RESEARCH

Implementation of the Single-Molecule Single-Nucleotide (SMSN) approach for the PacBio Sequel I sequencer with in-sillico control

Guillaume G. Delevoye
, Mathieu M. Bahin
and Eric E. Meyer*

Abstract

Background:

The PacBio SMRT sequencing allows de novo sequencing and DNA methylation analysis. The historical way of doing this methylation analysis consists in aggregating information from several distinct DNA molecules, which causes interpretability issues. A more recent and less common approach consists in treating each DNA molecule separately from the others, to detect modified nucleotides on both strands of a same DNA molecule. It is called the Single-Molecule Single-Nucleotide approach (SMSN).

The only bioinformatic pipeline available to date to handle SMSN data is called SMALR. It requires to use a PCR-amplified control to detected methylated nucleotides. For laboratories who cannot afford financially or technically to sequence such PCR-amplified control, Pacific Biosciences also provides an "in-silico control", which is a machine-learning modelling of the PCR-amplified control. While SMALR does allow to handle the newest .bam file formats, it does not allow to us this in-sillico control, which is a limitation for some research projects.

Results:

We implemented a new software that handles the more recent *.bam* formats starting from the Sequel I sequencer. It is compatible with the SMSN approach, and does not require any PCR-amplified control. We tested its ability to detect DNA n6-methyladenines in *E. coli*'s GATC sites. We found that the sensitivity of the pipeline is more than 92% while its specificity is more than 99.8%.

Conclusions:

The performances of this new pipeline are slightly below those of SMALR. This comes however, with the benefit of significant financial savings. Our python software is in the form of a simple and standard Command-Line-Interface (CLI) application. It takes in input 1) raw SMSN PacBio subreads and 2) a reference genome in input. It then outputs a tabular .csv file whose interpretability is straightforward. The code is freely-available. All its dependencies (including PacBio dependencies) can be installed easily via an anaconda environment, without root access. Its performances allow to process a whole SMRTCell in a day with a personal computer. We hypothesize that only minor modifications, if any, would be required to make it compatible with the latest Sequel II data or to study DNA 4-methylcytosine.

Keywords: PacBio; Sequel; DNA Methylation; n6-methyladenine; Pipeline; In-sillico-control

*Correspondence: emeyer@ens.fr
Institute of biology, Ecole Normale Supérieure - PSL, Paris, France
Full list of author information is available at the end of the article

Contents

List of Figures

List of Tables

1 Background

1.1 *De novo* assembly with PacBio SMRT sequencing

In 2013, Pacfic Bioscience© (Menlo Park, CA, USA) realeased the "RS II" DNA sequencer [4]. The RS II allowed to sequence individual DNA molecules without any PCR amplification, based on the "Single-Molecule Real-Time" (SMRT) technology.

SMRT-sequencing has been described extensively in the literature (See [1]) and is summarized in Figure 1. It works as follows :

- A flow-cell (called *"SMRT-Cell"*) of several thousands of microscopic wells (Figure 1 A.) is loaded with a library of fragmented and circularly ligated DNA molecules (Figure 2).
- The microscopic wells are called *"Zero-Mode-Waveguide"* (ZMW).
- The DNA molecules are called "*SMRTbell*".
- The loading process ensures that in each ZMW, only one SMRTbell can be fixed at a time.
- Once the flow-Cell is loaded, the sequencer starts an uninterrupted DNA polymerization, during which each of the 4 types of nucleotides emits a different fluorescence when incorporated (Figure 1 B, top).
- A camera records a movie of these light flashes, distinguishing each individual ZMW from the others (Figure 1 B, bottom).
- The resulting movie is analyzed to identify which nucleotide was sequenced, in which order, and recreate the sequence of the original DNA molecule.

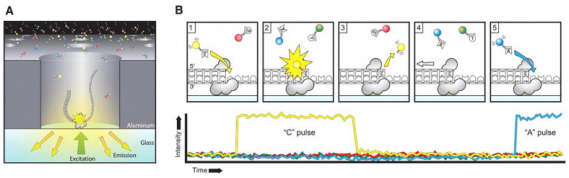


Figure 1: Sequencing via light pulses (Creative Commons, reproduction from [1], not modified)

A. A SMRTbell (gray) diffuses into a ZMW, and the adaptor binds to a polymerase immobilized at the bottom. B. Each of the four nucleotides is labeled with a different fluorescent dye (indicated in red, yellow, green, and blue, respectively for G, C, T, and A) so that they have distinct emission spectrums. As a nucleotide is held in the detection volume by the polymerase, a light pulse is produced that identifies the base. (1) A fluorescently-labeled nucleotide associates with the template in the active site of the polymerase. (2) The fluorescence output of the color corresponding to the incorporated base (yellow for base C as an example here) is elevated. (3) The dye-linker-pyrophosphate product is cleaved from the nucleotide and diffuses out of the ZMW, ending the fluorescence pulse. (4) The polymerase translocates to the next position. (5) The next nucleotide associates with the template in the active site of the polymerase, initiating the next fluorescence pulse, which corresponds to base A here.

Figure 2: SMRTbell template (Creative Commons, reproduction from [1], adapted description)

The sequencing library is obtained by taking double-stranded DNA molecules and circularize them with hairpin adapters (green), so that the polymerase (grey) can sequence the two strands of the molecules several times if the read is long enough. In the final continuous long read (CLR), the adapter sequences (green) can serve to separate the subreads orginating from one strand arbitrary called top (e.g orange) to those of the other strand (purple).

The result for each individual ZMW consists in a very long read of several kilobases called "Continuous Long Read" (CLR). These CLRs can be extremely long, up to $>$ 60 kb for the RS II sequencer [1]. In 2013, this property made the RS II a good candidate to make assemblies of long repeated regions [5].

The CLR produced by the RS II have only one yet important problem: their error rate (about 12 to 15% per nucleotide [6]).

When the template is short enough however, DNA molecules can be sequenced multiple times since they are circularized (Figure 2). When it happens, the adapter sequences in the CLR can be used to delimit "subreads" and count "the number of passes" - that is the number of time a DNA molecule is entirely sequenced.

When many subreads are available, it is bioinformatically possible to create a circular consensus (CCS) of the DNA molecule [7] with a near-100% accuracy. Of course the higher the number of subreads the higher the accuracy, which is why short DNA molecules tend to produce more subreads (and therefore more accurate CCS) than longer ones.

Depending on the purpose of the sequencing, it is up to the final user to find the right tradeoff between consensus size and consensus accuracy. In practice, the library preparation can allow to produce both long and short DNA molecules. The more accurate CCS can then be used bioinformatically as seeds against which less accurate ones can be aligned. This approach is called Hierarchical Genome Assembly Pro-

cess (HGAP) [8], and produces satisfying results for many model organisms (See Table 2 of A. Rhoads and K. Fai Au (2015) [1]).

1.2 Detection of methylated nucleotides with the PacBio SMRT sequencing

The RS II does not only measures the fluorescence emitted by nucleotides, but it also measures the time between the incorporation of two nucleotides - called Inter-Pulse-Duration (IPD).

Analyzing these IPDs allows to detect DNA modifications even without any base conversion treatment. The polymerase indeed tends to slow down just after it incorporates modified nucleotides (Figure 3).

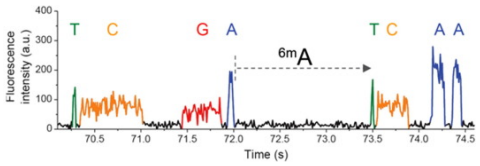


Figure 3: Detection of methylated bases using PacBio sequencing. (Creative Commons, reproduction from [1], adapted description)

PacBio sequencing can detect modified bases, including N^6-methyladenine (6mA), by analyzing variation in the time between base incorporations in the read strand. Here, the slowing of the polymerase after the adenine sequenced at t=72s can indicate the presence of 6mA.

Besides DNA modifications, the kinetics of the polymerase is also determined by the surrounding nucleotides ("the context"). On average in a window of -7/+2 nucleotide around a given nucleotide, roughly 80% of the IPD variation is explained by the context alone, while the rest is mostly due to the presence or absence of DNA modifications in this window [9].

By comparing the kinetic of the polymerase in a test DNA to that of a PCR-amplified DNA - where by definition no DNA modification is present, it is therefore possible to attenuate the noise produced by the context, to extract the signal corresponding to DNA modification.

In practice, this signal is composed mostly of a slowing of the polymerase when it passes over the modified nucleotide [10] [11]. Secondary slowing on the surrounding nucleotides (called "secondary peaks") can also be present, in which case they are often specific to one kind of DNA modification in particular.

Complex "kinetic signatures" composed of one main peak and one or several secondary peaks can be identified this way. It is notably the case for N^6 methyladenine (6mA), N^4 methylcytosine (4mC) , and to a lesser extend C^5-methylcytosine (5mC), that are the three most studied DNA modifications in biology.

1.3 AggSN approach versus SMSN approach to detect DNA modifications

Regarding the detection of DNA methylation with SMRT-sequencing, the first challenge to address is the weakness of the signal/noise ratio. To reliably detect DNA modifications, it is mandatory to measure the same IPDs multiple times and use the average instead of an isolated measure.

For 6mA and 4mC, the minimal number of IPDs is usually set to 25 in both the PacBio documentation [12] and the literature [13]. Other DNA modifications (e.g 5-methylcytosine (5mC) or 5-hydroxymethylcytosine (5-hmC)) can require up to 250 independent measures of the IPD to be detected correctly [12].

There are two ways to reach such numbers of IPDs :

1. An historical approach that consists in sequencing very long DNA molecules and pool together IPDs from physically distinct DNA molecules. This approach was later called "AggSN" [2])
2. A more recent approach that consists in sequencing short inserts, a very high number of time each. This second approach is called "SMSN" [2] - Single Molecule Single Nucleotide).

Figure 5, reproduced from [2], illustrates the difference between the two approaches.

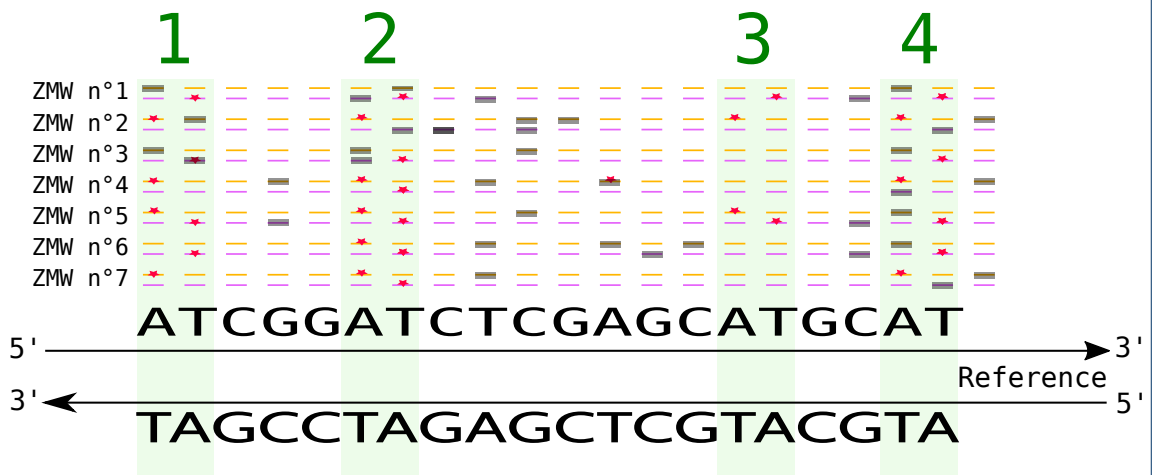
Only two softwares are available to date to perform a DNA methlation analysis from PacBio subreads : PacBio's ipdSummary [14] [15] for the AggSN approach and J. Beaulaurier's SMALR [16] [2] for the SMSN approach.

1.4 Description of ipdSummary (AggSN approach)

1.4.1 Alignment of subreads

In the AggSN approach, very long DNA molecules are sequenced. Then, IPDs from physically distinct DNA molecules are pooled together by the ipdSummary software to detect DNA modifications.

This aggregation is performed by first aligning all the subreads against the same reference genome assembly (Figure 4). This step is not done by ipdSummary itself, and it is the user's responsibility to provide a file with the aligned subreads. In general, PacBio's "BLASR" aligner can be used for this task.



Position	... 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 ...
Fraction 6mA (detected)	
strand +:	75% 100% ~28% 100%
strand -:	50% 100% ~28% 100%

Figure 4: AggSN approach implemented in ipdSummary

In the AggSN approach, all subreads are aligned against the same reference genome and the pipeline reports the fraction of methylated molecules for each positioon of the reference genome.

In this example, we consider a window of the reference genome where 7 different DNA molecules (ZMW) produced 7 different reads, each of them being composed of two subreads : One that aligns (orange) against the reference assembly (or "strand +") and one that aligns against (purple) its reverse complementary strand (or "strand -").

4 different AT sites are highlighted in pale yellow. Red stars represent 6mA. Grey boxes represent nucleotide whose IPD is ignored by the ipdSummary software because the subread does not match the reference. These 39 grey boxes correspond here to an error rate of about 13%, which is close to real-world error rates.

IpdSummary iterates over the positions in the reference genome, and computes for each of these positions the mean of the IPDs. IPDs of all the subreads are taken to perform this calculation, regardless of which ZMW they came from. At the end, a monte-carlo algorithm tries to estimate the fraction of the aligned ZMW that contained a modified nucleotide, and reports it in the final tabular output.

When it computes the mean IPD for a position, ipdSummary ignores not only the IPDs that are censored, but also the ones that are at position N+1 or N-1 of a censored IPD in the subread. For instance, the nucleotide at position 13 of ZMW n°3 (strand - / purple) is censored, which means that the nucleotides n°12 and n°14 of this subread are also ignored. Depending on the subread's error rate, this censoring can significantly lower the number of IPDs available for the analysis.

1.4.2 Preprocessing of IPDs

Once th subreads are aligned, ipdSummary preprocesses the IPDs. All IPDs that originated from a nucleotide of the subread that does not match the genome reference (grey boxes, Figure 4) is censored. The software also censors the IPDs from the positions N+1 and and the N-1 around nucleotides that don't match the reference, based on the principle that if the fluorescence was incorrectly measured at this location, then so was the IPD between the two light flashes.

There are also rare (but extremely long) random pause during the course of the polymerase, that have nothing to do with DNA modification and that also need to be removed. This problem is addressed by capping the highest percentiles of IPDs [17].

Because of the way IPDs are censored there is sometimes a significant difference between the number of sequenced subreads for a ZMW, and the number of IPDs actually retained for methylation analysis. of a given nucleotide. Unless otherwise stated, "coverage" will always refer in this article to the effective coverage that remains after IPDs have been preprocssed. This coverage can vary considerably within the same molecule, including between two neighboring nucleotides in the same molecule.

1.4.3 Per-position analysis

Once the preprocessing is done, ipdSummary iterates each position of the reference genome. Over the IPDs that are not censored, it computes (independantly for each strand) their normalized mean and standard error, as well as the normalized mean and standard error from the corresponding PCR-amplified control DNA.

In output, ipdSummary pipelines produces a tabular .csv file that contains the following columns for each covered nucleotide of the reference genome:

- Base: The nature of the nucleotide (A,T,C,G).
- refId/tpl: Its position in the reference genome.
- Strand: "+" if the nucleotide corresponds to the reference strand, and "-" for its reverse complement.
- controlCoverage/caseCoverage: Corresponds to the number of IPDs that remained after the preprocessing steps, and on which the DNA methylation analysis was performed ("effective coverage").
- caseMean/caseStd: Mean of normalized IPDs and standardized error in the test DNA.
- controlMean/controlStd: Same, but in the PCR-amplified control.
- ipdRatio: Ratio between the mean IPD in the test DNA, and the mean IPD in the control DNA. Sometimes, the $log(ipdRatio)$ is also used to get an easy visualization of the difference between modified and non-modified nucleotides.
- testStatistic: 1-sided t-test statistic, or "p-value" (H_0: No modification is present).
- score: A PHRED-transform [18] of this p-value. Also called "modification score", or "modQv". This score is positively correlated with the propensity of this position in the genome to carry DNA modifications.

