MINI REVIEW: Statistical methods for detecting differentially methylated loci and regions

Mark D Robinson [1,2,*], Abdullah Kahraman[1,2], Charity W Law[1,2], Helen Lindsay[1,2], Malgorzata Nowicka[1,2], Lukas M Weber[1,2] and Xiaobei Zhou[1,2]

[1]Institute of Molecular Life Sciences, University of Zurich, CH-8057 Zurich, Switzerland
[2]SIB Swiss Institute of Bioinformatics, University of Zurich, CH-8057 Zurich, Switzerland

[*] To whom correspondence should be addressed.
Tel: +41 44 635 48 48; Fax: +41 44 635 68 68;
Email: mark.robinson@imls.uzh.ch

Abstract

DNA methylation, and specifically the reversible addition of methyl groups at CpG dinucleotides genome-wide, represents an important layer that is associated with the regulation of gene expression. In particular, aberrations in the methylation status have been noted across a diverse set of pathological states, including cancer. With the rapid development and uptake of large scale sequencing of short DNA fragments, there has been an explosion of data analytic methods for processing and discovering changes in DNA methylation across diverse data types. In this mini-review, we aim to condense many of the salient challenges, such as experimental design, statistical methods for differential methylation detection and critical considerations such as cell type composition and the potential confounding that can arise from batch effects, into a compact and accessible format. Our main interests, from a statistical perspective, include the practical use of empirical Bayes or hierarchical models, which have been shown to be immensely powerful and flexible in genomics and the procedures by which control of false discoveries are made. Of course, there are many critical platform-specific data preprocessing aspects that we do not discuss here. In addition, we do not make formal performance comparisons of the methods, but rather describe the commonly used statistical models and many of the pertinent issues; we make some recommendations for further study.

1 Introduction

Epigenomics can be defined as the genome-wide investigation of stably heritable phenotypes resulting from changes in a chromosome without alterations in the DNA sequence [1]. DNA methylation (DNAme) is the most well-studied epigenetic mark and notably, the enzymatic mechanism for mitotically copying methylation status is well understood [2], unlike maintenance of chromatin state, for example [3]. In this review, we focus on differential methylation (DM) for methyl groups added to cytosines in the CpG dinucleotide context, since this is the predominant form observed in differentiated mammalian cells [4]. However, depending on the biological question and the technology used, some of the statistical methods discussed here can be applied more generally.

In the last decade, considerable progress in (observationally) characterizing epigenetic phenomena across a wide spectrum of normal and disease states has been made, predominantly due to emerging disruptive technologies, such as microarrays and large-scale sequencing of DNA. These studies have highlighted the important role of DNA methylation, perhaps for its causal associations with gene regulation, but also for its potential in diagnosing or stratifying patients according to their combined genomic/epigenomic molecular state. As mentioned, emerging technologies are driving these large-scale explorative studies and correspondingly, robust and efficient statistical and computational frameworks must be developed to facilitate interpretation of the growing masses of data. In the interrogation of CpG methylation, the main workhorse is treatment of DNA with sodium bisulphite [5], which preserves methylated cytosines while converting unmethylated cytosines to uracil. This transformation can allow high-throughput readouts, whether by hybridization of DNA to probes on a slide or through sequencing of DNA, to quantify the (relative) level of methylation (further discussion below).

While an individual's (pathologically normal) genome is almost completely static in all cells, the epigenome is highly dynamic both in time (e.g., through development) and across cell types. Since the epigenome is a combinatorial assembly of regulatory factors (e.g., DNA methylation, histone modifications, non-coding RNAs, etc.), comprehensively profiling the epigenome is orders of magnitude more difficult than its genomic counterpart. Therefore, accurately measuring DNA

methylation or other layers of the epigenome, unlike genome sequencing, require additional considerations to ensure that detected changes are not confounded with external factors, such as cell type.

Not surprisingly, the community has embraced consortium science to scale up data collection efforts. Some prominent projects that involve large-scale profiling of DNA methylation include the ENCODE Roadmap Epigenomics Consortium [6], The Cancer Genome Atlas and International Cancer Genome Consortium [7, 8], the BLUEPRINT project [9] and the International Human Epigenome Consortium [10].

2 Technologies

Present techniques for interrogating DNAme fall roughly into three categories: methylation-specific enzyme digestion (ED), affinity enrichment (AE) and chemical treatment with bisulphite (BS), in combination with a microarray or high-throughput sequencing readout; some of these techniques have been used in combination (e.g. ED+BS, commonly known as RRBS; see 11). There was some early demonstration of distinguishing methylcytosine from cytosine (and potentially other forms) with third-generation technologies [12], such as Pacific Biosciences, but no commercially viable offering has yet appeared. Importantly, depending on the interrogation technique and readout, DNAme information comes at low (∼100-200 base pair) or high resolution (individual CpG sites) and the costs vary widely. Each platform has its own limitations related to cost, resolution, scalability and the amount of starting DNA [13, 11, 14]. For example, ED studies remain dependent on the location and frequency of enzyme restriction sites, the prominent BS-based microarray platform is only available for human, the sensitivity of AE approaches depends on CpG density while genome-scale sequencing-based BS methods are costly and require sufficient computing resources. Depending on the biological question and resources available, a platform may be selected based on these tradeoffs.

Notably, BS-based methods cannot distinguish between methylcytosine and other variants, such as hydroxymethylcytosine [15], although additional treatment steps can readily allow this [16].

