG signal. This means that if we are given kinetic signal class models and a kinetic signal, we can compute the probability that a signal was sampled from either of the two classes.

Let R^+ be a region of size $|R^+|$bp on the forward strand and centered on a CpG. R^+ is extended on each side of the C of the CpG by Lbp, resulting in $|R^+| = 2L+1$bp. This window contains all positions r_i^+ with $i \in (-L, -L+1, ..., 0, ..., L+1, L)$ where r_0^+ and r_{+1}^+ corresponding to C and G respectively. The same can be applied to the CpG on the reverse strand and call the region R^-. Because $|R^+| = |R^-|$, in the text below we will refer to the size of the region as $|R|$.

Let $C = \{c^1, c^2, ..., c^n\}$ be the set of CCSs mapping perfectly - without missmatch nor indel - over R^+ and R^-. Each CCS c^j has a length of $|c^j|$bp.

A CCS is a consensus sequence obtained from the sequencing of both strands of the DNA. Hence, it contains kinetics for both R^+ and R^-. Let $i^{j,+}$ be defined as the vector of IPD kinetics of c^j corresponding to the forward strand. It is possible to extract a sub-vector $i_{-L}^{j,+}, i_{-L+1}^{j,+}, ..., i_0^{j,+}, ..., i_{L+1}^{j,+}, i_L^{j,+}$ corresponding to the IPD kinetics mapping over R^+. The same can be done for the reverse strand kinetics $i^{j,-}$, resulting in a sub-vector $i_{-L}^{j,-}, i_{-L+1}^{j,-}, ..., i_0^{j,-}, ..., i_{L+1}^{j,-}, i_L^{j,-}$ corresponding to the IPD kinetics mapping over R^-.

As previous work has shown it (Suzuki *et al.*, 2016), opposite strand CpGs share a consistent methylation state, because of the action of Dnmt1. Thus, we decided to combine the information on both strands in order to improve the kinetic information. For this end, we simply decided to link each position r_i^+ with its counterpart on the other strand r_i^-. From here onward, CpG will refer to the pair of complementary CpGs unless stated otherwise.

This allows us to construct a pileup W of IPD kinetics sub-vectors. W has $2 * |C|$ rows and $|R|$ columns. The left-most column of W corresponds to the 5'-most IPD kinetic of each sub-vector and the right-most column to the 3'-most kinetic.

Computing the per column averages of W provides us with the strand-aggregated average IPD signal $\bar{i}$ measured for this CpG. The same can be done with the PWD kinetic values to obtain the strand-aggregated average PWD signal $\bar{p}$

Predicting the presence of 5mC

Now let us assume we are provided with two sets of $|R|$ distributions. Let $M_{IPD}^1 = \{\phi_{-L}^1, \phi_{L-1}^1, ..., \phi_0^1, ..., \phi_{L-1}^1, \phi_L^1\}$ and $M_{IPD}^2 = \{\phi_{-L}^2, \phi_{L-1}^2, ..., \phi_0^2, ..., \phi_{L-1}^2, \phi_L^2\}$ describe the distribution of IPD kinetics expected at every position in a window if the CpG is methylated or not, respectively. Similarly, we can define $M_{PWD}^1 = \{\theta_{-L}^1, \theta_{L-1}^1, ..., \theta_0^1, ..., \theta_{L-1}^1, \theta_L^1\}$ and $M_{IPD}^2 = \{\theta_{-L}^2, \theta_{L-1}^2, ..., \theta_0^2, ..., \theta_{L-1}^2, \theta_L^2\}$. This allows to compute the likelihood of IPD or PWD kinetics originating from each of the signal classes as:

$$\mathcal{L}_{IPD}^k = \prod_{i=1} p(\bar{i}_i | \phi_i^k) \tag{1}$$

$$\mathcal{L}_{PWD}^k = \prod_{i=1} p(\bar{p}_i | \theta_i^k) \tag{2}$$

where $k = 1$ for the methylated case and $k = 2$ for the unmethylated case. Because both $\bar{i}$ and $\bar{p}$ were computed from the same window R, we can compute the overall likelihood of observing these IPD and PWD kinetics originate from a given class as:

$$\mathcal{L}^k = \mathcal{L}_{IPD}^k * \mathcal{L}_{PWD}^k \tag{3}$$

In turn, we can compute the posterior probability of a window containing a methylated CpG using Bayes'theorem:

$$p(k = 1 | \bar{i}, \bar{p}) = \frac{p_1 * \mathcal{L}^1}{p_1 * \mathcal{L}^1 + (1 - p_1) * \mathcal{L}^2} \tag{4}$$

where p_1 is the a-priori probability of this CpG to be methylated. In other word, this is the probability that this window bears 5mC CpG given its kinetic signal.

Accounting for pair-wise dependencies

The above models assume that the kinetic signal at each position is independent from the kinetic signal at any other position in the window. To account for this, we modified the formulation of kinetic signal class models. Let them contain 2-dimensional distributions. Each distribution models the kinetic signal at position x in the window with respect to the kinetic signal at another position y in the window with $x \neq y$.

We designed two types of models: 1) "di-position" (DP) models accounting only for adjacent position dependencies and 2) "pair-wise" (PW) kinetic models, accounting for all possible pair-wise dependencies in the window.

DP models accounts for the set of position dependencies $D_{MD} = \{(-L, -L+1), (-L+1, -L+2), ..., (L-1, L)\}$, for a total of $|D_{MD}| = |R| - 1$ interactions. The models for the methylated and unmethylated classes are defined as the sets $MD^k_{IPD} = \{\phi^k_{-L,-L+1}, \phi^k_{-L+1,-L+2}, ..., \phi^k_{L-1,L}\}$ where $\phi^k_{x,y}$ designates the distribution of IPD kinetics, for class k, for position x with respect to position y in the window. Similarly, the models for the PWD kinetics are $MD^k_{PWD}\theta^k_{-L,-L+1}, \theta^k_{-L+1,-L+2}, ..., \theta^k_{L-1,L}\}$.

PW models accounts for the set of position dependencies $D_{MP} = \{(-L, -L+1), (-L, -L+2), ..., (-L, L), (-L+1, -L+2), ..., (L-1, L)\}$, for a total of $|D_{MP}| = \frac{|R|^2 - |R|}{2}$ interactions. The models for the methylated and unmethylated classes are defined as the sets $MP^k_{IPD} = \{\phi^k_{-L,-L+1}, \phi^k_{-L,-L+2}, ..., \phi^k_{-L,L}, \phi^k_{-L+1,-L-2}, ..., \phi^k_{L-1,L}\}$. Similarly, the models for the PWD kinetics are $MP^k_{PWD} = \{\theta^k_{-L,-L+1}, \theta^k_{-L,-L+2}, ..., \theta^k_{-L,L}, \theta^k_{-L+1,-L-2}, ..., \theta^k_{L-1,L}\}$.

From this, we can modify equations 1 and 6 as follows:

$$\mathcal{L}^k_{IPD} = \prod_{x,y \in D} p(\bar{i}_x, \bar{i}_y | \phi^k_{x,y}) \tag{5}$$

$$\mathcal{L}^k_{PWD} = \prod_{x,y \in D} p(\bar{p}_x, \bar{p}_y | \theta^k_{x,y}) \tag{6}$$

where D is D_{MD} and D_{MP} for DP and PW models respectively. The overall class likelihood and the probability of CpG methylation are computed using equations 3 and 4.

Model training

Training a methylated kinetic signal class model requires a set of CCSs in which we know the CpGs are methylated (positive sample). Similarly, training an unmethylated kinetic signal class model requires CCSs in which we have the certainty that there is no epigenetic DNA modification (negative sample). We then trained one model from the positive sample to estimate the signal distributions when there is CpG methylation and one model from the negative sample to estimate the signal distribution when there is no CpG methylation. The training process requires to map the CCSs against a genome of reference and to list all CpGs within this genome.

The following steps are applied to each CpG of the genome independently (Fig.S5). A window of size $|R|$bp, around a CpG, centered on the C, can be built. The sub-vector of IPD kinetics mapping inside the window are extracted for both strands independently. The kinetic pileup W can be build from these kinetic sub-vectors. The per-column average yields the average per-position IPD $\bar{i}$ for this window. Finally each value $\bar{i}_i$ is used to estimate the distribution ϕ^k_i inside M^k_{IPD}. The exact same procedure is used to estimate the distributions θ^k_i inside M^k_{PWD} for the PWD kinetic.

Training DP and PW models follows the same logic (Fig.S6 and Fig.S7) until the creation of the per-position average IPD and PWD kinetic signal $\bar{i}$ and $\bar{p}$. Each pair of values at positions $x, y \in D_{DP}$ for DP models and $x, y \in D_{PW}$ for PW models is considered and used to estimate the corresponding distribution $\phi^k_{x,y}$ and $\theta^k_{x,y}$ respectively.

