NEW INSIGHTS INTO DNA METHYLATION MECHANICS USING HIGH THROUGHPUT COMPUTING ALGORITHMS FOR EPIGENOMICS DATA ANALYSIS

German Telmo Eizaguirre Suarez

BIOTECHNOLOGY FINAL DEGREE PROJECT

Academic tutor: Prof. María Jesús Torija Martínez

Degree in Biotechnology, Universitat Rovira I Virgili

Biochemistry and Biotechnology Department

mjesus.torija@urv.cat

In collaboration with: Computational Biology and Systems Biomedicine Group, IIS Biodonostia

Supervisor: Prof. Marcos Jesús Araúzo Bravo

Computational Biology and Systems Biomedicine Group, IIS Biodonostia,

marcos.arauzo@biodonostia.org

Defend date: July 15, 2020

I, German Telmo Eizaguirre Suarez, with DNI 49581396C, am aware of the guide to plagiarism prevention at the URV Prevenció, detecció i tractament del plagi en la docència: guia per a estudiants (approved in July 2017)

(http://www.urv.cat/ca/vidacampus/serveis/crai/que-us-oferim/formacio-competencies-nuclears/plagi/)

and I affirm that this TFG does not constitute any of the conducts considered as plagiarism by the URV.

Tarragona, 9th July 2020

Contents

Acknowledgements

I want to acknowledge Prof. Marcos J. Araúzo Bravo for his involvement in the project. I also thank the colleagues in the Computational Biology and Systems Biomedicine Group for their warm reception and support throughout my stay.

Thanks to Richard Meyers, from the HudstonAlpha Institute of Biotechnology[1] and Brad Bernstein, from the BROAD institute[2], for their methylome datasets publicly available in the ENCODE portal.

[1] **HudsonAlpha Institute for Biotechnology:** https://hudsonalpha.org/
[2] **BROAD institute:** https://bernstein.mgh.harvard.edu/broad-institute-epigenomics-program/

About the centre

This final degree project of Biotechnology is based on a research stay made by the author in the Computational Biology and Systems Biomedicine Group at the IIS Biodonostia (Donostia-San Sebastián, Guipúzcoa, Spain) from June to August of 2019.

IIS Biodonostia is one of the three main medical research institutes of the Basque public health system (Osakidetza). The institute has 26 research groups and more than 350 researchers. Their structure is divided into up to 9 platforms that give research assistance to inner groups, external collaborators, and local health organizations. We developed the current project in the Computational Biology Platform, in the Computational Biology and Systems Biomedicine Group headed by Prof. Marcos J. Araúzo-Bravo.

The Computational Biology and Systems Biomedicine Group focus on several objectives backed by research computational methods. Such objectives include the modelling of cell pluripotency and differentiation, the analysis of genetic and epigenetic networks, the development of computer vision algorithms for the study of cell structures or the analysis of clinical histories through artificial intelligence, among others. Currently, they work on several lines of research, for instance, the characterization of regulatory elements in pluripotent and stem cells or the development of computational frameworks for omics data analysis.

Our project belongs mainly to the line of research of the discovery of DNA sequence patterns for the generation of omics dictionaries. However, we developed our work in a transversal way, such that our results also relate to the discovery of regulatory elements involved in pluripotency. Also, we expect to implement a generic omics analysis software appliable to different scenarios and reusable in future projects.

Abstract

Epigenomic regulation is a complex process mediated by multiple factors. Understanding the functionality of DNA sequences involved in such process is crucial for clinical and research aims. DNA methylation is a crucial epigenetics mark responsible for gene silencing. Hence, motifs extracted from DNA methylation sites include specific DNA words relatable to many crucial biological processes, including cell reprogramming and differentiation. As methylation regulation is suggested to be a cooperative mechanism between different molecules and protein domains, methylation motifs do not necessarily need to appear centred on the methylation target. Instead, steric hindrance between different DNA binding domains should be considered when analysing DNA methylation motifs. If methylation regulator proteins are affected by steric hindrance, then DNA methylation motifs should be positioned at some bases of displacement from the methylation target. We implemented an extensive, genome-wide methylation motif discovery pipeline runnable in Slurm-based HPC clusters. We optimized the pipeline for the analysis of hundreds of displacements in one run. Our pipeline discloses motifs independently of their positioning relative to the DNA methylation target. We study the prevalence of displaced methylation motifs in cell lines at multiple differentiation levels and disclose valuable trends of motif quality at specific displacements from the methylation target. We relate our result to molecular mechanisms involved in differentiation and propose new models for the repression of genes involved in pluripotency.

Key words

DNA motifs, Methylation, Epigenomics, Bioinformatics, Differentiation

Introduction

Epigenetics and cell regulation

Cell regulation is a multi-factor process orchestrated by several types of molecules, including DNA, RNA, microRNA, or proteins. As a rule, coding regions of the DNA define the "how-to" of regulatory molecule synthesis, whereas non-coding regions, mainly gene promoters and enhancers, define its "when-to". Thus, in a clearly mechanistic point of view, we could consider the DNA as a universal Turing machine[3] codifying the instructions to building and keeping up and running itself, self-repairing, and self-dismantling. Far from the naïve and classical Mendelian perspective, cells are not only regulated by basic genetic traits. Instead, epigenetic marks play a crucial role in cell complexity and functionality.

The term epigenetics describes the study of heritable phenotypic features not directly related to changes in the DNA sequence. We call epigenome the collection of epigenetic modifications in the whole genome of a single organism. Epigenetic modifications comprise covalent modifications of histones and DNA [1] and are regulated by several factors that include three-dimensional genomic structure, nucleosome positioning and the influence of different proteins and signal molecules [2]. The crosstalk between such factors provides a highly flexible epigenomics profile for each organism and gives answer to many complex phenotypes that cannot be completely modelled by basic genetic traits. As a result, epigenetic biomarkers are gaining great interest for the clinical study of disease variations [1], specially focusing on DNA methylation and histone modification analysis.

Epigenetic studies are backed by academically successful projects such as the Encyclopaedia of DNA Elements (ENCODE) [3], an international consortium with a broad number of publicly available epigenomics datasets from various cell types. ENCODE offers a web server and a user-friendly portal[4] of great value for epigenomics studies, working as an encyclopaedia in the sense that it serves as a reference database for epigenetic knowledge and data. In this work, we hypothesize with the possibility of creating a dictionary of *DNA words* ruling epigenetic regulations. Such a dictionary would help, first, to navigate through such a vast database, and second, it

[3] **Turing machine:** abstract computational model and formal language developed by Alan Turing in 1936 for the interpretation of an infinite sequential input, a "tape", of symbols. It consists of a finite set of symbols and a finite number of states. It starts in the initial state and each transition from one symbol to another in the input "tape" implies a transition rule from one state of another. Thus, the global state of the system changes according to the sequence of the input symbols.

[4] https://www.encodeproject.org

would contribute to a mechanistic knowledge of epigenetics by providing relationships between basic DNA words and their possible regulatory meaning. A deeper insight into epigenetic regulatory units could as a result ease the use of current molecular tools such as CRISPR/Cas9 [2], for instance, in gene therapy and cell reprogramming experiments.

Previous research on identification and quantification of the activity of regulatory DNA patterns include transcription factor (TF) binding site (TFBS) classification [4], DNA methylation motif identification [2,5–7] and super enhancer search improvement [8], among others. DNA motifs are sequence patterns with significantly higher appearance frequency in the genome than random sequences of the same length. Frequently, DNA motifs are related to biological functionalities and show some degree of evolutive conservation.

Despite earlier work toward a full comprehension of the DNA language, such an objective is far from completed. Starting from the premise of interpreting epigenetic regulation as a language of DNA, we would associate DNA words with the semantics of the language, and the interaction rules between such words as the syntax. Based on this model, we find studies disclosing both semantic and syntactic components of epigenetics. From the semantic part, most of the results are focused on the retrieval of TF binding motifs (TFBMs) [4–6], motifs related to histone marks [2] and DNA methylation motifs [6,7]. From the syntactic part we find, for example, research on uncovering super enhancers [8], insulators [9] and combinatory patterns of TFBMs [10]. In this work, we target DNA methylation motifs as a continuation of the original work by Luu *et al.* [7].

Biological relevance of DNA methylation

DNA methylation is a crucial epigenomics modification. In eukaryotic cells, a methyl group is covalently bound to the carbon-5 position of a cytosine and yields a 5-methylcitosine residue [11]. Methylation values and patterns across the genome are highly variable between species and cell types. In animals, DNA methylation takes place mainly in CpG dinucleotides and less frequently in non-CpG DNA targets [12]. The number of CpG dinucleotides in an organism is massive, i.e. the human genome has 28 million CpG sites. However, to remain labelled as epigenetic mark, the methylation state of a CpG target must be stable through cell divisions [13]. Currently many genome-wide methylation mapping techniques are available for methylation state analysis, mostly based on whole genome bisulfite sequencing (WGBS). Yet, raw and annotated WGBS data from several species and cell lines can be downloaded from

various web servers, like the Roadmap Epigenomics Project [14] or the previously mentioned ENCODE portal [3].

Methylation of genomics sequences regulates gene translation. Genes with strongly methylated bodies tend to be highly expressed [4] and anomalous DNA methylation patterns have been established as indicators for many diseases [1,11,12]. Also, methylation level of active gene regulatory elements, such as promoters, is inversely correlated with gene expression [4]. The influence of DNA methylation on gene expression partly comes from the methylation-dependent interaction between TFs and DNA. TFs are proteins that recognise specific DNA sequences and regulate the expression level of genes associated with such sequences [7]. The binding of TFs with DNA can be dependent on DNA methylation levels, positively or negatively depending on the TF family [4]. Similarly, gene expression can be actively or passively regulated by DNA methylation. In active regulation, methylated DNA avoids the binding of TFs to their target sequence by steric hindrance. In passive regulation, instead, proteins related to gene transcription regulation detect methylated residues in their target sequence [4].

Besides the basic and fundamental relationship with gene expression, DNA methylation is essential in many higher-level biological processes. For instance, it guarantees the correct inactivation of the X chromosome during development. Also, it is the key regulator of genomic imprinting, that is, a non-Mendelian inheritance process in which one of the two inherited alleles for a concrete gene is silenced by DNA methylation [12]. Aberrant DNA methylation has been proven to work as an indicator of cancerous diseases, and specific CpG methylation profiles are currently being used as biomarkers for several cancers in preclinical phases [1]. Gains in methylation levels, for instance, can lead to tumour suppressor gene repression. However, due to the inherent plasticity of the epigenome and the fact that methylation patterns can vary during tumour progressions, it is not trivial to define which methylation traits are directly coupled with each case. Hence, a better knowledge of the DNA methylation mechanics could favour current preclinical diagnosis methods.

