

# A comprehensive overview of Infinium HumanMethylation450 data processing

Sarah Dedeurwaerder\*, Matthieu Defrance\*, Martin Bizet, Emilie Calonne, Gianluca Bontempi and François Fuks

Submitted: 25th April 2013; Received (in revised form): 5th July 2013

## Abstract

Infinium HumanMethylation450 beadarray is a popular technology to explore DNA methylomes in health and disease, and there is a current explosion in the use of this technique. Despite experience acquired from gene expression microarrays, analyzing Infinium Methylation arrays appeared more complex than initially thought and several difficulties have been encountered, as those arrays display specific features that need to be taken into consideration during data processing. Here, we review several issues that have been highlighted by the scientific community, and we present an overview of the general data processing scheme and an evaluation of the different normalization methods available to date to guide the 450K users in their analysis and data interpretation.

**Keywords:** Epigenomics; Genome-wide DNA methylation technology

## BACKGROUND

DNA methylation is involved in numerous physiological processes and also disease states, such as cancer [1]. This has raised wide interest in developing large-scale DNA methylation profiling technologies to improve our molecular understanding of diseases. The recently released Infinium HumanMethylation450 [2, 3] is a preferred technology for studying the DNA methylomes of various cell types in large-scale studies, and there is a current explosion of data generated with this technology [4]. Sequencing-based methods, although offering much higher genome coverage, are still not affordable by all laboratories, notably those with moderate budgets. Another reason for the success of DNA methylation arrays is the ease of reading and understanding

the data generated, notably because microarrays have been widely used over the past decades, particularly for gene expression profiling. Yet, accurate processing of Infinium HumanMethylation450 data remains difficult because of several confounding parameters. This is the subject of this review.

## The Infinium HumanMethylation450 technology

The Infinium HumanMethylation450 array makes it possible to assess the methylation status of >450 000 CpGs located throughout the genome [2]. Its particularity lies in the use of two different types of chemical assays (Infinium I and Infinium II) [3]. Both are based on a quantitative genotyping of the C/T polymorphism generated by DNA bisulfite

Corresponding author: François Fuks, Laboratory of Cancer Epigenetics, Université Libre de Bruxelles, Route de Lennik 808, 1070 Brussels, Belgium. Tel.: +32-0-2-555-62-45; Fax: +32-0-2-555-62-57; E-mail: ffuks@ulb.ac.be

\*These authors contributed equally to this work.

**Sarah Dedeurwaerder** is a post-doc at the Laboratory of Cancer Epigenetics, Université Libre de Bruxelles. Her projects focus on the study of the DNA methylome of breast cancer samples to identify potential biomarkers for diagnosis, prognosis or response to treatment.

**Matthieu Defrance** is a post-doc at the Laboratory of Cancer Epigenetics, Université Libre de Bruxelles. He is a computational scientist with strong expertise in the analysis of DNA methylation arrays, ChIP-seq data and next-generation sequencing data in general.

**Martin Bizet** is a PhD student at the Laboratory of Cancer Epigenetics and the Machine Learning Group, Université Libre de Bruxelles. He is a computational scientist.

**Emilie Calonne** is a lab technician at the Laboratory of Cancer Epigenetics, Université Libre de Bruxelles. She performs microarray and next-generation sequencing experiments.

**Gianluca Bontempi** is Director of the Machine Learning Group, Université Libre de Bruxelles. His research interests are the application of machine learning methods to many problems, including computational biology.

**François Fuks** is Director of the Laboratory of Cancer Epigenetics, Université Libre de Bruxelles. His research focuses on the role of epigenetic modifications, in particular DNA and histone modifications, in health and disease.

conversion, but the Infinium I assay resembles a single-channel microarray, whereas the Infinium II assay has a dual-color readout. The Infinium I assay uses two types of probes (one for the methylated allele and one for the unmethylated allele), and base extension is the same for both alleles. The Infinium II assay uses a single probe for both alleles, and base extension depends on the methylation state of the hybridized genomic DNA molecule (for a detailed illustration, see [3], Figure 2).