Sequence normalization

Previous studies have shown that the sequence content itself, irrespective of DNA modifications, creates fluctuations in the signal (Flusberg *et al.*, 2010; Suzuki *et al.*, 2016). This is obviously a problem for our classifier. In order to account for this, we devised a kinetic normalization method. For this, we needed to know the kinetic signal that is expected if there was nothing else but the DNA sequence to dictate kinetic fluctuations.

Let N be a map with 4^K entries corresponding to every sequence S of length K. Each sequence S is mapped to its expected IPD and PWD kinetics $n_{S,IPD}$ and $n_{S,PWD}$ respectively of size K. Each vector containing the average expected IPD and PWD kinetic at each position of s. $n^1_{S,IPD}$ corresponds to the IPD value expected when sequencing the 1st position of S, $n^2_{S,IPD}$ to the 2nd and so one.

To normalize the IPD kinetics, we need to create two independent pileups W^+ and W^- with the IPD kinetics mapping over R^+ and R^- respectively.

Let S^+ and S^- be be the sequences of R^+ and R^- respectively. There are $|R| - K + 1$ IPD sub-vectors of length K inside R^+ and R^- respectively, starting at positions $j = -L, -L+1, ..., L-K+1$ with respect to the CpG. Each of these IPD sub-vector is associated with a corresponding sub-sequence starting at the same position j. In turn, N contains the expected IPD and PWD values for each given sub-sequence of length K. Thus, it is possible to normalize each IPD and PWD sub-vector, in both R^+ and R^-, by dividing them term by term with their expected counters parts.

To normalize a stretch of kinetics Z of length $|Z| > K$, each smaller K-mer is considered independently. Each K-mer kinetic signal is normalized and only the central kinetic, at position $\lfloor \frac{K}{2} \rfloor$, is considered. Consequently, the edges of Z cannot be normalized and the normalized stretch is smaller and has length $|Z| - 2\lfloor \frac{K}{2} \rfloor$.

Sequence model training

We trained the background sequence models using the CCSs from the negative sample. unlike for kinetic models, for sequence models, the CCSs don't need to be mapped. To train a model, we take each strand of a CCS c and breaks it into $|c| - K + 1$ slices of length K. The IPD and PWD slices are then grouped according to their corresponding sub-sequences. Finally, for each sub-sequence, the average IPD and PWD kinetic signal is computed with the corresponding IPD and PWD sub-slices. Overall, we trained three different sequence background models with K of 7, 9 and 11bp. Low values of K are not informative whereas high values of K result in less precise average estimates because each sub-sequence is present a limited number of time or even absent from the dataset.

Results

To investigate the effect of CpG methylation on the kinetic signal, we used a SequelII dataset generated by PacBio. This dataset was obtained from the hg002/NA24385 individual.

The dataset is composed of two samples. In order to create a negative sample, some genomic DNA was extracted and subjected to a whole genome amplification (WGA) in order to remove any CpG methylation. A positive sample was then generated by subjecting some of the resulting DNA to CpG specific 5mC deposition using the SssI enzyme and sequenced.

5mC kinetic signatures

We verified whether a difference of kinetic signal could be detected between the WGA and SssI samples, around CpGs. We opted for a reference based approach and mapped the CCSs against the hg38p8 genome. We created 31bp windows centered on every CpG of the genome and considered the CCSs showing a perfect alignment with a window. We extracted the IPD (or PWD) stretches corresponding to the windows. According to previous work, a 21bp large window should be sufficient to capture any signature (Suzuki *et al.*, 2016). We then constructed a kinetic signal pileup using all the IPD (or PWD) stretches extracted. For each CpG, the mean kinetic signal (IPD or PWD) was computed by constructing a pile-up using the kinetic stretches falling within the CpG window and computing the per-column averages. As expected, the presence of 5mC created specific IPD (Fig. 1A) and PWD signatures (Fig.1B).

We observed an increase of IPD values at position -1, +1, +3 and +7 and a decrease of the PWD signal at positions +1 and +2 relative to the CpGs in the positive sample. Interestingly, the signature was clearer when considering the per-CpG kinetic averages (Fig.1) rather than simply the CCS kinetics (Fig.S1). The second corresponds to a reference-free method where it suffices to identify the CpGs within the unmapped CCSs and extract the IPD and PWD kinetics falling within the 31bp windows. We think that these differences underline the benefit of mapping the CCSs to study CpGs methylation.

We then sought to verify if there was any inter-position dependencies within the window in the positive sample. We measured this by computing the correlation between the kinetic signal seen at any pair of positions within the window (Fig.2). This revealed strong IPD kinetic dependencies between positions +2 and +3 and +6 and +7, which were also important positions for the signature. There were also less strong correlations between many other positions. Intriguingly, we observed this between pairs of positions zipping away from one another (diagonals in the matrix). Dependencies were also observed for the PWD kinetics. Finally, the inter-dependencies were stronger when considering the per-CpG averages rather than simply CCS kinetics (Fig.S2), once again suggesting there is more information when using a reference-based approach.

Sequence background effect

We studied the effect of the sequence on the kinetic signals. For this, we considered all 7-mers. Then, using the unmethylated sample CCSs, we computed the average kinetic signal at each position of the 7-mers. If the sequence background does not influence the kinetics, one expects a somewhat uniform kinetic distribution among the positions and among the 7-mers. The results clearly indicated that the IPD and PWD kinetics were influenced by the sequence backgound (Fig.3A). This was even clearer for PWD kinetics (Fig.S3A). . This prompted us to account for this by performing a kinetic normalization to account for sequence specific biases. To this end, we decided to create so-called "sequence background models" (Fig.S4). The model contains the mean expected kinetic signal, without 5mC, for any possible k-mer. To build this model, we broke down all CCSs from the WGA sample into 7, 9 or 11bp slices, extracted the corresponding IPD or PWD values and computed the average per k-mer values (Fig.3A). When then developed a method to normalize the kinetic signal using these models (Fig.S4B). Applying the normalization to the per k-mer average kinetic dramatically decreased the variance of the kinetic both across positions and across k-mers (Fig.3D and Fig.S3C). Applying the same normalization to the kinetic signal within the 31bp windows around CpGs lead to improved 5mC signatures (Fig.3E and Fig.S3C).

Kinetic signal models

Based on these information, we decided to create classifiers in order to predict the presence of 5mC at CpGs. For each classifier, we trained a pair of kinetic signal models: one from the WGA sample and one from the SssI sample. Both models are intended to capture the characteristic kinetic signatures of CpGs without and with methylation, respectively.

We created models with two types of characteristics: 1) with and without kinetics normalization using a sequence background model and 2) with or without accounting for inter-position dependencies. For the first characteristic, we trained 7, 9 and 11bp sequence background models from the WGA sample. For the latter, we implemented models without inter-dependencies or with the DP and PW model logic. We then combined all possible pairs of characteristics, for a total of 12 model pairs.

We decided to model the signal within a 17bp window (+/-8bp around the C of the CpG). We deemed this window size sufficient to capture most of the kinetic signature information since the most important distribution differences were observed between positions -1 and +7 (Fig.1 and Fig.2).

We modelled the kinetic signal distributions empirically and used 1-dimension histograms for models without inter-dependencies and 2-dimensional histograms otherwise.

CpG methylation prediction power

In order to assess the ability of the models to perform CpG methylation prediction, we used the Telomere-To-Telomere Consortium (T2T) generated data. This dataset contains SequelII data with a genome-wide average 42X coverage (Vollger *et al.*, 2020; Nurk *et al.*, 2020) as well as a methylation track constructed from ONT sequencing data (Gershman *et al.*, 2022). Recent work has demonstrated that ONT data can be used to confidently predict CpG methylation [REFERENCE]. Thus, we decided to use this track as our truth set. As genome reference, we decided to use the CHM13v1.0 genome assembly by T2T (Nurk *et al.*, 2022). To date, this assembly is the most complete human reference, with the most comprehensive delineation of complex/repeated regions.

We listed all the CpGs present in the CHM13v1.0 genome. For each CpG, we extracted the kinetics falling within a 17bp window centered on the CpG from every CCS mapping perfectly in the window. We then computed the window average IPD and PWD values and fed them to the predictors.

