PhD degree in Molecular Medicine

(Curriculum in Computational Biology)

European School of Molecular Medicine (SEMM)

University of Milan and Istituto Italiano di Tecnologia (IIT)

Settore disciplinare: Med/04

Development of Computational Tools to Study the Patterning of DNA and RNA Methylation in Healthy and Disease States

Kamal Kishore

Istituto Italiano di Tecnologia (IIT)

Matricola n. R09857

Supervisor: **Dr. Mattia Pelizzola**

Istituto Italiano di Tecnologia

Dr. Bruno Amati

Istituto Italiano di Tecnologia

Acknowledgement

The doctoral dissertation work presented in this thesis was made possible due to the contribution of many people, to all of whom I am indebted. My first vote of thanks goes to my major advisor, Dr. Mattia Pelizzola for providing me an opportunity to pursue PhD degree under his supervision at Istituto Italiano di Tecnologia, Milan. It has been an honor to be his first PhD student. Throughout the course of this research work, he has been immensely helpful, supportive and has given me great encouragement to pursue my research interests independently. His advice, support, guidance and enthusiasm towards my dissertation have been invaluable. He has always encouraged me to perform and present, and boosted my self -confidence over the years. It has been an honor to work with him and I would certainly cherish this experience for the rest of my life.

I am also grateful to my thesis committee members; Dr. Dirk Schübeler and Dr. Mark D. Robinson who have encouraged, guided and helped me meet several challenges. I am also indebted to SEMM for running a very competitive program, especially with regards to short courses and seminars that were critical to building a strong foundation of my research area. I have been lucky to receive prudent advice and support from them at several instances during my academic career.

I owe my deep regards to my supervisor Dr. Bruno Amati for providing me guidance during my PhD tenure. I appreciate all his contributions of time, ideas, and critical advice to make my Ph.D. experience productive and stimulating. His sincere suggestions helped me greatly in bringing out this work in its present shape. In

addition, I am thankful to my department colleagues for their support and thoughtful discussions.

Finally, I would like to give my deepest gratitude to my parents and sister Anupriya for their support, encouragement, patience and unwavering love. Their dedication and support throughout my life has provided the foundation for this work. Lastly but most importantly, I thank my wife Neha for her understanding, love and delicious food during the last year. Her unrelenting support and encouragement was critical in making this dissertation possible. Thank you for believing in me.

Table of Contents

Figures Index

Abstract

Epigenetics can be defined as the set of sequence independent processes that produces heritable changes in cellular information. These chromatin-based events such as covalent modification of DNA and histone tails are laid down by the co-ordinated action of chromatin modifying enzymes, thus altering the organisation of chromatin and its accessibility to the transcriptional machinery. Our understanding of epigenetic intricacies has considerably increased over the last decade owing to rapid development of genomic and proteomic technologies. This has resulted in huge surge in the generation of epigenomics data. Integrative analysis of these epigenomics datasets provides holistic view on the interplay of various epigenetic components and possible aberration in patterns in specific biological or disease states. Although, there are numerous computational tools available catering individually to each epigenomic data-type, a comprehensive computational framework for integrated exploratory analysis of these datasets was missing. We developed a suite of R packages methylPipe and compEpiTools that can efficiently handle whole genome base-resolution DNA methylation datasets and effortlessly integrate them with other epigenomics data. We applied these methods to the study of epigenomics landscape in B-cell lymphoma identifying a putative set of tumor suppressor genes. Moreover, we also applied these methods to explore possible associations between m6A RNA methylation, epigenetic marks and regulatory proteins.

Introduction

Biologists have been puzzled for decades to know how a single genome contained in a single cell fertilized egg can give rise to hundreds of different cell types found in an embryo or adult. Further, how it creates from the same genetic book of instructions different functional outcomes and phenotype for the cells. This outcome is attributed to the additional information on top of that genetic one, which allows the formation of each individual cell type, despite existence of the same genetic code within them. 'Epigenetics' a term coined by Edward Waddington[1] in 1942 is used to describe this additional layer of information on top of the genetic information. It is the study of mitotically heritable changes in gene expression that occur without changes to the DNA sequence. This epigenetic information allows the development and differentiation of the hundreds of different cell types, all from the same set of instructions[2, 3]. The genes that are expressed within each cell define its identity. Each cell type can actually express only a restricted subset of genes. In mammalian genome nowhere near 25000 (total) genes are expressed in each cell type, at each time. Each cell type expresses a combination of a few thousand genes out of the all the possible permutations and this particular set of expressed genes within each cell type enables the phenotype and the function of each cell.

Hence the question arises, if the genes that are expressed define each cell type, how can a different set of genes be expressed within each cell? How does the cell realize which genes to express? This is made possible by a combination of two things. Firstly, it is due to the activity of transcription factors that are specific for each cell lineage[4]. These proteins with sequence specificity in their binding to DNA, bind to the promoters of genes to activate or repress the expression of specific targets. But these lineage

specific transcription factors do not act alone. They need to work within the context of broader information that comes in the form of epigenetic marks within the genome.

Epigenetic marks provide structure to the chromosome by marking the beginning, center and the end of the whole chromosome. They alter how the gene information is read; so there are marks that are associated with both active (expressed) and inactive (silent) genes. So the reason for each cell type to have different sets of expressed genes is partly because of the epigenetic marks that are found on the genes that are expressed, or not being expressed. Hence, the epigenetic control allows or permits the differential expression of genes within each cellular lineage, and therefore permits the development and differentiation of hundreds of cell types from the single genome.

1.1 Epigenetics in development

Epigenetic control plays an important role throughout the development of an organism[5]. The diagram [**Figure 1**] depicts the epigenetic control during development from the single cell fertilized egg (zygote), pre-implantation to post-implantation development. In each stage of this life cycle different genomic regions are marked by particular epigenetic marks according to their particular functions. These epigenetic marks laid throughout development ensure accuracy and maintenance of the cellular identity. However, these marks are naturally removed globally two times during the life cycle of mammals. The first stage is the pre-implantation stage when the paternal (sperm) and maternal (oocyte) epigenomes are reset to form zygote. During this phase, DNA methylation and histone marks are removed throughout the genome. The

paternal genome is actively demethylated whereas the maternal genome is passively de-methylated (epigenetic marks are lost upon replication)[5]. This reaches a low at the blastocyst stage, just before the implantation. Thereafter, epigenetic marks are re-established in lineage specific manner. Concurrently, individual lineage specific processes undergo active re-modelling of these epigenetic marks. The second global remodeling of epigenetic marks happens during primordial germ cells development. These cells primed for somatic fate must be epigenetically remodeled to ensure totipotency in the next generation[6]. In the midst of this epigenetic resetting, parts of the genome particularly repeat, transposons and imprinted loci (during pre-implantation) remain unaltered[5].

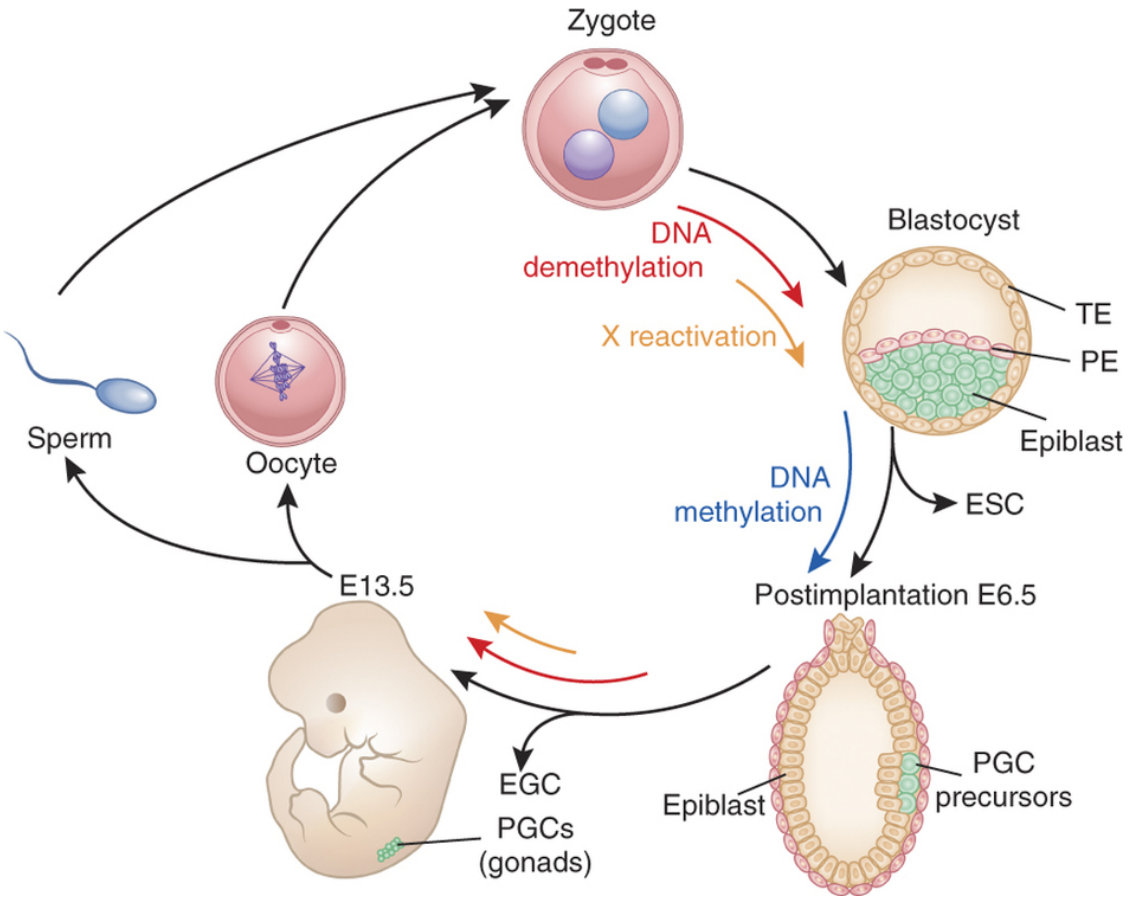


Figure 1: Epigenetic programming and reprogramming during the mouse life cycle.

Cantone & Fisher, Nature Structural & Molecular Biology 20, 282–289 (2013)

This prevents activation of retrotranposons and maintains chromosomal stability to ensure proper development. When this ubiquitously important process of epigenetic control goes wrong, it results in a wide array of different disorders like imprinting disorders, embryonic lethality and several types of tumors[7].

1.2 Chromosome structure

Chromatin inside the nucleus consists of DNA wrapped around histone proteins. This packaging compresses roughly two meters of DNA within the nucleus down to a ten-micrometer diameter. As DNA is accessed several times, this process must result in an ordered structure. The accessibility to the DNA is required during transcription as well as during DNA replication and repair mechanisms. Thus, how tightly the chromatin is packed relates to the need for accessibility. This further determines the accessibility to transcription factors and the RNA Polymerase. If the DNA is tightly packed around the histone proteins then it makes it inaccessible to the RNA Polymerase and the transcription factors. Consequently, it's less likely to be active or expressed and vice-versa.

The smallest unit of the DNA combined with histones is known as the nucleosome. It consists of a DNA wrapped around a histone octamer [**Figure 2**]. These 8 units are made up of 2 units each of histones H2A, H2B, H3, and H4. Together they make up the 8 histones molecules acting like a spool for the DNA. On the outside histone H1 locks together that spool and holds the DNA in place. The reason that this interaction can occur between the histones and the DNA is because histones are rich in positively charged amino acids, particularly lysine and arginine. Positively charged histones and negatively charged DNA interact via electrostatic interaction. The N-terminal tails of the histones tend to protrude out of the nucleosome and these protrusions are important for the modification of histones.

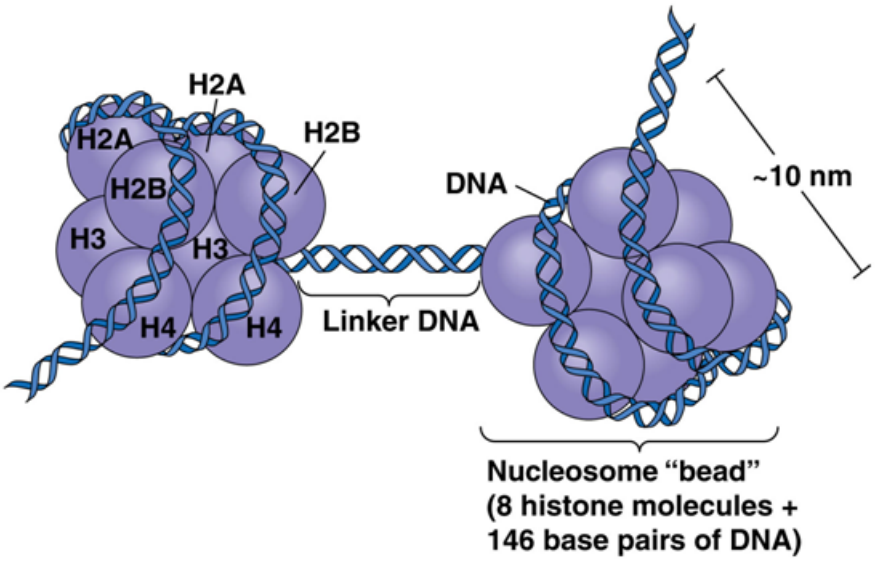


Figure 2: Histone proteins.

Adapted from Spreadscienceblog

The nucleosome has a bead-in-a-string structure. It can be further condensed into a higher order chromatin packaging into a 30 nanometer fiber via interaction between histone H1 molecules. If scaffolding proteins are added on top of this 30-nanometer fiber, it becomes an interphase chromosome. Subsequently, it is compacted further down at metaphase by adding a final layer of scaffolding proteins [**Figure 3**]. There are different forms of chromatin specially open and closed chromatin. Closed chromatin is known as heterochromatin whereas open chromatin is called euchromatin. Heterochromatin is further classified as facultative heterochromatin and constitutive heterochromatin. Facultative heterochromatin can differ between cell type thereby constituting tissue specific things that can be expressed in one cell type and not another.

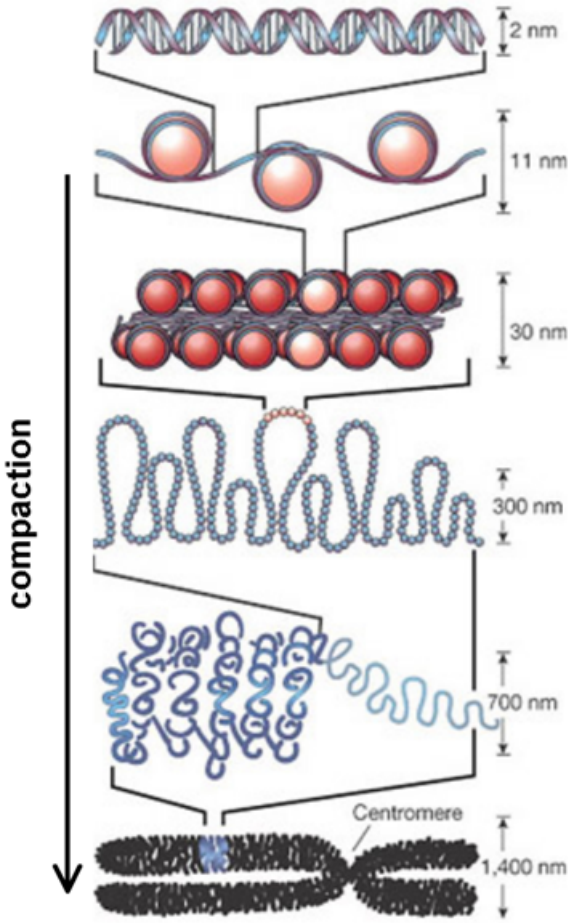


Figure 3: Chromosome compaction.

(https://beyondthedish.wordpress.com/tag/histone-proteins/

On the contrary, constitutive heterochromatin is stable in every cell and performs a structural role (telomeres and centromeres). Epigenetic marks at the beginning, end and even at center of the chromosomes is constitutive heterochromatin. This helps to maintain the structural integrity of the genome. Each of these different types of chromatin has a distinct set of epigenetic marks. The following section explains specific epigenetic modifications and their functional consequence.

1.3 Histone modifications

Histone proteins are one of the most important components of the epigenetics. These proteins are embellished by a number of post-translational histone modifications (called histone marks) such as acetylation, methylation, phosphorylation, SUMOylation and many more[8, 9]. These modifications tend to occur almost exclusively on the N-terminal tails of the histones that protrude out from the nucleosome. There are a plethora of different places where these histone modifications can occur with a variety of different residue[10] [**Figure 4**].

Each histone tail modification is associated with a different function. Histone methylation and acetylation are the best characterized of these histone modifications. Histone methylation, that occurs as mono, di or tri (number of added methyl group) at lysine or arginine residue tends to be associated with transcription[11]. Histone acetylation on the other hand is universally associated with actively transcribed gene[12]. Histones are acetylated by the histone acetyltransferases (HATs)[13] and these marks are removed by histone deacetylases (HDACs)[14]. Further, histone methylation can be associated either with gene activity or gene inactivity depending on the context. While H3K4me3 methylation marks active promoters, H3K9 methylation is mostly associated with constitutive heterochromatin[15]. H3K27 methylation is also an inactive mark spread over the entire gene but associated with facultative heterochromatin[15]. H3K27methylation is carried out by EZH2 that is part of polycomb repressive complex 2 (PRC2)[16]. Since this is found in facultative heterochromatin EZH2 and PRC2 complex plays an important role in tissue specific gene silencing[17].

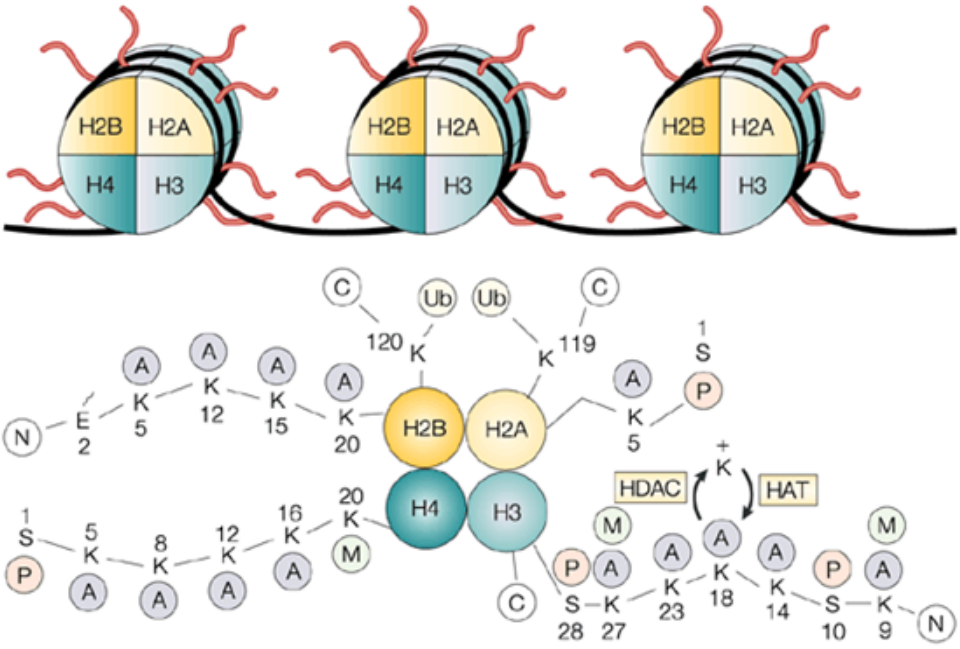


Figure 4: Histone modifications.

Marks et al, Nature Reviews Cancer 1, 194-202 (December 2001)

Histone modifications are recognized by specific effector proteins including chromatin regulators (CRs) to cause the desired downstream changes. This binding by different proteins under different contexts results in diverse functional outcomes. For example H3K4 methylation, H3K9 methylation and H3K27 methylation all are bound by a chromodomain. But each has specificity for particular residues: CHD1 binds H3K4 (me), HP1 binds H3K9 (me) and CBX2 binds H3K27 (me). CHD1 is an ATP dependent chromatin remodeler that can open or condense the nucleosomes[18]. On the other hand, HP1 can bring in DNA methyltransferase (DNMT1) that methylates the neighbouring CpG di-nucleotides[19] or recruit histone methyl-transferases to enable spreading of the H3K9 methylation to other neighbouring nucleosomes[20]. Further,

CBX2 is part of another polycomb repressive complex, PRC1[21]. To summarize, euchromatin is associated with high levels of acetylation and tri-methylated H3K4 or H3K36[22]. On the other hand, heterochromatin is associated with low levels of acetylation and high levels of H3K9 or H3K27 methylation[22]. However, an interesting case has emerged of co-existence of H4K4me3 and H3K27me3 (bivalent domains) that mark the developmentally important genes in embryonic stem cells[23]. These domains have crucial implications in maintaining totipotency in stem cells and are lost further upon cell differentiation.

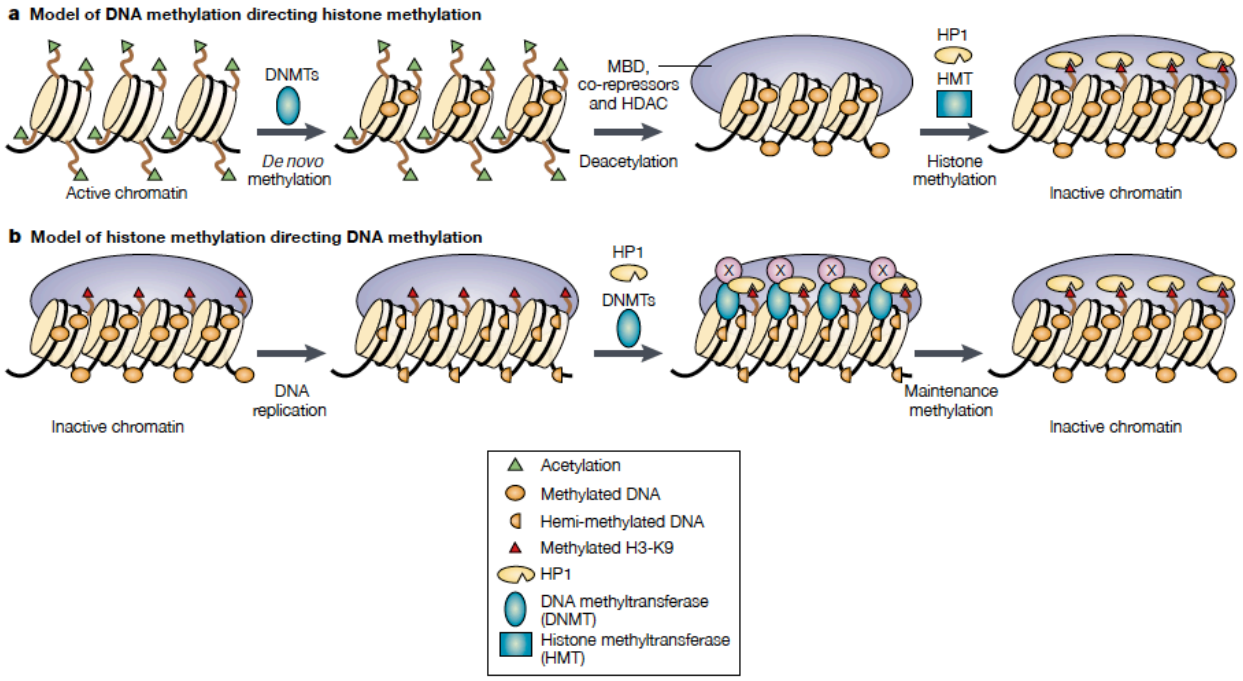


Figure 5: Links between DNA methylation and histone modification.

Adapted from En li, Nature Reviews Genetics 3, 662-673 (September 2002)

Histone marks cross talk with other epigenetic marks such as DNA methylation has been well studied [**Figure 5**]. The studies show the interplay of DNMT3L and H3K4 during early development[24]. According to this, DNMT3L interacts with histone H3

tails to recruit DNMT3A and cause de novo DNA methylation[24, 25]. However, this interaction is strongly inhibited at regions marked by H3K4me3 (active mark). Moreover, histone methyltransferase G9a (which catalyzes H3K9 methylation) can recruit DNMTs to perform targeted de-novo methylation to lock the silenced state established by the repressive histone marks[26, 27]. On the other hand, the instance of DNA methylation directed H3K9 di-methylation via recruitment of MeCP2 shows the bi-directionality of this cross talk[28–30]. Moreover, Strahl and Allis proposed "the histone code" hypothesis postulating that these multiple histone modifications act in a sequential or combinatorial manner to bring about distinct downstream changes[31]. Although experimental support for this hypothesis is lacking, many studies observing the recognition of multiple chromatin modification simultaneously by chromatin regulators supports this idea[32, 33]. Computational methods such as ChromaSig[34] and ChromHMM[35] have been developed to identify combinatorial association of these histone modifications. These methods could classify chromatin states according to the functional genomic constituents. The advancement of the high throughput sequencing technologies (leading to wealth of data) and computational methods has significantly increased our understanding of the combinatorial complexity of histone modifications.

1.4 DNA methylation

1.4.1 Laying and erasing DNA methylation

The covalent modification of DNA (DNA methylation) is one of the most extensively studied epigenetic marks in eukaryotes. It involves the covalent modification of cytosine bases of genomic DNA to 5-methylcytosine upon addition of a methyl group, catalysed by DNA methyl-transferase (DNMT) enzymes. The enzymes DNMT3A and DNMT3B lay down the methylation in a de novo fashion during development[36]. During cell division when the DNA is replicated, another DNA methyltransferase enzyme (DNMT1) specifically recognizes hemi-methylated DNA[37]. DNMT1 then lays down methylation on the un-methylated strand to restore fully methylated CpG dinucleotide. Hence, DNA methylation can be a stable epigenetic mark because at every cell division, this DNA methylation will be copied by DNMT1 onto a new daughter strand of DNA. A revised model of maintenance of DNA methylation proposes the participation of DNAMT3A and DNMT3B in addition to DNMT1 in this phenomenon[38]. According to this, the majority of DNA methylation during replication is copied by DNMT1 whereas DNMT3A and DNMT3B complete this process after replication. All the three methyltransferases are important for mammalian development and are required for maintaining viability of somatic[39] and cancer cells[40].

Although DNA methylation is a stable modification, its loss by either active or passive mechanisms has been observed in several biological contexts [**Figure 6**]. It happens in early development, in primordial germ cell development and during

differentiation[5]. In the passive demethylation, the DNA is replicated but fail to maintain the methylation in the absence of DNA methylation maintenance machinery[41]. Thus replication passively dilutes the DNA methylation. In contrast, active DNA demethylation is the enzymatic removal of the methyl group from 5mC or its modification. This enzymatic removal of the methyl group can happen in different ways via several chemical intermediates. The recent discovery of ten-eleven translocation (TET) family of oxidases has greatly advanced our understanding of the active DNA demethylation process[42]. These proteins catalyse oxidation of 5-methylcytosine (5mC) to the intermediate modified base 5-hydroxymethylcytosine (5-hmC)[43]. TET proteins further oxidize this intermediate to form 5-formylcytosine (5fC) and 5-carboxylcytosine (5caC)[44]. Active regulatory regions and actively transcribed genes show increased level of hydroxymethylation[45, 46] indicating its role in regulating enhancer activity in stem cells[47].

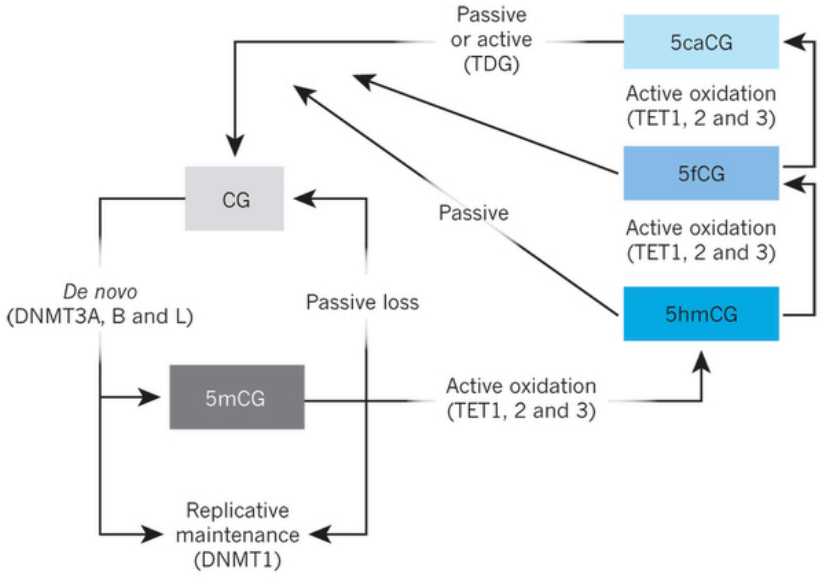


Figure 6: Active and passive demethylation.