We have shown that this particularity in the design of the 450K array has important consequences on the generated data [3]. We have notably observed that Infinium II probes show a reduced dynamic range of measured methylation values as compared with the Infinium I probes. Thus, an additional step is required to correct the performance of the Infinium II assay when preprocessing Infinium HumanMethylation450 data, and this processing already comprises several steps, notably filtering out defective probes, correcting dye bias and normalizing to eliminate a potential batch effect (Figure 1). Several pipelines and R packages have been developed or are under development for processing Infinium HumanMethylation450 data. It is difficult for 450K users to choose the best package and normalization method. Here, to guide 450K users in their analysis and data interpretation, we present an overview of the general data processing scheme (Figure 1) and an evaluation of the different normalization methods available to date.

## FILTERING OUT THE PROBLEMATIC PROBES

From our point of view, the first step when performing microarray data preprocessing is to filter out every probes that can generate artifactual data. Other scientists would perform this step at the end of the data preprocessing (i.e. after the normalization step) to avoid doing again the normalization step if someone wants to look at a different probe set than the one initially selected. Nevertheless, we think that it is more judicious to start by filtering out the problematic probes as values obtained from those probes appear not reliable, and therefore we do not wish to take these into account for further analyses. Also, we cannot exclude the possibility that these probes do not influence normalization.

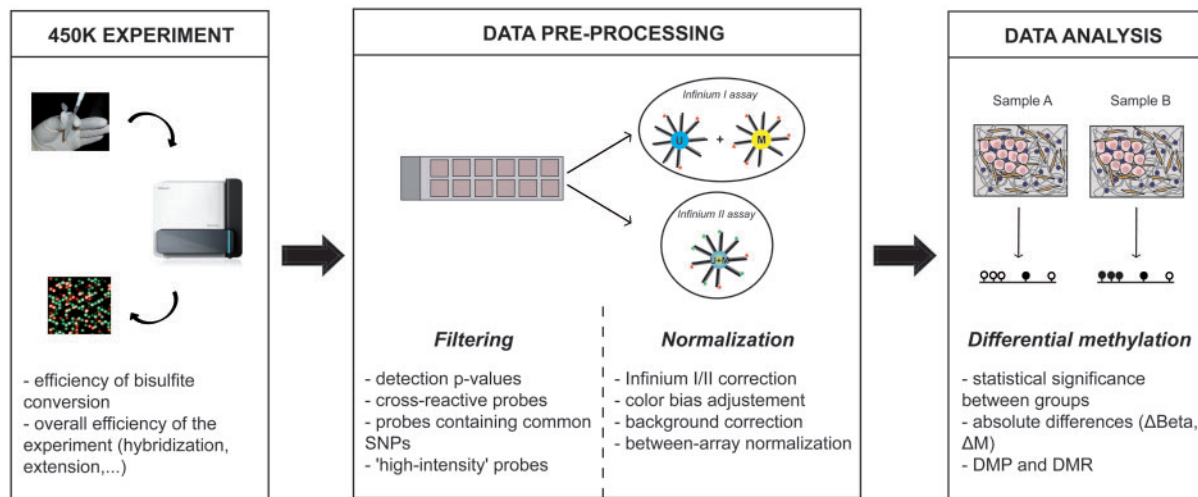
Several reasons can explain the generation of artifactual data. For example, the scanner can encounter

some difficulties to correctly read the signal for some probes owing to their low intensities or to some spatial artifacts on the array. This problem translates as a high detection  $P$ -value (i.e. a low quality signal) for the probes concerned. It is therefore strongly recommended to filter out probes displaying a high detection  $P$ -value (e.g.  $>0.05$ ) before performing downstream analyses. This problem is common to all microarray platforms. In this section, we focus on three other problems, specific to Infinium HumanMethylation450 arrays, which should lead to filtering out particular probes: (i) the cross-reactive probes mapping to multiple locations on the genome; (ii) the probes containing common single nucleotide polymorphisms (SNPs); and (iii) the probes displaying a very high average intensity.

### Cross-reactive probes

Infinium HumanMethylation450 uses bisulfite treatment to convert unmethylated cytosines, but not methylated one's, to uracils, generating at CpG sites after DNA amplification a C/T polymorphism that is readily detectable with the Infinium technology [5]. Another consequence of this bisulfite treatment is the generation—from an initial '4-letters genome' (A,T,G,C)—of an almost '3-letters genome' (A, T and G; the only remaining C's being methylated C's, i.e.  $\sim 3.5\%$  of the total number of the C's). This considerably increases the probability of probe cross-reactivity, i.e. the probability that some of the 50mer Infinium probes will co-hybridize at additional locations on the genome, different from the regions for which the probes were initially designed.