Then, four important issues remain :

1. The exact chemical nature of the identified modification is not known yet.
2. Secondary peaks from complex kinetic signatures might generate false positive detections around modified nucleotides.
3. There are also complex cases where, due to these secondary peaks, several combinations of different DNA modifications might explain the same kinetic signature.
4. Among all the molecules that align at a position N in the genome, not all of them must carry the same DNA modification for the nucleotide N.

To our understanding of its code, ipdSummary tackles the three first issue with a two-steps algorithm. The first step consists in listing all the suspect slowing of the polymerase in rolling windows of nucleotides. Then for each window, it performs a combinatory analysis of all the different possibilities of DNA modifications that can explain the observed kinetic signature in this window, to extract the most likely one with a likelihood-ratio test.

This process is entirely hidden in a blackbox for the final user. It outputs two additional columns :

- "modification": This column contains either "6mA", "4mC", "5mC" or "modified", depending on what is the most likely modification of this nucleotide, if it is indeed modified.
- "identificationQv": A score that is positively correlated with the confidence of the previous identification.

This way of reporting DNA modifications "per position in the reference genome" is convenient and can be easily interpreted when, between all the DNA molecules that originate from the same region in the reference genome, little to no heterogeneity is expected in terms of DNA modifications.

In real life however, it is possible that for each position N in the reference genome, not all corresponding DNA molecules might have the same DNA modification at position N. When it is asked by the user, ipdSummary can therefore produce three last columns to study this heterogeneity :

- frac: Among all the molecules that served for the methylation analysis at this position in the reference genome, what is the estimated fraction of them that carries the identified modification.
- fracUp/fracLow: Upper and lower boundaries of a 95% confidence interval around the aforementioned fraction.

1.4.4 Interpretation pitfalls

In the aggSN strategy, the fraction of methylated nucleotides reported by ipdSummary for each position N in the genome is the only metric through which heterogeneity between DNA molecules can be studied. The interpretation of this fraction however, is not straightforward. Many pitfalls can occur during data analysis, as can be illustrated with Figure 4, where hemi-methylated, symetrically-methylated or non-methylated AT sites are analyzed. It should can remarked for instance that :

- If an AT site is systematically hemi-methylated, this cannot be deduced from the fractions alone (Figure 4, site n°1)
- The same fractions can be obtained with very different situations (Figure 4, sites n°2 and n°4)
- The same fraction can correspond to different realities. (For instance, the site n°3 of Figure 4 could correspond to a reality of 28% of symetrically-methylated sites, or 56% hemi-methylated AT sites.)
- In more extreme situations, it is even possible that in an AT sites, the two adenines are reported methylated in 100% of times while there there is actually not a single DNA molecule that caries a symetrically-methylated site (Figure 4, site n°4).

Also the AggSN approach does not allow to study properly hemi-methylated molecules or strand-specific effects (e.g ZMW n°1 and 2 of Figure 4), or the correlation between positions in a same molecule.

It is on the basis of such observations that J. Beaulaurier et al proposed the SMSN approach in 2015, which they implemented in their own analysis software: SMALR.

1.5 Description of SMALR (SMSN approach)

The SMSN approach consists in using shorter SMRTBell and therefore, sequence them many times each to gather a sufficient number of IPDs. This allows to study each DNA molecule independantly from all the others, on its two strands, with a single-nucleotide precision.

SMALR analyzes these PacBio data in a way that is very similarly to ipdSummary, especially in the way it censors IPDs. There are however three notable differences with ipdSummary :

1 SMALR does not pool information from physically distinct molecules, which is the most important difference. ZMW are identified by a unique identifier. The pipeline reports one tabular .csv output per molecule sequenced, rather than one line per position in the reference genome.

2 Instead of computing the p-value of a one-sided t-test between the mean IPD of the native and the control DNA, SMALR computes its own "SMSN score" (1). The difference between the the metrics of ipdSummary and SMALR must be highlighted, because it can be misleading :

$$SMSN_{score} = \frac{Mean(log(IPDs_{native}))}{Mean(log(IPDs_{control}))} \quad (1)$$

$$log(ipdRatio) = log\left(\frac{Mean(IPDs_{native})}{Mean(IPDs_{control}}\right) \quad (2)$$

3 SMALR does not seek for complex kinetic signatures nor does it care about secondary peaks. It only reports nucleotides where the polymerase slows down, which turned to be sufficient for several real-life analysis. This is why unlike ipdSummary, SMALR does not output any identificationQV.

Incidentally, SMALR also re-implements the AggSN analysis of ipdSumamry. It does not, however, bring additional features to the original PacBio software.

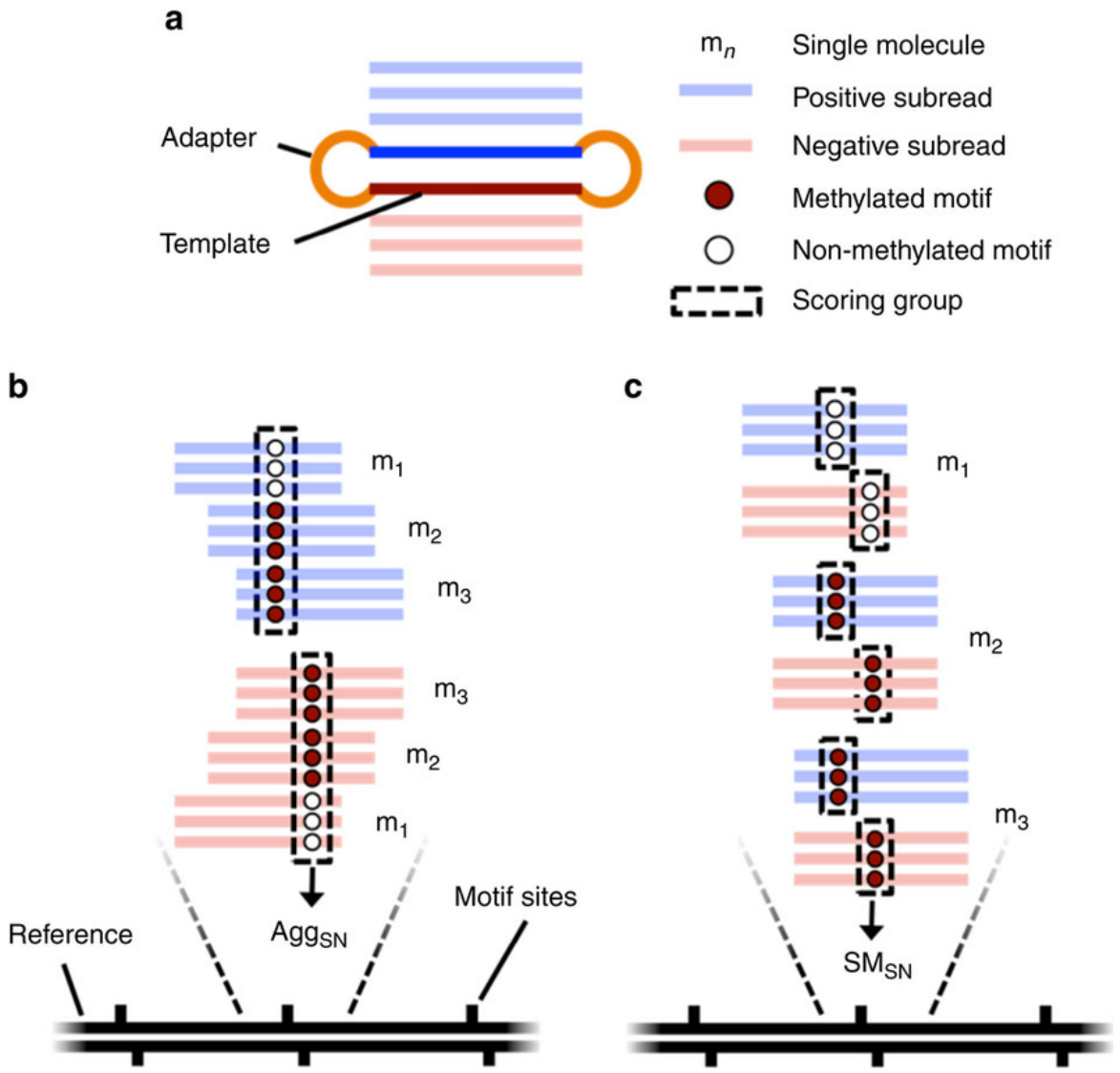


Figure 5: AggSN approach versus SMSN approach (Creative commons, reproduced from [2], description modified)

Schematic illustrating of the differences between the AggSN and the SMSN approaches in SMRT sequencing. (a) A single SMRT sequencing molecule (short DNA insert+adapters) and the subreads that are produced during sequencing. (b) The historical methylation detection method implemented by ipdSummary is based on a molecule-aggregated, single-nucleotide (AggSN) score. For a given strand and genomic position, the IPD values from all the subreads aligning to that strand and position are aggregated together across all molecules to infer the presence of a consensus methylated base. (c) The single molecule single nucleotide (SMSN) method for detecting DNA methylation implemented by SMALR relies instead on separate consideration of subreads from different molecules. The SMSN scores are calculated for each molecule, strand and genomic position.

1.6 The in-silico control

Sequencing DNA with PacBio sequencing is expensive, and this cost is even more important with the SMSN strategy than with the AggSN strategy. It is also more expensive to do a DNA methylation analysis than it is to realize a simple de novo assembly. DNA methylation analysis indeed requires to sequence a PCR-amplified DNA as a control, which doubles the sequencing cost the first time DNA methylation is studied in a new species.

For laboratories who, either for financial or technical limitations cannot afford to use a regular PCR-amplified control, PacBio also developed the so-called "in-silico control". The latter is a machine-learning modelling of what the speed of the polymerase should be for a given nucleotide in a given context (-8/+3 nt), in absence of DNA modifications [9] [19] [20]. It allows substantial financial savings, but it comes at the price of being less reliable than a real-world PCR-amplified control.

Although it was first mostly meant as a marketting argument, this in-silico control turns out to allow specific experimental designs, that would be impossible otherwise (see "Motivation").

The in-silico control produces for each position a point value without any standard error. This is why instead of a one-sided t-test, PacBio softwares compute a Aspin-Welch t-test when the in-silico control is used. The whole principle however, remains the same.

ipdSummary implements this in-silico control, whereas SMALR does not. In other words, the SMSN approach can only be used, to date, without the in-silico control.

1.7 Problematic and objectives

After the RS II sequencer, Pacific Biosciences released subsequently the Sequel I (2015) [21] and the Sequel II (2019) [22]. Each generation was released with new chemistries and binding kits, which resulted in different file formats, different Inter-Pulse durations, and different in-silico controls. These differences are summarized in Table 1.

Regularly, Pacific Biosciences released a new version of its open-source "SMRT analysis" [23] software suite to adapt to these changes. SMRT Analysis however, never included any implementation of the SMSN analysis pipeline. To date, the only software available to analyze SMSN data remains SMALR, which was developped for the RS II.

Although recent updates have been brought to SMALR to handle the new .bam file formats, its ability to work on post-RSII data has not been demonstrated and it does not implement the PCR-amplified control, which is an important limitation.

Our goal was therefore to implement an SMSN analysis pipeline in a way that is compatible with the more recent data Sequel data, and can be used with the in-silico control.

1.8 Motivation

The sole fact that using a PCR-amplified control requires to sequence twice the same DNA sample is already a good financial argument to favour the in-silico control over the PCR-amplified one.

Beyond this practical aspect however, there are also situations in which sequencing a PCR-amplified control is technically impossible, and we were placed in a such situation ourselves - which motivated the present paper.

At the origin of our work, we wanted to study DNA methylation in *P. tetraurelia*. This unicellular species possess two types of nuclei of very unequal ploidy: A macronucleus (MAC) of ploidy 800n, versus two diploïd micronuclei (MIC) - 4n in total, leading to a ploidy ratio of about 1:200 [24]. The genomes contained in these two tpes of nuclei vary little in size and content.

For reasons that are beyond the scope of this article, we had a particular interest in studying DNA methylation around the MIC-specific sequences, that is, the most rare sequences.

Our challenge was that it was impossible for us to experimentally purify the MIC DNA without risking severe DNA fragmentation, or without altering certain epigenetic marks – among which DNA methylation. This is why we could only sequence total, unpurified DNA where the MIC molecules were present in less than 1% of the total.

In this situation, having a PCR-amplified control of the whole MIC-specific sequences would have required to sequence our total DNA samples with coverage of about $25 \cdot 200 = 5000X$. This order of magnitude was financially impossible to reach since our genome of interest was $\sim 98Mb$ long [25].

The only viable PacBio sequencing option for us was therefore to use the SMSN approach on total DNA, isolate the $< 1\%$ of the total molecules that came physically from the MIC, to study their methylation, and analyze them with the in-sillico control. No available software allowed to do so, hence the present work.

Sequencer version	Year	Chemistry	File format	Consensus from CLR	Methylation analysis (AggSN - PCR-control)	methylation analysis (AggSN) - in-silico control	methylation analysis (SMSN-PCR-control)	methylation analysis (SMSN- in-silico control)
RS II	2013	P5-C3 orP6-C4	.h5	Available AggSN and SMSN (SMRTAnalysis)	Available (SMRTAnalysis)	Available (SMRTAnalysis)	Available (SMALR) - not tested	Not available
Sequel I	2015	SP2-C2	.bam	Available AggSN and SMSN (SMRTAnalysis)	Available (SMRTAnalysis)	Available (SMRTAnalysis)	available (SMALR) - not tested	Not available
Sequel II (v1.0)	2019	None	.bam	Available AggSN and SMSN (SMRTAnalysis)	Available (SMRTAnalysis)	Not available	available (SMALR) - not tested	Not available
Sequel IIe (v2.0)	2021	SP2-C2	.bam	Available AggSN and SMSN (SMRTAnalysis)	Available (SMRTAnalysis)	Available (SMRTAnalysis)	available (SMALR) - not tested	Not available

Table 1: Summary of the different PacBio sequencing versions, chemistries and software availability

The SMSN pipeline has only been implemented by SMALR for the RS II sequencer. It also requires to use a PCR-amplified control. For the subsequent versions of the sequencer, the documentation of SMALR claims that it has been implemented to handle the new .bam format. It could not however, be tested on real-world data produced by the new sequencers. This is why the Sensitivity and Specificity are just extrapolated from the initial work on the RS II sequencer - which migh lead to wrong estimations. To our knowledge, no software to date has implemented the SMSN analysis with the in-silico control.

2 Implementation and challenges

2.1 Dependencies

To implement this new pipeline, compatible with both the SMSN approach and the in-silico control, we decided to rely heavily on the already-existing PacBio software.

This decision was motivated by the fact that these tools already implemented complex or opaque algorithms that would be error-prone to recode ourselves including:

1. The parsing of the various PacBio chemistries and non-standard .bam formats [26]
2. The statistical treatments require to handle the random pauses of the polymerase [17]
3. The lossy encoding/decoding of the IPDs directly from the raw .bam files [26]
4. The identification of secondary peaks and complex kinetic signatures [20]
5. the in-silico control itself [20] which could not easily be retro-ingineered.

Another argument to rely on the PacBio softwares is that they are maintained by a large team of developers, whose capacity in software ingineering, continuous integration, code maintenance and client support necessarily outmatches by far that of a small academic team made of non-permanent members.

Among all the available PacBio softwares, our pipeline relies especially on using : *CCS* (create circular consensus from the subreads), *BLASR* (aligner), *ipdSummary/KineticsTools* (DNA modification analysis) and indirectly, the whole *pbcore* (fast toolbox implemented in C++) suite on which these three software are based.

All these softwares are usually shipped within a package called SMRTAnalysis [27]. To allow an easy installation for non-root users, ensure that no version conflict arises for the final users and facilitate the deployment, we created an anaconda [28] virtual environment containing all these dependencies.

The pipeline itself was implemented in pure python, to glue all the components correctly together.

2.2 Pipeline description

In one word, our pipeline consists simply in applying ipdSummary on each ZMW separately.

The first phase relies on two PacBio tools: CCS and BLASR. They are used respectively to build the circular consensus and align them against the reference genome. This phase starts from the sequencer's raw files and ends by producing a .bam file of aligned consensus that match the reference genome (Figure 6).

