UVIC UNIVERSITAT DE VIC

Master of Science in Omics Data Analysis

Master Thesis

Finding and characterizing Partially Methylated Domains in human haematopoietic cells

by

Martí Duran Ferrer

Supervisor: Simon Heath, Bioinformatics Development Group Leader, CNAG

Co-supervisor: Angelika Merkel, Statistical Genomics Team, CNAG

Department of Systems Biology

University of Vic – Central University of Catalonia

[25/09/2014]

Finding and characterizing Partially Methylated Domains in human haematopoietic cells

Martí Duran Ferrer

DNA cytosine methylation has been demonstrated to be a central epigenetic modification that has essential roles in a myriad of cellular processes. Some examples of these include gene regulation, DNA-protein interactions, cellular differentiation, X-inactivation, maintenance of genome integrity by suppressing transposable elements and viruses, embryogenesis, genomic imprinting and tumourigenesis. This list is increasingly growing thanks to recent advances in genome-wide technologies, like Whole Genome Bisulfite Sequencing (WGBS-Seq). The development of this technology in research has allowed the identification of new features of the DNA methylation landscape that was not possible using previous technologies, like Partially Methylated Domains (PMDs). PMDs have been found in several cell lines, as well as in both healthy and cancer primary samples. They have been described as regions with high variability in methylation levels across individual CpG sites and intermediate methylation levels on average with respect to the genome. Here, we performed an extensive search of PMDs in a big dataset of different haematopoietic primary cells from both myeloid and lymphoid lineages. We found and characterized significant PMDs in plasma B cells, confirming that PMDs are a phenomenon that is restricted to certain differentiated cells. Additionally, we found loci aberrantly hypomethylated in a myeloma sample which overlapped with plasma B cell PMDs. Genome-wide comparison of the myeloma and plasma B cell sample revealed that this is probably also the case for other loci.

INTRODUCTION

DNA methylation is a very common epigenetic mark, which involves a covalent attachment of a methyl group to the C5 position of the cytosine ring **(Cedar H, 2009).** The enzymes responsible for the methylation are grouped in the family of DNA methyltransferase enzymes DNMT **(Okano M, 1998)**, namely the catalytic active DNMT1, DNMT3A and DNMT3B, and DNMTL3, a catalytic inactive homologue to both DNMT3A and DNMT3B.

Each of the catalytically active DNMTs, have been demonstrated to be crucial for embryonic development, and the complete ablation of methylation provokes embryonic lethality **(Li E, 1992; Okano M, 1999).** DNMT1 is thought to be principally involved in the maintenance of pre-existing methylation, and DNMT3A and DNMT3B act as *de novo* methyltransferases, modifying unmethylated DNA. Embryonic stem (ES) cells lacking both DNMT3A and DNMT3B progressively lose differentiation potential with cell passage, although the potential for self-renewal is maintained **(Cheng T, 2003)**. Additionally, recent evidence shows that DNMT1 is necessary for self-renewal of Haematopoietic Stem Cells (HSC), and

also for its differentiation pattern, since its depletion promotes a dominant myeloid cell development **(Trowbridge J.J, 2009; Bröske A.M, 2009).** Finally, DNMT3A has been recently demonstrated to be essential for haematopoietic stem cell differentiation **(Challen GA, 2011)**. Collectively, these studies highlight the importance of DNA methylation for normal embryo development and cell fate commitment. In addition to this, DNA methylation has been demonstrated to play a major role in many diseases, for example in cancer **(Hanahan D, 2011).** The transcriptional start sites (TSS) of many genes encoding tumour suppressor genes have been found to be hypermethylated in cancer, for example retinoblastoma associated protein 1 (RB1), MLH1, p16 and BRCA1 (among others). This hypermethylation at TSS of tumour suppressor genes has been found to lead to gene silencing, thus promoting cancer progression **(Esteller M, 2008)**.

However, until recently, much of the work on DNA methylation in mammals has been focused on the 5-methylcytosine in the CpG sequence context, especially in CpG Islands (CGI) (regions with higher CpG density than expected) and at TSS. This work began in the middle of 80's with the studies of **Holliday R, 1975** and **Riggs AD, 1975**. In those studies, the authors suggested that DNA sequences could be methylated *de novo* by certain enzymes, that methylation could be inherited through somatic cell divisions and that DNA methylation directly silences genes. Although some of these claims hold true, the role of the DNA methylation in gene regulation has been demonstrated to be challenging to unravel. Recent improvements in genome-wide sequencing technology have shed light on new epigenetic mysteries, far beyond those of classical DNA methylation studies on a single locus. For example, DNA methylation located in the vicinity of TSS obstructs transcription initiation, but methylation in gene body not only does not block transcription, but might even favours transcription elongation. Even more excitingly, new evidences indicate that DNA methylation in gene bodies could regulate splicing **(Jones PA. 2012)**. Also, methylation in other sequence contexts than CpG, namely CHG or CHH (where H is A or T), have been proved to be widespread in plants and fungi **(Cokus SJ, 2008; Rountree MR, 1997)** and have been recently reported in H1 human embryonic stem **(Lister R, 2009)**. Additionally, differentially methylated sites have also been reported in contexts other than CGIs, called CpG islands shores **(Irizarry RA, 2009)**. Also, other types of methylation have been reported such as 5-hydroximethylation of 5C of cytosine **(Pastor WA, 2011).**

Furthermore, new methodologies in the study of DNA methylation have facilitated the finding of new phenomena at many different scales in human methylomes. Some examples are DNA methylation Valleys (DNMVs) **(Xie W, 2013)**, Unmethylated, Lowly Methylated and Fully Methylated Regions (UMR, LMR and FMR) **(Stadler MB, 2011)** or Partially Methylated Domains (PMDs). PMDs were first described by **Lister R et al (2009)** as regions of intermediate methylation levels with a mean of 153kb, constituting a medium-large scale methylation phenomenon. The authors found them

in differentiated cells (IMR90 lung fibroblasts) but not in human embryonic stem cells (H1-ESC). Since then, numerous studies have demonstrated their presence in different cultured and cancer cell lines, including foreskin fibroblasts (FF) and Adipose derived stem cells (ADS) **(Lister R, 2011)**, SH-SY5Y neuroblastoma **(Schroeder DI, 2011)** and human mammary epithelial cells (HMEC) **(Hon GC, 2012)**. They also have been found in human sample tumours **(Marzese DM, 2014; Hovestadt V, 2014; Berman BP, 2011 and Hansen KD, 2011)**, and finally, PMDs have been reported in human placenta **(Schroeder DI, 2013)**, representing the first human uncluttered and non-cancer tissue type having PMDs**.** Thus, PMDs are emerging as a new feature in methylation landscapes that cover huge portions of the genome, and their characterization is of paramount importance to understand their role in health and disease.