In principle, the methods we discuss below for determining differences in methylation status are agnostic to this technicality, aside from specific biological questions regarding the interplay between methylation states. Additionally, another biological phenomenon that we sidestep in this review is that of methylation in non-CpG contexts, which has shown to be prominent in pluripotent cells [4]. Interestingly, in a recent report using whole genome BS-seq (WGBS) data across a broad range of cell types and cell lines, it was shown that CpG methylation is only "dynamic" in approximately 20% of sites genome-wide, suggesting that the expense of WGBS could be more directed. For example, the combination of ED (and size selection) with BS-seq readout is already a favored method to reduce the amount of sequencing required, but would be difficult to tailor to specific genomic regions. An alternative reduced complexity strategy is to first capture fragments of interest, analogous to exome capture for sequencing protein-coding regions of genomic DNA; options are commercially available, such as the Agilent SureSelect system [17].

A popular, cost-efficient and scalable technology for profiling DNA methylation on a "genome-scale" is the Illumina 450k microarray. The platform can be thought of as genotyping BS-treated DNA to reveal the relative proportion of methylated and unmethylated alleles [18]. For every CpG site, measurements are either made with two separate physical beads (Type I) or through a single bead across two fluorescence channels (Type II); properties of these probe types are vastly different and require careful normalization [19, 20].

With the rapid advances and decreasing costs of single-base resolution DNA methylation data, AE techniques that capture methylated DNA fragments appear to have gone somewhat out of favor. Standard methylated DNA immunoprecipitation (MeDIP) or methyl-binding domain (MBD) enrichments share many features of chromatin immunoprecipitation experiments but are plagued by strong biases in enrichment according to CpG density, some of which can be fixed *in silico* [21, 22, 23]. Recently, a cost-effective combined MeDIP-seq and methylation-sensitive restriction enzyme sequencing approach (MRE-seq), termed M&M, has become available, promising to quickly compare methylomes at a fraction of the cost of WGBS-seq [24].

Table 1 summarizes the list of the methods reviewed and gives brief details on some of the important features of a statistical framework: i) the type of data that they operate on; ii) whether

they use predefined regions or are able to define differential regions themselves; iii) whether the methods support adjusting for covariates.

Method	Citation	Designed for	Determines regions or uses predefined	Accounts for covariates
Minfi	[20]	450k	determines	yes
IMA	[25]	450k	predefined	no
COHCAP	[26]	450k or BS-seq	predefined	yes
BSmooth	[27]	BS-seq	determines	no
DSS	[28]	BS-seq	determines	no
MOABS	[29]	BS-seq	determines	no
BiSeq	[30]	BS-seq	determines	yes
DMAP	[31]	BS-seq	predefined	yes
methylKit	[32]	BS-seq	predefined	yes
RADMeth	[33]	BS-seq	determines	yes
Bumphunter	[34]	General	determines	yes
ABCD-DNA	[35]	MeDIP-seq	predefined	yes
DiffBind	[36]	MeDIP-seq	predefined	yes
M&M	[24]	MeDIP-seq+MRE-seq	determines	no

3 Experimental design

Ultimately, the same experimental design concepts that apply broadly to any scientific investigation, such as sampling, randomization and blocking, are assumed. In experiments involving well-controlled cell culture conditions, populations of cells are typically uniform in their methylation status and no additional considerations are necessary. Although some single-cell DNA methylation studies are beginning to emerge [37], it is crucial to remember, especially in human epigenome-wide association studies (EWAS) that every experimental unit represents a population of cells. This implies that a consensus methylation estimate of 50% at any site or region could mean 50% of the alleles are methylated in all cells (e.g., allele-specific methylation) or 50% of the cells (e.g., mixtures of cell types) are fully methylated or various combinations in between. In particular, only BS-seq data can properly decompose this information through the methylation status from individual DNA fragments and at the same time, infer allele-specific patterns [38, 39]. But, there are limits: since small fragments are observed, it still remains challenging, without additional haplotype information

to relate the allelic methylation status at one loci to another genomically distant loci [40].

Unlike in genome sequencing studies, it is important in epigenome profiling projects to collect relevant populations of cell types. Many population-scale profiling studies may consider using readily accessible bodily fluids, such as blood, which itself represents a rich milieu of diverse cell types that may vary in their composition across the experimental units being studied. If it is known what cell types are of interest and cell surface markers are known, it may be beneficial to first sort populations of cells into subpopulations and profile each subpopulation individually [41]. Doing so will give a more focused interrogation of methylation and improved signal over noise. However, there are many situations where pre-sorting is not possible. Importantly, profiling mixtures of cell types and looking for changes in DNA methylation can be misleading when the cell composition is associated with an external factor, such as age of the patient [42]. However, there are now various emerging computational techniques to (estimate and) deconvolute the cell composition signals *in silico* (see below).

Another design consideration for BS-seq experiments is whether to spend money towards deeper sequencing or towards additional replicates. Because of the local smoothing frameworks available (e.g., BSmooth, 43), it is indeed considered more beneficial to sequence more replicates than to gather deep information on fewer samples, analogous to the available advice – replicates preferred over depth – for differential expression in RNA sequencing experiments [44].