As a comparison, we also run predicitons using PacBio's pb-CpG-tools for CpG methylation prediction. If a CpG was assigned a probability of methylation larger than 50%, it was considered methylated and unmethylated otherwise. We then compared the predictions of our predictors and of pb-CpG-tools against T2T methylation track (Fig.4). PacBio's pb-CpG-tools achieved an impressive Area Under the Curve (AUC) of 0.97

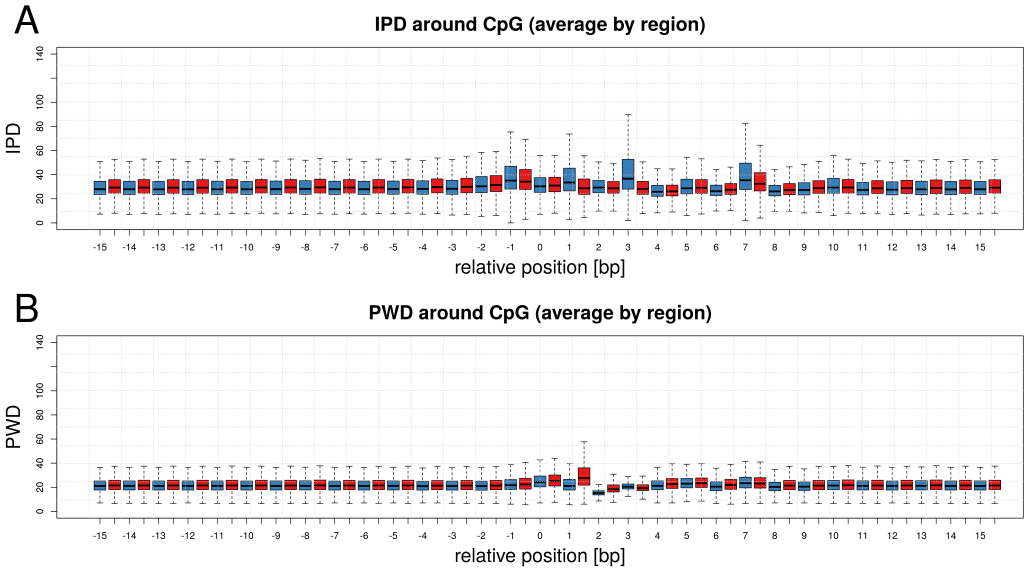


Fig. 1. Kinetic signal around CpGs A: Distribution of IPD signal in a +/-15bp window centered on all CpGs located on chromosome 1 (C is located at position +/-0). The IPD signal was extracted in a per CpG manner. For a given CpG, all CCSs mapping perfectly over the CpG window were considered. The IPD signal was extracted from the positions within the CCS corresponding to the window and the per position averages were computed. This process was repeated for all CpG on chromosome 1 and the distributions of values is displayed per position. The red and blue distributions indicate the WGA and SssI samples respectively. **B**: Same as A) but with the PDW signal.

with a high sensitivity (also called recall). Our models showed a wide variety of performances. The kinetic normalization clearly improved the performances, bringing an AUC of 0.83 for simplest model to an AUC of 0.93 when adding an 11bp background sequence normalization. Accounting for inter-position dependencies improved the specificity of the models, with all our models achieving a higher sensitivity than pb-CpG-tools. The best model on our side (11bp normalization and di-position modelling) achieved an AUC of 0.93, lower thanpb-CpG-tools, yet showing a superior specificity and precision with 0.94 and 0.93, respectively. This indicate that PacBio's predictor is better at finding methylated CpG but at the expense of falsely calling unmethylated CpG methylated.

Finally, we assessed the impact of coverage on the prediction performances of the best models (pb-CpG-tools and the 11bp normalized accounting for di-position models). For this, we down-sampled mapped CCSs down to a coverage of 20X, 10X, 5X, 2X and 1X and ran the predictions (Fig.5). Overall, the same trend was observed with decreasing coverage. pb-CpG-tools model showed better AUC and specificity but had a lower sensitivity and precision. Our model started showing a better AUC with extremely sparse sequencing coverage of 2X and 1X.

Discussion

In this paper, we showed that 5mC creates specific IPD and PWD kinetic signatures around CpGs (Fig.1). We also showed that specific position around the CpG resonate together in the kinetic signatures (Fig.2). The sequence background has a strong influence on the kinetics (Fig.3A and Fig.S3A). From this, we have designed a framework to model this signal and train the models from the data. Finally, we designed several types of signal models : normal models, DP models accounting for adjacent position inter-dependencies and PW models accounting for any two position inter-dependencies. Atop of this, we designed a way to normalize the signal for each model to account for the effect of the sequence on the kinetics.

PacBio's tool showed a better ability to predict the presence of 5mC at CpG, at the expense of falsly calling 5mC where there is none. However, our model showed a better propensity to distinguish unmethylated CpGs and made less mistakes than PacBio's tool when predicting 5mC. Overall, these models show an unexpected complementary behaviour.

We think there is still room to improve the predictor performances through the optimization of the model and parameters. For instance, different window size may provide better results. The same is true for the prior probability of a CpG to be methylated. In this work, we used a neutral choice of 0.5. However there may be situations in which adapting it can turn useful, for instance in case of deficient methylation deposition pathways, or on a specie-specific basis.

The approach proposed in this paper also has limitations. First, the 5mC is not really removed from the DNA in the WGA sample but is rather diluted out. Some CpG bearing 5mC, most likely on a single strand. Similarly, the SssI mediated methylation efficient is only $\tilde{9}5\%$ (Simpson *et al.*, 2016) and some CpG are thus left unmethylated. Consequently, the models that we trained contains a biased, mixture of methylated and unmethylated which may impact the classification performances. This can lead to discrepancies compared with WGBS inferred methylation. Second, the presented approach is not designed to be allele-aware. Therefore it does not distinguish potential alleles and likely miss-classifies CpG falling in differently methylated alleles by providing an intermediate score.

Finally, we think that this framework and this model are general enough to allow a wide variety of usages. First, no specie specific parameters that have the ability to modify the kinetic signal were introduced in the models. Thus PAPET could be used for non-human data as well. Second, this framework and

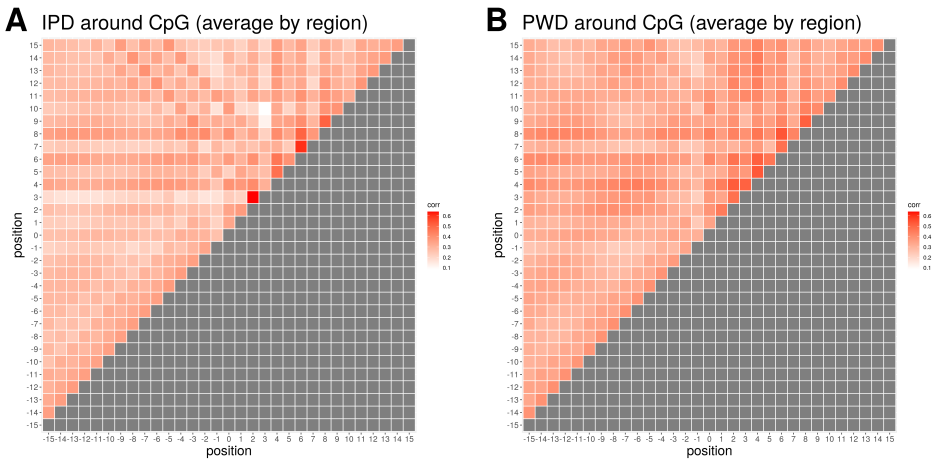


Fig. 2. Kinetic signal correlation around CpGs A: Correlation of the IPD signal, averaged per CpG, between all pairwise positions within a +/-15bp window centered on CpGs (C is located at position +/-0) from chromosome 1, 2 and 3, from the positive sample. For a given CpG, all CCSs mapping perfectly over the CpG window were considered. The IPD signal was extracted from the positions within the CCS corresponding to the window and the per position averages were computed. **B**: Same as A) but with the PDW signal.