Between 8.6 and 25% of the Infinium HumanMethylation450 probes have been identified as non-specific, i.e. cross-reactive, depending on the criteria used [6, 7]. This is particularly problematic, as a DNA methylation measurement from a cross-reactive probe is likely to represent a combination of the methylation levels of multiple genomic sites and not the methylation level of the initially targeted CpG site (for an illustration, see [8], Figure S2). Consequently, wrong methylation measurements are generated and can lead to detecting artifactual differentially methylated sites. For example, numerous sex-associated differences in methylation are reported to be technical artifacts created by autosomal probes cross-reacting with genomic regions on the sex chromosomes [8]. To avoid reporting artifactual differentially methylated



**Figure 1:** Overview of the general Infinium HumanMethylation450 data processing scheme with highlights on the different points to check during the processing to ensure an accurate analysis and interpretation. DMP, differentially methylated positions; DMR, differentially methylated regions.

sites, it is therefore recommended to disregard these non-specific probes and/or to use another approach such as bisulfite pyrosequencing (BPS) to check the DNA methylation measurements obtained from them.

### Probes containing common SNPs

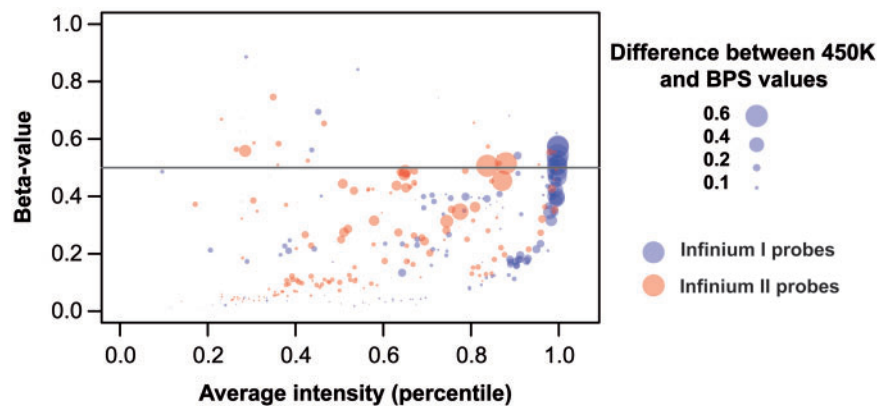
As aforementioned, the Infinium Methylation assay is based on the quantitative genotyping of C/T SNPs generated at CpG sites by bisulfite treatment of the DNA. A limitation of this method is that it can also detect C/T polymorphisms naturally present at the interrogated CpG sites (i.e. the genotype). DNA methylation measurements can thus be confounded by the actual DNA sequence [6, 8] (for an illustration, see [8], Figure S2). If one considers a fully methylated CpG site, for instance, in samples of genotype C/C the DNA methylation measurements approach 100% (as expected), whereas in samples of genotype T/T, the measurements will always be close to 0%. If a sample is heterozygous, the DNA methylation value measured will be  $\sim 50\%$ . Thus, in the case of probes containing SNPs at the targeted CpG site, Infinium measurements are more likely to reflect the genotype of the sample rather than a true DNA methylation value.

Some 4.3% of the Infinium HumanMethylation450 probes are reported to contain a known polymorphism specifically at the targeted C or G [6]. In the case of intra-individual studies (such as longitudinal studies) or ones involving monozygotic

twins, the presence of SNPs at some targeted CpG sites should not be an important confounder, but it should cause problems in inter-individual studies comparing, for instance, a group of healthy subjects with a group of patients suffering from a particular disease. The problem depends on the frequency of heterozygosity. Although 56.8% of these probes display infrequent SNPs, 43.2% have a polymorphism that is more frequent in the population (frequency of heterozygosity above 0.1) and are therefore more likely to confound the DNA methylation measurements [6]. In addition to SNPs located specifically at the targeted CpG site, SNPs can also be present within the remainder of the probe. Although it is known from other microarray platforms that the presence of one or more SNPs in a probe can affect its hybridization, it seems that DNA methylation measurements are not affected too much by the presence of such SNPs [6]. In conclusion, it seems important in inter-individual studies to filter out probes containing a frequent SNP at the targeted CpG site and/or to perform SNP genotyping in parallel of Infinium Methylation experiment. In intra-individual studies, filtering out these probes is probably not necessary.