Embryonic stem cells (ESC) are undifferentiated, pluripotent cells isolated from the inner cell mass of the early mammalian embryo. The term pluripotent describes those cells with the capacity to develop into any of the three germ layers of the embryo. Consequently, a pluripotent cell can develop into any differentiated cell type except for non-embryonic cell tissues (e.g. the placenta), which makes ESCs of great interest for medicament evaluation and regenerative medicine. ESCs are restricted for experimentation though, both for technical and ethical reasons. Consequently, cellular

reprogramming of somatic cells into induced pluripotent stem cells (iPSC) has gained attention instead. We find previous evidence about reprogramming being modulated by DNA methylation. TFs that are differentially inhibited by methylated CpG (mCpG) have been ontologically associated with cell differentiation, whereas TFs positively affected by mCpG are related to embryonic and organismal developmental processes [4]. Previous studies have found differentially methylated regions in specific tissues too [12]. Methylation-specific DNA binding domains (MBD) have also been linked with cell pluripotency and reprogramming. MBD2a and MBD2t, for example, regulate the expression of *OCT4* and *NANOG*, both crucial genes for pluripotency. Knocking down MBD3 in human fibroblasts also improved their reprogramming efficiency to iPSC [4]. Therefore, reprograming is mediated by both genetic and epigenetic factors, and although its procedure is extensively used today, we still do not have a global understanding of its mechanism. Our current knowledge about reprogramming would benefit from a broader insight into the DNA methylation process.

mCpG recognition and sequence context

The classical model of the interaction between proteins and methylated DNA does not support sequence-specific recognition of methylation patterns [12]. In plain words, proteins that directly interact with DNA methylation could be classified into “writers”, “editors” and “readers” [11]. ‘Writer” proteins settle and preserve methylation patterns across the genome during cell development and differentiation. DNA methyltransferases (DNMT), the enzymes that catalyse the previously explained DNA methylation reaction, belong to this family. “Editor” proteins change the state of methylated DNA. For instance, ten-eleven translocation (TET) enzymes pertain to DNA methylation “editors”. TET proteins oxidize the 5-methylcytosine into 5-hydroximethylcytosine, which yields to DNA demethylation. Finally, “reader” proteins bind mCpG and regulate the interaction between DNA methylation, histone modifications and chromatin topology. “Reader” proteins include methyl-CpG-binding domain (MBD) proteins, among others. This fundamental view of the methylated DNA-protein interaction holds that the binding of MBD proteins to methylated DNA is mostly non-sequence-specific, whereas TFs can sequence-specifically bind to non-methylated DNA at TFBSs [12]. Thus, DNA methylation motifs are not compatible with the classical interaction pattern, as the only way of TFs and other regulator proteins for interacting with mCpG is through MBDs. Recent evidence, however, suggests that such interaction is much more complex and could be driven in a sequence-specific manner.

The main proof of sequence-specific binding of proteins to methylated DNA are MBDs lacking mammalian TFs that can bind with mCpG [12]. Additionally, some proteins can

bind non-methylated or methylated sequences depending on the target-surrounding sequence. CEBPα, for instance, binds the sequence 5'-TGACGTCA, but it can bind 5'-mCGTGA too when DNA is methylated [12]. Other proteins, such as KLF4, have strictly different and not intersecting binding sequences if DNA is methylated or unmethylated [12].

Preceding research on DNA methylation motifs has also demonstrated sensitivity between methylation of CpGs and their sequence context. Luu *et al.* [7] proposed that if DNA methylation/unmethylation mechanisms could discriminate between CpG with different sequence context, CpG *loci* with disparate sequences in both strands would have higher variability in methylation distribution than CpG *loci* with similar sequence in both strands, that is, palindromic sequences. Luu *et al.* statistically validated such hypothesis in human fibroblasts, ESCs and iPSC and developed an algorithm for *de novo* discovery of DNA methylation words centred on CpGs. Not only such sequence specificity has been proven for DNA methylation motifs, but also for other epigenetic targets such as histone marks [2]. Yin *et al.* [4] demonstrated that TFs could pertain to one from up to five families depending on their type of interaction of methylated DNA, and evidenced sequence-specificity in the binding of some TFs to their methylated TFBS. They identified many TFBS motifs and classified TFs into "no CpG" (TFs with no CpG in their TFBSs), "little effect" (TFs with CpG in their TFBS but not affected by DNA methylation), "methyl-minus" (TFs with CpG in their TFBS that bind more weakly to their methylated TFBS) and "methyl-plus" (TFs with CpG in their TFBS that show higher affinity for their methylated TFBS). Different studies also obtained similar results [6,12]. Thus, sequence-specific recognition of mCpG would be feasible, and consequently, DNA methylation motif determination would be justified.

Different models have been proposed for the sequence-specific interaction of proteins with mCpG. Recent studies suggest that some proteins could directly bind to mCpG with specific binding domains and modulate methylation (Fig. 1A) [12], which would be coherent with TF families described by Yin *et al*. [4]. In fact, most MBD protein families, in which many TFs are included, present various sequence-specific DNA-binding domains such as transcription repressor domains (TRD), or DDT domains (responsible for transcription and chromosome remodelling) [11].

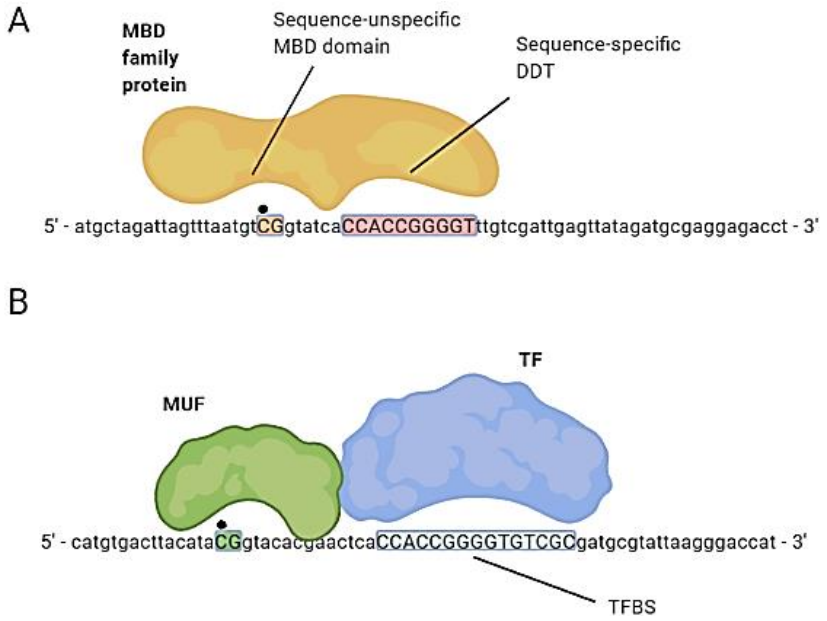

Figure 1 Examples for two hypothetic DNA methylation modulation models. We represent DNA methylation with a black filled circle. (A) A unique protein binds the DNA with a sequence-specific DNA binding domain (i.e., a DDT domain) and a sequence-unspecific domain (MBD) interacts with the CpG. The protein in (A) is based on the model for the MBD family protein TIP5/BAZ2A described by Du *et al.* [11]. (B) A first protein with a sequence-specific domain binds the DNA (a TF, in its TFBS) and recruits a second protein that interacts with a CpG and modulates its methylation state (MUF). In neither of the models the protein domain sequence-specifically interacting with DNA and the protein domain interacting with the CpG can bind in the same DNA *locus*, and thus there should be a gap between the targeted CpG and the sequence-specific binding site. Created with BioRender.com [15].

Luu *et al.* [7] proposed a cooperative model where proteins that sequence-specifically recognize CpGs interact with methylation “writers” and “editors” and recruit them to the target CpG (Fig. 1B). We could since label methylation “writers” and “editors”, which may interject with TFs, as methylation/unmethylation factors (MUFs): Some results sustain the coordinated model for the modulation of DNA methylation. There exists evidence that DNMTs can assemble with chromatin remodeler enzymes and TFs [12]. The nuclear receptor peroxisome proliferator-activated receptor-γ (PPARγ) can recruit TET1 too, unmethylating its surrounding mCpGs [12]. Even TFs have been suggested to work in clusters of multiple TFs, co-factors, and different protein complexes [5,10]. Cooperative regulation models are not rare, in fact, gene transcription is also cooperative, with DNA sequence-specific TFs recruiting transcription machinery to the gene promoter. Both models seem feasible and could coexist in the cell context.

Displaced DNA methylation motifs

The work by Luu *et al.* [7] that we take as reference only analysed CpG-centred methylation motifs. In our contemplated models, that would not necessarily be the only DNA methylation motif placement with respect to the CpG. In addition, steric hindrance must be considered, as MUFs and sequence-specific DNA readers cannot physically collocate on the same DNA target with different DNA binding domains. Thus, we would

expect some degree of displacement from the CpG *loci* (sequence-specifically or non-sequence-specifically interacting with MUFs) to the DNA methylation motif (sequence specifically interacting with proteins such as TFs) (see Fig. 1).

Here, we present a generalization and a readjustment of the pioneer *de novo* motif discovery algorithm by Luu *et al.* [7] capable of identifying DNA methylation patterns irrespective of the CpG target position from a FASTA[5] sequence of the genome and the annotated methylation results as input.

Our proposal differs from previous motif discovery methods. The Autoseed pipeline [4,5] is also a *de novo* motif discovery algorithm whose input sequences must be previously filtered. It is used, for instance, after a Chip-Seq assay, thus ignoring sequences not binding to the specific proteins being analysed. Our method allows a complete analysis of the methylation-related motifs in a genome, independently of the protein domains they bind. Epigram and Homer [2] have also been used exclusively with a previous Chip-Seq step. On the contrary, the Multiple Expression motifs for Motif Elicitation (MEME) [12] needs an input training dataset before motif recognition, and cannot be classified as a *de novo* motif discovery algorithm.

We propose a new perspective of omics experiment design and result analysis. Classical omics studies need to analyse at least two conditions, generally a wild type and a genetically or environmentally perturbed sample and evaluate Differentially Signalled Regions (DSR) between both. Inversely, we only need an input condition, as our algorithm searches similar signals in different positions of the input data without control samples. Classical methods compare *locus* in at least two different samples, whereas our method intrinsically uses the DSR states of different *loci* of a unique sample as control.

DSR methods are complementary and synergic to our algorithm, as they target different objectives. DSR methods are an appropriate perspective for the analysis of epigenetics signals in different experimental conditions. However, they are usually only capable of working on a unique region length and do not provide single nucleotide level information. We instead run an algorithm that self-adapts to region lengths and whose results are adjusted to the degree of similarity between different genomic regions, both advantageous features for the analysis of genome-intrinsic, variable length DNA words.