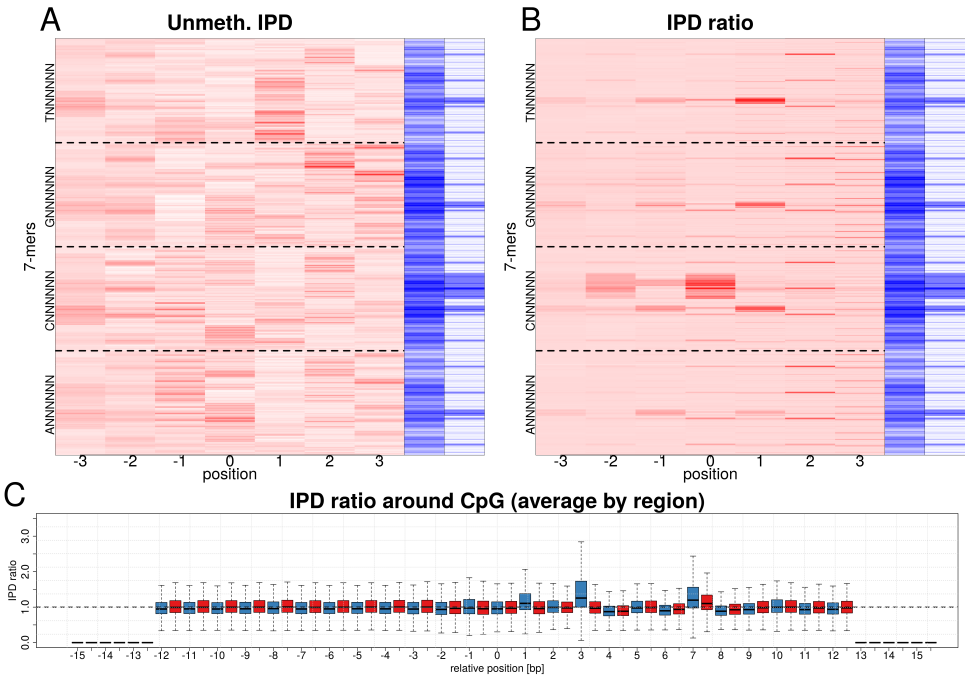


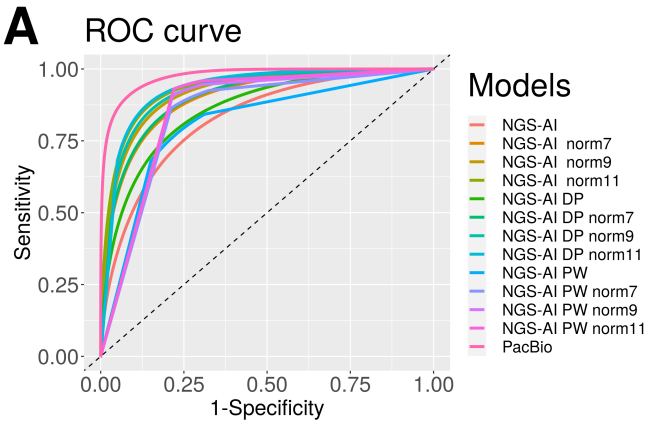
Fig. 3. Effect of the sequence on the IPD signal A: Heatmap of the mean IPD signal at each position of all possible 7-mers, from the WGA sample CCSs. The corresponding sequences and IPD signals were extracted and the mean IPD signal was computed for each 7-mer. The 7-mers are sorted in lexicographic order from bottom to top. The dashed lines indicate the limits between A, C, G and T starting 7-mers. The blue bars indicate the 7-mer CG content (center) and CpG content (right) respectively. **B**: Heatmap of the mean IPD ratios (IPDr) at each position of all possible 7-mers. The mean per 7-mer IPD values were computed from the SssI sample CCS as describe in A). Each 7-mer SssI mean IPD was then divided by the corresponding 7-mer WGS average. The ordering, dashed lines and blue bars are the same as in A). **C** Distribution of IPDr signal in a 31bp window centered on all CpGs located on chromosome 1 (C is located at position +/-0). The IPDr signal was extracted in a per CpG manner. For a given CpG, all CCSs mapping perfectly over the CpG window were considered. The IPD signal was extracted from the positions within the CCS corresponding to the window, normalized using a 7-mer backgound model and the per position averages were computed. This process was repeated for all CpG on chromosome 1 and the distributions of values is displayed per position. The first and last six positions in the window have no signal because the normalization process trims the signal, as explained above. The red and blue distributions indicate the WGA and SssI samples respectively.

tools could be used to train a predictor against any other epigenetics marks, such as N6-methyladenosine if comparable data are available.

Material and Methods

Data

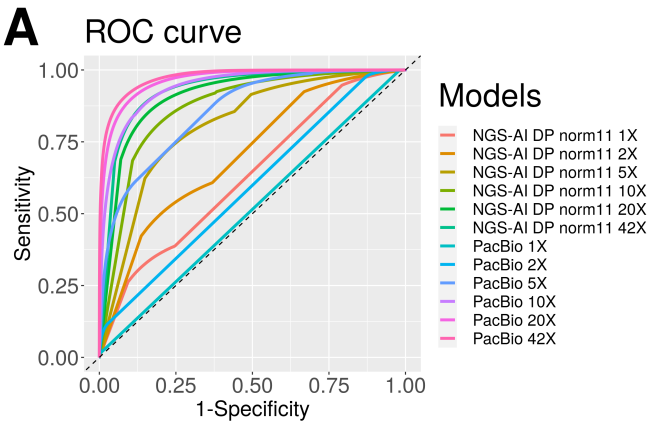
For this study, we used two publicly available datasets.



B

	AUC	Sensitivity	Specificity	Recall	Precision	F1
NGS-AI	0.83	0.74	0.76	0.74	0.77	0.75
NGS-AI norm7	0.9	0.71	0.9	0.71	0.88	0.79
NGS-AI norm9	0.92	0.71	0.93	0.71	0.92	0.8
NGS-AI norm11	0.93	0.71	0.95	0.71	0.94	0.81
NGS-AI DP	0.86	0.79	0.77	0.79	0.79	0.79
NGS-AI DP norm7	0.9	0.75	0.89	0.75	0.88	0.81
NGS-AI DP norm9	0.92	0.76	0.92	0.76	0.91	0.83
NGS-AI DP norm11	0.93	0.77	0.94	0.77	0.93	0.84
NGS-AI PW	0.81	0.78	0.77	0.78	0.78	0.78
NGS-AI PW norm7	0.84	0.74	0.9	0.74	0.88	0.8
NGS-AI PW norm9	0.86	0.74	0.93	0.74	0.92	0.82
NGS-AI PW norm11	0.86	0.74	0.95	0.74	0.94	0.83
PacBio	0.97	0.96	0.84	0.96	0.86	0.91

Fig. 4. Classification performances A: the ROC curves for each model. The classification performances were assessed using T2T methylation track as truth set. The models labelled as "norm-7", "norm-9" and "norm-11" used a sequence normalization background with 7-mers, 9-mers and 11-mers, respectively. Models labelled as "DP" and "PW" accounts for direct neighbour dependencies and all possible pairwise dependencies within the window, respectively. **B**: The detailed performance statistics per model.



B

	AUC	Sensitivity	Specificity	Recall	Precision	F1
NGS-AI DP norm11 1X	0.63	0.75	0.39	0.75	0.57	0.65
NGS-AI DP norm11 2X	0.7	0.63	0.61	0.63	0.63	0.63
NGS-AI DP norm11 5X	0.81	0.56	0.86	0.56	0.8	0.66
NGS-AI DP norm11 10X	0.87	0.62	0.92	0.62	0.89	0.73
NGS-AI DP norm11 20X	0.91	0.71	0.93	0.71	0.92	0.8
NGS-AI DP norm11 42X	0.93	0.77	0.94	0.77	0.93	0.84
PacBio 1X	0.51	0.99	0.02	0.99	0.52	0.68
PacBio 2X	0.59	0.97	0.12	0.97	0.54	0.69
PacBio 5X	0.86	0.88	0.61	0.88	0.71	0.78
PacBio 10X	0.94	0.89	0.83	0.89	0.85	0.87
PacBio 20X	0.96	0.94	0.84	0.94	0.86	0.9
PacBio 42X	0.97	0.96	0.84	0.96	0.86	0.91

Fig. 5. Effect of the coverage on the classification performances A: classification performances achieved by the best models according to the coverage. The classification performances were assessed using T2T methylation track as truth set. The whole dataset (42X) was down-sampled in order to obtain global average coverage of 1X, 2X, 10X, and 20X. A: the ROC curves for each model, with each coverage. The models labelled as "norm-7", "norm-9" and "norm-11" used a sequence normalization background of 7-mers, 9-mers and 11-mers respectively. Models labelled as "DP" accounts for direct neighbour dependencies within the window respectively. **B**: The detailed performance statistics per model.

The first dataset was created by PacBio. It originates from the hg002/NA24385 individual. It is composed of 2 samples, referred to as WGA and SssI samples. These samples contain 4 and 8 SMRT-cell each. The WGA sample contains genomic DNA that was subjected to WGA prior being sequenced on a PacBio SequelII sequencer. The positive sample is derived from the WGA sample by subjecting it to a SssI enzyme treatment prior to sequencing on a PacBio SequelII sequencer. The 8 resulting SMRT-cells are listed in the table below. The HiFi CCS reads were downloaded from `TODO`.