Adapted from Schubeler, Nature 517, 321-326 (2015)

1.4.2 Functional role and genomic context

DNA methylation has bimodal pattern with almost all CpG di-nucleotides being methylated except those located in CpG islands (CGIs). These CpG islands tend to be protected from methylation or only dynamically methylated in germ cells[48]. The vastly studied inactivation by DNA methylation is X chromosome inactivation in mammals that involves DNA methylation of the CpG islands[49]. There are a couple of mechanisms through which DNA methylation suppresses gene expression. The primary mechanism involves binding of methylated CpG by proteins known as MeCP1 and MeCP2 [50]. These proteins possess a transcriptional repression domain or alternatively, they can bring their own partner protein and condense the chromatin. The secondary mechanism, although not common, is that the methylated CpG itself will prevent transcription factor binding[51]. This secondary mechanism seems to be true for the transcription factors SP1[52], CTCF[53] and others.

DNA methylation is also found at other regions of the genome and it has different functions according to different genomic context [**Figure 7**]. DNA methylation at promoters is mostly associated with transcriptional repression. However, "transcriptional repression driven by promoter methylation is inversely related to the distance of mCpGs from the TSS (transcription start site) and increases with promoter CpG content"[54]. CG rich promoters are mainly repressed by Polycomb proteins, which mark them with H3K27me3 (repressive mark)[55]. Nonetheless, instances of their repression via DNA methylation do exist. It is usually associated with long-term repression as evident in case of imprinted genes on inactive X chromosome[56, 57].

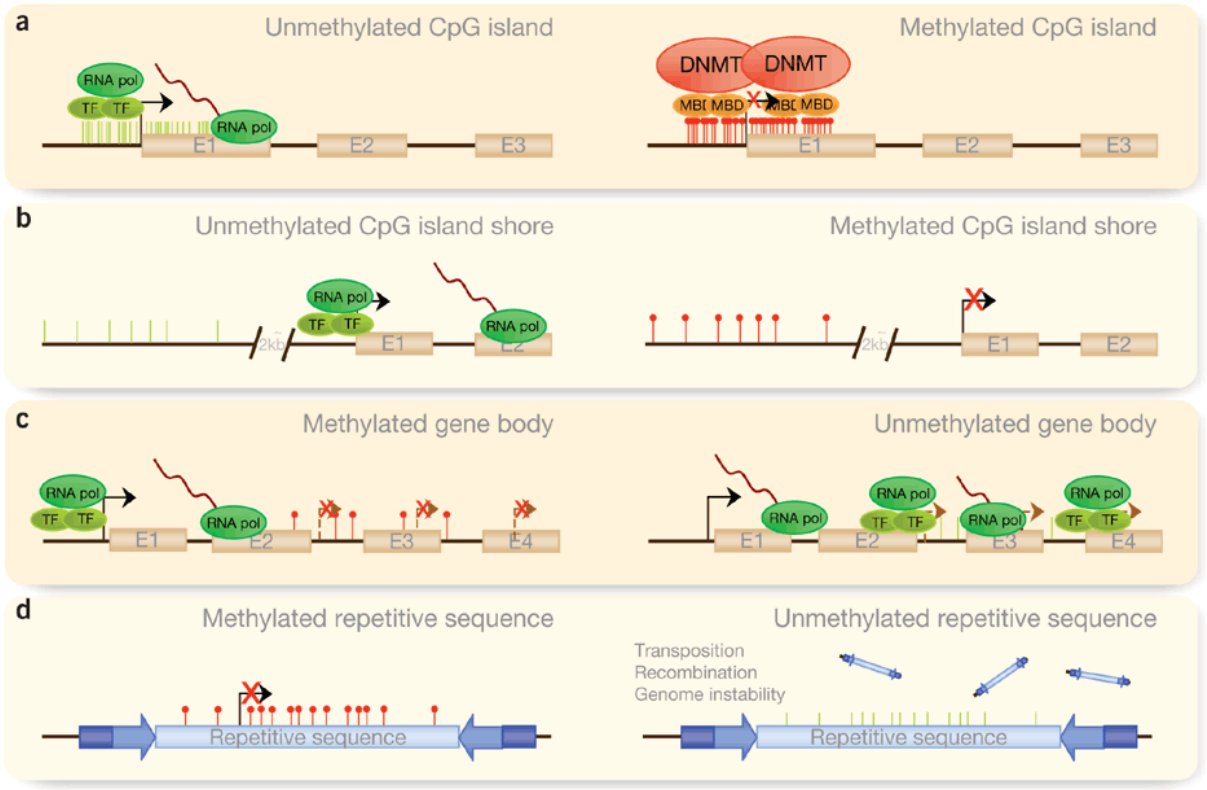


Figure 7: DNA methylation patterns.

Portela & Esteller, Nature Biotechnology 28, 1057–1068 (2010)

On the other hand, low CG content promoters are associated with dynamic DNA methylation[58]. It mostly confers tissue specificity by regulating the expression of genes required for particular lineage[59]. In contrast to promoters, gene bodies are CpG poor and mostly methylated. A recent genome wide study has shown that gene body methylation positively correlates with gene expression and is better determinant than promoter methylation[60]. DNA methylation is also notably found in inter-genic intervals and their repetitive elements[61] that are methylated to maintain genome integrity[62, 63]. In the absence of DNA methylation in these regions as shown by the cells lacking DNMT1, cells exhibit genomic instability[64, 65]. DNA methylation has a similar role at transposable elements of preventing them from autonomous

transposition within genome (which is clearly mutagenic)[66]. Subsequently, methylation of these repeats can also prevent illegitimate recombination[67]. The silencing of the repeats forms the basis of the genome defence model proposed by Tim Bestor[68]. He proposed that the primary function of DNA methylation is to protect the genome from these transposable elements and this is certainly predominantly where DNA methylation is found.

Cell differentiation is the process of forming specialized cells from pluripotent embryonic stem cells. This occurs by regulation of gene expression via selective activation of lineage specific genes and silencing of pluripotency genes. DNA methylation plays a key role in it by regulating the expression of important transcriptional factors OCT-4 and NANOG during early embryogenesis[69]. These factors selectively activate genes responsible for pluripotency while repressing genes required for cell differentiation. OCT-4 gene, necessary for pluripotency is expressed in embryonic stem cells but not in trophoblast stem cells. DNA methylation analysis revealed that the OCT-4 enhancer/promoter region is hypomethylated in ES cells but hypermethylated in TS cells[70]. In addition, de-novo DNA methyltransferase are also required for cellular differentiation by systematically regulating gene expression via promoter methylation. DNMTs show high expression in undifferentiated state while reduced expression in differentiated state[71]. Moreover, genes involved in pluripotency and differentiation undergo aberrant DNA methylation[72]. Extensive methylation changes were also found in regulatory regions outside of the core promoters during cellular differentiation[73]. The cytosine methylation also occurs in the non-CpG context at CHG and CHH sites (where H is A, C or T) in pluripotent cells where nearly one-quarter of all methylation occurs at non-CpG sites, especially in gene bodies[72] and in brain, where it is even more frequent than CpG methylation[74]. Non-

CpG methylation is lost with differentiation and is restored in induced pluripotent stem cells[72]. This suggests an important role of non-CpG methylation in the origin and maintenance of pluripotent state. In conclusion, the maintenance of DNA methylation pattern plays an important role in cellular development by modulating gene expression during entire lifetime of an organism.

1.4.3 DNA methylation in cancer

Cancer is not just a genetic disease, but rather it is genetic abnormalities partnered with epigenetic abnormalities. Epigenetic aberrations in cancer include locus-specific hypermethylation of CpG islands (CGIs) and genome-wide hypomethylation[75]. The genome in general is methylated in their repetitive elements through the intergenic regions and even in the introns of genes. In contrast, in cancer cell CpG islands are more likely (not all) to be methylated than in a normal cell[76–78]. The rest of the genome including the repetitive elements in these intergenic regions, and the introns are hypomethylated[79–81]. Hypermethylated CpG islands are found in the promoters of tumor suppressor genes thereby inactivating them. This acts as one of the hits of Knudson's "two-hit" hypothesis[82]. This phenomenon occurs frequently in a number of different tumor type studied till date[83–85]. Genome-wide studies have also shown that the identity of hyper-methylated CGIs varies by tumor types/sub-type[86] and there is an increase in methylation level with cancer progression[87, 88].

In addition to de-novo hypermethylation, cancer cells are also characterized by genome wide hypomethylation outside of CpG islands, which are generally methylated in normal cells. This decrease is mostly attributed to loss of methylation at repetitive

elements and retro-transposons. This results in illegitimate deletions, insertions or reciprocal translocations, and even duplication of chromosomes or loss of chromosomes; in general genomic instability. In ICF (Immunodeficiency, Centromere instability and Facial) syndrome, DNMT3B is mutated determining hypomethylation at centromeric repeats and leading to genomic instability[89]. Moreover, a number of studies have linked the absence of DNMT1 to genomic instability in cancer[64, 90]. Hypomethylation also occurs in other regions of the genome (apart from repeats) but it is less frequent. One such example is hypomethylation of CpG poor promoters associated with activation of oncogene RAS in gastric cancer[91].

DNA methylation in cancer has both hypomethylation genome-wide and hypermethylation of the tumor suppressor genes. But the driving role of DNA methylation is context dependent and different tumors have different dependencies on each of these particular aberrant methylation profiles. Some tumors are driven by tumor suppressor hypermethylation that enables them to continue to dividing rapidly[92]. While some types of cancers are driven by chromosomal instability due to hypomethylation[93]. Further several others change their dependencies throughout the lifespan of tumor [79, 94].

1.4.4 Technologies for measurement

The experimental methodologies to profile DNA methylation have been revolutionized over the last decade. Profiling DNA methylation was previously restricted to one particular loci, while it can now be performed over entire genomes at single base resolution. These technologies vary according to sample throughput and genome coverage. The methods for profiling DNA methylation can broadly be classified into

three categories: I) digestion with methylation-sensitive restriction enzymes to fractionate methylated/un-methylated fragments[95], II) enrichment of methylated DNA fragments by methyl-cytosine binding proteins[96] or antibodies[97] and III) bisulfite treatment of DNA which converts un-methylated cytosines leaving methylated one unaffected[73]. Subsequently, the DNA methylation patterns information extracted through these methods is coupled with high throughput sequencing.

Endonuclease digestion

DNA methyltransferases as part of restriction/modification systems protect host DNA from cleavage by methylating bases in the recognition site of the restriction endonucleases. The methylation sensitivity of these restriction enzymes is used to elucidate read-out of DNA methylation from the patterns of their cutting. HpaII and SmaI are the most widely used restriction enzymes for DNA methylation studies[98]. The common methodology to discern differences in methylation is by detecting differences in the pattern of restriction fragments generated by digestion with a methylation-sensitive restriction enzyme separated by two-dimensional gel electrophoresis. Over the years different techniques have been developed that couple enzymatic methods to array-based analysis. The most optimized workflow in this category is comprehensive high-throughput arrays for relative methylation (CHARM) [99]. Further, restriction enzyme enrichment techniques are adapted to obtain methylation read-out by next-generation sequencing techniques. Methyl–seq is the most common method in which sequencing-by-synthesis of libraries constructed from size-fractionated HpaII or MspI digests are compared with randomly sheared fragments[100]. Sequence-based analysis provides more genome coverage with less

input DNA, avoids hybridization artefacts and also allows allele-specific DNA methylation analysis.

Affinity enrichment

In affinity enrichment of methylated regions specific antibodies for 5mC or methyl-binding proteins with affinity for methylated native genomic DNA are used for profiling DNA methylation. In these methods methylated regions are enriched by immunoprecipitation of genomic DNA with an antibody specific for methylated cytosine, followed by hybridization to either a tiling array or microarray referred to as MeDIP[97, 101]. These techniques have been extensively used in studying the plant[102], mouse[59] and human methylomes[97, 99, 101]. In addition, other approaches uses antibody specific for higher affinity methyl-binding proteins (MBD) domains[103]. These affinity-enrichment methods are mostly combined with array hybridization in which the input DNA and enriched DNA are labelled with different fluorescent dyes. These methods allows for rapid and efficient assessment of genome-wide DNA methylation. However, it does not yield information on individual CpG and requires adjustments for varying CpG density of different genomic regions.

Bisulphite conversion

Bisulphite conversion based methods revolutionized DNA methylation analysis. The denatured genomic DNA is treated with sodium bisulphite that chemically deaminates unmethylated cytosine residues while leaving methylated cytosines intact[104]. This chemical treatment of DNA converts the unmethylated Cs into Ts (by uracil) thereby

enabling DNA methylation detection at base-pair resolution[105]. Further, it is comprised of three different bases instead of four resulting in reduced sequence complexity. This makes their adaption to array-hybridization techniques difficult and requires dedicated arrays based on the bisulphite-converted genome. Illumina adapted its 'BeadChip' technology on Infinium I platform to generate a comprehensive genome wide profiling of human DNA methylation known as Illumina 27k methylation array[106, 107]. Infinium I has two beads per probe, one in the red channel and one in the green channel. The technology is further extended to 450k array that include an additional second assay type, Infinium II[108]. The 450k array contains 485,512 probes covering about 99% of RefSeq genes[108]. In addition to array- based methods, bisulphite-converted DNA is also combined with sequencing-based approaches. Reduced representation bisulfite sequencing (RRBS) is an efficient high-throughput technique to analyze the genome-wide methylation profiles on a single nucleotide level[109]. It combines restriction enzymes and bisulfite sequencing to enrich for high CpG content genomic regions. The size-fractionated DNA fragments are selected after BglII digestion or MspI digestion. These fragments include the majority of promoters and CpG rich regions but do not target specific regions of interest in the genome. The most comprehensive single-base-pair resolution DNA methylation analysis technique is whole-genome bisulphite sequencing. This has been utilized for profiling small eukaryotic genomes, such as A. thaliana[110], and for mammalian DNA[72].

1.4.5 Computational Tools for Epigenomics data analysis

To accelerate the research in the area of epigenetics, many international consortia have been formed during the last decade which explore its various aspects including DNA methylation, histone marks, chromatin accessibility and binding of regulatory proteins[111]. Integrative analysis of these epigenomics datasets provides holistic view on the interplay of various epigenomics component during biological processes and possible aberration in patterns in specific biological states. As a consequence, a multitude of data types, generated by different experimental methods is often combined in the same study. Numerous computational tools have been developed for the analysis of epigenomics data, typically focusing on specific analysis steps and data types[112, 113]. However, it remains difficult to understand the relative merits and performance of all the available approaches. Regarding DNA methylation, a comparative study has been performed on methods for identifying differentially methylated regions (DMRs) discussing the importance of experimental design along with confounding factors such as batch effects and cell type composition[114]. Many early WGBS studies used Fisher's Exact test to identify DMRs[72]. However, recently beta-binomial is the most preferred statistical method implemented in several recent packages, such as BiSeq[115], DSS[116], RADMeth[117] and methylSig[118]. For other epigenomics data, Galaxy and Bioconductor resources presents a series of functionality critical for dealing with the complexity of their analysis [119, 120]. While the former is very intuitive to use, it is dependent on a limited set of embedded tools. On the other hand, Bioconductor currently offers more than 900 packages for the analysis of high-throughput data, but it requires greater experience for the identification and use of the available resources.

1.5 Objectives

1.5.1 Development of Computational tools for the integrative analysis of epigenomics data

Our understanding of epigenetic intricacies has considerably increased over the last decade owing to rapid development of genomic and proteomic technologies. The coupling of chromatin immuno-precipitation with next-generation sequencing (NGS) platforms (ChIP-Seq) has presented us with a comprehensive view of the epigenome. Earlier studies on the role of epigenetics in cancer were limited primarily to gene expression and DNA methylation, while recently more comprehensive epigenomics maps have shed light on the interplay between DNA methylation and histone modifications, and the subsequent impact on transcriptional programs[72, 74, 121]. Moreover, the results from International Cancer Genome Consortium about recurring somatic mutation in various cancer types have highlighted the presence of driver mutations associated to key epigenetic players[122]. These studies highlight the role of integrative analysis for deeper understanding of epigenomics regulatory mechanisms.

Numerous computational tools have been developed for the analysis of epigenomics data but it requires greater experience for the identification and use of the available resources. Eventually, while experienced bioinformaticians are able to combine different tools and accommodate various input format requirements, simple and comprehensive tools for an integrative analysis of these various data types are missing. This leaves biologists generating high-throughput sequencing data to depend on help from other computational scientists. Hence, we intended to develop specific

methods that can efficiently handle whole genome base-resolution DNA methylation datasets and perform integrative analysis with other epigenomics components.

To this purpose, we developed two companion packages for the integrative analysis of the most common epigenomics data types, providing easy access to what in our knowledge are the features most commonly requested by biologist, and facilitating the execution of routine tasks for bioinformaticians. Briefly, methylPipe supports the analysis of methyl- and hydroxymethyl-cytosines (mC and hmC, respectively) in both the CpG and non-CpG sequence contexts, derived from any methodology providing base- or low-resolution data. compEpiTools is a companion R package for the analysis, integration and simultaneous visualization of various epigenomics data types across multiple genomic regions in multiple samples.

1.5.2 Studying epigenomics landscape of B-cell lymphoma

Myc is a transcription factor and proto-oncogene that can activate or repress gene expression of target genes via dimerization with the partner protein Max and additional cofactors. *Myc* activation leads to induction of the tumor suppressor ARF, stabilization of p53, and subsequently induction of apoptosis[123]. Hence subverting the apoptotic response during tumor progression sets the selective pressure to mutate ARF or p53 in *Myc*-induced lymphoma[124]. However, considerable fraction of the lymphomas arising in Eμ-*myc* transgenic mice (a commonly used mouse model of Myc-induced B cell lymphoma) does not show mutations in those genes. This suggests that to allow the full development of *Myc*-induced lymphoma additional genes must be targeted. We

hypothesized that among these exist a set of tumor suppressor genes silenced via promoter CpG methylation that are positively selected during tumor progression. Hence, we applied the methods we developed (methylPipe & compEpiTools) for the identification of tumor suppressor genes regulated by DNA methylation in B-cell lymphoma.

1.5.3 Epigenomics and genomics determinants of RNA methylation

The methylation of RNA at the level of the N6 position of the adenosine (m6A), is emerging as an important area of investigation, thanks to the identification of enzymes critical for the establishment and regulation of this mark and the development of methodologies for its genome-wide investigation[125]. Commonly used methods are based on the immuno-precipitation with an antibody specific for m6A followed by high-throughput sequencing (MeRIP-seq), providing genome-wide low-resolution data about the patterning of this mark[126]. Based on this technique, a number of cell lines and tissues were profiled in mouse and human, identifying thousands of transcripts marked by m6A mostly in the 3'UTR, long internal exons and around the transcription start site (TSS), with a remarkable similarity between mouse and human[127]. The presence of this mark influences the splicing, translation, stability, localization and export of the transcripts, and the patterning of this mark can be dynamically established and removed, emphasizing the potential plasticity of this regulatory layer[127]. Recently, it has been found that m6A deposition is regulated by miRNA, and is critical for the processing of miRNA[128, 129]. Finally, RNA methylation has recently been shown to be relevant for the regulation of stem cells differentiation[130, 131].

Most of the key actors so far associated with the establishment and regulation of m6A are enriched in nuclear speckles, where they have the potential to directly interact with nascent RNA in an environment rich in components of the splicing machinery[127]. Epigenetic marks, such as DNA methylation and histone post-translational modifications, were also described to be involved in the regulation of the splicing process[132]. Nevertheless, there are currently no reports investigating on possible associations between RNA methylation and epigenomics or regulatory proteins such as transcription factors (TFs). In this study we explore this possibility, taking advantage of the fact that some of the cell lines used for studying m6A through MeRIP-seq experiments were extensively profiled through genomics and epigenomics approaches in independent studies. Specifically, we focus on the human embryonic stem cell (H1) and mouse embryonic fibroblast (MEFs). For the former, base-resolution bisulfite sequencing (for DNA methylation) data[72], a number of histone marks, and a number of transcriptional features (in ENCODE) are available, while for the latter low-resolution MeDIP-seq (for DNA methylation) data are available[133].

Materials and Methods

2. methylPipe

methylPipe provides memory efficient analysis of base resolution DNA methylation data in both the CpG and non-CpG sequence context. It allows integration of DNA methylation data derived from any methodology providing base- or low-resolution data.

2.1 Data Input and Processing

The input for methylPipe is base-resolution DNA methylation data that can be provided as tabular text files containing, for each profiled cytosine, the genomic positions and the number of reads with C or T (where the cytosine was protected or converted by the action of sodium bisulfite treatment[134], respectively). Alternatively, these data can be generated providing the path to SAM files of aligned reads such as, but not limited to, the alignment files obtained with the popular Bismark aligner[135]. Data are stored as Tabix-indexed compressed files[136] enabling compact representation and fast access to post-alignment processed DNA methylation data (*BSprepare* function). For example, the size of a compressed WGBS experiment for IMR90 and H1 human cell lines is 269MB and 380MB, respectively. Not only the limited size of this file results in a reduced disk space requirements (the size of the uncompressed flat files is 372MB and 554MB, respectively): these data can also be directly accessed from the disk, thanks to the Tabix indexing, further saving on the memory usage. Through this strategy methylPipe can easily accommodate data from

multiple WGBS experiments or any combination of WGBS and targeted base-resolution datasets. In addition to the methylPipe package, the complete set of mCs mapped in the IMR90 and H1 cell lines in the first human base-resolution DNA methylomes profiled by Lister and colleagues[72] are included in the ListerEtAlBSseq Bioconductor metadata package. The WGBS data available in this package were processed compressed and indexed with Tabix through methylPipe and can directly be accessed using the package functionalities.

When post-alignment tabular DNA methylation data are processed in methylPipe through the *BSprepare* function, the confidence of calling a mC is determined for each C through a binomial test[72]. Briefly, sodium bisulfite treatment of DNA specifically converts unmethylated C to U (ultimately read as T) without affecting methylated C. For a given cytosine in the reference genome, the more sequencing reads have a C, the higher is the likelihood of that C being methylated. The binomial test is performed taking into account both the bisulfite conversion rate, which is typically calculated by sequencing of an unmethylated spike-in, and the sequencing error rate[72]. The resulting multiple-testing corrected p-values are stored on the disk in the Tabix compressed and indexed file, and are available in methylPipe through the *BSdata* class. This class has methods to easily access and filter the base-resolution data based on sequencing depth and statistical significance of the mC call. While using the binomial test to measure the confidence of a methylation event is straightforward in case of cell lines or very pure cell populations, its interpretation could be less direct in case contaminants or subpopulations with mixed epigenetic states are present in the sample. In those cases, the number of reads with C (#C, supporting the methylation call at a given cytosine), the number of reads with T (#T, not supporting the methylation call),

and the combined methylation level summary #C/(#C + #T) are available in the *BSdata* class and should be used for evaluating this heterogeneity. Further methods are being developed to resolve distinct epigenomes in mixed population[137, 138].

2.2 Descriptive Statistics of DNA methylation

methylPipe allows checking the basic stats about the methylation data such as range, mean and quantile distribution of methylation and assess sample similarity with correlation and clustering analysis. The *methstats* method computes pairwise correlation coefficients (Pearson) between the methylation profiles across all the samples in *BSdataSet* object. This function plots scatter plot and correlation coefficients, also outputs a correlation matrix **[Figure 8]**. In addition clustering of samples is performed based on the similarity of their methylation profiles and is displayed as a dendrogram **[Figure 8]**.

These results can be used as quality controls on the data distribution of each individual sample and on the expected correlation structure between multiple samples: in an ideal experiment, replicates should be clustering close to each other, while samples from different conditions should be assigned to those specific groups.

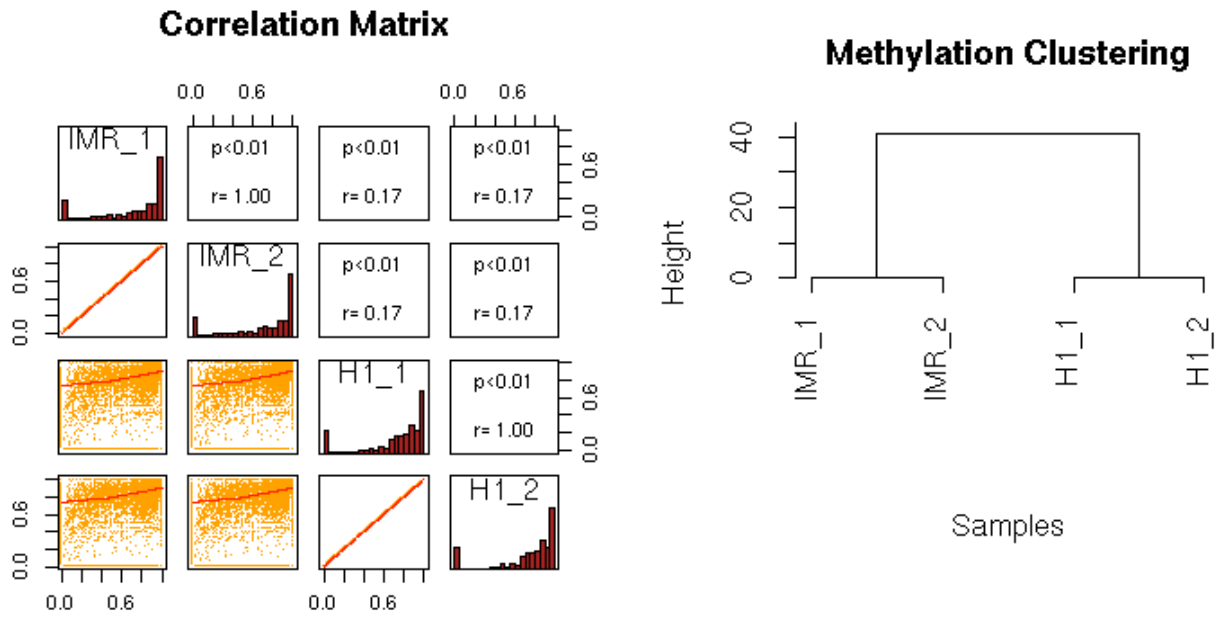


Figure 8: Quality control metrics.

Plot on the left pane shows the distribution of the methylation levels for each sample together with the pairwise scatter plot of these values comprehensive of p-values and R-values. Plot on the right pane reports the hierarchical clustering (Euclidean distance, complete linkage) of the samples based on their methylation levels, highlighting patterns and groups in the provided dataset. Distribution information and pair-wise correlation values are also returned at the command line for further processing of these results.

In addition, methylPipe provides multiple methods for filtering the methylated cytosines based on any combination of the following: sequencing depth, methylation level and binomial p-value. This is available in the methods used to profile methylation patterns, identify differentially methylated regions or determine quality checks.

2.3 Profiling DNA methylation on genomic ranges

DNA methylation data are typically profiled over a set of regions of interest (ROIs) such as CpG Islands or gene promoters. A class inheriting from the Bioconductor SummarizedExperiment class was defined for this task (*GEcollection* class) and has data slots specific for the absolute and relative density of DNA methylation events, which can be populated with the *profileDNAmetBin* method. The absolute methylation level of a genomic region (or bins thereof) is determined as the number of mCs per base-pair, while the relative methylation level is determined as the proportion of mCs over the total number of potential methylation sites in that region. While the methylPipe *GEcollection* class is designed to profile DNA methylation in a set of genomic regions for a single sample, the class *GElist* can be conveniently used to collect the same information for a number of samples and pass it to other methylPipe or compEpiTools methods.

Particular attention was dedicated to the strategy used for incorporating base-resolution information about unmethylated and uncovered (unsequenced) cytosines. While unmethylated Cs are the vast majority of the cytosines in all profiled genomes[139], the amount of uncovered Cs depends on the experimental technique adopted. Different techniques are available for the acquisition of base-resolution DNA methylation profiles. These can target the whole genome (WGBS) or only a subset of it, focusing on CpG-rich (Reduced Representation of Bisulfite Sequencing, RRBS[73]) or custom regions (such as the approach based on padlock probes[140]). In WGBS, most of the Cs are profiled, thus there is a limited number of uncovered Cs, and a majority of unmethylated Cs. On the other hand, RRBS or padlock experiments only cover a limited portion of the genome, resulting in data where few unmethylated Cs are vastly

outnumbered by a large majority of uncovered Cs. As a trade-off between these extremes, we decided to include in the BSdata class only the cytosines where at least one sequencing read has a C, thus building support on the methylation call. In addition, we provide a function to generate a GRanges including all the unmapped regions (with no sequence data) based on a BAM file, as most of the uncovered cytosines tend to occur in a limited number of refractory regions in WGBS experiments or in a relatively small number of regions complementary to the RRBS or padlock targeted regions. In this way, (i) considering Cs with #C >0 and (ii) having the list of the regions containing uncovered Cs, we can exactly recover the methylation status for each C in the genome as methylated (at a significant or not significant level), unmethylated or unmapped. Importantly, this piece of information is critical for the identification of the differentially methylated regions by methylPipe. Eventually, this results in a compression of WGBS experiments in relatively small files, while maintaining the ability of efficiently accommodating and integrating experiments with any level of genome targeting.