In this study, we systematically searched for PMDs in several haematopoietic primary cells included in myeloid and lymphoid branches. We found that PMDs were prominent in the lymphoid branch and in particular in terminally differentiated B cells, plasma B cells (plBC). Impressively, PMDs found in plasma B cells covered about 70% of the genome representing the second human uncultured and non-cancerous tissue type reported to have PMDs. We show that plBC-PMDs coincide with histone modification marks associated with heterochromatic regions, and are strongly depleted for active genes and DNAse hypersensitivity sites (DHS). Additionally, they are regions showing depletion of CpG Islands (CGI), a moderate enrichment in Lamina Associated Domains (LADs) and no enrichment for any particularly repetitive element. Gene Ontology (GO) enrichment analysis of Biological Processes (BP) in genes present in PMDs revealed enrichment in developmental, cell fate and morphogenesis biological processes. Finally, we describe a locus in which plBC-PMDs overlapped with hypomethylated regions in a hypomethylated myeloma sample, with lamina associated domains (LADs), and with late replication regions. Our study adds new evidence to the growing body of literature for PMDs as an important feature of human methylomes. Understanding how these large domains are formed, what their functions are and how they are maintained through cell divisions will be of paramount importance to understand normal cell fate commitment, differentiation, development and ultimately its role in health and diseases like cancer.

RESULTS

Finding PMDs across haematopoietc cell types.

We have analysed WGBS-Seq of a series of haematopoietic samples in the context of the generated BLUEPRINT epigenome project **(Adam D, 2012)**. Samples included cells at various stages of myeloid and lymphoid commitments, and a B cell differentiation dataset **(Supplementary Table 1).** Additionally, we included data from a very recent study **(Hansen KD, 2014)**, in which the authors found hypomethylated regions associated to the infection of EBV to B-cells.

To search PMDs in this huge dataset, we took advantage of a recently published algorithm available as the MethySeekR package **(Burger L, 2013, Materials and Methods)**. Briefly, the algorithm estimates the distribution of polarized (unmethylated, fully methylated) and intermediate methylation values across a small training set (sliding window on chr22) and applies a two-state HMM to segment the genome into PMDs and non-PMDs

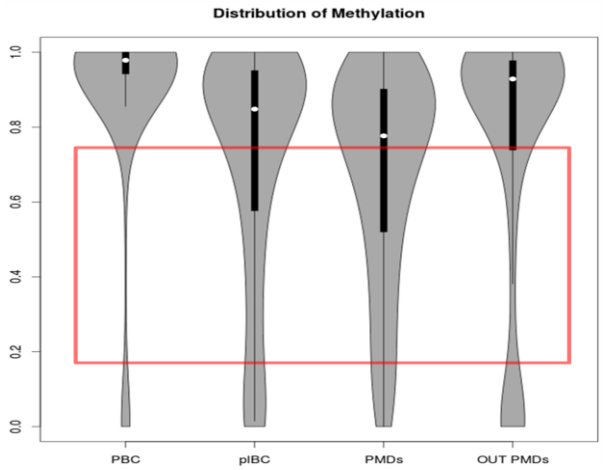


Figure 1. Density distribution of methylation of PBC, plBC, and the PMDs found in plBC and outside of PMDs. PBC do not show intermediate methylation, which is in contrast to plBC (marked with red rectangle). PMDs regions contain an increase in those intermediate levels of methylation and, importantly, outside PMDs it can be shown a depletion of those levels which resembles the PBC genome. PBC, Progenitor B Cell, plBC, plasma B Cell, respectively.

(sexual chromosomes were removed from the analysis to avoid confounding effects of imprinted genes). After applying the software to our dataset, we did only confidently found PMDs in one of the samples, the plBC (B cell differentiation dataset) **(Supplementary Figure 1A-C, Materials and Methods, Figure 1)**. This can be explained by two plausible reasons: First, the majority of the mammalian published methylomes are divided in a fully methylated state and an unmethylated one, and do not contain a significant amount of intermediate DNA methylation **(Lister R 2009 and Burge L, 2013)**. Consequently, this polarized distribution of DNA methylation throughout the genome departs from the search of the algorithm, which looks at regions with high variability in the methylation levels (and an average intermediate methylation). Second, although some of samples from **Hansen KD et al, (2014)** displayed some preliminary parameters indicative for PMDs presence in the methylome (i.e., a bimodal or long-tailed distribution of α-values with a significant fraction of windows with $\alpha > 1$, **see Materials and Methods**), we discarded those results, since the data was sequenced at low coverage, below the desired one to use the algorithm. Consequently, from there on, we focused or analysis of PMDs in plasma B cells in the context of B cell differentiation dataset.

Global loss of methylation upon differentiation.

One implication of finding PMDs in a particular genome, is that it has to have regions with on averaged low methylation values and high variability of methylation **(Burger L, 2013).** Thus, given that we have only found PMDs in the plasma B cell, which is the most differentiated cell type, it suggests that progenitor B cell have lost methylation through differentiation. Indeed, when we plotted CpG pairwise genome-wide comparisons of the levels of methylation between PBC and the other samples of the B cell differentiation dataset, we clearly appreciate this global loss of methylation during cell differentiation **(Figure 2 and Figure 1)**. This is in agreement with some other studies of