### Other problematic probes

It is easy to understand why the cross-reactive probes and probes containing SNPs at the targeted CpGs can generate artifactual data, but other probe measurements can also be problematic for more obscure



**Figure 2:** CpGs with high average signal intensity display lower concordance with BPS data. Plot illustrating the difference between methylation values obtained from Infinium HumanMethylation450 and BPS as a function of the average signal intensity and the  $\beta$ -value. The absolute difference between the two techniques is proportional to the circle radius (blue: type I probes; red: type II probes). The plot was generated using 450K and matched BPS data from 22 tissues described in [9] (352 points). A colour version of this figure is available at BIB online: <http://bib.oxfordjournals.org>.

reasons. For example, we have looked at the relation between signal intensities and DNA methylation measurements [ $\beta$ -values, defined as the ratio of the methylated signal over the total signal (methylated + unmethylated)] and have observed that probes displaying a high average intensity (i.e. a high average of the methylated and unmethylated signals) are more prone than probes displaying a lower average intensity to provide DNA methylation measurements inconsistent with measurements obtained with other approaches, such as BPS (Figure 2). They have a tendency to provide values close to 0.5, independently of their true methylation state. Of note, type II Infinium probes seem to be less prone to this phenomenon (Figure 2).

We wish to warn the Infinium HumanMethylation450 users about these ‘high-intensity probes’ providing measurements close to 0.5. They might have to be filtered out before downstream analysis or at least their measurements need to be checked with another approach. In general, we would recommend being cautious with any probe displaying extreme values of any parameter, i.e. a high average intensity (as described earlier in the text) or also a high standard deviation between bead replicates, for instance.

## NORMALIZING THE DATA

The second and key step of microarray data preprocessing consists in removing any source of variation that is not related to biology but rather to technical limitations, such as dye bias or batch effect. This step is

called data normalization. Although Infinium HumanMethylation450 is a two-color channel microarray, the methods developed previously for gene expression arrays cannot be used as such. The rationale behind this has already been reviewed elsewhere [10]. As briefly described later in the text, Infinium HumanMethylation450 displays specific properties. First, the two color channels are used on the Infinium array to measure the methylation state of a single sample, whereas in gene expression arrays, each color channel is associated with a different sample. Second, normalization methods developed for gene expression arrays frequently assume that the experimental condition alters the expression of only a small number of genes. Based on this assumption, the sum of the fluorescence across all genes for each microarray experiment should be the same. This hypothesis is not verified in a methylation context, as the global methylation level can vary from one sample to another. Third, the particular design of the 450K array makes it necessary to perform a normalization between the Infinium I and Infinium II probes. For all these reasons, Infinium HumanMethylation450-specific normalization methods are required. A lot of methods are already available (see Table 1), and it is not that easy to know which one, or which combination, is the most suitable. In this section, we review the different normalization methods developed for Infinium HumanMethylation450, distinguishing within-array and between-array methods, and we try to guide the 450K users in their choice of a normalization method.

