Computational Analysis and Integration of MeDIP-seq Methylome Data

Gareth A. Wilson and Stephan Beck

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/61207

Abstract

The combinatorial number of possible methylomes in biological time and space is astronomical. Consequently, the computational analysis of methylomes needs to cater for a variety of data, throughput and resolution. Here, we review recent advances in 2nd generation sequencing (2GS) with a focus on the different methods used for the analysis of MeDIP-seq data. The challenges and opportunities presented by the integration of methylation data with other genomic data types are discussed as is the potential impact of emerging 3rd generation sequencing (3GS) based technologies on methylation analysis.

Keywords: DNA methylation, methylome, immuno precipitation, analysis pipeline

1. Introduction

For many years it's been widely known by scientists that, despite possessing the same DNA sequence, not all genes can be active in all cells within an organism all of the time. It is through the regulation of genes that we are able to see phenotypic differences between cells with identical genotypes. In the late 1930's, Conrad Waddington introduced the term 'epigenetic landscape' to provide a metaphor for the cellular mechanisms leading to this regulation [1]. These regulatory, or epigenetic, patterns can be seen to persistently influence gene expression levels through cell division. Hence, epigenetics involves the study of marks and mechanisms that control gene expression in a mitotically and potentially meiotically heritable manner [2].

One such mechanism is DNA methylation (or more specifically cytosine methylation), an important epigenetic modification. DNA methylation, in conjunction with histone modifications, remodeling complexes and non-coding RNAs, plays a vital role in regulating genome dynamics. In combination with these other modifications, DNA methylation can control the accessibility of the underlying DNA to the transcriptional machinery through the modulation of chromatin density. As a result DNA methylation is involved in a diverse range of processes including embryogenesis, genomic imprinting, cellular differentiation, DNA protein interactions and gene regulation [3].

In mammalian genomes, DNA methylation occurs almost exclusively at palindromic CpG dinucleotides. CpG dinucleotides are found throughout the genome but are significantly depleted (21% of that expected in the human genome [4]) in comparison to other dinucleotide combinations. This is due to the hypermutability of methylated cytosines [5] where spontaneous deamination to thymine occurs. However as a result of chance or potentially due to their functional importance, a minority of CpGs are maintained against this loss.

The surviving CpGs are often found at a high density in localised genomic regions termed CpG islands (CGIs) [3]. Unlike the majority of CpGs, these regions, of approximately 1kb in length (though different algorithms produce different CGI predictions [6]), are largely unmethylated and have been found to overlap the promoter regions of 60–70% of all human genes, representing all constitutively expressed genes and approximately 40% of those displaying tissue specific expression patterns [7, 8]. Unmethylated CGIs are able to recruit CpG binding proteins such as Cfp1 [9], these in turn lead to the modification of histone tails [10] and the formation of permissive chromatin domains, potentially enabling the initiation of transcription [11]. In contrast, methylated CGIs are associated with gene silencing. This silencing can occur via various routes such as inhibiting the recruitment of DNA binding proteins from their target sites [12] or alternatively through the recruitment of methyl-CpG-binding domain (MBD) proteins that in turn recruit histone modifying complexes to the methylated sites [13].

Whilst methylation changes at CGIs is perhaps the most studied region, methylation occurs in other genomic locations as well. CpG island shores represent regions of lower CpG density flanking a CGI. They are generally defined as reaching 2kb upstream and downstream of an island. It has been found that most tissue specific methylation occurs in these shore regions rather than the islands [14, 15]. Additionally, high levels of DNA methylation can be found in repetitive genomic regions. Rather than directly regulating the transcriptional potential of a gene, this methylation is seen to prevent chromosomal instability [16-18].

Although DNA methylation is largely found in the CpG dinucleotide, it has also been reported in humans and mouse at CHG and CHH sites [19, 20]. In comparison with a methylated CpG site, methylated non-CpG sites display a much lower level of methylation within a cell population [21] and show lower conservation between cell lines [22]. The mechanisms and functionality of non-CpG methylation are currently unclear but the levels appear to decrease during differentiation whilst being restored in induced pluripotent stem cells. This potentially suggests a role in the origin and maintenance of the pluripotent state [19, 23, 24].

DNA methylation changes have been associated with numerous conditions. Many cancers have shown hypomethylation at repetitive sequences thus promoting chromosomal instability. Examples include the LINE repeat L1 in a range of tumours [25] and satellite repeats ALRα and SATR1 in peripheral nerve sheath tumours [26]. Hypomethylation at specific promoters can lead to aberrant expression of oncogenes, whilst in contrast hypermethylation at specific island or shore sites can lead to transcriptional inactivation of genes involved in pathways such as DNA repair and apoptosis [2, 13]. Neurological disorders such as Alzheimer's and Multiple sclerosis have been associated with aberrant DNA methylation as have autoimmune diseases such as ICF syndrome and rheumatoid arthritis [2].

2. Methods for the study of genome-wide DNA methylation

Even within the relatively new field of second-generation (or next-generation) sequencing (2GS), a plethora of methods exist for the exploration of DNA methylation and the analysis of the ensuing data (Table 1). Such methods include the use of restriction endonucleases, or the bisulphite conversion of DNA. Here we discuss in detail the analysis of affinity enrichment techniques, specifically MeDIP-seq. For a full review of other methods see [27].

