Accurate measurement of DNA methylation: Challenges and bias correction

Eguzkine Ochoa[1,2], Verena Zuber[3,4] and Leonardo Bottolo[1,4,5,*]

[1]Department of Medical Genetics, University of Cambridge, Cambridge CB2 0QQ, UK [2]Cambridge NIHR Biomedical Research Centre, Cambridge CB2 0QQ, UK [3]Department of Epidemiology and Biostatistics, Imperial College London, London W2 1PG, UK [4]MRC Biostatistics Unit, University of Cambridge, Cambridge CB2 0SR, UK [5]The Alan Turing Institute, London NW1 2DB, UK *corresponding author: lb664@cam.ac.uk

Abstract

DNA methylation is a key epigenetic modification involved in gene regulation whose contribution to disease susceptibility is still not fully understood. As the cost of genome sequencing technologies continues to drop, it will soon become commonplace to perform genome-wide quantification of DNA methylation at a single base-pair resolution. However, the demand for its accurate quantification might vary across studies. When the scope of the analysis is to detect regions of the genome with different methylation patterns between two or more conditions, e.g., case vs control; treatments vs placebo, accuracy is not crucial. This is the case in epigenome-wide association studies used as genome-wide screening of methylation changes to detect new candidate genes and regions associated with a specific disease or condition. If the aim of the analysis is to use DNA methylation measurements as a biomarker for diseases diagnosis and treatment [1, 2], it is instead recommended to produce accurate methylation measurements. Furthermore, if the objective is the detection of DNA methylation in subclonal tumour cell populations or in circulating tumour DNA or in any case of mosaicism, the importance of accuracy becomes critical. The aim of this chapter is to describe the factors that could affect the precise measurement of methylation levels and a recent Bayesian statistical method called MethylCal and its extension that have been proposed to minimize this problem.

Keywords: Incomplete bisulfite conversion; PCR, region and coverage bias; Calibration technical replicates; Corrected methylation degree; Differential calibrated analysis.

1. Introduction

The deviation between the observed and the expected methylation level is mainly due to the chemical treatment used to discriminate between methylated and unmethylated cytosine. The detection of a methyl group at the 5'-carbon of cytosines in a CpG dinucleotide is challenging because the methylation marks will be lost after one cycle of PCR and the limitation of the large majority of high-throughput technologies to detect only four bases (C, T, G, A).

In 1970, two independent groups Shapiro et al. [3] and Hayatsu et al. [4] discovered a chemical interaction between sodium bisulfite and pyrimidines capable to deaminate cytosine into uracil. Ten years later, Wang et al. [5] observed significant differences in the speed of this reaction between 5-methylcytosine (5mC) and cytosines (C). Inspired by these works, Frommer et al. [6] proposed in 1992 to use the reaction to distinguish 5mC from C in the analysis of DNA methylation levels in genomic DNA. The deamination into uracil is much faster at cytosine residues than 5mC. Since then, this chemical treatment is called "bisulfite conversion" and consists in the conversion of unmethylated cytosines into uracil, maintaining methylated cytosines unchanged. The conversion begins with the denaturalization of the genomic DNA at 95°C in order to separate DNA strands and make cytosines accessible to bisulfite treatment (**Fig. 1**). After denaturalization, single strand genomic DNA suffers a nucleophilic addition of bisulfite to the C-6 position of cytosine, allowing the deamination into 5,6-dihydrouracil-6-sulfonate. The treatment with an alkaline solution eliminates the sulfonate group and regenerates the double bond, obtaining uracil (**Fig. 1**). Uracil will be converted into thymidine after one polymerase chain reaction cycle. Finally, after bisulfite conversion, DNA becomes single strand fragmented DNA no longer complementary.

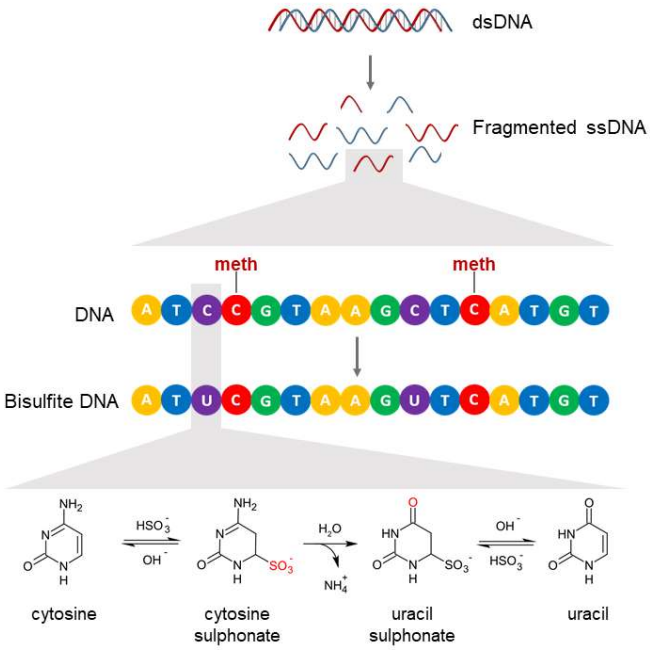

Fig. 1 Bisulfite conversion reaction. During the treatment of double-strand genomic DNA with bisulfite, DNA suffers denaturalization and conversion of unmethylated cytosines in uracil. This conversion begins with the addition of bisulfite to cytosine producing cytosine sulphonate, which is deaminated into uracil sulfonate. After the treatment with an alkaline solution, uracil sulfonate is converted into uracil.

The revolution of bisulfite conversion treatment in the detection of DNA methylation has led in recent years to an increasing number of DNA methylation studies as well as the identification of new epigenetic biomarkers and epigenetic signatures associated with specific diseases. All high-throughput technologies, capable to detect DNA methylation, including pyrosequencing, amplicon-based bisulfite sequencing, capture-based bisulfite sequencing, genome-wide methylation array, whole-genome bisulfite sequencing (WGBS) and reduced representation bisulfite sequencing (RRBS) use bisulfite conversion for the discrimination of methylated cytosine from unmethylated cytosines. However, the bisulfite conversion of the genomic DNA also introduces some technical problems in the measurement of methylation levels that should be taken into account in downstream analyses.

1.1. Incomplete bisulfite conversion

During bisulfite conversion, only unmethylated cytosines will be transformed in uracil, so any incomplete bisulfite conversion will be read in the downstream analysis as methylated cytosines, a false positive. The main reason for incomplete bisulfite conversion is the incomplete denaturation or partial renaturation. Since double-strand DNA is unreactive to bisulfite, an incomplete denaturation will result in cytosine fail to react with bisulfite instead of the conversion of unmethylated cytosines into uracil (**Fig. 2**).

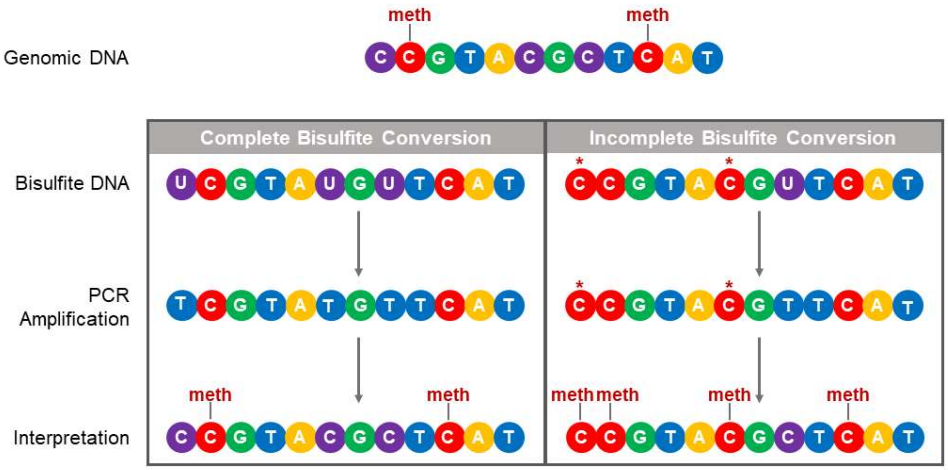

Fig. 2 Incomplete bisulfite conversion (left) produces a different sequence of bisulfite DNA compared to the sequence produced after complete conversion (right). Also, the PCR product

obtained after PCR amplification will have a different sequence. In the example, two unmethylated cytosines (red asterisk) elude the conversion and produce bias results that lead to misinterpretation in downstream analyses. In this case, the incomplete bisulfite conversion of two cytosines will be read as two false methylated cytosines.

Commercial bisulfite conversion kits like EZ DNA Methylation-Gold (Zymo©) assures a conversion efficiency rate greater than 99%. The conversion rate is calculated as the number of non-CpG cytosines converted into thymidine over the total number of non-CpG sites analysed [7, 8]. Moreover, it is estimated in non-CpG sites because the majority of non-CpG sites are unmethylated and the conversion should be closed to 100%.

To assess the bisulfite conversion rate, another option is the inclusion of completely unmethylated DNA controls. In the past, the most commonly unmethylated DNA control used was Lambda phage DNA, but now there are many companies that provide commercial unmethylated DNA controls especially for human and mouse studies. There are multiple methods to produce unmethylated DNA like the DNA purification from cells knockout for both DNA methyltransferases DNMT1 (-/-) and DNMT3b (-/-) or DNA from whole genome amplification. None of these methods produces exact values and all have a similar error rate, around 5%.

In summary, the conversion rate gives a general overview of the efficiency of the bisulfite reaction. A conversion rates greater than 97% is considered good, while less than 90% is considered very poor for methylation analysis. However, since it is estimated in non-CpG sites, it may fail to provide an accurate assessment of the bisulfite conversion in CpG sites.