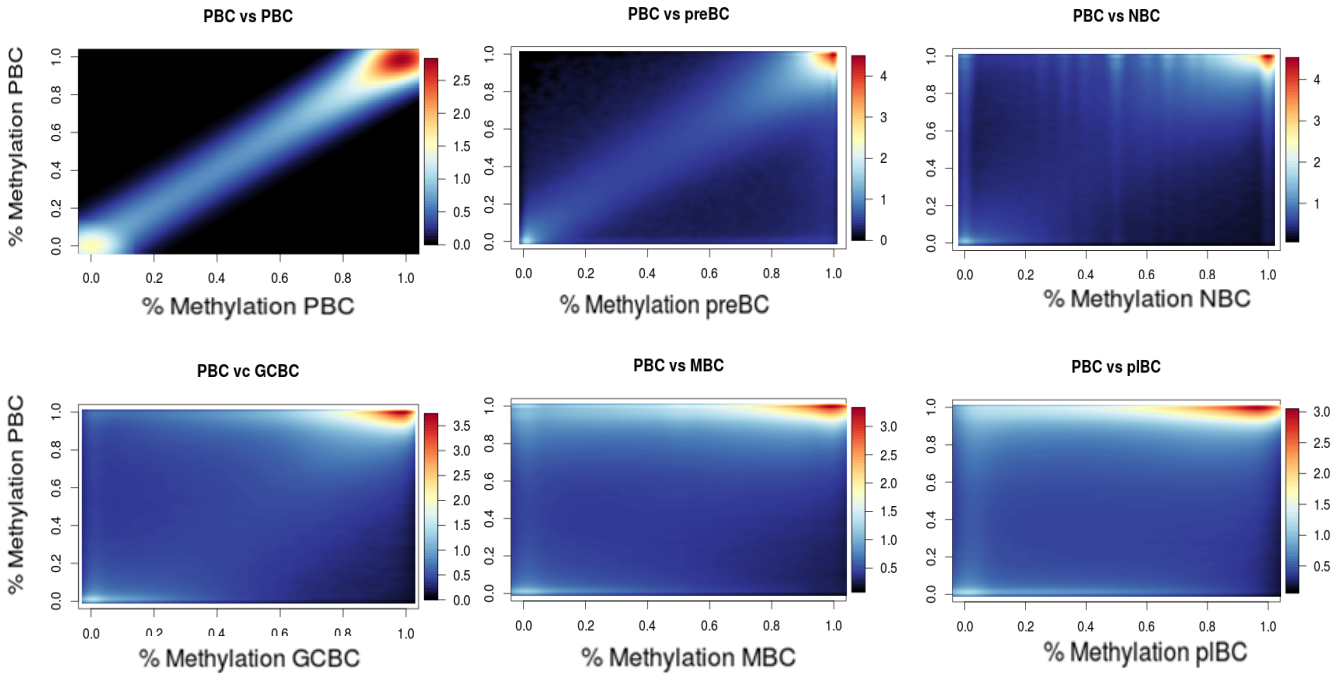


Figure 2. Global methylation loss during B cell differentiation. PCB loss methylation in CpGs upon differentiation. This loss is more accentuated form GCBC onwards. PBC (progenitor B cell), preBC (pre B cell), NBC (Naïve B cell), GCBC (Germinal center B cell), MBC (Memory B cell) and plBC (Plasma B cell), The bars show the density of CpGs methylation.

methylation upon differentiation and cell fate commitment **(Ji H, 2010; Kulis M, 2012).**

Characterization of plBC-specific PMDs.

Once we found regions of partially methylation in the plasma B cell genome, we sought to characterize them. Examination of the PMDs, revealed that they covered about 70% of plasma B cell genome, spanning from 0,8kb to 22MB, with a median of 50kb and a mean of 245kb, which are metrics greater than those of the first report **(Lister R, 2009)**. In fact, within the plBC genome these domains were visible at chromosome level **(Supplementary Figure 2)**. Surprisingly, PMDs are not homogeneously distributed across the genome, but range from 38% of chromosome 22 to 83 % of chromosome 21 **(Supplementary table 2)**. This difference in coverage of each autosome can not be explained by the different Pearson's correlation of -0.096, adjusted R^2**=0.04, Supplementary Figure 3).** Thus, we speculate that PMDs in plBC genome are domains which can be involved in many genomic contexts, from small-scale phenomena like promoters or transcription start sites, to large-scale ones, like LADs or LOCKs **(Gulen L, 2008; Wen B, 2009)**. However, for a simplistic approach we followed our analysis with the whole set of PMDs, including PMDs of all sizes.

Next, to get insights into the functionality of these regions, we examined the enrichment of various chromatin states inside and outside of PMDs. We used the chromatin segmentation data of the lymphoblastoid cell line GM12878 **(Ernst, J, 2011)** available at UCSC Genome Browser **(Kent WJ, 2002)**. We performed two different combinations of enrichment analysis: First, we