We have generalized the mathematical description of the method by Luu *et al*. [7] to detect motifs at variable displacements from the CpG target. Due to limitations related to computation resources, data analytic libraries and the algorithm itself, it was

[5] FASTA: text file format for the representation of nucleotide or amino acidic sequences.

unfeasible for the original algorithm to compute hundreds of displacements at one run, an essential capability for our study. Hence, we have delineated parallelization techniques, optimized the source code with low-level routines and adapted the program for running on High Performance Computing (HPC) clusters (see *Annex B: Computational optimizations of the original algorithm*).

Hypothesis

Previous work has demonstrated that DNA methylation is sensitive to the sequence around CpG targets. Recent results suggest that methylation is a complex, coordinated mechanism between various proteins, were factors as steric hindrance between molecules take relevance, and thus DNA methylation motifs could emerge displaced from the CpG target. However, only CpG-centred motifs have been studied. Here**, we hypothesize that the optimal distance to modulate the DNA methylation of a CpG target is different from zero**.

Objectives

The overall objectives of the current work are the following:

1. Analyse the quality of DNA methylation motifs through variable displacements from CpG methylation targets. Determine the optimal displacement from the target for DNA methylation motifs.
2. Contribute to the current molecular model for the modulation of DNA methylation.
3. Study methylation motif discovery dynamics with cells at many differentiation stages and relate our results to common knowledge about cell reprogramming.
4. Provide a generic *de novo* motif discovery software primarily targeting DNA methylation but extensible to any epigenomic signal.

The long-term objectives and linked to the continuity of the project are the subsequent:

5. Obtain epigenetics DNA words from variably differentiated cells pertaining to each of the three germ layers, for different epigenomic signals. Relate DNA words to their functional meaning and uncover the crosstalk between them.
6. Serve as the landmark for a future public resource (implemented as a DNA word dictionary) to design DNA targets for CRISPR/Cas9 editing. Ease the design of new cellular reprogramming and therapeutic methods.

Materials and methods

Methylation annotation data retrieval

We analysed three methylation datasets from three different cell lines: foreskin fibroblast (FF)-derived iPSCs, fibroblasts (FB) and HUES64 line ESCs[6]. For the correct interpretation of raw WGBS data, it must be correctly processed and converted into a standardized annotation format, mainly Browser Extensible Data (BED)[7]. The WGBS data processing pipeline generally consists of an indexing of the reference genome, an alignment of the indexed genome with the bisulphite reads and trimming of the results, and annotation, signal generation, quantification, quality assessment and final formatted file generation [3,16]. We used two different sources to obtain annotated methylomes. For iPSCs we downloaded the raw DNA methylomes [17] (fastq files) from the Sequence Read Archive (SRA) database[8] and processed them with an automatic BSseq data analysis software, P3BSseq [16]. For FB and ESC datasets we directly downloaded annotated datasets from the ENCODE portal [3]. The ENCODE portal provides annotated WGBS data processed with the ENCODE WGBS analysis pipeline, which is based on Bismark [18,19]. For further details about the analysed datasets, see *Annex A: Detailed dataset features*.

Software implementation details

Our program is structured as a R library (compatible with R 2.7.1 or above). For lower level source code optimizations, we used C. We have implemented an encapsulating bash script to adapt the software to a Slurm[9]-based distributed HPC cluster execution (see *Annex B: Computational optimizations of the original algorithm*).

CpG methylation motif detection algorithm

We depict the overall pipeline for motif discovery in Fig. 2. The algorithm receives the genomic sequence and methylome annotation files as input and executes the same pipeline on both DNA strands independently. However, to simplify the explanation we only describe the procedure for the positive DNA strand.

[6] **HUES64**: human embryonic stem cells derived from human blastocyst.
[7] **BED**: text file format based on spaced columns for the representation of genomic regions. It includes the coordinates and domain-specific annotations of each region.
[8] https://www.ncbi.nlm.nih.gov/sra
[9] **Slurm:** Workload manager framework for distributed computing (https://slurm.schedmd.com/).

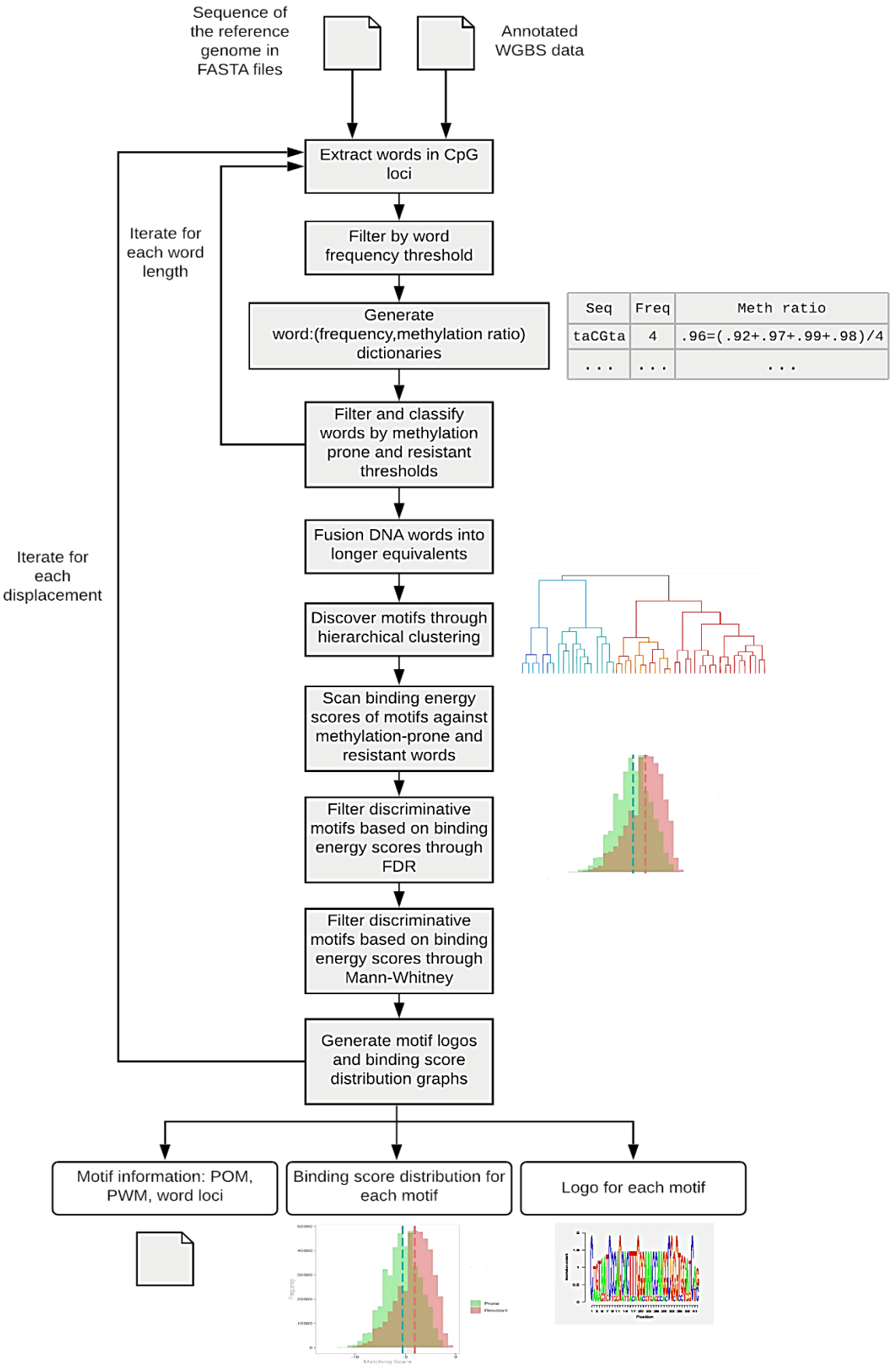

Figure 2. Pipeline of the target position-irrespective CpG methylation motif discovery algorithm.

Generation of CpG word dictionaries

Luu *et al.* [7] proposed the analysis of $w_{\min} = 12$ to $w_{max} = 44$ word lengths based on probabilistic reasoning. The minimum width of 12 was properly justified and demonstrated to encompass most of the possible sequences in the genome. Studying previous findings on DNA motifs though [2,4,6,7], we see that real motifs hardly reach 44 bp lengths. Thus, we limit the range of lengths and execute the analysis on DNA words from $w_{\min} = 12$ to $w_{max} = 32$.

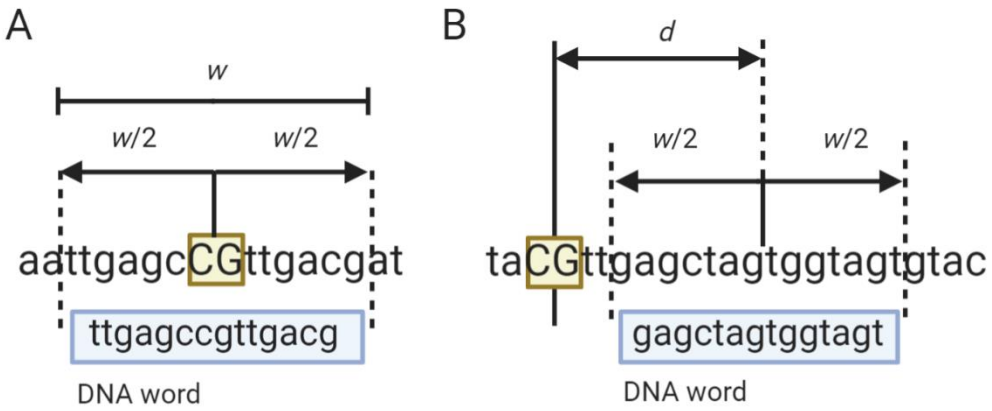

Figure 3 Extraction method for a *w*-length word from a CpG target if (A) displacement is 0, that is, the CpG target is centred in the word (same procedure as Luu *et al.* [7]) (B) displacement is not 0 (we only illustrate a positive displacement in the example, the method is symmetric for a negative displacement though). Created with BioRender.com [15].

We extract the word corresponding to each CpG in the methylome and assign the CpG's methylation rate to the word. We manage the current word displacement according to the method depicted in Fig. 3, and we iterate the process for lengths from $w_{\min}$ to $w_{\max}$. We group equal sequences and select only those with a frequency of at least 4 in the collected word collection. When grouping equal sequences, we assign them the average of methylation rates of the CpGs of each word, as we assume CpGs associated to similar sequences have similar methylation rates (see Fig. 2). Both the minimum word frequency and the similar methylation rate between similar sequences assumptions were previously reasoned by Luu *et al.* [7].