1.2. PCR bias

Some technologies, like amplicon-based bisulfite sequencing or pyrosequencing, require the amplification of the region of interest to detect methylation levels. These technologies may suffer an additional problem called “PCR bias” which is a preferential amplification of an allele and strand in the PCR due to methylation state [9]. After bisulfite conversion, bisulfite DNA is a single strand fragmented DNA no longer complementary with a reduced number of cytosines which results in a low diversity sequence. Depending on the methylation status in each allele, the amplification of the region of interest could show variable efficiency which will affect the results and the accuracy.

In order to obtain accurate results, it is important to minimize the PCR-bias effect as much as possible. To this end, investigators [10, 11] have proposed to redesign primers by looking at strand-specific as well as bisulfite-specific flanking primers, but this solution is expensive and time-consuming and might not solve the problem completely. Instead, PCR bias can be calculated and corrected *in silico* [9, 12, 13] by using standard controls with known methylation

levels (**Fig. 3**). The apparent level of methylation after PCR in standard controls is used to correct the observed methylation levels in the case and control samples.

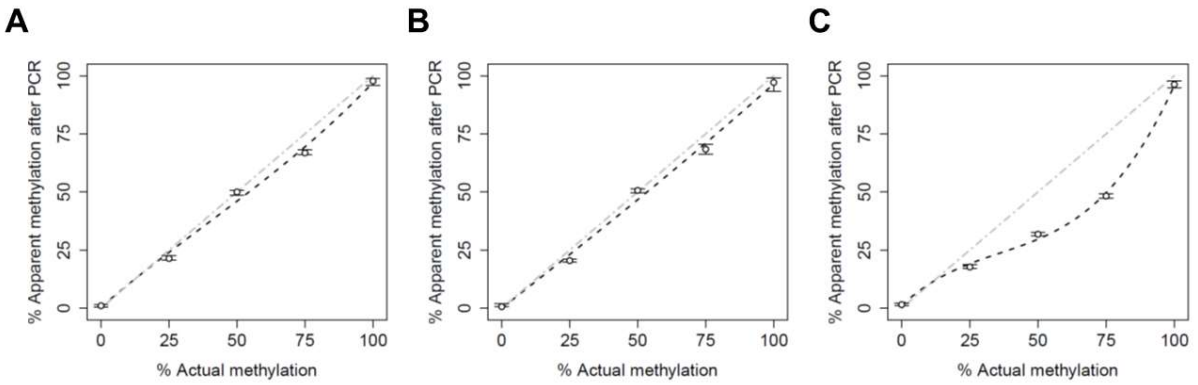

Fig. 3 The observed degree of bias introduced by PCR amplification in (**A**) *KCNQ1OT1*, (**B**) *H19*/*IGF2* and (**C**) *PLAGL1* DMRs by amplicon-based bisulfite sequencing presented in Ochoa et al. [13]. The apparent level of methylation observed after amplification (*y*-axis) is plotted as a function of the actual methylation percentage (*x*-axis). The black dashed line represents Moskalev's cubic polynomial interpolation curve [12] on the average (across CpGs) apparent level of methylation. The grey dot-dashed line represents an unbiased plot. The deviation of observed DNA methylation from the expected DNA methylation is similar in the first two target regions (low degree of preferential PCR amplification of unmethylated alleles), but it is different in *PLAGL1* DMR (high degree of methylated DNA loss).

The effect of the PCR bias in amplicon-based approaches will depend on the region and should be calculated during the development of new assays. The rest of technologies (pyrosequencing, methylation array, capture-based BS**,** WGBS and RRBS) only used PCR for the enrichment of the library and it is considered to have a small contribution in the deviation between the observed and the expected DNA methylation. Despite the cost, the analysis of at least three standard controls with known methylation level (0%, 50% and 100% actual methylation) is desirable to obtain accurate calibrated methylation measurements.

1.3. Region-specific bias

A study of the influence of DNA sequence and methylation status on bisulfite conversion revealed non-B DNA structures that could form monomolecular double-stranded regions resistant to bisulfite treatment [14]. In addition to this, the same authors [15] described that the bisulfite reactivity of neighbouring cytosines, within a DNA sequence that has the potential to form secondary structures, was influence by certain methyl groups. The majority of repetitive DNA motifs in the genome forms the canonical right-handed B-form, but some may also adopt

alternative conformations, called non-B DNA structures like hairpins/cruciforms, Z-DNA, triplexes (H-DNA), G-quadruplex and slipped DNA (**Fig. 4**). These non-B DNA structures are more resistant to denaturation, and therefore resistant to bisulfite conversion. In these cases, a high bisulfite conversion rate, which demonstrates a general success of the bisulfite conversion, will not warn about region-specific bisulfite conversions errors. In 2008 Clark et al. [16] described an inaccessible region for bisulfite treatment in a DNA methylation hotspot for breast cancer produced by a non-B DNA structure form by tandem repeat elements.

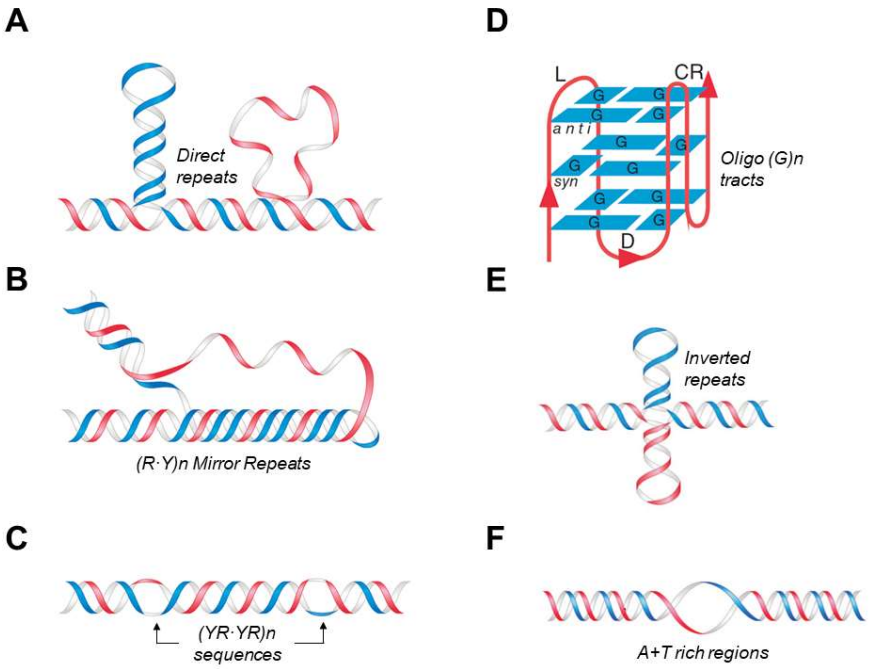

Fig. 4 Non-B DNA structures formed by genomic repetitive sequences (Figure adapted from Bacolla et al. [17]). (**A**) Slipped (hairpin) structure; (**B**) Triplex; (**C**) Left-handed Z-DNA; (**D**) Quadruplex; (**E**) Cruciform; (**F**) DNA unwinding element.

Repetitive motifs folded in non-B structures are responsible for repeat expansion or genomic rearrangements associated with neurodegenerative disorders and cancer [18]. These structures are chromosomal targets for DNA repair, recombination and aberrant DNA synthesis indicating the crucial role of these regions in human disease [17]. Repeated elements folded in non-B conformations are also frequent in imprinted regions and there are considered to take part in the regulation of these regions [19]. The majority of the DNA methylation studies focus their effort in CpG sites (28 million in the human genome) which are predominantly methylated [20] and heterogeneously distributed across the genome. However, only 21.8% of autosomal CpGs show dynamic regulation and most of which are located distal to the transcription start sites co-localizing with gene regulatory elements (enhancers and transcription factor-binding sites) and repetitive elements [20]. Therefore, there is a high

probability that some genomic regions which are clinically relevant can show resistance to denaturation during bisulfite treatment which, in turn, will affect methylation measurements. These observations suggest that regions with repetitive elements should be carefully studied to avoid misinterpretation of the observed methylation levels. The analysis of standard controls with different actual methylation levels could help understand the response of these regions to bisulfite conversion and improve the interpretation of the results.

1.4. *Coverage bias*

The accuracy of methylation detection also depends on the coverage or read-counts obtained per CpG. Recently, the comparison between whole-genome bisulfite sequencing (WGBS) and methylation arrays data revealed that minimum coverage of 100X for WGBS is recommended to achieve a level of precision broadly comparable to the methylation array [21] (**Fig. 5**).

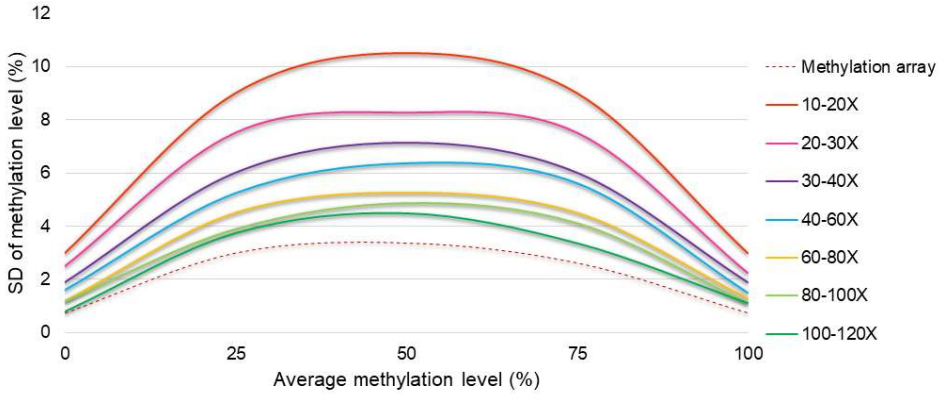

Fig. 5 Comparison of the standard deviation (SD) (*y*-axis) of methylation levels between methylation array and WGBS at different levels of the average read depths (*x*-axis) (Figure adapted from Zhou et al. [21]).