Software	Method	Summary	Publication
Batman	MeDIP-seq	Bayesian tool for methylation analysis of MeDIP profiles	[33]
Bismark	Bisulphite	Maps bisulfite treated sequencing reads through in-silico bisulfite conversion of both reads and genome. Performs methylation calling in a quick and easy-to-use fashion.	[81]
Bis-SNP	Bisulphite	Estimates methylation probabilities of different cytosine context to determine genotypes and methylation levels simultaneously.	[61]
BSMAP	Bisulphite	Mapping software for bisulphite sequencing. BSMAP aligns the Ts in the reads to both Cs and Ts in the reference by building a "seed" index of the reference genome.	[82]
BS-Seeker	Bisulphite	Accurate and fast mapping of bisulfite-treated short reads through in-silico bisulfite conversion of both reads and genome.	[83]
EpiExplorer	Various	Web tool that allows you to use large reference epigenome datasets for your own analysis without complex scripting or preprocessing.	[58]
Epigenome Browser	Various	Resource for visualizing and interacting with whole-genome datasets. The browser currently hosts Human Epigenome Atlas data produced by the Roadmap Epigenomics project.	[84]

Software	Method	Summary	Publication
MEDIPS	MeDIP-seq	Bioconductor package providing a comprehensive approach for normalizing and analyzing MeDIP-seq data	[38]
MeDUSA	MeDIP-seq	Performs a full analysis of MeDIP-seq data, including sequence alignment, QC and determination and annotation of DMRs	[40]
MeQA	MeDIP-seq	Pipeline for the pre-processing, quality assessment, read distribution and methylation estimation for MeDIP-seq datasets	[39]
MethMarker	Validation	Implements a systematic workflow for design, optimization and (computational) validation of DNA methylation biomarkers.	[85]
Methylcoder	Bisulphite	Software pipeline for bisulfite-treated sequences	[86]
MethylSeekR	Bisulphite	Accurately identifies the footprints of active regulatory regions from bisulfite-sequencing data	[87]
Metmap	Methyl-seq	Produces corrected site-specific methylation states from MethylSeq experiments and annotates unmethylated islands across the genome.	[88]
Sherman	Validation	Simulates ungapped high-throughput datasets for bisulfite sequencing. Allows for evaluation of the influence of common problems observed in many sequencing experiments.	http://tinyurl.com/bwkttgh

Table 1. Examples of software available for the analysis of 2GS methylation data.

Buoyed by the success of combining chromatin immunoprecipitation with second generation sequencing for genome-wide studies of histone modifications and transcription factor binding sites [28] (termed ChIP-seq), similar techniques were adopted for methylation. These methods generally involve either enrichment through methylcytosine-specific protein domains (e.g. MethylCap[29], MBD-seq[30]) or through antibody-mediated immunoprecipitation (e.g. MeDIP[31], MRE-seq[32]) prior to sequencing[33, 34]. Such approaches, whilst not offering the resolution of bisulphite sequence data, are both genome-wide and increasingly affordable. Concordance in methylation calls between different enrichment and bisulphite methods have been shown to be high[35, 36]. In methylated DNA immunoprecipitation (MeDIP), an antibody capable of recognizing 5mC is utilized to immunoprecipitate the methylated fraction of the genome. One issue that has been highlighted with enrichment methods such as MeDIP, is the necessity to take the sequencing to saturation in order to confirm lack of methylation at a CpG site. Such a policy would be costly and would generate a vast amount of redundant data and as such saturation has not been reached with these methods. Methylation-sensitive restriction enzymes (MRE) target unmethylated CpGs for sequencing thus one alternative suggestion is to integrate the MRE-seq method with MeDIP-seq. Such integration will have the benefit of reducing the need for saturation sequencing and will highlight regions of intermediate methylation, which would be difficult to detect using a single method. Going a step further, if coupled with single nucleotide polymorphism (SNP) profiling, it would also be possible to detect potential allele-specific epigenetic states[35].

MeDIP-seq is a popular enrichment technique for interrogating the methylation status of cytosines across entire genomes. It has been used in numerous studies including the first mammalian methylome [33] and the first cancer methylome [26]. In the next section, approaches for the analysis of MeDIP-seq data will be discussed in greater detail.

3. Computational approaches for the analysis of MeDIP-seq data

A number of computational tools have been developed for the analysis of MeDIP data (Table 1), including Batman [33], MEDME [37], MEDIPS [38], MeQA [39] and MeDUSA [40]. The method to use depends very much on the questions you want to ask of the data, and as a result the type of analysis performed can be described as analyzing absolute methylation or, alternatively, relative methylation.

3.1. Absolute methylation

The efficiency of immunoprecipitation in MeDIP is largely dependent on the density of methylated CpG sites. Therefore, it is difficult to distinguish true variation in enrichment, and hence methylation, from confounding effects caused by fluctuations in CpG density. This bias needs to be corrected for in order to perform accurate and biologically relevant comparisons of methylation states between different genomic regions.

The first method to try and correct for this bias was called Batman (Bayesian Tool for Methylation Analysis)[33]. This tool was originally written to analyse MeDIP-chip data, but can also be applied to 2GS. Batman, distributed as a suite of Java scripts, models the effect of varying densities of methylated CpGs on MeDIP enrichment, resulting in the transformation of the count of the aligned sequence read depth within a 100bp region into a quantitative measure of DNA methylation. Such data can then be used to compare global methylation scores between methylomes or between feature types (e.g. CpG islands, exons) within a methylome. Batman was used for the analysis of the first mammalian methylome[33] and also the first cancer methylome[26]. Unfortunately, Batman was disproportionately time consuming to run, even when running with multiple processors. The R BioConductor package[41] MEDIPS v1.8 [38] attempted to utilize much of the methodology used in the Batman approach whilst outperforming the computation time by orders of magnitude. By implementing MEDIPS as an R package, this method is also more approachable for the majority of users, requiring less computational knowledge to run the methods. In addition to generating genome-wide methylation scores, MEDIPS sought to provide MeDIP-seq specific quality control metrics such as calculating the degree of enrichment of CpG-rich sequenced reads relative to genomic background. Finally, MEDIPS provided a methodology for determining the location of differentially methylated regions (DMRs) between samples. Whilst MEDIPS, building on the strengths of Batman, undoubtedly provided an important step forward in the analysis of MeDIP-seq data, it also had significant issues that need to be considered both before use and when interrogating output from the program. For example, the DMR calling algorithm requires an input sample to be sequenced in addition to the immunoprecipitated sample, thus effectively doubling costs.