The second dataset was created as part of the T2T Consortium work. It originates from the human CHM13 cell line. The genomic DNA was extracted from the cells and sequenced with a PacBio SequelII sequencer. It is composed of 3 Hifi SMRT-cells. The SMRT-cells are listed in the table below. The SMRT-cell are available in the SRA archive under accession number SRR11292120, SRR11292121 and SRR11292123. A corresponding CpG methylation track derived from ONT sequencing data for the CHM13v1.0 assembly was downloaded from here. The bigWig file was then converted to bedgraph with

```
bigWigToWig methyFreq.bigWigmethyFreq.bedgraph
```

Data owner	SMRT-cell ID	origin	treatment
Pacific Biosciences	m64168_200820_000733	hg002	WGA
Pacific Biosciences	m64168_200822_124644	hg002	WGA
Pacific Biosciences	m64169_200821_062603	hg002	WGA
Pacific Biosciences	m64169_200823_191821	hg002	WGA
Pacific Biosciences	m64168_200821_062040	hg002	SssI
Pacific Biosciences	m64168_200823_191315	hg002	SssI
Pacific Biosciences	m64169_200820_001235	hg002	SssI
Pacific Biosciences	m64169_200822_125206	hg002	SssI
Pacific Biosciences	m64169_200827_002456	hg002	SssI
Pacific Biosciences	m64169_200828_063824	hg002	SssI
Pacific Biosciences	m64169_200829_130438	hg002	SssI
Pacific Biosciences	m64169_200830_193054	hg002	SssI
Pacific Biosciences	m64168_200821_062040	hg002	SssI
T2T Consortium	m64062_190803_042216	CHM13	None
T2T Consortium	m64062_190804_172951	CHM13	None
T2T Consortium	m64062_190806_063919	CHM13	None

Extraction of PacBio CCS kinetic data for analysis

The negative and positive sample CCSs were mapped against hg38p8 genome assembly with `pbmm2` (PacBio, 2022a) using:

```
pbmm2 align --preset CCS \
 --sample <sample-name> \
 -j 30 --log-level INFO \
 <file.in.bam> \
 <file.mmi> \
 <file.out.bam>
```

where `<sample-name>` is the sample label, `<file.in.bam>` is the path to the input bam file, `<file.mmi>` is the path to the genome index file and `<file.out.bam>` is the path to the output bam file.

The coordinates of the CpGs in the hg38p8 genome were found with `getCpG.py`, a program of ours, using:

```
python3.6 getCpG.py --fasta hg38\_p8.fa | \
 sort -k1,1V -k2,2n -k 3,3 \
     > hg38\_p8\_cpg.bed
```

Then kinetic signal of each CCS, around the CpG sites, were extracted with `papet kinetics`, a program of ours, using:

```
papet kinetics \
 --bed hg38\_p8\_cpg.bed \
 --bam <file_list> \
 --winSize 31 > kinetics.tsv
```

where `<file\_list>` is a coma-separated list of bam files containing the CCS from which the kinetic will be extracted. `papet kinetics` was run once with all the negative sample bam files and once with all the positive sample bam files.

The same procedure was used to extract normalized kinetic signal:

```
papet kinetics \
 --bed hg38\_p8\_cpg.bed \
 --bam <file_list> \
 --winSize 31 \
 --model <file\_model\_bckg> > kinetics\_normalized.tsv
```

where `<file\_model\_bckg>` contains the background sequence model to use to normalize the kinetics. Three different models were used : 7bp 9bp and 11bp.

Background sequence training

The background sequence models were trained using the negative sample data with `papet model-sequence`, a program of ours, using:

```
papet model-sequence \
 --bam <file_list> \
 --kmer <k> \
 --out <file\_model\_bckg>
```

where `<file\_list>` is a coma-separated list of bam files containing non-methylated CCS, `<k>` is the kmer size to use and `<file\_model\_bckg>` is the path where the sequence background model will be written.

Kinetic model training

To train the kinetic signal model, we used the Pacific Bioscience dataset. The models were always created by pairs. One model from the negative sample data describing the expected signal without CpG methylation and one model from the positive sample data describing the expected signal with CpG methylation.

The simplest models, neither accounting for inter-position dependencies nor the DNA sequence effect, were trained with `papet model-kinetic`, a program of ours, using:

```
papet model-kinetic \
 raw \
 --bed hg38\_p8\_cpg.bed \
 --bam <file\_list> \
 --size 17 \
 --nbin 952 \
 --xmin 0 \
 --xmax 952 \
 --pseudocount 1 \
 --out <file_model>
```

where `<file\_list>` is a coma separated list of all bam file paths from either the negative or positive sample, `--nbin` is the number of bins inside the histograms, `--xmin` is the lower boundary of the lower bin in the histograms, `--xmax` is the upper boundary of the upper bin in the histograms, `--pc` is a pseudocount value added once to each histogram bin to avoid 0 count bins and `<file_model>` is the path where the kinetic model will be written. `--nbin`, `--xmin` and `--xmax` were set to 952, 0 and 952 because raw IPD and PWD values can take any integer value from 0 to 952.

Normalized kinetic models, accounting for the DNA sequence effect only, were trained with `model-kinetic`, using:

```
papet model-kinetic \
 raw-norm \
 --bed hg38\_p8\_cpg.bed \
 --bam <file\_list> \
 --background <file\_model\_bckg> \
 --size 17 \
 --nbin 952 \
 --xmin 0 \
 --xmax 5 \
 --pseudocount 1 \
 --out <file_model>
```

where `<file\_model\_bckg\>` is the path to a sequence background model created with `model-sequence`. We used three different sequence models : 7bp, 9bp and 11bp. The values of `--xmin` and `--xmax` were set to 0 and 5 because normalized IPD and PWD values cannot be smaller than 0 and were mostly smaller than 5. In addition, two bins per histogram dimension

were automatically added to cover the inverals]−∞,0] and]5,∞]. Finally, `--nbin` was set to 952 in order to obtain the same level of granularity as with the simplest model.

PD kinetic models, accounting for neighbouring position inter-dependencies only, were trained with `model-kinetic`, using:

```
model-kinetic \
 diposition \
 --bed hg38\_p8\_cpg.bed \
 --bam <file\_list> \
 --size 17 \
 --nbin 100 \
 --xmin 0 \
 --xmax 100 \
 --pseudocount 1 \
 --out <file_model>
```

The value of `--xmin`, `--xmax` `--nbin` were set to 0, 100 and 100 in order to limit the number of bins. Since these histograms are 2-dimensional, they contain 100 ∗ ∗2 bins. Since both the IPD and the PWD are smaller than 100, we created integer bins from 0 to 100. In addition, two bins per histograms were automatically added to account for the values passed the min and max of the histograms.

Normalized PD kinetic models, accounting for neighbouring position inter-dependencies and DNA sequence effect, were trained with `model-kinetic`, using:

```
model-kinetic \
 diposition-norm \
 --bed hg38\_p8\_cpg.bed \
 --background <file\_model\_bckg> \
 --bam <file\_list> \
 --size 17 \
 --nbin 100 \
 --xmin 0 \
 --xmax 5 \
 --pseudocount 1 \
 --out <file_model>
```

where `<file\_model\_bckg\>` is the path to a sequence background model created with `model-sequence`. We used three different models: 7bp, 9bp and 11bp. The values of `--xmin` and `--xmax` were set to 0 and 5 because normalized IPD and PWD values cannot be smaller than 0 and were mostly smaller than 5. In addition, two bins per histogram dimension were automatically added to cover the inverals]−∞,0] and]5,∞]. Finally, `--nbin` was set to 100 in order to limit the number of bins per histogram. Indeed these histograms are 2-dimensional. They contain 100 ∗ ∗2 bins.

PW kinetic models, accounting for all position inter-dependencies only, were trained with `model-kinetic`, using:

```
model-kinetic \
 pairwise \
 --bed hg38\_p8\_cpg.bed \
 --bam <file\_list> \
 --size 17 \
 --nbin 100 \
 --xmin 0 \
 --xmax 100 \
 --pseudocount 1 \
 --out <file_model>
```

The value of `--xmin`, `--xmax` `--nbin` were set to 0, 100 and 100 in order to limit the number of bins. Since these histograms are 2-dimensional, they contain 100 ∗ ∗2 bins. Since both the IPD and the PWD are smaller than 100, we created integer bins from 0 to 100. In addition, two bins per histograms were automatically added to account for the values passed the min and max of the histograms.

Normalized PW kinetic models, accounting for all position inter-dependencies and DNA sequence effect, were trained with `model-kinetic`, using:

```
model-kinetic \
 pairwise-norm \
 --bed hg38\_p8\_cpg.bed \
 --background <file\_model\_bckg> \
 --bam <file\_list> \
 --size 17 \
 --nbin 100 \
 --xmin 0 \
 --xmax 5 \
 --pseudocount 1 \
 --out <file_model>
```

where `<file\_model\_bckg\>` is the path to a sequence background model created with `model-sequence`. We used three different models : 7bp, 9bp and 11bp. The values of `--xmin` and `--xmax` were set to 0 and 5 because normalized IPD and PWD values cannot be smaller than 0 and were mostly smaller than 5. In addition, two bins per histogram dimension were automatically added to cover the inverals]−∞,0] and]5,∞]. Finally, `--nbin` was set to 100 in order to limit the number of bins per histogram. Indeed these histograms are 2-dimensional. They contain 100 ∗ ∗2 bins.

CpG methylation predictions

For this part, we used the CHM13 SequelII whole genome dataset generated by the T2T consortium. We downloaded the subreads corresponding to entries SRR11292120 (m_64062_190806_063919), SRR11292121 (m_64062_190803_042216) and SRR11292123 (m_64062_190804_172951) on the SRA and also available on T2T github page.

Fist, we computed the CCSs from the subreads with PacBio's `ccs` using: `ccs -j 4 \`

```
 --all \
 --hifi-kinetics \
 --report-file ccs_report.txt \
 --report-json ccs_report.json \
 --metrics-json ccs_metric.json\
 subreads.bam \
 ccs.bam\
```

where `subreads.bam` is the path to the file containing the subreads and `ccs.bam` is the path to the file that will contain the CCSs. This step was performed for each subread file independently.