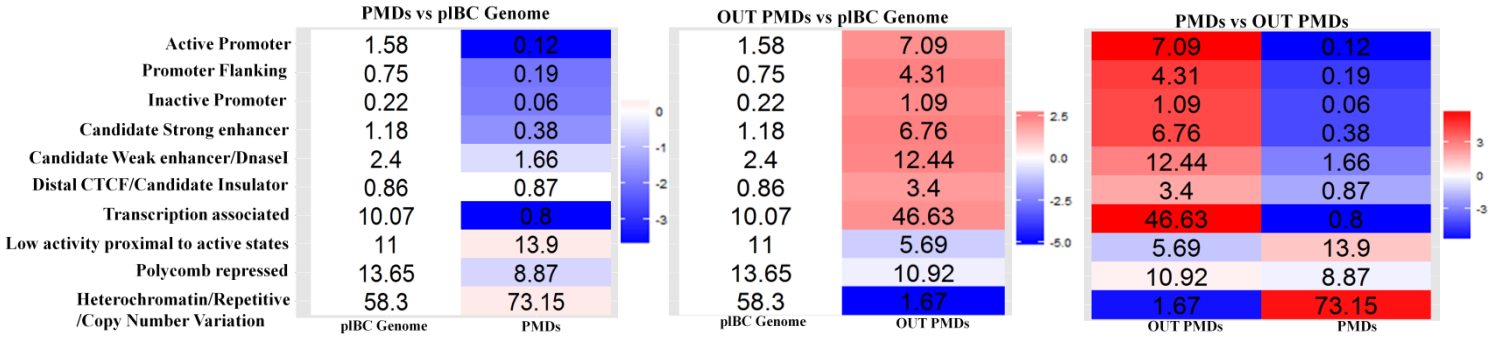
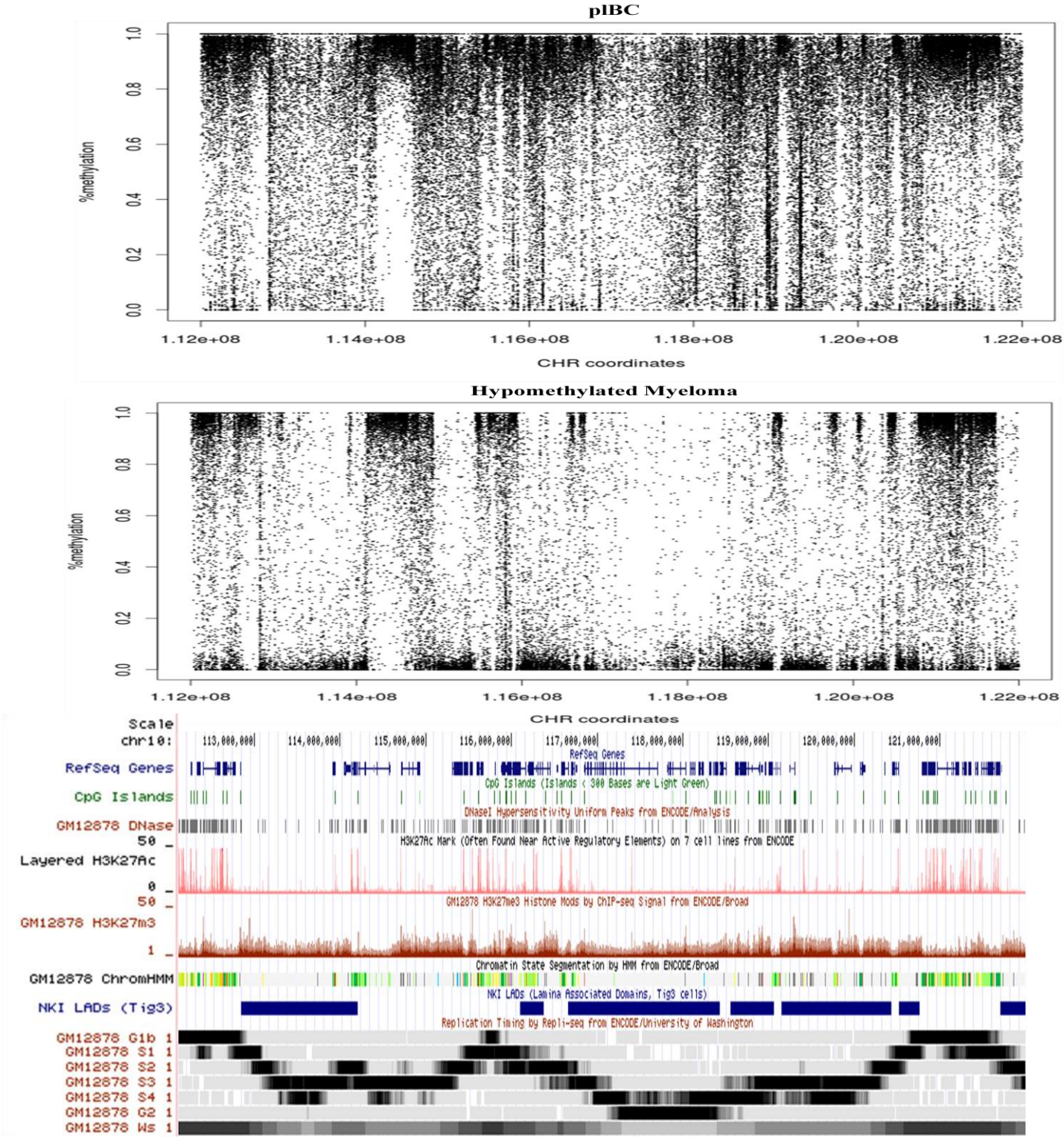


Figure 3. Distribution and enrichment analysis of chromatin states in PMDs and OUT PMDs groups compared to plBC genome (left and central table) and with themselves (right table). Numbers inside each cell represent the percentage of each feature in each group. The colors represent the log2 fold change enrichment of PMDs with respect to the group compared.

calculated the proportions of each chromatin state in PMDs and outside PMDs, and compared them with the coverage across the whole genome. Second, we compared chromatin segments inside PMDs vs segments outside PMDs **(Figure 3).** Compared to the entire genome, we found that PMDs are strongly depleted of gene activity **(Figure 3, left table)**. We also found that they reside primarily in heterochromatic regions. Conversely, outside PMDs showed an enrichment in states related to gene activity, supported by the enrichment of states like "Active Promoter" or "Transcription Associated", and a large depletion of heterochromatic regions **(Figure 3, central table).** The direct comparison between the two groups highlighted the differences even stronger **(Figure 3, right table)**. Finally, to get clearer insights into the establishment of those PMDs, we looked at the chromatin states of human embryonic stem cell precursor (H1-hESC) in PMDs regions found in the plBC genome. This comparison revealed that PMDs were regions in H1 undergoing extensive transcription and extremely depleted of histone repressive marks associated with heterochromatin **(Supplementary Table 3).**

Following the notion that PMDs represent heterochromatic transcriptionally inactive regions, we sought to support our analysis by examining the overlap of PMDs regions with DNAase I hypersensitivity sites (DHS). DHS are regions which have an accessible chromatin state **(Thurman RE, 2012),** representing another mean to further decipher the functional signature of PMDs regions. We found that DHS was 2.3 times depleted in PMDs compared to plBC genome, whereas they are 2.15 enrich outside PMDs (compared to the genome). Directly compared we found DHS 5.3 times enrich inside PMDs compared to regions outside PMDs. This further suggests that PMDs are regions with low gene activity **(Supplementary table 4)**.