3.2. Relative methylation

Methods for calculating absolute methylation have proven to be useful when identifying large global changes, for example hypomethylation of satellite repeats in peripheral nerve sheath tumours[26]. Additionally, transforming MeDIP-seq data from read counts to a methylation score has assisted in validating experiments against bisulphite data[33]. However, as yet, these methods have not provided a framework for determining the location of DMRs in a statistically rigorous manner. To achieve this, relative changes in DNA methylation between cohorts can be determined, rather than absolute changes within a cohort. As such the problem has much in common with other sequencing protocols, such as identifying differential expression between RNA-seq cohorts or identifying peaks from a ChIP-seq sample. This commonality opens up an abundance of methods that can be used or adapted for MeDIP-seq sample analysis, for example peak finding using MACS[42, 43], or DMR finding using DESeq [44] or edgeR [45].

There are several hurdles to cross when analysing MeDIP-seq data, particularly during the identification of DMRs. Read counts need to be normalized to eliminate biases as a result of variability in sequencing depth between samples. Whilst global read count normalization can help address this problem, it does not account for 'competition' effects. RNA-seq provides an example of such effects, in which condition specific highly expressed genes can lead to a depressed read count in other genes and hence a bias when comparing samples[46]. An analogous situation can be found in MeDIP-seq, where sample-specific repeat methylation could potentially diminish reads in other genomic regions and introduce bias to analyses, particularly given the large amount of repetitive sequence methylated in the genome. Further, despite falling sequencing costs, MeDIP-seq experiments will often have few biological replicates. As a result, it can be difficult to obtain reliable estimates of model parameters to fit statistical models and thereby locate real differences between samples. By using methods such as DESeq that estimate variance in a local fashion, it is possible to remove potential selection biases [44]. Additionally, DESeq estimates a flexible, mean-dependent local regression rather than attempting to reliably estimate both the variance and mean parameters of the distribution from limited numbers of replicates. Typically, there is enough data available in these experiments to allow for sufficiently precise local estimation of the dispersion [44] and hence avoid bias towards certain areas of the dynamic range when identifying DMRs. Finally, accurate biological interpretation could be compromised by differences in DNA fragment size distributions between samples. Performing fragment length normalization through read subsampling to equalize the distributions can eliminate this potential bias.

Additionally, the methods developed for absolute methylation calculation are unable to take account of non-CpG methylation and, due to the models used being based on local CpG density, the presence of non-CpG methylation could adversely skew the output. In contrast, a relative methylation approach should be able to locate differences in methylation driven by asymmetric non-CpG methylation[47], taking advantage of the affinity of the MeDIP-seq antibody for methylated cytosine (rather than exclusively selecting for methylated CpGs).

The first pipeline to provide a comprehensive methodology for analyzing MeDIP-seq data, with the focus on accurate and statistically rigorous identification of DMRs, was MeDUSA (Methylated DNA Utility for Sequence Analysis) (https://www.ucl.ac.uk/cancer/medical-genomics/medusaproject) [40]. MeDUSA (v1) utilized a number of software packages to perform a complete analysis of MeDIP-seq data. This included sequence alignment, quality control (QC), and determination and annotation of DMRs. The novel aspect of MeDUSA was the approach to DMR calling. It utilized the USeq suite of tools, specifically MultipleReplicaScanSeqs (MRSS) and EnrichedRegionMaker [48]. MRSS formatted data for use in the BioConductor package DESeq [44]. DESeq determined significant differential counts between cohorts. These regions are passed to EnrichedRegionMaker to determine if multiple regions can be combined to create single larger regions. MeDUSA proceeded to provide initial annotation of these DMR regions.

More recently new versions of both MEDIPS (v1.10) and MeDUSA (v2) have been released. The MEDIPS package now incorporates methods from the edgeR [45] bioconductor package to provide a DMR calling methodology analogous to that used in MeDUSA. However, the approach and implementation employed by MEDIPS is more efficient (both time and computational) than the DMR calling method used in MeDUSA v1. As a consequence, MeDUSA (v2) now utilises MEDIPS for the DMR calling stage of the pipeline.

4. Data integration

As more studies are published and sequencing costs fall, the opportunity to integrate methylation datasets with other data types increases[49]. Whilst being able to detect changes in methylation is interesting, it is more interesting, and indeed more likely to be of functional importance, if this change associates with other detectable biological signals. For example, the potential of associating a methylation change with a corresponding change in transcription of a particular splice variant[50-52] from RNA-seq, or with an increase in binding of a specific transcription factor using ChIP-seq data[53].

In addition to the published sequence and array based datasets stored in public repositories such as GEO[54], a number of datasets are pre-loaded in public Genome Browsers. For example, the UCSC Genome Browser provides access to data from the ENCODE project[55], including expression data in the form of RNA-seq and regulatory data generated through ChIP-seq representing several different cell lines and various primary tissue types. Compressed file formats such as bigWig and bigBed[56] make it relatively simple to load and visualize multiple data types (Figure 1) whilst software such as bedTools[57] allow for quick intersections between data to be determined. EpiExplorer functions as a user-friendly web-based solution for providing initial annotations of feature sets [58], such as differentially methylated regions. It enables exploratory analysis of user-uploaded data and provides links to many external public datasets. As datasets become larger and more complex, other methods of integration may be required, for example an unsupervised clustering approach may be useful [49, 59].