**Table I:** Freely available packages/pipelines for Infinium 450K data preprocessing and analysis

Package	Description	References
<i>IMA</i>	R-Package. Pipeline function available: allows probe filtering (detection p-values, SNPs, ...) and identification of differentially methylated sites. Proposed normalization methods: type I/II bias correction PBC and quantile normalization between arrays (pipeline options).	[11] <a href="http://ima.r-forge.r-project.org/">http://ima.r-forge.r-project.org/</a> PBC: [3]
<i>Lumi</i>	Bioconductor R-package. No pipeline function. Proposed normalization methods: background corrections (using or not negative controls), color bias adjustment and between-array normalization methods ('smooth quantile' or 'shift and scaled normalization').	[12] <a href="http://www.bioconductor.org/">http://www.bioconductor.org/</a>
<i>Minfi</i>	Bioconductor R-package. No pipeline function. Proposed normalization methods: type I/II bias correction SWAN and dye bias equalization (originally proposed in the Genome Studio software).	<a href="http://www.bioconductor.org/">http://www.bioconductor.org/</a> SWAN: [13]
<i>wateRmelon</i>	Bioconductor R-package. Pipeline function available: pipeline proposed by Touleimat and Tost. Other proposed normalization methods: wide-range of within-array normalization methods, including the type I/II bias corrections PBC, BMIQ and SWAN, and between-array normalization methods, such as Nasen.	<a href="http://mammoet.iop.kcl.ac.uk/i450/INSTALL.html">http://mammoet.iop.kcl.ac.uk/i450/INSTALL.html</a> <a href="http://www.bioconductor.org/">http://www.bioconductor.org/</a> Pipeline proposed by Touleimat and Tost: [14] PBC: [3]; SWAN: [13]; BMIQ: [15] (also available independently at <a href="http://code.google.com/p/bmiq/">http://code.google.com/p/bmiq/</a> ) Nasen: [16]
<i>Methylumi</i>	Bioconductor R-package. No pipeline function. Proposed normalization methods: dye bias equalization and several background correction methods, including the recently developed Noob method.	<a href="http://www.bioconductor.org/">http://www.bioconductor.org/</a> Noob: [17]
<i>RnBeads</i>	R-package. Pipeline function available: allows probe filtering, quality control, estimation of batch effect and identification of differentially methylated sites. Proposed normalization methods: type I/II bias correction SWAN and dye bias equalization.	<a href="http://rnbeads.bioinf.mpi-inf.mpg.de/">http://rnbeads.bioinf.mpi-inf.mpg.de/</a> SWAN: [13]
<i>NIMBL</i>	Mathlab package dedicated to the identification of DNA methylation markers. No pipeline function. Proposed normalization methods: type I/II bias correction PBC.	[18] <a href="https://sites.google.com/site/emesbioinformatics/group-software/nimbl">https://sites.google.com/site/emesbioinformatics/group-software/nimbl</a> PBC: [3]

### Within-array normalization

For Infinium HumanMethylation450, within-array normalization concerns three main points: background correction, color bias (or dye bias) adjustment and Infinium I/II-type bias correction. Actually, at least a part of the Infinium I/II-type bias is a combination of the two first-mentioned biases. Indeed, because the Infinium II assay uses the same bead to measure both the methylated and unmethylated signals, the measurement of one of these two signals is disturbed by the residual emission of the other dye, therefore likely resulting in a higher background for Infinium II probes than for Infinium I probes, hence contributing in the reduction of the

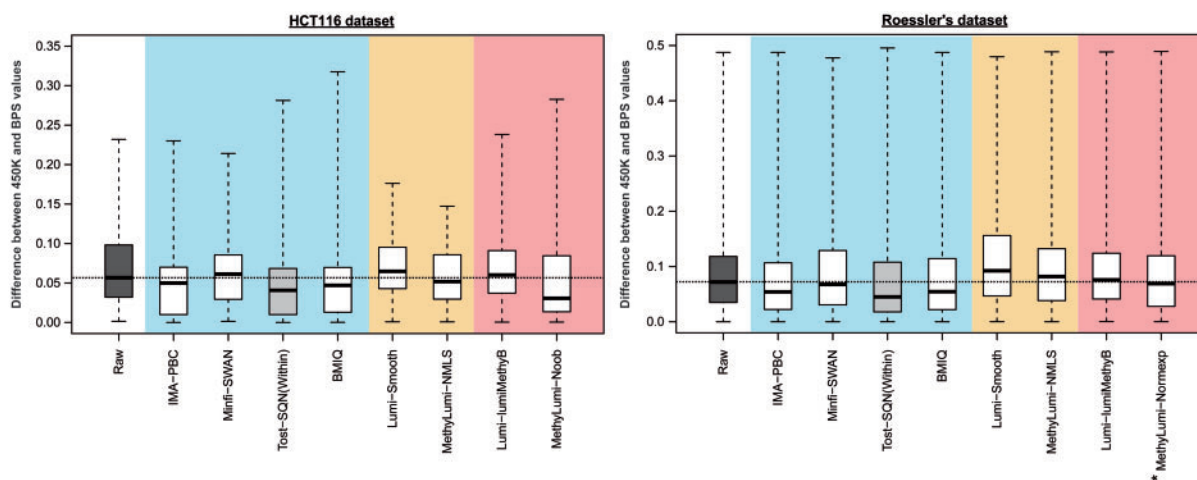
dynamic range of  $\beta$ -values for Infinium II probes as compared with Infinium I probes. Moreover, the color bias is related to the difference in intensity measurement fidelity between the two dyes. As the methylated and unmethylated states of each CpG are evaluated in the same color channel in the Infinium I assay, the dye bias has little impact on the  $\beta$ -values obtained from Infinium I probes. Nevertheless, there is a notable difference between the  $\beta$ -value range for Infinium I probes using the red or the green channel that is probably due to the different backgrounds of the two color channels. On the contrary, for the Infinium II assay, the methylated and unmethylated states of each CpG are evaluated in different channels