Next, we sought to determine the presence of CpG Islands (CGIs) in PMDs. CGIs are regions found in the 5' region of the majority of human genes, and are related to the regulation of the preceding genes **(Illingworth RS, 2009).** We found a depletion in CGIs in PMDs compared to the plBC genome (2.3 fold), and about the same amount of enrichment outside of PMDs. Again, comparing PMDs and regions outside of PMDs



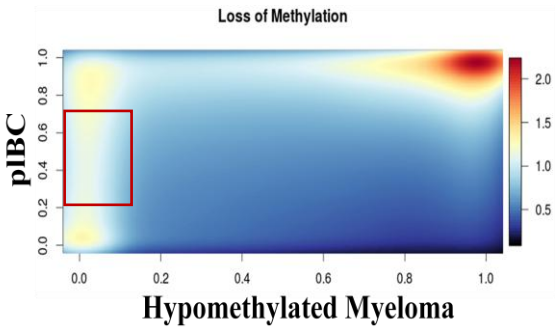


Figure 4. Top: Single CpG methylation of an amplified locus of plBC genome containing PMDs. Middle figures: same region as plBC is displayed for the hypomethylation myeloma, with UCSC genome browser annotation under them. Bottom left: CpG genome-wide comparison of methylation between plBC and hypomethylated myeloma. PMDs in plBC displayed in chr10 region are aberrantly hypomethylated in hypomethylated myeloma. Those regions are marked by heterochromatin marks (e.g H3K27me3) and depleted for active (H3K27ac), are late replicating regions and contain LADs. Looking at genome-wide comparison, this might be the case for other locis with partially methylated state in plBC and hypomethylated in this cancer sample (Red rectangle).

against each other, revealed difference, - a 5-fold depletion of CGIs in PMDs **(Supplementary table 5)**. Interestingly, the number of CpGs, the length of CpGs and the GC content in CGI in each group did not show any clear difference **(Supplementary Figure 4)**.

Then, since a significant proportion of PMDs are very large (data not shown), we sought to determine if these regions overlap with Lamina Associated Domains (LADs). These domains are large regions (0.1-10MB) that are characterized by low gene-expression levels and represent zones of contact between chromosome and the nuclear lamina **(Guelen L, 2008).** Thus, given that LADs are regions rich in heterochromatin and have poor gene expression, we sought to determine whether PMDs were enriched in LADs. We found that LADs cover about 42% of the plBC genome and in PMDs about 49%. Conversely, outside PMDs they only covered 27%. Finally, comparing PMDs and regions outside PMDs revealed an increase of 76% of LADs. This data suggest that PMDs have a moderate enrichment in LADs **(Supplementary table 6)**, although we suggest that this enrichment becomes greater when filtering PMDs of larger sizes (discussed hereafter and data not shown). Afterwards, we asked if PMDs regions were enriched in any particular repetitive element, since heterochromatic regions are rich in repetitive elements. In general, we found instead the same proportions of repeat elements in PMDs with respect to the genome, although some of them showed a slight increase or depletion. The same was true for elements in regions outside of PMDs with the exception of Satellite elements, which showed depletion with respect to the genome and also to the PMDs. **(Supplementary table 7).**

Following the functional characterization, we next sought to determine the biological processes (BP) related with PMDs. To do so, we performed an Gene Ontology (GO) enrichment analysis of BP of curated Refeq mRNA genes belonging to three different groups: First, we selected genes totally overlapping PMDs, second the same but for genes that fell outside PMDs, and finally, we selected those genes overlapping the border of PMDs. It has been demonstrated that the border of LADs can be enriched in promoters of genes being transcribed outside of LADs **(Guelen L, 2008)**. Although we did not found an astonishing enrichment of LADs in PMDs (although the enrichment becomes larger if one filters for large PMDs), we wanted to explore PMD boundaries to see if there was any trace of a possible enrichment in a particular BP for those genes. We found 14246 Refseq mRNA genes inside PMDs, 26358 outside PMDs and 8959 at PMDs borders, confirming that PMDs are regions depleted of genes. Interestingly, we found an impressive enrichment in functional gene annotations related to development and cell fate commitment in PMDs (e.g. "multicellular organismal development", "cell fate commitment" and "stem cell differentiation") **(Supplementary Table 8).** Conversely, outside of PMDs, we did not find any BP related to development, but the majority were related to cell activity (e.g. "cell cycle", "DNA damage checkpoint", etc.) and also to the immune system, like "immune response-activating signal transduction", different BPs associated with toll-like receptors or MHC-II.

(Supplementary Table 9). On the other hand, for genes overlapping the borders of PMDs, we found more variability in BPs, thus it is likely that those genes are not a separate category **(Supplementary Table 10)**.

PMDs in cancer.

We found and characterized PMDs in a genome-wide manner. However, we wanted to investigate the phenomena in a locus specific manner. Thus, we visualized the methylation status of single CpGs in a 10 Mb region on chromosome 10, a region previously analysed in fibroblast (IMR90 cell line) by **Gaidatzis D, (2014) (Figure 4, top)**. We found a high concordance of PMDs published by that study and our data in that region (80% of overlap), although plBC-PMDs were larger than those in IMR90, in agreement with the greater overall coverage of PMDs in the plBC genome. Accordingly, we found that in the plBC genome the methylation levels in PMDs were also highly variable **(Figure 4, top)**, suggesting that this variability in the methylation accompanied by intermediate average methylation levels might be the most distinctive features of these domains, and thus this should be taken into account when searching for them. Curiously, when visualizing the same region in a hypomethylated myeloma sample, we found regions where PMDs coincided with hypomethylated regions in cancer **(Figure4, middle).** As we have found previously with our genome-wide analysis **(Figure 3)**, PMDs in this loci also coincided with histone modifications associated with repressive chromatin states (H3K27me3) and with a depletion in histone modifications related to gene active regions (H3K27ac) **(Figure 4, middle bottom)**. Furthermore, they were depleted of CpG islands and DHS. Interestingly, these PMDs also coincided with LADs and with regions that replicate late in the cell **(Figure 4, middle bottom).** Importantly, the overlap of hypomethylation in the myeloma and the PMDs in plBC might not only be the case for the genomic loci shown, but the same might hold true for other genomic regions which display partial methylation levels in the plBC genome and hypomethylation in the hypomethylated myeloma cancer **(Figure 4, bottom left).**