Methylome data provides methylation rates data for the forward and reverse strands of the genome. As both strands are not necessarily equally methylated, we execute a separate analysis on each strand. For DNA word retrieval in the reverse strand, the negative value of each displacement must be applied. In Fig. 4 we represent the extraction of a DNA word at a certain displacement from the forward and reverse strands.

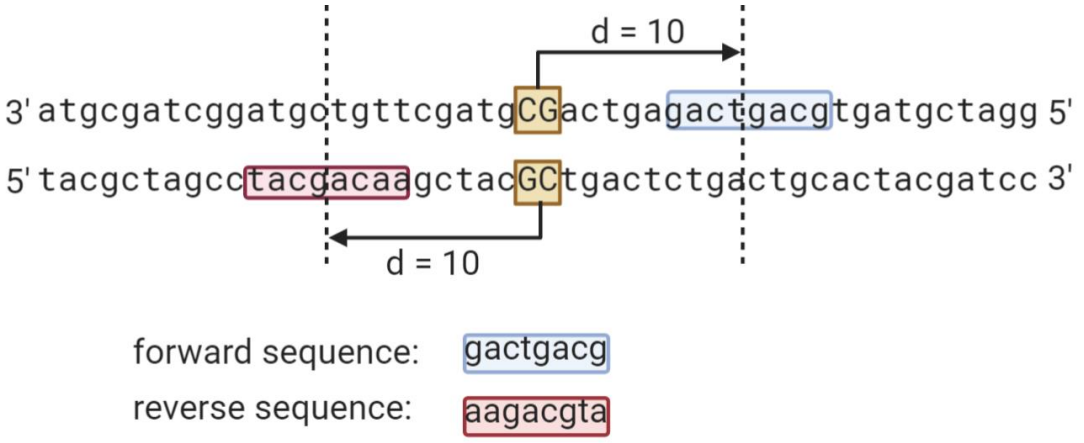

Figure 4 Example of DNA word retrieval from the forward and reverse DNA strands. In the picture, $d = 10$ and $w = 8$. The extraction of DNA words from the forward strand is straightforward. As the reverse strand runs in the opposite sense of the forward strand, the reverse DNA word is the reverse complementary of the forward DNA word at the negative value of the analysed displacement.

Methylation-prone and methylation-resistant CpG word classification

We filter and classify all the processed sequences into methylation-prone (their methylation ratio is greater or equal to Thr_{prone}) or methylation-resistant (their methylation ratio is greater or equal to $Thr_{resistant}$), and discard the remaining sequences. Luu *et al*. [7] established $Thr_{prone} = 0.85$ and $Thr_{resistant} = 0.5$. We adapted such values if necessary, for each dataset, based on empirical tests and histograms of methylation rates for each methylome. By now, we do such adjustment manually as we explain below. We expect a future increment in the algorithm to automatically perform threshold adaptation. Fig. 5 illustrates the methylation rate histograms for each sample. Generally, differentiated cells show higher methylation rates than their undifferentiated counterparts [4]. However, in Fig. 5 iPSC and ESC show a significantly higher number of CpG sites with methylation rate above 0.85 than FB. Hence, we assume such methylation trends come from the intrinsic functioning of the annotation pipeline, as methylomes were processed with two different algorithms.

As we depict in Fig. 5, methylation rates in FB (unipotent cells) are more uniformly distributed than in ESC and iPSC (pluripotent cells). We tested the method with $Thr_{prone} = 0.85$ and $Thr_{resistant} = 0.5$ for the three datasets, and ESC and iPSCs gave no prone motifs. This is probably due to the strong bimodality and the high difference between the number of methylation-prone and methylation-resistant CpGs. Having a higher number of words classified as prone means that heterogeneity between prone words increases and the conservation rate of prone words declines. Thus, the average binding score of prone motifs with the set of prone words is likely to decrease, as the probability to find words that do not match their Position Occurrence Matrix (POM) augments.

To solve the inexistence of prone motifs in iPSC and ESC we had reduced the number of prone words adjusting Thr_{prone} to 0.95. In FB we used $Thr_{prone} = 0.85$. In all datasets we set $Thr_{resistant}$ to 0.5.

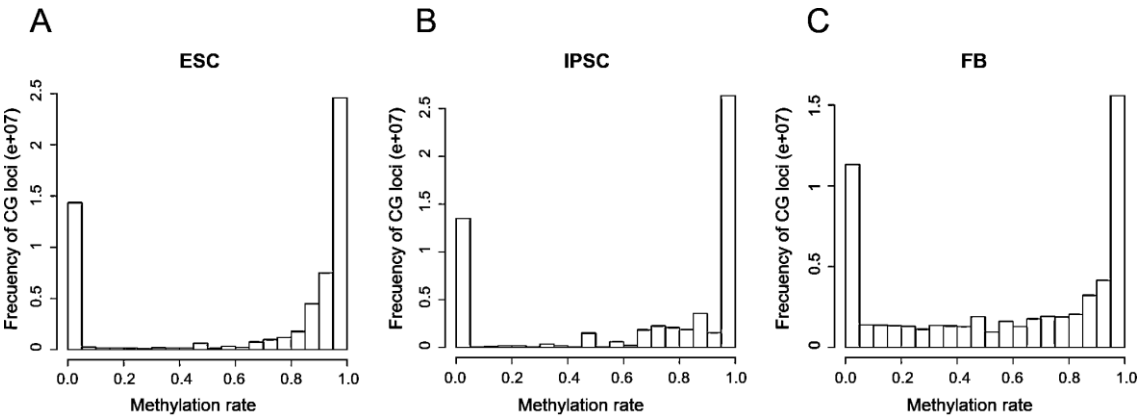

Figure 5 Histograms of methylation rates for (A) ESC methylome (B) iPSC methylome (C) FB methylome. ESC and iPSC histograms are highly bimodal and show a difference of more than a million loci between CpGs with methylation rate >= 0.85 and CpGs with methylation rate <= 0.5 (methylation thresholds stablished by Luu *et al.*[7]). The FB histogram has methylation rates more distributed and the difference of frequencies between methylation-prone and methylation-resistant CpGs is significantly lower.

Fusion of CpG sub words with longer CpG words

DNA words of consecutive lengths extend in both directions, so the centre of the word is conserved. Sequences of length w are thus likely to appear in sequences of length $w + 2$. We fusion short sub sequences into longer ones iteratively if they are perfectly contained, and we remove the fused sub sequences from the data structure. To conserve individual frequencies and methylation rates, we vectorize all sequences before the fusion step and frequency and methylation rates are assigned at nucleotide-level rather than sequence-level. When the fusion of a subsequence of length w into a sequence of length $w + 2$ is performed, methylation rates and frequencies of each of the central nucleotides of the container sequence are updated as we show in Eq. (1) and (2), respectively. The new methylation rates and frequencies are a function of the methylation rate $MR_{s^w}(i')$ and the frequency $F_{s^w}(i')$ of each nucleotide at the subsequence and the methylation rate $MR_{s^{w+2}}(i)$ and the frequency $F_{s^{w+2}}(i)$ of the corresponding nucleotide in the container sequence. The methylation rate of each nucleotide of the fused sequence $MR_{s^{w+2}}^{updated}(i)$ is calculated as the weight averaged methylation ratio of the original nucleotides. The frequency of each nucleotide in the fused sequence $F_{s^{w+2}}^{updated}(i)$ is the sum of the frequencies of the corresponding nucleotides in the subsequence and the container sequence.

$$MR_{s^{w+2}}^{updated}(i) = \frac{F_{s^w}(i') \times MR_{s^w}(i') + F_{s^{w+2}}(i) \times MR_{s^{w+2}}(i)}{F_{s^w}(i') + F_{s^{w+2}}(i)} \qquad (1)$$

$$F_{s^{w+2}}^{updated}(i) = F_{s^w}(i') + F_{s^{w+2}}(i) \quad (2)$$

Clustering for motif definition

For sequence clustering[10] and motif identification we transform each vectorized word into a Position Occurrence Matrix (POM)[11] of 4 rows and *w* columns. Initially, we assign at each column the frequency of its corresponding position of the sequence only into the row of the nucleotide at that position of the sequence. The rest of the positions are filled with zeros and the POM is flattened into a unidimensional array.

We perform a hierarchical clustering[12] with the function `hclust` from the R library `fastcluster` with complete linkage[13] mode. We use the cosine metric[14] for the dissimilarity matrix[15] calculation. Unlike Luu *et al.* [7], we established a dynamic cut-off[16] parameter based on the clustering results, choosing the cut-off value with the best average silhouette[17] for all the analysed objects. After the clustering, we rearrange the output vectors into bidimensional POMs.

Measurement of binding scores of motifs against CpG words

POMs obtained from the clustering phase must be filtered to select those that significantly resemble one distribution of sequences (prone or resistant) and differ from the other. We replicate the method of TF-DNA binding energy used by Luu *et al.* [7], based on the Berg-von Hippel method [20]. We consider a motif to match a sequence if its matching score with that sequence is high. First, we normalize all the POM values into Position Weight Matrices (PWM), see Eq. (3). Each position in the PWM $PWM(i,j)$

[10] **Clustering:** Arranging a collection of elements into groups (*clusters*) based on a similarity function.
[11] **POM:** Bidimensional matrix that, for a specific pattern, describes the probability of having each nucleotide (rows) at each position (columns).
[12] **Hierarchical clustering:** Specific method of clustering that generates a hierarchy of groups or *clusters*. The results of a hierarchical clustering can be represented in a dendrogram (see Fig. 2).
[13] **Complete linkage:** Hierarchical clustering method that starts by assigning an individual group to each of the elements into the collection to analyse and sequentially joins subgroups into broader groups.
[14] **Cosine metric:** Similarity measurement between non-zero two vectors, based on calculating the cosine of the angle between both vectors.
[15] **Dissimilarity matrix**: Being *N* the number of elements in a clustering process, the dissimilarity matrix is bidimensional array representing the similarity of each of elements with every other element in the collection.
[16] **Cut-off parameter**: Level of a dendrogram at which it must be cut, thus returning as a result the *clusters* defined at that point.
[17] **Silhouette (clustering):** Cluster quality measurement and graphical representation that shows the closeness of an object with its assigned *cluster* (cohesion) and the remoteness with the rest of *clusters* (separation).

is the quotient between the same position in the POM $POM(i,j)$ and the sum of the values of the four rows at the given column $\sum_k^4 POM(k,j)$.

$$PWM(i,j) = \frac{POM(i,j)}{\sum_k^4 POM(k,j)} \quad (3)$$

Second, we calculate the matching score MS of each PWM against every sequence Seq_w from the methylation prone and methylation sets obtained after the first methylation prone and resistant classification (see *Methylation-prone and methylation-resistant CpG word classification*), following Eq. (4). β is summed to the divisor to avoid a possible division by 0 and to the dividend to maximize the dynamic range of the resulting score, as justified by Luu *et al*. [7].

$$MS(PWM_w, Seq_w) = \sum_i^w \ln\left(\frac{PWM_w(Seq_w(i), i) + \beta}{\max\left(PWM_w(:, i)\right) + \beta}\right) \quad (4)$$

Filtering of discriminative motifs

Finally, we filter those motifs that effectively discriminate between prone and resistant sequences. Methylation prone motifs should statistically show higher matching scores over methylation prone sequences, whereas methylation resistant motifs should do the opposite.