2.4 Identification of differentially methylated regions

The *findDMR* function uses the Wilcoxon or Kruskal-Wallis paired non-parametric tests for the identification of differentially methylated regions (DMRs) between two groups of samples or between multiple samples, respectively, by comparing the mC methylation levels. Briefly, the algorithm adopts a dynamic sliding window approach that identifies regions suitable for testing depending on mC density and relative distance, and possibly excluding regions with no, or negligible, variation between groups. More in detail:

i. A cytosine identified as methylated by the binomial test, having a minimum sequencing depth and a minimum methylation difference among the samples to compare (all these cutoffs are user-defined), is identified as the seed.

ii. Downstream Cs, satisfying the same criteria, within a maximum distance (D) from the seed is then considered. A minimum number of Cs (data points) is required within that window, while allowing a maximum number of missing data (unmapped Cs).

iii. The methylation level of the considered Cs is then compared between the samples using the statistical tests described above.

iv. The first mC call downstream of the position of the first seed C incremented of D/2 bp is considered as the seed of the next window and the process repeats from (i).

Alternatively, the *findDMR* function can utilize average methylation levels in discrete genomic regions. This could be useful in case of non-CpG methylation events, which are more unevenly distributed in the genome compared to mCpGs. Importantly, it is possible to upload GRanges with a list of Cs associated to known SNPs, which could confound the DMR analysis and that are consequently discarded. Moreover, when comparing differentiated to pluripotent cells it is suggested to include a GRanges object defining the partially methylated domains (*findPMDs*), which are large regions of partial methylation typical of differentiated cells that could cover up to 30% of their genome[72]. These regions are, by definition, differentially methylated when compared to pluripotent cells, and should be skipped in the DMR analysis presented here, which is targeted to smaller differential regions. Eventually, the *consolidateDMRs* function applies multiple testing correction to p-values, and significant genomic regions that are

close enough by an user-adjustable threshold are merged (their corresponding p-values are combined using Fisher's method).

The *findDMR* method was proven useful for the identification of DMRs in a number of studies[141, 121], resulting in the identification of genomic regions confirmed by other independent studies[140, 142]. In addition, the method was successfully adopted for the simultaneous analysis of hundreds of *A. thaliana* WGBS methylomes[143, 144], which are characterized by mosaic DNA methylation patterns[110]. Thus, methylPipe has been shown to be effective not only in the analysis of very large datasets, but also in managing DNA methylomes of various species, and in particular DNA methylomes with peculiar mC patterning compared to mammals.

Recently, an additional type of cytosine methylation was discovered, the 5-hydroxy-methylcytosine (hmC), which was proven to be a critical intermediary in active de-methylation pathways[145]. Bisulfite sequencing experiments do not distinguish hmC from mC. Specific experimental methods for the identification of this mark at the base-resolution were developed, and MLML is a popular computational method for a first analysis of these data[146]. methylPipe includes the *process.hmc* function to parse the MLML output and create a BSdata object specific for the hmCs data, which can then be combined with any other kind of DNA methylation data using the package functionalities.

2.5 Data integration and Visualization

A wide array of alternative methods providing low-resolution DNA methylation estimates is available, including MeDIP- or MBD-seq; these assays are based on the binding of mCs by a specific antibody or methyl-binding protein, respectively.

Computational methods are available to convert these low-resolution data into high-resolution estimates[147–149]. These high-resolution data can be subsequently imported in methylPipe as if they were native high-resolution bisulfite data. Alternatively, low-resolution data describing the methylation of genomic regions can be incorporated in methylPipe at the level of *GEcollection* and *GElist* objects, which were designed to summarize mC density in genomic regions.

Finally, the *plotMeth* method was developed to visualize DNA methylation base-resolution data, as well as their summary over genomic regions (or low-resolution DNA methylation data) along with gene models and other *omics* data or annotation tracks. This method, which allows a genome-browser like visualization of a specific genomic region, takes advantage of the Gviz Bioconductor library **[Figure 9]**.

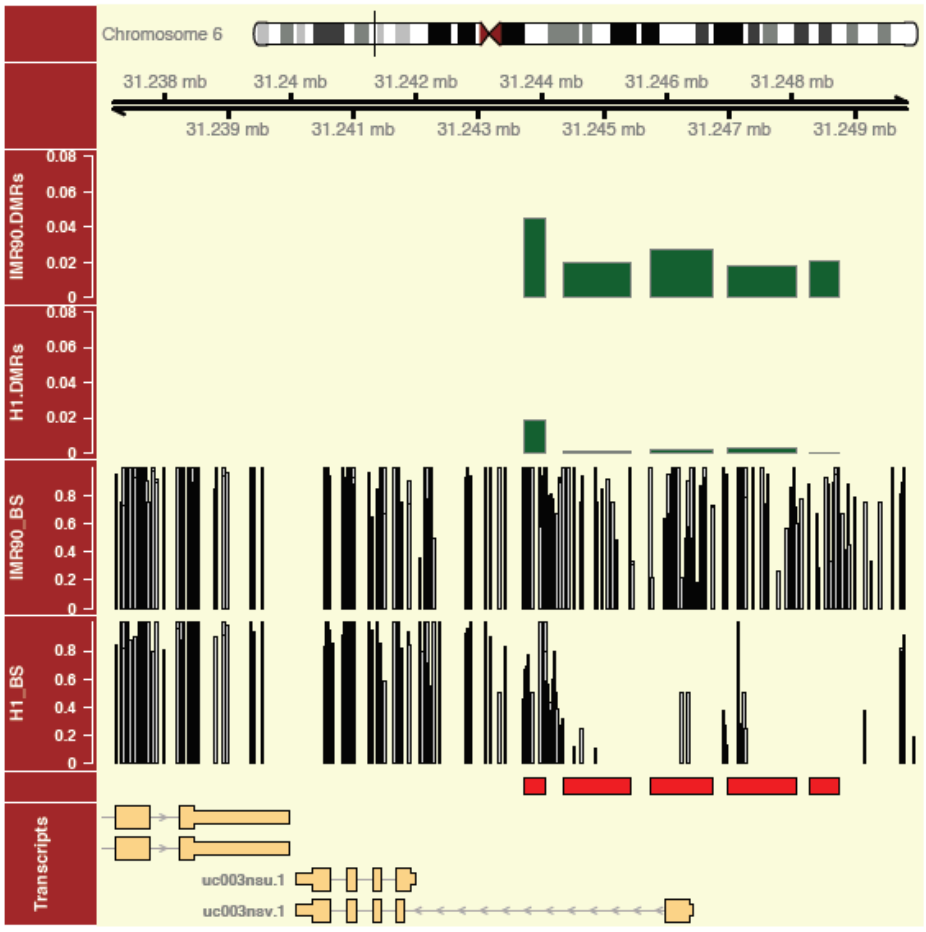


Figure 9: Base-resolution DNA methylation patterns displayed for a specific gene locus.

In this plot, DNA methylation is profiled for H1 and IMR90 samples in the identified DMRs. The "red" track represents the DMRs. The base resolution data is represented in H1_BS and IMR_BS tracks with methylation level ranging from 0-1. For each genomic region the absolute methylation density (mC/bp), the density of possible methylation sites (C/bp) and the relative methylation level (mC/C) is determined. The relative methylation of the DMRs in H1 and IMR90 are displayed in "green" track. This information is overlaid with genomic annotation using plotMeth. These tracks are visualized for the OCT4 locus, a key developmental gene, whose promoter is well known to be hypo-methylated in pluripotent cells, such as H1.

2.6 Computational performance of methylPipe

The computational performance of methylPipe was tested on whole genome bisulfite sequencing data from Lister et al 2009 (Table 1).

a) Data Processing: TABIX compression, indexing and computation of binomial tests, which are the steps necessary to create a new *BSdata* object in methylPipe, took 30 minutes with 1 core (max 4GB RAM peak usage) for the human H1 stem cells[72].
b) Profiling methylation data: After the data processing step, access to the data is fast: with one core, it can profile 100 human promoters in a sample in about 50 seconds (max 1GB RAM peak usage).
c) DMRs identification (pairwise): DMRs identification between 2 WGBS samples took 20 minutes with 1 core on chromosome 1 (max 4GB RAM peak usage), and 45 minutes with 10 cores for a genome-wide analysis (max 28GB RAM peak usage on a cluster).
d) DMRs identification (multi-sample): Finally, the most computationally intense task, i.e. the identification of DMRs among 8 WGBS methylomes[141], on chromosome 1

took 40 minutes with a single core (max 4.9GB RAM peak usage), proving to be manageable even on a laptop computer. Parallelization is implemented in the package, and it automatically adjusts to the available number of cores and RAM. The same DMR analysis of 8 WGBS methylomes could in fact be completed in a similar time on a cluster by assigning 10 cores.

Table 1: Computational performance of methylPipe

Functions	Time Taken	Memory Required
Data Processing[1]	30 mins	4 GB
Profiling promoters (100)	50 secs	1 GB
Chromosome-wise DMRs Identification (pairwise)[72]	20 mins	4 GB
Whole genome DMRs Identification (pairwise)	45 mins	28GB (10 cores)
Chromosome-wise DMRs Identification (multiple)[141]	40 mins	5 GB

2.7 Comparison with other tools

We compared the differentially methylated regions (DMRs) identified by methylPipe with other published tools using WGBS dataset (Lister el al, 2009). To assess the DMRs identified by methylPipe, we compared them to those identified by methylKit[150], methylSig[118], RADMeth[117], methPipe[151] and DSS[116] in terms of number of regions, overlap, and annotation. Some important issues complicating this comparison are following:

i. Some tools identify differentially methylated sites (DMCs)
ii. Some tools would not work with genome-wide WGBS for the DMR identification
iii. The difference in method of DMR identification

Keeping the above issues in mind, we tried to reach an approximate comparative estimation of DMRs by setting threshold on methylation difference (> 30%) and statistical significance (q-val< 0.05). We compared all the tools able to perform DMR (DMC) identification on genome-wide WGBS together with those only able to work with a very limited subset of the data. To this purpose, we only considered H1 vs. IMR90 methylation data for chr1. The following heatmap **[Figure 10]** (generated by *overlapOfGRanges* method of compEpiTools) reports the number and % overlap between the DMRs (DMCs) identified from the various considered methods:

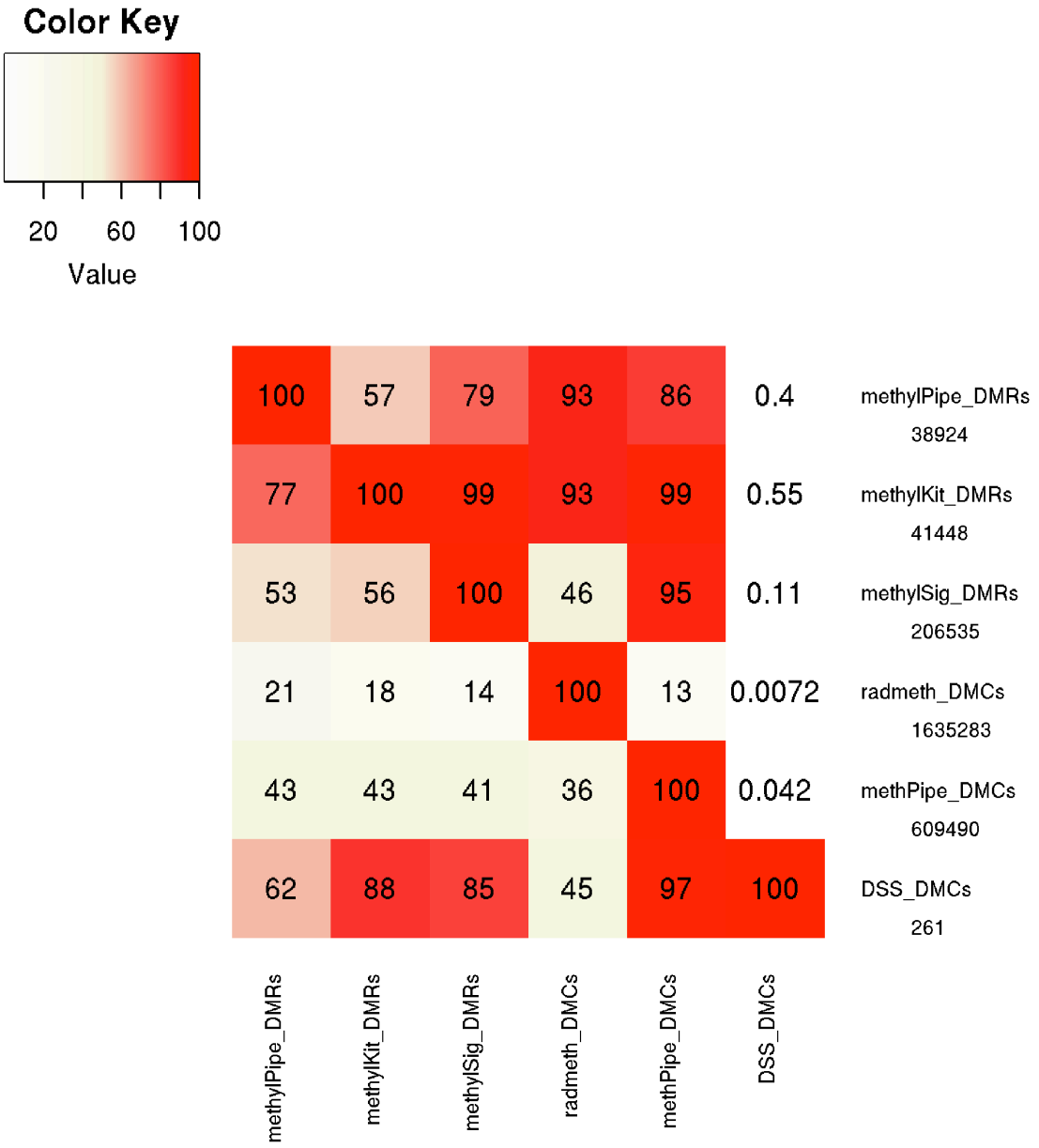


Figure 10: Heatmap of percentage overlap of DMRs/DMCs identified by various methods.

The figure displays the percentage overlap of DMRs/DMCs identified by various software tools between H1 and IMR90 cell line. Each box in the heatmap represents the pairwise percentage overlap of the regions identified as DMRs by both packages.

methylPipe, methylKit and methylSig shows reasonable percentage overlap of the identified DMRs regions. However, RADMeth, methPipe and DSS provide very discordant results. RADMeth identified 1635k individual DMCs, methPipe identified 609k DMCs whereas DSS only identified 261 individual DMCs. The overlap is calculated in respect to the method reported on the vertical axis (for example the first heatmap row is about the percentage of methylPipe detected regions confirmed by the other methods). While all this does not reassures about the bona-fide DMR identification it can be noticed that the first three methods have a good agreement. For example 57% and 79% of the methylPipe DMRs are confirmed by methylKit and methylSig, respectively, while methylKit basically provides a subset of the regions identified by methylSig. The comparison with the other three tools (RADMeth, methPipe and DSS) is greatly complicated by the very discordant number of sites identified. However, methylPipe shows good overlap with all the tools except DSS. All the identified DMRs/DMCs are further annotated (using *GRannotateSimple* method of compEpiTools) according to their genomic location **[Figure 11]**. As expected, methylPipe, methylKit and methylSig show similar annotation patterns of identified DMRs where RADMeth, methPipe and DSS show discordant genomic annotation patterns.

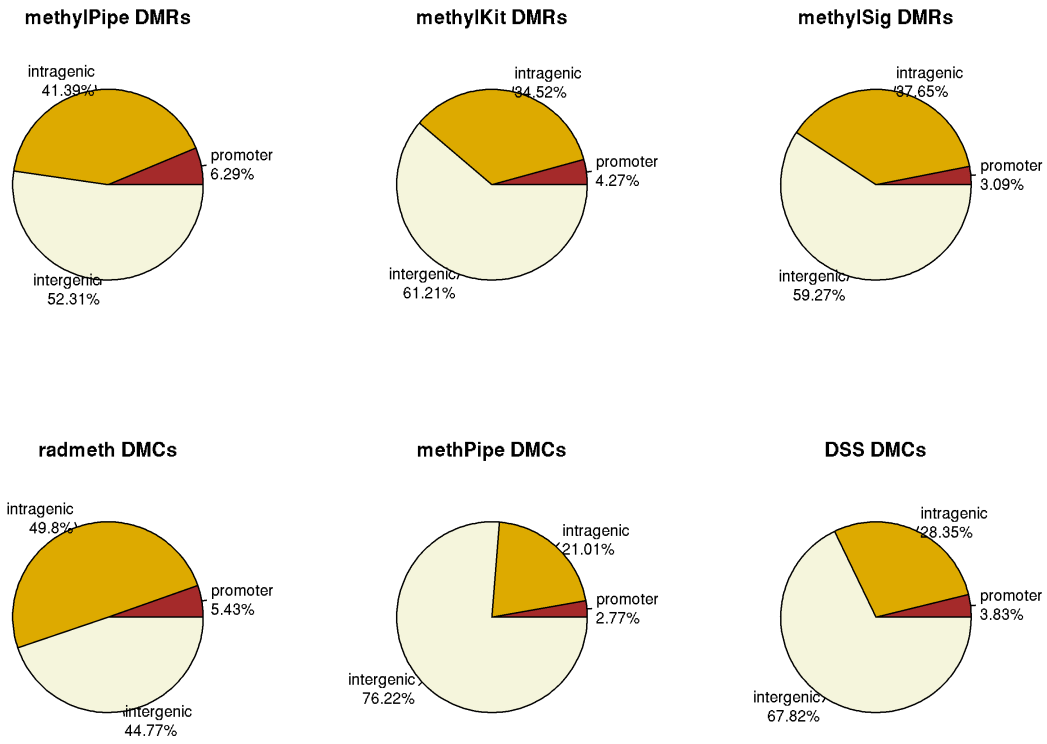


Figure 11: Genomic annotation of identified DMRs/DMCs.

The figure displays the percentage genomic annotation of DMRs/DMCs identified by various software tools between H1 and IMR90 cell line based on overlap with promoters, intergenic and intragenic regions derived from UCSC based annotation

.

3. compEpiTools

compEpiTools functionalities can be grouped in three main categories: (i) computing various read counts metrics in genomic regions, (ii) performing functional/genomic annotation, and (iii) integrated visualization of heterogeneous data-types.

3.1 Computing read counts metrics

compEpiTools facilitates several common operations associated with the quantification of the sequencing signal in a set of genomic regions. The base-level or overall count of reads within a set of regions can be determined based on BAM files, using the *GRbaseCoverage* and *GRcoverage* methods, respectively. Resulting counts can be normalized by library size and/or region length. In addition, specifically for ChIP-seq experiments, the peak summit position and the overall region enrichment given a matched input sample can be determined, through the *GRcoverageSummit* and *GRenrichment* method, respectively. Regarding RNA Polymerase II (RNAPII) ChIP-seq experiments, the *stallingIndex* and *plotStallingIndex* functions are available to compute and visualize the cumulative distribution of the stalling index (SI), thus estimating the degree of RNA Polymerase stalling[152]. The SI is defined as the ratio of the RNAPII signal in the promoter and genebody region. When comparing different samples, the SI could increase significantly because of either an increase at the level of RNAPII in the promoter or a decrease of in the genebody, or because of differential dynamics of RNAPII in these two regions. For this reason, to better dissect the dynamics of

differential SI, the cumulative SI distribution is conveniently integrated with the analysis of promoters and genebody RNAPII read densities.

3.2 Genomic Annotation

To ascertain the biological significance of a set of ROIs (ChIP-seq peaks, DMRs etc.), it is important to consider them in their genomic context. compEpiTools allows an effortless, rich and fast annotation of genomic regions based on Bioconductor standard annotation libraries derived from the UCSC genome database, using the *GRannotate* method **[Figure 12]**. Specifically, for each ROI, the distance from the nearest transcription start site and its location are determined. The region location is also annotated based on the overlap with promoters, intragenic and intergenic genomic regions, and the corresponding transcript and gene id(s) and symbol(s) are reported. The resulting GRanges conveniently embed the annotation for all the isoforms that might occur in correspondence of a given ROI. Notably, the user is provided the flexibility to supply additional sources of annotation, which could results from other *omics* analyses. These might also be any list of ROIs taken from the literature, or obtained within R using the ucscTableQuery function from the rtracklayer package to access UCSC Tables (e.g. the list of CpG Islands).

```
GRanges object with 5 ranges and 8 metadata columns:
        seqnames             ranges strand | nearest_tx_name distance_fromTSS
           <Rle>          <IRanges>  <Rle> |     <character>        <integer>
  18777     chr1 [4797174, 4797174]      + |      uc007afg.1              799
  18777     chr1 [4797174, 4797174]      + |      uc007afg.1              799
  21399     chr1 [4846975, 4846975]      + |      uc007afi.2              799
  21399     chr1 [4846975, 4846975]      + |      uc007afi.2              799
  21399     chr1 [4847609, 4847609]      + |      uc007afi.2              165
        nearest_gene_id nearest_gene_symbol                  location
            <character>         <character>               <character>
  18777           18777              Lypla1          promoter;promoter
  18777           18777              Lypla1          promoter;promoter
  21399           21399               Tcea1 promoter;promoter;promoter
  21399           21399               Tcea1 promoter;promoter;promoter
  21399           21399               Tcea1 promoter;promoter;promoter
                        location_tx_id  location_gene_id location_gene_symbol
                           <character>       <character>          <character>
  18777            uc007afg.1;uc007afh.1       18777;18777        Lypla1;Lypla1
  18777            uc007afg.1;uc007afh.1       18777;18777        Lypla1;Lypla1
  21399 uc007afi.2;uc011wht.1;uc011whu.1 21399;21399;21399    Tcea1;Tcea1;Tcea1
  21399 uc007afi.2;uc011wht.1;uc011whu.1 21399;21399;21399    Tcea1;Tcea1;Tcea1
  21399 uc007afi.2;uc011wht.1;uc011whu.1 21399;21399;21399    Tcea1;Tcea1;Tcea1
  -------
```

Figure 12: Genomic annotation output for the region of interest.

The method GRannotate based on a GRanges and a TxDb, returns the GRanges with a series of annotations. The program annotates the genomic regions of interest based on Bioconductor standard annotation libraries derived from the UCSC genome database. For each ROI, the distance from the nearest transcription start site and its location are determined. The location is further annotated based on the overlap with promoters, intragenic and intergenic genomic regions, and the corresponding transcript and gene id(s) and symbol(s) are reported.

3.3 Functional Annotation

Genomic regions can also be annotated in a number of epigenomic relevant states. The *enhancers* method can be used to identify enhancers based on promoter-distal H3K4me1 (indicative of enhancers that could be either active or poised) or H3K27ac (indicative of active enhancers) ChIP-seq peaks, possibly excluding CpG Islands regions [**Figure 13**]. Using the *matchEnhancers* method, enhancers can conveniently be matched to the most proximal genes, and possibly stratified based on the association to transcription factor (TF) binding events to study the TF-dependent activity of enhancer regions and putative target regions.

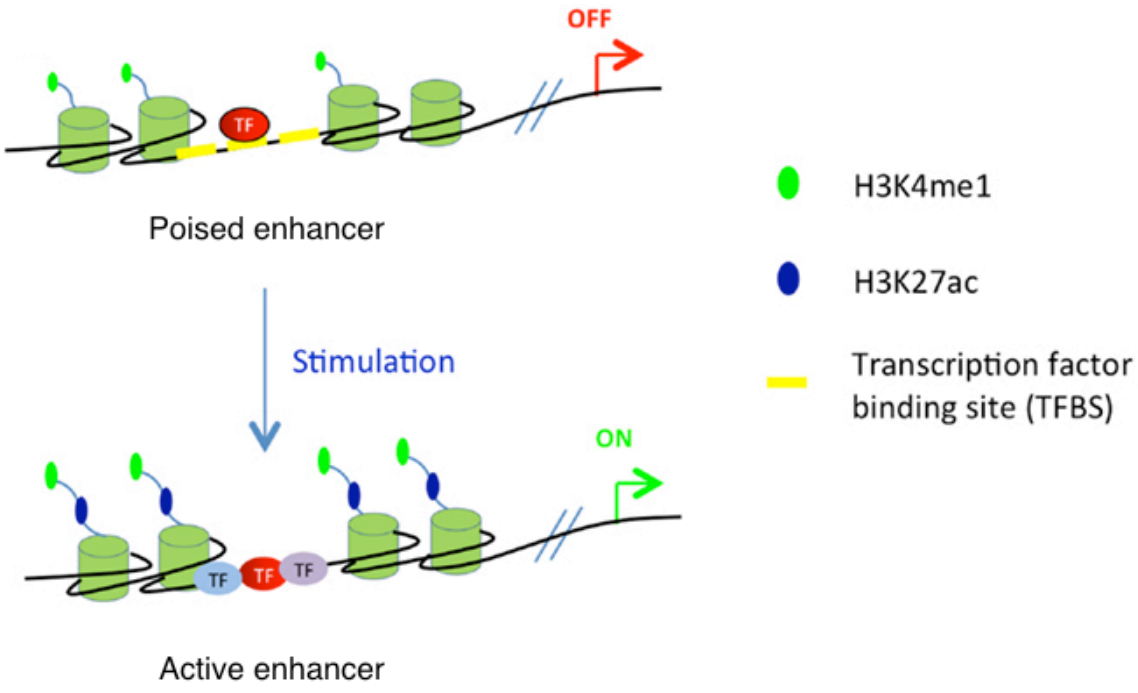


Figure 13: Enhancers identification.

Enhancers are defined as distal H3K4me1 peaks not overlapping with CGI, to avoid un-annotated transcriptional units. Alternatively H3K27ac peaks could be used to identify active enhancers.

Particularly relevant for DNA methylation is the concept of promoter CpG-content, which is critical for the epigenetic control on the downstream gene expression: variation in the absolute or relative methylation at the level of intermediate or high-CpG density promoters was proven to be associated to differential expression of the downstream gene, compared to low CpG content promoters[153]. The promoter CpG context can be determined with compEpiTools through a sliding window scoring approach proposed by[101], implemented in the *getPromoterClass* function [**Figure 14**].

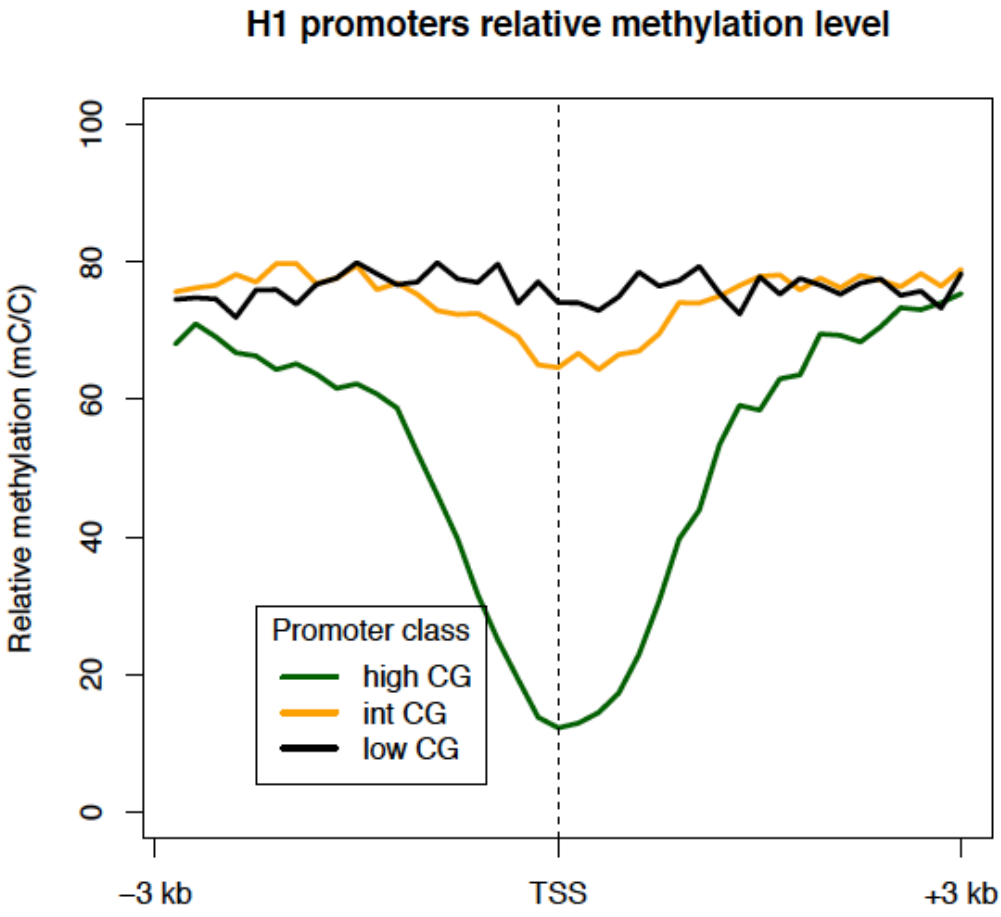


Figure 14: Promoter class by CG content.

getPromoterClass can be used to classify promoters according to their CpG content. In the plot, the promoter CpG content is determined and promoters are classified into low-, intermediate- and high-CpG content. Then, relative DNA methylation (mC/C) is profiled for promoters in the three classes. As can be expected, highCG content promoters have lower methylation level whereas; lowCG and intCG promoters have considerably higher relative methylation around the TSS.