DISCUSSION

Almost 5 years have passed since the first human base-pair resolution methylome was published, and during these years, a plethora of new methylation landscapes have been described, including PMDs. Here, we performed an extensive search for PMDs in haematopoietic methylomes including cells from both myeloid and lymphoid branches. We only found PMDs in a single cell type, plBC, which represents the most differentiated cell of the B cell lineage. Importantly, this is the second report of PMDs in human uncultured and non-cancerous tissue type, suggesting that PMDs might be also methylation phenomena of adult tissues and not only of developmental tissues **(Schroeder DI, 2013)**. We found that the plBC genome is covered by almost 70% of PMDs and they are unequally distributed along the genome and range from 0,8kb to 22Mb in size (with a median of 50kb and a mean of 245kb). Those metrics are greater than those originally reported by **Lister R, 2009**. However, it should be noted that **Lister R, 2009** used a different approach to find PMDs in the IMR90

and H1 cell lines, calculating average methylation levels in sliding windows across the genome and selecting for regions with methylation values less than 0.6 or 0.7. As demonstrated here and in **Gaidatzis D et al (2014)**, PMDs are domains with high variability in methylation which is not taken into account by the sliding window approach. Thus, it might be the case that using a sliding window approaches regions could be wrongly identified as PMDs (for example, imprinted regions). In any case, those differences in the approaches to find PMDs should be borne in mind, and they might influence the differences in the metrics found in those other studies. Another possible explanation is the fact the other studies were primarily made on cluttered cells rather than on primary tissues samples. On the other hand, we decided to study PMDs including of all sizes. As commented above, this implies that PMDs might be present in different genomic scale phenomena, from small to large ones. Thus, it will be needed in the future to split PMDs according to their length and study each subgroup as a single entity. In fact, our preliminary results demonstrate that there are differences in the distribution of some genomic features studied here for example, as commented above, an increase in LADs when PMDs are filtered as larger regions.

We have extensively characterized PMDs with several analyses, studying the distribution of different genomic features in PMDs with respect to non-PMDs and the whole plBC genome. Chromatin segmentation analysis revealed enrichment in heterochromatic regions and a large depletion in gene activity marks in PMDs **(Figure 3)**. This is in agreement with all the previous studies addressing the chromatin signature of PMDs. (**Lister R, 2009; Hon GC, 2012; Marzese DM, 2014 and Hovestadt V, 2014)**. Furthermore, **Berman NP, 2011**, found that PMDs of colorectal cancer cell line overlapped with LADs, another established repressive chromatin mark **(Gulen L, 2008)**. We found a moderate genome-wide enrichment of LADs in PMDs, and an impressive overlap in the locus analyzed **(Figure 4).** However, as commented above, further splitting PMDs according to size could clarify the presence of LADs in PMDs of the plBC genome. All these correlations of PMDs with inactive chromatin states have been complemented with microarray and RNA-seq experiments, verifying that PMDs are regions with inactive genes **(Schroeder DI, 2013, Lister R 2009)**. Thus, it is clear that PMDs are regions characterized by chromatin repressive marks and poor gene expression. In addition to this, we found a strong decrease in CGIs. This depletion in CGIs can be explained by the observation that CGIs of the human genome exist mostly in a fully or unmethylated state **(Deaton AM, 2011)**, and thus are excluded from the search for PMDs. Furthermore, CGIs have been shown to colocalise with the promoters of constitutively expressed genes **(Illingworth RS, 2009)**, whereas PMDs are marked with repressed states **(Figure 3)**. Accordingly, we found a decrease in gene density in PMDs, and genes outside PMDs were enriched in BP related to active metabolism and B cell commitment **(Supplementary Table 7)**.

When inducing a pluripotent state (iPSCc), **Lister R, 2011** found that PMDs of differentiated cells were transformed to a fully methylated state, and the reprogramming process

was also able to reverse the transcriptional repression associated with the PMDs. Thus, PMDs are established in differentiated cells but not in embryonic or pluripotent cells. As previously stated, we found PMDs in a differentiated plasma B cell, and the genes inside PMDs showed a great enrichment in BPs involved in development and cell fate commitment. These results are consistent with recent studies reporting expanded repressed chromatin domains (H3K9me2, H3K9me3 and H3K27me3) in lineage-committed cells that selectively affect genes involved in pluripotency and development, suggesting that epigenetic mechanisms play a critical role in cellular differentiation and maintenance of the differentiated state **(Wen B, 2009; Hawkins RD, 2010)**. Additionally, they demonstrated that the expansion of those repressive modifications were associated with a decrease in DNA methylation in differentiated cells. We found that PMDs are enriched for chromatin repressive marks, and by definition are regions with an averaged lower DNA methylation. Furthermore, we have found that PMDs were regions in H1-hESC, (which have a fully methylated genome, **Figure 1**) extensively transcribed, extremely poor in repressive chromatin marks, and contained a great proportion of Polycomb genes, which are well known to be involved in developmental processes. Thus, all this data strongly suggest that the majority of demethylation in PBC during differentiation occurs in the form of PMDs **(Figure 2),** take place in regions with genes enriched in developmental BP, and those regions become blocks of heterochromatic regions in the committed cells. Our enrichment analysis also supports this notion, since we found that genes outside PMDs (which are marked with active state) are related to active metabolism and in BPs typical of committed B **(Supplementary Table 7)**. We also demonstrated that PMDs are strongly depleted of DHS. In turn, regions depleted in DHS have been associated with heterochromatin and late replicating regions **(Hansen RS, 2009),** and late replication regions have been found to lose gradually DNA methylation **(Aran D, 2010)**. Thus, it might be the case that PMDs, since they are heterochromatic regions depleted of gene activity and might replicate late in the cell **(Figure 4)**, experience a gradually demethylation process which ultimately leads to a partially methylated state. Results from **Gaidatzis D 2014**, supports this notion. They found that in PMDs, and not outside, the methylation is, to a significant extent, determined by the specific local DNA sequence. They proposed that in PMDs there is a point where the concentration of DNA machinery becomes rate limiting, and it is at this point that DNA sequence preference becomes evident. It might be also the case that PMDs are the result of reducing the cost of methylation maintenance at the inactive and repressed chromatin portion of the genome, where a precise methylation level is presumably not essential. Collectively, these studies suggest that is likely that PMDs are secondary to, rather than causative of, heterochromatin formation.