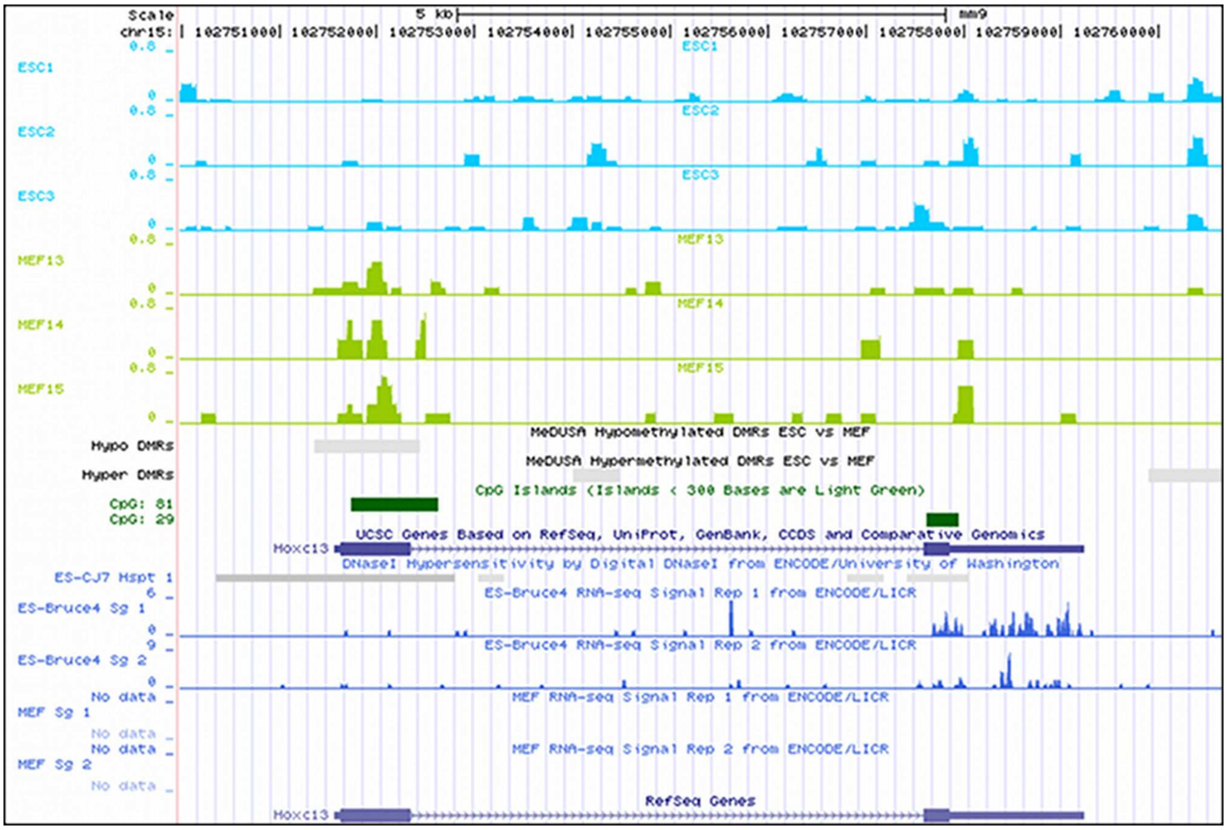
Scale
5 kb
mm9
chr15: | 102751000| 102752000| 102753000| 102754000| 102755000| 102756000| 102757000| 102758000| 102759000| 102760000|
ESC1
ESC2
ESC3
MEF13
MEF14
MEF15
0.8
0
MeDUSA Hypomethylated DMRs ESC vs MEF
Hypo DMRs
MeDUSA Hypermethylated DMRs ESC vs MEF
Hyper DMRs
CpG Islands (Islands < 300 Bases are Light Green)
CpG: 81
CpG: 29
UCSC Genes Based on RefSeq, UniProt, GenBank, CCDS and Comparative Genomics
Hoxc13
DNaseI Hypersensitivity by Digital DNaseI from ENCODE/University of Washington
ES-CJ7 Hspt 1
ES-Bruce4 RNA-seq Signal Rep 1 from ENCODE/LICR
6
ES-Bruce4 Sg 1
ES-Bruce4 RNA-seq Signal Rep 2 from ENCODE/LICR
9
ES-Bruce4 Sg 2
No data
MEF RNA-seq Signal Rep 1 from ENCODE/LICR
MEF Sg 1
No data
No data
MEF RNA-seq Signal Rep 2 from ENCODE/LICR
MEF Sg 2
No data
RefSeq Genes
Hoxc13

Figure 1. Visualising MeDUSA output in UCSC Genome Browser. MeDIP tracks are shown for 3 embryonic stem cell (ESC) replicates and 3 Mouse embryonic fibroblasts (MEF) replicates over the Hoxc13 gene. The CpG island in the promoter region is hypomethylated in the ESC samples, suggesting more permissible chromatin in ESCs than in MEFs. This is supported by the ES-CJ7 DNase I Hypersensitivity track. Additionally the RNA-seq tracks show transcriptional differences in this gene between ESCs and MEFs.

In addition to transcriptomic and regulatory data, it is also possible to integrate methylation data with genomic information. A perceived difference in methylation at a given CpG dinucleotide between samples could be caused by one sample possessing a methylated cytosine whilst the other sample possesses an unmethylated cytosine. Alternatively, the methylation difference could be due to the presence of a SNP, seeing the cytosine replaced with an alternative base. Therefore, the use of genotype profiling can clarify whether a methylation difference is a result of genetic or epigenetic changes. The need to consider both genetic and epigenetic changes came to the fore with the release of the Illumina Infinium HumanMethylation450 BeadChip. This chip allows for the interrogation of 485000 potential sites of methylation. However, a significant proportion of these sites are also sites of known SNPs[60]. Thus, any difference detected at these sites could be driven by epigenetic or genetic factors. Whilst this is an issue for the array analysis, tools such as Bis-SNP are able to make

SNP calls from bisulphite sequencing data, in doing so allowing for both accurate quantification of methylation levels and for identification of allele-specific epigenetic events such as imprinting [61].