We first apply a False Discovery Rate (FDR) test, where we measure the number of false positives from the matching score results Eq. (5). For instance, for a methylation-prone motif a false positive would be a methylation-resistant sequence with matching score above a certain threshold. We calculate the threshold as shown in Eq. (6). μ is the mean of the "equivalent" matching score distribution (prone sequences for prone motifs, resistant sequences for resistant motifs) and σ is the standard deviation of the "equivalent" matching score distribution. λ is an adjustment parameter set to 2 as previously done by Luu *et al*. [7]. We only select those motifs with FDR above or equal to 0.05 and we discard the rest.

$$FDR = \frac{FalsePositives}{FalsePositives + TruePositives} \quad (5)$$

$$Threshold = \mu + \sigma \times \lambda \quad (6)$$

Once applied the FDR step, we filter the motifs with a Mann-Whitney test. Mann-Whitney is a statistical test to validate the null hypothesis that groups are equally

distributed. For each motif, we compare the matching scores with the “equivalent” distribution against the matching scores against the other distribution. Originally, Luu *et al.* used a Kolmogorov-Smirnoff test accompanying the FDR. We use Mann-Whitney because the Kolmogorov-Smirnoff test works better on continuous values, whereas Mann-Whitney is used for discrete ones. Although matching scores are mathematically floating-point values, they are virtually discrete, as after the scanning phase we histogram them into a fixed number of breaks. For the Mann-Whitney test we use the R function `wilcox.test` from the stats package. We filter those motifs with *p*-value above significance threshold 0.00001.

Final motif retrieval

At the end of the pipeline, we get motifs for each considered length and each strand, classified into methylation-prone and methylation-resistant, for each analysed displacement. We depict such complexity in Fig. 6.

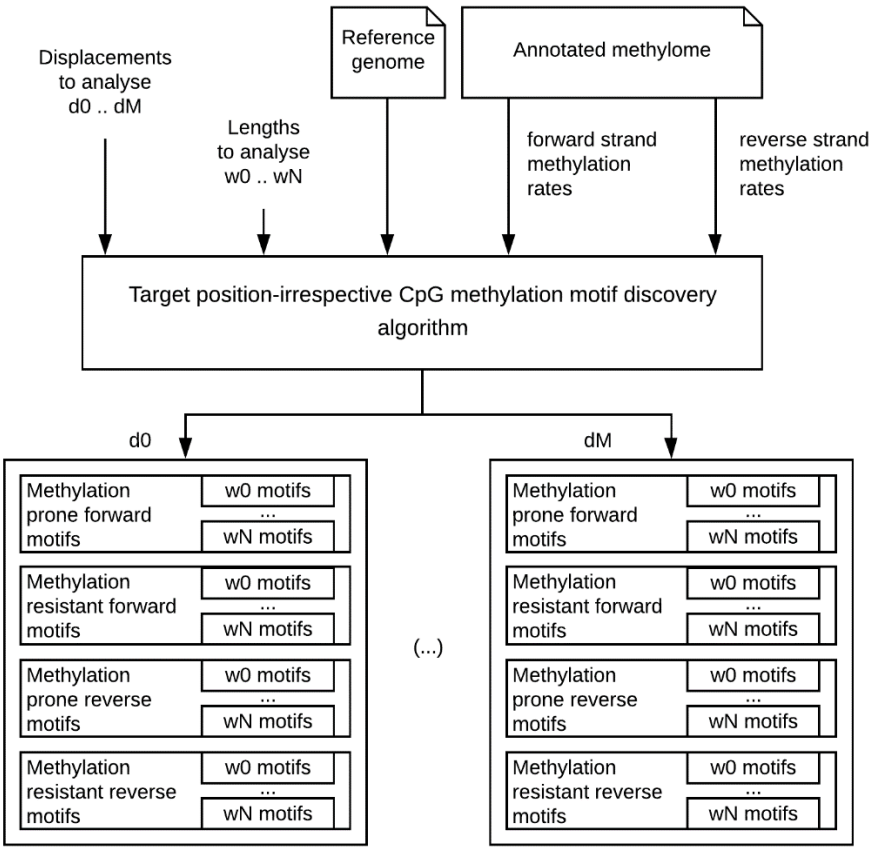

Figure 6 Input arguments and every type of motif returned by our algorithm. For each motif, it also returns the average binding score for equivalent and non-equivalent sequences and its FDR value.

Results

We studied displacements from $d_{min} = -50$ to $d_{max} = 50$ for ESC, iPSC and FB. The three studied cell lines pertain to different potency states. The analysed ESC and iPSC are pluripotent whereas FB is unipotent. iPSC are FB derived. Thus, we could infer a methylation motif history across the three samples. For each dataset, we extracted three quality measurements for the obtained motifs at each displacement, (1) the distribution of FDR values of all the motifs (as explained in *Filtering of discriminative motifs* a lower FDR value relates to a better motif), (2) the average matching score difference between the two distributions (an effectively discriminative motif should get a high difference between its "equivalent" sequence distribution and the other sequence distribution) and (3) the total number of obtained motifs.

Quality of the motifs is dependent on the displacement value

FDR distributions across the analysed displacements show variable behaviour across the three samples (see Fig. 7). Although tendencies are similar, the overall distribution of motif quality differs between samples. FDR values in ESC prone motifs and iPSC are more distributed across their possible spectrum of values than those in ESC resistant motifs and FB. Such differences probably come from the dissimilar methylation histograms described in *Methylation-prone and methylation-resistant CpG word classification*.

FDR values tend to decrease when the distance to the 0 displacement increases. In iPSC prone motif results we perceive a subtle oscillation at both sides of $d = 0$. Also, FDR values seem to improve symmetrically just before $d = 25$ and $d = -25$. Theoretically, such a result would favour our hypothesis that methylation motifs appear at some distance from the methylation target. In most cases, FDR distributions are symmetrical to both sides of $d = 0$, although we see some exceptions. In FB prone motif results, for instance, we see a minor change in FDR dynamics at about 18 nucleotides of positive displacement. *A priori*, motifs at that point would be less discriminative than the rest, as FDR values slightly increase compared to their surrounding results. However, FDR values are only a statistical indicative of motif quality not directly related to biological consequences. Therefore, such variations in the FDR distribution could reflect an implicit feature of motifs at that position that remains to be studied from other perspectives.

Our results show motif quality local maximums at $d \neq 0$. Obtaining discriminative motifs distant from $d = 0$ reaffirms the possibility of a cooperative or multiple-domain model for DNA modulation.

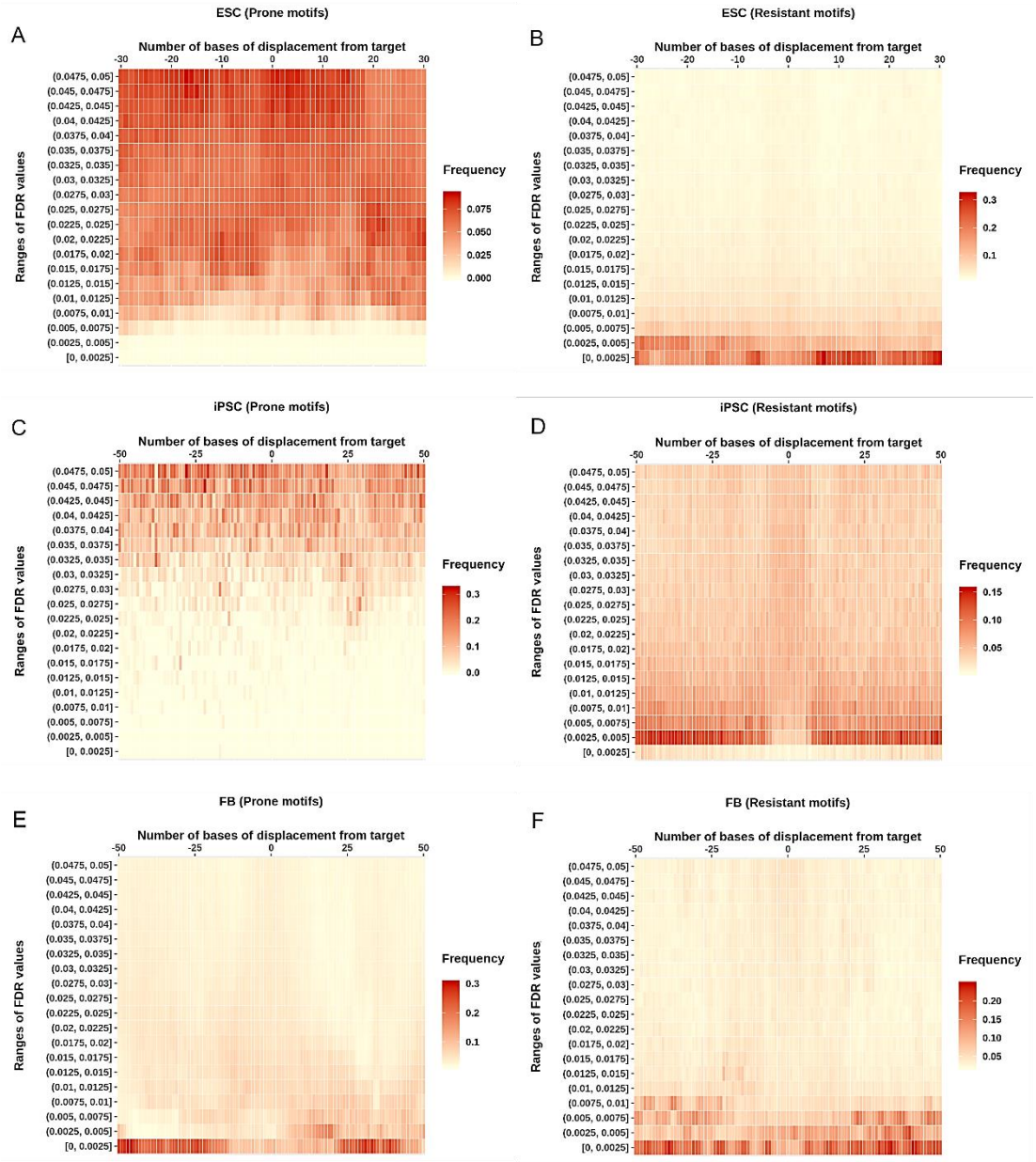

Figure 7 Heatmaps of the distribution of motif FDR values for each of the three samples analysed. We show resistant motif and prone motif results separately. (A) Prone motifs in ESC. (B) Resistant motifs in ESC. (C) Prone motifs in iPSC. (D) Resistant motifs in iPSC. (E) Prone motifs in FB. (F) Resistant motifs in FB. We could not perform the full displacement range analysis in ESC for time issues.