(the methylated signal is measured in the green channel and the unmethylated signal in the red channel), making the color bias more problematic in this case: it notably skews the  $\beta$ -values obtained from Infinium II probes and also contributes to the reduction of their dynamic range. Hence, a correction of the Infinium I/II-type bias should correct in a similar manner the dye bias and the background than a dye bias correction combined with a background correction (even if the background is balanced between the two types of Infinium probes but not completely eliminated). Of note, these two approaches are unlikely to provide exactly the same effect as the Infinium I/II-type bias brings into play a third component that is the different probe design of Infinium I and II. Indeed, Infinium I assumes - for loci with flanking CpGs - that methylation is regionally correlated and therefore underlying CpGs are in phase with the methylated or unmethylated query sites, whereas Infinium II uses 'degenerate' bases [2]. During these last months, several Infinium I/II-type bias correction methods have been developed. We therefore detail here mainly these methods, before presenting briefly the color bias and background correction methods also available to date.

The first method used to correct the Infinium I/II type bias was developed by our laboratory and is called peak-based correction (PBC) [3]. As the methylation level distribution is bimodal (one peak corresponding to the unmethylated sites and the other to the methylated sites), the PBC proposes to rescale the methylation levels of the Infinium II probes to obtain the same modes for the distribution of methylation values obtained from the Infinium II probes as for the distribution of methylation values obtained from the Infinium I probes (which is kept unmodified). To illustrate the effectiveness of this method and of the others presented below, we applied them to two data sets for which we obtained BPS data that we used as reference values. The first data set has been previously described in [3] and consists of 90 measurements providing from three replicates of HCT116 WT cells and three replicates of HCT116 DKO cells (Double Knock-Out for DNMT1 and DNMT3B displaying a low global level of methylation as compared with HCT116 WT cells). Roessler and coworkers kindly provided the second data set comprising 352 measurements from 22 tissue samples [9]. As already shown [3, 19], the PBC method

results in better agreement between 450K and other technologies (such as BPS, as in the present case), and thus proves effective (Figure 3). It is worth noting that this method is sensitive to variations in the shape of the methylation density curves and is therefore less robust when applied to samples that do not display clear methylated or unmethylated peaks.

Two other proposed methods are derived from quantile normalization. Yet, because Infinium I and Infinium II probes do not interrogate the same CpG population, the two types of probes are not expected to have the same distribution [3], and classic quantile normalization methods cannot be applied as such. Subset quantile approaches have thus been proposed. Touleimat and Tost have developed a categorical Subset Quantile Normalization method (SQN) based on the assumption that CpGs having the same biological properties should have the same distribution [14]. For this purpose, they separated the target CpGs into different classes based on their location with respect to CpG islands (CGIs) and then applied quantile normalization between the Infinium I and Infinium II probes, independently for each different class of CpGs. Of note, this method is applied to all samples simultaneously, thus performing a between-array normalization at the mean time. Maksimovic and coworkers have proposed a similar method, called Subset quantile for Within Array Normalization (SWAN) [13]. Instead of classifying the target CpGs on the basis of their location with respect to CGIs, they classified the probes on the basis of the number of CpGs they contain, assuming that probes having the same number of CpGs in their sequences should reside in a similar region (CGIs or open sea) and thus have the same profile. When applied to the two aforementioned data sets, the SWAN method does not seem to improve the data quality (Figure 3). If one considers only the within-array normalization component of the method, the categorical SQN correction of Touleimat and Tost reduces the difference between methylation measurements obtained with Infinium HumanMethylation450 and BPS (Figure 3). Yet, if the complete method is applied (i.e. if the data are also subjected to the between-array normalization component of the technique), the data quality can be strongly degraded in some cases (see the next part 'Between-array Normalization' and Figure 4). The SWAN and categorical SQN methods thus have drawbacks.