A recent study utilised a combination of SNP, expression and methylation data to determine whether methylation has a passive or active role in gene regulation [62]. Three models were considered for the relationship between methylation and regulation. The first model described how a SNP would independently influence expression and methylation, for example through SNP modification of a transcription factor binding site (the impact on methylation of small changes to nucleotides constituting a TFBS have been explored in a recent tri-primate methylome study [89]). In the second model, a SNP would impact upon methylation, which, in turn, would modify expression. The final model shows a SNP affecting expression that consequently alters the methylation state. It was found that, in reality, each of these models occurs in different contexts with the frequency of the model varying according to cell type [62, 63]. Such studies underline the complexity inherent in, and the difficulty in deciphering, regulatory interactions and should serve as a warning to those seeking overly simplistic interpretations [63].

Extending the genetic effect out from a single site to an entire region, it is possible that methylation levels could be strongly influenced by the haplotypic phase[64]. Haplotype specific methylation (HSM) is a result of the cumulative methylation effect driven by the phase of a number of CpG-SNPs within the haplotype. This signal was strong enough to be identified across the 47kb FTO linkage disequilibrium block[65]. Such a finding is only possible through the integration of DNA methylation data and genome wide association study data. It is also worth remembering at this juncture that whether a measured methylation difference is due to a SNP or not, the downstream impact on the transcriptional potential of the chromosomal region in question could be the same.

5. Future perspectives

The field of epigenetics and specifically the study of DNA methylation have emerged as major areas of research in recent years. This rise can be largely attributed to the impact of emerging technologies, particularly 2GS. Projects that would have been perceived as impossible just a few years ago have been completed and more are underway. The International Human Epigenome Consortium (IHEC) (http://www.ihec-epigenomes.org/) was established to provide high-resolution reference epigenome maps to the research community by coordinating large-scale international efforts. The grand aim of which is to generate 1000 reference epigenomes. Various initiatives worldwide have joined IHEC in an attempt to complete the goal. In Europe, the BLUEPRINT Project[66] will take the IHEC goal forward and in doing so improve our understanding of the human epigenome – of which the methylome is a key constituent.

There are still many questions associated with the role of DNA methylation. Some with regards to the biology, and some the techniques used. It is important to know, for example, if using an enrichment based technique, what the specificity of your antibody is. Different antibodies appear to show differing levels of repeat enrichment when performing MeDIP[67]. It would be of benefit to standardize these analyses. Similarly, different bisulphite conversion protocols may lead to differing conversion success. Global CpG methylation levels obtained from WGBS for 3 human embryonic stem cell (HESC) lines showed surprising variability (72% - 85%)[68]. This could be due to unstable gain and loss of methylation as previously reported in embryonic stem cells (ESCs)[69, 70], but it could also be a result of pre-analysis protocol and lab specific differences in sample preparation. Equally, it will be interesting to discover more about the biological roles and genomic location of the different cytosine modifications (5-hydroxymethylcytosine[47], 5-Formylcytosine and 5-Carboxylcytosine[71]) and also non-CpG methylation.

New technologies with the potential for adaption for the analysis of DNA methylation are being developed constantly. For example, improved methods of methylation validation would be highly beneficial. Often hundreds or thousands of potential candidate regions are generated from a multi-sample MeDIP-seq comparison, and similar numbers could be produced by future EWAS (Epigenome-Wide Association Studies)[72]. Ideally, many of these regions would be validated using a different technology. Targeted bisulphite sequencing is often used, however this can often be laborious and time-consuming. Combining new technologies such as microdroplet-based PCR target enrichment (e.g. RainDance Technologies) with 2GS has recently been developed into a high-throughput platform termed RainDropBS-seq [73], providing an excellent option to remove the validation bottle-neck. There is also the emergence of third generation sequencing on the horizon. Third generation sequencing (3GS) theoretically promises many advantages over existing 2GS methods including higher throughput, longer read lengths, improved accuracy and requiring smaller amounts of starting material[74], indeed some companies e.g. Oxford Nanopore Technologies, are promising single molecule sequencing[75, 76]. The potential of single molecule nanopore sequencing is particularly exciting for researchers working in the field of DNA methylation. Theoretically, it should be possible to sequence complex mammalian genomes and determine any base modifications such as methylation[77], potentially including hitherto undiscovered modifications, without the need for any of the treatments or enrichments discussed above.

As the large-scale projects, such as IHEC, BLUEPRINT and increasingly clinically oriented projects such as OncoTrack progress, it is expected that many methods and tools will become standardized. This will be an important step in translating epigenomic knowledge from the bench to the clinic[78, 79]. In the future, it is hoped that a patient will be treated with drugs tailored to their particular condition – this is of particular relevance for cancer patients. Preliminary work using whole genome, exome and RNA-seq has demonstrated the potential for treating a real patient in a relatively short time period (24 days) and a relatively low cost (~$3600)[80]. Adding reliable epigenetic information, utilising the IHEC reference genomes, to this diagnostic toolbox is a logical next step. Extrapolating from these advances, it is quite clear that the bottleneck is shifting from logistics and data generation to computational analysis.

Acknowledgements

Research in the Beck lab was supported by: Wellcome Trust (99148), Royal Society Wolfson Research Merit Award (WM100023), IMI-JU OncoTrack (115234) and EU-FP7 projects, EPIGENESYS (257082) and BLUEPRINT (282510).