Ziller et al. [22] showed that at the previously recommended sequencing depth of 30x, the technical variability for WGBS is two to three-fold higher than for the methylation array. An insufficient read-counts will generate lower precision quantification. This problem will affect mainly WGBS because pyrosequencing, amplicon-based bisulfite sequencing and capture-based bisulfite sequencing and RRBS can easily achieve very high sequencing depth for a reasonable price (> 100X).

In summary, there are some limitations to obtain accurate methylation measurements. Some of these limitations affect equally all technologies like the incomplete BS conversion or region bias and others only affect some technologies like the PCR bias and the coverage bias (**Table**

1). Therefore, DNA methylation studies should consider these factors during design, analysis and interpretation of DNA methylation levels.

Table 1 Factors that could affect methylation measurements stratified by technology. BS: bisulfite sequencing; WGBS: whole-genome bisulfite sequencing; RRBS: reduced representation bisulfite sequencing.

	BS conversion	PCR bias	Region bias	Coverage bias
Pyrosequencing	✓	✓	✓	
Amplicon-based BS	✓	✓	✓	
Methylation array	✓		✓	
Capture-based BS	✓		✓	
WGBS	✓		✓	✓
RRBS	✓		✓	

2. Material

- MethylCal is available as an R package on https://github.com/lb664/MethylCal.
- On line R help files are associated with MethylCal's functions. They provide a detailed description of the syntax and the arguments of the functions as well as the class of the values returned by each function.
- A vignette included in the R package illustrates, by means of simple examples, data pre-processing, input data format and output generated by MethylCal.
- A markdown step-by-step user's documentation that explains how to perform the analysis and produce the plots presented in this book chapter is also available on https://github.com/lb664/MethylCal.

3. Methods

The first step to improve the accuracy of the methylation levels requires the control of the most critical point, the bisulfite conversion, and design the study using an appropriate coverage per region. The second step is the use of methylation controls to determine the risk of each region to suffer incomplete conversion. The third step is the estimation and correction of the deviation of the observed DNA methylation from the expected DNA methylation using appropriate statistical algorithms. The efficiency of the conversion reaction and the structure of the genomic region will influence the "observed:expected" (O:E) ratio. If this quantity is small, the correction produced by the calibration will be negligible, but if the ratio is large the calibration will produce results that could depart significantly from the observed methylation levels.

3.1. Statistical methods to correct methylation levels

PCR bias can be calculated and corrected *in silico* by using standard controls with known methylation levels. Different statistical models have been proposed to correct the observed methylation levels, specifically the best-fit hyperbolic [9], cubic polynomial curve [12] and a recently proposed Bayesian calibration model called MethylCal [13].

MethylCal can analyse jointly all CpGs within a CpG island (CGI) or a differentially methylated region (DMR), avoiding "one-at-a-time" CpG calibration or the calibration of the average methylation level across CpGs that ignores the variability and dependence across CpGs [23]. The second key feature of MethylCal is the inclusion of random effects [24] that capture the unobserved heterogeneity (variability) of the apparent methylation levels at each actual methylation percentage or CpG or a combination of both. This enables more precise modelling of the methylation measurements observed in the standard controls compared to alternative methods [9, 12] since they assume a unique calibration curve across all CpGs or independent calibration curves for each CpG.

MethylCal's modularity allows also the analysis of complex experimental designs. While the derivation of the calibration curves obtained from a single standard control sample is the typical experimental design in clinical diagnostic and research laboratories, technical replicates might also be available. Rather than taking the average across them, MethylCal can be extended, with negligible computational costs, to include a further random effect that models the variability of the apparent level of methylation across repeated control experiments (technical replicates), allowing precise quantification of biological and technical variability.

3.2. Model outline

MethylCal's regression model and its extension for repeated control experiments can be described as follows:

$$y_{ijk} = \beta_{0k} + \beta_{1k}x_{ijk} + \beta_{2k}x_{ijk}^2 + \beta_{3k}x_{ijk}^3 + \mathrm{RE}_{ij} + \mathrm{RE}_k + \epsilon_{ijk}, \quad (1)$$

where y_{ijk} is the apparent level of methylation after PCR at the ith actual methylation percentages (AMP) ($i = 1, \cdots, I$), jth CpG ($j = 1, \cdots, J$) and kth technical replicate ($k = 1, \cdots, K$), x_{ijk} is the ith AMP ($x_{1jk} = 0\%$ and $x_{Ijk} = 100\%$) which is assumed constant across CpGs and technical replicates and $\beta_{0k}, \cdots, \beta_{3k}$ are the intercept and coefficients of the cubic polynomial regression. In a typical experimental design the percentage control points x_{ijk} are equally spaced and placed at least at 0%, 50% and 100% actual methylation. Finally, RE_{ij} and RE_k are random effects and $\epsilon_{ijk} \sim \mathrm{N}(0, \sigma^2)$.

The specification of the random-effects RE_{ij} is key in MethylCal's formulation (1) (full details regarding the technical replicates random-effects RE_k are presented in Section 3.3). Specifically,

- **Simple random-effects model**, $\mathrm{RE}_{ij} = \mathrm{AMP}_i$: the random-effects AMP_i are introduced to model the variability of the apparent level of methylation after PCR at different AMPs not explained by the cubic polynomial regression;
- **Crossed random-effects model**, $\mathrm{RE}_{ij} = \mathrm{AMP}_i + \mathrm{CpG}_j$: besides the random-effects AMP_i, the crossed random-effects CpG_j are added to capture the heterogeneity of the apparent level of methylation across CpGs;
- **Paired random-effects and latent Gaussian field model**, $\mathrm{RE}_{ij} = \mathrm{AMP}_i + \mu_j$: the crossed random-effects CpG_j are replaced by the latent Gaussian field $\boldsymbol{\mu} = \left(\mu_1, \cdots, \mu_J\right)^T$ to smooth the spatial dependence of the apparent levels of methylation across CpGs;
- **Crossed random-effects with random-intercepts and random-slopes model**, $\mathrm{RE}_{ij} = \mathrm{AMP}_i + \mathrm{CpG}_j + \mathrm{CpG}_j^* \times x_{ij}$: the random-slopes CpG_j^* are added to model the variability of the apparent level of methylation after PCR across CpGs at lower/higher AMPs (we dropped the indexes k in x_{ijk} since we assume that they are constant across technical replicates).

The priors distributions on all unknown quantities (intercept and coefficients of the cubic polynomial regression, random effects and the latent Gaussian field) are discussed in detail in Ochoa et al. [13]. MethylCal uses INLA [25] for fast inference and model selection of the random-effects RE_{ij} specification by selecting the RE_{ij} model that fits the data best based on the DIC [26]. **Fig. 6** and **Fig. 7** show the departure of MethylCal from the simple hypothesis of cubic polynomial regression (black dots) with different specifications of the random-effects RE_{ij}.

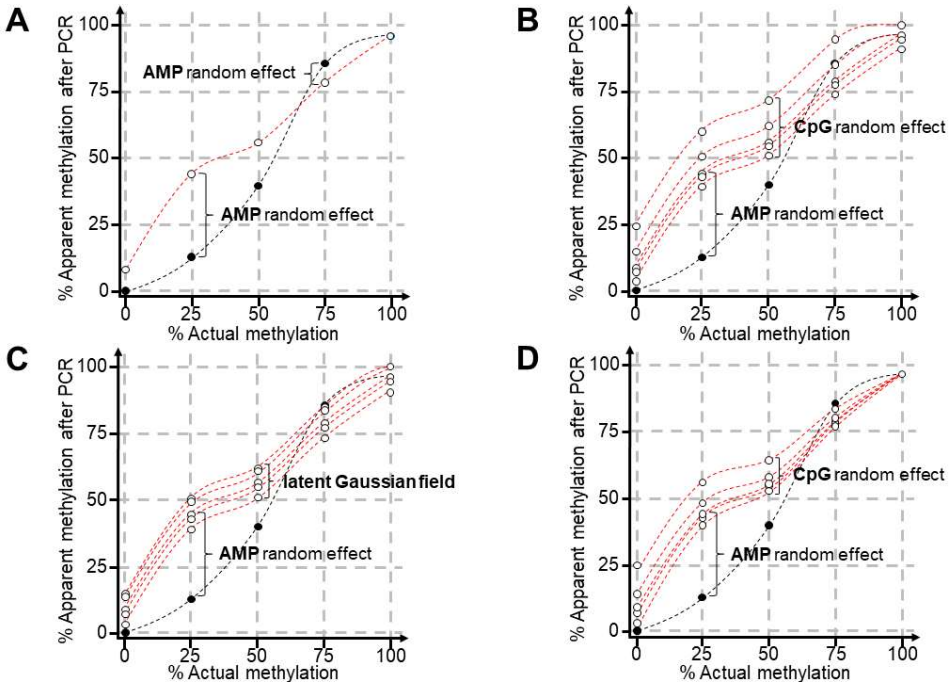

Fig. 6 Schematic representation of MethylCal's models. The apparent level of methylation observed after amplification (*y*-axis) is plotted as a function of the actual methylation percentages (*x*-axis). The black dots and the black dotted line represent the apparent level of methylation predicted by the best-fit cubic polynomial regression across all CpGs, whereas dots and red dotted lines depict the level of methylation predicted by MethylCal's models. Panels (**A**), (**B**), (**C**) and (**D**) correspond to different specifications of the random-effects RE_{ij}, simple random effects, crossed random effects, paired random effects and latent Gaussian field and crossed random effects with random intercepts and random slopes, respectively. In (**A**), the calibration curves for different CpGs overlap.