Then, the CCSs were mapped against CHM13 genome assembly v1.0 released by T2T with PacBio's pbmm2 using:

```
pbmm2 \
 align \
 --preset CCS \
 --sample t2t-1 \
 -j 30 \
 --log-level INFO \
 ccs.bam \
 chm13.draft_v1.0.mmi \
 ccs.chm13v1_0.bam \
```

where ccs.bam is the path to the file containing the CCSs, chm13.draft_v1.0.mmi is the path to the indexed genome and ccs.chm13v1_0.bam the path to the file that will contain the mapped CCSs. This step was performed for each subread file independently. Then all mapped CCSs were pooled together with samtools (Li *et al.*, 2009) using:

```
samtools merge all.bam \
 m64062_190803_042216.ccs.chm13v1_0.bam \
 m64062_190804_172951.ccs.chm13v1_0.bam \
 m64062_190806_063919.ccs.chm13v1_0.bam ; \
samtools sort -o all.sorted.bam
```

The coordinates of the CpGs in the hg38p8 genome were found with our program getCpG.py, using:

```
python3.6 getCpG.py --fasta chm13.draft_v1.0.fasta | \
 sort -k1,1V -k2,2n -k 3,3 \
     > chm13.draft_v1.0.cpg.bed
```

where chm13.draft_v1.0.cpg.bed is the file containing the CHM13 v1.0 genome assembly chromosome sequences and chm13.draft_v1.0.cpg.bed the file containing the genomic coordinates of the CpGs found in the genome.

The predictions using our models were computed with papet predict, a program of ours, using:

```
papet predict \
 --bam all.sorted.bam \
 --bed chm13.draft_v1.0.cpg.bed \
 --modelMeth <file_model_meth> \
 --modelUnmeth <file_model_meth> \
 --prob 0.5 \
 > <file_pred>\
```

where <file_model_meth> and <file_model_unmeth> are two files containing models for methylated and unmethylated kinetic signals respectively. Predictions were run using the simplest, the DP, the DP normalized, the PW and the PW normalized models independently.

Predictions using PacBio's pb-CpG-tools were computed using:

```
primrose \
 --keep-kinetics \
 --num-threads 30 \
 all.sorted.bam \
 all.sorted.primrose.bam && \
samtools index \
 -@ 30 \
 all.sorted.primrose.bam && \
python /pb-CpG-tools/aligned_bam_to_cpg_scores.py \
 -b all.sorted.primrose.bam \
 -f chm13.draft_v1.0.fasta \
 -o <label> \
 --pileup_mode model \
 --model_dir /pb-CpG-tools/pileup_calling_model \
 --modsites reference
```

where ccs.bam is the path to the file containing all CCSs, all.sorted.primrose.bam is the file containing the per read methylation predictions computed by primrose, chm13.draft_v1.0.fasta is the file containing CHM13 genome assembly v1.0 as released by T2T. We also tried to run CpG methylation predictions using ccsmeth (Ni *et al.*, 2022). Unfortunately, we were not successful in running this software.

CCS sampling

CCS down sampling to achieve lower genome coverage were performed using:

```
sambamba view \
 -f bam \
 -h \
 -s <ratio> \
 --subsampling-seed <seed> \
 all.sorted.primrose.bam > all.sorted.sub.bam \
```

where <ratio> is a ratio corresponding to the final desired coverage. The 3 pooled SMRT-cells accounted for a total of 42x genome-wide average coverage. In order to sample down to 20x, 10x, 5x, 2x and 1x we used 20/42, 10/42, 5/42, 2/42 and 1/42 respectively. <seed> is a seed to set random processes for reproducibility purposes.

Predictor performances

All prediction files were merged into one file containing, for each CpG, all the predictions computed. If a predictor did not produce a prediction for a CpG while at least one other did, the value was set to NA for the missing predictors. For this, we used:

```
python3.6 bedToTable.py <files_pred> > tmp.bed &&
sort -k1,1 -k2,2n -k3,3n  tmp.bed > tmp.sorted.bed &&
mv tmp.sorted.bed tmp.bed &&
bedtools map \
    -a tmp.bed \
    -b methyFreq.bedgraph" \
    -c 4 -o mean > tmp.tmp.bed &&
sed -E 's/\.$/NA/' tmp.tmp.bed > tmp.bed &&
rm tmp.tmp.bed &&
bedtools closest \
    -D ref -io -a tmp.bed -b tmp.bed \
    cut -f ${echo {1..103} | sed 's/ /,/g'},207\
    > chm13.draft_v1.0_cpg_5mC.bed
```

where bedToTable.py is a program of ours, <files_pred> is a coma-separated list of all prediction files created by <CpGMethylator> and pb-CpG-tools and methyFreq.bedgraph is the methylation track released by T2T for CHM13 v1.0. T2T methylation track was considered as the truth set.

The table was the imported in R. Any CpG for which no truth set methylation value was available were discarded. CpGs with at least one missing prediction were also discarded. The remaining CpGs were separated into methylated and unmethylated based on the truth set methylation probability. If the probability was > 0.5, the CpG was considered methylated. The CpG was considered unmethylated otherwise. The classifier performances were then measured with the precrec R library (Saito and Rehmsmeier, 2017) using:

```
pred = mmdata(scores=data,                labels=labels)
auc  = evalmod(pred)
```

where data is a dataframe of methylation prediction probabilities with the CpGs on the rows and the predictors on the columns. Labels is a vector of boolean indicating whether a CpG is considered methylated or not according to the truth set.

Competing interests

All authors are employed by JSR Life Sciences incorporated.

Author contributions statement

R.G. designed and implemented the algorithms. R.G. and E.S ran the benchmark and analysed the results. R.G. wrote the manuscript. R.G., I.X. and E.S reviewed the manuscript.

Acknowledgments

The authors thank Richard Hall and the Pacific Bioscience bioinformatics team for sharing their WGA/Sssl training dataset. Also, the authors express their gratitude to the first author's mom for sharing the best recipe of papet ever.

References

Adams, D. *et al.* (2012). BLUEPRINT to decode the epigenetic signature written in blood. *Nature Biotechnology*, **30**(3), 224–226. Number: 3 Publisher: Nature Publishing Group.

Bernstein, B. E. *et al.* (2010). The NIH Roadmap Epigenomics Mapping Consortium. *Nature Biotechnology*, **28**(10), 1045–1048. Number: 10 Publisher: Nature Publishing Group.

Breiling, A. and Lyko, F. (2015). Epigenetic regulatory functions of DNA modifications: 5-methylcytosine and beyond. *Epigenetics & Chromatin*, **8**(1), 24.

Deamer, D. *et al.* (2016). Three decades of nanopore sequencing. *Nature Biotechnology*, **34**(5), 518–524.

Deans, C. and Maggert, K. A. (2015). What Do You Mean, "Epigenetic"? *Genetics*, **199**(4), 887–896.

Deaton, A. M. and Bird, A. (2011). CpG islands and the regulation of transcription. *Genes & Development*, **25**(10), 1010–1022.

Dunham, I. *et al.* (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**(7414), 57–74. Number: 7414 Publisher: Nature Publishing Group.

Eid, J. *et al.* (2009). Real-time DNA sequencing from single polymerase molecules. *Science (New York, N.Y.)*, **323**(5910), 133–138.

Flusberg, B. A. *et al.* (2010). Direct detection of DNA methylation during single-molecule, real-time sequencing. *Nature Methods*, **7**(6), 461–465.

Gershman, A. *et al.* (2022). Epigenetic patterns in a complete human genome. *Science*, **376**(6588), eabj5089. Publisher: American Association for the Advancement of Science.

Holliday, R. (1994). Epigenetics: an overview. *Developmental Genetics*, **15**(6), 453–457.

Hotchkiss, R. D. (1948). The quantitative separation of purines, pyrimidines, and nucleosides by paper chromatography. *The Journal of Biological Chemistry*, **175**(1), 315–332.

Héberlé, and Bardet, A. (2019). Sensitivity of transcription factors to DNA methylation. *Essays in Biochemistry*, **63**(6), 727–741.

IHEC (n.d.). IHEC Consortium web site.

Jain, M. *et al.* (2016). The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome Biology*, **17**(1), 239.

Lenhard, B. *et al.* (2012). Metazoan promoters: emerging characteristics and insights into transcriptional regulation. *Nature Reviews Genetics*, **13**(4), 233–245. Number: 4 Publisher: Nature Publishing Group.

Li, H. *et al.* (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**(16), 2078–2079.

Liu, Q. *et al.* (2019). Detection of DNA base modifications by deep recurrent neural network on Oxford Nanopore sequencing data. *Nature Communications*, **10**(1), 2449. Number: 1 Publisher: Nature Publishing Group.