Long non-coding RNAs (lncRNAs) can be identified in compEpiTools using the *findLncRNA* function based on their epigenetic signatures. Briefly, H3K4me3 peaks distal from promoters and associated with a lower H3K4me1 reads density (thus avoiding enhancer regions), are identified as seed for the identification of potential lncRNA promoters. Evidence for RNA transcription in the regions downstream and upstream these H3K4me3 peaks are evaluated by computing the read density of RNA-seq, H3K79me2 and/or RNAPII. Random genomic regions of the same size, not overlapping with promoters, are used as a background to determine the random expected density of these marks of transcriptional activity. Regions with a signal for these marks greater than the 95th percentile of the background are then selected as putative regions expressing lncRNAs.

Finally, a convenient wrapper (*topGOres*) is provided to perform GeneOntology (GO) enrichment analysis on a set of Entrez gene ids (query) based on the topGO Bioconductor package [**Figure 15**]. A common problem in the analysis of GO enrichments that complicates the interpretation of the results is the redundancy between enriched GO terms that are very close in the considered ontology. For this purpose, compEpiTools provides the *simplifyGO*terms function for pruning poorly informative and redundant enriched terms. The rationale behind this pruning is that often parent and a child enriched-terms point to very similar GO terms, associated to a very similar set of genes. Iteratively, for each enriched term T, the parent of T is searched within the set of enriched terms, based on the specified ontology. If both a parent and a child terms were identified as enriched and if they match to a set of genes overlapping more than a user-adjustable threshold within the query, the parent term is discarded in favor of the more specific child term.

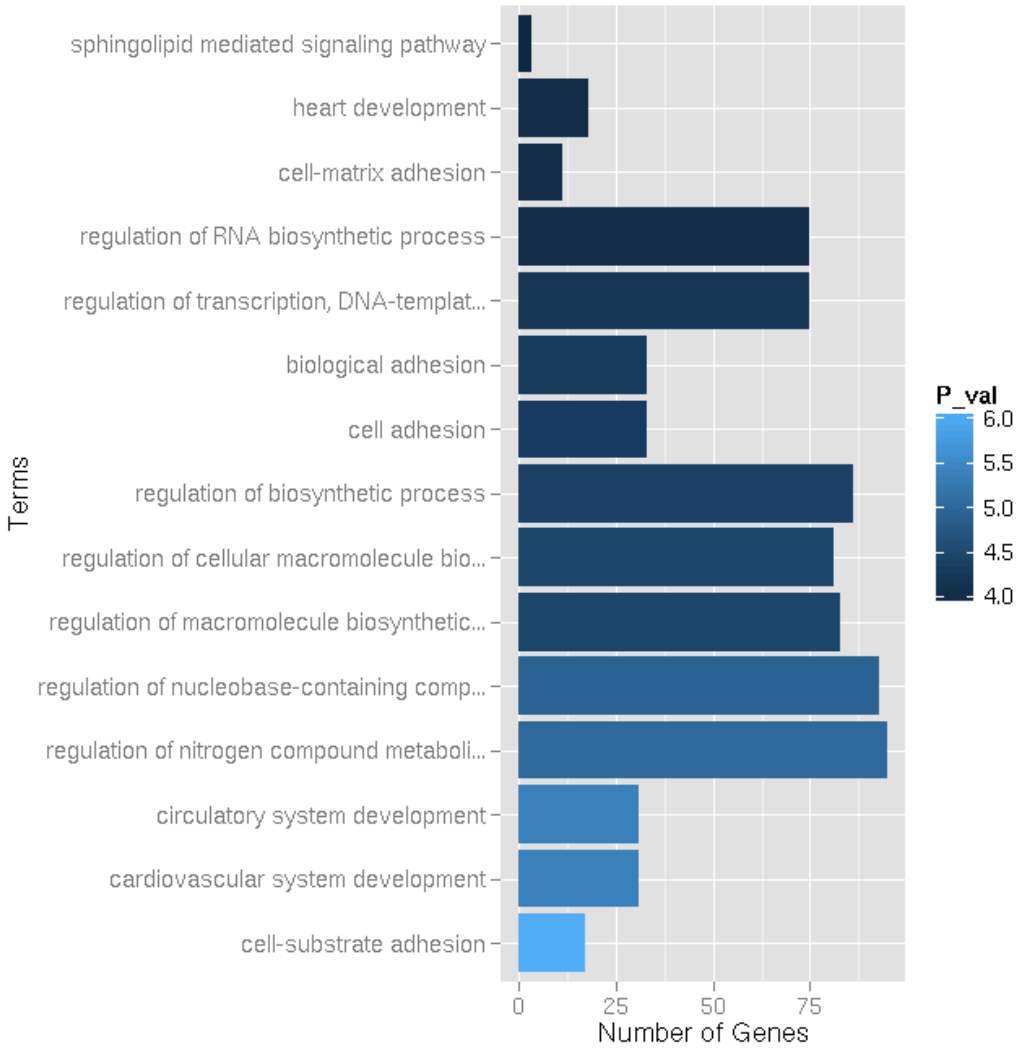


Figure 15: GeneOntology annotation plot.

topGOres is used to identify significantly enriched GeneOntology terms. The plot (output of topGOres) displays the top ranking terms with number of genes and associated p-value (-log10 (p-value)).

3.4 Data integration and visualization

The integration of heterogeneous data types is a challenging task, and explorative analyses based on the generation of heatmaps are frequently used to highlight patterns in combined datasets. In our experience, the creation of such heatmaps requires an extensive number of processing steps, especially when applied to datasets composed of heterogeneous data types and annotation tracks, discouraging the repeated use of these tools. Rather, heatmaps are typically iteratively generated until a satisfactory combination of data tracks, clustering and normalization settings is identified. A powerful and efficient visualization system based on heatmaps is provided in compEpiTools, based on the *heatmapData* and *heatmapPlot* functions. Heatmap rows represent ROIs and columns represent data tracks. Every track can be assigned to any of the supported data types: GRanges, GRanges metadata, BAM files, and GElist and GEcollection objects generated by methylPipe. Thus, any combination of base-resolution or low-resolution DNA methylation data, histone marks, TF binding, RNA-seq expression and genomic annotations, including gene models, is accommodated. Quantile or thresholding-based normalization methods can be activated independently for each track, to emphasize patterns in the combined dataset and adjust the signal range of the track (for example to exclude outliers or underweight data tracks that are overall poorly scoring in the ROIs). Clustering of rows can be activated, including data from all or selected tracks. Dividing each ROI in a user-defined number of uniformly sized bins can control the resolution of the displayed data. Importantly, each track can be supplied with significance scores, which can be conveniently used to progressively dim the color of low-scoring (less significant) hits, while maintaining full brightness for the significant ones. The data matrix underlying the heatmap is returned together with

the dendrogram structure, allowing further analysis of the clusters of interest [**Figure 16**].

In the following plot, the regions of interest (plotted regions) is determined 10 kb upstream and downstream of each of the DMRs identified as hypo-methylated in H1 compared to IMR90 on chromosome 6 (methylation difference greater than 25 percent). For these regions (ROIs), the *heatmapData* function is used to integrate data from various data types. The relative density of DNA methylation is computed using *profileDNAmetBin* function of methylPipe. For H3K4me1 and H3K4me3 data tracks (in both H1 and IMR90 samples) are provided as GRanges containing pre-determined ChIP-seq peaks (while they could contain any kind of genomic regions). These tracks are represented in the heatmap, as presence or absence of a ChIP-seq peak for each ROI (or bins thereof). For H3K27me3 and H3k36me3 data tracks the path to the aligned reads stored in BAM files (limited to chr6) are provided. The *heatmapData* function consequently computes the reads density (coverage) for these BAM files in the ROIs. The reads density is further normalized by library size. Finally, the RNA-seq data tracks, the pre-computed reads density is provided for both the H1 and IMR90. The resolution of the data to be displayed for each track is defined based on the number of bins (nbins) that each ROI is uniformly divided into (20 in this example).

The list resulting from the *heatmapData* function is passed to *heatmapPlot* to display the heatmap. This is convenient in case one would repeat this step testing several plotting and normalization settings, saving the time needed to determine the raw data underlying the heatmap. When calling *heatmapPlot* the data can be normalized

independently for each track based on a specific signal percentile (which is set as the maximum saturation value and displayed as 1, corresponding to full red in this example), or based on a specific arbitrarily chosen threshold. In this case a hybrid approach is used. Several tracks for the same data type (for example histone reads density) are temporarily combined and their overall 85th percentile is set as the maximum value. The gene annotation information for the forward and reverse strand can be automatically extracted from a *TranscriptDb* object and overlaid in the heatmap, reporting exons in red and intron in pink. This offers the possibility of adding a commonly desired annotation track, using a custom graphic representation to highlight introns and exons. The clustering of rows can be activated specifying the index of the tracks to be used for clustering; in this case all of them are used including gene annotation tracks. This option could be useful to direct the clustering to use only a subset of the data or annotation tracks, increasing the flexibility and allowing emphasizing various patterns in the data.

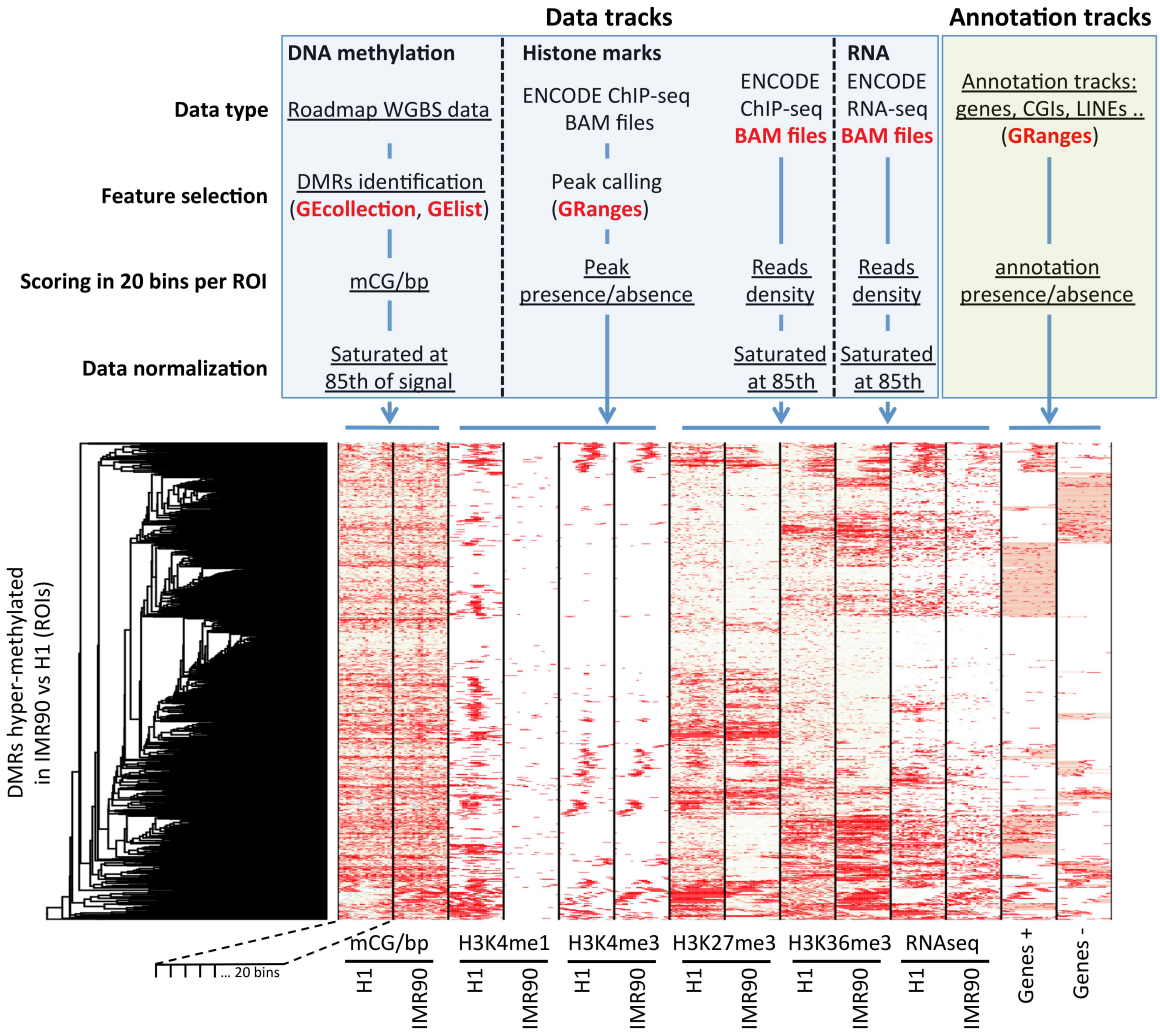


Figure 16: The integrative heatmap generated by heatmapData and heatmapPlot functions.

Heatmaps can easily be obtained incorporating any mixture of data and annotation tracks. Heatmap rows represent ROIs, while columns represent tracks profiled over those ROIs (or bins thereof). Data and annotation tracks might contain either quantitative (e.g. normalized reads counts) or categorical (e.g. presence/absence of a ChIP-seq peak) data. If available, the significance of associated data can be incorporated affecting color brightness. In this example, generated as described in detail in the supplemental material, NIH Roadmap DNA methylation data where visualized together with ENCODE histone marks for a set of differentially methylated regions. ROIs were clustered based on the data available in all the displayed tracks including gene models annotations. The schema on the top of the figure depicts the workflow leading to the heatmap. A set of standard Bioconductor objects, listed in red, is the input for the heatmapData and heatmapPlot compEpiTools functions. The underlined text points to the key analysis steps automatically performed internally to the functions generating the heatmap, calling routines available in the same packages.

3.5 Computational performance and comparison with other tools

Most of the functionalities offered by compEpiTools can be run on the order of minutes or less. For example, profiling the normalized number of reads from a GRanges of 40.000 ROIs and a typical BAM ChIP-seq file takes less than 40 seconds. Dividing each region in a number of bins does not require much additional time, because the binning is performed after the initial count, which is the most time consuming step. Building heatmaps with a dozen of tracks is typically performed in a few minutes, mostly depending on the number of ROIs to be clustered. To achieve maximum efficiency, optimized clustering routines, as implemented in the fastcluster R package, are adopted[154]. The only tool which to our knowledge is comparable in functionality with compEpiTools is the Bioconductor RepiTools package[155]. While RepiTools provides a useful set of tools for the integrative analysis of epigenomics data, mostly focused on statistical testing, integration with gene expression data and visualization, it is tailored to enrichment-based epigenomics data only, and it is unable to provide most of the compEpiTools functionalities listed in **Table 2**.

		BiSeq	M3D	Bsseq	DSS	*methylKit*	*methPipe*	*RADMeth*	*methylSig*	*WBSA*	*Methy-pipe*
methylPipe	Targeted BS-seq data	+	+	+	+	+	+	+	+	+	+
	WGBS data	-	-	-	+	-	+	+	-	(a)	-
	Multi-WGBS dataset	-	-	-	-	-	+	+	-	-	-
	Non-CpG mCs	-	-	-	-	-	-	-	-	+	+
	hmCs	-	-	-	-	+	+	-	-	-	-
	Low-resDNA methylation	-	-	-	-	-	-	-	-	-	-
	Absolute methylation	-	-	-	-	-	-	-	-	-	-

	Relative methylation	+	-	+	-	+	+	-	+	+	+
	Binning of ROIs	-	-	-	-	-	-	-	-	-	-
	Pairwise DMR analysis (45′)	+ NA	+ NA	+ NA	+ 3d	+ NA	+ 36′	+ 90′	+ NA	+ NA	+ NA
	Multi-groups DMR analysis	-	-	+	-	-	-	+	-	-	-
	Browser-like data plot	-	-	-		-	-	-	+	-	-
compEpiTools	Promoter-CpG content	-	-	-	-	-	-	-	-	-	-
	Reads counting	-	-	-	-	-	-	-	-	-	-
	Signal enrichment	-	-	-	-	-	-	-	-	-	-
	ROIs Annotation	+	-	-	-	+	-	-	+	+	-
	RNAPII stalling index	-	-	-	-	-	-	-	-	-	-
	Non-redundant GO enrichment	-	-	-	-	-	-	-	-	-	-
	Enhancers	-	-	-	-	-	-	-	-	-	-
	lncRNAs	-	-	-	-	-	-	-	-	-	-
	Integrative heatmaps	-	-	-	-	-	-	-	-	-	-
Reference		[115]	[156]	[157]	[116]	[150]	[151]	[117]	[118]	[158]	[159]

Table 2: Comparison of methylPipe and compEpiTools features with functionalities offered by other similar tools. The first column lists the key features offered by methylPipe and compEpiTools. Column headers report the tool name and reference. A "+" sign indicates that the feature is provided by a given tool, while a "–" sign indicates that it is not available. The "Pairwise DMR analysis" row includes in parenthesis the time (in minutes or days) needed for a complete WGBS differential analysis between two samples; NA is reported if this analysis is not supported for WGBS data. (a) WBSA is an online web-service imposing a limitation or 2GB for the upload of fastq files, which is clearly insufficient for the analysis of a WGBS dataset; the software can be installed locally although this requires significant effort (requiring Perl, R,

MySQL, Java and C compiler) and it is only available for Linux; the analysis of the H1 and IMR90 WGBS was reported by the Authors to be completed in one week.

Table 2 provides a comparison of the features offered by methylPipe and compEpiTools with those offered by other computational tools able to analyze epigenomics data. Most of the tools available for the analysis of these data are implemented as R packages, focusing only on the analysis of DNA methylation data. Currently, 38 packages are available in Bioconductor associated with the DNA methylation assay domain. The large majority of these packages were developed for the analysis of data generated with the 450K Illumina platform, which is able to profile only about 1% of the cytosines that are typically found in a complete human DNA methylome. Only four of these packages (BiSeq, M3D, bsseq and DSS) are potentially able to manage WGBS data. These packages provide a very limited subset of the functionalities offered by the methylPipe / compEpiTools packages, most of these tools were developed and tested for the identification of DMRs on RRBS data (**Table 2**), and claim to be able to analyze WGBS data. We tested whether they could perform three specific tasks that we consider necessary in the analysis of WGBS data:

I. Uploading a single WGBS dataset and profiling a set of ROIs,
II. Identifying DMRs between 2 conditions,
III. Identifying DMRs between multiple conditions.

BiSeq[115] and M3D[156] are designed to upload the entire dataset into memory, and we failed with both in uploading an entire WGBS dataset even when 80GB of memory was provided (we could only upload and work with data for chromosome 1). Consequently, we were unable to perform any of the three proposed testing operations.

Regarding the remaining programs, bsseq[157] only provides a smoothing-based method to identify DMRs, without offering additional functionalities, and DSS[116] is not specific for DNA methylation data. Neither of these tools delivered satisfactory results: DSS completed the DMR identification in a 2-group comparison in about 3 days, while after the same amount of time bsseq returned an error. In addition to these Bioconductor packages, few additional stand-alone tools or web-services are available to manage base-resolution DNA methylation data. Among these, only methPipe[151] and RADMeth[117], both developed by the Smith Lab, can analyze WGBS datasets and perform DMR analyses. The time needed by these tools for the identification of DMRs (36 and 90 minutes, using 1 and 10 cores respectively) is similar or slightly higher than methylPipe (45 minutes using 10 cores) (**Table 2**). To our knowledge, the examined tools do not provide an extensive set of supplemental functionalities beyond to those listed in the Table 2. In this regard, methPipe is the only exception, providing additional routines that are complimentary to those offered by methylPipe. In summary, only methPipe and RADMeth were able to efficiently complete the proposed tasks with standard resources. Importantly, neither these tools nor the other software packages, limited on the analysis of targeted DNA methylation data, could match the complete set of functionalities offered by methylPipe and compEpiTools (**Table 2**).

Results

4. Epigenomics landscape of B-cell lymphoma

This study is being carried out in Dr. Bruno Amati lab where Dr. Alessandra Majorana (post-doc) has designed the experiment, prepared the RRBS library and is further carrying out the experimental validation of the computational results. We have analyzed the RRBS methylation data, developed methodology for the integrative analysis of DNA methylation data with ChIP-seq and RNA-seq data and finally identified a list of putative tumor suppressor genes for experimental validation.

4.1 Experimental Methods

To identify genes silenced by CpG methylation we collected samples from Eu-myc mice in tumor stages (four biological replicates) and B-cells from normal mice (three biological replicates) that we used as control and subjected them to genome wide methylation analysis through Illumina sequencing. We used Reduced Representation Bisulfite Sequencing (RRBS) a bisulfite-based protocol that enriches CG-rich regions of the genome, thereby reducing the amount of sequencing required while capturing the majority of promoters[160]. The advantage of this method is that it provides single-nucleotide resolution with high sensitivity and limited cost.

4.2 Data processing

We developed the computational pipeline for the analysis of RRBS data. Firstly, we performed preprocessing of bisulfite reads by removing bad sequencing and low quality reads. The processed reads were thereafter aligned to the reference genome mm9 using the Bismark aligner (v0.5.4). The DNA methylation information in terms of percentage methylation with coverage at each cytosine is then extracted from SAM file generated from Bismark. Since a high enough read coverage would increase the power of the statistical tests, the bases with low read coverage (< 5) were discarded from further analysis. To remove PCR bias, the bases that have more than 99.9th percentile of coverage in each sample were also discarded.

4.3 Coverage Statistics

To assess the empirical genome coverage of the method, we calculated the number of reads for each of the following regions: I) CpG Islands II) Gene promoters (defined as ±1kb of TSS). The results show **[Figure 17]** that RRBS covered most of the CpG islands and promoters. The coverage is consistent across all the samples reassuring on the lack of coverage biases in any of the samples that could have affected down-stream analyses.

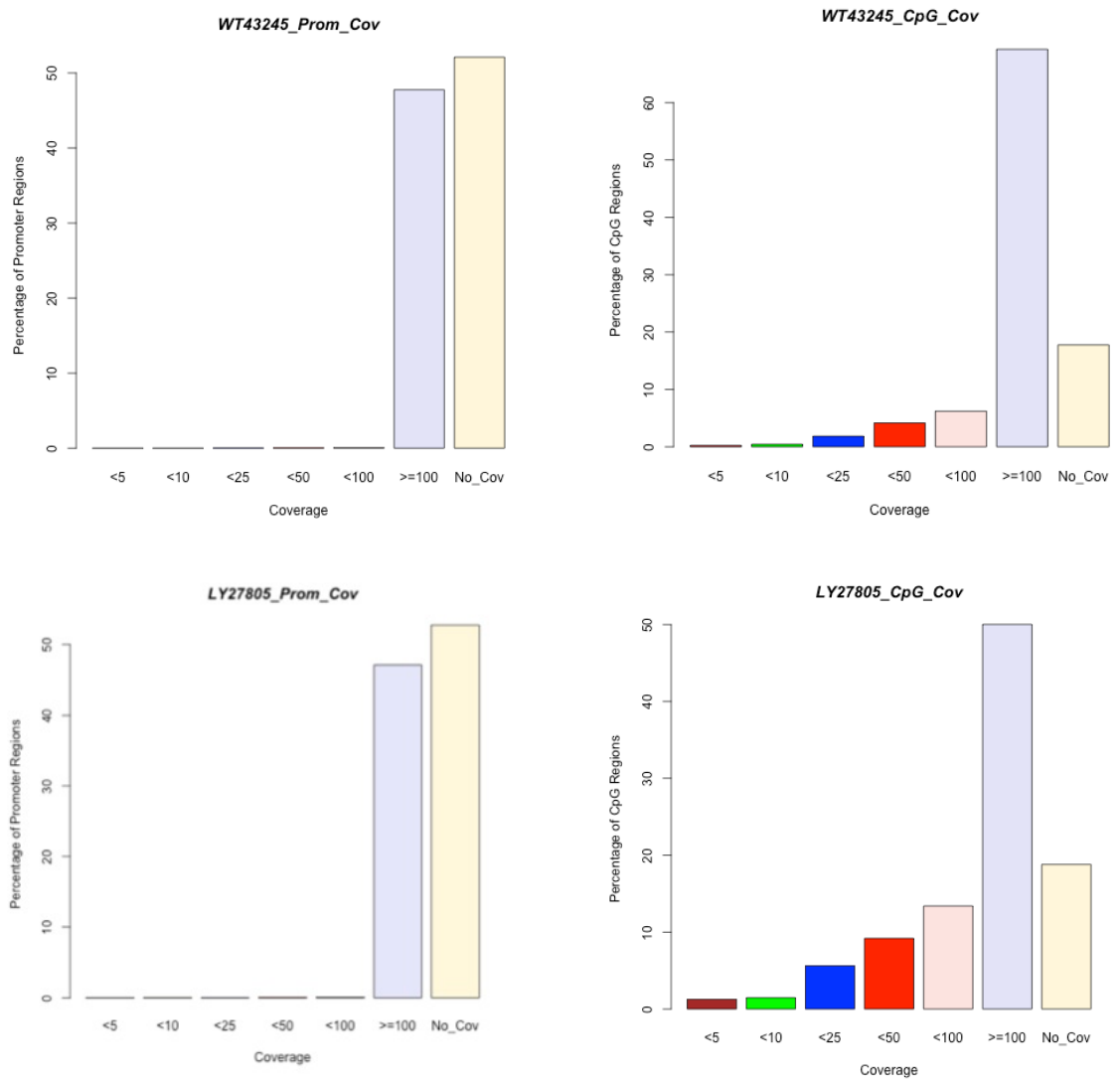


Figure 17: Coverage statistics.

The figure above shows number of individual CpG methylation measurements for the CpG Islands (top) and promoters (bottom) for control sample (WT43245) and tumor sample (LY27805) and. Promoter regions were calculated based on Ensembl gene annotations, such that the region starts 1 kb upstream of the annotated transcription start site (TSS) and extends to 2 kb downstream of the TSS. CpG islands information were obtained from the UCSC browser.

4.4 Identification of Differentially methylated Regions (DMRs)

DMRs were identified by computing the number of methylated versus un-methylated CpGs and tested for statistically significant differences between two samples using

Fisher's exact test. The replicates within the normal group are pooled together and compared with tumor samples. The complete tiling of the genome was performed for genome-wide DMR detection. The genome was tiled with windows 100bp length and 100bp step-size. This generated a list of DMRs for each comparison with percentage methylation and p-value. Following the differential methylation test and calculation of P-values, sliding linear model (SLIM) method was used to correct P-values to q-values[161], which corrects for the problem of multiple hypothesis testing. Thereafter, the hyper-methylated and hypo-methylated regions that have q-value<0.05 and percent methylation difference larger than 20% were selected. In total, 26585 hyper-methylated DMRs were selected from four tumor samples compared to control. These regions were put into genomic context by annotating them according to their genomic location.

4.5 Annotation of DMRs

To ascertain the biological significance of each differential methylation events, it was put into its genomic context for subsequent analysis. Thus, each list of DMRs was annotated according to their location with regard to CpG islands, proximity to the nearest transcription start site (TSS) and gene components, e.g. intron and exons. The annotation was performed using *GRannotate* function of compEpiTools.

4.6 Associating DNA methylation with RNA-seq Information

To discern the tumor suppressor genes that are silenced through cytosine methylation within CpG elements of their promoters, it was important to associate gene expression

information with the differential methylated regions. Therefore, all the DMRs within the promoter region (-1kb upstream and +2kb downstream of TSS) of each gene that were simultaneously hyper-methylated were combined. Differential methylation percentage is computed for each gene by computing average of all methylation differences of the combined regions. A combined q-value of these regions is computed by Fisher method. The gene expression information (RNA-seq) was mapped to the differential methylation information for each gene. Only the genes hypermethylated by more than 20% and displaying change in gene expression greater than two fold were considered for the analysis.

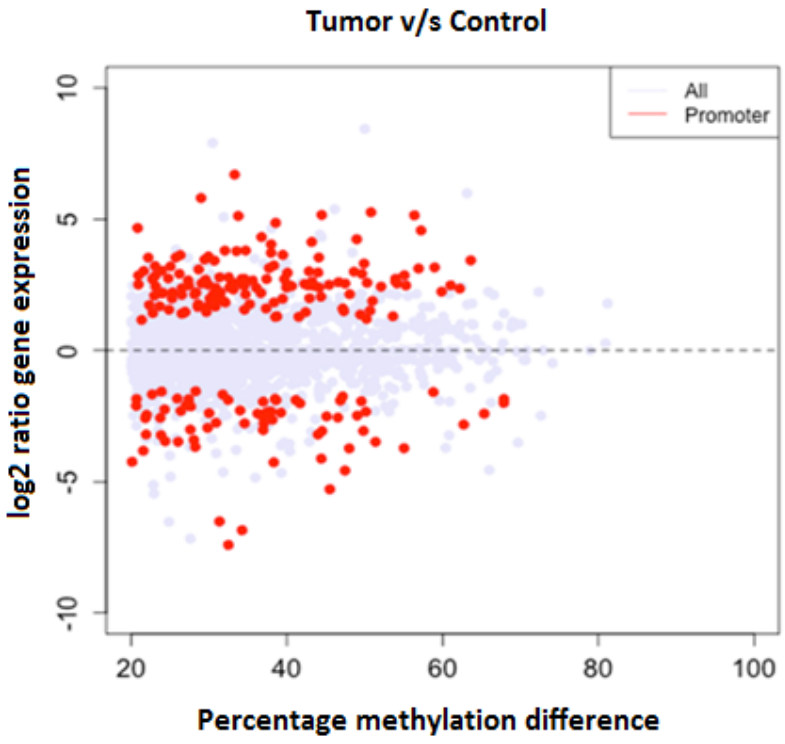


Figure 18: Associating DNA methylation with RNA-seq information.