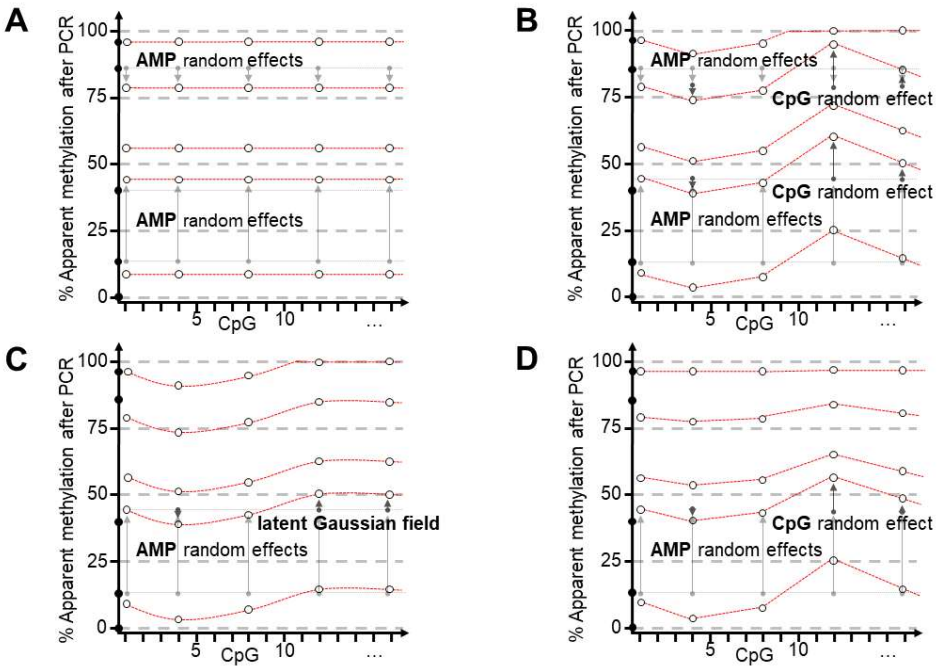

Fig. 7 The apparent level of methylation observed after amplification (*y*-axis) is plotted as a function of the CpGs (*x*-axis) in the target region. The black dots at $x = 0$ represent the level of methylation predicted by the best-fit cubic polynomial regression across all CpGs, whereas dots and red dotted lines depict the level of methylation predicted by MethylCal's models. Panels (**A**), (**B**), (**C**) and (**D**) correspond to different specifications of the random-effects RE_{ij} as in **Fig. 6**.

3.3. Calibration experiment with technical replicates

While the cost of running a control experiment may prevent its wide applicability, the effects of the of PC bias, as well as BS conversion and region bias (**Table 1**), can be greatly reduced. We advise to include at least three percentage control points at 25%, 50% and 75% actual methylation, besides at 0% and 100%, prepared by mixing different ratios of nonmethylated (0%) and methylated (100%) control DNA bisulphite converted.

To raise the sensitivity of the correction curve, rather than increasing the number of AMPs, a different strategy would be to generate technical replicates, i.e., repeated measurements of the same control experiment that represent independent measures of the random noise associated with the preparation of different control DNA ratios. The main problem connected to the generation of these ratios is that currently only 0% and 100% methylated DNA are commercially available in humans. The rest, 25%, 50% and 75%, are manually generated by mixing an appropriate volume of 0% and 100%. During this process, any pipetting error will create an incorrect expected ratio.

In this scenario, MethylCal's regression model (1) is extended to include the random effects RE_k $(k = 1, \cdots, K)$ that model the heterogeneity (variability) of the levels of apparent

methylation across replicates. Instead of fitting and estimate a separate MethylCal's model for each control experiment, all technical replicates are jointly analysed using the whole vector of observations $\mathbf{y}$ of dimension $n = I \times J \times K$.

3.4. Corrected methylation degree

The specification of a further random effect for the technical replicates leads to a modification of the original MethylCal's algorithm for the derivation of the corrected methylation degree. Following Ochoa et al. [13], for each CpG ($j = 1, \cdots, J$) and for each technical replicate ($k = 1, \cdots, K$), the PCR-bias corrected methylation degree $\hat{x}_{jk} \in [0\%, 100\%]$ is the solution of

$$\hat{x}_{jk} = \arg\min_{x_{jk}} \left[y_j^{\text{obs}} - \mathrm{E}(\boldsymbol{\beta}_k|\mathbf{y}) \boldsymbol{x}_{jk} + \eta\left(x_{jk}\right) + C_{jk} \right]^2, \tag{2}$$

where $y_j^{\text{obs}} \in [0\%, 100\%]$ is the observed level of methylation at the jth CpG, $\mathrm{E}(\boldsymbol{\beta}_k|\mathbf{y})$ is the posterior mean of the intercept and regression coefficients of the cubic polynomial regression in the kth control experiment, $\boldsymbol{x}_{jk} = \left(1, x_{jk}, x_{jk}^2, x_{jk}^3\right)$ and $\eta\left(x_{jk}\right)$ is the predicted value evaluated in x_{jk} of the smoothed posterior mean of the random-effects AMP_i obtained from a cubic spline that interpolates $\mathrm{E}(\mathrm{AMP}_i|\mathbf{y})$ ($i = 1, \cdots, I$). Finally, $C_{jk} = \mathrm{E}(\mathrm{RE}_k|\mathbf{y})$, $C_{jk} = \mathrm{E}\left(\mathrm{CpG}_j\middle|\mathbf{y}\right) + \mathrm{E}(\mathrm{RE}_k|\mathbf{y})$, $C_{jk} = \mathrm{E}\left(\mu_j\middle|\mathbf{y}\right) + \mathrm{E}(\mathrm{RE}_k|\mathbf{y})$ and $C_{jk} = \mathrm{E}\left(\mathrm{CpG}_j\middle|\mathbf{y}\right) + \mathrm{E}\left(\mathrm{CpG}_j^*\middle|\mathbf{y}\right) \times x_{jk} + \mathrm{E}(\mathrm{RE}_k|\mathbf{y})$ according to the specification of the random-effects RE_{ij} where $\mathrm{E}\left(\mathrm{CpG}_j\middle|\mathbf{y}\right)$, $\mathrm{E}\left(\mathrm{CpG}_j^*\middle|\mathbf{y}\right)$, $\mathrm{E}\left(\mu_j\middle|\mathbf{y}\right)$ and $\mathrm{E}(\mathrm{RE}_k|\mathbf{y})$ are the posterior mean of the random-intercepts CpG_j, random-slopes CpG_j^*, latent Gaussian field μ_j and the technical replicates random-effects RE_k, respectively. Note that C_{jk} depends on the unknown PCR-bias corrected methylation degree only when the random-slopes CpG_j^* is specified in RE_{ij}.

The solution of the minimisation problem (2) produces k corrected methylation degrees $\hat{x}_{jk}$ for each observed level of methylation y_j^{obs} at each CpG. This is expected since there is a different correction depending on which technical replicate is considered, although the posterior estimates are obtained by using all n observations at once. To get a unique calibration curve, we interpolate $\hat{x}_{jk}$ by a cubic spline across the technical replicates and along CpGs and finally obtain a single corrected value $\hat{x}_j \in [0\%, 100\%]$ $(j = 1, \cdots, J)$.

3.5. Real case applications

The interest in DNA methylation and its correlation with diagnostic, prognostic and treatment is constantly growing. There are well-known DNA methylation biomarkers that support the diagnostic of imprinting disorder [27], the clinical decisions in various cancers [28–32] and are

also used for non-invasive prenatal testing [33]. In the context of clinical practice, accurate DNA methylation levels are crucial because they can influence treatment decisions or further actions.

3.5.1. Application in clinical diagnostic of celiac patients

Celiac disease (CD) is a common, immune-mediated enteropathy induced by the exposure to gluten. Multiple studies in the last decade exposed the important relationship between autoimmunity and DNA methylation [34, 35]. Recently, a methylation array study in celiac disease was published confirming the implication of DNA methylation in the pathogenesis of this disease [36]. The validation and identification of new diagnostic, prognostic and/or therapeutic biomarkers related to DNA methylation in celiac disease will require approaches to interrogate specific target like pyrosequencing or amplicon-based bisulfite sequencing. As we explained earlier in this chapter, these methodologies may suffer PCR bias which should be considered and corrected in order to obtain accurate DNA methylation measurements.

In the following example, we show pyrosequencing results from three assays: *NFKBIA, RELA* and *TNFAIP3* genes previously associated with susceptibility to celiac disease. **Fig. 8** presents the calibration curves for the target regions considered. It is apparent that there is a large variability of the apparent level of methylation in the control experiment, especially at medium and high percentages of actual methylation. **Fig. 8B**, **D** and **F** show that this is mainly due to abnormal low values of the apparent level of methylation measured in CpGs located at the end of each target region, in particular in *NFKBIA* and *RELA* assays. MethylCal is able to identify several of these observations and automatically mark them as possible outliers (black dots). As a result, they can be removed from the analysis (including the corresponding CpGs from the observed levels of methylation $\boldsymbol{y}^{\mathrm{obs}}$), or the data generation process should be further investigated to understand the possible causes of this unusual pattern, including biological and biochemical factors. For instance, several factors reported in **Table 1** that could affect methylation measurements, may play an important role also in the derivation of the calibration curves, exerting their influence on the apparent level of methylation measured in the control experiment. A combination of these factors can explain the observed anomalous pattern of the methylation levels at CpGs located at the end of the target regions.