Curiously, when looking at a particular locus on chromosome 10, we found that PMDs of the plBC genome overlapped hypomethylated regions in hypomethylaed myeloma cancer **(Figure 4)**. This locus displayed the same features found in our genome-wide analysis of PMDs in

plBC, though with a greater overlap of LADs and also with late replication regions. Additionally, pairwise comparison of the plBC methylome and the hypomethylated myeloma revealed that a considerable fraction of hypomethylated sites of the myeloma were in a partially methylated state in plBC **(Figure 4, bottom left)**, suggesting that our finding could be also the case for many different loci. This notion is consistent with other studies showing abnormalities in cancer related to PMDs. In fact, PMDs have been recently reviewed and noted as a prominent feature of tumour methylomes **(Shen H, 2013)**. For instance, **Berman BP, 2011**, found PMDs in colorectal tumours relative to adjacent normal colon, and those PMDs were also found to overlap with LADs. In another study, **Hansen KD, 2011** identified regions with increased variability in methylation levels across cancer types, and these regions overlapped with PMDs. On the other hand, **Hon GC et al (2012)** found that PMDs in HMEC were the most likely regions to either gain or lose DNA methylation during tumorigenesis. In another study conducted by **Hovestadt V, 2014,** the authors found PMDs subgroup specificity in WNT and Group 3 medulloblasomas, and PMDs were found to have a significant increased somatic mutation rate compared to regions outside PMDs. Finally, PMDs were found in melanoma brain metastasis (MBM) patients, and a MBM-specific PMD was able to predict the prognosis of patients with melanoma **(Marzese DM, 2014)**. Collectively, these studies manifest a major role of PMDs in cancer, altering the normal methylomes landscape and promoting tumorigenesis. Further studies will be needed to clarify the mechanisms by which these regions appear in some cancers and why PMDs present in normal cells are hotspots to either gain or loss methylation during tumorigenesis, and what are the consequences of losing or gaining methylation at that sides.

MATERIALS AND METHODS

Samples analyzed.

WGBS-Seq of all the samples was performed as in **Kulis M et al, (submitted)** at the Centro Nacional d'Anàlisis Genòmica (CNAG) in Barcelona, Spain. Details for all samples are in **Supplementary table 1**. The data from **Hansen KD, 2014** can be found in the original study. Briefly, we analyzed WGBS-Seq of the following samples: 6 monocytes, 4 granulocytes, 1Naïve CD8+ T cell, 3 naïve B cell,1effector memory T cell, 3 adult Nautral killer Cells, 1 memory B cells,1 class switch B cell, 2 central memory T cells, 1 hypermethylated myeloma, and 1 hypomethylated myeloma and finally the differentiation dataset: progenitor B cell (PBC), a pre B cell (preBC),a naïve B cell (NBC),a germinal center B cell (GCBC), a memory B cell (MBC) and a plasma B cell (plBC)

Data processing

Analysis throughout the study were performed in R programming language (www.r-project.org), using core packages from Bioconductor. Custom Awk and bash scripts as well as Bedtools suite **(Quinlan AR, 2010)** were also used to perform the analysis and manage the data.

Searching of PMDs in the methylomes

As commented above, we used the MethylSeekR package **(Burger L, 2013)** to find PMDs in all the samples. Briefly, the distribution of α-values is calculated for one chromosome. α characterizes the distribution of methylation levels in sliding windows containing 101 consecutive CpGs along the genome. The α parameter is derived from modelling methylation by the number of BS-converted and BS-unconverted reads based on a symmetric beta binomial distribution:

$$P(f_i|\alpha) = \frac{1}{B(\alpha,\alpha)} f_i^{\alpha-1}(1-f_i)^{\alpha-1}$$

α-values smaller than 1 reflect a bimodal distribution of methylation, favouring either unmethylated or methylated states. On the other hand, α-values equal or greater than 1 indicate distributions of methylation values that are rather uniform or polarized towards

intermediate methylation levels, as in PMDs. Importantly, we used the shape of the distribution of α-values instead of looking only at the proportion of windows greater than 1. Bimodal or long-tailed distributions of α-values are indicative of PMDs. Once the distribution is calculated, a two-state HMM is trained on the distribution and PMDs are found using the Viterbi algorithm.

Functional characterization of PMDs.

Chromatin segmentations of the GM12878 and H1-hESC cell lines were downloaded from the European Bioinformatics Institute (EMBL) http://www.ebi.ac.uk/ .DHS data for GM12878 was downloaded from the UCSC Genome Browser **(Kent WJ, 2002**) (https://genome-euro.ucsc.edu/) as well as CGI data. LADs data was also downloaded from UCSC (from TIG3 fibroblasts), as well as repetitive element annotation, with RepeatMasker. Uncertain assignments of repetitive elements were removed for the analysis. Finally, RefSeq gene annotation was downloaded also from UCSC.

Gene Ontology enrichment analysis.

We used the Bedtools suite to retrieve genes falling inside and outside of PMDs and those ones overlapping the borders of PMDs. GO enrichment analysis of genes present in each of the three groups were done using GOstats and biomaRt packages from Bioconductor. Genes without GO Terms were removed from the analysis. We used a cutoff of the FDR of 0.001, and a conditional overenrichment analysis was performed for all the three groups.

ACKNOWLEDGEMENTS

I thank to all the members of Simon Heath's lab (Angelika Merkel, Emanuele Raineri, Ron Schuyler, Marc Dabad and Anna Esteve) for their advises, useful discussions and meetings that have aided the development of this work, with especial mention to my co-supervisor Angelika Merkel and my supervisor Simon Heath. I would like also thank to Justin Whalley and Meritxell Oliva from Ivo Gut lab for their useful discussions of the topic.

REFERENCES

Adams, D. et al. (2012). BLUEPRINT to decode the epigenetic signature written in blood. Nat. Biotechnol. 30, 224–226.

Aran D et al. (2010). Replication timing-related and gene body-specific methylation of active human genes. Hum Mol Genet. 15;20(4):670-80.

Berman BP, Weisenberger DJ, Aman JF et al. (2012). Regions of focal DNA hypermethylation and long-range hypomethylation in colorectal cancer coincide with nuclear lamina-associated domains. Nat. Genet. 44(1), 40–46.

Bröske, A.M et al. (2009). DNA methylation protects hematopoietic stem cell multipotency from myeloerythroid restriction. Nat. Genet. 41, 1207–1215

Burger L, Gaidatzis D, Schubeler D, Stadler MB (2013) Identification of active regulatory regions from DNA methylation data. Nucleic acids research 41: e155.

Challen, G. A. et al. (2011). Dnmt3a is essential for hematopoietic stem cell differentiation. Nature Genet. 44, 23–31

Cedar H, Bergman Y. (2012) Linking DNA methylation and histone modification: patterns and paradigms. Nat Rev Genet.10(5):295-304-5.

Chen, T., Ueda, Y., Dodge, J.E., Wang, Z. & Li, E. (2003). Establishment and maintenance of genomic methylation patterns in mouse embryonic stem cells by Dnmt3a and Dnmt3b. Mol. Cell. Biol. 23, 5594–5605.