Author details

Gareth A. Wilson* and Stephan Beck

*Address all correspondence to: gareth.wilson@crick.ac.uk

Medical Genomics, UCL Cancer Institute, University College London, London, UK

References

[1] Waddington CH. An introduction to modern genetics. New York,: The Macmillan company; 1939. 2 p.l., 7 -441 p. p.

[2] Portela A, Esteller M. Epigenetic modifications and human disease. Nat Biotechnol. 2010;28(10):1057-68.

[3] Bird A. DNA methylation patterns and epigenetic memory. Genes Dev. 2002;16(1): 6-21.

[4] Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. Initial sequencing and analysis of the human genome. Nature. 2001;409(6822):860-921.

[5] Duncan BK, Miller JH. Mutagenic deamination of cytosine residues in DNA. Nature. 1980;287(5782):560-1.

[6] Illingworth RS, Bird AP. CpG islands--'a rough guide'. FEBS Lett. 2009;583(11): 1713-20.

[7] Weber M, Hellmann I, Stadler MB, Ramos L, Paabo S, Rebhan M, et al. Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome. Nat Genet. 2007;39(4):457-66.

[8] Saxonov S, Berg P, Brutlag DL. A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. Proc Natl Acad Sci U S A. 2006;103(5):1412-7.

[9] Thomson JP, Skene PJ, Selfridge J, Clouaire T, Guy J, Webb S, et al. CpG islands influence chromatin structure via the CpG-binding protein Cfp1. Nature. 2010;464(7291): 1082-6.

[10] Blackledge NP, Zhou JC, Tolstorukov MY, Farcas AM, Park PJ, Klose RJ. CpG islands recruit a histone H3 lysine 36 demethylase. Mol Cell. 2010;38(2):179-90.

[11] Blackledge NP, Klose R. CpG island chromatin: A platform for gene regulation. Epigenetics. 2011;6(2):147-52.

[12] Kuroda A, Rauch TA, Todorov I, Ku HT, Al-Abdullah IH, Kandeel F, et al. Insulin gene expression is regulated by DNA methylation. PLoS One. 2009;4(9):e6953.

[13] Esteller M. Epigenetic gene silencing in cancer: the DNA hypermethylome. Hum Mol Genet. 2007;16 Spec No 1:R50-9.

[14] Doi A, Park IH, Wen B, Murakami P, Aryee MJ, Irizarry R, et al. Differential methylation of tissue- and cancer-specific CpG island shores distinguishes human induced pluripotent stem cells, embryonic stem cells and fibroblasts. Nat Genet. 2009;41(12): 1350-3.

[15] Irizarry RA, Ladd-Acosta C, Wen B, Wu Z, Montano C, Onyango P, et al. The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores. Nat Genet. 2009;41(2):178-86.

[16] Esteller M. Cancer epigenomics: DNA methylomes and histone-modification maps. Nat Rev Genet. 2007;8(4):286-98.

[17] Gaudet F, Hodgson JG, Eden A, Jackson-Grusby L, Dausman J, Gray JW, et al. Induction of tumors in mice by genomic hypomethylation. Science. 2003;300(5618):489-92.

[18] Walsh CP, Chaillet JR, Bestor TH. Transcription of IAP endogenous retroviruses is constrained by cytosine methylation. Nature genetics. 1998;20(2):116-7.

[19] Lister R, Pelizzola M, Dowen RH, Hawkins RD, Hon G, Tonti-Filippini J, et al. Human DNA methylomes at base resolution show widespread epigenomic differences. Nature. 2009;462(7271):315-22.

[20] Stadler MB, Murr R, Burger L, Ivanek R, Lienert F, Scholer A, et al. DNA-binding factors shape the mouse methylome at distal regulatory regions. Nature. 2011;480(7378): 490-5.

[21] Dyachenko OV, Schevchuk TV, Kretzner L, Buryanov YI, Smith SS. Human non-CG methylation: are human stem cells plant-like? Epigenetics : official journal of the DNA Methylation Society. 2010;5(7):569-72.

[22] Chen PY, Feng S, Joo JW, Jacobsen SE, Pellegrini M. A comparative analysis of DNA methylation across human embryonic stem cell lines. Genome Biol. 2011;12(7):R62.

[23] Laurent L, Wong E, Li G, Huynh T, Tsirigos A, Ong CT, et al. Dynamic changes in the human methylome during differentiation. Genome Res. 2010;20(3):320-31.

[24] Lister R, Pelizzola M, Kida YS, Hawkins RD, Nery JR, Hon G, et al. Hotspots of aberrant epigenomic reprogramming in human induced pluripotent stem cells. Nature. 2011;471(7336):68-73.

[25] Wilson AS, Power BE, Molloy PL. DNA hypomethylation and human diseases. Biochim Biophys Acta. 2007;1775(1):138-62.

[26] Feber A, Wilson GA, Zhang L, Presneau N, Idowu B, Down TA, et al. Comparative methylome analysis of benign and malignant peripheral nerve sheath tumors. Genome research. 2011;21(4):515-24.

[27] Laird PW. Principles and challenges of genome-wide DNA methylation analysis. Nat Rev Genet. 2010;11(3):191-203.

[28] Park PJ. ChIP-seq: advantages and challenges of a maturing technology. Nature reviews Genetics. 2009;10(10):669-80.

[29] Cross SH, Charlton JA, Nan X, Bird AP. Purification of CpG islands using a methylated DNA binding column. Nature genetics. 1994;6(3):236-44.

[30] Serre D, Lee BH, Ting AH. MBD-isolated Genome Sequencing provides a high-throughput and comprehensive survey of DNA methylation in the human genome. Nucleic Acids Res. 2010;38(2):391-9.

[31] Weber M, Davies JJ, Wittig D, Oakeley EJ, Haase M, Lam WL, et al. Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed human cells. Nat Genet. 2005;37(8):853-62.