In all assays analysed, the crossed random-effects with random-intercepts and random-slopes model $\mathrm{RE}_{ij} = \mathrm{AMP}_i + \mathrm{CpG}_j + \mathrm{CpG}_j^* \times x_{ij}$ is chosen as the best-fit model. By looking at **Fig. 8A**, **C** and **E**, this is expected since there is an increasing level of variability of the apparent level of methylation after PCR across CpGs at higher AMPs.

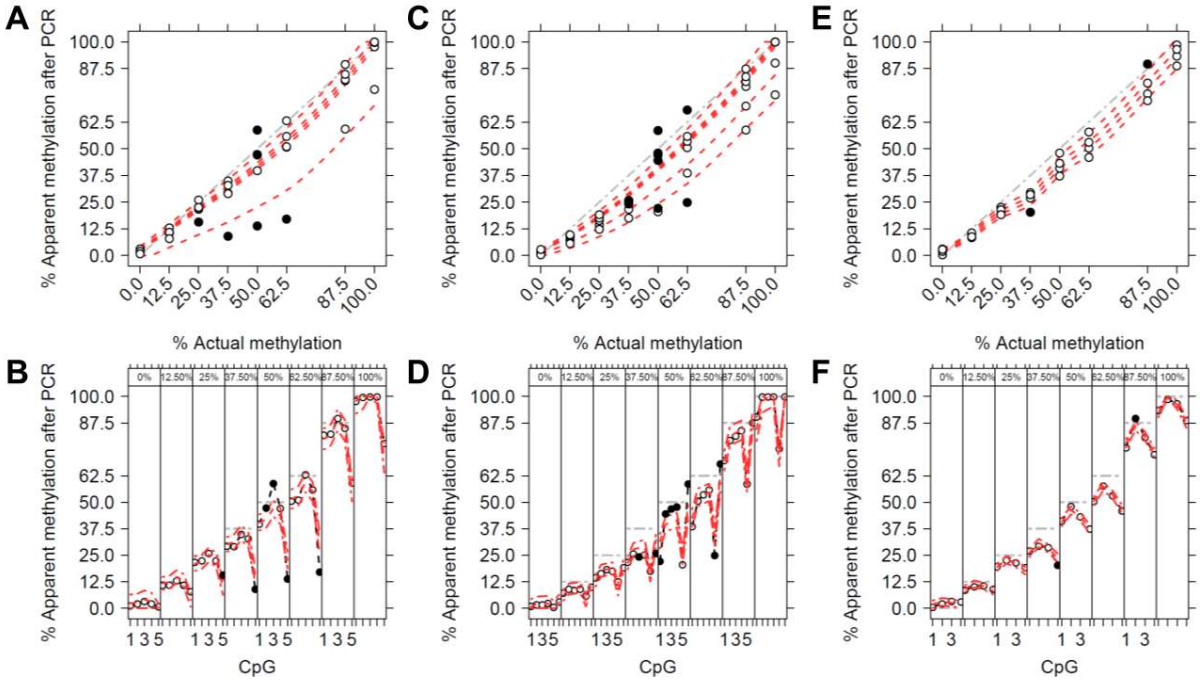

Fig. 8 Calibrated methylation levels of (**A**, **B**) *NFKBIA*, (**C**, **D**) *RELA* and (**E**, **F**) *TNFAIP3* assays in celiac patients using MethylCal. (**A**, **C**, **E**) The apparent level of methylation observed after PCR (*y*-axis) is plotted as a function of the actual methylation percentage (AMP) (*x*-axis). Circles depict the apparent level of methylation for each CpG at different AMPs, whereas red dotted lines show the predicted level of methylation. Black dots highlight potential outliers. (**B**, **D**, **F**) The apparent level of methylation (*y*-axis) is plotted (circles) against the CpGs in the target region (*x*-axis), stratified by AMP (top panel box and grey dot-dashed line). For each stratum, red dotted and dot-dashed lines show the level of methylation and the 95% prediction credible interval. All panels have been generated using MethylCal's R package. For example, (**A**) and (**B**) are obtained by typing in R command line: `MethylCalCalibrationOutlierPlot(Celiac_data, Target="NFKBIA")`.

Once the calibration curves have been estimated, we applied MethylCal PCR-bias correction to obtain the methylation degree $\hat{x}_j \in [0\%, 100\%]$ $(j = 1, \cdots, J)$ as the solution to the minimization problem (2) with $k = 1$. **Fig. 9** shows the corrected methylation degree of *NFKBIA*, *RELA* and *TNFAIP3* assays in healthy controls (top panels) and celiac patients (bottom panels). The statistical test used to declare individuals having undergone loss/gain of methylation is based on the mean of the corrected methylation degree across CpGs for each individual in the case and control groups. Patients with an average corrected methylation level below healthy individuals' three standard deviations (SD) confidence interval were considered to undergo loss of methylation and those with a level above the 3SD confidence interval were considered to experience gain of methylation. To avoid false positives, we chose healthy

individuals' ±3SD confidence interval since it guarantees very low type-I error ($\alpha = 2.7 \times 10^{-3}$).

The results of differential calibrated analysis displayed in the bottom panels of **Fig. 9** show that while *RELA* assay is not clinically relevant to distinguish patients from healthy controls, in both *NFKBIA* and *TNFAIP3* assays, we were able to classify patients 09D, 12D and 14D and 09D, 10D and 11D as having undergone gain of methylation. For *NFKBIA* and *TNFAIP3* assays, it is also evident that almost all celiac patients considered were borderline cases that are difficult to diagnose. More sophisticated *in silico* calibration statistical tools and consequently more reliable corrections of the methylation levels in target regions are therefore essential for improving diagnosis based on methylation data.

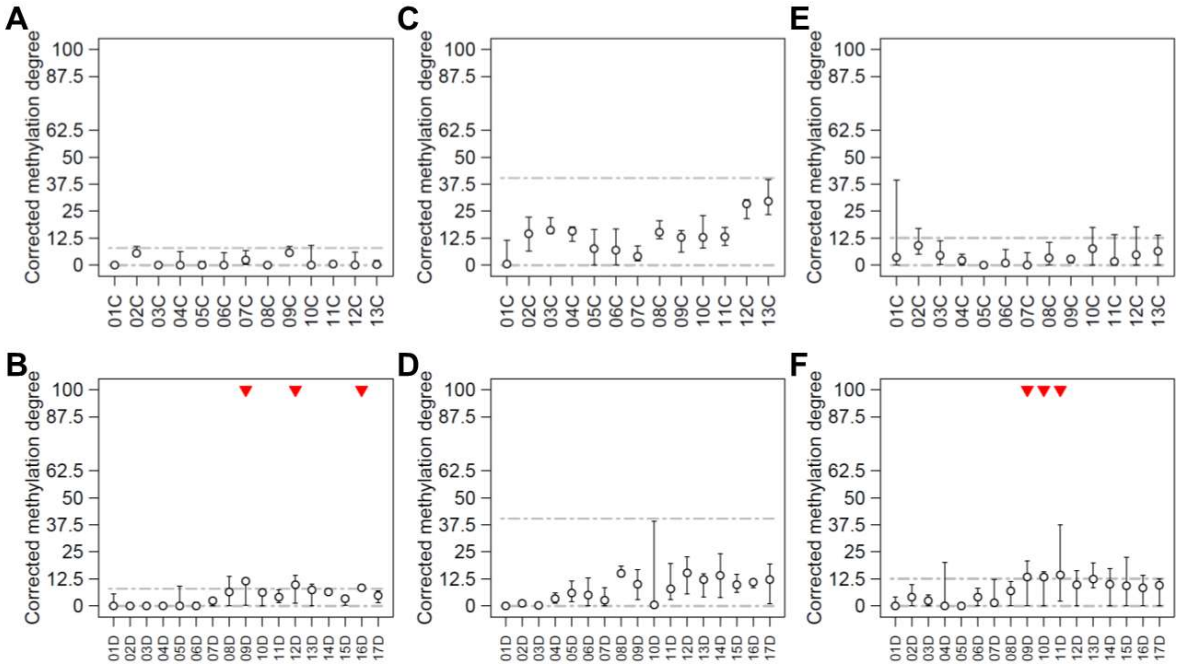

Fig. 9 Corrected methylation degree of (**A**, **B**) *NFKBIA*, (**C**, **D**) *RELA* and (**E**, **F**) *TNFAIP3* assays in healthy controls (top panels) and celiac patients (bottom panels) using MethylCal. For each individual (*x*-axis), the boxplot depicts the range and the mean (circle) of the corrected methylation degree (*y*-axis) across CpGs, while the dashed-dotted grey lines show the healthy controls' ±3SD confidence interval. Top red triangles in bottom panels indicate patients classified as having undergone gain of methylation. All panels have been generated using MethylCal's R package. For example, (**A**) and (**B**) are obtained by typing in R command line: `MethylCalCorrection(Celiac_data, Target="NFKBIA", n_Control=13, n_Case=17)`.

3.5.2. *Application in clinical diagnostic: Mosaic Beckwith–Wiedemann syndrome*

Beckwith–Wiedemann syndrome (BWS) is an imprinting disorder caused by genetic and epigenetic abnormalities that dysregulate imprinted genes on 11p15.5 [37]. This rare disorder is characterised by overgrowth, macroglossia, hemihyperplasia, exomphalos and embryonal tumour produced by an increased expression of the growth factor, *IGF2*, or reduced expression of the growth suppressor, *CDKN1C*. The most frequent epigenetic alterations on BWS are: Loss of methylation of maternal *KCNQ1OT1*, gain of methylation of maternal *H19*/*IGF2* and paternal uniparental disomy of chromosome 11. Furthermore, BWS presentation is frequently as mosaic due to alterations during embryonic development. In mosaic cases, methylation measurements will depend on the percentage of cells with the alteration in the tissue tested. Depending on the level of mosaicism, methylation levels may be borderline for diagnostic criteria and the accuracy may be critical in these cases.