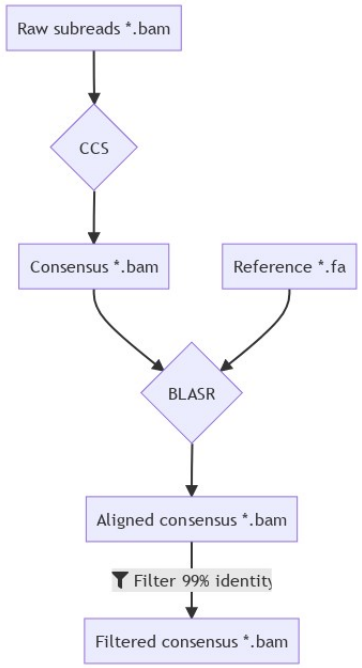


Figure 6: Illustration of the proposed pipeline (phase I)

The proposed pipeline start by building the circular consensus from the subreads with CCSn and aligns these consensus against the reference genome with BLASR.

The parameters of CCS are set to allow a maximum number of consensus to be produced while the parameters of BLASR are set to ensure that it reports only the best alignment for each consensus.

The consensus whose percentage of identity (see Equation (3)) is greater than a user's specified threshold (default: 99%) will be ignored in the next steps. This serves as a quality control.

Consensus that don't reach a user-defined percentage of identity ((3)) will be marked to be ignored in Phase II (Figure 7 and 8).

$$Identity = 2 \cdot \frac{Number\ matches}{length_{Alignment} + length_{molecule}} \quad (3)$$

Phase II then basically consists in iterating over the subreads *.bam file (which is sorted by ZMW identifier), and apply ipdSummary to each ZMW, one after the other, without mixing data from distinct ZMWs.

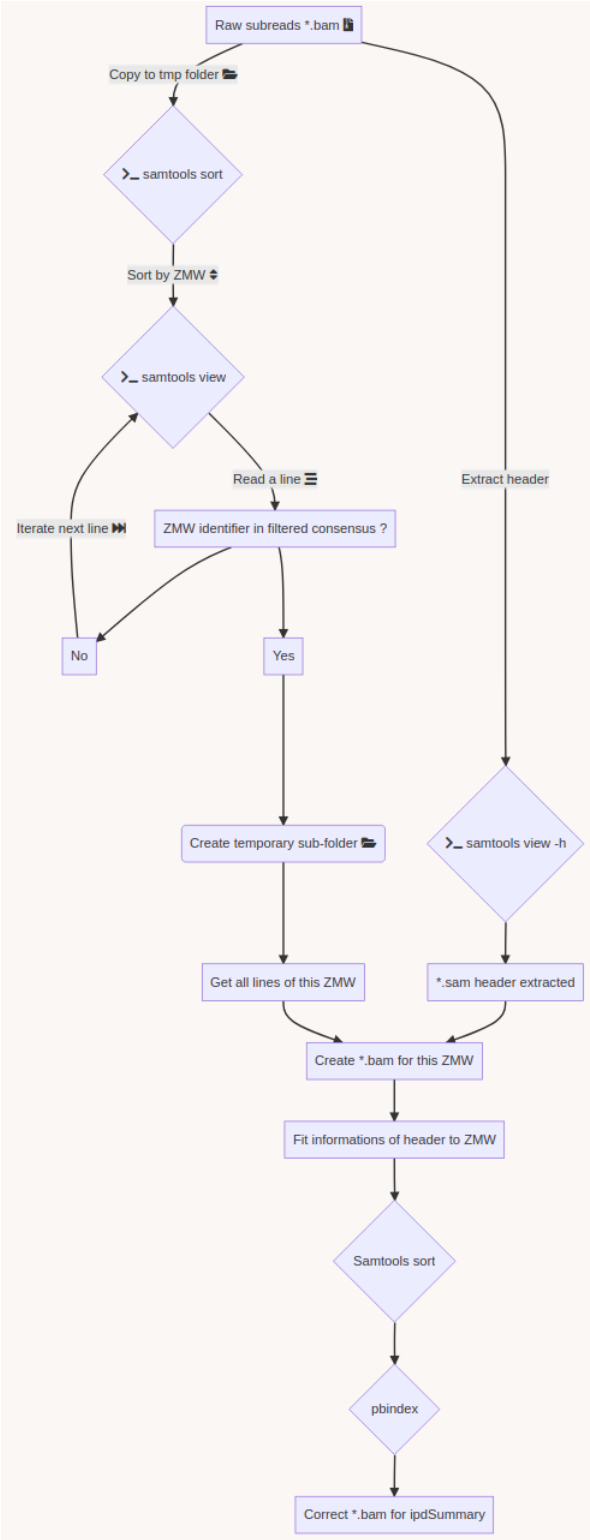


Figure 7: Illustration of the proposed pipeline (Phase II A)

During the second phase, we iterate over the .bam file of the subreads, where they are sorted by ZMW identifier. Whenever we find a ZMW identifier that fits to those that must be analyzed, we recreate a separated .bam file that contains all the subreads of this ZMW. We ensure in particular that its header is compatible to be fed to ipdSummary in phase II.B. 8).

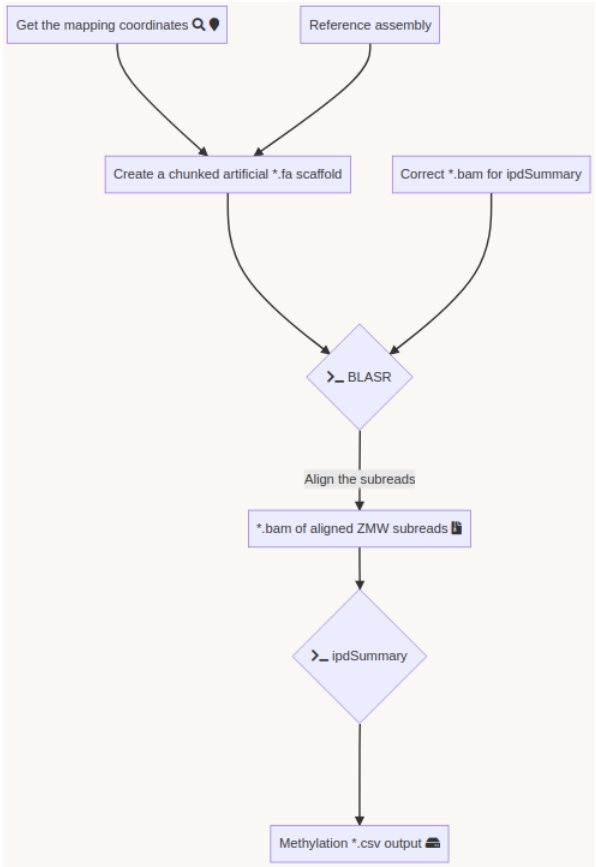


Figure 8: Illustration of the second phase of the proposed pipeline (B)

We take the *.bam files recreated in Phase II.A, and we align the subreads with BLASR (separately for each ZMW) against the reference genome.

Then, we launch the DNA methylation analysis with ipdSummary, separately for each ZMW.

Here, the analysis is made by aligning the subreads against a "chunked" version of the original genome assembly rather than the full file. The reasons behind this choice are described in details page 14 (See "Concerns about performances and parallelism").

Three additional operations are also performed but not mentioned in Figure 6 7 or 8:

- identificationQv are actually not produced by ipdSummary in its .csv output and must be gathered from a separated .gff file.
- We mark the last and first 15 nucleotides of ech molecule as "DNA molecule extremity", so that they can be censored in downstream analysis [1].
- ZMWs are actually processed in parallel batches.

2.3 Concerns about performances and parallelism

In theory it is possible individually to align the subreads against the whole genome assembly for each ZMW, and then launch ipdSummary against the whole genome assembly too. In practice, doing so would expose us to four pitfalls :

1 When mapped against the whole genome, the different subreads of a same ZMW will not necssarily align in the same region of the reference. This can be the case, for instance, where the ZMW originates from a duplicated region in the genome, because the subreads are prone to a high-error rate.
2 Before aligning any subread, BLASR computes a table of K-mers counts on the whole genome to optimize its search. This optimization is expensive and counter-productive in our case because we already know from Phase I where the consensus aligned.
3 ipdSummary is designed to analyze a single genome entirely. It therefore iterates all the positions in the reference genome to seek DNA methylation, including those where not a single subread is aligned.
4 When trying to make the analysis run in parallel, data-race and even deadlocks can occur because in each process, BLASR tries to access to the same file on the hard drive (the genome assembly).

Pitfalls n°2, n°3 et n°4 in particular, represent an extreme waste of computing time.

To address these 3 issues in Phase II.B, we only align the subreads of each ZMW against only a chunked version of the reference fasta. This truncated version of the original fasta is limited to the region of -50/+50 nucleotides where the CCS already mapped in Phase I. This simple fix allows to speed up the analysis by more than a factor 10. To keep the genomes coordinates correct in the final .csv file, we also re-shift them at the end (not shown in Figure 8)

Finally, up to than 350.000 ZMW can be analyzed in a Sequel I SMRTCell, and even more in a Sequel II SMRTCell. Analyzing them in an humanly acceptable time required to implement parallelization, which we did so using the python library *pandarallel* [29].

With *pandarallel*, The ZWMs are processed in batches of a user-defined size (default : 5000 ZWMs). Batching the analysis simplifies the output compilation, in addition to greatly improve the performances in our test configuration, as should also be the case for filesystems whose speed depend on the number of existing directories.

3 Results

In this section, we present the performances of our pipeline to detect 6mA, and what was our rationale behind our benchmarking method. Everywhere in this section, counts are indicated as counts of adenines whose effective coverage is superior or equal to 25X. Our pipeline's ability to detect 4mC, 5mC or any other kind of DNA modification has not been benchmarked.

3.1 Benchmarking rationale

To our knowledge, the only Sequel I - SMSN data available to date is our own, and we are not aware of any other post-RSII publication that used the SMSN strategy. We could therefore not count on other data in the literature to benchmark our pipeline.

Our data consists in 10 DNA samples of *P. tetraurelia*, with 10 different experimental conditions whose purpose is described in Table 4 Page 16. This *P. tetraurelia* DNA could not be used to benchmark our pipeline, since we had only little information about its methylation

Fortunately, our *P. tetraurelia* cells were fed with *E. coli* cells. We took advantage of the abundant presence of *E. coli* DNA to benchmark our pipeline's ability to detect 6mA, since the latter is very well described in *E. coli* [30][31][32][33] :

- The two adenines in GATC sites must almost always be n6-methylated *E. coli*, and the methylation is kept symmetrical after each DNA replication by the *Dam* methyltransferase. Dam^- strains are not viable enough to be found in nature or

[1] This subtlety has originally been introduced in [2] to correctly handle the fact that the context near the adapter sequence is not the one of the reference genome, which thus biases the analysis.

even in laboratory outside of specific protocols, which is why we expect that nearly 100% of GATC sites should be symmetrically methylated, let aside a negligible steady-state after DNA replication. These $GA^{N6m}TC$ sites represent a vast majority of the n6mA in *E.coli*.
- A vast majority of *E.coli* laboratory strains also have secondary motifs methylated by EcoK1 methyltransferases: **AAC(N6)GTGC** and its reverse complementary **GCAC(N6)GTT**. These sites are also expected to be symmetrically methylated. They represent a much smaller amount in the total of n6mA in *E. coli*.
- Other strain-specific or experiment-specific motifs of 6mA can also exist. Online tools such as REBASE [34] reference these variations. They are usually marginal in the total of n6mA.

Based on this paradigm, our rationale was simple : check if our pipeline is capable to detect n6mA in these sites, and capable of not reporting 6mA elsewhere.

It is possible that not all GATC and EcoK sites are symmetrically methylated in our DNA, or that 6mA also exist outside of these sites. These objections are ignored for the moment because they will be debated later (see "Discussion").

3.2 Identification of *E. coli* DNA molecules

The first step to benchmark our pipeline was to create the consensus, and identify the ZMW that corresponded to *E. coli* contaminants.

For each of our 10 sequencing datasets described in Table 4, we counted the number of reads available in the raw PacBio files produced by the sequencer and we created as much consensus as possible (Table 5).

Among the revendicated 1M ZWMs of a PacBio SMRTCell, we were able to create a consensus for about 300-350K ZMWs per SMRTCell. This number fits with the expectations.

The medium/mean size and coverage also suggest that the data can be analyzed with the SMSN approach.

From these 350K consensus per SMRTCell, we wanted to isolate the ones that corresponded to DNA molecules from *E. coli*.

Unfortunately, we dit not have access to the reference genome of our *E. coli* strain. Empirical BLAST [35] analysis however (data not shown), indicated that a large amount of these consensus did align correctly against various *E. coli* assemblies described in the literature. Among all the candidate reference assemblies, we chose to use an O157:H7 reference assembly (Z_AVCD01000005.1) as a proxy for the reference genome of our *E. coli* strain.

We aligned our consensus against this *E. coli* genome reference, and considered that any consensus that mapped with a percentage of identity superior to 95% (See equation (3)) should be treated as a DNA molecule from *E. coli*. This threshold of 95% was intentionnally set lower than the actual expected accuracy of the consensus, to take account from the fact that the reference genome might be slightly different to the real genome of our strain.

With this method we identified a total of 3955 *E. coli* molecules in which at least one adenine was covered at 25X. This corresponds to 555571 adenines covered with more than 25X (Table 2).

Experiment	Number of E.coli CCS with at least one adenine covered at 25X	Number of adenines covered at least at 25X
HT2	1	286
HT6	0	0
HTVEG	942	155447
MAB	493	66135
MT1A-1B	1147	155838
MT1A-1B-2	224	30558
MT2	171	23526
NM4	477	64488
NM4-9-10	141	22214
NM-9-10	359	48661

Table 2: Summary of number of *E. coli* consensus identified in each of the 10 DNA samples, and number of adenines with more than 25X of effective coverage

Number of the circular consensus that map on the O157:H7 NZ AVCD01000005.1 genome reference of *E.coli* with more than 95% of identity, and have at least one adenine covered by at least ≥ 25 exploitable IPDS.

As expected, the samples HT2 and HT6 (which were starved to induce autogamy in the *P. tetraurelia* cells) contained almost no identifiable *E.coli* DNA (resp. 1 and 0 DNA molecule).

When adding together the data of our 10 samples, the total amount of *E. coli* adenines for which a DNA methylation analysis can be carried was 9656 adenines located in a GATC site, 252 for the EcoK sites, and 557245 in other sites (See Table 3).

other	557245
GaTC	9656
AaC(6N)GTGC	125
GCaC(6N)GTT	127

Table 3: Raw counts of adenines per motif (All *E. coli* molecules, ¿25X)

Among the 9656 adenines located in a GATC site, 9258 belong to a GATC site where the adenine is covered with $> 25X$ on the two strands of the site.

Sample name	P. tetraurelia phase (cell cycle)	Biological description	IPTG-induced E. coli	P. tetraurelia starvation	Expected number of E. coli DNA Molecules
HTVEG	Vegetative	WT	No	Not starved	+ + +
HT2	Autogamy	WT	No	Starved	-
HT6	Autogamy	WT	No	Starved (longer)	- - -
MAB	Vegetative	Control silencing	Yes	Not starved	+ + +
MT2	Vegetative	Silencing	Yes	Not starved	+ + +
MT1A-1B	Vegetative	Silencing	Yes	Not starved	+ + +
MT1A-1B-2	Vegetative	Silencing	Yes	Not starved	+ + +
NM4	Vegetative	Silencing	Yes	Not starved	+ + +
NM9_10	Vegetative	Silencing	Yes	Not starved	+ + +
NM4_9_10	Vegetative	Silencing	Yes	Not starved	+ + +

Table 4: Biological summary of the 10 sequenced DNA samples

We sequenced 10 samples that contained mostly *P. tetraurelia* cells. Our primary goal was to analyze the DNA methylation of these *P. tetraurelia* DNA molecules, but the cells were fed with an unsequenced strain of *E. coli*, so that we expect the latter DNA to represent a significant proportion of the sequenced molecules. In HT2 and HT6, the paramecia were starved to induce autogamy (see [36] [37] [38]), this is why no *E. coli* cell is expected to remain in the middle. This is especially true in HT6, where the cells were starved several hours longer (~ 4 more hours). In the silenced conditions, *E. coli* cells were genetically engineered (IPTG-Induced - see [39]) to produce non-coding RNAs in order to silence specific methylase candidate genes in *P. tetraurelia* (See [40]). When wild-type cells were sequenced, the *E. coli* cells in the medium are also wild type cells.