**Figure 3:** Comparison of the different within-array normalization methods using BPS data as referential data. Boxplots show the distribution of the absolute difference between DNA methylation measurements obtained from Infinium HumanMethylation450 and BPS, when Infinium data are subjected (white) or not (dark gray) to within-array normalization, for HCT116 and Roessler's data sets. Blue, orange and red indicate Infinium type/I bias correction methods, color bias adjustment and background correction methods, respectively. Raw: Infinium raw data; IMA-PBC: PBC from the *IMA* package; Minfi-SWAN: Subset quantile Within-Array Normalization from the *minfi* package; Tost-SQN(within): categorical SQN from Touleimat and Tost pipeline (this boxplot is highlighted in light gray to indicate that each sample has been normalized individually to apply only the within-array normalization component of this method); BMIQ: Beta-Mixture Quantile Normalization; Lumi-Smooth: color bias adjustment from the *lumi* package (smooth quantile normalization); MethyLumi-NMLS: dye bias equalization (*normalizeMethyLumiSet*) of the *methylumi* package (method originally proposed in the Genome Studio software); Lumi-lumiMethyB: background correction from the *lumi* package; MethyLumi-Noob: background correction based on normal exponential convolution model using out-of-band intensities as controls from the *methylumi* package; MethyLumi-Normexp: same as MethyLumi-Noob but controls used are negative probes present on the array (\*On the Roessler's data set, this method was applied instead of the 'Noob' method because we do not have access to the IDAT files of these samples). A colour version of this figure is available at BIB online: <http://bib.oxfordjournals.org>.

A fourth method also based on quantile normalization was published more recently to correct the Infinium I/II type bias [15]. It is called Beta Mixture Quantile normalization (BMIQ). This method decomposes the density profile of Infinium I and Infinium II probes into two mixtures of three  $\beta$ -distributions based on the three methylation states: unmethylated (close to 0), partially methylated (close to 0.5) and fully methylated (close to 1). Then it uses a quantile normalization to fit each  $\beta$ -distribution of the Infinium II profile to the corresponding  $\beta$ -distribution of the Infinium I profile. Unlike SWAN and the categorical SQN method of Touleimat and Tost, BMIQ (like PBC) is assumption-free and does not depend on arbitrary choices of biological characteristics to be used to perform SQN. This method thus seems to us more suitable than the SWAN and Touleimat methods. Our comparison with the BPS data for our two test data sets tends to confirm this view, with global improvement of the data quality (Figure 3: for both data sets, the

median of the boxplot is lower after BMIQ correction than for the raw data). A drawback, however, is that some points appear worse after correction (Figure 3: for the HCT116 data set, the maximum of the boxplot whisker is higher after BMIQ correction than for the raw data).

In addition to these methods for correcting the Infinium I/II-type bias, other within-array normalization methods have also been developed, such as methods for correcting dye bias or eliminating the background. For instance, the popular *lumi* and *methylumi* R packages proposed a color bias adjustment based on smooth quantile or shift-and-scaling normalization and on the method proposed in Genome Studio, respectively. These corrections improve marginally the quality of the HCT116 data and seem to decrease the quality of the Roessler's data (Figure 3). Nevertheless, of potential interest, in the HCT116 data set, these methods (notably the dye bias equalization of the *methylumi* package) decrease the range of differences between the