Nanopore (2022). nanoporetech/megalodon. original-date: 2019-06-20T15:09:43Z.

Ni, P. *et al.* (2019). DeepSignal: detecting DNA methylation state from Nanopore sequencing reads using deep-learning. *Bioinformatics*, **35**(22), 4586–4595.

Ni, P. *et al.* (2022). DNA 5-methylcytosine detection and methylation phasing using PacBio circular consensus sequencing. Pages: 2022.02.26.482074 Section: New Results.

Nurk, S. *et al.* (2020). HiCanu: accurate assembly of segmental duplications, satellites, and allelic variants from high-fidelity long reads. *Genome Research*, **30**(9), 1291–1305. Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab.

Nurk, S. *et al.* (2022). The complete sequence of a human genome. *Science*, **376**(6588), 44–53. Publisher: American Association for the Advancement of Science.

PacBio (2022a). PacificBiosciences/pbmm2. original-date: 2017-11-12T13:19:24Z.

PacBio (2022b). pb-CpG-tools. original-date: 2022-03-21T19:13:37Z.

Saito, T. and Rehmsmeier, M. (2017). Precrec: fast and accurate precision–recall and ROC curve calculations in R. *Bioinformatics*, **33**(1), 145–147.

Saxonov, S. *et al.* (2006). A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. *Proceedings of the National Academy of Sciences of the United States of America*, **103**(5), 1412–1417.

Schreiber, J. *et al.* (2013). Error rates for nanopore discrimination among cytosine, methylcytosine, and hydroxymethylcytosine along individual DNA strands. *Proceedings of the National Academy of Sciences*, **110**(47), 18910–18915. Publisher: Proceedings of the National Academy of Sciences.

Simpson, J. T. *et al.* (2016). Detecting DNA Methylation using the Oxford Nanopore Technologies MinION sequencer. Pages: 047142 Section: New Results.

Stoiber, M. *et al.* (2017). De novo Identification of DNA Modifications Enabled by Genome-Guided Nanopore Signal Processing. Pages: 094672 Section: New Results.

Suzuki, Y. *et al.* (2016). AgIn: measuring the landscape of CpG methylation of individual repetitive elements. *Bioinformatics*, **32**(19), 2911–2919.

Tse, O. Y. O. *et al.* (2021). Genome-wide detection of cytosine methylation by single molecule real-time sequencing. *Proceedings of the National Academy of Sciences*, **118**(5), e2019768118. Publisher: Proceedings of the National Academy of Sciences.

Vinson, C. and Chatterjee, R. (2012). CG methylation. *Epigenomics*, **4**(6), 655–663.

Vollger, M. R. *et al.* (2020). Improved assembly and variant detection of a haploid human genome using single-molecule, high-fidelity long reads. *Annals of Human Genetics*, **84**(2), 125–140. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/ahg.12364.

Wu, X. and Zhang, Y. (2017). TET-mediated active DNA demethylation: mechanism, function and beyond. *Nature Reviews Genetics*, **18**(9), 517–534. Number: 9 Publisher: Nature Publishing Group.

Yuen, Z. W.-S. *et al.* (2021). Systematic benchmarking of tools for CpG methylation detection from nanopore sequencing. *Nature Communications*, **12**(1), 3438. Number: 1 Publisher: Nature Publishing Group.

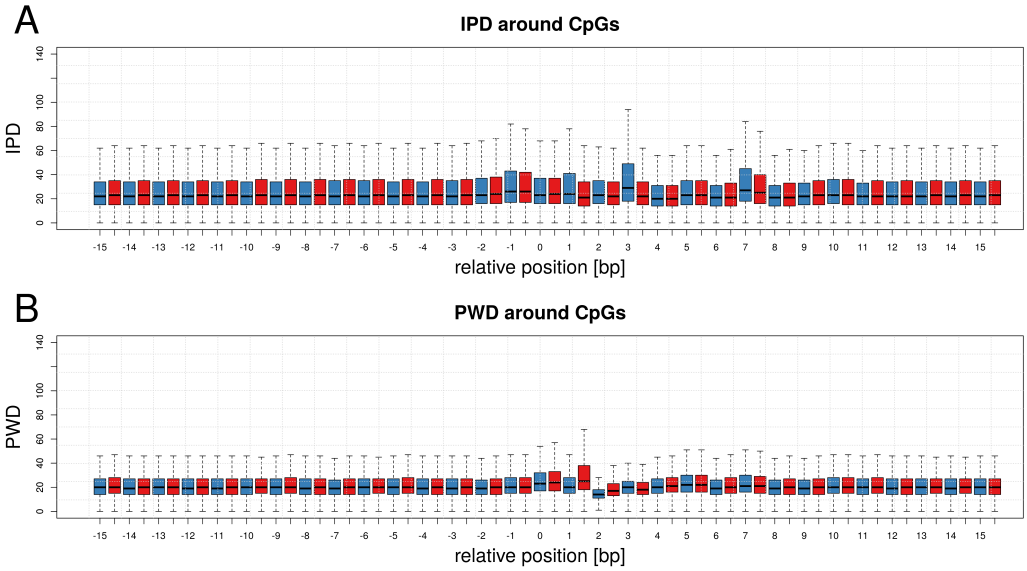


Fig. S1. Kinetic signal around CpGs A distribution of IPD signal in a +/-15bp window centered on all CpGs located on chromosome 1 (C is located at position +/-0). The IPD signal was extracted from the positions within the CCS corresponding to the window. The red distributions indicate CCSs without 5mC, blue distributions CCSs with 5mC. **B**: Same as A but with the PDW signal.

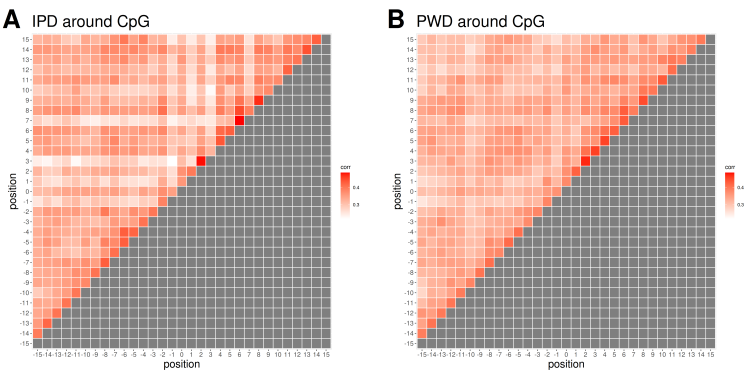


Fig. S2. Kinetic signal correlation around CpGs A: Correlation of the IPD signal between all pairwise positions within a +/-15bp window centered on CpGs (C is located at position +/-0) from chromosome 1, 2 and 3. The IPD signal was extracted from the positions within the CCS corresponding to the window and the per position averages were computed. **B**: Same as A but with the PDW signal.