The figure above represents the log2 fold changes of gene expression and DNA methylation for tumor sample v/s control samples. Each point on the plot represents a gene with red ones are the genes, which are statistically significantly differentially expressed (q-value < 0.05) with hypermethylated promoter (q-value < 0.05).

We did not observe strong overlap between alteration in DNA methylation and expression differences in tumor **[Figure 18]**. This observation can be reconciled with the hypothesis that DNA methylation plays an important role in transcriptional silencing of tumor suppressor genes but it affects only a moderate number of genes, rather than a large and unspecific set of genes. It may be reasoned that most genes most of the time are not actively regulated by DNA methylation. Given the relatively small number of genes with overlapping DNA methylation and gene expression changes, we reasoned that genes exhibiting consistently negative association between these two properties might constitute potential tumor suppressor gene candidates [**Figure 19**].

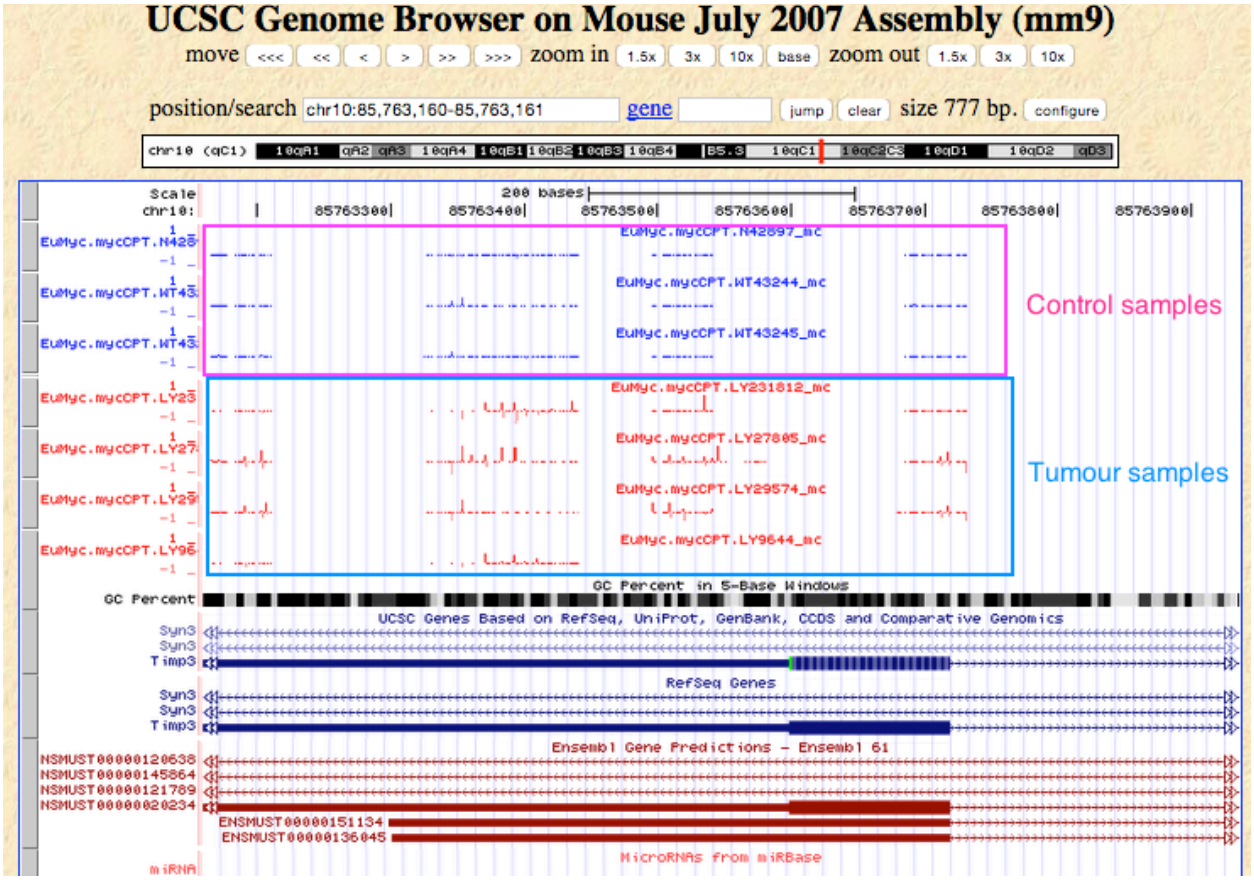


Figure 19: UCSC view of hypermethylated gene promoter.

The figure displays tracks of base resolution DNA methylation level for control and tumor samples. Timp3 gene is hypermethylated and down-regulated (gene expression) in tumor as compared to control samples. Timp3 is an inhibitor of the matrix metalloproteinases, a group of peptidases involved in degradation of the extracellular matrix (ECM) and is hypermethylated in gastric cancer.

4.7 Associating DNA methylation with Histone marks Information

Histone modifications play an important role in mediating patterns of DNA methylation[162]. We thought it would be interesting to understand how this relation between histone modifications and DNA methylation provides us insight into the aberrant gene expression pattern in our tumor model. To investigate this relationship, we implemented the methodology to determine the overlap between enrichment/depletion of H3K4me3 and Pol II with the DNA methylation pattern at the gene promoter region in tumor samples compared to normal ones. Firstly, we computed the normalized (by library size) reads coverage (of H3K4me3 and Pol II) at the promoter region (+-1kb of TSS) in both tumor and control samples. We then used Mann Whitney non-parametric statistical test on these computed promoter regions to ascertain the genes showing significant enrichment/depletion of H3K4me3 and Pol II at the promoter regions (tumor v/s control). Thereafter, log2 fold change (of histone mark enrichment/depletion) is computed between tumor v/s control to give quantitative score to these changes. Finally, this quantitative histone information was mapped to the differential methylation information for each gene. Only the genes hyper-methylated by more than 20 % and displaying significant alteration in histone modification enrichment was considered for the analysis. We observed stronger overlap between alterations in histone modification enrichment and DNA methylation changes in tumor **[Figure 20]**.

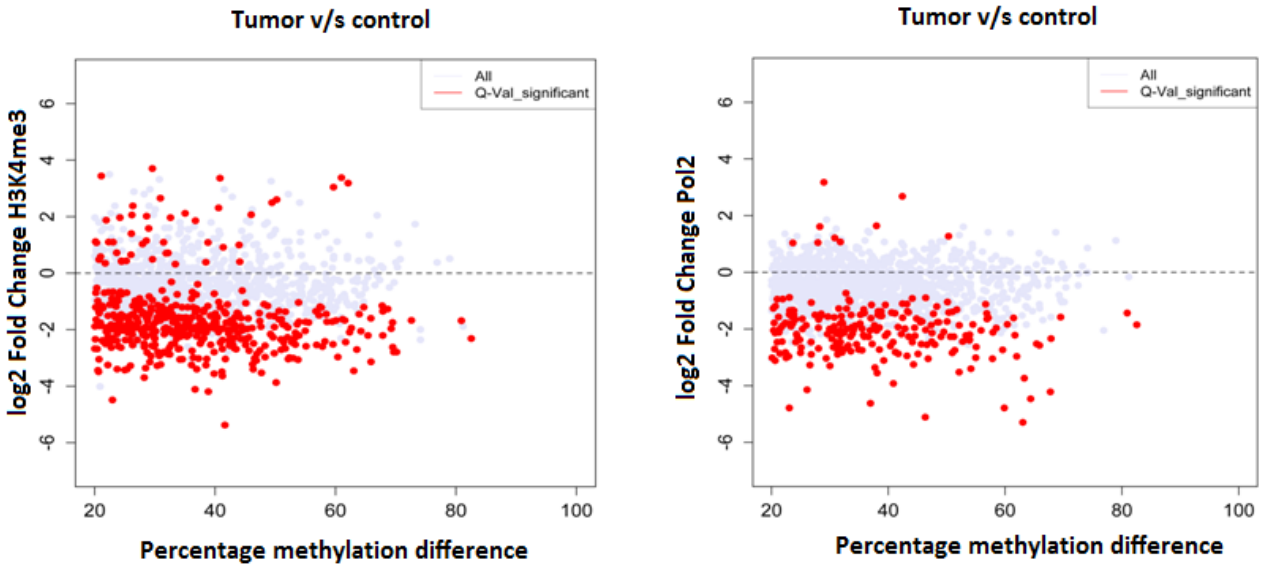


Figure 20: Association of DNA methylation with histone modifications

The figure represents the log2 fold change of H3K4me3 (left panel) and Pol2 (right panel) for tumor sample v/s control samples.

Eventually, we selected 316 genes classified based on the epigenomics data relations: Group 1 consisting of all genes hypermethylated in promoter and down-regulated in gene expression; Group 2 consisting of all genes hypermethylated in promoter and down-regulated in H3K4me3 or Pol II; Group 3 consisting of all genes that are down-regulated (gene expression), whose promoters are hyper-methylated and shows H3K4me3 and Pol II depletion in tumor as compared to control samples. These candidate tumor suppressor genes are presently being screened in-vivo by sh-RNA library for experimental validation.

5. Epigenomic and genomic determinants of RNA methylation

5.1 Background

The role of epigenetic modifications of histones and DNA has been extensively studied in various biological contexts but similar studies about the RNA modifications are still in nascent stage. The N6-methyladenosine (m6A) is the most abundant RNA modification whose genomic distribution and biological significance has only recently been researched [163]. This modification is laid down by a multi-component methyltransferase complex comprising of methyltransferase-like 3 (METTL3), METTL14 and Wilms tumor 1-associating protein (WTAP). The m6A modification can be removed by α-ketoglutarate-dependent dioxygenase FTO and alkylated DNA repair protein alkB homolog 5 (ALKBH5). The presence of m6A RNA demethylation and demethylases illustrates that m6A modification is dynamic and reversible in nature similar to other epigenetic modifications. Till now, two YTH domain-containing proteins, YTHDF2 and YTHDC1, have been identified as the readers of m6A-modification. YTHDF2 mediated mRNA degradation is dependent on methylation of target RNAs and thereby regulates mRNA stability and localization. Mutational studies and transcriptome-wide profiling of m6A have revealed RRACU as its primarily modified consensus sequence. This modification is mostly enriched around stop codons, in 3' un-translated regions (3' UTRs) and within long internal exons. Moreover, it affects various components of RNA metabolism including translation, splicing, RNA stability, transport and localization. Till now the primary mechanism suggested for its biological roles is the modulation of the RNA-protein interaction. The key components associated with the establishment and regulation of m6A are enriched in nuclear speckles. In this environment rich in

components of the splicing machinery they could directly interact with nascent RNA [127]. On the other hand, the role of other epigenetic marks, such as DNA methylation and histone post-translational modifications in the splicing mechanism and transcription modulation has been well established. So we investigated the possible associations between RNA methylation and epigenomics or regulatory proteins by integrative analysis of publicly available datasets of some of the cell lines that are profiled for RNA methylation. We focused on the human embryonic stem cell (H1) for which base-resolution DNA methylation data; a number of histone marks, and transcription factors data are available. In addition we also used data from the mouse embryonic fibroblast (MEFs) cell lines for which low-resolution MeDIP-seq (DNA methylation) data are available.

5.2 Materials and Methods

5.2.1 Source of the publicly available datasets

RNA methylation data for the MEF cells: RNA methylation peaks were retrieved for GSM1518036 GEO sample, and specifically from the GSE61995_final_peaks.xlsx file available on the corresponding GSE61995 GEO series[131].

DNA methylation data for the MEF cells: MeDIP-seq peaks were retrieved from the GSM886553 GEO sample; peaks were called using MACS on mm9 by the authors[133].

CTCF peaks for MEF cells: ChIP-seq peaks for CTCF were retrieved from the GSM918743 GEO sample; peaks were called using MACS on mm9 by the authors as part of the mouse ENCODE project[164].

RNA methylation data for the H1 cells: RNA methylation data were downloaded for the GSM1272365 (H1 m6A) and GSM1272366 (input) GEO samples; the corresponding SRR1035222.sra and SRR1035221.sra files containing MeRIP-seq raw reads were retrieved from the SRA for H1 m6A and input, respectively[130].

DNA methylation data for the H1 cells: the H1 whole-genome bisulfite data available in the ListerEtAlBSseq Bioconductor package were considered[72].

Histone marks and TF ChIP-seq data for the H1 cells are downloaded from ENCODE project.

5.2.2. Processing of public data

MEF m6A released peaks coordinates were mapped to mm10. Peaks were consequently moved to mm9 using the LiftOver UCSC online tool.

H1 m6A peaks from[130] were called again from the raw data. Sra files were converted to fastq files using the fastq-dump tool. Raw reads were aligned to the hg18 genome using topHat (version 2.0.6) with default settings for un-stranded data and single-end reads[165]. Peaks were called on the resulting BAM files (H1 m6A vs input) using MACS (version 2.0.9) using default settings and the --auto-bimodal flag[166].

Histone marks and H1 TF ChIP-seq data were converted to hg18 using the liftOver tool implemented in the rtracklayer Bioconductor package.

5.2.3. Integration with low-resolution DNA methylation data

Mouse TSS coordinates were identified using the *TSS* method of the compEpiTools Bioconductor package, based on the mm9 TxDb Bioconductor metadata package containing transcripts definitions. mm9 CpG Islands coordinates were obtained from UCSC using the rtracklayer Bioconductor package. 20Kbp regions centered at the mm9 TSS coordinates were considered (55419 regions), and within this set we identified 2230 regions overlapping with m6A and DNA methylation peaks in MEF cells (overlap>=1bp).

5.2.4. Integration with base-resolution DNA methylation data

H1 m6A peaks mapping to autosomal and sex chromosomes were retained. The summit (point of highest MeRIP-seq enrichment) was determined using the *GRcoverageSummit* from the compEpiTools package. m6A summits were divided into intragenic, intergenic and promoter summits using the *GRannotateSimple* from the compEpiTools package. Intergenic peaks were further filtered removing those summits laying closer than 5Kbp to an hg18 genebody. Peaks into exons and 3'UTRs were also identified based on the information extracted from the TxDb hg18 Bioconductor package. Exons matching 3'UTRs are removed from the exons set. At the same time, random exons and 3'UTRs were identified, sampling from the overall set of exons and 3'UTRs not associated with m6A a number of regions equal to the number of those associated with m6A. 4Kbp regions centered at the m6A summit were defined.

The m6A enrichment reported in **Figure 22** is determined using the compEpiTools package (normalized number of reads in the m6A IP subtracted of the normalized number of reads in the control sample; where normalized indicates that is divided by the total number of aligned reads in the corresponding BAM file).

Absolute and relative DNA methylation within these 4Kbp regions (divided in 20 equally sized bins) were determined using the methylPipe Bioconductor package, focusing on cytosine positions covered by at least 5 reads reading C (unconverted by bisulfite, thus supporting the methylation call), and binomial corrected p-value lower than 0.01. Intragenic regions and those mapped to promoters, exons and 3'UTRs for genes on the minus strand were reverted, so that the signal -2Kbp upstream the m6A peaks indicate the signal on the 5' of the RNA methylation event.

5.2.5. Prediction of m6A peaks from epigenetic and regulatory features

To measure how much (epi) genomic features are predictive of m6A we used m6A peaks and ENCODE datasets of histone marks and regulatory proteins for H1 cell lines. Before building a model, it was imperative to generate a simulated dataset (representing background) of the m6A-regions. The majority of m6A peaks lies in exon regions (nearly 70%). To consider the genomic prevalence of this modification we generated m6A- regions by random sampling of exon regions not overlapping with m6A peaks. This would mean that no genomic composition bias is included while comparing foreground (m6A+) and background regions (m6A-). We had 28871 m6A peaks (m6A+) and 28871 random exon regions devoid of any m6A peaks (m6A-), totaling 57742 genomic regions. These genomic regions were limited to a length of 288 bases determined from the average length of m6A peak. m6A+ regions co-ordinates are defined as ± 144b from the m6A peak summit whereas m6A- regions co-ordinates are defined as ± 144b from the midpoint of random exon region. We also added CpG island information and RRACT motif information to these set of features. A classification matrix was computed representing the presence or absence (1 or 0) of the associated (epi)genomic features. Only DNA methylation values was represented in continuous numerical scale determined as relative methylation level by *profileDNAmetBin* function of methylPipe. The association of each of the marks with m6A positivity was assessed using univariate logistic regression and penalized linear regression method LASSO.

5.3 Results

5.3.1 RNA and DNA methylation

In order to evaluate if RNA methylation peaks (m6A) could be associated to any corresponding pattern on the methylation of DNA (5-methyl cytosine, 5mC), we looked for high-throughput experiments profiling these data types in the same cell type. We were able to identify two independent studies profiling m6A and 5mC in mouse embryonic fibroblast (MEF) cells[131, 133], and other two independent studies analyzing these marks in human embryonic stem cells (H1 cell line)[72, 130]. While both studies profiled m6A using the MeRIP-seq methodology, DNA methylation was profiled in the MEF cells using the MeDIP-seq methodology (providing low-resolution data) and in the H1 cells using whole-genome bisulfite sequencing (providing base-resolution data).

5.3.2 Integration with low-resolution DNA methylation data

RNA and DNA methylation peaks in MEF cells were retrieved from the original publication, and accounted for 10167 and 187860 peaks, respectively. For a first qualitative analysis on the patterning of these two marks, we focused on 20Kbp genomic regions centered at mouse TSS, including the genomic counterpart for most of the expected m6A sites (5'UTRs and the longest internal exons). We specifically selected those regions containing peaks for both marks (2230 regions), and RNA and DNA methylation peaks within these regions are displayed in **Figure 21** together with CpG

Islands (CGIs) and gene annotation. Considering that each region was divided into 20 bins (1Kbp/bin), the spearman correlation for all the m6A and 5mC bins in all the regions is 0.35. When these two marks are co-occurring, their patterning is remarkably similar, as is the correspondence with the exons.

Both m6A and 5mC are implicated in splicing processes[127, 132], but no evidence of joint action of these marks in this process has been reported yet. We reasoned that genomic enrichment of 5mC within the same exons marked by m6A in the corresponding transcripts could be relevant for proper RNA splicing, and we speculated that the exons in the 2230 regions where m6A and 5mC are co-occurring and highly correlated might represent a set of genes whose splicing is actively controlled. Several regulatory proteins are involved in the control of splicing mediated by DNA methylation, including CTCF, MeCP2 and HP1, and we could identify an ENCODE ChIP-seq experiment targeting CTCF in MEF cells[164]. Overall, taking into account exons not simultaneously marked by m6A and 5mC, only 7% of them are associated to CTCF in MEF cells. The same proportion of exons is associated to CTCF even when restricting this set to the exons marked by 5mC MeDIP peaks. Rather, when focusing on the exons that are marked by both m6A and 5mC, 14% of them are associated with CTCF, thus supporting the idea that DNA methylation and RNA methylation co-occur since they are cooperatively involved in active splicing of these transcripts.

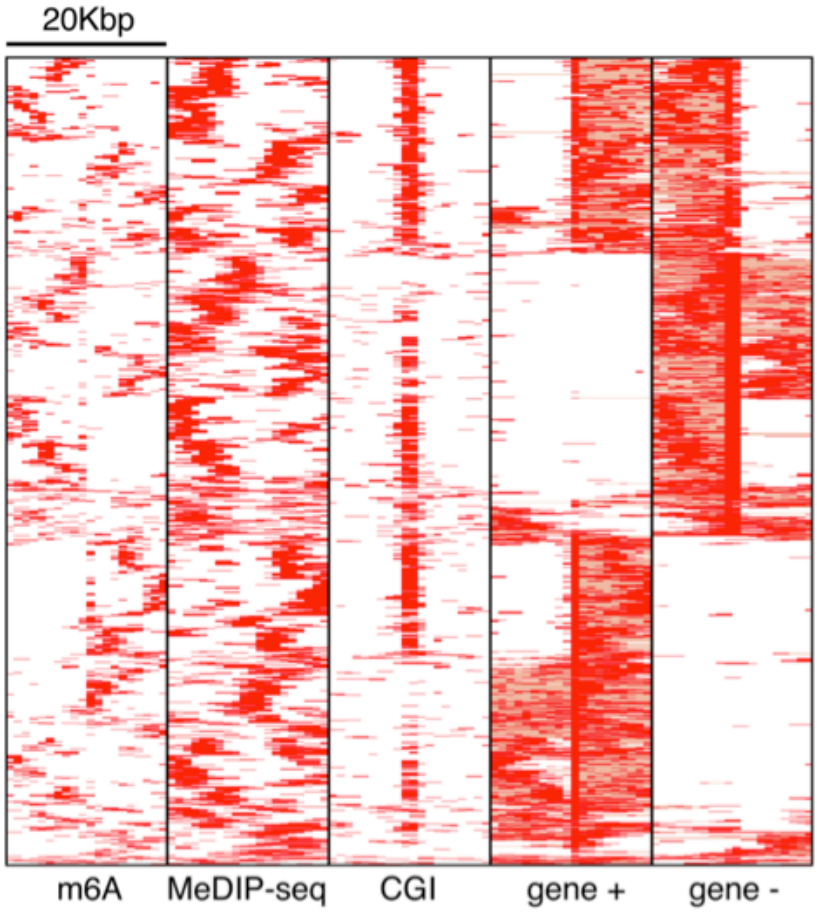


Figure 21: Association of m6A peaks with MeDIP-seq DNA methylation peaks in MEF cells.

Each row of the heatmap refers to a 20Kbp mouse genomic region, centered at a mouse TSS, containing at least one m6A peak and one MeDIP-seq peak. Each region is divided in 20 equally sized bins. For the m6A and MeDIP-seq tracks, a bin is colored in red if it matches a peak of the corresponding data type. For the CGI track, red bins indicate the presence of CpG Islands. For the gene + and gene – tracks, red bins indicate exons and pink bins indicate introns on the forward and reverse strand, respectively.

5.3.3. Integration with base-resolution DNA methylation data

There are two possible reasons for the high density of 5mC events in correspondence of m6A marks: (i) there is an increased frequency of 5mC out of the total number of cytosines possibly methylated, or (ii) there is an increased density in the total number of

cytosines possibly methylated. These two scenarios are not mutually exclusive, and MeDIP-seq data do not easily allow shedding light on this point[147, 167]. To clarify this aspect and explore DNA methylation patterns at the base-level in correspondence of m6A peaks, we took advantage of the availability of RNA methylation and base-resolution DNA methylation data in human H1 cells[72, 130].

We could identify 28872 m6A peaks in H1 cells by re-analyzing MeRIP-seq raw data[130]. We stratified these peaks into different genomic functional units, including intergenic regions, promoters, intragenic regions, exons, and 3'UTRs. For each m6A peak, we considered a 4Kbp region centered at the peak summit, and we determined MeRIP-seq enrichment together with average absolute (mC/bp) and relative (mC/C) DNA methylation profiles. We considered 5mCs in both the CpG and non-CpG context (divided in CHG and CHH, where H is any DNA base but G), since they are all highly frequent in embryonic stem cells[72, 141]. The CpG content (mC/bp in the CpG context) increases at the level of the m6A peak for all these regions. While the density of 5mCs follows the increasing CpG density in all the considered regions with the exception of the promoter, the resulting relative DNA methylation (mC/C) indicates that the number of 5mCs does not keep up with the number of potential sites particularly in promoter regions and exons [**Figure 22**]. While for the promoter regions this might be expected, this is not necessarily the case for exons. This finding is not in contrast with the result obtained in MEF cells, since the observed increasing density for the mCpG/bp mirrors the analogous trend observed for the MeDIP-seq signal in MEF cells. Complimentary to the pattern of increased mCpG density described in mouse (and confirmed in human), the base-resolution analysis of 5mCs in the CpG context in human reveals incomplete DNA methylation of the CpGs. This depletion might suggest

that the regions that are marked by m6A in the transcripts are interaction spots for specific regulatory proteins on the genome. Interestingly, the same pattern of depletion of mCpGs reported for promoters and exons can be found for 5mCs in the non-CpG context in all considered regions [**Figure 23**]. Differential patterning between 5mCs in these two contexts were already reported, suggesting different functionality and control of these marks[72, 141]. Importantly, all these patterns of 5mC depletion in the CpG and non-CpG context are specific for exons and 3'UTRs associated with m6A peaks, and cannot be found in control exons and 3'UTRs devoid of that mark.

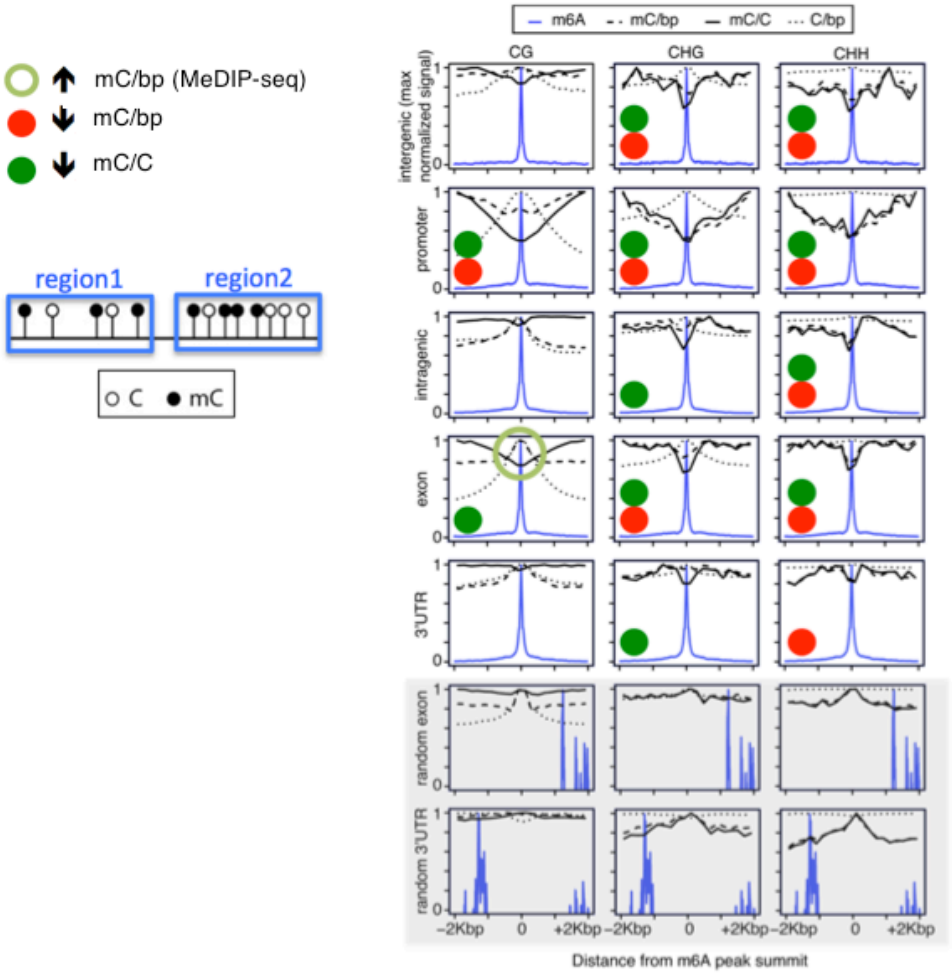


Figure 22: Broad depletion of DNA methylation in correspondence of m6A peaks in H1 cells.

MeRIP-seq enrichment (blue lines) in 4Kbp regions centered at the m6A peak summit in H1 cells is stratified by a set of human genomic functional units. For each functional unit, the absolute density of 5mC events (mC/bp), the density of potential DNA methylation sites (C/bp), and the relative DNA methylation (mC/C) are determined for 20 equally sized bins. 5mC in different sequence context (CG, CHG and CHH, where H is any base but G) are separately plotted. Random exon regions do not contain m6A peaks and their number matches the number of exon regions containing m6A peaks. Random 3'UTR regions do not contain m6A peaks and their number matches the number of 3'UTR regions containing m6A peaks. Each data series in these plots is normalized dividing by each series maximum value.

m6As were found to be predominantly associated to adenosines in the RRACT context[127], where R could match with either G (more frequent for this motif) or A. We took advantage of the base-resolution DNA methylation data to explore the genomic prevalence of 5mC in correspondence of the C within the motif (a potential non-CpG methylation site), and in the bases surrounding the motif occurrences. 20120 m6A peaks can be associated to the RRACT motif on the genome, and in case of multiple occurrences we focused on the motif that is closest to the m6A peak summit. We considered a 20bp region centered on the motif cytosine, and we counted for each position the number of 5mCs normalized for the total number of potential 5mC sites, stratifying by sequence context (**Figure 23A**). In general, 36% of the m6A peaks associated with RRACT have at least one 5mC. The methylation levels of CpGs are remarkably depleted at RRACT sites associated with m6A peaks in correspondence to those not associated with m6A peaks (compare **Figure 23B** with **Figure 23C**).