[32] Maunakea AK, Nagarajan RP, Bilenky M, Ballinger TJ, D'Souza C, Fouse SD, et al. Conserved role of intragenic DNA methylation in regulating alternative promoters. Nature. 2010;466(7303):253-7.

[33] Down TA, Rakyan VK, Turner DJ, Flicek P, Li H, Kulesha E, et al. A Bayesian deconvolution strategy for immunoprecipitation-based DNA methylome analysis. Nat Biotechnol. 2008;26(7):779-85.

[34] Brinkman AB, Simmer F, Ma K, Kaan A, Zhu J, Stunnenberg HG. Whole-genome DNA methylation profiling using MethylCap-seq. Methods. 2010;52(3):232-6.

[35] Harris RA, Wang T, Coarfa C, Nagarajan RP, Hong C, Downey SL, et al. Comparison of sequencing-based methods to profile DNA methylation and identification of monoallelic epigenetic modifications. Nat Biotechnol. 2010;28(10):1097-105.

[36] Bock C, Tomazou EM, Brinkman AB, Muller F, Simmer F, Gu H, et al. Quantitative comparison of genome-wide DNA methylation mapping technologies. Nat Biotechnol. 2010;28(10):1106-14.

[37] Pelizzola M, Koga Y, Urban AE, Krauthammer M, Weissman S, Halaban R, et al. MEDME: an experimental and analytical methodology for the estimation of DNA methylation levels based on microarray derived MeDIP-enrichment. Genome research. 2008;18(10):1652-9.

[38] Chavez L, Jozefczuk J, Grimm C, Dietrich J, Timmermann B, Lehrach H, et al. Computational analysis of genome-wide DNA methylation during the differentiation of human embryonic stem cells along the endodermal lineage. Genome research. 2010;20(10):1441-50.

[39] Huang J, Renault V, Sengenes J, Touleimat N, Michel S, Lathrop M, et al. MeQA: a pipeline for MeDIP-seq data quality assessment and analysis. Bioinformatics. 2012;28(4):587-8.

[40] Wilson GA, Dhami P, Feber A, Cortazar D, Suzuki Y, Schulz R, et al. Resources for methylome analysis suitable for gene knockout studies of potential epigenome modifiers. GigaScience. 2012;1(1).

[41] Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, et al. Bioconductor: open software development for computational biology and bioinformatics. Genome Biol. 2004;5(10):R80.

[42] Sati S, Tanwar VS, Kumar KA, Patowary A, Jain V, Ghosh S, et al. High resolution methylome map of rat indicates role of intragenic DNA methylation in identification of coding region. PLoS One. 2012;7(2):e31621.

[43] Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, et al. Model-based analysis of ChIP-Seq (MACS). Genome Biol. 2008;9(9):R137.

[44] Anders S, Huber W. Differential expression analysis for sequence count data. Genome Biol. 2010;11(10):R106.

[45] Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010;26(1): 139-40.

[46] Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol. 2010;11(3):R25.

[47] Ficz G, Branco MR, Seisenberger S, Santos F, Krueger F, Hore TA, et al. Dynamic regulation of 5-hydroxymethylcytosine in mouse ES cells and during differentiation. Nature. 2011;473(7347):398-402.

[48] Nix DA, Courdy SJ, Boucher KM. Empirical methods for controlling false positives and estimating confidence in ChIP-Seq peaks. BMC Bioinformatics. 2008;9:523.

[49] Hawkins RD, Hon GC, Ren B. Next-generation genomics: an integrative approach. Nature reviews Genetics. 2010;11(7):476-86.

[50] Chodavarapu RK, Feng S, Bernatavichute YV, Chen PY, Stroud H, Yu Y, et al. Relationship between nucleosome positioning and DNA methylation. Nature. 2010;466(7304):388-92.

[51] Hodges E, Smith AD, Kendall J, Xuan Z, Ravi K, Rooks M, et al. High definition profiling of mammalian DNA methylation by array capture and single molecule bisulfite sequencing. Genome research. 2009;19(9):1593-605.

[52] Lyko F, Foret S, Kucharski R, Wolf S, Falckenhayn C, Maleszka R. The honey bee epigenomes: differential methylation of brain DNA in queens and workers. PLoS Biol. 2010;8(11):e1000506.

[53] Shukla S, Kavak E, Gregory M, Imashimizu M, Shutinoski B, Kashlev M, et al. CTCF-promoted RNA polymerase II pausing links DNA methylation to splicing. Nature. 2011;479(7371):74-9.

[54] Barrett T, Edgar R. Gene expression omnibus: microarray data storage, submission, retrieval, and analysis. Methods Enzymol. 2006;411:352-69.

[55] A user's guide to the encyclopedia of DNA elements (ENCODE). PLoS Biol. 2011;9(4):e1001046.

[56] Kent WJ, Zweig AS, Barber G, Hinrichs AS, Karolchik D. BigWig and BigBed: enabling browsing of large distributed datasets. Bioinformatics. 2010;26(17):2204-7.

[57] Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26(6):841-2.

[58] Halachev K, Bast H, Albrecht F, Lengauer T, Bock C. EpiExplorer: live exploration and global analysis of large epigenomic datasets. Genome Biol. 2012;13(10):R96.

[59] Wang Z, Zang C, Rosenfeld JA, Schones DE, Barski A, Cuddapah S, et al. Combinatorial patterns of histone acetylations and methylations in the human genome. Nature genetics. 2008;40(7):897-903.

[60] Wang D, Yan L, Hu Q, Sucheston LE, Higgins MJ, Ambrosone CB, et al. IMA: an R package for high-throughput analysis of Illumina's 450K Infinium methylation data. Bioinformatics. 2012;28(5):729-30.