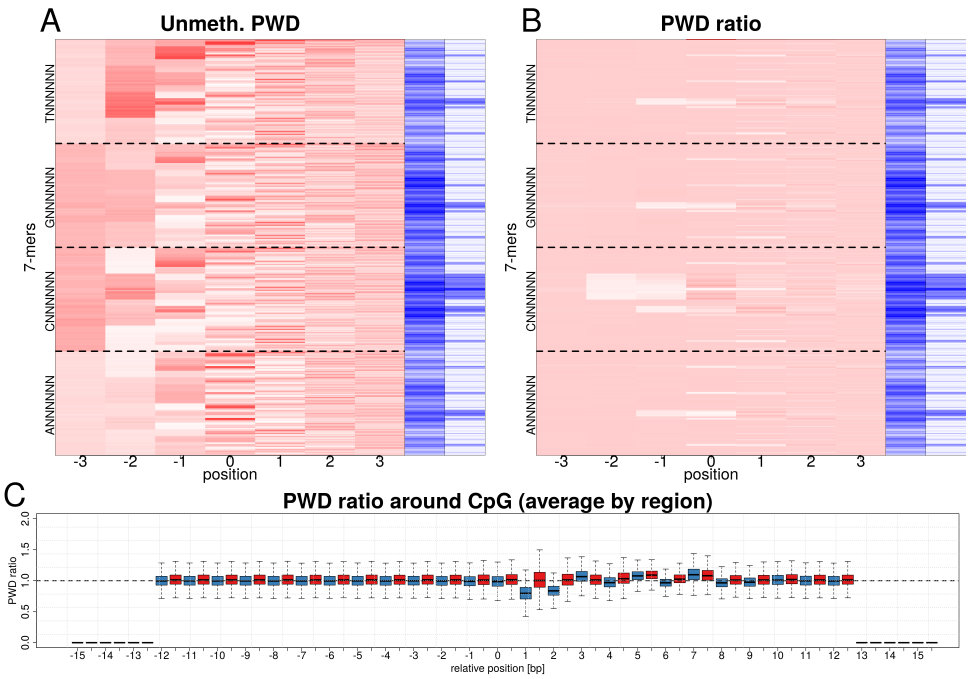


Fig. S3. Effect of the sequence on the PWD signal A: Heatmap of the mean PWD signal at each position of all possible 7-mers, from the WGA sample CCSs. The corresponding sequences and PWD signals were extracted and the mean PWD signal was computed for each 7-mer. The 7-mers are sorted in lexicographic order from bottom to top. The dashed lines indicate the limits between A, C, G and T starting 7-mers. The blue bars indicate the 7-mer CG content (center) and CpG content (right) respectively. **B**: Heatmap of the mean PWD ratios (PWDr) at each position of all possible 7-mers. The mean per 7-mer PWD values were computed from the SssI sample CCS as describe in A). Each 7-mer SssI mean PWD was then divided by the corresponding 7-mer WGS average. The ordering, dashed lines and blue bars are the same as in A). **C** Distribution of PWDr signal in a 31bp window centered on all CpGs located on chromosome 1 (C is located at position +/-0). The PWDr signal was extracted in a per CpG manner. For a given CpG, all CCSs mapping perfectly over the CpG window were considered. The PWD signal was extracted from the positions within the CCS corresponding to the window, normalized using a 7-mer backgound model and the per position averages were computed. This process was repeated for all CpG on chromosome 1 and the distributions of values is displayed per position. The first and last six positions in the window have no signal because the normalization process trims the signal, as explained above. The red and blue distributions indicate the WGA and SssI samples respectively.

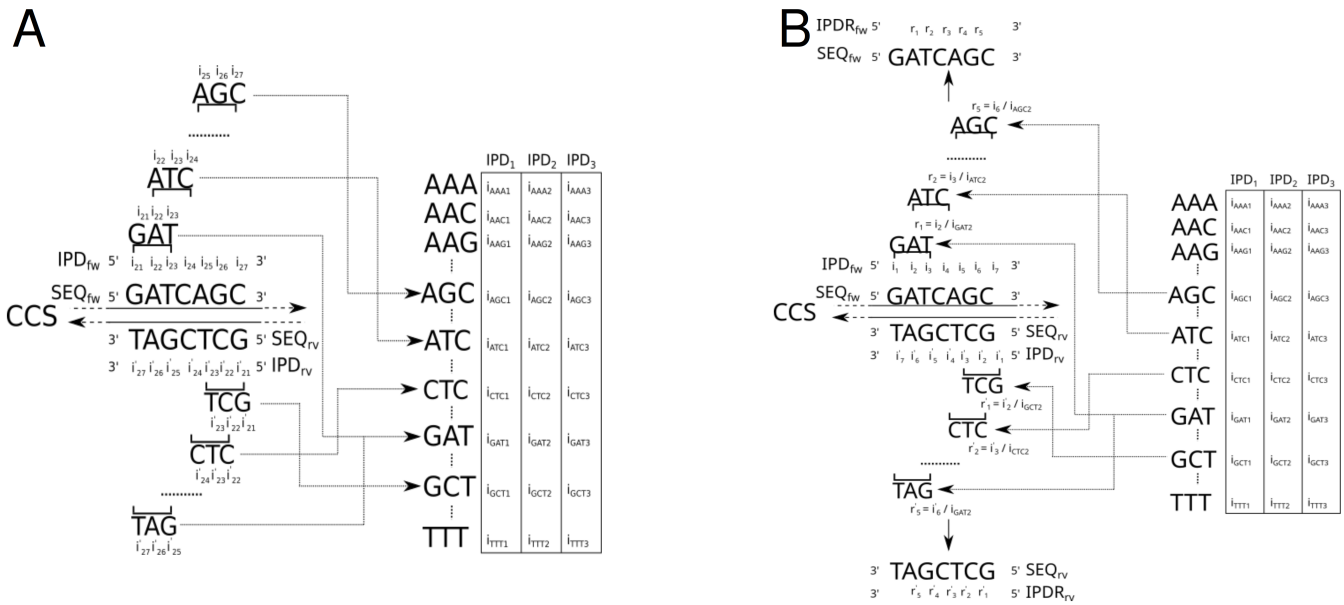


Fig. S4. Sequence background model A: A map contains all kmer sequences as keys and the expected per kmer position IPD (or PWD) values without any DNA modification is stored as value. The sequence background model is trained from each kmer in a WGA dataset and computing the per kmer IPD (or PWD) average signal. **B**: A stretch of IPD (or PWD) kinetics can be normalized using the map. The stretch is broken in smaller sub-slices of length K. The central position of each sub-slice in the kinetic signal is divided by the central expected IPD (or PWD) value expected for this kmer without DNA modification. This procedure is repeated for every possible sub-slice.

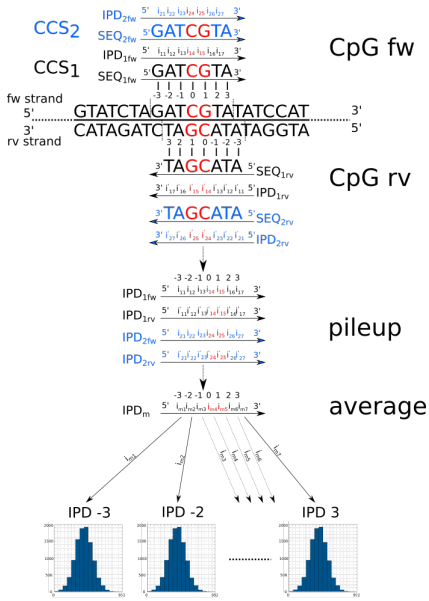


Fig. S5. Training of a kinetic signal model: the different steps of the training of a kinetic signal model are displayed. From top to bottom: 1) CCSs mapping over a CpG window - here the window is of size 7bp - are considered, 2) the sub-vector of kinetic values - the IPD only are displayed for clarity - overlapping the CpG window are extracted, 3) a pileup of kinetic vectors is constructed, 4) the strand-aggregated average per-position kinetics are computed from the pileup and 5) each value is stored in the corresponding per-position histogram.

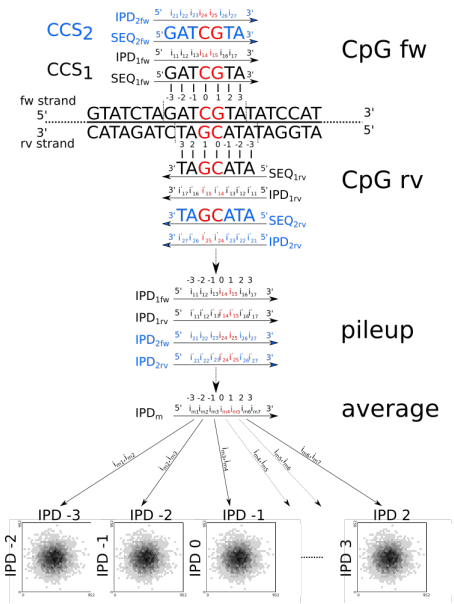


Fig. S6. Training of a DP kinetic signal model: the different steps of the training of a kinetic signal model are displayed. From top to bottom: 1) CCSs mapping over a CpG window - here the window is of size 7bp - are considered, 2) the sub-vector of kinetic values - the IPD only are displayed for clarity - overlapping the CpG window are extracted, 3) a pileup of kinetic vectors is constructed, 4) the strand-aggregated average per-position kinetics are computed from the pileup and 5) each pair of neighboring values with index $x, y \in D_{DP}$ is fed in the corresponding per-position histogram.

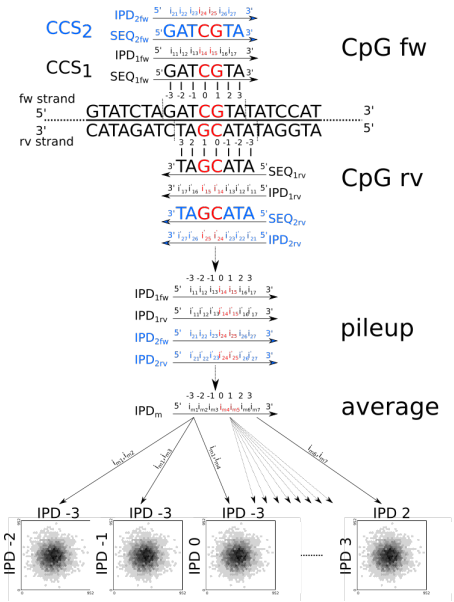


Fig. S7. Training of a PW kinetic signal model: the different steps of the training of a kinetic signal model are displayed. From top to bottom: 1) CCSs mapping over a CpG window - here the window is of size 7bp - are considered, 2) the sub-vector of kinetic values - the IPD only are displayed for clarity - overlapping the CpG window are extracted, 3) a pileup of kinetic vectors is constructed, 4) the strand-aggregated average per-position kinetics are computed from the pileup and 5) each pair of neighboring values with index $x, y \in D_{PW}$ is fed in the corresponding per-position histogram.