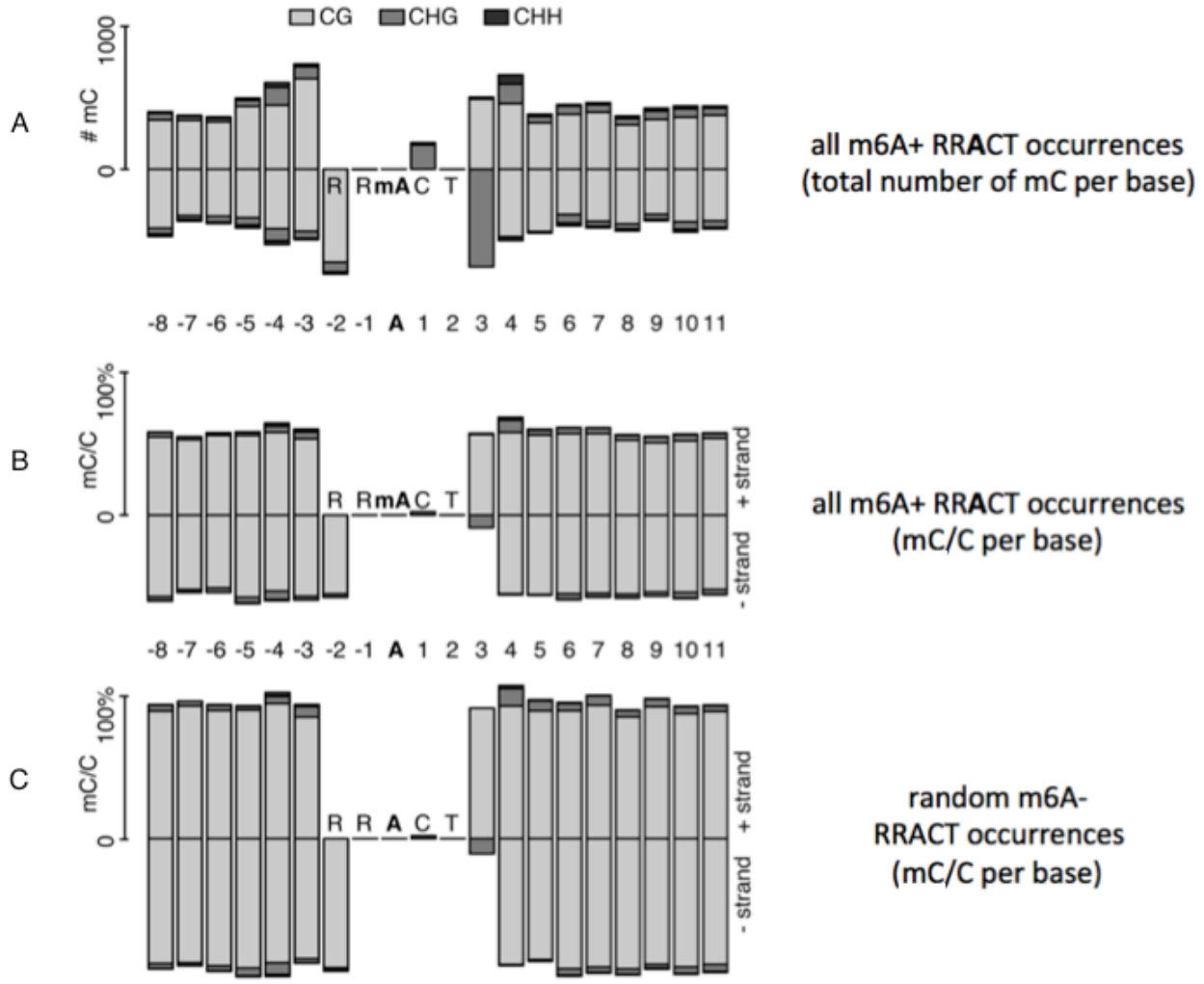


Figure 23: Depletion of DNA methylation around the RRACT m6A motif in H1 cells.

(A) The count of 5mC events, in a given sequence context (CG, CHG or CHH, where H is any base but G) and for a given DNA strand, is determined for each human genomic base in 20bp genomic regions centered at the RRACT motif associated to a m6A peak. (B) For each base the relative methylation is determined dividing the data reported in panel (A) by the number of regions having a potential DNA methylation site of the considered context at that position and strand. (C) As in panel (B) but for an equal number of random 20bp regions containing the RRACT motif not matching any H1 m6A peak.

5.4 Association of m6A with various epigenomics and regulatory features

To investigate the association of (epi) genomic and regulatory features with m6A used different kinds of analysis were used: multivariate analysis and univariate analysis. The analysis was performed in collaboration with Dr. Lara Lusa (Biostatistician, University of Ljubljana). The multivariate analysis performed using penalized linear regression method LASSO builds a model to test combined predictive ability of epigenomics/regulatory features to m6A positivity. A 5-fold cross-validation analysis was performed to test the classification ability of this model. The cross validation AUC of the model built using LASSO was about 0.70 [**Figure 24**].

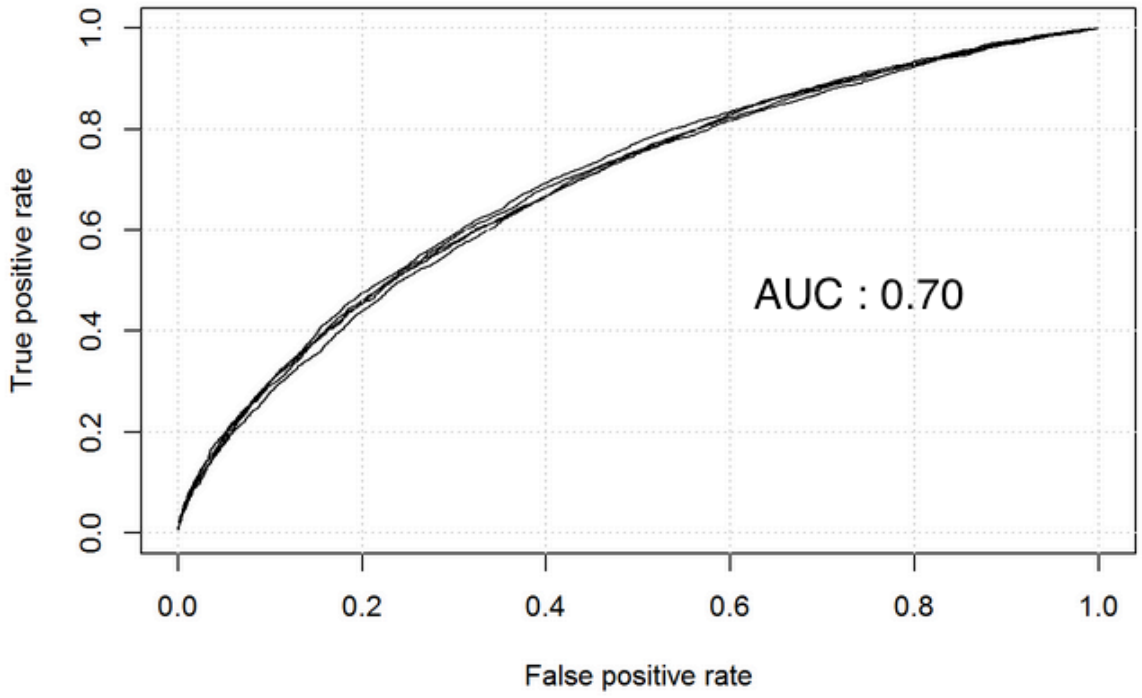


Figure 24: AUC curves obtained from LASSO method

The cross-validation curves indicate that it would not be possible to achieve a large sensitivity (true positive rate) without making several false positive assignments. For

example, in order to achieve an approximately 90% accuracy for the prediction of m6A+ samples, one would expect a 20% accuracy for the m6a- samples (0.80 false positive rate), indicating that most of the m6a+ samples would be predicted correctly, but most of the m6a- samples would be misclassified. Hence, the multivariate model built by combining all the features was not strong enough to properly classify m6A presence or absence.

We then performed univariate analysis using logistic regression to determine the association of individual feature with m6A positivity. Further we selected the top ranking marks according to the odds ratio determined by univariate regression [Table 3]. Interestingly, RRACT (binding motif of m6A modification) was not good in predicting m6a+ and many other variables performed better. Next we examined the association of these top-ranking marks with the m6A peaks using various plots.

Marks	A) Neg. Marker	B) Pos. Marker	C) Total Numbers	D) RR	E) OR	F) UL	G) LL
ZNF274	**0.47**	**0.93**	**284.00**	1.98	14.91	23.49	9.47
E2F6	**0.45**	**0.74**	**3905.00**	1.64	3.43	3.69	3.18
CGI	**0.43**	**0.72**	**7604.00**	1.66	3.30	3.48	3.13
CTBP2	**0.47**	**0.73**	**888.00**	1.57	3.13	3.63	2.69
CHD1	**0.47**	**0.72**	**1464.00**	1.55	2.95	3.31	2.63
KDM4A	**0.42**	**0.68**	**11301.00**	1.62	2.90	3.03	2.77
RBBP5	**0.46**	**0.71**	**2931.00**	1.55	2.87	3.12	2.65

POLR2ApS5	**0.46**	**0.67**	**3149.00**	1.47	2.44	2.64	2.26
POLR2A	**0.46**	**0.67**	**3882.00**	1.47	2.44	2.62	2.28
EZH2	**0.47**	**0.68**	**608.00**	1.45	2.43	2.88	2.05
SAP30	**0.43**	**0.65**	**10783.00**	1.51	2.42	2.53	2.32
TAF7	**0.46**	**0.68**	**1933.00**	1.45	2.40	2.64	2.18
TAF1	**0.46**	**0.67**	**3428.00**	1.46	2.40	2.58	2.23
CREB1	**0.46**	**0.67**	**3074.00**	1.45	2.36	2.55	2.19
GTF2F1	**0.47**	**0.67**	**805.00**	1.44	2.35	2.72	2.02
H3K9ac	**0.40**	**0.60**	**19896.00**	1.53	2.33	2.41	2.24
RFX5	**0.47**	**0.67**	**141.00**	1.43	2.32	3.29	1.63
SIN3A	**0.46**	**0.66**	**3475.00**	1.44	2.31	2.49	2.15
TBP	**0.46**	**0.66**	**2776.00**	1.44	2.31	2.50	2.13
BACH1	**0.47**	**0.67**	**945.00**	1.43	2.28	2.61	1.99
REST	**0.47**	**0.67**	**622.00**	1.42	2.26	2.68	1.91
JUND	**0.47**	**0.66**	**1141.00**	1.42	2.25	2.55	1.99
PHF8	**0.41**	**0.61**	**16346.00**	1.48	2.23	2.32	2.15
RRACT	**0.48**	**0.51**	**39664.00**	1.05	1.1	1.14	1.06

Table 3: Variable association with m6A. The table reports, for each of the marks: A) the proportion of m6A+ samples in the group of samples with negative marker (mark not present), B) the proportion of m6A+ samples in the group of samples with positive marker (mark present), C) Total number of occurrences of these marks at the m6A positive regions, D) RR is defined as the ratio of the two proportions (proportion of m6A+ samples in the positive marker

group divided by the proportion of m6A+ samples in the negative marker group), E) The OR is defined as the ratio of the odds for m6A+ in these groups, F)UL and G) LL denotes the upper and lower limit of the 95% confidence intervals for the OR respectively (estimated with univariate logistic regression, where the outcome is m6A positivity and the covariate is the marker). OR=1 indicate that the probability of being m6A+ is the same in the group of marker positive and negative samples. OR>1 indicate that marker positive samples have a larger probability of being m6A+ compared to marker negative samples, OR<1 that marker negative samples have larger probability of being m6A+ compared to marker positive samples. Large confidence intervals are observed for markers that have a low positivity proportion.

5.4.1 Combinatorial association and overlap with TSS regions

To investigate the patterning of our top rankings features with RNA methylation events, we generated heatmaps of the m6A+ regions marked by the top rankings features (determined by odds ratio). Separate heatmaps were drawn for features having extensive overlap (associated with thousands of m6A peaks) and those with few overlaps (associated with hundreds of m6A peaks) with m6A peaks. Further, these regions are annotated according to overlaps with the TSS: leftmost TSS, internal TSS, no TSS and 3′ UTRs. The region defined for overlap was m6A peaks summit (+-144b) same as the one used for LASSO model building.

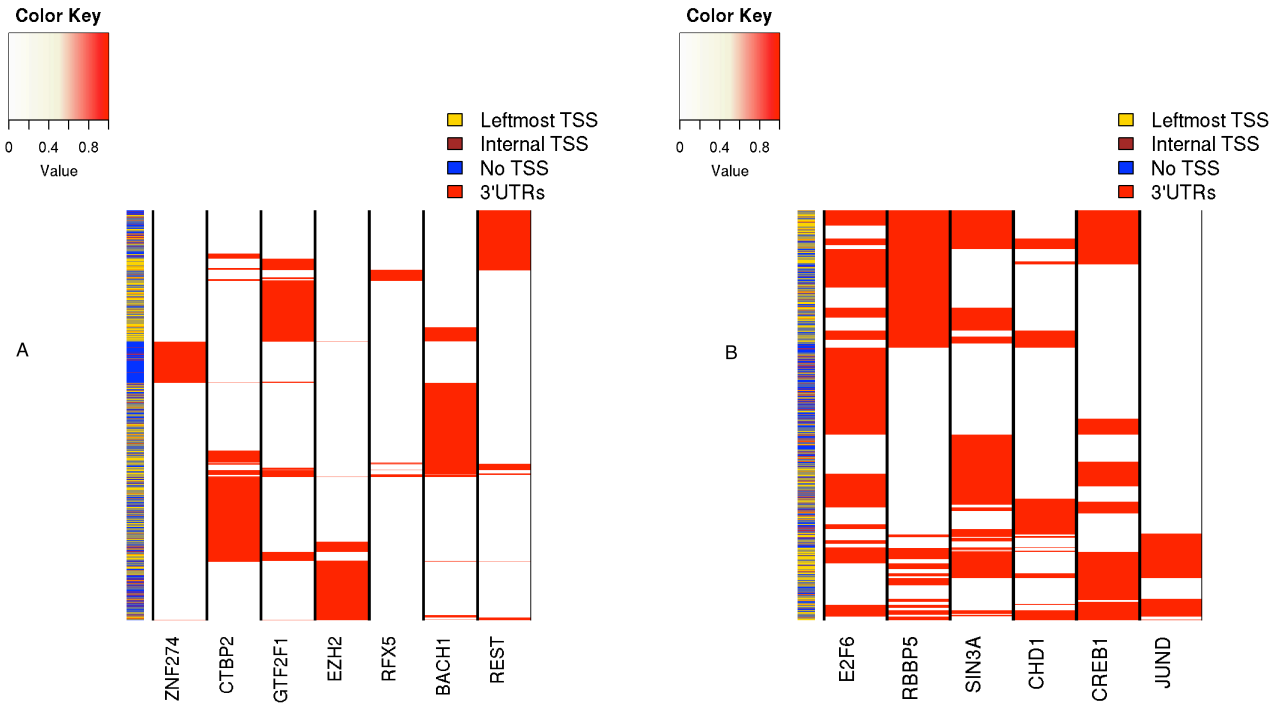


Figure 25: Heatmaps of top ranking features.

A) Patterns of selected marks with few hundred events corresponding with m6A peaks. B) Patterns of selected marks with thousands of events corresponding with m6A peaks.

The heatmap for features having few overlaps with m6A peaks (associated with hundreds of m6A peaks) displays that the binding of these marks is mutually exclusive in genome (in regions marked by m6A) [**Figure 25A**]. Similarly, most of the features of the other group (associated with thousands of m6A peaks) also show mutual exclusivity in binding albeit not of same degree as the other [**Figure 25B**]. Within our top rankings features were a set of marks associated with Pol2 machinery: POLR2Ap5S, POLR2A, TAF7, TAF1 and TBP. The heatmap of these features show nice overlap in their binding pattern with each other [**Figure 26A**]. Although a majority of the overlap exists in regions around the TSS (where it is mostly expected), interestingly a good number of these overlap exists in regions away from TSS as well. All these heatmaps (the three

different groups) show that the majority overlaps of these features are at the genomic regions closer to the TSS. A protein-protein interaction map displaying interaction between components of Pol2 and the RNA methylation machineries was obtained using STRING database [**Figure 26B**]. It displays that the major component of RNA methylation machineries METTL3 and METTL4 (writers of m6A modification) shows interaction with Pol2 machinery.

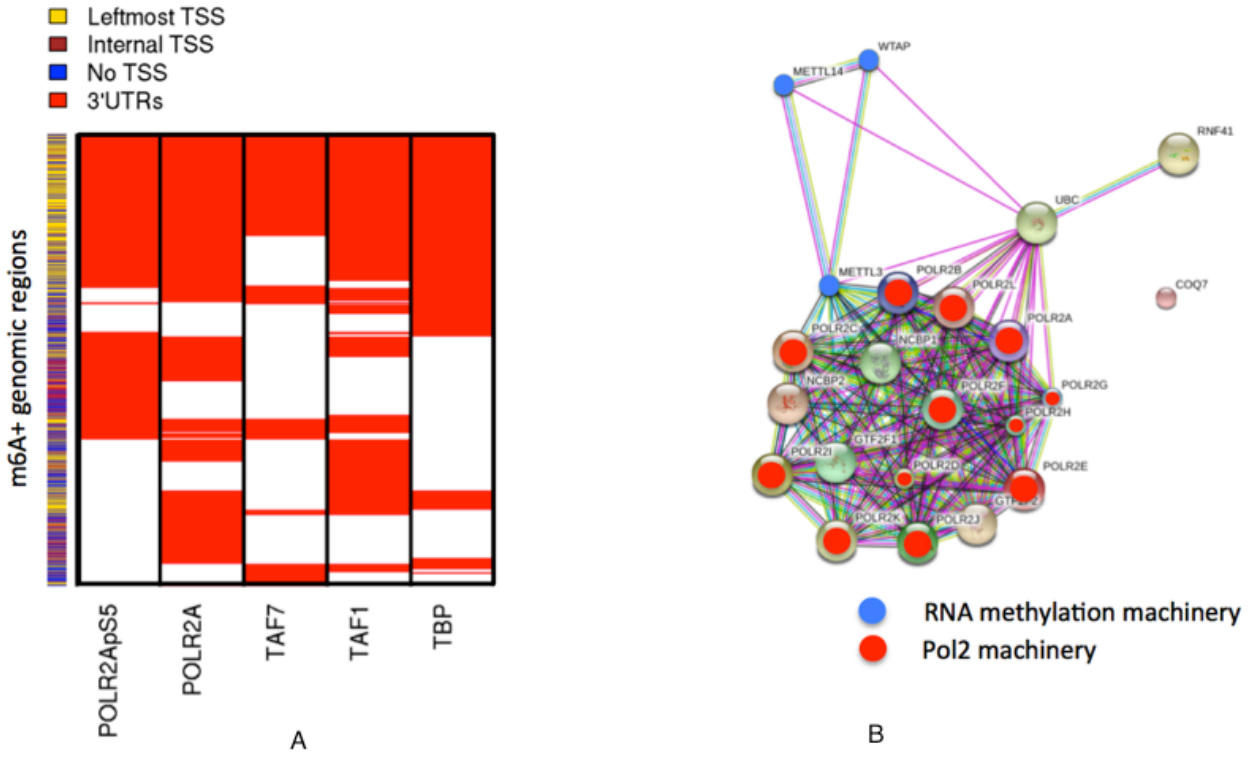
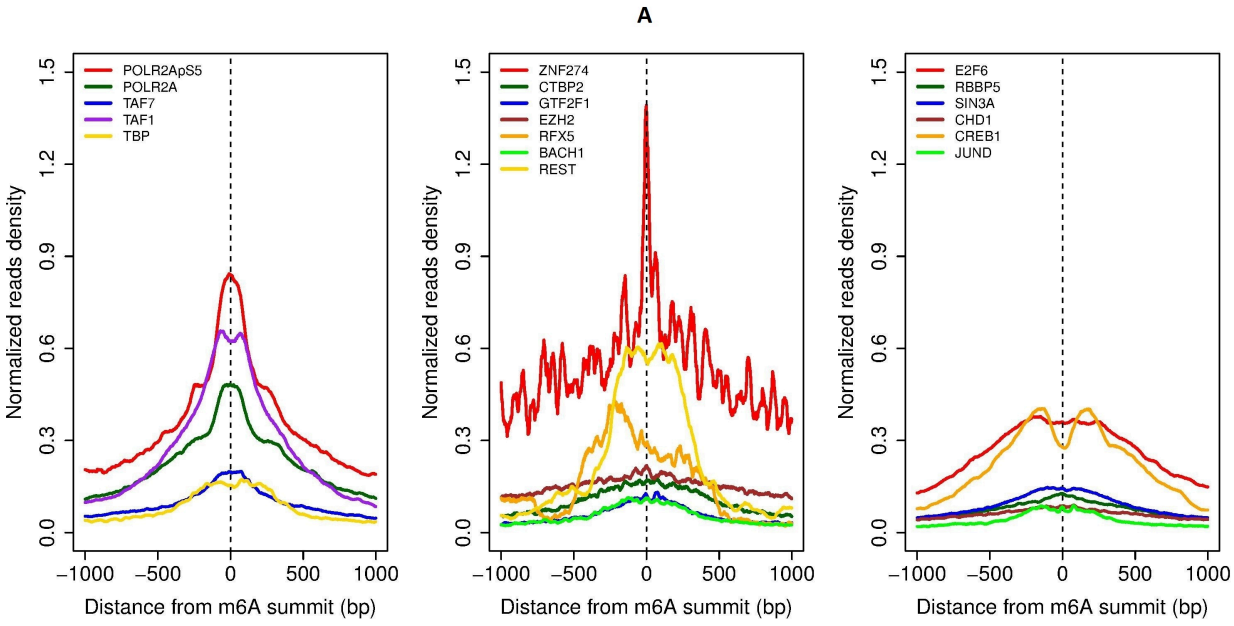


Figure 26: Pol2 and RNA methylation machinery.

A) Patterns of components of the Pol2 machinery corresponding with m6A peaks. B) Protein-protein interaction between components of the Pol2 and the RNA methylation machineries.

To further examine the spatial association of the top ranking marks and m6A peaks, we plotted reads density of these features around (+-1kb) of m6A peaks summit [**Figure 27A**]. Most of the marks show increase in the enrichment (reads coverage) around the m6A peak summit while some remain uniformly enriched throughout the defined region. ZNF274 shows sharp increase in enrichment at the m6A peak summit and also marks of the Pol2 machinery have nice overlap in the enrichment at these regions. These plots quantitatively confirm the prevalence of these marks around the m6A-binding region (summit) as previously shown by the heatmap analysis. A similar pattern emerged at the regions of m6A peaks distal from TSS implying this association is not driven by the proximity to TSS [**Figure 27B**].



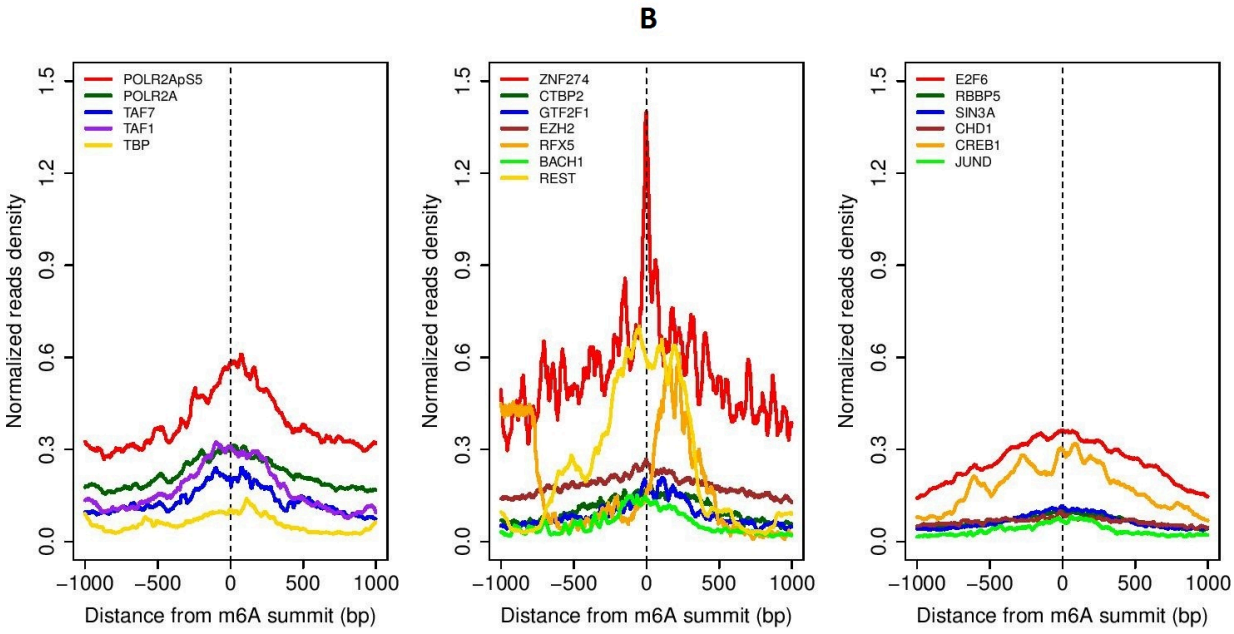


Figure 27: Spatial association between selected marks and m6A peaks.

A) Spatial relationship between selected marks and all m6A peaks. B) Spatial relationship between selected marks and m6A peaks distal from TSS.

5.5 Marks associated with transcriptional repression

Many studies have described the influence of m6A mark on nearly all aspects of RNA metabolism [127]. The two mechanisms currently ascribed to define the function of m6A mark are: a) recruitment of proteins or b) bringing conformational changes in RNA[163]. The underlying theme behind both mechanisms is m6A binding altering the RNA-protein interaction. The binding of m6A could either make an effector protein bind to those transcripts or could alter sequence specific binding of other proteins in its vicinity. One such example is YTHDF protein (which binds to m6A) mediated decay of transcripts that requires methylation of RNA and could be dynamically regulated by methylation or demethylation mechanisms[168]. Furthermore, the levels of m6A modification have been inversely correlated with mRNA stability implying its role as a transcriptional repressor[169]. Additional reader proteins might also exist that affect various components of RNA metabolism. So it was interesting to find many top rankings features having roles in biological phenomenon such as transcriptional repression, splicing, chromatin modification and RNA Pol II machinery [**Figure 28**]. We discuss below some of these interesting features.

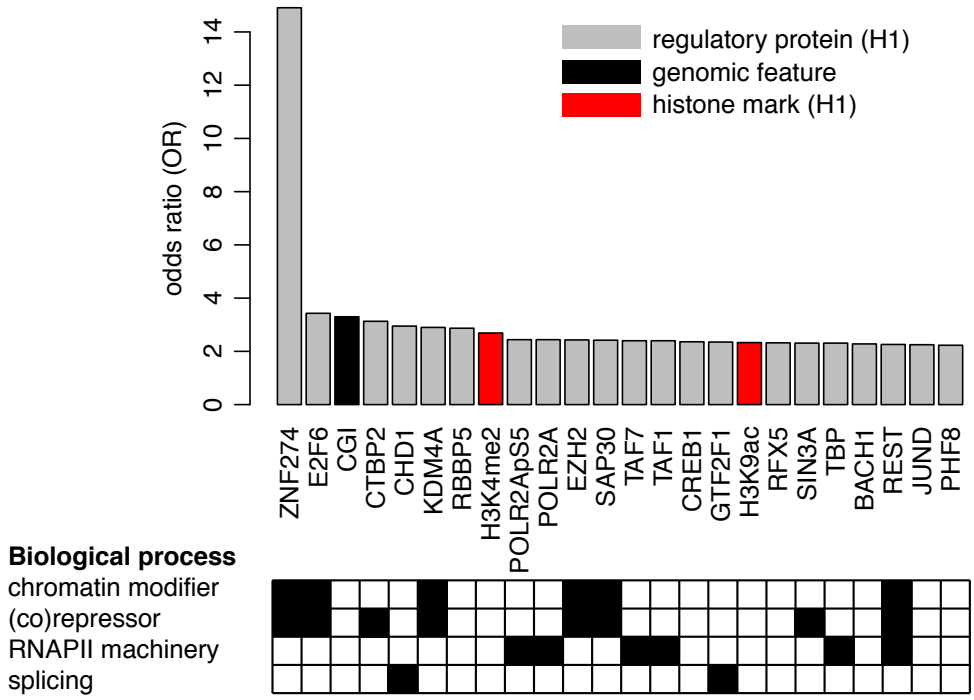


Figure 28: Biological process of top rankings marks.