Cokus SJ, Feng S, Zhang X, Chen Z, Merriman B, Haudenschild CD, Pradhan S, Nelson SF, Pellegrini M, Jacobsen SE. (2008) Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning. Nature 452, 215–219

Deaton AM, Bird A. (2011). CpG islands and the regulation of transcription. Genes Dev. 15;25(10):1010-22

Esteller M. (2008) Epigenetics in cancer. N Engl J Med. 13;358(11):1148-59.

Ernst, J. et al. (2011) Mapping and analysis of chromatin state dynamics in nine human cell types. Nature 473, 43–49.

Gaidatzis D, Burger L, Murr R, Lerch A, Dessus-Babus S, Schübeler D, Stadler MB. (2014) DNA sequence explains seemingly disordered methylation levels in partially methylated domains of Mammalian genomes. PLoS Genet. 2014 Feb 13;10(2):e1004143.

Guelen L, Pagie L, Brasset E, Meuleman W, Faza MB, Talhout W, Eussen BH, de Klein A, Wessels L, de Laat W, van Steensel B. (2008) Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions Nature. June 12;453:948-951.

Hanahan D, Weinberg RA. (2011) Hallmarks of cancer: the next generation. Cell. 4;144(5):646-74.

Hansen KD, Timp W, Bravo HC et al. (2011). Increased methylation variation in epigenetic domains across cancer types. Nat. Genet. 43(8), 768–775

Hansen KD, Sabunciyan S, Langmead B, Nagy N, Curley R, Klein G, Klein E, Salamon D, Feinberg AP. (2014). Large-scale hypomethylated blocks associated with Epstein-Barr virus-induced B-cell immortalization. Genome Res. 24(2):177-84.

Hansen RS et al. (2009) Sequencing newly replicated DNA reveals widespread plasticity in human replication timing. Proc Natl Acad Sci U S A.;107(1):139-44.

Hawkins RD et al. (2010). Distinct epigenomic landscapes of pluripotent and lineage-committed human cells. Cell Stem Cell. 7;6(5):479-91.

Holliday, R. & Pugh, J. E. (1975). DNA modification mechanisms and gene activity during development. Science 187, 226–232

Hon GC, Hawkins RD, Caballero OL et al. (2012). Global DNA hypomethylation coupled to repressive chromatin domain formation and gene silencing in breast cancer. Genome Res. 22(2), 246–258 (2012).

Hovestadt V et al. (2014). Decoding the regulatory landscape of medulloblastoma using DNA methylation sequencing. Nature 26;510(7506):537-41.

Illingworth RS, Bird AP (2009). CpG islands-'A rough guide'. FEBS Lett. 5;583(11):1713-20.

Irizarry RA, Ladd-Acosta C, Wen B et al. (2009).The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores. Nat. Genet.41(2), 178–186

Ji H, et al (2010) Comprehensive methylome map of lineage commitment from haematopoietic progenitors. Nature. 16;467(7313):338-42.

Jones PA. (2012) Functions of DNA methylation: islands, start sites, gene bodies and beyond. Nat Rev Genet. 2012 May 29;13(7):484-92.

Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. (2002) The human genome browser at UCSC. Genome Res. (6):996-1006.

Kulis M, et al. (2012) Epigenomic analysis detects widespread gene-body DNA hypomethylation in chronic lymphocytic leukemia. Nature Genetics. 44(11):1236-42.

Li, E., Bestor, T. H. & Jaenisch, R. (1992).Targeted mutation of the DNA methyltransferase gene results in embryonic lethality. Cell 69, 915–926

Lister R, Pelizzola M, Dowen RH, Hawkins RD, Hon G, et al. (2009) Human DNA methylomes at base resolution show widespread epigenomic differences.
Nature 462: 315–322.

Lister R, Pelizzola M, Kida YS, Hawkins RD, Nery JR, et al. (2011) Hotspots of aberrant epigenomic reprogramming in human induced pluripotent stem cells. Nature 471: 68–73.

Marzese DM et al. (2014). Epigenome-wide DNA methylation landscape of melanoma progression to brain metastasis reveals aberrations on homeobox D cluster associated with prognosis. Hum Mol Genet. 1;23(1):226-38

Okano, M., Xie, S. & Li, E. (1998). Cloning and characterization of a family of novel mammalian DNA (cytosine-5) methyltransferases. Nat. Genet. 19, 219–220

Okano, M., Bell, D. W., Haber, D. A. & Li, E. (1999). DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development. Cell 99, 247–257.

Pastor WA et al (2011) Genome-wide mapping of 5-hydroxymethylcytosine in embryonic stem cells. Nature. 19;473(7347):394-7.

Quinlan AR and Hall IM,(2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 26, 6, pp. 841–842.

Riggs, A. D. (1975).X inactivation, differentiation, and DNA methylation. Cytogenet. Cell Genet. 14, 9–25

Rountree MR, Selker EU. (1997) DNA methylation inhibits elongation but not initiation of transcription in Neurospora crassa. Genes Dev. 15;11(18):2383-95. 8.

Stadler MB, Murr R, Burger L et al. (2011) DNA-binding factors shape the mouse methylome at distal regulatory regions.Nature 480(7378), 490–495

Schroeder DI, Lott P, Korf I, LaSalle JM. (2011). Large-scale methylation domains mark a functional subset of neuronally expressed genes. Genome Res. 21(10), 1583–1591

Schroeder DI, Blair JD, Lott P et al. (2013). The human placenta methylome. Proc. Natl Acad. Sci. USA 110(15), 6037–6042.

Shen, H. & Laird, P. W (2013.)Interplay between the cancer genome and epigenome. Cell 153, 38–55.

Trowbridge, J.J., Snow, J.W., Kim, J. & Orkin, S.H. (2009). DNA methyltransferase 1 is essential for and uniquely regulates hematopoietic stem and progenitor cells. Cell Stem Cell 5, 442–449.

Thurman RE et al. (2012). The accessible chromatin landscape of the human genome.Nature. 6;489(7414):75-82.

Wen, B., Wu, H., Shinkai, Y., Irizarry, R.A. & Feinberg, A.P. (2009). Large histone H3 lysine 9 dimethylated chromatin blocks distinguish differentiated from embryonic stem cells. Nat. Genet. 41, 246–250

Xie W et al. (2013) Epigenomic analysis of multilineage differentiation of human embryonic stem cells Cell. 23;153(5):1134-48