[61] Liu Y, Siegmund KD, Laird PW, Berman BP. Bis-SNP: Combined DNA methylation and SNP calling for Bisulfite-seq data. Genome Biol. 2012;13(7):R61.

[62] Gutierrez-Arcelus M, Lappalainen T, Montgomery SB, Buil A, Ongen H, Yurovsky A, et al. Passive and active DNA methylation and the interplay with genetic variation in gene regulation. Elife. 2013;2:e00523.

[63] Muers M. Gene expression: Disentangling DNA methylation. Nature reviews Genetics. 2013;14(8):519.

[64] Bell CG. Integration of genomic and epigenomic DNA methylation data in common complex diseases by haplotype-specific methylation analysis. Personalized Medicine. 2011;8(3):243.

[65] Bell CG, Finer S, Lindgren CM, Wilson GA, Rakyan VK, Teschendorff AE, et al. Integrated Genetic and Epigenetic Analysis Identifies Haplotype-Specific Methylation in the FTO Type 2 Diabetes and Obesity Susceptibility Locus. PLoS One. 2010;5(11):e14040.

[66] Adams D, Altucci L, Antonarakis SE, Ballesteros J, Beck S, Bird A, et al. BLUEPRINT to decode the epigenetic signature written in blood. Nat Biotechnol. 2012;30(3):224-6.

[67] Matarese F, Carrillo-de Santa Pau E, Stunnenberg HG. 5-Hydroxymethylcytosine: a new kid on the epigenetic block? Mol Syst Biol. 2011;7:562.

[68] Flicek P, Amode MR, Barrell D, Beal K, Brent S, Chen Y, et al. Ensembl 2011. Nucleic Acids Res. 2011;39(Database issue):D800-6.

[69] Ooi SK, Wolf D, Hartung O, Agarwal S, Daley GQ, Goff SP, et al. Dynamic instability of genomic methylation patterns in pluripotent stem cells. Epigenetics Chromatin. 2010;3(1):17.

[70] Humpherys D, Eggan K, Akutsu H, Hochedlinger K, Rideout WM, 3rd, Biniszkiewicz D, et al. Epigenetic instability in ES cells and cloned mice. Science. 2001;293(5527):95-7.

[71] Ito S, Shen L, Dai Q, Wu SC, Collins LB, Swenberg JA, et al. Tet proteins can convert 5-methylcytosine to 5-formylcytosine and 5-carboxylcytosine. Science. 2011;333(6047):1300-3.

[72] Rakyan VK, Down TA, Balding DJ, Beck S. Epigenome-wide association studies for common human diseases. Nature reviews Genetics. 2011;12(8):529-41.

[73] Guilhamon P, Eskandarpour M, Halai D, Wilson GA, Feber A, Teschendorff AE, et al. Meta-analysis of IDH-mutant cancers identifies EBF1 as an interaction partner for TET2. Nat Commun. 2013;4:2166.

[74] Schadt EE, Turner S, Kasarskis A. A window into third-generation sequencing. Hum Mol Genet. 2010;19(R2):R227-40.

[75] Mason CE, Elemento O. Faster sequencers, larger datasets, new challenges. Genome Biol. 2012;13(3):314.

[76] Cherf GM, Lieberman KR, Rashid H, Lam CE, Karplus K, Akeson M. Automated forward and reverse ratcheting of DNA in a nanopore at 5-A precision. Nat Biotechnol. 2012;30(4):344-8.

[77] Clarke J, Wu HC, Jayasinghe L, Patel A, Reid S, Bayley H. Continuous base identification for single-molecule nanopore DNA sequencing. Nat Nanotechnol. 2009;4(4): 265-70.

[78] Lyon GJ. Personalized medicine: Bring clinical standards to human-genetics research. Nature. 2012;482(7385):300-1.

[79] Scudellari M. Genomics contest underscores challenges of personalized medicine. Nat Med. 2012;18(3):326.

[80] Roychowdhury S, Iyer MK, Robinson DR, Lonigro RJ, Wu YM, Cao X, et al. Personalized oncology through integrative high-throughput sequencing: a pilot study. Sci Transl Med. 2011;3(111):111ra21.

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

[82] Xi Y, Li W. BSMAP: whole genome bisulfite sequence MAPping program. BMC Bioinformatics. 2009;10:232.

[83] Chen PY, Cokus SJ, Pellegrini M. BS Seeker: precise mapping for bisulfite sequencing. BMC Bioinformatics. 2010;11:203.

[84] Zhou X, Maricque B, Xie M, Li D, Sundaram V, Martin EA, et al. The Human Epigenome Browser at Washington University. Nat Methods. 2011;8(12):989-90.

[85] Schuffler P, Mikeska T, Waha A, Lengauer T, Bock C. MethMarker: user-friendly design and optimization of gene-specific DNA methylation assays. Genome Biol. 2009;10(10):R105.

[86] Pedersen B, Hsieh TF, Ibarra C, Fischer RL. MethylCoder: software pipeline for bisulfite-treated sequences. Bioinformatics. 2011;27(17):2435-6.

[87] Burger L, Gaidatzis D, Schubeler D, Stadler MB. Identification of active regulatory regions from DNA methylation data. Nucleic Acids Res. 2013.

[88] Singer M, Boffelli D, Dhahbi J, Schonhuth A, Schroth GP, Martin DI, et al. MetMap enables genome-scale Methyltyping for determining methylation states in populations. PLoS Comput Biol. 2010;6(8):e1000888.

[89] Wilson GA, Butcher LM, Foster HR, Feber A, Roos C, Walter L, et al. Human-specific epigenetic variation in the immunological Leukotriene B4 Receptor (LTB4R/BLT1) implicated in common inflammatory diseases. Genome medicine. 2014;6(3):19.