Each bar represents odds ratio of association of each feature with m6A mark. The top features identified by odds ratio are enriched in biological processes to which the role of m6A has also been assigned.

Zinc Finger Protein 274 (ZNF274) has highest association with m6A+ peaks. Although it has only 284 occurrences (among the total tested regions), 93% of these regions co-occur with m6A. We also explore this association in HepG2 cells [**Figure 29**] and HeLa cells and observed similar overlap between m6A and ZNF274. ZNF274 is suggested to play the role of a transcriptional repressor. H3K9me3 is one of the histone modification associated with gene silencing especially zinc finger genes[170]. SETDB1 and G9a histone methyltransferases are involved in mediating this process[170]. SETDB1 is recruited to specific genomic locations via interaction with the co-repressor TRIM28 (KAP1), which is in turn recruited to the genome via interaction with ZNF274

[170]. Eventually SETDB1 promotes the increase of H3K9me3 at the 3' ends of zinc finger genes leading to their transcription repression[170].

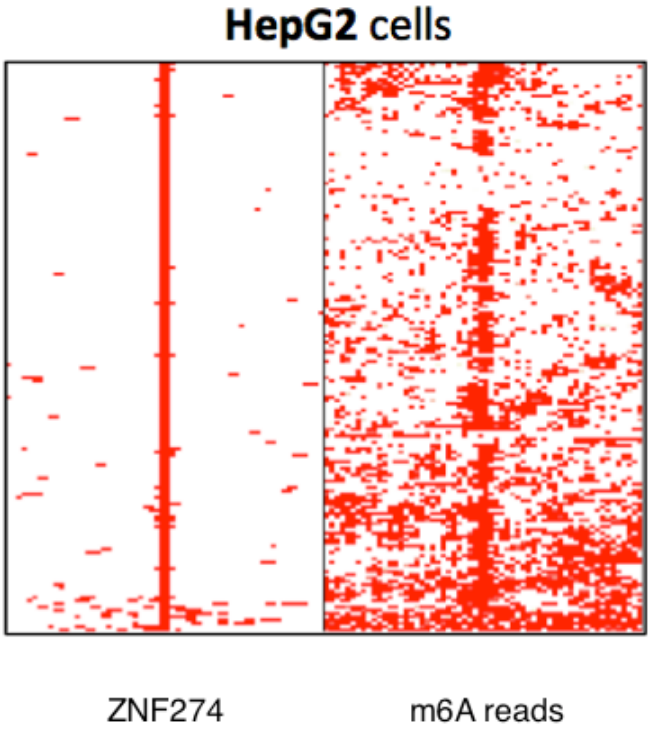


Figure 29: Overlap of ZNF274 with m6A in HepG2 cells.

The figure displays density of ZNF274 reads around the region of m6A peaks summit in HepG2 cells. It represents association of ZNF274 at the m6A peak sites.

E2F6 is a transcriptional repressor [171] binding to 3905 regions in H1, co-occurring with m6A peaks in 74% of cases. It regulates a subset of E2F-dependent genes whose products are required for entry into the cell cycle (but not for normal cell cycle progression). It may silence expression via the recruitment of a chromatin remodeling complex containing histone H3K9 methyltransferase activity[172]. Similarly, CTBP2 is a co-repressor targeting diverse transcription regulators. In H1, it binds to 888 regions, co-occurring with m6A peaks in 73% of cases. CHD1 is a chromatin-remodeling factor that function as substrate recognition component of the transcription regulatory histone

acetylation (HAT) complex SAGA. In H1, it binds to 1464 regions, co-occurring with m6A peaks in 72% of cases. It functions to modulate the efficiency of pre-mRNA splicing in part through physical bridging of spliceosomal components to H3K4me3[173].

KDM4A also known as JMJD2A demethylates H3K9me3 and H3K36me3. In H1, it binds to 11301 regions co-occurring with m6A peaks in 68% of cases. The protein belongs to a family that includes the ALKBH5 and the FTO RNA demethylases [174]. It participates in transcriptional repression of ASCL2 gene (an imprinted gene essential for proper placental development) by targeting the N-CoR complex[175]. RBBP5 is a ubiquitously expressed nuclear protein that binds directly to retinoblastoma protein thereby regulating cell proliferation. In mouse embryonic stem (ES) cells it plays a crucial role in differentiation along the neural lineage by regulating gene induction and H3K4 methylation at key developmental loci[176].

SAP30 is a component of histone deacetylase complex. "It is involved in the functional recruitment of the Sin3-histone deacetylase complex (HDAC) to a specific subset of N-CoR corepressor complexes"[177]. SIN3A acts as a transcriptional co-repressor utilized by the Mad-Max family of DNA-binding transcriptional repressors[178]. It is also recruited by PER complex to repress Per1 transcription thereby modulating circadian rhythm[179]. Similarly various circadian RNAs contain the m6A modification and inhibition of this modification increase nuclear time exit[180]. SIN3A is also recruited by REST to carry out transcriptional repression of neuronal genes[181]. REST is a transcriptional repressor binding neuron-restrictive silencer element (NRSE) and repressing neuronal gene transcription in non-neuronal cells[181]. In H1, it binds to 622 regions, co-occurring with m6A peaks in 67% of cases. It

restricts the expression of neuronal genes by associating with two distinct corepressors, mSin3 and CoREST, which in turn recruit histone deacetylase to the promoters of REST-regulated genes[182]. In addition to the above mechanism, it also mediates repression by recruiting the BHC complex at RE1/NRSE sites [183].

Out of these features currently follow up experimental validation (in collaboration with Dr. Bruno Amati Lab) is being carried out on ZNF274 by knockdown studies. ZNF274 has the highest association with the m6A marks and its role as a transcriptional repressor has been well established. The knockdown experiment will determine if lowering of its expression reduces the m6A methylation levels in the targeted mRNAs.

Discussion

The field of epigenomics is rapidly progressing with new information coming out from multiple studies and various international consortia. The massive amount of genome-wide data accumulated presents computational challenges owing to the complex nature of relationship of the various components of epigenetics. These players act in a coordinated and combinatorial way to modulate various biological processes. To understand the complexity of cross talk of these components it is imperative to develop computational methods that can perform integrative analysis of all these datasets though conscious of the biases of technologies profiling them. There are many tools available catering specifically to different technologies but a common platform for integrative analysis was missing. Considering this lacuna we intended to develop specific methods that can efficiently handle whole genome base-resolution DNA methylation datasets and perform integrative analysis with other epigenomics component.

The methylPipe and compEpiTools companion libraries offer a comprehensive system for the integrative analysis of heterogeneous epigenomics data types. methylPipe provides a set of classes, methods and functions that are tailored to DNA methylation high-throughput data, while accommodating data highly different in terms of resolution and genome coverage. To our knowledge, methylPipe is the first software package allowing the analysis and manipulation of multiple WGBS experiments while also being compatible with targeted or low-resolution DNA methylation experiments. Furthermore, compEpiTools includes a series of methods and functions that are

commonly used in the integrative analysis of epigenomics, genomics and regulatory datasets. Importantly, compEpiTools is compatible with methylPipe classes thus allowing an effortless combination of these data with other epigenomics data. Lower-level versions of few of these functionalities are already available albeit dispersed in various Bioconductor packages, such as the routines for counting reads. For these tasks methylPipe and compEpiTools provide simplified and more homogeneous access to lower-level routines, adding an extensive number of new functionalities for DNA methylation and other epigenomics and regulatory data types. Altogether, this suite of packages provides a clear reference entry-point for scientists focusing on the analysis of epigenomics data. This set of tools is currently being successfully used to build pipelines for the most common omics data types. Even more importantly, in our hands this approach is proving to be an excellent resource to effectively provide to experimental scientists with very basic R skills a complete toolkit for the comprehensive analysis of their own generated data. In conclusion, the Bioconductor-compliant methylPipe and compEpiTools packages provide a comprehensive suite of tools for the integrative analysis of epigenomics data, covering most of the functionalities commonly required in the joint analysis of DNA methylation and epigenomics data.

We applied these methods to identifying novel tumor suppressor genes. New tumor suppressor genes identified in this project shall shed new light on the genes, pathways and mechanisms that are critical for Myc-induced lymphomas. The Eμ-myc model is ideally suited not only for the gene discovery phase of this project but also for the dissection of the cellular and molecular events that underlie tumor progression. The relevance of the genes and pathways identified will be addressed in human samples. Finally, an important translational objective is to understand whether and how we may

specifically re-activate altered tumor suppressor pathways through pharmacological approaches in order to curb tumor progression.

Epigenetic modifications on DNA and histones are subjected to reversible regulation affecting cell differentiation and development. Recently, the writers and erasers of N6-methyladenosine (m6A) the most abundant RNA modification have been identified illustrating the possibility of its regulation in different biological contexts. Its role in various aspects of RNA metabolism has already been illustrated. However, no study has looked at the association of this post-transcriptional modification with epigenomics components and regulatory features. In our study we investigated the possible associations between RNA methylation and epigenomics or regulatory proteins by integrative analysis of publically available datasets of some of the cell lines that are profiled for RNA methylation.

First we evaluated the association of RNA methylation peaks (m6A) to the corresponding pattern on the methylation of DNA (5-methyl cytosine, 5mC) in MEF cells. The statistical results show low correlation (spearman correlation 0.35) between the two features but the patterning was remarkably similar at co-occurring regions especially at the exons. Since, both m6A and 5mC are implicated in splicing processes we hypothesize that genomic enrichment of 5mC within the same exons marked by m6A in the corresponding transcripts could be relevant for proper RNA splicing. The low resolution MeDIP-seq data could not clarify if high density of 5mC events in correspondence of m6A marks is due to an increase in true 5mC events or an increase in potential methylation site (increased CpG density). To shed light into it we explored DNA methylation patterns at the base-resolution in correspondence of m6A peaks in human H1 cells. The base-resolution analysis of 5mCs in human reveals incomplete

DNA methylation of the CpGs which might suggest m6A binding sites in transcripts as interaction spots for specific regulatory proteins on the genome. Importantly, depletion was specific for exons and 3'UTRs associated with m6A peaks but not in random exons and 3'UTRs devoid of that mark. Since the predominant consensus sequence of m6A enrichment is RRACU, we further explored the genomic prevalence of 5mC in correspondence of the C within the motif and the bases surrounding the motif occurrences. The methylation levels of CpGs are remarkably depleted at RRACT sites associated with m6A peaks in correspondence to those not associated with m6A peaks.

To investigate the association of (epi) genomic and regulatory features with m6A we performed multivariate LASSO analysis and univariate logistic regression analysis. The model built using LASSO was not strong enough to classify the m6A presence or absence. Hence, we focused our attention to univariate analysis to identify individual features associated with m6A peaks. Among them, DNA methylation was the top ranking predictor of m6A (associated with lower level of DNA methylation). On the contrary RRACU motif was not among the top rankings features. We then investigated the combinatorial association of the top rankings features determined by odds ratio with m6A marked regions. It showed mutually exclusive binding patterns of a number of features apart from the components of Pol2 machinery that were showing nicely overlapping binding patterns. Further examination of the spatial association of these features (using reads density) and m6A peaks quantitatively confirmed these findings.

Numerous studies have described the role of m6A mark on several RNA metabolism processes primarily by altering the RNA-protein interaction. The important challenge ahead is to elucidate the underlying mechanism and players carrying out this role. The first step towards it would be to identify additional reader or participatory

proteins that affects this mechanism. So in light of this it was interesting to find many top rankings features identified by us having roles in biological phenomenon such as transcriptional repression, splicing and chromatin modification. We are conducting follow up experimental validation on ZNF274 (a transcriptional repressor) that has the highest association with the m6A marks. We believe our study has provided a preliminary insight into the association of epigenetics modifications and RNA methylation and anticipate that the follow up studies will provide stimulus to this exciting new realm of research.

References

1. Esteller M: **Epigenetics in evolution and disease**. *Lancet* 2008, **372**(December 2008):S90–S96.

2. Rideout WM, Eggan K, Jaenisch R: **Nuclear cloning and epigenetic reprogramming of the genome.** *Science* 2001, **293**:1093–1098.

3. Fraga MF, Ballestar E, Paz MF, Ropero S, Setien F, Ballestar ML, Heine-Suñer D, Cigudosa JC, Urioste M, Benitez J, Boix-Chornet M, Sanchez-Aguilera A, Ling C, Carlsson E, Poulsen P, Vaag A, Stephan Z, Spector TD, Wu Y-Z, Plass C, Esteller M: **Epigenetic differences arise during the lifetime of monozygotic twins.** *Proc Natl Acad Sci U S A* 2005, **102**:10604–10609.

4. Yamamizu K, Piao Y, Sharov A a., Zsiros V, Yu H, Nakazawa K, Schlessinger D, Ko MSH: **Identification of transcription factors for lineage-specific ESC differentiation**. *Stem Cell Reports* 2013, **1**:545–559.

5. Cantone I, Fisher AG: **Epigenetic programming and reprogramming during development.** *Nat Struct Mol Biol* 2013, **20**:282–9.

6. Ohinata Y, Ohta H, Shigeta M, Yamanaka K, Wakayama T, Saitou M: **A Signaling Principle for the Specification of the Germ Cell Lineage in Mice**. *Cell* 2009, **137**:571–584.

7. Portela A, Esteller M: **Epigenetic modifications and human disease.** *Nat Biotechnol* 2010, **28**:1057–1068.

8. Ruthenburg AJ, Li H, Patel DJ, Allis CD: **Multivalent engagement of chromatin modifications by linked binding modules.** *Nat Rev Mol Cell Biol* 2007, **8**:983–994.

9. Martin C, Zhang Y: **Mechanisms of epigenetic inheritance**. *Curr Opin Cell Biol* 2007, **19**:266–272.

10. Tan M, Luo H, Lee S, Jin F, Yang JS, Montellier E, Buchou T, Cheng Z, Rousseaux S, Rajagopal N, Lu Z, Ye Z, Zhu Q, Wysocka J, Ye Y, Khochbin S, Ren B, Zhao Y: **Identification of 67 histone marks and histone lysine crotonylation as a new type of histone modification**. *Cell* 2011, **146**:1016–1028.

11. Zhang Y, Reinberg D: **Transcription regulation by histone methylation: Interplay between different covalent modifications of the core histone tails**. *Genes Dev* 2001, **15**:2343–2360.

12. Verdone L, Agricola E, Caserta M, Di Mauro E: **Histone acetylation in gene regulation.** *Brief Funct Genomic Proteomic* 2006, **5**:209–221.

13. Lee KK, Workman JL: **Histone acetyltransferase complexes: one size doesn't fit all.** *Nat Rev Mol Cell Biol* 2007, **8**:284–295.

14. Leipe DD, Landsman D: **Histone deacetylases, acetoin utilization proteins and acetylpolyamine amidohydrolases are members of an ancient protein superfamily**. *Nucleic Acids Res* 1997, **25**:3693–3697.

15. Rose NR, Klose RJ: **Understanding the relationship between DNA methylation and histone lysine methylation**. *Biochim Biophys Acta - Gene Regul Mech* 2014, **1839**:1362–1372.

16. Schwartz YB, Pirrotta V: **Polycomb complexes and epigenetic states**. *Curr Opin Cell Biol* 2008, **20**:266–273.

17. Golbabapour S, Majid NA, Hassandarvish P, Hajrezaie M, Abdulla MA, Hadi a H a: **Gene silencing and Polycomb group proteins: an overview of their structure, mechanisms and phylogenetics.** *OMICS* 2013, **17**:283–96.

18. Narlikar GJ, Sundaramoorthy R, Owen-Hughes T: **Mechanisms and Functions of ATP-Dependent Chromatin-Remodeling Enzymes**. *Cell* 2013, **154**:490–503.

19. Smallwood A, Estève PO, Pradhan S, Carey M: **Functional cooperation between HP1 and DNMT1 mediates gene silencing**. *Genes Dev* 2007, **21**:1169–1178.

20. Fuks F, Hurd PJ, Deplus R, Kouzarides T: **The DNA methyltransferases associate with HP1 and the SUV39H1 histone methyltransferase**. *Nucleic Acids Res* 2003, **31**:2305–2312.

21. Tardat M, Albert M, Kunzmann R, Liu Z, Kaustov L, Thierry R, Duan S, Brykczynska U, Arrowsmith CH, Peters AHFM: **Cbx2 Targets PRC1 to Constitutive Heterochromatin in Mouse Zygotes in a Parent-of-Origin-Dependent Manner**. *Mol Cell* 2015, **58**:157–171.

22. Bártová E, Krejcí J, Harnicarová A, Galiová G, Kozubek S: **Histone modifications and nuclear architecture: a review.** *J Histochem Cytochem* 2008, **56**:711–721.

23. Bernstein BE, Mikkelsen TS, Xie X, Kamal M, Huebert DJ, Cuff J, Fry B, Meissner A, Wernig M, Plath K, Jaenisch R, Wagschal A, Feil R, Schreiber SL, Lander ES: **A Bivalent Chromatin Structure Marks Key Developmental Genes in Embryonic Stem Cells**. *Cell* 2006, **125**:315–326.

24. Ooi SKT, Qiu C, Bernstein E, Li K, Jia D, Yang Z, Erdjument-Bromage H, Tempst P, Lin S-P, Allis CD, Cheng X, Bestor TH: **DNMT3L connects unmethylated lysine 4 of histone H3 to de novo methylation of DNA.** *Nature* 2007, **448**:714–717.

25. Jia D, Jurkowska RZ, Zhang X, Jeltsch A, Cheng X: **Structure of Dnmt3a bound to Dnmt3L suggests a model for de novo DNA methylation.** *Nature* 2007, **449**:248–251.

26. Feldman N, Gerson A, Fang J, Li E, Zhang Y, Shinkai Y, Cedar H, Bergman Y: **G9a-mediated irreversible epigenetic inactivation of Oct-3/4 during early embryogenesis**. *Nat Cell Biol* 2006, **8**:188–194.

27. Epsztejn-Litman S, Feldman N, Abu-Remaileh M, Shufaro Y, Gerson A, Ueda J, Deplus R, Fuks F, Shinkai Y, Cedar H, Bergman Y: **De novo DNA methylation promoted by G9a prevents reprogramming of embryonically silenced genes.** *Nat Struct Mol Biol* 2008, **15**:1176–1183.

28. Jones PL, Jan Veenstra GC, Wade PA, Vermaak D, Kass SU, Landsberger N, Strouboulis J, Wolffe AP: **Methylated DNA and MeCP2 recruit histone deacetylase to repress transcription** . *Nat Genet* 1998, **19**:187–191.

29. Nan X, Ng HH, Johnson C a, Laherty CD, Turner BM, Eisenman RN, Bird a: **Transcriptional repression by the methyl-CpG-binding protein MeCP2 involves a histone deacetylase complex.** *Nature* 1998, **393**:386–389.

30. Hashimshony T, Zhang J, Keshet I, Bustin M, Cedar H: **The role of DNA methylation in setting up chromatin structure during development**. *Nat Genet* 2003, **34**:187–192.

31. Strahl BD, Allis CD: **The language of covalent histone modifications.** *Nature* 2000, **403**:41–45.

32. Morinière J, Rousseaux S, Steuerwald U, Soler-López M, Curtet S, Vitte A-L, Govin J, Gaucher J, Sadoul K, Hart DJ, Krijgsveld J, Khochbin S, Müller CW, Petosa C: **Cooperative binding of two acetylation marks on a histone tail by a single bromodomain.** *Nature* 2009, **461**:664–668.

33. Ruthenburg AJ, Li H, Milne T a., Dewell S, McGinty RK, Yuen M, Ueberheide B, Dou Y, Muir TW, Patel DJ, Allis CD: **Recognition of a mononucleosomal histone modification pattern by BPTF via multivalent interactions**. *Cell* 2011, **145**:692–706.

34. Hon G, Ren B, Wang W: **ChromaSig: A probabilistic approach to finding common chromatin signatures in the human genome**. *PLoS Comput Biol* 2008, **4**.

35. Ernst J, Kellis M: **ChromHMM: automating chromatin-state discovery and**

characterization. *Nat Methods* 2012, **9**:215–216.

36. Okano M, Bell DW, Haber D a., Li E: **DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development**. *Cell* 1999, **99**:247–257.

37. Goll MG, Bestor TH: **Eukaryotic cytosine methyltransferases.** *Annu Rev Biochem* 2005, **74**:481–514.

38. Jones P a, Liang G: **Rethinking how DNA methylation patterns are maintained.** *Nat Rev Genet* 2009, **10**:805–811.

39. Jackson-Grusby L, Beard C, Possemato R, Tudor M, Fambrough D, Csankovszki G, Dausman J, Lee P, Wilson C, Lander E, Jaenisch R: **Loss of genomic methylation causes p53-dependent apoptosis and epigenetic deregulation.** *Nat Genet* 2001, **27**:31–39.

40. Chen T, Hevi S, Gay F, Tsujimoto N, He T, Zhang B, Ueda Y, Li E: **Complete inactivation of DNMT1 leads to mitotic catastrophe in human cancer cells.** *Nat Genet* 2007, **39**:391–396.

41. Chen T, Ueda Y, Dodge JE, Wang Z, Li E: **Establishment and Maintenance of Genomic Methylation Patterns in Mouse Embryonic Stem Cells by Dnmt3a and Establishment and Maintenance of Genomic Methylation Patterns in Mouse Embryonic Stem Cells by Dnmt3a and Dnmt3b**. *Society* 2003, **23**:5594–5605.

42. Tahiliani M, Koh KP, Shen Y, Pastor W a, Bandukwala H, Brudno Y, Agarwal S, Iyer LM, Liu DR, Aravind L, Rao A: **Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1.** *Science* 2009, **324**:930–935.

43. Ito S, D'Alessio AC, Taranova O V, Hong K, Sowers LC, Zhang Y: **Role of Tet proteins in 5mC to 5hmC conversion, ES-cell self-renewal and inner cell mass specification.** *Nature* 2010, **466**:1129–1133.

44. Ito S, Shen L, Dai Q, Wu SC, Collins LB, Swenberg JA, He C, Zhang Y: **Tet Proteins Can Convert 5-Methylcytosine to 5-Formylcytosine and 5-Carboxylcytosine**. *Science (80-)* 2011, **333**:1300–1303.

45. Pastor W a, Aravind L, Rao A: **TETonic shift: biological roles of TET proteins in DNA demethylation and transcription.** *Nat Rev Mol Cell Biol* 2013, **14**:341–56.

46. Song CX, Szulwach KE, Dai Q, Fu Y, Mao SQ, Lin L, Street C, Li Y, Poidevin M, Wu H, Gao J, Liu P, Li L, Xu GL, Jin P, He C: **Genome-wide profiling of 5-formylcytosine reveals its roles in epigenetic priming**. *Cell* 2013, **153**:678–691.

47. Stroud H, Feng S, Morey Kinney S, Pradhan S, Jacobsen SE: **5-Hydroxymethylcytosine is associated with enhancers and gene bodies in human embryonic stem cells**. *Genome Biol* 2011, **12**:R54.

48. Smallwood S a, Tomizawa S-I, Krueger F, Ruf N, Carli N, Segonds-Pichon A, Sato S, Hata K, Andrews SR, Kelsey G: **Dynamic CpG island methylation landscape in oocytes and preimplantation embryos.** *Nat Genet* 2011, **43**:811–814.

49. Jones P a.: **Functions of DNA methylation: islands, start sites, gene bodies and beyond**. *Nat Rev Genet* 2012, **13**:484–492.

50. Meehan RR, Lewis JD, McKay S, Kleiner EL, Bird a P: **Identification of a mammalian protein that binds specifically to DNA containing methylated CpGs.** *Cell* 1989, **58**:499–507.

51. Medvedeva Y a, Khamis AM, Kulakovskiy I V, Ba-Alawi W, Bhuyan MSI, Kawaji H, Lassmann T, Harbers M, Forrest ARR, Bajic VB: **Effects of cytosine methylation on transcription factor binding sites.** *BMC Genomics* 2014, **15**:119.

52. Brandeis M, Frank D, Keshet I, Siegfried Z, Mendelsohn M, Nemes a, Temper V, Razin a, Cedar H: **Sp1 elements protect a CpG island from de novo methylation.** *Nature* 1994:435–438.

53. Wang H, Maurano MT, Qu H, Varley KE, Gertz J, Pauli F, Lee K, Canfield T, Weaver M,

Sandstrom R, Thurman RE, Kaul R, Myers RM, Stamatoyannopoulos J a.: **Widespread plasticity in CTCF occupancy linked to DNA methylation**. *Genome Res* 2012, **22**:1680–1688.

54. Koga Y, Pelizzola M, Cheng E, Krauthammer M, Sznol M, Ariyan S, Narayan D, Molinaro AM, Halaban R, Weissman SM: **Genome-wide screen of promoter methylation identifies novel markers in melanoma.** *Genome Res* 2009, **19**:1462–70.

55. Taberlay PC, Kelly TK, Liu CC, You JS, De Carvalho DD, Miranda TB, Zhou XJ, Liang G, Jones P a.: **Polycomb-repressed genes have permissive enhancers that initiate reprogramming**. *Cell* 2011, **147**:1283–1294.

56. Mohandas T, Sparkes RS, Shapiro LJ: **Reactivation of an inactive human X chromosome: evidence for X inactivation by DNA methylation**. *Sci* 1981, **211** (4480):393–396.

57. Allen RC, Zoghbi HY, Moseley a B, Rosenblatt HM, Belmont JW: **Methylation of HpaII and HhaI sites near the polymorphic CAG repeat in the human androgen-receptor gene correlates with X chromosome inactivation.** *Am J Hum Genet* 1992, **51**:1229–1239.

58. Han H, Cortez CC, Yang X, Nichols PW, Jones P a., Liang G: **DNA methylation directly silences genes with non-CpG island promoters and establishes a nucleosome occupied promoter**. *Hum Mol Genet* 2011, **20**:4299–4310.

59. Farthing CR, Ficz G, Ng RK, Chan CF, Andrews S, Dean W, Hemberger M, Reik W: **Global mapping of DNA methylation in mouse promoters reveals epigenetic reprogramming of pluripotency genes**. *PLoS Genet* 2008, **4**.

60. Lou S, Lee H-M, Qin H, Li J-W, Gao Z, Liu X, Chan LL, Lam V, So W-Y, Wang Y, Lok S, Wang J, Ma R, Tsui S, Chan J, Chan T-F, Yip KY: **Whole-genome bisulfite sequencing of multiple individuals reveals complementary roles of promoter and gene body methylation in transcriptional regulation.** *Genome Biol* 2014, **15**:408.

61. Su J, Shao X, Liu H, Liu S, Wu Q, Zhang Y: **Genome-wide dynamic changes of DNA methylation of repetitive elements in human embryonic stem cells and fetal fibroblasts**. *Genomics* 2012, **99**:10–17.

62. Suter CM, Martin DI, Ward RI: **Hypomethylation of L1 retrotransposons in colorectal cancer and adjacent normal tissue**. *Int J Colorectal Dis* 2004, **19**:95–101.

63. Negrini S, Gorgoulis VG, Halazonetis TD: **Genomic instability--an evolving hallmark of cancer.** *Nat Rev Mol Cell Biol* 2010, **11**:220–228.

64. Kim M, Trinh BN, Long TI, Oghamian S, Laird PW: **Dnmt1 deficiency leads to enhanced microsatellite instability in mouse embryonic stem cells**. *Nucleic Acids Res* 2004, **32**:5742–5749.

65. Elliott EN, Sheaffer KL, Schug J, Stappenbeck TS, Kaestner KH: **Dnmt1 is essential to maintain progenitors in the perinatal intestinal epithelium**. *Development* 2015:2163–2172.

66. Yoder J a., Walsh CP, Bestor TH: **Cytosine methylation and the ecology of intragenomic parasites**. *Trends Genet* 1997, **13**:335–340.

67. Maloisel L, Rossignol JL: **Suppression of crossing-over by DNA methylation in Ascobolus**. *Genes Dev* 1998, **12**:1381–1389.

68. Bestor TH, Tycko B: **Creation of genomic methylation patterns**. *Nat Genet* 1996, **12**:363–367.

69. Loh Y-H, Wu Q, Chew J-L, Vega VB, Zhang W, Chen X, Bourque G, George J, Leong B, Liu J, Wong K-Y, Sung KW, Lee CWH, Zhao X-D, Chiu K-P, Lipovich L, Kuznetsov VA, Robson P, Stanton LW, Wei C-L, Ruan Y, Lim B, Ng H-H: **The Oct4 and Nanog transcription network regulates pluripotency in mouse embryonic stem cells**. *Nat Genet* 2006, **38**:431–440.

70. Hattori N, Nishino K, Ko Y-G, Hattori N, Ohgane J, Tanaka S, Shiota K: **Epigenetic control of mouse Oct-4 gene expression in embryonic stem cells and trophoblast stem cells.** *J Biol Chem* 2004, **279**:17063–17069.