Sample	HTVEG	HT2	HT6	MAB	MT1A-1B	MT2	MT1A-1B-2	NM4	NM9-10	NM4-9-10
Number of subreads	28363462	8610515	7474650	11026315	14534167	13228479	13322003	13027953	12825328	**24371865**
Nb of reads	374901	255148	292660	136045	178285	168147	166270	144195	149254	**326125**
Mean number of subreads per read	75,66	33,75	25,54	81,05	81,52	78,67	80,12	90,35	85,93	**74,73**
Median number of subreads per read	58	20	13	63	58	56	62	72	68	**54**
Number of reads with >= 50 subreads	202752	66010	51482	76628	96462	89054	93308	85610	87447	**170359**
% reads with >= 50 subreads	54,08 %	25,87 %	17,59 %	56,33 %	54,11 %	52,96 %	56,12 %	59,37 %	58,59 %	**52,24 %**
Nb of reads with CCS	350120	246555	271903	121907	142773	160415	130116	134918	141668	**309490**
% of reads with CCS	93,39 %	96,63 %	92,91 %	89,61 %	80,08 %	95,40 %	78,26 %	93,57 %	94,92 %	**94,90 %**
CCS median size	381	378	368	319	309	378	318	315	302	**369**
CCS average size	415,96	432,93	436,49	358,51	341,76	347,84	346,65	641,49	338,36	**403,49**

Table 5: Summary of the circular consensus creation step (10 samples)

4 samples were sequenced on individual SMRTCells : HTVEG, HT2, HT6, NM4_9_10. About the third of the theoretical SMRTCell capacity (1 million ZWM) produced continuous long reads for these samples. The other SMRTCells on the other hand were loaded similarly but were multiplexed two by two, hence their halved amount of reads. Starting from the subreads, consensus were built for each of the 10 DNA sequencing, with the most laxist parameters allowed by PacBio's CCS tool. The parameters used were the following : –minLength 50 –maxLength 50000 –minPasses 0 –minPredictedAccuracy 0.5
In all conditions there is an important amount of reads for which no consensus could not be built. This can be due either to chimeric DNA molecules between physically distinct molecules when building the library, or reads for which the CCS software is not capable of separating the subreads into two reverse-complementary categories. It can also correpond to cases where the read had less that one full pass, or reads with a predicted accuracy that is extremely low. Alltogether, the number of consensus produced, their size, and the number of subreads per ZMWs are in the expected order of magnitude.

3.3 Inspection of the p-values produced by ipdSummary

Because the p-values produced by ipdSummary were originally designed for the AggSN approach, using them in the SMSN approach as we do raises the question of whether or not they can be used, and how they should be interpreted.

In the AggSN approach, their interpretation is not straightforward: AggSN p-values are only correlated with the propensity of a given site to carry DNA methylation. Whether this methylation is abundant or not, if it even exists, must be determined in a second step by estimating the fraction of methylated molecules.

In the SMSN approach however, things are much easier to interpret : either a nucleotide is methylated, or it is not.

Because of this, we simply expect methylated nucleotides to produce low p-values, and non-methylated nucleotides to produce uniformely distributed p-values. We show in Figure 9 that in our SMSN data, the first tendancy is confirmed but the second is not.

3.4 p-values and FDR control

The distribution of the p-values in our SMSN data is U-shaped, which is pathological (Figure 9).

Usually, this shape indicates that a one-sided test was performed whereas the alternative hypothesis is actually two-sided.

The main danger statisticians fear in this situation, occurs when using adaptative False Discovery Rate (FDR) control procedures. Indeed, adaptative FDR control first estimates the proportion of H_0 in all tested hypothesis, to then reject H_0 with an optimal power. If a second alternative hypothesis yields p-values close to 1, then the proportion of H_0 can be overestimated, leading to be too conservative.

Here, the situation is more complex. Indeed, p-values close to 1 correspond to cases where the polymerase went significantly *faster* than expected by the in-silico model. Biologically speaking, these cases truly correspond to truly unmethylated nucleotides, that is to H_0. This is why depending on the algorithm used to estimate the proportion of H_0, our pathological distribution may or may not lead to overestimate it.

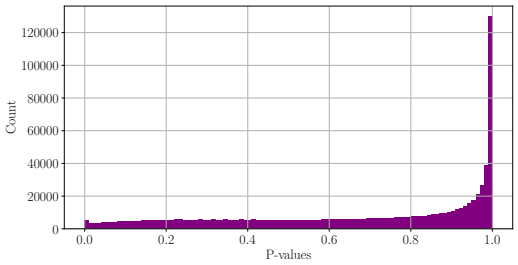
In Table 6, we compare different popular FWER and FDR control procedures on our *E. coli* SMSN data, using the python package *multipy* [41]. The Storey's estimator cannot converge, which is due to the pathological shape of the p-value. Other adaptative methods do not suffer from this problem, which underlines that the U-shaped distribution can be (but is not necessarily) an issue.

Strategy	Motif	#H0 rejected	#H0 !rejected	FDR
Adaptative BKY	AaC(6N)GTGC	111	14	13,30 %
	GATC	8898	758	
	GCaC(6N)GTT	112	15	
	other	1399	555846	
Adaptative Two-step LSU	AaC(6N)GTGC	111	14	13,15 %
	GATC	8898	758	
	GCaC(6N)GTT	112	15	
	other	1381	555864	
FWER Bonferroni	AaC(6N)GTGC	66	59	2,45 %
	GATC	6633	3023	
	GCaC(6N)GTT	56	71	
	other	170	557075	
FWER Hochberg	AaC(6N)GTGC	66	59	2,45 %
	GATC	6636	3020	
	GCaC(6N)GTT	56	71	
	other	170	557075	
FWER Holm Bonferroni	AaC(6N)GTGC	66	59	2,45 %
	GATC	6636	3020	
	GCaC(6N)GTT	56	71	
	other	170	557075	
FWER Sidak	AaC(6N)GTGC	66	59	2,45 %
	GATC	6638	3018	
	GCaC(6N)GTT	56	71	
	other	170	557075	
Non adaptative BH-LSU	AaC(6N)GTGC	111	14	13,30 %
	GATC	8898	758	
	GCaC(6N)GTT	112	15	
	other	1399	555846	
Adaptative Storey	??	??	??	??

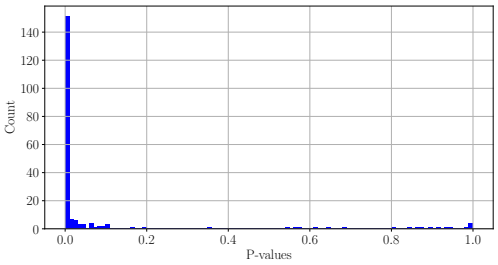
Table 6: Number of rejections of H_0 with different FDR-control and FWER-control procedures for q=5%, in *E. coli*, depending on the motif

Several popular FWER and FDR control procedures are compared on the adenines of *E. coli*. The FWER-control procedures are too conservatives, which is one of their usual characteristics. If we consider that all GATC and EcoK are methylated and that there is no other methylated site in *E. coli*, the computed FDR is about 13% with both adaptative and non-adaptative FDR-control methods. In theory, this FDR should be less than 5% by definition. One obvious hypothesis to explain this result is that other secondary methylation sites might exist in this *E. coli* strain, and the FDR *is* really less than 5%. Another hypothesis is that some nucleotides can generate long IPDs even if they are not methylated, or that the in-silico model can give biased values in some some contexts. Because the p-values are severely U-shaped, adaptative methods do not provide a significant power gain over non-adaptative ones and the Storey estimator cannot converge.

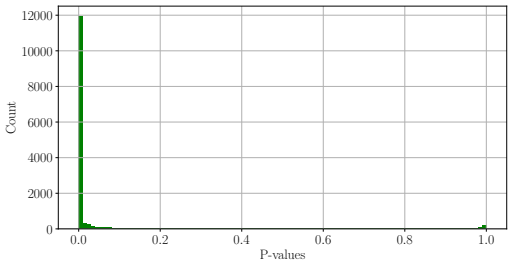
Asides from the p-values, ipdSummary also outputs their PHRED-transform score, which can be interesting alternatives to the p-values.



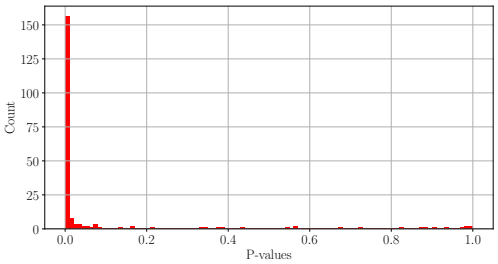
(a) All adenines outside of GATC and EcoK sites



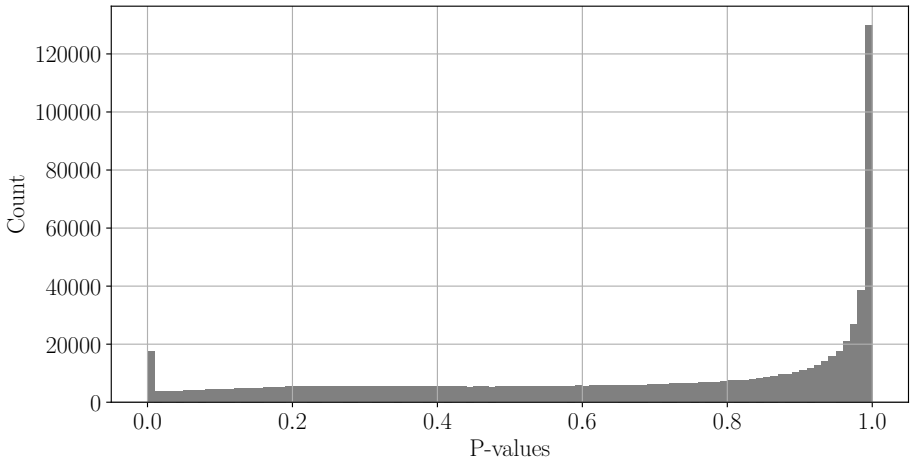
(b) Adenines within AaC(6N)GTGC sites



(c) Adenines within GaTC sites



(d) Adenines within GCaC(6N)GTT



(e) All adenines together

Figure 9: Compared distributions of p-values produced by PacBio's ipdSummary between EcoK1 sites, GATC sites and other sites in *E.coli*, when used in a SMSN approach.

A) Outside of GATC and EcoK sites, many p-values are close to 1 because they correspond to cases where the polymerase went faster than expected - hence a very thin probability that the nucleotide is modified. Otherwise, the quasi-absence of p-values near 0 and the relatively flat distribution otherwise is expected when the adenines are not methylated. B) C) D) Among the GATC and EcoK sites, adenines have p-values very close to 0, as expected when adenines are methylated. E) The global distribution is a pathological U-shaped. This distribution can be incompatible with some FDR control procedures.

3.5 Interpretation of the AggSN scores in the SMSN approach

The scores produced by ipdSummary are just PHRED-transforms of the original p-values.

In the AggSN approach, one key property of the modification score is to be linearly correlated with the coverage in presence of 6mA. This correlation is usually visualized by plotting the scores against the coverage, as is done in Figure 10, reproduced from [3].

To determine if this correlation still exists and can be used to call DNA methylation on SMSN data, we compared the different cov-score plots in the adenines of our *E. coli*, depending on which motif they are located (GATC sites, EcoK sites, others) - Figure 12. This works shows that the linear correlation indeed holds for our SMSN data too.

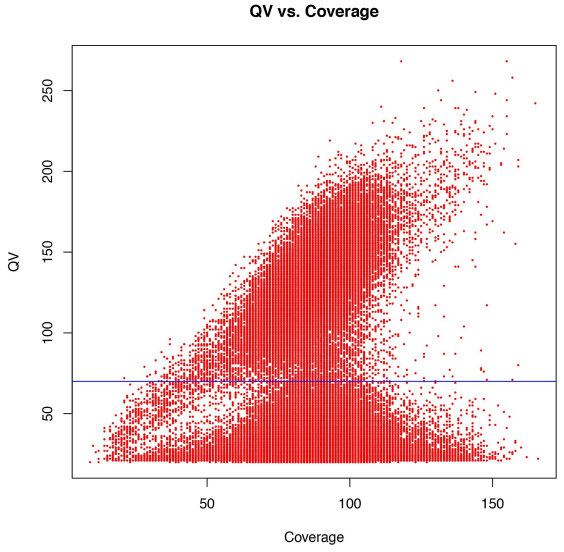


Figure 10: Typical AggSN Cov-score plot in presence of 6mA (reproduced from [3], Creative Commons, description modified)

Red dot represent adenines. In the AggSN approach, the scores produced by ipdSummary tend to be linearly correlated with their coverage in presence of 6mA. Here, the authors ([3]) called methylated all the adenines whose score is above a flat threshold represented in blue.

Despite their apparent simplicity, cov-score plots can be deceiving because both the scores and the coverages are discrete values. As a consequence, it is impossible to tell visually when several nucleotides occupy the same position in the graph. This can be addressed by transforming the plot with a kernel density estimation (KFE) - Figure 11.

In [3], the authors called methylated adenines with a flat threshold on the score that turns out to be an arbitrary choice. Another possibility is to pass a straight line through the density trough between the two clouds of points clouds (dashed lines Figure 11) to separate methylated and non-methylated nucleotides.

Here we propose to use the threshold defined by Equation (4) :

$$score \geq \frac{80}{115} \cdot coverage + \left(20 - 15 \cdot \frac{80}{115}\right) \qquad (4)$$

Equation (4) was determined empirically and subjectively by hand from the total cov-score of the adenines in *P. tetraurelia* (HTVEG). Figures 11 and 12 however, give an insight of how well this equation can be applied to other sequencing data and motifs.