4 Finding differential loci

For single-base resolution assays (BS-seq, 450k array), we first focus on the methodology for discovering individual differentially methylated CpG **sites** using statistical criteria. BS-seq data can be summarized as counts of methylated and unmethylated reads at any given site. Many early BS-seq studies profiled cells in the absence of replicates and used Fisher's exact test (FET) to discern DM [45]. While this strategy may be sufficient in comparing cell lines, we stress that the use of FET should be avoided in the general case; most systems have inherent biological variation and FET does not account for it. For example, in a two-condition comparison, FET requires the data to

be condensed to counts for each condition, which completely ignores the within-condition variability. This will underestimate the variability and overstate the evidence of the detected differential regions, leading to a high false positive rate. Likewise, using the binomial distribution, such as a logistic regression framework (e.g., methylKit; 32), also does not facilitate estimation of biological variability, unless an overdispersion term is used. While BSmooth uses a "signal-to-noise" statistic to quantify DM evidence at CpG site, this statistic is not used directly for inference; it is used within a framework to discover differential regions (see below).

Perhaps the most natural statistical model for replicated BS-seq DNA methylation measurements is the beta-binomial (BB) distribution. Conditional on the methylation proportion at a particular site, the observations are binomial distributed, while the methylation proportion itself can vary, according to a beta distribution, across experimental units (e.g., patients). It is therefore no surprise that BB assumptions are made in several recently proposed packages, such as BiSeq [30], MOABS [29], DSS [28] and RADMeth [33]. Similarly, empirical Bayes (EB) methods fit naturally for modeling and inference across many types of genomic data and DNA methylation assays are no different. MOABS and DSS both implement hierarchical models and use the whole dataset to estimate the hyperparameters of the beta distribution; RADMeth and BiSeq use standard maximum likelihood without any moderation. While BiSeq and RADMeth do not moderate parameter estimates, they provide facilities for complex designs through design matrices, which MOABS and DSS do not currently offer. Inference for parameters of interest (i.e., changes in methylation) are conducted using standard techniques, such as Wald tests (DSS, BiSeq) and likelihood ratio tests (RADMeth). Notably, MOABS introduces a new metric, called credible methylation difference, which is a conservative estimate of the true methylation difference, calibrated by the statistical evidence available.

DNA methylation arrays, such as Illumina's 27k or 450k array, give fluorescence intensities that quantify relative abundance of methylated and unmethylated loci, in contrast to the count-based modeling assumptions for BS-seq based profiling. In particular, the data used for downstream analyses can be either i) log-ratios of methylated to unmethylated intensities, analogous to two-channel expression arrays, or ii) the so-called beta-value, which gives the ratio of the methylated to the

total of methylated and unmethylated intensities. Previous comparisons suggest that statistical inferences based on log-ratios are preferred [46], perhaps not surprisingly since they can rely on the successful moderated statistical testing strategies used for expression arrays (e.g., limma; 47). Much of the recent effort for the 450k array has been dedicated to normalization and filtering (e.g., 20, 48) and despite this early knowledge of statistical inferences on log-ratios versus beta-values, various options for inferring DM sites from 27k/450k array data have been proposed. To test for DM, IMA proposes Wilcoxon rank-sum tests on beta-values [25]. Similarly, COHCAP is a method that operates both on methylation array data or BS-seq data, using beta-values or methylation proportions, respectively, as input; they offer FET (see comment above), t-tests and ANOVA analyses (without moderation), depending on the study design [26]. Ultimately, our intuition suggests that moderated t/F-statistics on the normalized log-ratios of intensities seems most rigorous.

Many of the packages mentioned above and below provide tools to integrate with other genomic layers, such as annotation databases, however these additional functionalities are not reviewed here.

5 Finding differential regions

Although there are examples where researchers are interested in relating single CpG sites to a phenotype (e.g., 49), many methods, including downstream analysis of the statistics discussed in the last section, are designed to detect DMRs as a more representative predictive (or diagnostic/prognostic) feature. Another advantage is that while methylation differences at any individual site may be small, if they are persistent across a region, there may be greater statistical power to detect them at the region-level. In this space, one must distinguish between methods that *operate* on predefined regions, with those that *define* regions of DM. The latter is considerably more difficult because ensuring control of the false discovery rate (FDR) at the region-level is non-trivial; in particular, controlling false discoveries at the site-level does not give a direct way to controlling false discoveries at the region-level when the region itself is also to be defined [50].

Therefore, the most straightforward approach is to use predefined regions, such as CpG islands, CpG shores, UTRs and so on; statistical testing can be conducted fairly routinely at a region-level.

Many of the packages mentioned above, such as IMA, COHCAP, DMAP and methylKit, do exactly this. A special case is DMAP, which can operate on fragments (using the sampled MspI-digested fragments as the region of interest) or according to predefined regions [31].

There are now a wide array of proposals for *defining* DMRs. Developed originally for specialized methylation microarrays, Bumphunter can be applied quite generally across data types [34], perhaps after transformation in the case of count data. Notably, it also integrates a surrogate variable analysis [51] to simultaneously account for potential batch effects while permutation tests are used to assign FDR at the region-level; notably, users are required to set a smoothing window size and a threshold on the percentile of the smoothed effect sizes (or t-statistics) [34]. Similarly, BSmooth searches for runs of smoothed absolute t-like scores beyond a threshold, however, does not suggest a permutation strategy to control region-level FDR. Also from the same authors, minfi wraps bumphunting into the suite of methods available for Illumina 450k arrays; in addition, a module for *block finding*, essentially bumphunting with a much greater window size (e.g., 250kbp), is also made available [20]. BiSeq proposes, via a Wald test statistic from the beta-binomial regression fit, a hierarchical testing strategy that first considers target regions and controls error using a cluster-wise weighted FDR strategy [30]; secondly, the differential clusters are trimmed using a second stage of testing, analogous to that used for spatial signals [52]. A clustering method, A-clust, proposes first to cluster CpG sites according to correlation in methylation signal across samples; within the clusters, associations can be modeled with correlated error and fit using a generalized estimating equation framework [53]. The DSS authors simply set some thresholds on the P-values, number of CpG sites and length of regions, but they do not pursue a FDR control story [28]. The MOABS authors suggest grouping DM sites into DMRs using a hidden Markov model or alternatively testing of predefined regions, but no specific details are given. Most recently, RADMeth proposes a transformation of P-values (from a standard likelihood ratio test) into a weighted Z test that builds in the correlation of neighboring probes [33].