Fig. 10 shows the observed degree of bias introduced by PCR amplification in three target regions chosen for their BWS clinical diagnostic power (*KCNQ1OT1* and *H19*/*IGF2* DMRs) and the high degree of methylated DNA loss (*PLAGL1* DMR) at intermediate percentages of actual methylation (25%, 50% and 75%). In these target regions, rather than increasing the number of AMPs, we generated data from two independent technical replicates to assess the variability associated with the preparation of different control DNA ratios. By looking at **Fig. 10**, significant differences were observed at 25%, 50% and 75% actual methylation in *KCNQ1OT1* DMR, while values slightly diverge between replicates at 75% and 100% and 25% in *H19*/*IGF2* and *PLAGL1* DMRs, respectively.

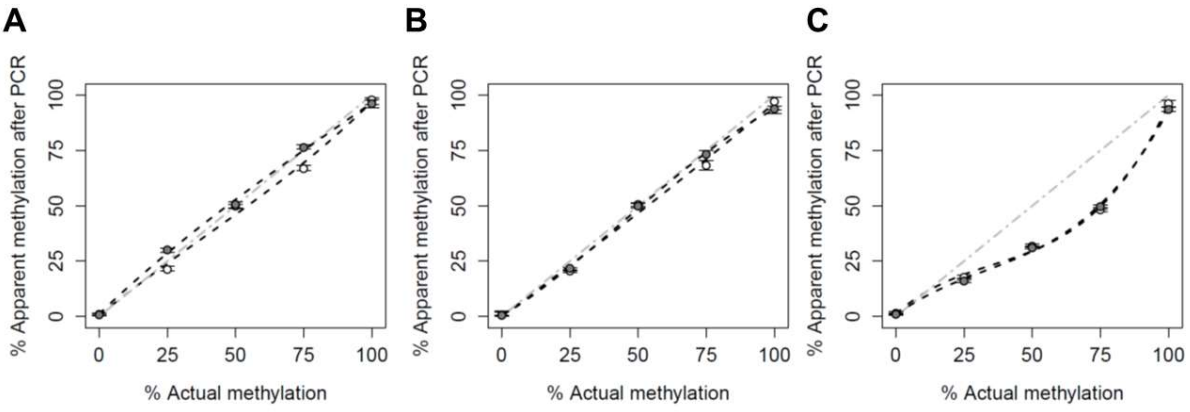

Fig. 10 The observed degree of bias introduced by PCR amplification in (**A**) *KCNQ1OT1*, (**B**) *H19*/*IGF2* (**C**) *PLAGL1* DMRs by amplicon-based bisulfite sequencing. The apparent level of methylation observed after amplification (*y*-axis) is plotted as a function of the actual methylation percentage (*x*-axis). Colour-coded circles (white and grey) indicate two independent calibration replicates. The black dashed line represents Moskalev's cubic polynomial interpolation curve of the average (across CpGs) apparent level of methylation for each technical replicate. The grey dot-dashed line represents an unbiased plot. The deviation

of observed DNA methylation from the expected DNA methylation is significantly different between the technical replicates at 25% and 75% actual methylation in *KCNQ1OT1* DMR while differences between replicates are apparent at 75% and 100% and 25% and 100% in *H19*/*IGF2* and *PLAGL1* DMRs, respectively. All panels have been generated using MethylCal's R package. For example, (**A**) is obtained by typing in R command line: `ExploratoryPlot(BWS_data, Target="KCNQ1OT1").`

The straightforward modification of MethylCal's algorithm presented in Section 2.1, when independent technical replicates are considered in the control experiment, allows us to estimate the calibrated methylation levels in this new experimental design. **Fig. 11** displays the output of MethylCal's algorithm in the three target regions considered. While there is a negligible difference between the replicate-specific calibration curves (dark red dotted lines) and the cubic spline interpolant (red dashed line) in *H19*/*IGF2* and *PLAGL1* DMRs (**Fig. 11C** and **E**), in *KCNQ1OT1* DMR the cubic spline interpolant lies between the calibration curves obtained in each technical replicate (**Fig. 11A**). A closer inspection of the results shows that in order to generate well-separated replicate-specific calibration curves, there should be discordant values of the apparent level of methylation between replicates in at least two AMPs (**Fig. 11B** at 25% and 75%) and the overall magnitude of the observed degree of bias at different AMPs should be the same across technical replicates in contrast to *H19*/*IGF2* DMR where at 75% and 100% the levels of the apparent methylation after PCR are flipped between technical replicates (**Fig. 11D**). Finally, in *PLAGL1* DMR abnormal low values of the apparent level of methylation are measured in CpGs located nearly at the end of each target region (**Fig. 11F**) and this phenomenon is reproduced in both technical replicates. The reason that causes this behaviour is not completely clear and may depend on a combination of factors, but the additional information obtained by standard controls at different AMPs shows that it is CpG specific suggesting a possible region bias. In addition, the behaviour observed in **Fig. 11F** indicates that the observed bias grows with the level of apparent methylation. The most plausible explanation of this pattern could be a secondary structure of DNA form in the presence of methyl groups that avoid the access to MssI methyltransferase enzyme used to produce commercial fully-methylated DNA.

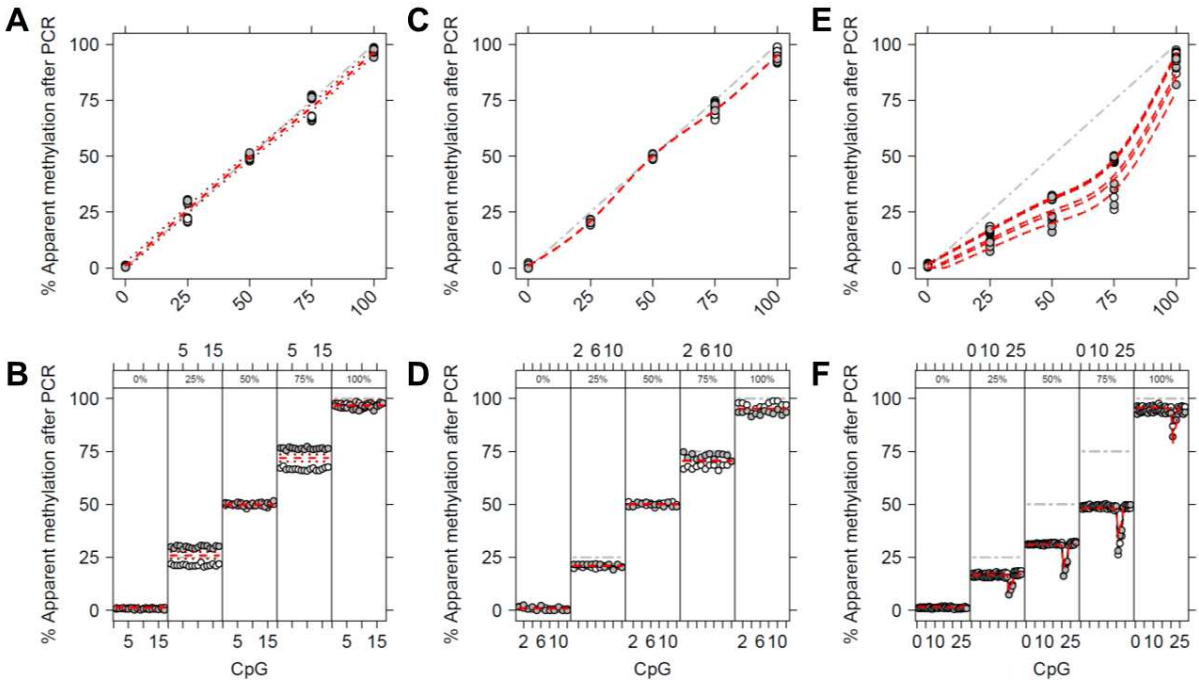

Fig. 11 Calibrated methylation levels of (**A**, **B**) *KCNQ1OT1*, (**C**, **D**) *H19*/*IGF2* and (**E**, **F**) *PLAGL1* DMRs in mosaic Beckwith–Wiedemann syndrome using MethylCal. (**A**, **C**, **E**) The apparent level of methylation observed after PCR (*y*-axis) is plotted as a function of the actual methylation percentage (AMP) (*x*-axis). Colour-coded circles (white and grey) depict the apparent level of methylation for each CpG at different AMPs in two independent calibration replicates. In each panel, dark red dotted lines show the predicted level of methylation for each technical replicate while the red dashed line displays the cubic spline interpolant of the predicted level of methylation curves. (**B**, **D**, **F**) The apparent level of methylation (*y*-axis) is plotted (circles) against the CpGs in the DMR (*x*-axis), stratified by AMP (top panel box and grey dot-dashed line) with different colours (white and grey) for the independent calibration replicates. For each stratum, dark red dotted lines show the predicted level of methylation for each technical replicate while the red dashed line displays the cubic spline interpolant of the predicted level of methylation curves. All panels have been generated using MethylCal's R package. For example, (**A**) and (**B**) are obtained by typing in R command line: `MethylCalCalibrationPlot(BWS_data, Target="KCNQ1OT1")`.

For *KCNQ1OT1* and *H19*/*IGF2* DMRs, the best MethylCal fit-model is the simple random-effects model $\mathrm{RE}_{ij} = \mathrm{AMP}_i$; for *PLAGL1* assay the crossed random-effects with random-intercepts and random-slopes model $\mathrm{RE}_{ij} = \mathrm{AMP}_i + \mathrm{CpG}_j + \mathrm{CpG}_j^* \times x_{ij}$ is selected to fit the increasing level variability of the apparent level of methylation after PCR across CpGs at higher AMPs. **Fig. 12** presents the results of the differential calibrated analysis. Interestingly, the higher sensitivity of the calibration curve obtained by adding a technical replicate allows a different classification of patient B5B37 (**Fig. 12B**) as undergoing loss of methylation in

KCNQ1OT1 DMR compared to the single calibration curve presented in Ochoa et al. [13] where the same patient was considered as undergoing gain of methylation in *H19*/*IGF2* DMR. This final example highlights the importance of independent calibration replicates for the quantification of the observed degree of bias introduced after PCR since they can reduce the variability associated with the preparation of different control DNA ratios that, in turn, can have a big impact in the derivation of the corrected methylation degree and classification of borderline cases.