Discriminative motifs emerge at the centred displacement

In Fig. 8 we depict both the number of motifs obtained from the pipeline and the difference of matching scores between prone and resistant word distributions for our 3 samples. In iPSC and ESC resistant motifs and in FB, the number of filtered motifs

increases around $d = 0$. This tendency is especially abrupt in iPSC resistant motifs. Although the increase of the number of motifs at $d = 0$ could contradict our hypothesis, it could also be the consequence of a possible bias produced by the CpG placement. Most words around $d = 0$ contain the CpG in their sequence and thus carry an implicit, basal level of conservation. As words move away from the centre, the probability of having an CpG decreases asymptotically until it reaches the average frequency of CpG dinucleotides in the genome. Thus, "central" words would be more conserved and thus the matching score of generated motifs with word distributions would increase, hence increasing the number of effectively discriminative motifs.

FB resistant motifs show oscillatory quality peaks

Matching score difference measurements in FB resistant motifs display an oscillatory trend at both sides of the centred displacement that suggest that motifs could concentrate at certain specific displacements from the CpG target. ESC resistant motif quality metrics also show a slight oscillation. Relating such displacements to biological features would partially enforce our hypothesis, as there could exist local maximums of motif optimality at determined displacements from $d = 0$, rather than an absolute maximum at one specific d.

Binding score difference is a better indicator for motif quality trends than FDR

Although matching scores and FDR distributions do not perfectly correlate, they follow similar trends in all samples. The asymmetry in iPSC prone motif matching scores is also slightly present in its FDR distribution, and the valley in the iPSC resistant motif matching scores around $d = 0$ is represented in FDR distribution as an abrupt general increase of FDR values. Also, the oscillatory trend in FB resistant motifs and the improvement of the overall quality at $d \cong 35, -35$ in prone motifs can be deducted from the FDR distributions. In general, mean matching scores seem to represent more precise quality trends of motifs than FDR distributions.

FB prone motif and iPSC resistant motif graphs show similar shape

In iPSC and ESC, resistant motifs tend to improve as d increases. FB prone motifs follow a similar trend, although displaced motifs do not reach the quality of centred motifs. Considering iPSC cells are FB derived, this equivalence is feasible to be related to a real biological event.

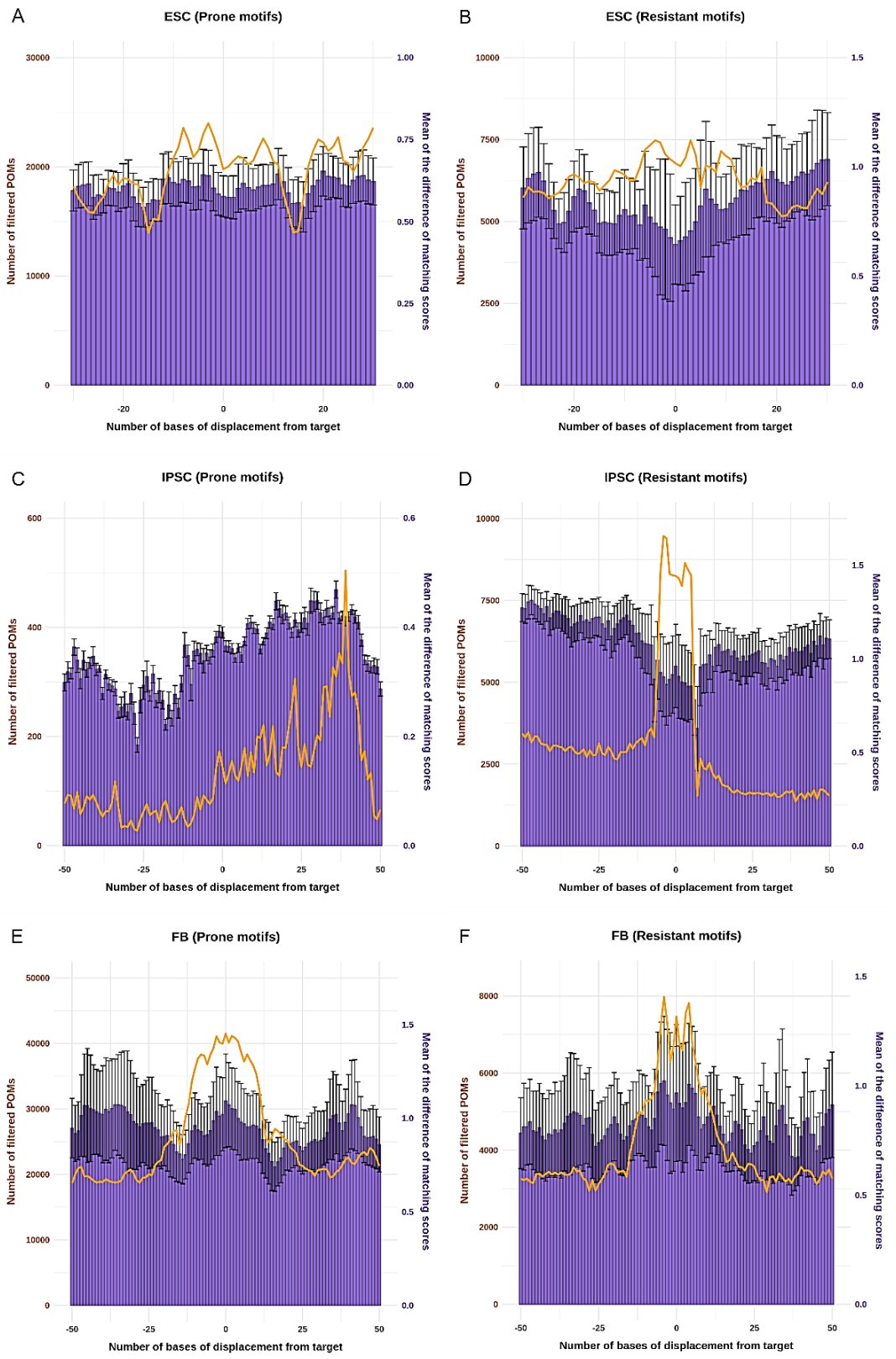

Figure 8 Number of filtered motifs and average of the difference of matching scores for each analysed displacement. The left axis (orange line chart) represent the total number of POMs (motifs) obtained at the end of the pipeline. The right axis (bar charts) represent the average of the difference for each motif between its matching score with its “equivalent” distribution and the other distribution. (A) Prone motifs in ESC. (B) Resistant motifs in ESC. (C) Prone motifs in iPSC. (D) Resistant motifs in iPSC. (E) Prone motifs in FB. (F) Resistant motifs in FB. We could not perform the full displacement range analysis in ESC for time issues.

The execution time of the algorithm has been optimized

The original algorithm by Luu *et al*.[7] took a few days to perform the motif discovery analysis on a single displacement. Our algorithm without optimizations performs the analysis of a single displacement in about 26 hours, whereas the optimized algorithm takes about 15 hours for each displacement.

The cluster execution script is critical for performance. In our specific case, analysing 101 displacements for a unique sample in a cluster of 7 nodes took 217 hours (about 9 days). Without the cluster execution script, the algorithm would take 1515 hours (about 63 days) for the analysis.

Discussion

In this work, we study the role of DNA words in epigenomic regulation by targeting DNA methylation motifs. We propose a hypothesis, that the optimal displacement from a CpG target for DNA motif presence is different to 0, and we suggest two possible molecular models for DNA methylation modulation that support displaced DNA motifs. To validate our hypothesis, we detect DNA motifs at several positive and negative displacements from the CpG target in ESC, iPSC and FB samples.

The presence of discriminative motifs centred on the CpG, apart from being influenced by the CpG positioning bias explained in *Results*, is consistent with previous results on DNA methylation motifs. From the total of the human TFs, for instance, about a 67% have been associated with TFBSs containing CpG dinucleotides at variable positioning within their sequence [4].

Although motifs containing a CpG dinucleotide seem to contradict our hypothesis, they match other DNA methylation molecular mechanisms that could be complementary to our proposal. Methylation resistant motifs containing a CpG may be associated to proteins, such as TFs, sequence-specifically binding unmethylated DNA and preventing it from methylation. TFs binding to DNA to avoid methylation have previously been suggested [7,12] and could coexist with the cooperative and multiple-domain models.

The local maximum of methylation prone motifs with CpG inside its sequence is more difficult to justify in molecular terms, as sequence specific recognition of methylated DNA *in situ* has not been probed. However, previous results support this fact. “Methyl-plus” TFs described by Yin *et al*. [4] have higher affinity for methylated DNA binding sites rather than unmethylated DNA binding sites. Combining the mEpigram methylation motif discovery pipeline and Chip-seq analysis, Ngo *et al.* demonstrated that the enrichment of some methylated TFBS decreased significantly after DNA demethylation [6]. The existence of a molecular model for sequence-specific methylated DNA recognition by TFs would therefore be plausible.

The equivalence between FB prone motifs and iPSC and ESC resistant motifs can be related to molecular mechanisms for reprogramming, enforced by the fact that the iPSC are FB-derived. For instance, *OCT4* and *NANOG* are crucial genes for cell pluripotency, silenced through promoter methylation in non-pluripotent cells. It is known that in cell reprogramming promoters of such pluripotency genes are sequence-specifically recognised for unmethylation, activating their expression [7]. As previously

cited, there exists no evidence of any molecular process for sequence-specific recognition of methylated DNA for unmethylation. Then, *OCT4* and *NANOG* activation could occur through specific recognition of unmethylated sequences in their promoters, displaced from their methylated CpG, and recruitment of DNA demethylation proteins (see Fig. 9A). This would explain the similar trends between iPSC and ESC resistant motifs and FB prone motifs. Discriminative prone motifs displaced from the centre in FB could therefore appear unmethylated in iPSC. The same reasoning could be applied to the rest of genes involved in pluripotency.