Enrichment assays, such as MBD-seq and MeDIP-seq, are by their very nature of capturing fragments, only capable of finding *regions* of DM. Such packages, including MEDIPS [54, 55], ABCD-DNA [35] and DiffBind [36], compare relative abundance of fragment counts by repackaging

RNA sequencing statistical inference frameworks. In a related assay, the M&M algorithm models normalized methylated counts (MeDIP-seq) and unmethylated counts (MRE-seq) as jointly Poisson distributed, for non-overlapping genomic bins, with a shared parameter [24]. However, analogous to the FET, the Poisson model does not account for biological variability [24].

6 Reducing the impact of batch, cell type composition or other confounding effects

As briefly mentioned above, researchers need to carefully design studies that associate phenotypes with DNA methylation (e.g., using blood). However, some aspects, such as cell type composition cannot be readily controlled by design; patients and therefore individual DNA samples simply differ in their cell type composition. A recent report has highlighted that many of the DNA methylation markers that have been associated to age, are actually driven by changes in cell composition, which are also associated with age [42]. Whole blood is a mixture of several cell types; using an independent dataset of methylation profiles of the dominant cell types (Monocytes, CD4+ and CD8+ T cells, Granulocytes, B cells, natural killer cells) from flow sorting, patient profiles were deconvoluted using a reimplementation of the Houseman algorithm [42, 41]. First, from the methylation profiles of pure cell populations, cell-type-specific markers are selected and then used to "calibrate" a regression model that associates methylation observations to a response of interest [41, 56]. Of course, this approach requires advance knowledge of the dominant cell types and methylation profiles for them, preferably across multiple replicates to seed the deconvolution algorithm with appropriate methylation markers. However, a recent study has highlighted that advanced statistical modeling can correct for cell type composition without the need for pure-cell profiles; starting from uncorrected standard model fits, the method regresses principal components within a linear mixed model until control for the inflation of the test statistics (e.g., relative to a uniform distribution of P-values) is achieved [57].

7 Discussion

In this review, we briefly explored the various methodologies available for deciphering differentially methylated regions across the main data types and highlighted some of the common themes and current challenges. The tradeoffs made by method developers are apparent. In fact, its a lot to ask of a single statistical framework to be able to do everything: moderate parameter estimates using genome-wide information or accurately and robustly smooth local estimates, accommodate low coverage data, account for batch effects and cell type composition, allow complex experimental designs and accurately control the FDR at the site- and/or region-level. In addition, identification of DMRs is only the discovery step; validating these detections, perhaps by associating them with other biological outcomes *in silico* requires additional frameworks, some of which have already been integrated alongside the packages reviewed here.

On the statistical and computational side, the field is moving fast and several advanced methods have been proposed. One of the next challenges will be to comprehensively compare method performance, in terms of power to detect changes and ability to control FDR, robustness and scalability to large datasets (e.g., whole genome BS-seq) and large studies (e.g., EWAS). Representative simulation frameworks will be fundamental for this task. Given the large number of methods available, this will already be a large undertaking. To avoid bias, these comparisons should be done either independently of the method development process, or collectively with all method developers. Advanced deconvolution algorithms and batch effect removal strategies are, at present, targeted at 450k array data. The development and vetting of similar techniques that can be readily applied to count data, such as BS-seq data, are well underway [58, 59].

Disclosure/Conflict-of-Interest Statement

The authors declare no conflicts of interest.

Author Contributions

MDR wrote the main text with critical contributions from all co-authors: AK, CWL, HL, MN, LMW and XZ. All authors read and approved the final manuscript.

Acknowledgement

We would like to thank various members of the Twitterverse, and in particular: @timtriche and @PeteHaitch for suggesting additional citations. We also wish to thank colleagues for reading an earlier version of this review posted on arXiv.

Funding MDR acknowledges financial support from SNSF project grant (143883) and from the European Commission through the 7th Framework Collaborative Project RADIANT (Grant Agreement Number: 305626).

References

[1] Shelley L Berger, Tony Kouzarides, Ramin Shiekhattar, and Ali Shilatifard. An operational definition of epigenetics. *Genes & development*, 23(7):781–783, 2009.

[2] Adrian Bird. DNA methylation patterns and epigenetic memory. *Genes & development*, 16(1):6–21, January 2002.

[3] Danesh Moazed. Mechanisms for the inheritance of chromatin states. *Cell*, 146(4):510–518, 2011.

[4] Katherine E Varley, Jason Gertz, Kevin M Bowling, Stephanie L Parker, Timothy E Reddy, Florencia Pauli-Behn, Marie K Cross, Brian a Williams, John a Stamatoyannopoulos, Gregory E Crawford, Devin M Absher, Barbara J Wold, and Richard M Myers. Dynamic DNA methylation across diverse human cell lines and tissues. *Genome research*, 23(3):555–67, 2013.