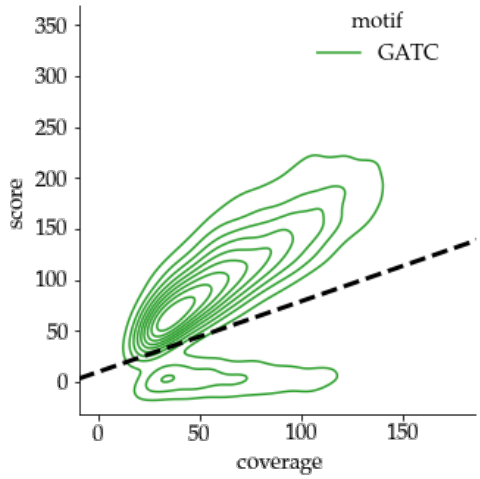
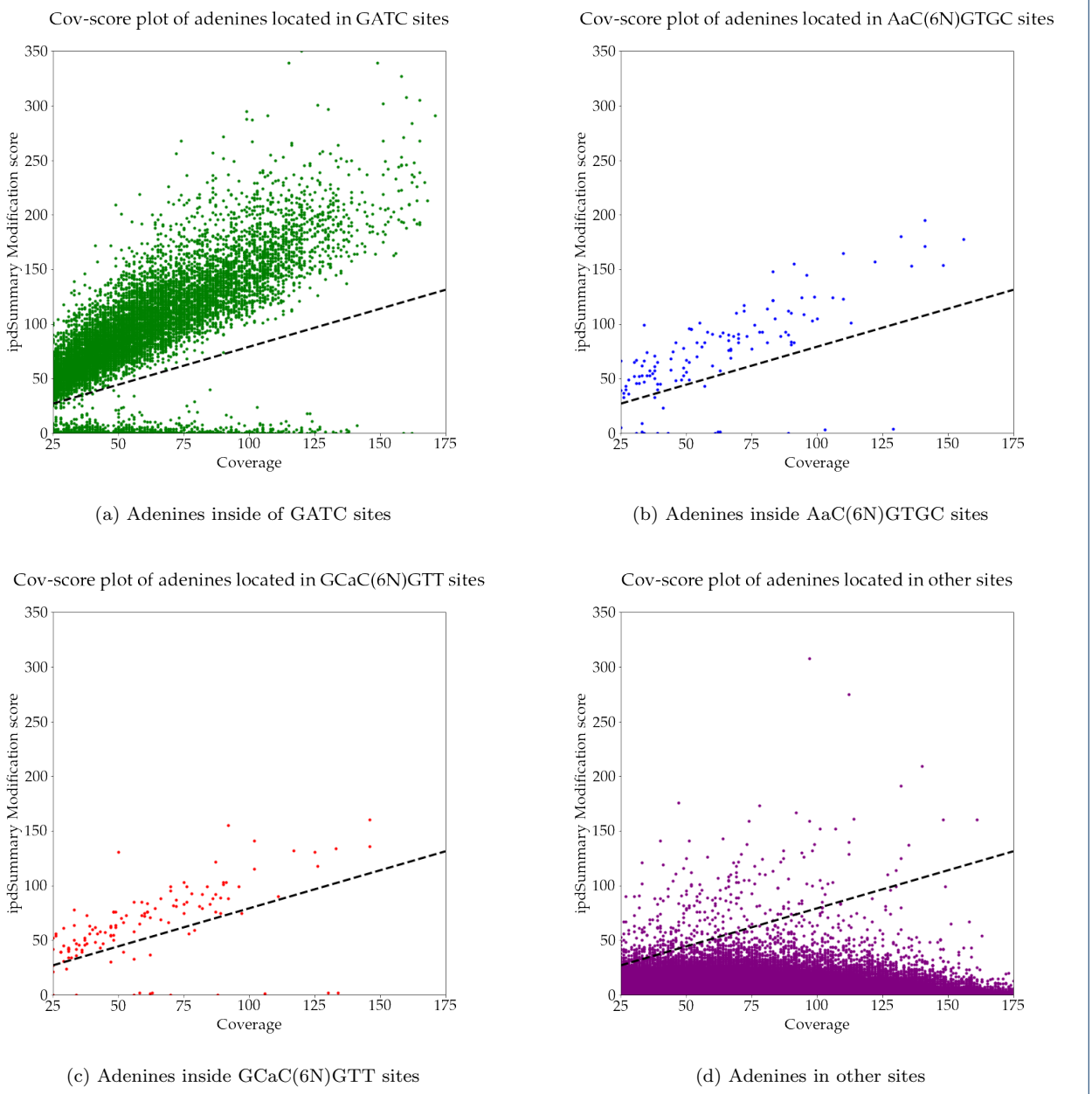


Figure 11: Density visualization of a cov-score plot)

Here, a plot similar to Figure 10 is done on *E. coli*'s adenines in the GATC sites. Instead of individual dots, the density has been estimated with a Kernel Density Estimation (KDE). The plot was realized only using adenines with $> 25X$ of effective coverage, even if because of its nature the KDE extends artificially to areas of negative scores and $< 25X$ coverages. The interest of such plots compared to pure dot plots is that it can highlight the differences in density between the two cloud of points (methylated adenines on top, versus the non-methylated adenines at the bottom). Here for instance, the density is much higher above the linear threshold than under it.



(a) Adenines inside of GATC sites

(b) Adenines inside AaC(6N)GTGC sites

(c) Adenines inside GCaC(6N)GTT sites

(d) Adenines in other sites

Figure 12: Per-motif comparison of the Cov-score plots for adenines in *E. coli*.

In the AggSN strategy, scores produced by ipdSummary are linearly correlated with the coverage in presence of 6mA, as shown in Figure 10. In a-b-c) We confirm the linear correlation between coverage and score for EcoK and GATC sites in our *E. coli* SMSN data. The dashed line represents the threshold defined by equation (4). d) Because the density of adenines is very unequal on this graph, it seems at first sight that many adenines are above the linear threshold (4) even for nucleotides that are supposed to be not methylated. It is however an optical illusion due to the fact that the density of adenines is extremely unequal, and the extreme majority of adenines have actually very low scores. More than 99.9% of the adenines are under the threshold (4).

Concerning the second kind of scores produced by ipdSummary (the "identificationQv"), it cannot be used alone to detect 6mA. It first requires that a nucleotide is first called "modified" with the "modificationQv". Only then, the specificity can be improved by adding a second threshold on the identificationQv, if necessary.

3.6 A worst-case estimation of Se and Sp

The main advantage of using p-values is that they it can be used alongside with FDR-control procedures. The scores used with a linear threshold on the other hand, have the very significant advantage that one can compute the Sensitivity (Se) and Specificity (Sp).

The sensitivity Se is the probability that a nucleotide is detected as methylated, when it really is :

$$Se = P(D^+|M^+) \tag{5}$$

The specificity Sp is the probability that a nucleotide is detected as not-methylated, when it is really not-methylated :

$$Sp = P(D^-|M^-) \tag{6}$$

Se and Sp can in turn be used to compute the true prevalence Π of methylation. Indeed when analyzing N nucleotides, the number p of positive detections equals to the sum of false positives and true positives :

$$p = [Se \cdot \Pi + (1 - Sp) \cdot (1 - \Pi)] \cdot N \tag{7}$$

In other words, when Se and Sp are known, we can have a precise estimate of what is the true fraction Π of methylated nucleotides :

$$(7) \implies \Pi \approx \frac{\frac{p}{N} - 1 + Sp}{Se - 1 + Sp} \tag{8}$$

In practice, Π is already an important quantity by itself. But additionally, it can also be re-used to compute the False Discovery Rate (FDR) and the False Non-discovery Rate (FNDR), two other important derived metrics.

Unfortunately, Se and Sp cannot be determined precisely without a proper gold standard, which we do not have because we can never be sure of what is the true DNA methylation of our molecules in our case. Fortunately, we can still have a good estimate of their lowest possible value.

Indeed, let $\widehat{Se}$ be the estimation of the real Se:

- We first compute $\widehat{Se}$ based on the fact that all GATC sites and EcoK sites are symmetrically methylated in *E. coli*.
- If our assumption is wrong and some adenines are in fact not methylated in EcoK/GATC sites, then $\widehat{Se} < Se$: We have a worst-case estimate of Se

Similarly, let $\widehat{Sp}$ be the estimation of the real Sp:

- We first compute $\widehat{Sp}$ based on the fact that there is no DNA methylation outside of EcoK/GATC sites
- If our assumption is wrong and some adenines are in fact methylated outside of EcoK/GATC sites, then $\widehat{Sp} < Sp$: We have a worst-case estimate of Sp

Using this reasoning, we provide worst-case estimates of Se and Sp for different threshold strategies (flat, coverage-dependant, etc) in Table 7.

These estimates indicate that 6mA can be detected with a great sensitivity and specificity with our pipeline.

Two important elements must be remarked:

- When we benchmark the flat thresholds of PacBio scores, our estimates of $\widehat{Se}$ and $\widehat{Sp}$ cannot be extrapolated from our data to other sequencing. Indeed, these scores are function of the coverage, which varies from one experiment to another.
- Among all the $\widehat{Se}$ and $\widehat{Sp}$ computed in Table 7, only the linear threshold can be extrapolated to other sequencing data.

In addition to this, 6mA is usually rare which means that investigating it without suffering too many false positives often requires very high specificity ($Sp > 99.9\%$). Among all the thresholds that allow this specificity, the linear-dependant coverage is by far the one that offers the best Se.

These observations are the reason why we advocate that a coverage-dependant threshold on PacBio's scores should always be preferred over a flat threshold. The performances of the linear threshold seem comparable to those of the p-values after FDR-control (see Figure 12), but the determination of Se and Sp with the linear threshold on scores offers a more direct probabilistic interpretation.

score	idqv	TP	FP	TN	FN	p	n	$\widehat{Se}$	$\widehat{Sp}$	$\widehat{FDR}$	$\widehat{FNDR}$	total
20	0.0	9153	5134	552111	755	14287	552866	92.38 %	99.08 %	35.93 %	0.14 %	567153
20	20.0	8363	652	556593	1545	9015	558138	84.41 %	99.88 %	7.23 %	0.28 %	567153
20	30.0	7861	373	556872	2047	8234	558919	79.34 %	99.93 %	4.53 %	0.37 %	567153
20	40.0	7206	248	556997	2702	7454	559699	72.73 %	99.96 %	3.33 %	0.48 %	567153
20	50.0	6441	178	557067	3467	6619	560534	65.01 %	99.97 %	2.69 %	0.62 %	567153
30	0.0	9124	1508	555737	784	10632	556521	92.09 %	99.73 %	14.18 %	0.14 %	567153
30	20.0	8360	504	556741	1548	8864	558289	84.38 %	99.91 %	5.69 %	0.28 %	567153
30	30.0	7858	325	556920	2050	8183	558970	79.31 %	99.94 %	3.97 %	0.37 %	567153
30	40.0	7205	230	557015	2703	7435	559718	72.72 %	99.96 %	3.09 %	0.48 %	567153
30	50.0	6441	172	557073	3467	6613	560540	65.01 %	99.97 %	2.6 %	0.62 %	567153
40	0.0	8974	670	556575	934	9644	557509	90.57 %	99.88 %	6.95 %	0.17 %	567153
40	20.0	8307	383	556862	1601	8690	558463	83.84 %	99.93 %	4.41 %	0.29 %	567153
40	30.0	7839	281	556964	2069	8120	559033	79.12 %	99.95 %	3.46 %	0.37 %	567153
40	40.0	7198	204	557041	2710	7402	559751	72.65 %	99.96 %	2.76 %	0.48 %	567153
40	50.0	6439	157	557088	3469	6596	560557	64.99 %	99.97 %	2.38 %	0.62 %	567153
50	0.0	8558	362	556883	1350	8920	558233	86.37 %	99.94 %	4.06 %	0.24 %	567153
50	20.0	8074	277	556968	1834	8351	558802	81.49 %	99.95 %	3.32 %	0.33 %	567153
50	30.0	7698	228	557017	2210	7926	559227	77.69 %	99.96 %	2.88 %	0.4 %	567153
50	40.0	7145	178	557067	2763	7323	559830	72.11 %	99.97 %	2.43 %	0.49 %	567153
50	50.0	6421	143	557102	3487	6564	560589	64.81 %	99.97 %	2.18 %	0.62 %	567153
>linear	*	9109	379	556866	799	9488	557665	91.94 %	99.93 %	3.99 %	0.14 %	567153

Table 7: Estimation of the Se and Sp of various thresholds on ipdSummary scores

Here, we compared the estimated sensitivity Se and specificity Sp when 6mA was called using different thresholds on the score and the identificationQv. This estimate was realized based on the principle that the vast majority of GATC and EcoK sites should methylated in our *E. coli* strain, and that these two sites should represent nearly 100% of the 6mA in the cells. TP: True Positives, FP: False Positives, TN: True Negatives, FN: False Negatives, p: Positive detections n: Negative detections. $\widehat{Se}$, $\widehat{Sp}$, $\widehat{FDR}$ and $\widehat{FNDR}$ represent respectively the estimate of the sensitivity, specificity, False Discovery Rate, and False non-discovery Rate based on the aforementioned assumptions.

score	identificationQv	#D=0	#D=1	#D=2	total	frac. non-methylated	frac. hemi-methylated	frac. sym-methylated
20	0.0	69	555	4140	4764	1.45 %	11.65 %	86.9 %
20	20.0	225	952	3587	4764	4.72 %	19.98 %	75.29 %
20	30.0	355	1149	3260	4764	7.45 %	24.12 %	68.43 %
20	40.0	600	1278	2886	4764	12.59 %	26.83 %	60.58 %
20	50.0	916	1366	2482	4764	19.23 %	28.67 %	52.1 %
30	0.0	73	570	4121	4764	1.53 %	11.96 %	86.5 %
30	20.0	226	952	3586	4764	4.74 %	19.98 %	75.27 %
30	30.0	356	1149	3259	4764	7.47 %	24.12 %	68.41 %
30	40.0	600	1279	2885	4764	12.59 %	26.85 %	60.56 %
30	50.0	916	1366	2482	4764	19.23 %	28.67 %	52.1 %
40	0.0	97	640	4027	4764	2.04 %	13.43 %	84.53 %
40	20.0	246	956	3562	4764	5.16 %	20.07 %	74.77 %
40	30.0	366	1146	3252	4764	7.68 %	24.06 %	68.26 %
40	40.0	603	1279	2882	4764	12.66 %	26.85 %	60.5 %
40	50.0	917	1366	2481	4764	19.25 %	28.67 %	52.08 %
50	0.0	194	812	3758	4764	4.07 %	17.04 %	78.88 %
50	20.0	326	1006	3432	4764	6.84 %	21.12 %	72.04 %
50	30.0	434	1145	3185	4764	9.11 %	24.03 %	66.86 %
50	40.0	634	1268	2862	4764	13.31 %	26.62 %	60.08 %
50	50.0	928	1361	2475	4764	19.48 %	28.57 %	51.95 %
>linear	*	73	577	4114	4764	1.53 %	12.11 %	86.36 %

Table 8: Estimation of the fraction of hemi-methylated, non-methylated or symmetrically methylated GATC sites with different thresholds on ipdSummary scores

Here, we compared the estimated fraction of non-methylated, hemi-methylated and symmetrically-methylated GATC sites using different thresholds on the PacBio scores and identificationQv. #D=0, #D=1 and #D=2 stand respectively for the number of GATC sites where 0, 1 or 2 adenines are detected as methylated with the given thresholds. To make these statistics, we polled together all GATC sites of *E. coli* in all our 10 DNA samples, which represents a total of 4764 GATC sites where both adenines were covered with $\geq 25X$ of effective coverage.

3.7 Effects of coverage and motif on ipdRatio

As we already mentioned, PacBio's scores can vary due to the coverage. Figure 14 illustrates the importance of this effect with the log(ipdRatio), in a way that is visually even more clearly than in Figures 11 or 11.

Figure 13 shows a second source of variability that is also important to take into account : the motif. At equal coverage, two methylated adenines in two different motifs will not have the same ipdRatio. As a consequence, it is easier to detect DNA methylation in some motifs than others. From Figure 13 we can guess that the methylated adenines in the EcoK sites for instance, are less easy to distinguish than the ones in the GATC sites.

In practice, this means that the estimates of Se and Sp we made earilier (Table 7), including the ones for the coverage-dependent threshold, cannot be used as an absolute truth when analyzing DNA methylation in new motifs.

Fortunately, the same observation was already made in Beaulaurier et al 2015 [2] where this motif-dependent effects could be considered neglictible in most cases.

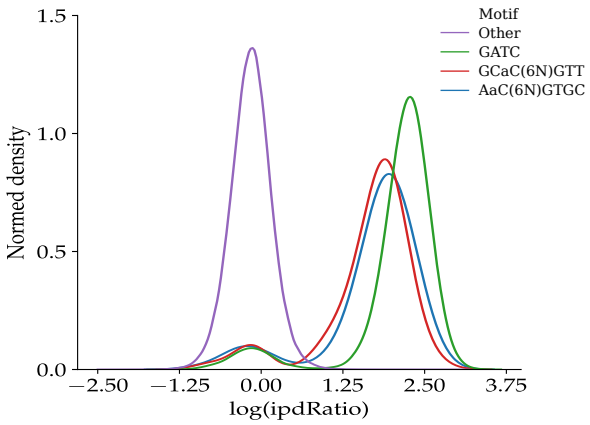


Figure 13: Normed KDE distribution of the log(ipdRatio) in adenines of *E. coli* covered at $> 25X$ of effective coverage, depending on the motif in which they are located.

4 Discussion

Our results demonstrate our pipeline's ability to detect 6mA. Several details however, must be taken into account correctly when using the pipeline to study DNA methlation on new data.

4.1 Guidelines to call DNA methylation with the proposed pipeline

As already mentioned, calling 6mA with flat thresholds on the scores produced by our pipeline should always be avoided. Low flat threshold are not associated with a sufficient specificity and produce too many false positives. High flat thresholds on the other hand do not provide sufficient power.