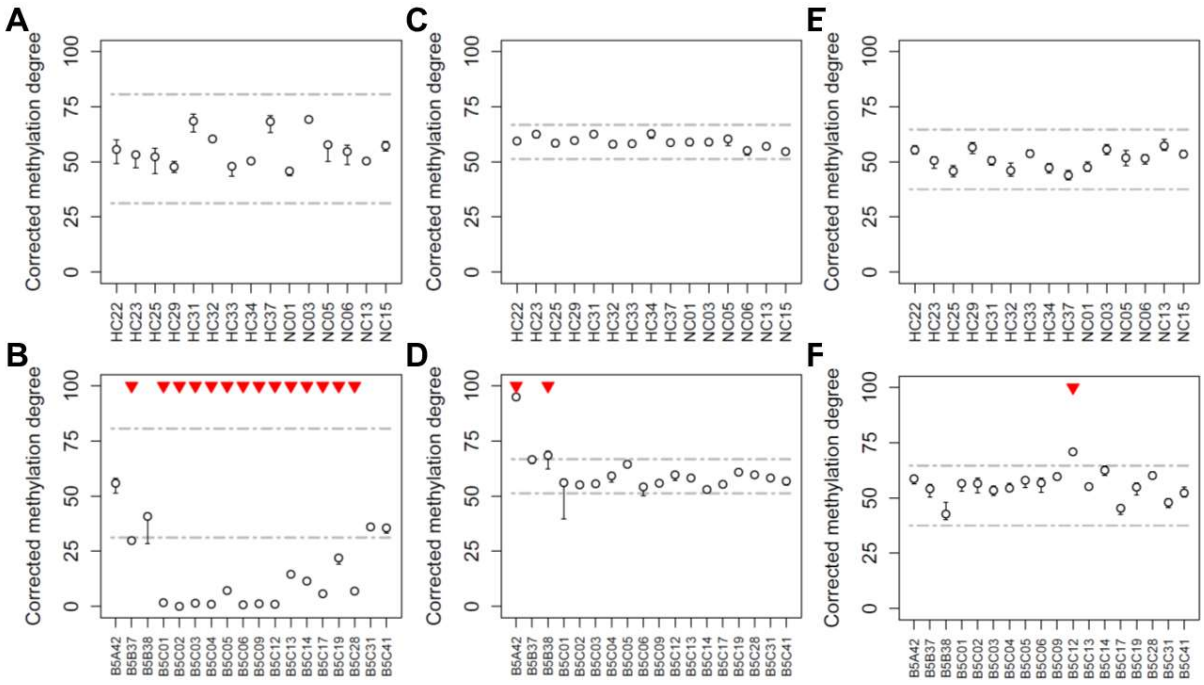

Fig. 12 Corrected methylation degree of (**A**, **B**) *KCNQ1OT1*, (**C**, **D**) *H19*/*IGF2* and (**E**, **F**) *PLAGL1* DMRs in healthy controls (top panels) and alleged Beckwith–Wiedemann syndrome patients (bottom panels) using MethylCal. For each individual (*x*-axis), the boxplot depicts the range and the mean (circle) of the corrected methylation degree (*y*-axis) across CpGs, while the dashed-dotted grey lines show the healthy controls' confidence interval. Top red triangles in bottom panels indicate patients classified as having undergone gain/loss of methylation. All panels have been generated using MethylCal's R package. For example, (**A**) and (**B**) are obtained by typing in R command line: `MethylCalCorrection(BWS_data, Target="KCNQ1OT1", n_Control=15, n_Case=18)`.

4. Notes

1. MethylCal's R functions detect automatically from the input data if independent control replicates have been included in the analysis. However, they assume that the same levels of actual methylation are used across all technical replicates. Different sets of

CpGs are allowed across technical replicates although MethylCal's R functions consider only CpGs that are in common.

2. The automatic detection of outliers may inform regarding possible factors that cause the observed bias. While MethylCal is robust to outliers (see calibration curves in **Fig. 8A** and **C**), they are not removed automatically from the analysis. Corrected methylation degree in **Fig. 9B** and **D** are obtained using all standard controls. If CpGs are manually removed from standard controls, in the same way, they should be excluded from the observed methylation levels.
3. When technical replicates are included in the derivations of the calibration curves, the automatic detection of possible outliers has not been implemented in MethylCal's R package. Instead, we suggest checking for outliers independently in each control replicate and mark observations as outliers if they are detected in more than half of the technical replicates. Similarly to 2. above, they can be removed manually from the technical replicates and observed methylation levels before running MethylCal.
4. In MethylCal's formulation (1), we made the assumption that the error term is normally distributed. While this assumption seems to be reasonable in the pyrosequencing real data example (continuous observations although both y_{ijk}, and $y_j^{\mathrm{obs}} \in [0\%, 100\%]$), in the second example with amplicon-based bisulfite sequencing we obtained beta intensities by taking the ratio between the read counts and the read depth for each CpG in the target regions. The same strategy could be employed for WGBS and RRBS data, although a better approach would be to use a negative-binomial likelihood. However, in count data regression models, the inclusion of crossed or nested random-effects or a latent Gaussian field may produce biased estimates and overly narrow confidence intervals [38] when posterior inference is performed by using numerical integration or Laplace approximation [25].

5. Conclusions

DNA methylation is an epigenetic mechanism responsible for multiple biological processes including genomic imprinting, X-inactivation, tissue-specific gene-expression and transgenerational effects. Likewise, cumulative evidence indicates also a critical role of DNA methylation in human diseases [39] like rheumatoid arthritis, obesity or autism [40–42]. The interest in the study of DNA methylation changes and its relationship with the disease is constantly growing. However, the technologies and methodologies available at the moment are still evolving and right now the analysis of DNA methylation requires a deep understanding of the limitations and possible solutions.

Acknowledgements

Alan Turing Institute under the Engineering and Physical Sciences Research Council [EP/N510129/1 to L.B.]. The Authors thank Jose Ramon Bilbao for providing the celiac data as well as Nora Fernandez-Jimenez and Eamonn Maher for useful discussion regarding the results of the celiac data and mosaic Beckwith–Wiedemann syndrome.

Conflict of interest statement. The views expressed are those of the authors and not necessarily those of the NHS or Department of Health.

References

1. Laird PW (2003) The power and the promise of DNA methylation markers. Nat Rev Cancer 3:253–266 . https://doi.org/10.1038/nrc1045
2. Bock C (2009) Epigenetic biomarker development. Epigenomics 1:99–110 . https://doi.org/10.2217/epi.09.6
3. Shapiro R, Weisgras JM (1970) Bisulfite-catalyzed transamination of cytosine and cytidine. Biochem Biophys Res Commun 40:839–843 . https://doi.org/10.1016/0006-291X(70)90979-4
4. Hayatsu H, Wataya Y, Kai K (1970) The Addition of Sodium Bisulfite to Uracil and to Cytosine. J Am Chem Soc 92:724–726 . https://doi.org/10.1021/ja00706a062
5. Wang RYH, Gehrke CW, Ehrlich M (1980) Comparison of bisulfite modification of 5-methyldeoxycytidine and deoxycytidine residues. Nucleic Acids Res 8:4777–4790 . https://doi.org/10.1093/nar/8.20.4777
6. Frommer M, McDonald LE, Millar DS, Collis CM, Watt F, Grigg GW, Molloy PL, Paul CL (1992) A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. Proc Natl Acad Sci U S A 89:1827–1831 . https://doi.org/10.1073/pnas.89.5.1827
7. Yu H, Hahn Y, Yang I (2015) Reference materials for calibration of analytical biases in quantification of DNA methylation. PLoS One 10:8–19 . https://doi.org/10.1371/journal.pone.0137006
8. Masser DR, Berg AS, Freeman WM (2013) Focused, high accuracy 5-methylcytosine quantitation with base resolution by benchtop next-generation sequencing. Epigenetics and Chromatin 6:1 . https://doi.org/10.1186/1756-8935-6-33
9. Warnecke PM, Stirzaker C, Melki JR, Millar DS, Paul CL, Clark SJ (1997) Detection and measurement of PCR bias in quantitative methylation analysis of bisulphite-treated DNA. Nucleic Acids Res 25:4422–4426 . https://doi.org/10.1093/nar/25.21.4422
10. Wojdacz TK, Borgbo T, Hansen LL (2009) Primer design versus PCR bias in methylation independent PCR amplifications. Epigenetics 4:231–234 . https://doi.org/10.4161/epi.9020
11. Wojdacz TK, Hansen LL, Dobrovic A (2008) A new approach to primer design for the control of PCR bias in methylation studies. BMC Res Notes 1:3–5 . https://doi.org/10.1186/1756-0500-1-54
12. Moskalev EA, Zavgorodnij MG, Majorova SP, Vorobjev IA, Jandaghi P, Bure I V., Hoheisel JD (2011) Correction of PCR-bias in quantitative DNA methylation studies by means of cubic polynomial regression. Nucleic Acids Res 39: . https://doi.org/10.1093/nar/gkr213