[5] Susan J Clark, Aaron Statham, Clare Stirzaker, Peter L Molloy, and Marianne Frommer. DNA methylation: bisulphite modification and analysis. *Nature protocols*, 1(5):2353–64, January 2006.

[6] Bradley E Bernstein, John a Stamatoyannopoulos, Joseph F Costello, Bing Ren, Aleksandar Milosavljevic, Alexander Meissner, Manolis Kellis, Marco A Marra, Arthur L Beaudet, Joseph R Ecker, Peggy J Farnham, Martin Hirst, Eric S Lander, Tarjei S Mikkelsen, and James a Thomson. The NIH Roadmap Epigenomics Mapping Consortium. *Nature Biotechnology*, 28(10):1045–8, October 2010.

[7] Lynda Chin, Jannik N Andersen, and P Andrew Futreal. Cancer genomics: from discovery science to personalized medicine. *Nature medicine*, 17(3):297–303, 2011.

[8] Thomas J Hudson, Warwick Anderson, Axel Artez, Anna D Barker, Cindy Bell, Rosa R Bernabé, M K Bhan, Fabien Calvo, Iiro Eerola, Daniela S Gerhard, Alan Guttmacher, Mark Guyer, Fiona M Hemsley, Jennifer L Jennings, David Kerr, Peter Klatt, Patrik Kolar, Jun Kusada, David P Lane, Frank Laplace, Lu Youyong, Gerd Nettekoven, Brad Ozenberger, Jane Peterson, T S Rao, Jacques Remacle, Alan J Schafer, Tatsuhiro Shibata, Michael R Stratton, Joseph G Vockley, Koichi Watanabe, Huanming Yang, Matthew M F Yuen, Bartha M Knoppers, Martin Bobrow, Anne Cambon-Thomsen, Lynn G Dressler, Stephanie O M Dyke, Yann Joly, Kazuto Kato, Karen L Kennedy, Pilar Nicolás, Michael J Parker, Emmanuelle Rial-Sebbag, Carlos M Romeo-Casabona, Kenna M Shaw, Susan Wallace, Georgia L Wiesner, Nikolajs Zeps, Peter Lichter, Andrew V Biankin, Christian Chabannon, Lynda Chin, Bruno Clément, Enrique de Alava, Françoise Degos, Martin L Ferguson, Peter Geary, D Neil Hayes, Amber L Johns, Arek Kasprzyk, Hidewaki Nakagawa, Robert Penny, Miguel a Piris, Rajiv Sarin, Aldo Scarpa, Marc van de Vijver, P Andrew Futreal, Hiroyuki Aburatani, Mónica Bayés, David D L Botwell, Peter J Campbell, Xavier Estivill, Sean M Grimmond, Ivo

Gut, Martin Hirst, Carlos López-Otín, Partha P Majumder, Marco Marra, John D McPherson, Zemin Ning, Xose S Puente, Yijun Ruan, Hendrik G Stunnenberg, Harold Swerdlow, Victor E Velculescu, Richard K Wilson, Hong H Xue, Liu Yang, Paul T Spellman, Gary D Bader, Paul C Boutros, Paul Flicek, Gad Getz, Roderic Guigó, Guangwu Guo, David Haussler, Simon Heath, Tim J Hubbard, Tao Jiang, Steven M Jones, Qibin Li, Nuria López-Bigas, Ruibang Luo, Lakshmi Muthuswamy, B F Francis Ouellette, John V Pearson, Victor Quesada, Benjamin J Raphael, Chris Sander, Terence P Speed, Lincoln D Stein, Joshua M Stuart, Jon W Teague, Yasushi Totoki, Tatsuhiko Tsunoda, Alfonso Valencia, David a Wheeler, Honglong Wu, Shancen Zhao, Guangyu Zhou, Mark Lathrop, Gilles Thomas, Teruhiko Yoshida, Myles Axton, Chris Gunter, Linda J Miller, Junjun Zhang, Syed a Haider, Jianxin Wang, Christina K Yung, Anthony Cross, Yong Liang, Saravanamuttu Gnaneshan, Jonathan Guberman, Jack Hsu, Don R C Chalmers, Karl W Hasel, Terry S H Kaan, William W Lowrance, Tohru Masui, Laura Lyman Rodriguez, Catherine Vergely, David D L Bowtell, Nicole Cloonan, Anna DeFazio, James R Eshleman, Dariush Etemadmoghadam, Brooke a Gardiner, James G Kench, Robert L Sutherland, Margaret a Tempero, Nicola J Waddell, Peter J Wilson, Steve Gallinger, Ming-Sound Tsao, Patricia a Shaw, Gloria M Petersen, Debabrata Mukhopadhyay, Ronald a DePinho, Sarah Thayer, Kamran Shazand, Timothy Beck, Michelle Sam, Lee Timms, Vanessa Ballin, Youyong Lu, Jiafu Ji, Xiuqing Zhang, Feng Chen, Xueda Hu, Qi Yang, Geng Tian, Lianhai Zhang, Xiaofang Xing, Xianghong Li, Zhenggang Zhu, Yingyan Yu, Jun Yu, Jörg Tost, Paul Brennan, Ivana Holcatova, David Zaridze, Alvis Brazma, Lars Egevard, Egor Prokhortchouk, Rosamonde Elizabeth Banks, Mathias Uhlén, Juris Viksna, Fredrik Ponten, Konstantin Skryabin, Ewan Birney, Ake Borg, Anne-Lise Bø rresen Dale, Carlos Caldas, John a Foekens, Sancha Martin, Jorge S Reis-Filho, Andrea L Richardson, Christos Sotiriou, Giles Thoms, Laura van't Veer, Daniel Birnbaum, Hélène Blanche, Pascal Boucher, Sandrine Boyault, Jocelyne D Masson-Jacquemier, Iris Pauporté, Xavier Pivot, Anne Vincent-Salomon, Eric Tabone, Charles Theillet, Isabelle Treilleux, Paulette Bioulac-Sage, Thomas Decaens, Dominique Franco, Marta Gut, Didier Samuel, Jessica Zucman-Rossi, Roland Eils, Benedikt Brors, Jan O Korbel, Andrey Korshunov, Pablo Landgraf, Hans Lehrach, Stefan Pfister, Bernhard Radlwimmer, Guido Reifenberger, Michael D Taylor, Christof von Kalle, Paolo Pederzoli, Rita a Lawlor, Massimo Delledonne, Alberto Bardelli, Thomas Gress, David Klimstra, Giuseppe Zamboni, Yusuke Nakamura, Satoru Miyano, Akihiro Fujimoto, Elias Campo, Silvia de Sanjosé, Emili Montserrat, Marcos González-Díaz, Pedro Jares, Heinz Himmelbaue, Silvia Bea, Samuel Aparicio, Douglas F Easton, Francis S Collins, Carolyn C Compton, Eric S Lander, Wylie Burke, Anthony R Green, Stanley R Hamilton, Olli P Kallioniemi, Timothy J Ley, Edison T Liu, Brandon J Wainwright, The International Cancer Genome Consortium, A Introduction, and B Consortium Goals. International network of cancer genome projects. *Nature*, 464(7291):993–8, April 2010.