71. Watanabe D, Suetake I, Tada T, Tajima S: **Stage- and cell-specific expression of Dnmt3a and Dnmt3b during embryogenesis**. *Mech Dev* 2002, **118**:187–190.

72. Lister R, Pelizzola M, Dowen RH, Hawkins RD, Hon G, Tonti-Filippini J, Nery JR, Lee L, Ye Z, Ngo Q-M, Edsall L, Antosiewicz-Bourget J, Stewart R, Ruotti V, Millar a H, Thomson J a, Ren B, Ecker JR: **Human DNA methylomes at base resolution show widespread epigenomic differences.** *Nature* 2009, **462**:315–22.

73. Meissner A, Mikkelsen TS, Gu H, Wernig M, Hanna J, Sivachenko A, Zhang X, Bernstein BE, Nusbaum C, Jaffe DB, Gnirke A, Jaenisch R, Lander ES: **Genome-scale DNA methylation maps of pluripotent and differentiated cells.** *Nature* 2008, **454**:766–70.

74. Lister R, Mukamel EA, Nery JR, Urich M, Puddifoot CA, Johnson ND, Lucero J, Huang Y, Dwork AJ, Schultz MD, Yu M, Tonti-Filippini J, Heyn H, Hu S, Wu JC, Rao A, Esteller M, He C, Haghighi FG, Sejnowski TJ, Behrens MM, Ecker JR: **Global Epigenomic Reconfiguration During Mammalian Brain Development**. *Science (80-)* 2013, **341**:1237905–1237905.

75. Kulis M, Esteller M: **2 - DNA Methylation and Cancer**. In *Epigenetics and Cancer, Part A*. *Volume Volume 70*. Edited by Genetics ZH and TUBT-A in. Academic Press; 2010:27–56.

76. Huang TH, Perry MR, Laux DE: **Methylation profiling of CpG islands in human breast cancer cells.** *Hum Mol Genet* 1999, **8**:459–470.

77. Van Vlodrop IJH, Niessen HEC, Derks S, Baldewijns MMLL, Van Criekinge W, Herman JG, Van Engeland M: **Analysis of promoter CpG island hypermethylation in cancer: Location, location, location!** *Clin Cancer Res* 2011, **17**:4225–4231.

78. Curtin K, Slattery ML, Samowitz WS: **CpG island methylation in colorectal cancer: past, present and future.** *Patholog Res Int* 2011, **2011**:902674.

79. Ehrlich M: **DNA methylation in cancer: too much, but also too little.** *Oncogene* 2002, **21**:5400–5413.

80. Ehrlich M, Woods CB, Yu MC, Dubeau L, Yang F, Campan M, Weisenberger DJ, Long T, Youn B, Fiala ES, Laird PW: **Quantitative analysis of associations between DNA hypermethylation, hypomethylation, and DNMT RNA levels in ovarian tumors.** *Oncogene* 2006, **25**:2636–2645.

81. Ehrlich M: **DNA hypomethylation in cancer cells**. *Epigenomics* 2009, **1**:239–259.

82. Knudson a G: **Mutation and cancer: statistical study of retinoblastoma.** *Proc Natl Acad Sci U S A* 1971, **68**:820–823.

83. Drexler HG, Uphoff CC, Gaidano G, Carbone a: **Lymphoma cell lines: in vitro models for the study of HHV-8+ primary effusion lymphomas (body cavity-based lymphomas).** *Leuk Off J Leuk Soc Am Leuk Res Fund, UK* 1998, **12**:1507–1517.

84. Herman JG, Baylin SB: **Gene Silencing in Cancer in Association with Promoter Hypermethylation**. *N Engl J Med* 2003, **349**:2042–2054.

85. Fazzari MJ, Greally JM: **Epigenomics: beyond CpG islands.** *Nat Rev Genet* 2004, **5**:446–455.

86. Sproul D, Meehan RR: **Genomic insights into cancer-associated aberrant CpG island hypermethylation**. *Brief Funct Genomics* 2013, **12**:174–190.

87. Peng DF, Kanai Y, Sawada M, Ushijima S, Hiraoka N, Kitazawa S, Hirohashi S: **DNA methylation of multiple tumor-related genes in association with overexpression of DNA methyltransferase 1 (DNMT1) during multistage carcinogenesis of the pancreas**. *Carcinogenesis* 2006, **27**:1160–1168.

88. Arai E, Kanai Y, Ushijima S, Fujimoto H, Mukai K, Hirohashi S: **Regional DNA hypermethylation and DNA methyltransferase (DNMT) 1 protein overexpression in both renal tumors and corresponding nontumorous renal tissues**. *Int J Cancer* 2006, **119**:288–296.

89. Jin B, Tao Q, Peng J, Soo HM, Wu W, Ying J, Fields CR, Delmas AI, Liu X, Qiu J, Robertson KD: **DNA methyltransferase 3B (DNMT3B) mutations in ICF syndrome lead to altered epigenetic modifications and aberrant expression of genes regulating development, neurogenesis and immune function**. *Hum Mol Genet* 2008, **17**:690–709.

90. Barra V: **DNMT1 and Genomic Instability in Cancer**. *J Mol Genet Med* 2014, **8**:1–2.

91. Nishigaki M, Aoyagi K, Danjoh I, Fukaya M, Yanagihara K, Sakamoto H, Yoshida T, Sasaki H: **Discovery of aberrant expression of R-RAS by cancer-linked DNA hypomethylation in gastric cancer using microarrays**. *Cancer Res* 2005, **65**:2115–2124.

92. Esteller M: **Epigenetic gene silencing in cancer: The DNA hypermethylome**. *Hum Mol Genet* 2007, **16**:50–59.

93. Rodriguez J, Frigola J, Vendrell E, Risques R-A, Fraga MF, Morales C, Moreno V, Esteller M, Capellà G, Ribas M, Peinado M a: **Chromosomal instability correlates with genome-wide DNA demethylation in human primary colorectal cancers.** *Cancer Res* 2006, **66**:8462–9468.

94. Ehrlich M, Jiang G, Fiala E, Dome JS, Yu MC, Long TI, Youn B, Sohn O-S, Widschwendter M, Tomlinson GE, Chintagumpala M, Champagne M, Parham D, Liang G, Malik K, Laird PW: **Hypomethylation and hypermethylation of DNA in Wilms tumors.** *Oncogene* 2002, **21**:6694–6702.

95. Oda M, Glass JL, Thompson RF, Mo Y, Olivier EN, Figueroa ME, Selzer RR, Richmond T a, Zhang X, Dannenberg L, Green RD, Melnick A, Hatchwell E, Bouhassira EE, Verma A, Suzuki M, Greally JM: **High-resolution genome-wide cytosine methylation profiling with simultaneous copy number analysis and optimization for limited cell numbers.** *Nucleic Acids Res* 2009, **37**:3829–39.

96. Brinkman AB, Simmer F, Ma K, Kaan A, Zhu J, Stunnenberg HG: **Whole-genome DNA methylation profiling using MethylCap-seq.** *Methods* 2010, **52**:232–6.

97. Weber M, Davies JJ, Wittig D, Oakeley EJ, Haase M, Lam WL, Schübeler D: **Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed human cells.** *Nat Genet* 2005, **37**:853–62.

98. Laird PW: **Principles and challenges of genomewide DNA methylation analysis.** *Nat Rev Genet* 2010, **11**:191–203.

99. Ladd-Acosta C, J. Aryee M, Ordway JM, Feinberg AP: **Comprehensive High-Throughput Arrays for Relative Methylation (CHARM)**. *Curr Protoc Hum Genet* 2010(Table 1):780–790.

100. Brunner AL, Johnson DS, Kim SW, Valouev A, Reddy TE, Neff NF, Anton E, Medina C, Nguyen L, Chiao E, Oyolu CB, Schroth GP, Absher DM, Baker JC, Myers RM: **Distinct DNA methylation patterns characterize differentiated human embryonic stem cells and developing human fetal liver**. 2009:1044–1056.

101. Weber M, Hellmann I, Stadler MB, Ramos L, Pääbo S, Rebhan M, Schübeler D: **Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome.** *Nat Genet* 2007, **39**:457–466.

102. Zhang X, Yazaki J, Sundaresan A, Cokus S, Chan SW-L, Chen H, Henderson IR, Shinn P, Pellegrini M, Jacobsen SE, Ecker JR: **Genome-wide High-Resolution Mapping and Functional Analysis of DNA Methylation in Arabidopsis**. *Cell* 2006, **126**:1189–1201.

103. Cross SH, Charlton JA, Nan X, Bird AP: **Purification of CpG islands using a methylated DNA binding column**. *Nat Genet* 1994, **6**:236–244.

104. Hikoya B: **Review Discovery of bisul te-mediated cytosine conversion to uracil, the key reaction for DNA methylation analysis — A personal account**. 2008, **84**:2–11.

105. Pomraning KR, Smith KM, Freitag M: **Genome-wide high throughput analysis of DNA**

methylation in eukaryotes. *Methods* 2009, **47**:142–150.

106. Bibikova M, Lin Z, Zhou L, Chudin E, Garcia EW, Wu B, Doucet D, Thomas NJ, Wang Y, Vollmer E, Goldmann T, Seifart C, Jiang W, Barker DL, Chee MS, Floros J, Fan J: **High-throughput DNA methylation profiling using universal bead arrays**. 2006:1–11.

107. Bibikova M, Le J, Barnes B, Saedinia-Melnyk S, Zhou L, Shen R, Gunderson KL: **Genome-wide DNA methylation profiling using Infinium® assay.** *Epigenomics* 2009, **1**:177–200.

108. Bibikova M, Barnes B, Tsan C, Ho V, Klotzle B, Le JM, Delano D, Zhang L, Schroth GP, Gunderson KL, Fan JB, Shen R: **High density DNA methylation array with single CpG site resolution**. *Genomics* 2011, **98**:288–295.

109. Meissner A: **Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis**. *Nucleic Acids Res* 2005, **33**:5868–5877.

110. Lister R, O'Malley RC, Tonti-Filippini J, Gregory BD, Berry CC, Millar a. H, Ecker JR: **Highly Integrated Single-Base Resolution Maps of the Epigenome in Arabidopsis**. *Cell* 2008, **133**:523–536.

111. Rivera CM, Ren B: **Mapping human epigenomes.** *Cell* 2013, **155**:39–55.

112. Bock C, Lengauer T: **Computational epigenetics**. *Bioinformatics* 2008, **24**:1–10.

113. Bock C: **Analysing and interpreting DNA methylation data.** *Nat Rev Genet* 2012, **13**:705–19.

114. Robinson MD, Kahraman A, Law CW, Lindsay H, Nowicka M, Weber LM, Zhou X: **Statistical methods for detecting differentially methylated loci and regions**. *Front Genet* 2014, **5**(September):1–7.

115. Hebestreit K, Dugas M, Klein HU: **Detection of significantly differentially methylated regions in targeted bisulfite sequencing data**. *Bioinformatics* 2013, **29**:1647–1653.

116. Feng H, Conneely KN, Wu H: **A Bayesian hierarchical model to detect differentially methylated loci from single nucleotide resolution sequencing data**. *Nucleic Acids Res* 2014, **42**:1–11.

117. Dolzhenko E, Smith AD: **Using beta-binomial regression for high-precision differential methylation analysis in multifactor whole-genome bisulfite sequencing experiments.** *BMC Bioinformatics* 2014, **15**:215.

118. Park Y, Figueroa ME, Rozek LS, Sartor M a: **MethylSig: a whole genome DNA methylation analysis pipeline**. *Bioinformatics* 2014, **30**:1–8.

119. Giardine B, Riemer C, Hardison RC, Burhans R, Elnitski L, Shah P, Zhang Y, Blankenberg D, Albert I, Taylor J, Miller W, Kent WJ, Nekrutenko A: **Galaxy: A platform for interactive large-scale genome analysis**. *Genome Res* 2005, **15**:1451–1455.

120. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, Hornik K, Hothorn T, Huber W, Iacus S, Irizarry R, Leisch F, Li C, Maechler M, Rossini AJ, Sawitzki G, Smith C, Smyth G, Tierney L, Yang JYH, Zhang J: **Bioconductor: open software development for computational biology and bioinformatics.** *Genome Biol* 2004, **5**:R80.

121. Dowen RH, Pelizzola M, Schmitz RJ, Lister R, Dowen JM, Nery JR, Dixon JE, Ecker JR: **Widespread dynamic DNA methylation in response to biotic stress.** *Proc Natl Acad Sci U S A* 2012, **109**:E2183–91.

122. Forbes S a, Bindal N, Bamford S, Cole C, Kok CY, Beare D, Jia M, Shepherd R, Leung K, Menzies A, Teague JW, Campbell PJ, Stratton MR, Futreal PA: **COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer.** *Nucleic Acids Res* 2011, **39**(Database issue):D945–50.

123. Zindy F, Eischen CM, Randle DH, Kamijo T, Cleveland JL, Sherr CJ, Roussel MF: **Myc signaling via the ARF tumor suppressor regulates p53-dependent apoptosis**

and immortalization. *Genes Dev* 1998, **12**:2424–2433.

124. Schmitt C a, McCurrach ME, de Stanchina E, Wallace-Brodeur RR, Lowe SW: **INK4a/ARF mutations accelerate lymphomagenesis and promote chemoresistance by disabling p53.** *Genes Dev* 1999, **13**:2670–7.

125. Fu Y, Luo G-Z, Chen K, Deng X, Yu M, Han D, Hao Z, Liu J, Lu X, Doré LC, Weng X, Ji Q, Mets L, He C: **N6-Methyldeoxyadenosine Marks Active Transcription Start Sites in Chlamydomonas**. *Cell* 2015:879–892.

126. Dominissini D, Moshitch-Moshkovitz S, Salmon-Divon M, Amariglio N, Rechavi G: **Transcriptome-wide mapping of N(6)-methyladenosine by m(6)A-seq based on immunocapturing and massively parallel sequencing.** *Nat Protoc* 2013, **8**:176–89.

127. Fu Y, Dominissini D, Rechavi G, He C: **Gene expression regulation mediated through reversible m^{6}A RNA methylation.** *Nat Rev Genet* 2014, **15**:293–306.

128. Chen T, Hao Y-J, Zhang Y, Li M-M, Wang M, Han W, Wu Y, Lv Y, Hao J, Wang L, Li A, Yang Y, Jin K-X, Zhao X, Li Y, Ping X-L, Lai W-Y, Wu L-G, Jiang G, Wang H-L, Sang L, Wang X-J, Yang Y-G, Zhou Q: **m6A RNA Methylation Is Regulated by MicroRNAs and Promotes Reprogramming to Pluripotency**. *Cell Stem Cell* 2015, **16**:289–301.

129. Alarcon CR, Lee H, Goodarzi H, Halberg N, Tavazoie SF: **N6-methyladenosine marks primary microRNAs for processing**. *Nature* 2015, **519**:482–485.

130. Batista PJ, Molinie B, Wang J, Qu K, Zhang J, Li L, Bouley DM, Lujan E, Haddad B, Daneshvar K, Carter AC, Flynn RA, Zhou C, Lim K-S, Dedon P, Wernig M, Mullen AC, Xing Y, Giallourakis CC, Chang HY: **m6A RNA Modification Controls Cell Fate Transition in Mammalian Embryonic Stem Cells**. *Cell Stem Cell* 2014, **15**:707–719.

131. Geula S, Moshitch-Moshkovitz S, Dominissini D, Mansour AA, Kol N, Salmon-Divon M, Hershkovitz V, Peer E, Mor N, Manor YS, Ben-Haim MS, Eyal E, Yunger S, Pinto Y, Jaitin DA, Viukov S, Rais Y, Krupalnik V, Chomsky E, Zerbib M, Maza I, Rechavi Y, Massarwa R, Hanna S, Amit I, Levanon EY, Amariglio N, Stern-Ginossar N, Novershtern N, Rechavi G, et al.: **m6A mRNA methylation facilitates resolution of naïve pluripotency toward differentiation**. *Science (80-)* 2015, **347**:1002–1006.

132. Lev Maor G, Yearim A, Ast G: **The alternative role of DNA methylation in splicing regulation**. *Trends Genet* 2015, **31**:274–280.

133. Chang G, Gao S, Hou X, Xu Z, Liu Y, Kang L, Tao Y, Liu W, Huang B, Kou X, Chen J, An L, Miao K, Di K, Wang Z, Tan K, Cheng T, Cai T, Gao S, Tian J: **High-throughput sequencing reveals the disruption of methylation of imprinted gene in induced pluripotent stem cells.** *Cell Res* 2014, **24**:293–306.

134. Frommer M, McDonald LE, Millar DS, Collis CM, Watt F, Grigg GW, Molloy PL, Paul CL: **A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands.** *Proc Natl Acad Sci U S A* 1992, **89**:1827–1831.

135. Krueger F, Andrews SR: **Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications.** *Bioinformatics* 2011, **27**:1571–2.

136. Li H: **Tabix: fast retrieval of sequence features from generic TAB-delimited files.** *Bioinformatics* 2011, **27**:718–9.

137. Peng Q, Ecker JR: **Detection of allele-specific methylation through a generalized heterogeneous epigenome model**. *Bioinformatics* 2012, **28**:163–171.

138. Shao X, Zhang C, Sun M, Lu X, Xie H: **Deciphering the heterogeneity in DNA methylation patterns during stem cell differentiation and reprogramming**. 2014, **15**:1–10.

139. Pelizzola M, Ecker JR: **The DNA methylome**. *FEBS Lett* 2011, **585**:1994–2000.

140. Ball MP, Li JB, Gao Y, Lee J-H, LeProust EM, Park I-H, Xie B, Daley GQ, Church GM: **Targeted and genome-scale strategies reveal gene-body methylation signatures in human cells.** *Nat Biotechnol* 2009, **27**:361–368.

141. 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 J a, Evans RM, Ecker JR: **Hotspots of aberrant epigenomic reprogramming in human induced pluripotent stem cells.** *Nature* 2011, **471**:68–73.

142. Zhu J, Adli M, Zou JY, Verstappen G, Coyne M, Zhang X, Durham T, Miri M, Deshpande V, De Jager PL, Bennett D a., Houmard J a., Muoio DM, Onder TT, Camahort R, Cowan C a., Meissner A, Epstein CB, Shoresh N, Bernstein BE: **Genome-wide chromatin state transitions associated with developmental and environmental cues**. *Cell* 2013, **152**:642–654.

143. Oudenaarden a Van, Young JW, Alon U, Swain PS, Elowitz MB, Desplan C, Dalal CK, Losick R, Brody MS, Price CW, Savageau M a, Gross C a, Mithoe SC, Boor KJ, Wiedmann M, Haldenwang WG, Huxley a F, Liberman LM, Leibler S, Kulkarni RP, Dworkin J, Levy S, Barkai N, Shilo BZ, Liu F, Zhang XP, Wang W, Igoshin O a, Duncan L, Becskei a, et al.: **Activation on Growth Rate in Some Conditions, Even Under Energy Stress (27), These Results Suggest That Cells Balance the Benefits and Costs of S**. *Science (80-)* 2011, **334**(October):369–373.

144. Schmitz RJ, Schultz MD, Urich M a, Nery JR, Pelizzola M, Libiger O, Alix A, McCosh RB, Chen H, Schork NJ, Ecker JR: **Patterns of population epigenomic diversity.** *Nature* 2013, **495**:193–8.

145. Guo JU, Su Y, Zhong C, Ming GL, Song H: **Hydroxylation of 5-methylcytosine by TET1 promotes active DNA demethylation in the adult brain**. *Cell* 2011, **145**:423–434.

146. Quy J, Zhouy M, Song Q, Hong EE, Smith AD: **MLML: Consistent simultaneous estimates of DNA methylation and hydroxymethylation**. *Bioinformatics* 2013, **29**:2645–2646.

147. Down T a, Rakyan VK, Turner DJ, Flicek P, Li H, Kulesha E, Gräf S, Johnson N, Herrero J, Tomazou EM, Thorne NP, Bäckdahl L, Herberth M, Howe KL, Jackson DK, Miretti MM, Marioni JC, Birney E, Hubbard TJP, Durbin R, Tavaré S, Beck S: **A Bayesian deconvolution strategy for immunoprecipitation-based DNA methylome analysis.** *Nat Biotechnol* 2008, **26**:779–785.

148. Riebler A, Menigatti M, Song JZ, Statham AL, Stirzaker C, Mahmud N, Mein C a, Clark SJ, Robinson MD: **BayMeth: improved DNA methylation quantification for affinity capture sequencing data using a flexible Bayesian approach.** *Genome Biol* 2014, **15**:R35.

149. Stevens M, Cheng JB, Li D, Xie M, Hong C, Maire CL, Ligon KL, Hirst M, Marra M a., Costello JF, Wang T: **Estimating absolute methylation levels at single-CpG resolution from methylation enrichment and restriction enzyme sequencing methods**. *Genome Res* 2013, **23**:1541–1553.

150. Akalin A, Kormaksson M, Li S, Garrett-bakelman FE, Figueroa ME, Melnick A, Mason CE: **methylKit : a comprehensive R package for the analysis of genome-wide DNA methylation profiles**. *Genome Biol* 2012, **13**:R87.

151. Song Q, Decato B, Hong EE, Zhou M, Fang F, Qu J, Garvin T, Kessler M, Zhou J, Smith AD: **A reference methylome database and analysis pipeline to facilitate integrative and comparative epigenomics**. *PLoS One* 2013, **8**.

152. Rahl PB, Lin CY, Seila AC, Flynn R a, McCuine S, Burge CB, Sharp P a, Young R a: **c-Myc regulates transcriptional pause release.** *Cell* 2010, **141**:432–45.

153. Koga Y, Pelizzola M, Cheng E, Krauthammer M, Sznol M, Ariyan S, Narayan D, Molinaro AM, Halaban R, Weissman SM: **Genome-wide screen of promoter methylation identifies novel markers in melanoma**. *Genome Res* 2009, **19**:1462–1470.

154. Müllner D: **fastcluster : Fast Hierarchical , Agglomerative**. *J Stat Softw* 2013, **53**:1–18.

155. Statham AL, Strbenac D, Coolen MW, Stirzaker C, Clark SJ, Robinson MD: **Repitools: an R package for the analysis of enrichment-based epigenomic data.** *Bioinformatics* 2010, **26**:1662–3.

156. Mayo TR, Schweikert G, Sanguinetti G: **M3D: a kernel-based test for spatially correlated changes in methylation profiles**. *Bioinformatics* 2014, **31**(November 2014):809–816.

157. Hansen KD, Langmead B, Irizarry RA: **BSmooth : from whole genome bisulfite sequencing reads to differentially methylated regions**. *Genome Biol* 2012, **13**:R83.

158. Liang F, Tang B, Wang Y, Wang J, Yu C, Chen X, Zhu J, Yan J, Zhao W, Li R: **WBSA: Web service for bisulfite sequencing data analysis**. *PLoS One* 2014, **9**:1–9.

159. Jiang P, Sun K, Lun FMF, Guo AM, Wang H, Chan KCA, Chiu RWK, Lo YMD, Sun H: **Methy-pipe: an integrated bioinformatics pipeline for whole genome bisulfite sequencing data analysis.** *PLoS One* 2014, **9**:e100360.

160. Gu H, Smith ZD, Bock C, Boyle P, Gnirke A, Meissner A: **Preparation of reduced representation bisulfite sequencing libraries for genome-scale DNA methylation profiling.** *Nat Protoc* 2011, **6**:468–81.

161. Wang H-Q, Tuominen LK, Tsai C-J: **SLIM: a sliding linear model for estimating the proportion of true null hypotheses in datasets with dependence structures.** *Bioinformatics* 2011, **27**:225–31.

162. Cedar H, Bergman Y: **Linking DNA methylation and histone modification: patterns and paradigms.** *Nat Rev Genet* 2009, **10**:295–304.

163. Meyer KD, Jaffrey SR: **The dynamic epitranscriptome: N6-methyladenosine and gene expression control.** *Nat Rev Mol Cell Biol* 2014, **15**:313–26.

164. Consortium ME, Tables S: **A comparative encyclopedia of DNA elements in the mouse genome**. *Nature* 2014, **515**:355–364.

165. Trapnell C, Pachter L, Salzberg SL: **TopHat: Discovering splice junctions with RNA-Seq**. *Bioinformatics* 2009, **25**:1105–1111.

166. Zhang Y, Liu T, Meyer C a, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W, Liu XS: **Model-based analysis of ChIP-Seq (MACS).** *Genome Biol* 2008, **9**:R137.

167. Pelizzola M, Koga Y, Urban AE, Krauthammer M, Weissman S, Halaban R, Molinaro AM: **MEDME : An experimental and analytical methodology for the estimation of DNA methylation levels based on microarray derived**. *Genome Res.* 2008: 18(10): 1652–1659.

168. Wang X, Lu Z, Gomez A, Hon GC, Yue Y, Han D, Fu Y, Parisien M, Dai Q, Jia G, Ren B, Pan T, He C: **N6-methyladenosine-dependent regulation of messenger RNA stability.** *Nature* 2014, **505**:117–20.

169. Liu J, Yue Y, Han D, Wang X, Fu Y, Zhang L, Jia G, Yu M, Lu Z, Deng X, Dai Q, Chen W, He C: **A METTL3-METTL14 complex mediates mammalian nuclear RNA N6-adenosine methylation.** *Nat Chem Biol* 2014, **10**:93–5.

170. Frietze S, O'Geen H, Blahnik KR, Jin VX, Farnham PJ: **ZNF274 recruits the histone methyltransferase SETDB1 to the 39 ends of ZNF genes**. *PLoS One* 2010, **5**.

171. Trimarchi JM, Fairchild B, Verona R, Moberg K, Andon N, Lees J a: **E2F-6, a member of the E2F family that can behave as a transcriptional repressor.** *Proc Natl Acad Sci U S A* 1998, **95**:2850–2855.

172. Ogawa H, Ishiguro K-I, Gaubatz S, Livingston DM, Nakatani Y: **A complex with chromatin modifiers that occupies E2F- and Myc-responsive genes in G0 cells.** *Science* 2002, **296**:1132–1136.

173. Sims RJ, Millhouse S, Chen CF, Lewis B a., Erdjument-Bromage H, Tempst P, Manley JL, Reinberg D: **Recognition of Trimethylated Histone H3 Lysine 4 Facilitates the Recruitment of Transcription Postinitiation Factors and Pre-mRNA Splicing**. *Mol Cell* 2007, **28**:665–676.

174. Xu C, Liu K, Tempel W, Demetriades M, Aik W, Schofield CJ, Min J: **Structures of human ALKBH5 demethylase reveal a unique binding mode for specific single-stranded N6-methyladenosine RNA demethylation**. *J Biol Chem* 2014, **289**:17299–17311.

175. Zhang D, Yoon H, Wong J: **JMJD2A Is a Novel N-CoR-Interacting Protein and Is Involved in Repression of the Human Transcription Factor Achaete JMJD2A Is a Novel N-CoR-Interacting Protein and Is Involved in Repression of the Human Transcription Factor Achaete**. *Mol Cell Biol* 2005, **25**:6404–6414.

176. Jiang H, Shukla A, Wang X, Chen WY, Bernstein BE, Roeder RG: **Role for Dpy-30 in ES cell-fate specification by regulation of H3K4 methylation within bivalent domains**. *Cell* 2011, **144**:513–525.

177. Laherty CD, Billin a N, Lavinsky RM, Yochum GS, Bush a C, Sun JM, Mullen TM, Davie JR, Rose DW, Glass CK, Rosenfeld MG, Ayer DE, Eisenman RN: **SAP30, a component of the mSin3 corepressor complex involved in N-CoR-mediated repression by specific transcription factors.** *Mol Cell* 1998, **2**:33–42.

178. Laherty CD, Yang WM, Jian-Min S, Davie JR, Seto E, Eisenman RN: **Histone deacetylases associated with the mSin3 corepressor mediate Mad transcriptional repression**. *Cell* 1997, **89**:349–356.

179. Duong H a., Robles MS, Knutti D, Weitz CJ: **A Molecular Mechanism for Circadian Clock Negative Feedback**. *Science* 2011, **332**:1436–1439.

180. Fustin JM, Doi M, Yamaguchi Y, Hida H, Nishimura S, Yoshida M, Isagawa T, Morioka MS, Kakeya H, Manabe I, Okamura H: **XRNA-methylation-dependent RNA processing controls the speed of the circadian clock**. *Cell* 2013, **155**:793–806.

181. Huang Y, Myers SJ, Dingledine R: **Transcriptional repression by REST: recruitment of Sin3A and histone deacetylase to neuronal genes**. *Nat Neurosci* 1999, **2**:867–872.

182. Ballas N, Mandel G: **The many faces of REST oversee epigenetic programming of neuronal genes**. *Curr Opin Neurobiol* 2005, **15**:500–506.

183. Hakimi M-A, Bochar D a, Chenoweth J, Lane WS, Mandel G, Shiekhattar R: **A core-BRAF35 complex containing histone deacetylase mediates repression of neuronal-specific genes.** *Proc Natl Acad Sci U S A* 2002, **99**:7420–7425.