13. Ochoa E, Zuber V, Fernandez-Jimenez N, Bilbao JR, Clark GR, Maher ER, Bottolo L (2019) MethylCal: Bayesian calibration of methylation levels. Nucleic Acids Res 47:e81 . https://doi.org/10.1093/nar/gkz325
14. Rother KI, Silke J, Georgiev O, Schaffner W, Matsuo K (1995) Influence of DNA Sequence and methylation status on bisulfite conversion of cytosine residues. Anal Biochem 231:263–265 . https://doi.org/10.1006/abio.1995.1530
15. Zhao J, Bacolla A, Wang G, Vasquez KM (2010) Non-B DNA structure-induced genetic instability and evolution. Cell Mol Life Sci 67:43–62 . https://doi.org/10.1007/s00018-009-0131-2
16. Clark J, Smith SS (2008) Secondary structure at a hot spot for DNA methylation in DNA from human breast cancers. Cancer Genomics and Proteomics 5:241–252
17. Bacolla A, Cooper DN, Vasquez KM, Tainer JA (2018) Non-B DNA Structure and Mutations Causing Human Genetic Disease. In: eLS. John Wiley & Sons, Ltd, pp 1–15
18. Bacolla A, Wells RD (2009) Non-B DNA conformations as determinants of mutagenesis and human disease. Mol Carcinog 48:273–285 . https://doi.org/10.1002/mc.20507
19. Walter J, Hutter B, Khare T, Paulsen M (2006) Repetitive elements in imprinted genes. Cytogenet Genome Res 113:109–115 . https://doi.org/10.1159/000090821
20. Ziller MJ, Gu H, Müller F, Donaghey J, Tsai LTY, Kohlbacher O, De Jager PL, Rosen ED, Bennett DA, Bernstein BE, Gnirke A, Meissner A (2013) Charting a dynamic DNA methylation landscape of the human genome. Nature 500:477–481 . https://doi.org/10.1038/nature12433
21. Zhou L, Ng HK, Drautz-Moses DI, Schuster SC, Beck S, Kim C, Chambers JC, Loh M (2019) Systematic evaluation of library preparation methods and sequencing platforms for high-throughput whole genome bisulfite sequencing. Sci Rep 9:1–16 . https://doi.org/10.1038/s41598-019-46875-5
22. Ziller MJ, Hansen KD, Meissner A, Aryee MJ (2015) Coverage recommendations for methylation analysis by whole-genome bisulfite sequencing. Nat Methods 12:230–232 . https://doi.org/10.1038/nmeth.3152
23. Zhang W, Spector TD, Deloukas P, Bell JT, Engelhardt BE (2015) Predicting genome-wide DNA methylation using methylation marks, genomic position, and DNA regulatory elements. Genome Biol 16: . https://doi.org/10.1186/s13059-015-0581-9
24. (2006) Linear Mixed-Effects Models: Basic Concepts and Examples. In: Mixed-Effects Models in S and S-PLUS. Springer-Verlag, pp 3–56
25. Rue H, Martino S, Chopin N (2009) Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. J R Stat Soc

Ser B Stat Methodol 71:319–392 . https://doi.org/10.1111/j.1467-9868.2008.00700.x

26. Spiegelhalter DJ, Best NG, Carlin BP, Van Der Linde A (2002) Bayesian measures of model complexity and fit. J R Stat Soc Ser B Stat Methodol 64:583–616 . https://doi.org/10.1111/1467-9868.00353
27. Monk D, Mackay DJG, Eggermann T, Maher ER, Riccio A (2019) Genomic imprinting disorders: lessons on how genome, epigenome and environment interact. Nat Rev Genet 20:235–248 . https://doi.org/10.1038/s41576-018-0092-0
28. Claus R, Lucas DM, Ruppert AS, Williams KE, Weng D, Patterson K, Zucknick M, Oakes CC, Rassenti LZ, Greaves AW, Geyer S, Wierda WG, Brown JR, Gribben JG, Barrientos JC, Rai KR, Kay NE, Kipps TJ, Shields P, Zhao W, Grever MR, Plass C, Byrd JC (2014) Validation of ZAP-70 methylation and its relative significance in predicting outcome in chronic lymphocytic leukemia. Blood 124:42–48 . https://doi.org/10.1182/blood-2014-02-555722
29. DeVos T, Tetzner R, Model F, Weiss G, Schuster M, Distler J, Steiger K V., Grützmann R, Pilarsky C, Habermann JK, Fleshner PR, Oubre BM, Day R, Sledziewski AZ, Lofton-Day C (2009) Circulating methylated SEPT9 DNA in plasma is a biomarker for colorectal cancer. Clin Chem 55:1337–1346 . https://doi.org/10.1373/clinchem.2008.115808
30. Giovannetti E, Codacci-Pisanelli G, Peters GJ (2012) TFAP2E-DKK4 and chemoresistance in colorectal cancer. N Engl J Med 366:966 . https://doi.org/10.1056/NEJMc1201170
31. Hegi ME, Diserens AC, Gorlia T, Hamou MF, De Tribolet N, Weller M, Kros JM, Hainfellner JA, Mason W, Mariani L, Bromberg JEC, Hau P, Mirimanoff RO, Cairncross JG, Janzer RC, Stupp R (2005) MGMT gene silencing and benefit from temozolomide in glioblastoma. N Engl J Med 352:997–1003 . https://doi.org/10.1056/NEJMoa043331
32. Queirós AC, Villamor N, Clot G, Martinez-Trillos A, Kulis M, Navarro A, Penas EMM, Jayne S, Majid A, Richter J, Bergmann AK, Kolarova J, Royo C, Russiñol N, Castellano G, Pinyol M, Bea S, Salaverria I, López-Guerra M, Colomer D, Aymerich M, Rozman M, Delgado J, Giné E, González-Díaz M, Puente XS, Siebert R, Dyer MJS, López-Otín C, Rozman C, Campo E, López-Guillermo A, Martín-Subero JI (2015) A B-cell epigenetic signature defines three biologic subgroups of chronic lymphocytic leukemia with clinical impact. Leukemia 29:598–605 . https://doi.org/10.1038/leu.2014.252
33. Papageorgiou EA, Karagrigoriou A, Tsaliki E, Velissariou V, Carter NP, Patsalis PC (2011) Fetal-specific DNA methylation ratio permits noninvasive prenatal diagnosis of trisomy 21. Nat Med 17:510–513 . https://doi.org/10.1038/nm.2312

34. McDermott E, Ryan EJ, Tosetto M, Gibson D, Burrage J, Keegan D, Byrne K, Crowe E, Sexton G, Malone K, Harris RA, Kellermayer R, Mill J, Cullen G, Doherty GA, Mulcahy H, Murphy TM (2016) DNA methylation profiling in inflammatory bowel disease provides new insights into disease pathogenesis. J Crohn's Colitis 10:77–86 . https://doi.org/10.1093/ecco-jcc/jjv176
35. Imgenberg-Kreuz J, Almlöf JC, Leonard D, Sjöwall C, Syvänen AC, Rönnblom L, Sandling JK, Nordmark G (2019) Shared and Unique Patterns of DNA Methylation in Systemic Lupus Erythematosus and Primary Sjögren's Syndrome. Front Immunol 10:1686 . https://doi.org/10.3389/fimmu.2019.01686
36. Fernandez-Jimenez N, Garcia-Etxebarria K, Plaza-Izurieta L, Romero-Garmendia I, Jauregi-Miguel A, Legarda M, Ecsedi S, Castellanos-Rubio A, Cahais V, Cuenin C, Degli Esposti D, Irastorza I, Hernandez-Vargas H, Herceg Z, Bilbao JR (2019) The methylome of the celiac intestinal epithelium harbours genotype-independent alterations in the HLA region. Sci Rep 9:1–13 . https://doi.org/10.1038/s41598-018-37746-6
37. Maher ER, Reik W (2000) Beckwith-Wiedemann syndrome: Imprinting in clusters revisited. J. Clin. Invest. 105:247–252
38. Lea AJ, Tung J, Zhou X (2015) A Flexible, Efficient Binomial Mixed Model for Identifying Differential DNA Methylation in Bisulfite Sequencing Data. PLoS Genet 11: . https://doi.org/10.1371/journal.pgen.1005650
39. Jin Z, Liu Y (2018) DNA methylation in human diseases. Genes Dis 5:1–8 . https://doi.org/10.1016/j.gendis.2018.01.002
40. Liu Y, Aryee MJ, Padyukov L, Fallin MD, Hesselberg E, Runarsson A, Reinius L, Acevedo N, Taub M, Ronninger M, Shchetynsky K, Scheynius A, Kere J, Alfredsson L, Klareskog L, Ekström TJ, Feinberg AP (2013) Epigenome-wide association data implicate DNA methylation as an intermediary of genetic risk in rheumatoid arthritis. Nat Biotechnol 31:142–147 . https://doi.org/10.1038/nbt.2487
41. Dick KJ, Nelson CP, Tsaprouni L, Sandling JK, Aïssi D, Wahl S, Meduri E, Morange PE, Gagnon F, Grallert H, Waldenberger M, Peters A, Erdmann J, Hengstenberg C, Cambien F, Goodall AH, Ouwehand WH, Schunkert H, Thompson JR, Spector TD, Gieger C, Trégouët DA, Deloukas P, Samani NJ (2014) DNA methylation and body-mass index: A genome-wide analysis. Lancet 383:1990–1998 . https://doi.org/10.1016/S0140-6736(13)62674-4
42. Andrews S V., Ellis SE, Bakulski KM, Sheppard B, Croen LA, Hertz-Picciotto I, Newschaffer CJ, Feinberg AP, Arking DE, Ladd-Acosta C, Fallin MD (2017) Cross-tissue integration of genetic and epigenetic data offers insight into autism spectrum disorder. Nat Commun 8:1–10 . https://doi.org/10.1038/s41467-017-00868-y