[9] Joost HA H.A. Martens and Hendrik G. Stunnenberg. BLUEPRINT: mapping human blood cell epigenomes. *Haematologica*, 98(10):1487–1489, 2013.

[10] J B Bae. Perspectives of international human epigenome consortium. *Genomics Inform*, 11(1):7–14, 2013.

[11] Peter W Laird. Principles and challenges of genome-wide DNA methylation analysis. *Nature Reviews Genetics*, 11(3):191–203, March 2010.

[12] Benjamin A Flusberg, Dale R Webster, Jessica H Lee, Kevin J Travers, Eric C Olivares, Tyson A Clark, Jonas Korlach, and Stephen W Turner. Direct detection of DNA methylation during single-molecule, real-time sequencing. *Nature Methods*, 7(6):461–465, 2010.

[13] Christoph Bock. Analysing and interpreting DNA methylation data. *Nature Reviews Genetics*, 13(10):705–19, October 2012.

[14] Mark D Robinson, Aaron L Statham, Terence P Speed, and Susan J Clark. Protocol matters : which methylome are you actually studying ? *Epigenomics*, 2(4):587–598, August 2010.

[15] Yun Huang, William A. Pastor, Yinghua Shen, Mamta Tahiliani, David R. Liu, and Anjana Rao. The behaviour of 5-hydroxymethylcytosine in bisulfite sequencing. *PLoS ONE*, 5(1), 2010.

[16] M. J. Booth, M. R. Branco, G. Ficz, D. Oxley, F. Krueger, W. Reik, and S. Balasubramanian. Quantitative Sequencing of 5-Methylcytosine and 5-Hydroxymethylcytosine at Single-Base Resolution, 2012.

[17] S. T. Borno, A. Fischer, M. Kerick, M. Falth, M. Laible, J. C. Brase, R. Kuner, A. Dahl, C. Grimm, B. Sayanjali, M. Isau, C. Rohr, A. Wunderlich, B. Timmermann, R. Claus, C. Plass, M. Graefen, R. Simon, F. Demichelis, M. A. Rubin, G. Sauter, T. Schlomm, H. Sultmann, H. Lehrach, and M. R. Schweiger. Genome-wide DNA Methylation Events in TMPRSS2-ERG Fusion-Negative Prostate Cancers Implicate an EZH2-Dependent Mechanism with miR-26a Hypermethylation, 2012.

[18] Ruth Pidsley, Chloe C Y Wong, Manuela Volta, Katie Lunnon, Jonathan Mill, Leonard C Schalkwyk, and Chloe C Y Wong. A data-driven approach to preprocessing Illumina 450K methylation array data. *BMC genomics*, 14(1):293, May 2013.

[19] Jovana Maksimovic, Lavinia Gordon, Alicia Oshlack, and Jovana Makismovic. SWAN: Subset-quantile within array normalization for illumina infinium HumanMethylation450 BeadChips. *Genome Biology*, 13(6):R44, January 2012.

[20] Martin J Aryee, Andrew E Jaffe, Hector Corrada-Bravo, Christine Ladd-Acosta, Andrew P Feinberg, Kasper D Hansen, and Rafael a Irizarry. Minfi: A flexible and comprehensive Bioconductor package for the analysis of Infinium DNA Methylation microarrays. *Bioinformatics (Oxford, England)*, pages 1–8, 2014.

[21] Andrea Riebler, Mirco Menigatti, Jenny Z Song, Aaron L Statham, Clare Stirzaker, Nadiya Mahmud, Charles a Mein, Susan J Clark, and Mark D Robinson. BayMeth: improved DNA methylation quantification for affinity capture sequencing data using a flexible Bayesian approach. *Genome biology*, 15(2):R35, 2014.