Coverage biases are also a very important bias when dealing with the scores produced by ipdSummary and cannot be ignored, which is why DNA modifications should always be done either using the p-values (and a proper FDR-control procedure) or a linear coverage-dependant threshold. The $\widehat{Se}$ and $\widehat{Sp}$ computed for the flat thresholds in Figure 7 in particular, cannot be extrapolated to other sequencing in the future.

When calling 6mA with a coverage-dependent threshold, it is the user's choice to define the linear equation of this coverage-dependent threshold. This choice can be done using a cov-score scatterplot (Figure 11). When possible, we advise to prefer using Kernel Density Estimate (KDE) transformation instead, because in general the density cannot be appreciated correctly with a scatter plot.

When using Sequel I data, our results indicate that the equation (4), although very empirical, produces a satisfying result even in data from physically distinct sequencing.

When calling 6mA with the p-values, one should be aware that the pathological distribution of the p-values under H_0 might break adaptative FDR-control procedures (e.g Storey's estimator) whereas non-adaptative FDR control procedures and FWER procedures work as expected.

As general advice, rare DNA methylation must always be studied with the most stringent methods to limit the FDR.

We also advocate from practical experience to respected the bare minimum of $25X$ effective recommended by PacBio to detect DNA methylation. When a particular experiment requires very high sensitivity and specificity, one can also limit the analysis to the most covered nucleotides; as the separability between methylated and unmethylated nucleotides always goes up with the coverage.

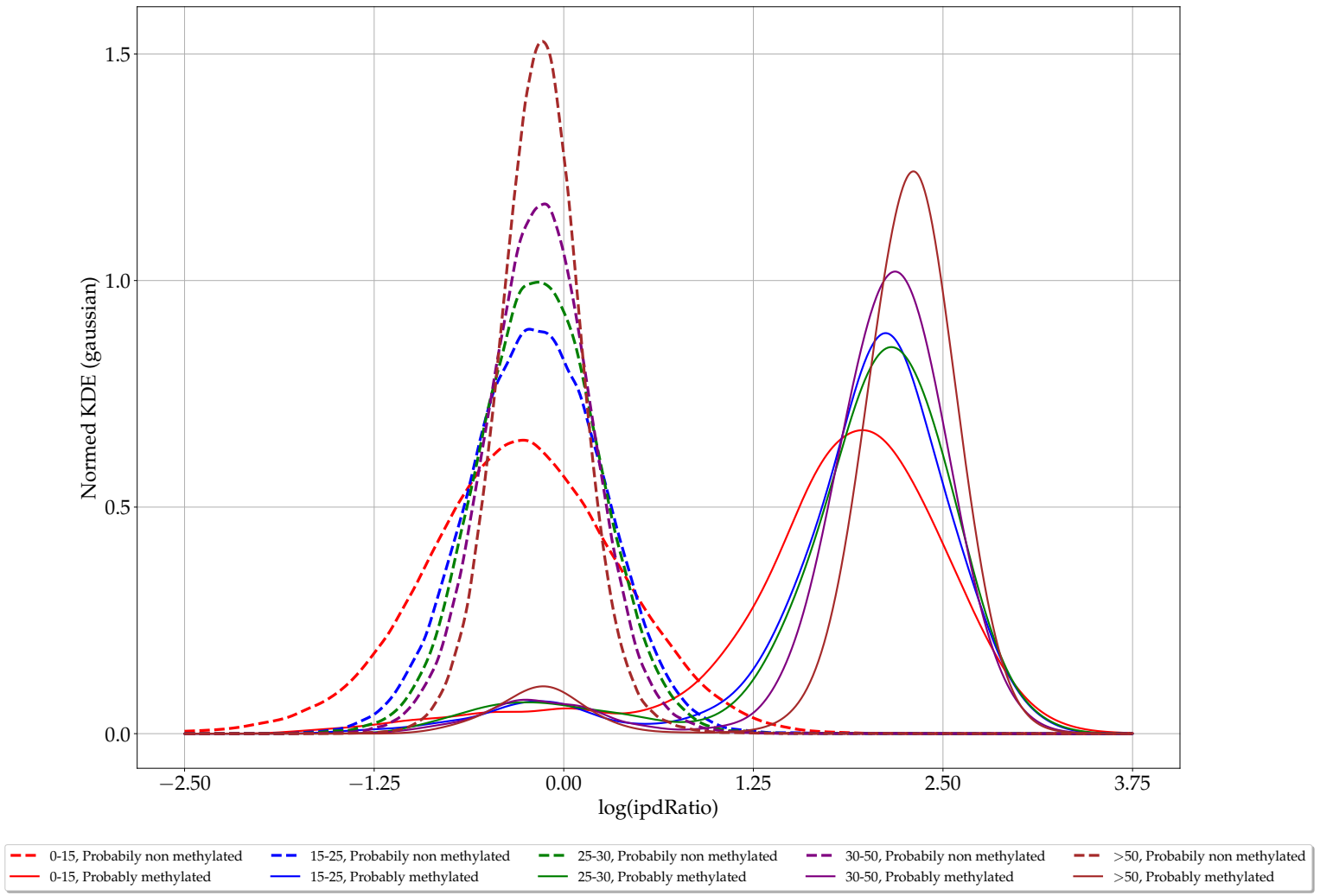


Figure 14: Distribution of the log(ipdRatio) of adenines depending of their coverage and likely methylation.

E. coli's adenines of different effective coverages are compared : from 0X to 15X (red), 15X to 25X (blue), 25X to 30X (green), 30X to 50X (purple) and >50X (brown). Plain lines correspond to adenines in GATC or EcoK sites, whereas dashed lines correspond to other adenines. Instead of histograms, normed densities (obtained with a gaussian Kernel Density Estimation (KDE)) are plotted, to allow easier comparison between the different cases. We can see that in terms of (log) ipdRatio, the separability increases between methylated and un-methylated adenines when the coverage increases. Some adenines that should be methylated in theory are actually concentrated around x=0. This can correspond either to cases of truly not-methylated adenines, or correpond to cases of truly methylated for which the sequencer did not observe high IPDs.

4.2 About the reliability of our estimates of Se

To determine our pipeline's sensitivity (Se) and specificity (Sp) toward 6mA when 6mA is called with our coverage-dependent threshold, we used the simplifying assumption that all GATC and EcoK1 sites were symmetrically-methylated, and that no other DNA methylation was present elsewhere in *E. coli*'s DNA.

In fact, our *E. coli* cells are not Wild-Type cells in long-stationnary phase. They are instead genetically engineered (IPT-induced) and used as food for our *P. tetraurelia* cells. This is why we can actually expect that a significant fraction of GATC and EcoK sites could be really not methylated in our sequencing data.

This hypothesis is supported by the data from Figure 12 and Figure 12, where the distribution of the scores (or ipdRatio) of supposedely methylated adenines is clearly bimodal, with one part of the distribution being clearly identical to unmethylated nucleotides, which suggests that these nucleotides are not methylated.

However, even if some GATC and EcoK sites are indeed not methylated, this would just imply that $\widehat{Se} < Se$ which can be considered as a good news.

The main concerns about the sensitivity come in fact rather from the fact that it is slightly motif-dependent (Figure 12), and that the total number of adenines on which this metric was computed (9908) is rather low to estimate precisely Se.

Despite these details, one can reliably expect the real Se to always be in at least an order of magnitude of $\sim 90\%$, with slight variations depending on the motif considered.

4.3 About the reliability of our estimates of Sp

Our assumption to estimate Sp was that no DNA methylation exists outside of GATC and EcoK sites in our *E. coli* strain. This point is not certain because the methylome of our *E. coli* strain is not well-known, and it is frequent to find strain-dependent methylations in *E. coli*.

Here again, our data tend to show that our assumption is false, and that these secondary sites of 6mA indeed exist in our *E. coli* strin.

The existence of these secondary sites is supported notably by Table 6, where we the expected FDR appears to be 13% with our simplifying assumptions whereas the mathematics theory tells us that the real FDR is less than 5%. It is also supported by our supplumental Figures (Appendix - Figures 17 and 18) where we show that the relative frequency of candidate secondary sites augments when we raise the specificity.

Here again, even if there is 6mA in secondary motifs, this would just imply that $\widehat{Sp} < Sp$, which can be considered as a good news.

Contrary to Se, Sp was determined on an extremely high number of adenines ($> 500K$), from a great variety of motifs. This ensures that the specificity is very well characterized, and subject to almost no statistical noise. When called correctly with a coverage-dependent threshold, one should expect the specificity to be reliably higher than 99.9%, which is sufficient for almost any application.

4.4 Impact of $\widehat{Se}$ and $\widehat{Sp}$ when estimating hemi-methylation

The most important interest of the SMSN sequencing over the AggSN sequencing is that it allows to study both strands of a same physical molecule independantly. This is especially useful when studying heterogeneity of DNA modifications between molecules.

Another major application of interest (whose feasibility was studied briefly here) is to quantify the fraction of sites corresponding to a given motif that are not-methylated, hemi-methylated, or non-methylated.

As shown in Table 7 (Which can also be visualized graphically in Figures of Appendix), such studies can be done with our method but the final interpretation is hard and should be subject to further methodological developments. Indeed, the results vary heavily with even the slightest change of Se and Sp. We believe a specific statistical methodology should also be developped to debiase such analysis.

5 Conclusions

Our pipeline is a proof of concept that n6mA can be deteted with the Sequel I sequencr, using the SMSN strategy described by J. Beaulaurier et al [2], even when no PCR-amplified control of the full genome is available. Such situation can occur for instance when doing metaepigenomics studies, when facing financial constraints, when sequencing DNA methylation of an unknown organism, or even when sequencing total DNA in ciliates (our motivational example).

The proposed pipeline's Sensitivity and Specificity toward 6mA are respectively above 92% and 99.8% when 6mA is called with a coverage-dependant threshold. These performances are sufficient for most applications.

The sensitivity (power) can be increased by putting higher thresholds on the coverage than 25X to call

DNA methylation, whereas the specificity can be increased by rejecting H_0 starting from lower p-values or by using higher thresholds on the modification scores and identification scores. The p-values produced are compatible with FDR-control procedures.

The final pipeline heavily relies on PacBio tools that were originally meant for the AggSN strategy. All the required softwares are embedded in a conda virtual environment for the final user, that can be installed easily without root access. The pipeline itself is implemented as a standardized Command-Line-Interface (CLI) python application, which can be installed easily with pip.

We hypothesize that only few modifications, if any, would be required to adapt this pipeline to the newest Sequel IIv2 sequencer, or to detect 4mC.

6 Availability, technical usage and requirements

6.1 Main specifications

- Platform : Computer x86-64 bits
- OS : Tested on Ubuntu LTS 18 and 20
- RAM requirement :
  - $\sim 2GB$ of RAM per processor (default)
  - 0.15Mb per ZWM
  - Both conditions must be met, but not added
- Format : Command-Line-Interface (CLI)
- Programming Language : Python 3.7
- Software requirements : Conda 4.6
- Parallelism : Yes (one machine only)
- Behaviour : Deterministic
- Requires root to be installed : No
- License : MIT 2.0
- Availability (Code and documentation) : `https://github.com/EMeyerLab/SMSN`

6.2 Installation procedure

Known or possible installation issues are documented on the GitHub README of the project. The following bash commands will install the pipeline :

```
git clone https://github.com/GDelevoye/SMSN.git
conda env create -n smsn -f ./SMSN/environment.yml
conda activate smsn
pip install ./SMSN/
```

6.3 Available chemistries

The in-sillico models are created and published by Pacific Biosciences under a specific copyright license at `https://github.com/PacificBiosciences/kineticsTools`.

Our software redistributes three of these models for the final user: P6-C4, SP2-C3 and SP3-C3.

SP2-C2 is recommended for Sequel I chemistries, and is the model that was tested in this article. SP3-C2 is recommanded for Sequel IIv2 chemistries, and was not tested in this article.

6.4 Output files

The pipeline outputs a tabular.csv file that contains the following columns:

- HoleID: Unique identifier of the DNA molecule
- nucleotide: Which nucleotide was sequenced
- scaffold: Its scaffold and position in the genome
- strand: Its DNA strand relatively to the reference assembly
- coverage: Its effective coverage [2]
- SMSN: Its Beaulaurier's SMSN scores
- context: Its surrounding nucleotide context
- ipdModel:The synthetic IPDs corresponding to this context if no methylation was present
- meanIPD: The mean of the experimental IPD, after pauses of the DNA polymerase were capped.
- ModQV/idQv: The usual PacBio modification and identificiation scores.

6.5 Computation time

Our implementation allows to process data from an entire Sequel I SMRTCell of 300.000 ZWM in less than a day, on an average consumer computer. The creation of the Circular Consensus (CCS) represents a significant part of this total. This is why we implemented a mode where the user can feed its own consensus, in addition to the raw subreads.

6.6 RAM requirement and final concatenation

The RAM indicated in the requirements is indicative. Indeed, it is function of the batch size, and of the average number of lines corresponding to each molecule in the final output (Longer molecules will require more RAM per molecule in the final modification.csv file).

[2] The effective coverage is the number of subreads that were of sufficient quality to be included in the analysis for this nucleotide

6.7 Temporary files and behaviour in case of crash

As previously said, the time to generate some intermediate files (e.g : CCS) or the methylation analysis itself is significant.

At the end of the pipeline, the program tries to compile all the outputs of all the batches into a single output .csv file.

If it fails to do so (for instance: Insufficient RAM), then all intermediary batch results are preserved to be compiled manually by the user, usually with a simple bash command [42].

6.8 Test configuration

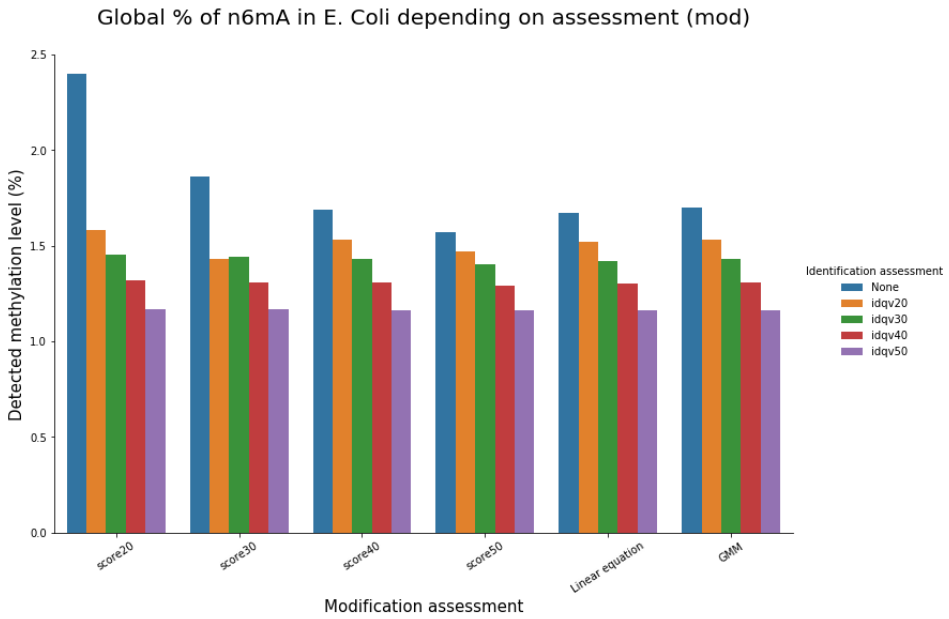
OS: Ubuntu LTS 20.04 CPU: AMD Ryzen R5 2600x (12 threads, 6 physical cores) Processor @ 3200MHz RAM: 32 GB

The functional and unit testing also worked properly on a Docker image of Ubuntu LTS 18 in circleCI. Other configurations might also work correctly, but were not tested.

6.9 Command-Line-Interface

The Command-line interface was implemented with argparse and follows the standards in the domain. When asked to display the help, users will face the standardized documentation described in Figure 19 - Appendix.