We find further evidence for similar models in previous literature. The MBD family protein MBD2a is directly related to differentiation and reprogramming, as it binds *OCT4* and *NANOG*. Overexpressed MBDa can cause differentiation in ESC, silencing *OCT4* and *NANOG* through the recruitment of the chromatin remodelling and histone deacetylase complex NuRD [11]. Histone deacetylation removes an acetyl group from the histone tail, tightening the wrapping of the DNA around the nucleosome, reducing DNA accessibility, and thus silencing its expression. MBD2 proteins have an MBD domain and sequence specific TRDs. Thus, MBD2a could match both the cooperative and the single protein, multiple domains models. Once a pluripotency gene is methylated in the differentiation process, MBD2a would sequence-specifically recognise the pluripotency promoter through its TRD and the methylated CpG through its MBD. Then, MBD2a would recruit the repression machinery for pluripotency genes, in this case NuRD (see Fig. 9B). Unlike the previous model, MBD2a recruit chromatin remodelling factors rather than MUFs.

The two pluripotency gene regulation models could be complementary and consecutive. During differentiation, specific promoter sequences would be recognized and methylated. Once differentiated, expression of pluripotency genes would be repressed through sequence-specific histone modification and chromatin remodelling (see Fig. 9).

Oscillatory motif quality trends in FB and ESC resistant motifs can be the consequence of multiple factors. The effect of steric hindrance when recruiting other proteins would depend on the specific structure and size of each TF family [4]. Hence, it is logical to think that different protein families would have optimal motifs at different displacements from $d = 0$. TFs have also been suggested to work in clusters [6,10], and different proteins in a same cluster could concurrently bind to DNA. For instance, pioneer TFs bind chromosomes and recruit other TFs to open the chromatin region [12]. In this context, several motifs could emerge in a small range of nucleotides, as it is the case of FB resistant motif results.

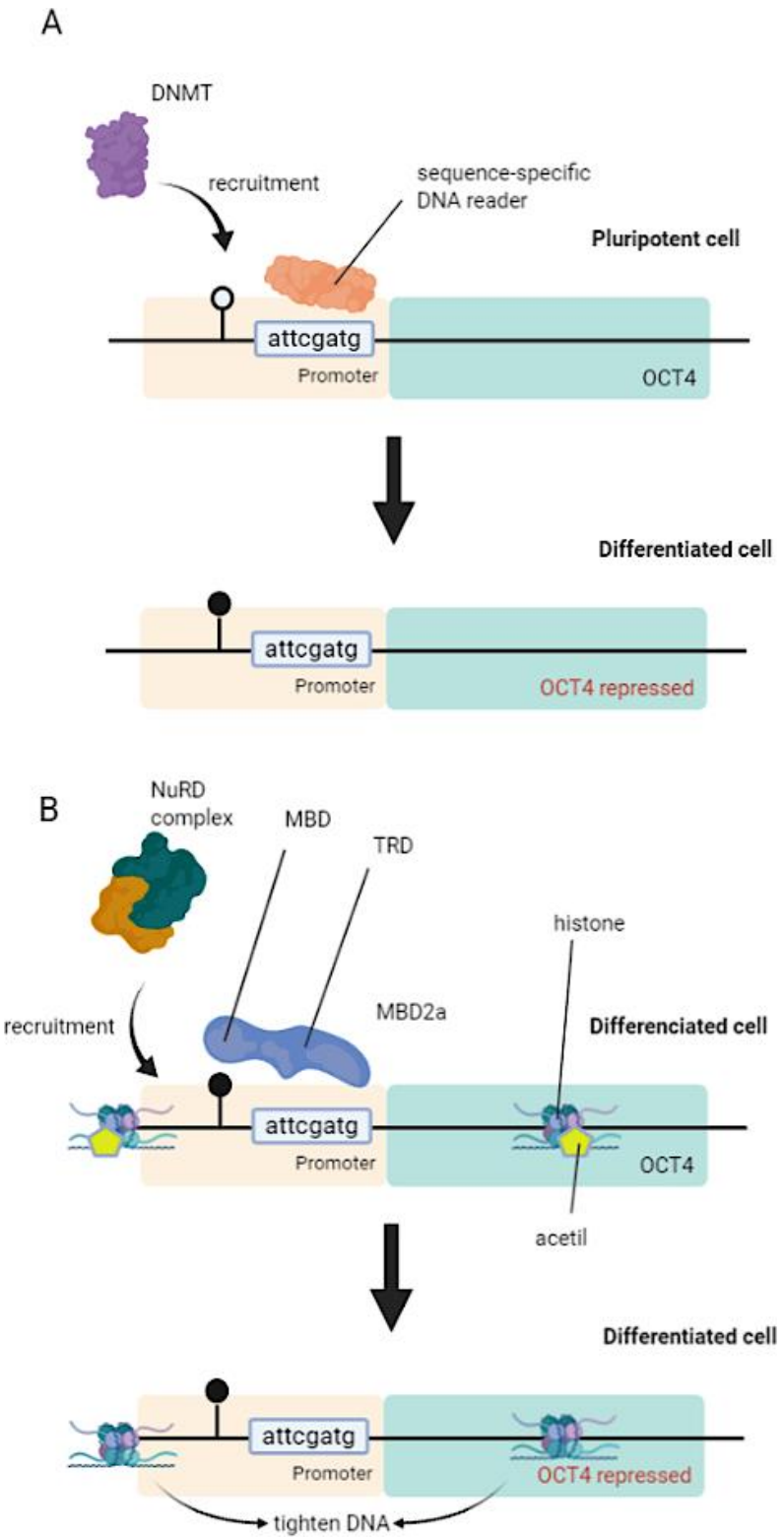

Figure 9 Possible regulation models for the expression of pluripotency genes during differentiation. In the figure, *OCT4* is taken as an example. A methylated CpG is represented with a filled lollipop, whereas an unfilled lollipop represents an unmethylated CpG. (A) In a pluripotent cell, a sequence-specific DNA reader binds the *OCT4* promoter and recruits DNMT proteins for methylation. (B) In a differentiated cell, MBD2a sequence-specifically recognises the *OCT4* promoter through its TRD and not-specifically binds the methylated CpG through its MBD. MBD2a then recruits the NuRD complex for histone modification and

chromatin remodelling. DNA around nucleosomes tightens and loses accessibility, thus repressing the expression of *OCT4*. Created with BioRender.com [15].

The possible relationship between methylation motifs and histone modification makes us think that displaced epigenomic words are not only feasible for methylation motifs. For example, displaced motif have also been reported for histone marks [2]. The co-occurrence of DNA methylation motifs and histone marks has also been previously studied [7] and it previous research suggested that DNA methylation regulates discrimination of enhancers from promoters through a H3K4me1-H3K4me3 seesaw mechanism [21]. Our algorithm is generic for the discovery of DNA motifs related to any measured signal. Thus, our pipeline could be extensible to the study of target position-irrespective motifs related to histone marks, as histone modifier proteins would be affected by similar steric hindrance effects as methylation modulators.

Next steps

Once implemented and tested the pipeline on cells at variable differentiation levels, our short-term continuation is the analysis of methylation motifs on additional cell samples. A key insight into methylation motif placement would come from the analysis of cells from the three primary germ layers. Determining common and differential motif quality patterns in such cells would extend our present results and would contribute to the current knowledge about reprogramming.

Concerning the long-term objectives of our project, we could extract, sort, and perform a top-k selection on motifs at local maximums of quality. Then, we would perform ontology tests on such motifs to define their biological functionality. Complementary assays for DNA word functionality definition could also include testing their co-occurrence with other epigenomic marks or performing a motif-comparison assays between our extracted motifs and TFBS databases, such as JASPAR [22].

Biological functionality tests should be validated *in vivo*. Once determined differential motifs for each cell line, we could perform CRISP-Cas9 genome editing on their matching DNA words. We could measure their cell type-specific phenotypical features before and after DNA word splicing. A significant variation between the pre-splicing and the post-splicing states would be an indicator of the biological relevance of our motifs. Validating our results both *in vivo* and *in vitro* would provide the enough consistency to our research for the publication of a public server of DNA words for CRISPR/Cas9 genome editing.

Conclusion

Identifying and profiling functional DNA words in epigenomic regulation is crucial for a better understanding of processes such as differentiation and reprogramming. Such research would favour gene therapy treatment designed, epigenomics-based disease diagnosis and cell reprogramming for research, among other. Methylation is an essential epigenetic mark that works as cause and consequence of many biological mechanisms. Although DNA methylation motifs have been previously studied, the results about genome-wide motifs intrinsic to the DNA sequence are scarce. In addition, previous studies only contemplate motifs centred in the methylation target. Current molecular models for DNA methylation regulation suggest that the interpretation of methylated DNA could be a cooperative mechanism between multiple proteins. Hence, we could think that the binding of proteins with DNA is influenced by steric hindrance between different molecules. Consequently, DNA motifs could emerge at some displacement from the methylation target.

We present a pipeline for genome-wide analysis of target position-irrespective methylation motifs. Our results on ESC, iPSC and FB show that motifs have peaks of quality at determinate displacements from the methylation target. We suggest that displaced motifs could come from pluripotency gene repression mechanisms. Also, we see indications of cooperative binding of multiple proteins to DNA at once, with oscillatory motif peaks separated by a low number of nucleotides. We conclude that DNA motifs prevail at specific displacements from the methylation target that include the methylation target centre.