[22] Mattia Pelizzola, Yasuo Koga, Alexander Eckehart Urban, Michael Krauthammer, Sherman Weissman, Ruth Halaban, and Annette M Molinaro. MEDME: An experimental and analytical methodology for the estimation of DNA methylation levels based on microarray derived MeDIP-enrichment. *Genome Research*, 18(10):1652–1659, 2008.

[23] Thomas A Down, Vardhman K Rakyan, Daniel J Turner, Paul Flicek, Heng Li, Eugene Kulesha, Stefan Gräf, Nathan Johnson, Javier Herrero, Eleni M Tomazou, Natalie P Thorne, Liselotte Bäckdahl, Marlis Herberth, Kevin L Howe, David K Jackson, Marcos M Miretti, John C Marioni, Ewan Birney, Tim J P Hubbard, Richard Durbin, Simon Tavaré, and Stephan Beck. A Bayesian deconvolution strategy for immunoprecipitation-based DNA methylome analysis. *Nature biotechnology*, 26(7):779–85, July 2008.

[24] B Zhang, Y Zhou, N Lin, R F Lowdon, C Hong, R P Nagarajan, J B Cheng, D Li, M Stevens, H J Lee, X Xing, J Zhou, V Sundaram, G Elliott, J Gu, T Shi, P Gascard, M Sigaroudinia, T D Tlsty, T Kadlecek, A Weiss, H O'Geen, P J Farnham, C L Maire, K L Ligon, P A Madden, A Tam, R Moore, M Hirst, M A Marra, J F Costello, and T Wang. Functional DNA methylation differences between tissues, cell types, and across individuals discovered using the M&M algorithm. *Genome Res*, 23(9):1522–1540, 2013.

[25] Dan Wang, Li Yan, Qiang Hu, Lara E. Sucheston, Michael J. Higgins, Christine B. Ambrosone, Candace S. Johnson, Dominic J. Smiraglia, and Song Liu. IMA: An R package for high-throughput analysis of Illumina's 450K Infinium methylation data. *Bioinformatics*, 28(5):729–730, 2012.

[26] Charles D. Warden, Heehyoung Lee, Joshua D. Tompkins, Xiaojin Li, Charles Wang, Arthur D. Riggs, Hua Yu, Richard Jove, and Yate Ching Yuan. COHCAP: An integrative genomic pipeline for single-nucleotide resolution DNA methylation analysis. *Nucleic Acids Research*, 41(11), 2013.

[27] Kasper D Hansen, Rafael A Irizarry, and Zhijin Wu. Removing technical variability in RNA-seq data using conditional quantile normalization. *Biostatistics*, 13(2):204–216, 2012.

[28] Hao Feng, Karen N Conneely, and Hao Wu. A Bayesian hierarchical model to detect differentially methylated loci from single nucleotide resolution sequencing data. *Nucleic acids research*, 42(8):e69, 2014.

[29] Deqiang Sun, Yuanxin Xi, Benjamin Rodriguez, Hyun Jung Park, Pan Tong, Mira Meong, Margaret a Goodell, and Wei Li. MOABS: model based analysis of bisulfite sequencing data. *Genome biology*, 15(2):R38, 2014.

[30] Katja Hebestreit, Martin Dugas, and Hans Ulrich Klein. Detection of significantly differentially methylated regions in targeted bisulfite sequencing data. *Bioinformatics*, 29(13):1647–1653, 2013.

[31] Peter a Stockwell, Aniruddha Chatterjee, Euan J Rodger, and Ian M Morison. DMAP: differential methylation analysis package for RRBS and WGBS data. *Bioinformatics*, pages 1–9, 2014.

[32] Altuna Akalin, Matthias Kormaksson, Sheng Li, Francine E Garrett-Bakelman, Maria E Figueroa, Ari Melnick, and Christopher E Mason. methylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles. *Genome Biology*, 13(10):R87, 2012.

[33] Egor Dolzhenko and Andrew D Smith. Using beta-binomial regression for high-precision differential methylation analysis in multifactor whole-genome bisulfite sequencing experiments. *BMC bioinformatics*, 15:215, 2014.

[34] Andrew E Jaffe, Peter Murakami, Hwajin Lee, Jeffrey T Leek, M Daniele Fallin, Andrew P Feinberg, and Rafael A Irizarry. Bump hunting to identify differentially methylated regions in epigenetic epidemiology studies. *International journal of epidemiology*, 41(1):200–9, February 2012.

[35] Mark D Robinson, Dario Strbenac, Clare Stirzaker, Aaron L Statham, Jenny Song, Terence P Speed, and Susan J Clark. Copy-number-aware differential analysis of quantitative DNA sequencing data. *Genome Research*, 22(12):2489–96, December 2012.

[36] Caryn S Ross-Innes, Rory Stark, Andrew E Teschendorff, Kelly A Holmes, H Raza Ali, Mark J Dunning, Gordon D Brown, Ondrej Gojis, Ian O Ellis, Andrew R Green, Simak Ali, Suet-Feung Chin, Carlo Palmieri, Carlos Caldas, and Jason S Carroll. Differential oestrogen receptor binding is associated with clinical outcome in breast cancer. *Nature*, 2012.

[37] Hongshan Guo, Ping Zhu, Xinglong Wu, Xianlong Li, Lu Wen, and Fuchou Tang. Single-cell methylome landscapes of mouse embryonic stem cells and early embryos analyzed using reduced representation bisulfite sequencing. *Genome research*, 23(12):2126–35, 2013.