7 Appendix



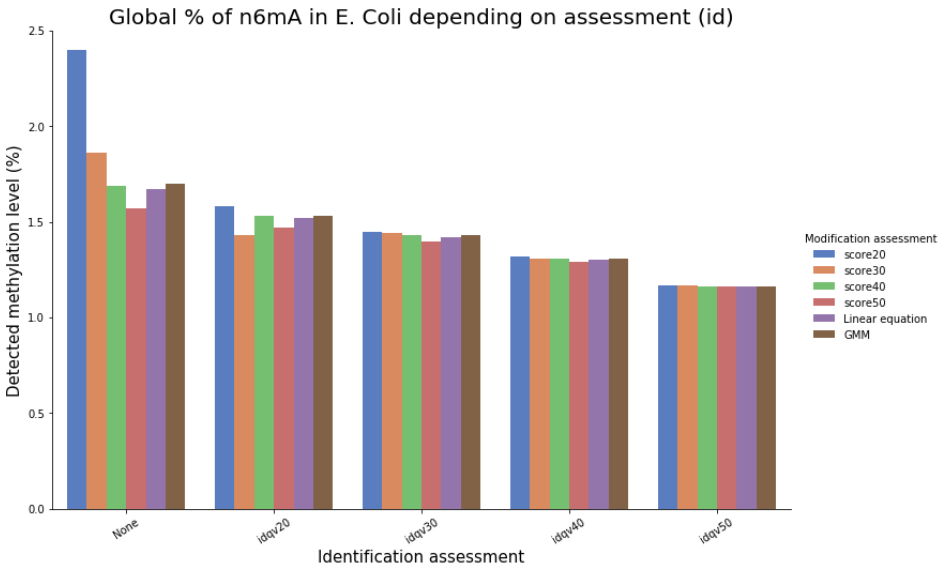


Figure 15: Impact of the identificationQv in the estimation of the global level of 6mA in E. coli (SMSN).

This figure is the graphical representation of the data presented in Table 7.

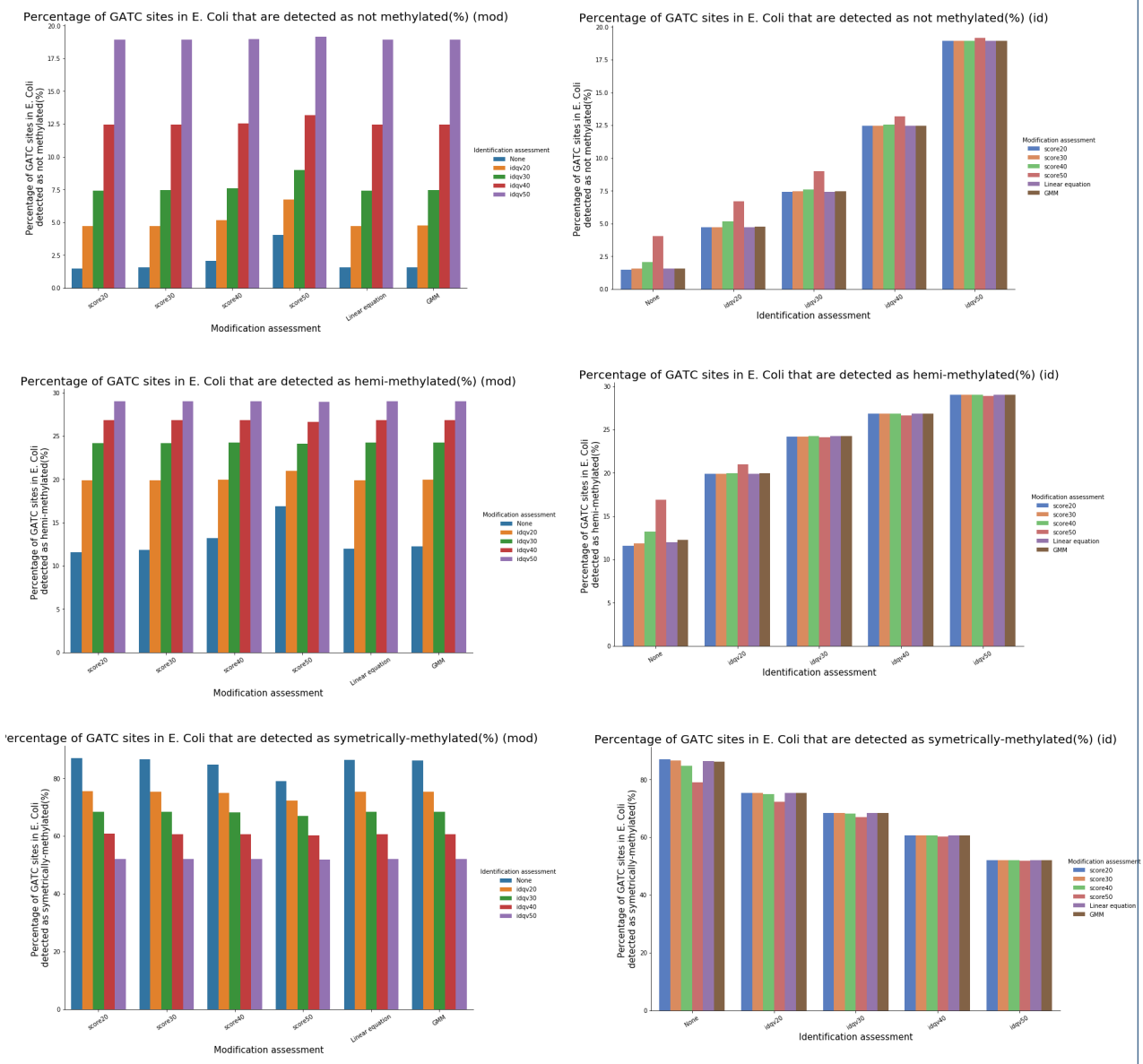


Figure 16: Impact of the identificationQv in the estimation of the level of 6mA in the GATC sites of E. coli (SMSN).

This figure is the graphical representation of the data presented in Table 8.

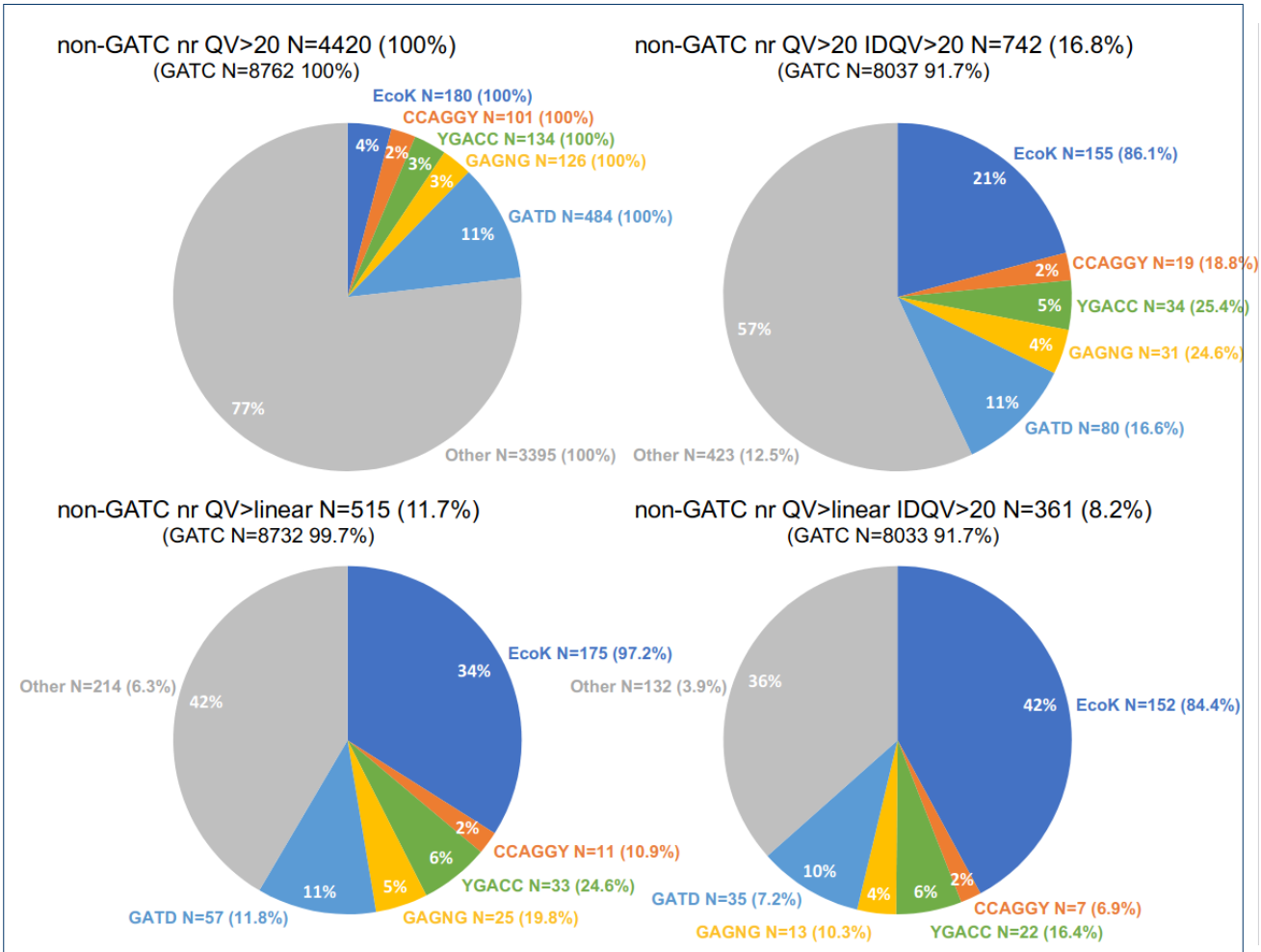


Figure 17: Fraction of the different motifs among the methylated, depending on the criteria chosen to call DNA methylation (all coverages).

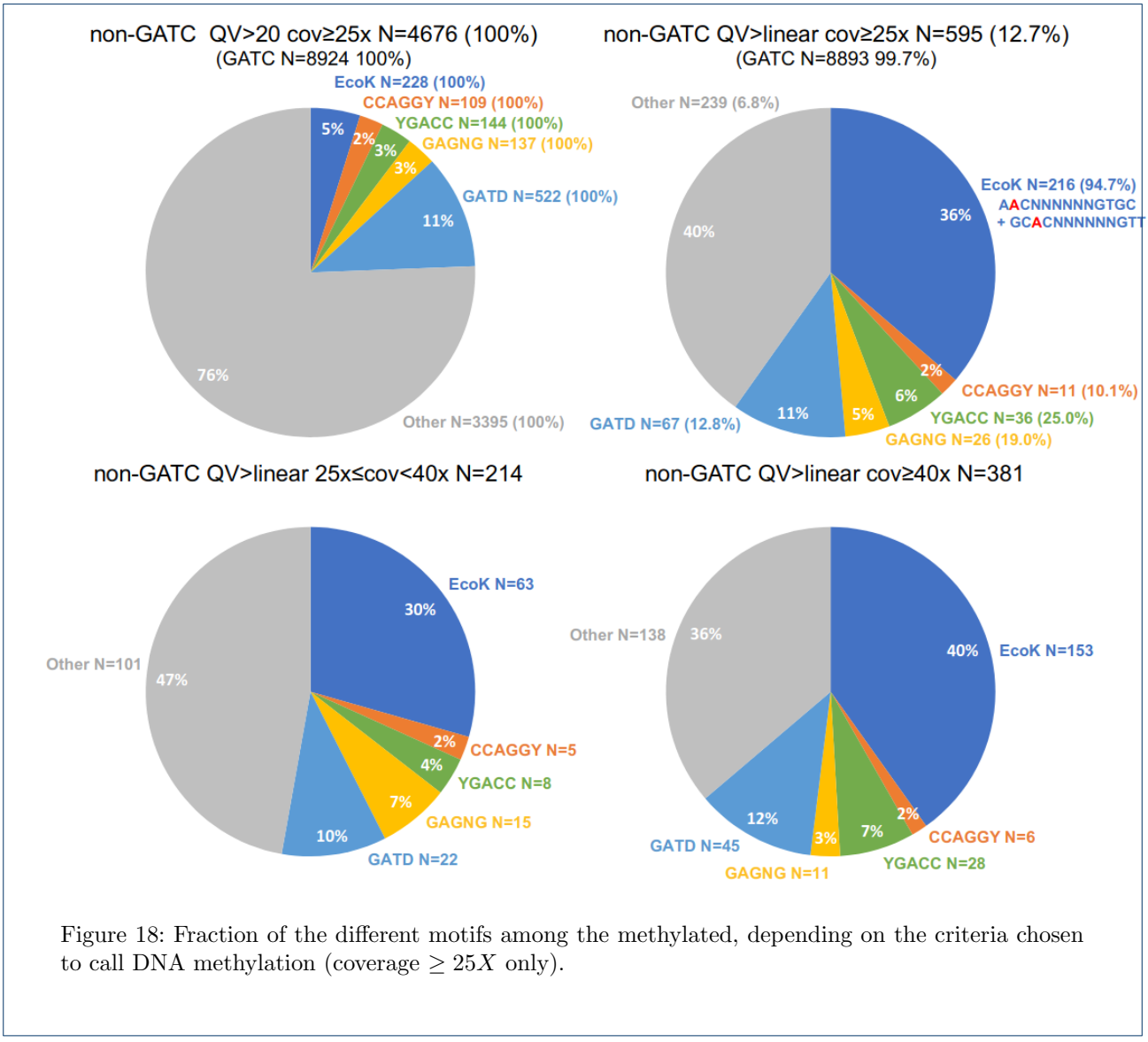


Figure 18: Fraction of the different motifs among the methylated, depending on the criteria chosen to call DNA methylation (coverage $\geq 25X$ only).

```
usage: smsn [-h] --bam BAM --reference REFERENCE --model {SP2-C2,SP3-C3,P6-C4}
            --output_csv OUTPUT_CSV [--CCS CCS] [--min_identity MIN_IDENTITY]
            [--min_subreads MIN_SUBREADS] [--tmpdir TMPDIR]
            [--verbosity {DEBUG,INFO,WARNING,ERROR,CRITICAL}]
            [--progress_bar PROGRESS_BAR] [--nb_proc NB_PROC]
            [--sizechunks SIZECHUNKS] [--add_context ADD_CONTEXT]
            [--idQvs IDQVS]

This software implements the SMSN approach of PacBio sequencing in similar way
as Beaulaurier et al 2015, except that it relies on PacBio ipdSummary and can
be used even in the absence of a PCR-amplified control.

 Tested on Sequel I data only.DELEVOYE Guillaume 2020.
 https://github.com/EMeyerLab/SMSN

required arguments:
  --bam BAM, -b BAM     Path to a .bam file with all the subreads (adapters
                        sequences must be already removed)
  --reference REFERENCE, -r REFERENCE
                        Path to a genome reference (fasta file).
  --model {SP2-C2,SP3-C3,P6-C4}, -m {SP2-C2,SP3-C3,P6-C4}
                        Choose the model for IPD prediction. [DEFAULT: auto
                        (PacBio's kineticsTools softwarechoses after it has
                        parsed the input file)]
  --output_csv OUTPUT_CSV, -o OUTPUT_CSV
                        Ouput file (csv) of the methylation analysis. See the
                        README.md for further details on the output's format.

optional arguments:
  --CCS CCS, -c CCS     [FACULTATIVE] Path to the circular consensus
                        corresponding to the .bam subreads. Default = CCS will
                        be recreated from the subreads provided.
  --min_identity MIN_IDENTITY, -i MIN_IDENTITY
                        minimum identity (percentage) of the CCS required to
                        launch analysis on a hole.[DEFAULT : 0.99]. Must be in
                        [0;1]
  --min_subreads MIN_SUBREADS
                        Minimum number of subreads required to launch analysis
                        on the hole. DEFAULT = 50 (so that its possible to
                        have >=25X per strand on at least one position).
  --tmpdir TMPDIR, -t TMPDIR
                        Tmp directory (DEFAULT : smsn_tmpdir_[DATE_HOUR])
  --verbosity {DEBUG,INFO,WARNING,ERROR,CRITICAL}, -v {DEBUG,INFO,WARNING,ERROR,CRITICAL}
                        Choose your verbosity. Default: INFO
  --progress_bar PROGRESS_BAR, -p PROGRESS_BAR
                        Displays a progress bar. Disabled automatically if
                        verbosity is set to debug[DEFAULT]: False
  --nb_proc NB_PROC, -n NB_PROC
                        Multiprocessing on n CPU. Default: 1.
  --sizechunks SIZECHUNKS, -k SIZECHUNKS
                        The subreads will often not fit entirely in RAM and
                        the methylation analysis itself generates lots of I/O
                        usage. Because of it, smsn will pause and perform
                        intensive I/O operation every S holes it has analyzed.
                        Lower values are better for machines that are limited
                        in RAM. The optimal nb_proc/sizehunks ratio will vary
                        from one computer to another. In case sizechunks <
                        nb_proc, SMSN will use sizechunks = 20x nb_proc
                        instead. DEFAULT : 5000
  --add_context ADD_CONTEXT
                        In the output .csv file, displays the +12/-12 context
                        around the nucleotide.(Files generated can be
                        sensitively heavier) [DEFAULT: True, choices = True or
                        False]
  --idQvs IDQVS         Outputs PacBio's identificationQV [DEFAULT: True,
                        choices = True or False]
```

Figure 19: Standardized help of the CLI.