References

1. Berdasco M, Esteller M. Clinical epigenetics: seizing opportunities for translation. Nat Rev Genet. 2019;20(2):109–27.
2. Whitaker JW, Chen Z, Wang W. Predicting the human epigenome from DNA motifs. Nat Methods. 2015;12(3):265–72.
3. Davis CA, Hitz BC, Sloan CA, Chan ET, Davidson JM, Gabdank I, Hilton JA, Jain K, Baymuradov UK, Narayanan AK, Onate KC, Graham K, Miyasato SR, Dreszer TR, Strattan JS, Jolanki O, Tanaka FY, Cherry JM. The Encyclopedia of DNA elements (ENCODE): data portal update. Nucleic Acids Res. 2017;46(D1):D794–801.
4. Yin Y, Morgunova E, Jolma A, Kaasinen E, Sahu B, Khund-Sayeed S, Das PK, Kivioja T, Dave K, Zhong F, Nitta KR, Taipale M, Popov A, Ginno PA, Domcke S, Yan J, Schübeler D, Vinson C, Taipale J. Impact of cytosine methylation on DNA binding specificities of human transcription factors. Science. 2017;356(6337).
5. Jolma A, Yin Y, Nitta KR, Dave K, Popov A, Taipale M, Enge M, Kivioja T, Morgunova E, Taipale J. DNA-dependent formation of transcription factor pairs alters their binding specificity. Nature. 2015;527(7578):384–8.
6. Ngo V, Wang M, Wang W. Finding de novo methylated DNA motifs. Bioinformatics. 2019;35(18):3287–93.
7. Luu PL, Schöler HR, Araúzo-Bravo MJ. Disclosing the crosstalk among DNA methylation, transcription factors, and histone marks in human pluripotent cells through discovery of DNA methylation motifs. Genome Res. 2013;23(12):2013–29.
8. Ascensión AM, Arrospide-Elgarresta M, Izeta A, Araúzo-Bravo MJ. NaviSE: Superenhancer navigator integrating epigenomics signal algebra. BMC Bioinform. 2017;18(1):296.
9. Tarjan DR, Flavahan WA, Bernstein BE. Epigenome editing strategies for the functional annotation of CTCF insulators. Nat Commun. 2019;10(1):1–8.
10. Müller-Molina AJ, Schöler HR, Araúzo-Bravo MJ. Comprehensive Human Transcription Factor Binding Site Map for Combinatory Binding Motifs Discovery. PLoS One. 2012;7(11).
11. Du Q, Luu PL, Stirzaker C, Clark SJ. Methyl-CpG-binding domain proteins: Readers of the epigenome. Vol. 7, Epigenomics. Future Medicine Ltd.; 2015. p. 1051–73.
12. Zhu H, Wang G, Qian J. Transcription factors as readers and effectors of DNA methylation. Nat Rev Genet. 2016;17(9):551–65.
13. Ovkvist CL¨, Dodd IB, Sneppen K, Haerter JO. DNA methylation in human epigenomes depends on local topology of CpG sites. Nucleic Acids Res. 2016;44(11):5123–32.
14. Roadmap Epigenomics Consortium, Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, Heravi-Moussavi A, Kheradpour P, Zhang Z, Wang J, Ziller MJ, Amin V, Whitaker JW, Schultz MD, Ward LD, Sarkar A, Quon G, Sandstrom RS, Eaton ML, Wu YC, Pfenning AR, Wang X, Claussnitzer M, Liu Y, Coarfa C, Harris RA, Shoresh N, Epstein CB, Gjoneska E, Leung D, Xie W, Hawkins RD, Lister R, Hong C, Gascard P, Mungall AJ, Moore R, Chuah E, Tam A, Canfield TK, Hansen RS, Kaul R, Sabo PJ, Bansal MS, Carles A, Dixon JR, Farh KH, Feizi S, Karlic R, Kim AR, Kulkarni A, Li D, Lowdon R, Elliott G, Mercer TR, Neph SJ, Onuchic V, Polak P, Rajagopal N, Ray P, Sallari RC, Siebenthall KT, Sinnott-Armstrong NA, Stevens M, Thurman RE, Wu J, Zhang B, Zhou X, Beaudet AE, Boyer LA, De Jager PL,

Farnham PJ, Fisher SJ, Haussler D, Jones SJM, Li W, Marra MA, McManus MT, Sunyaev S, Thomson JA, Tlsty TD, Tsai LH, Wang W, Waterland RA, Zhang MQ, Chadwick LH, Bernstein BE, Costello JF, Ecker JR, Hirst M, Meissner A, Milosavljevic A, Ren B, Stamatoyannopoulos JA, Wang T, Kellis M. Integrative analysis of 111 reference human epigenomes. Nature. 2015;518(7539):317–29.

15. BioRender [Internet]. BioRender. [cited 2020 Aug 1]. Available from: https://biorender.com/

16. Luu P-L, Araúzo-Bravo MJ, Gerovska D, Arrospide-Elgarresta M, Retegi-Carrión S, Schöler HR. P3BSseq: parallel processing pipeline software for automatic analysis of bisulfite sequencing data. Bioinformatics. 2017;33(3):428–31.

17. Lister R, Pelizzola M, Kida YS, Hawkins RD, Nery JR, Hon G, Antosiewicz-Bourget J, O'Malley R, Castanon R, Klugman S, Downes M, Yu R, Stewart R, Ren B, Thomson JA, Evans RM, Ecker JR. Hotspots of aberrant epigenomic reprogramming in human induced pluripotent stem cells. Nature. 2011;471(7336):68–73.

18. Tsuji J, Weng Z. Evaluation of preprocessing, mapping and postprocessing algorithms for analyzing whole genome bisulfite sequencing data. Brief Bioinform. 2015;17(6):938–52.

19. Krueger F, Andrews SR. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. Bioinformatics. 2011;27(11):1571–2.

20. Berg OG, von Hippel PH. Selection of DNA binding sites by regulatory proteins. Statistical-mechanical theory and application to operators and promoters. J Mol Biol. 1987;193(4):723–43.

21. Sharifi-Zarchi A, Gerovska D, Adachi K, Totonchi M, Pezeshk H, Taft RJ, Schöler HR, Chitsaz H, Sadeghi M, Baharvand H, Araúzo-Bravo MJ. DNA methylation regulates discrimination of enhancers from promoters through a H3K4me1-H3K4me3 seesaw mechanism. BMC Genom. 2017;18(1):964.

22. JASPAR - a database of transcription factor binding profiles [Internet]. [cited 2020 Jul 3]. Available from: http://jaspar.genereg.net/

Self-evaluation

The project progressed based on the expectations and the available time. We could implement the algorithm, tune its parameters up for coherent biological results, optimize its execution time and perform the analysis on different samples. Initially, objectives were broader than the ones reached, as we planned to perform some complementary assays for result consistency. However, as each analysis takes several hours to complete, the tuning of the algorithm took longer than expected. However, I personally consider the research reached interesting results that allowed a coherent argumentation for the presented hypothesis.

Personally, my stay was greatly profitable both in personal and academic terms. In the professional domain, my supervisor presented the project and its objectives to me from the start, clearly and without ambiguities. A remarkable fact about the proposal was that it was highly individualized, as my supervisor and I were the main researchers of the project. Although I was constantly supervised and we performed follow up meetings periodically, I was given a wide range of freedom. Frequently, they gave me the possibility to make my own decisions during the research path and considered all my contributions. Thus, my stay was exigent concerning responsibility but favourable in terms of applying both my own criteria and the knowledge acquired during my degree.

Regarding the domain of the project, I found it of great value for the transversal development of my whole Double Degree in Biotechnology and Computer Engineering. Despite the research was focused on the retrieval of biological results, it had a strong component of informatics. In fact, I had to learn and dominate R, a completely programming language for me. I could therefore develop new skills for bioinformatics programming and algorithmics.

Finally, and concerning my personal experience in the group, it was a great introduction to the research community. I was constantly backed by my teammates and we collaborated with each other even if we were working on different projects. A positive aspect about the group was its heterogeneity in researcher profiles, from computer engineers to biochemists. The coherence of the group hence compensated the individuality of the project.

Annex A: Detailed dataset features

Table 1 Details of the analysed datasets.

Cell line	Cell type	Library	Reference genome	CG coverage (%)	WGBS annotation pipeline	Experiment reference	File reference
FF-iPSC 19.11	iPSC	SRA	hg19	84.7	PB3Seq	[17]	Annotated by us
HUES64	ESC	ENCODE	hg38	78.6	ENCODE	ENCSR354DMU	ENCFF331VRY
GM23248	FB	ENCODE	Hg38	83.6	ENCODE	ENCSR625HZA	ENCFF116DGM

Annex B: Computational optimizations of the original algorithm

R optimizations

We used the R library `doMC` to parallelize the filtering through FDR and Mann-Whitney tests across multiple CPUs. For low latency R data frame operations, we used pipelining from the `magrittr` library.

Low level routines

Some steps of the pipeline show excessively high latency in their implementation in R, even using parallelization libraries. We thus implemented C routines for some phases of the algorithm of heavy computational burden. The initial extraction of DNA words from the genome, the generation of the similarity matrix for clustering, the scanning of motifs against sequences for matching score calculation and some complementary routines are written in C.

For C calls from R we used the `.Call` base R function.

HPC execution script

As the analysis of each displacement is independent from the others, we implemented an encapsulating bash script for the execution of our pipeline on slurm-based clusters. We use a bash FIFO queue for the enqueuing of the displacements to be analysed and an automatic recognition of available nodes in the slurm context. Thus, the algorithm sequentially assigns idle nodes to displacement analysis, detects the end of analysed displacements and manages the sequential repartition of nodes from finished displacements to displacements in the analysis stack.

Annex C: Additional software design features

Motifs are usually represented with logos that visually show the relevance of each of the four possible nucleotides at each position of the motif. Our algorithm can be parameterized by the user to optionally generate logos for all the filtered motifs at the end of the pipeline. Logos are generated with the R function `seqLogo` from the `seqLogo` library. We depict an example for a motif logo in Fig. 10.

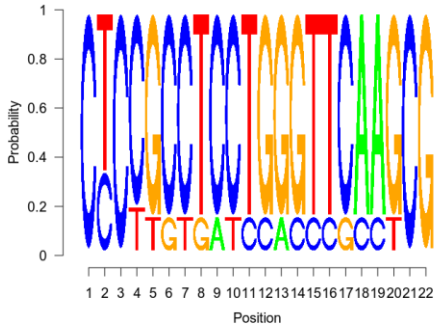

Figure 10 Example of a logo for a 44-nucleotide-length motif.

We also provide HTML files that contain matching score histograms for both distributions for each of the filtered motifs, as depicted in Fig. 11. Such information files are divided into methylation-prone and resistant motifs and positive and negative DNA strands.

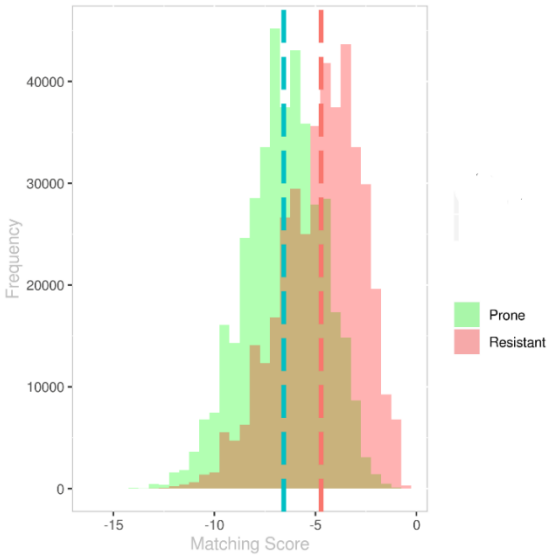

Figure 11 Matching score distribution representation for a specific motif.

Finally, the user can optionally perform gene assignment of the filtered motifs by their closeness to an annotated gene. The user only must provide a gene annotation file of the reference genome and turn the gene assignment parameter on.

UNIVERSITAT ROVIRA i VIRGILI