[38] Fang Fang, Emily Hodges, Antoine Molaro, Matthew Dean, Gregory J Hannon, and Andrew D Smith. Genomic landscape of human allele-specific DNA methylation. *Proceedings of the National Academy of Sciences of the United States of America*, 109(19):7332–7, May 2012.

[39] Qiang Song, Benjamin Decato, Elizabeth E. Hong, Meng Zhou, Fang Fang, Jianghan Qu, Tyler Garvin, Michael Kessler, Jun Zhou, and Andrew D. Smith. A reference methylome database and analysis pipeline to facilitate integrative and comparative epigenomics. *PLoS ONE*, 8(12), 2013.

[40] Volodymyr Kuleshov, Dan Xie, Rui Chen, Dmitry Pushkarev, Zhihai Ma, Tim Blauwkamp, Michael Kertesz, and Michael Snyder. Whole-genome haplotyping using long reads and statistical methods. *Nature biotechnology*, 32(3):261–6, 2014.

[41] EA Houseman, WP Accomando, DC Koestler, BC Christensen, CJ Marsit, HH Nelson, JK Wiencke, and KT Kelsey. DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC bioinformatics*, 13:86, 2012.

[42] A E Jaffe and R A Irizarry. Accounting for cellular heterogeneity is critical in epigenome-wide association studies. *Genome Biol*, 15(2):R31, 2014.

[43] Kasper D Hansen, Benjamin Langmead, and Rafael A Irizarry. BSmooth: from whole genome bisulfite sequencing reads to differentially methylated regions, 2012.

[44] Franck Rapaport, Raya Khanin, Yupu Liang, Mono Pirun, Azra Krek, Paul Zumbo, Christopher E Mason, Nicholas D Socci, and Doron Betel. Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. *Genome biology*, 14(9):R95, 2013.

[45] Ryan Lister, Mattia Pelizzola, Robert H Dowen, R David Hawkins, Gary Hon, Julian Tonti-Filippini, Joseph R Nery, Leonard Lee, Zhen Ye, Que-Minh Ngo, Lee Edsall, Jessica Antosiewicz-Bourget, Ron Stewart, Victor Ruotti, A Harvey Millar, James A Thomson, Bing Ren, and Joseph R Ecker. Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature*, 462(7271):315–22, November 2009.

[46] Pan Du, Xiao Zhang, Chiang-Ching Huang, Nadereh Jafari, Warren A Kibbe, Lifang Hou, and Simon M Lin. Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. *BMC bioinformatics*, 11(1):587, January 2010.

[47] G K Smyth. Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology*, 3(1):Article3, 2004.

[48] Magda E Price, Allison M Cotton, Lucia L Lam, Pau Farré, Eldon Emberly, Carolyn J Brown, Wendy P Robinson, and Michael S Kobor. Additional annotation enhances potential for biologically-relevant analysis of the Illumina Infinium HumanMethylation450 BeadChip array. *Epigenetics & chromatin*, 6(1):4, January 2013.

[49] Ian C G Weaver, Nadia Cervoni, Frances A Champagne, Ana C D'Alessio, Shakti Sharma, Jonathan R Seckl, Sergiy Dymov, Moshe Szyf, and Michael J Meaney. Epigenetic programming by maternal behavior. *Nature neuroscience*, 7(8):847–854, 2004.

[50] Aaron TL Lun and Gordon K Smyth. De novo detection of differentially bound regions for ChIP-seq data using peaks and windows: controlling error rates correctly. *Nucleic Acids Research*, 42(11):e95, 2014.

[51] Jeffrey T Leek and John D Storey. Capturing Heterogeneity in Gene Expression Studies by Surrogate Variable Analysis. *PLoS Genetics*, 3(9):12, 2007.

[52] Yoav Benjamini and Ruth Heller. False Discovery Rates for Spatial Signals. *Journal of the American Statistical Association*, 102(480):1272–1281, 2007.

[53] Tamar Sofer, Elizabeth D. Schifano, Jane A. Hoppin, Lifang Hou, and Andrea A. Baccarelli. A-clustering: A novel method for the detection of co-regulated methylation regions, and regions associated with exposure. *Bioinformatics*, 29(22):2884–2891, 2013.

[54] Lukas Chavez, Justyna Jozefczuk, Christina Grimm, Jörn Dietrich, Bernd Timmermann, Hans Lehrach, Ralf Herwig, and James Adjaye. Computational analysis of genome-wide DNA methylation during the differentiation of human embryonic stem cells along the endodermal lineage. *Genome research*, 20(10):1441–50, October 2010.

[55] Matthias Lienhard, Christina Grimm, Markus Morkel, Ralf Herwig, and Lukas Chavez. MEDIPS: Genome-wide differential coverage analysis of sequencing data derived from DNA enrichment experiments. *Bioinformatics*, 30(2):284–286, 2014.

[56] Raymond J Carroll. *Measurement error in nonlinear models: a modern perspective*. Chapman & Hall/CRC, Boca Raton, 2006.

[57] J Zou, C Lippert, D Heckerman, M Aryee, and J Listgarten. Epigenome-wide association studies without the need for cell-type composition. *Nat Methods*, 11(3):309–311, 2014.

[58] Jeffrey Leek. svaseq: removing batch effects and other unwanted noise from sequencing data. *bioRxiv*, page http://dx.doi.org/10.1101/006585, 2014.

[59] Davide Risso, J Ngai, Terence P Speed, and Dudoit Sandrine. Normalization of RNA-seq data using factor analysis of control genes or samples. *Nature Biotechnology*, page in press., 2014.