This documentation is printed by our CLI software when the user asks for help. The in-sillico model must be specified, and models from other sequencers than the Sequel I should work correctly, but were not tested. It is notably possible for the user to change the format of the output, set the number of parallel processes, or to provide its own-made circular consensus (CCS), as this step is computationnaly expensive and it is frequent that PacBio users already have their own ready for any analysis.

Acknowledgements

We would like to thank the following people whose help and interactions, although punctual, were particularly useful to us:

- Pierre Vincens - Advices (Conda environment)
- Mael Lefeuvre - Advices (software deployment, bioinformatics platform)
- Leandro Quadrana - Advices (Biology)
- Chunlong Chen - Advices (PacBioo)

Computations were realized on the facilities of the bioclust platform, in the Institute of Biology of Ecole Normale Supérieure Paris (IBENS).

Funding

This work was funded by the French Agence Nationale de la Recherche (ANR) under the LaMarque project.

Abbreviations

- FDR: False Discovry Rate
- FWERR: Family-wise error-rate
- PCR: Polymerase Chained Reaction
- IPD: Inter Pulse Duration
- AggSN: Historical way of doing a DNA methylation analysis with PacBio-SMRT, that consists in sequencing very long inserts, and then aggregate the Inter-Pulse Durations of physically distinct DNA molecules to call DNA modifications.
- SMSN: A second way of doing a DNA methylation analysis with PacBio-SMRT sequencing, proposed by Beaulaurier et al 2015. This approach consists in sequencing shorter molecules many times, to gather many measurements for the inter pulse duration of the same nucleotide.
- SMRT: Single Molecule Real-Time; designates the ability of the PacBio sequencers to sequence each DNA molecule independantly of the others, without any amplification.
- SMALR: Software developped by J. Beaulaurier to analyze Single-Molecule SIngle-Nucleotide data
- CLI: Command-Line Interface
- n6mA: N6-methyladenine
- 5mC: 5-methylcytosine
- 4mC: 4-methylcytosine
- CLR: Continuous Long Read

Availability of data and materials

Data are not published yet, but should be released for a real peer-review publication.

Competing interests

The authors declare that they have no competing interests.

Consent for publication

This pre-print is not meant to be published as it is, but only serves as a chapter in a PhD thesis manuscript. It was proofread only by Eric Meyer. Mathieu Bahin did not read the present manuscript. However, I would like to recognize his significant contribution.

Authors' contributions

1 DELEVOYE Guillaume (PhD student) - Programming, data analysis, redaction
2 MEYER Eric (Research director) - Biology advisor, corrections, proofreading,
3 BAHIN Mathieu - Informatics and programming advisor, debugging

All authors shared weekly scientific meetings for about a year on the subject.

Author details
Institute of biology, Ecole Normale Supérieure - PSL, Paris, France.

References

1. Rhoads, A., Au, K.F.: PacBio sequencing and its applications. Genomics, Proteomics & Bioinformatics **13**(5), 278–289 (2015). doi:10.1016/j.gpb.2015.08.002
2. Beaulaurier, J., Zhang, X.-S., Zhu, S., Sebra, R., Rosenbluh, C., Deikus, G., Shen, N., Munera, D., Waldor, M.K., Chess, A., Blaser, M.J., Schadt, E.E., Fang, G.: Single molecule-level detection and long read-based phasing of epigenetic variations in bacterial methylomes. Nature Communications **6**(1) (2015). doi:10.1038/ncomms8438
3. Pirone-Davies, C., Hoffmann, M., Roberts, R.J., Muruvanda, T., Timme, R.E., Strain, E., Luo, Y., Payne, J., Luong, K., Song, Y., Tsai, Y.-C., Boitano, M., Clark, T.A., Korlach, J., Evans, P.S., Allard, M.W.: Genome-wide methylation patterns in salmonella enterica subsp. enterica serovars. PLOS ONE **10**(4), 0123639 (2015). doi:10.1371/journal.pone.0123639
4. Biosciences, P.: Photo Release - Pacific Biosciences Launches the PacBio(R) RS II Sequencing System - PacBio. `https://www.pacb.com/pres\_releases/photo-release-pacific-biosciences-launches-the-pacbior-rs-ii-sequencing-system/`. (Accessed on 01/17/2022)
5. English, A.C., Richards, S., Han, Y., Wang, M., Vee, V., Qu, J., Qin, X., Muzny, D.M., Reid, J.G., Worley, K.C., Gibbs, R.A.: Mind the gap: Upgrading genomes with pacific biosciences RS long-read sequencing technology. PLoS ONE **7**(11), 47768 (2012). doi:10.1371/journal.pone.0047768
6. Biosciences, P.: Perspective - Understanding Accuracy in SMRT Sequencing. `https://www.pacb.com/wp-content/uploads/2015/09/Perspective\_UnderstandingAccuracySMRTSequencing1.pdf`. (Accessed on 01/17/2022)
7. Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., Peluso, P., Rank, D., Baybayan, P., Bettman, B., Bibillo, A., Bjornson, K., Chaudhuri, B., Christians, F., Cicero, R., Clark, S., Dalal, R., deWinter, A., Dixon, J., Foquet, M., Gaertner, A., Hardenbol, P., Heiner, C., Hester, K., Holden, D., Kearns, G., Kong, X., Kuse, R., Lacroix, Y., Lin, S., Lundquist, P., Ma, C., Marks, P., Maxham, M., Murphy, D., Park, I., Pham, T., Phillips, M., Roy, J., Sebra, R., Shen, G., Sorenson, J., Tomaney, A., Travers, K., Trulson, M., Vieceli, J., Wegener, J., Wu, D., Yang, A., Zaccarin, D., Zhao, P., Zhong, F., Korlach, J., Turner, S.: Real-time DNA sequencing from single polymerase molecules. Science **323**(5910), 133–138 (2009). doi:10.1126/science.1162986
8. Chin, C.-S., Alexander, D.H., Marks, P., Klammer, A.A., Drake, J., Heiner, C., Clum, A., Copeland, A., Huddleston, J., Eichler, E.E., Turner, S.W., Korlach, J.: Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. Nature Methods **10**(6), 563–569 (2013). doi:10.1038/nmeth.2474
9. Feng, Z., Fang, G., Korlach, J., Clark, T., Luong, K., Zhang, X., Wong, W., Schadt, E.: Detecting DNA modifications from SMRT sequencing data by modeling sequence context dependence of polymerase kinetic. PLoS Computational Biology **9**(3) (2013). doi:10.1371/journal.pcbi.1002935
10. Clark, T.A., Murray, I.A., Morgan, R.D., Kislyuk, A.O., Spittle, K.E., Boitano, M., Fomenkov, A., Roberts, R.J., Korlach, J.: Characterization of DNA methyltransferase specificities using single-molecule, real-time DNA sequencing. Nucleic Acids Research **40**(4), 29–29 (2011). doi:10.1093/nar/gkr1146
11. Flusberg, B.A., Webster, D.R., Lee, J.H., Travers, K.J., Olivares, E.C., Clark, T.A., Korlach, J., Turner, S.W.: Direct detection of DNA methylation during single-molecule, real-time sequencing. Nature Methods **7**(6), 461–465 (2010). doi:10.1038/nmeth.1459
12. PacBio: White Paper - Detecting DNA Base Modifications Using SMRT Sequencing. `https://www.pacb.com/wp-content/uploads/2015/09/WP_Detecting_DNA_Base_Modifications_Using_SMRT_Sequencing.pdf`. (Accessed on 01/20/2022)
13. Ye, P., Luan, Y., Chen, K., Liu, Y., Xiao, C., Xie, Z.: MethSMRT: an integrative database for DNA n6-methyladenine and n4-methylcytosine generated by single-molecular real-time sequencing. Nucleic Acids Research **45**(D1), 85–89 (2016). doi:10.1093/nar/gkw950
14. PacBio: GitHub - PacificBiosciences/kineticsTools: Tools for detecting DNA modifications from single molecule, real-time sequencing data. `https://web.archive.org/web/20200917205747/https://github.com/PacificBiosciences/kineticsTools`. (Accessed on 01/19/2022)
15. PacBio: kineticsTools/manual.rst at master · PacificBiosciences/kineticsTools · GitHub. `https://web.archive.org/web/20170813085437/https://github.com/PacificBiosciences/kineticsTools/blob/master/doc/manual.rst`. (Accessed on 01/19/2022)
16. Fang Lab, B.e.a.: GitHub - fanglab/SMALR: SMALR: a framework for single-molecule level interrogation of the methylation status of SMRT reads. `https://github.com/fanglab/SMALR`. (Accessed on 01/17/2022)
17. PacBio: kineticsTools/manual.rst at master · PacificBiosciences/kineticsTools · GitHub. `https://web.archive.org/web/20170813085437/https://github.com/PacificBiosciences/kineticsTools/blob/master/doc/manual.rst`. (Accessed on 01/19/2022)
18. Shi, H., Li, W., Xu, X.: Learning the comparing and converting method of sequence phred quality score. In: Proceedings of the 6th International Conference on Management, Education, Information and Control (MEICI 2016). Shenyang, China. P, pp. 260–263 (2016)
19. Schadt, E.E., Banerjee, O., Fang, G., Feng, Z., Wong, W.H., Zhang, X., Kislyuk, A., Clark, T.A., Luong, K., Keren-Paz, A., *et al.*: Modeling kinetic rate variation in third generation dna sequencing data to detect putative modifications to dna bases. Genome research **23**(1), 129–141 (2013)
20. PacBio: kineticsTools/kinetics.pdf at master · PacificBiosciences/kineticsTools · GitHub. `https://web.archive.org/web/20220119141308/https://github.com/PacificBiosciences/kineticsTools/blob/master/doc/whitepaper/kinetics.pdf`. (Accessed on 01/19/2022)
21. PacBio: Introducing the Sequel System: The Scalable Platform for SMRT Sequencing - PacBio. `https://www.pacb.com/blog/introducing-the-sequel-system-the-scalable-platform-for-smrt-sequencing/`. (Accessed on 01/17/2022)
22. PacBio: Pacific Biosciences Launches New Sequel II System, Featuring ~8 Times the DNA Sequencing Data Output - PacBio. `https://www.pacb.com/press\_releases/pacific-biosciences-launches-new-sequel-ii-system-featuring-8-times-the-dna-sequencing-data-output/`. (Accessed on 01/17/2022)
23. PacBio: SMRT Analysis Software - PacBio. `https://www.pacb.com/products-and-services/analytical-software/smrt-analysis/`. (Accessed on 01/17/2022)
24. Duharcourt, S., Sperling, L.: Chapter five - the challenges of genome-wide studies in a unicellular eukaryote with two nuclear genomes. In: Carpousis, A.J. (ed.) High-Density Sequencing Applications in Microbial Molecular Genetics. Methods in Enzymology, vol. 612, pp. 101–126. Academic Press, ??? (2018). doi:10.1016/bs.mie.2018.08.012. https://www.sciencedirect.com/science/article/pii/S0076687918302908
25. Arnaiz, O., Meyer, E., Sperling, L.: ParameciumDB 2019: integrating genomic data across the genus for functional and evolutionary biology. Nucleic Acids Res. **48**(D1), 599–605 (2020)
26. PacBio: BAM format specification for PacBio — PacBioFileFormats 3.0 documentation. `https://web.archive.org/web/20200918232456/https://pacbiofileformats.readthedocs.io/en/3.0/BAM.html`. (Accessed on 01/19/2022)
27. PacBio: SMRT Analysis Software - PacBio. `https://web.archive.org/web/20220109065045/https://www.pacb.com/products-and-services/analytical-software/smrt-analysis/`. (Accessed on 01/19/2022)

28. foundation, A.: Anaconda Software Distribution. Anaconda Inc. (2020). https://docs.anaconda.com/
29. nalepae: Pandarallel - A simple and efficient tool to parallelize Pandas operations on all available CPUs. https://web.archive.org/web/20201128040916/https://github.com/nalepae/pandarallel
30. Marinus, M.G., Morris, N.R.: Isolation of deoxyribonucleic acid methylase mutants of Escherichia coli K-12. J. Bacteriol. **114**(3), 1143–1150 (1973)
31. Geier, G.E., Modrich, P.: Recognition sequence of the dam methylase of Escherichia coli K12 and mode of cleavage of Dpn I endonuclease. J. Biol. Chem. **254**(4), 1408–1413 (1979)
32. Russell, D.W., Zinder, N.D.: Hemimethylation prevents DNA replication in e. coli. Cell **50**(7), 1071–1079 (1987). doi:10.1016/0092-8674(87)90173-5
33. Urig, S., Gowher, H., Hermann, A., Beck, C., Fatemi, M., Humeny, A., Jeltsch, A.: The escherichia coli dam DNA methyltransferase modifies DNA in a highly processive reaction. Journal of Molecular Biology **319**(5), 1085–1096 (2002). doi:10.1016/s0022-2836(02)00371-6
34. Roberts, R.J., Vincze, T., Posfai, J., Macelis, D.: REBASE—a database for DNA restriction and modification: enzymes, genes and genomes. Nucleic Acids Research **43**(D1), 298–299 (2014). doi:10.1093/nar/gku1046
35. McGinnis, S., Madden, T.L.: BLAST: at the core of a powerful and diverse set of sequence analysis tools. Nucleic Acids Research **32**(Web Server), 20–25 (2004). doi:10.1093/nar/gkh435
36. Mikami, K., Koizumi, S.: Microsurgical analysis of the clonal age and the cell-cycle stage required for the onset of autogamy in paramecium tetraurelia. Developmental Biology **100**(1), 127–132 (1983). doi:10.1016/0012-1606(83)90203-8
37. Berger, J.D.: Autogamy in paramecium cell cycle stage-specific commitment to meiosis. Experimental Cell Research **166**(2), 475–485 (1986). doi:10.1016/0014-4827(86)90492-1
38. SONNEBORN, T.M.: The relation of autogamy to senescence and rejuvenescence in paramecium aurelia. The Journal of Protozoology **1**(1), 38–53 (1954). doi:10.1111/j.1550-7408.1954.tb00792.x
39. Galvani, A., Sperling, L.: Rna interference by feeding in paramecium. TRENDS in Genetics **18**(1), 11–12 (2002)
40. Beisson, J., Bétermier, M., Bré, M.-H., Cohen, J., Duharcourt, S., Duret, L., Kung, C., Malinsky, S., Meyer, E., Preer, J.R., Sperling, L.: Silencing specific paramecium tetraurelia genes by feeding double-stranded RNA. Cold Spring Harbor Protocols **2010**(1), 5363 (2010). doi:10.1101/pdb.prot5363
41. Puoliväli, T., Palva, S., Palva, J.M.: Influence of multiple hypothesis testing on reproducibility in neuroimaging research: A simulation study and python-based software. Journal of Neuroscience Methods **337**, 108654 (2020). doi:10.1016/j.jneumeth.2020.108654
42. Mashelkar, P.: Concatenating Text Files into a Single File in Linux — Baeldung on Linux. https://web.archive.org/web/20220112125340/https://www.baeldung.com/linux/concatenate-files. (Accessed on 01/20/2022)