methylation measurements obtained with 450K and BPS (Figure 3: the maximum of the boxplot whisker is lower after the color bias corrections than for the raw data). Concerning the background correction methods, most of them (such as those implemented in the *lumi* package) involve a simple background subtraction and do not significantly improve data quality (Figure 3, example shown for the ‘lumiMethyB’ method of the *lumi* package). Nevertheless, new background correction methods have recently been developed by Triche and coworkers and implemented in the *methyllumi* package (such as ‘Noob’ and ‘Normexp’) [17]. They are based on convolution models and use, most of the time, the out-of-band intensities, i.e. the Infinium I probes in the color channel opposite their designed base extension, to measure the background. Globally, these methods outperform the previously developed background correction methods and in some cases seem to improve data quality almost as well as the best Infinium I/II-type bias corrections (Figure 3, examples shown for the ‘Noob’ method on the HCT116 data set and for the ‘Normexp’ method on Roessler’s data set). Nevertheless, as with the BMIQ correction, some points appear worse after the background correction (Figure 3: for the HCT116 data set, the maximum of the boxplot whisker is higher after the ‘Noob’ correction than for the raw data). It is worth noting that the methods proposed by Triche and coworkers tend to increase the dynamic range of the  $\beta$ -values and to have the greatest reduction in bias at the extremes of the  $\beta$ -value distribution (i.e. close to 0 and 1), resembling, in this sense, our PBC method. But in this case, rescaling of the  $\beta$ -value range occurs as a natural consequence of background correction and affects both the Infinium I and Infinium II probes [17].

In conclusion as regards within-array normalization, the key point that we wished to highlight here (in agreement with the conclusions of [20]) is that the Infinium I/II type bias seems to be the one it is most crucial to correct, as all techniques that adequately address this bias improve 450K data more significantly than the others. We would therefore recommend using PBC, which seems to be the only method giving a true global benefit without generating any worse data. Yet, if samples display no clear methylated or unmethylated peaks, BMIQ or Triche and coworkers’ background correction methods based on convolution models can be good alternative methods. Also, for purposes of clarity, we

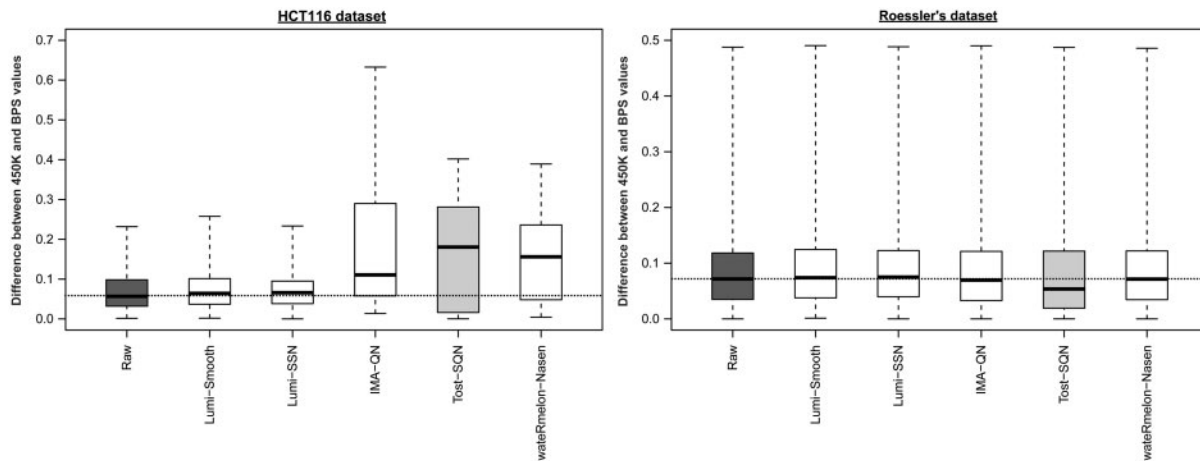
tested here all the methods separately, but we do not want to exclude the possibility of using some of them in combination. For instance, applying the dye bias equalization of the *methyllumi* package in combination with the background correction ‘Noob’ could improve the benefit obtained when these methods are used separately [17].

### Between-array normalization

In addition to technical biases linked to the array design itself, other sources of non-biological variations related to external parameters, such as unequal quantities of starting material, differences in labeling or detection efficiencies can lead to misleading results. Between-array normalization methods have been developed to reduce these array-to-array variations by adjusting measurements at a global level. Of note, non-biological technical variations tend actually to be less pronounced for Infinium data than for gene expression data because DNA methylation measurement used for sample comparisons ( $\beta$ -value) is a ratio of intensities, whereas gene expression measurement corresponds directly to the signal intensity.

To our knowledge, all between-array normalization methods proposed to date in the different packages for 450K data processing are derived from normalization methods initially developed for gene expression arrays. The *ima* package offers quantile normalization on the  $\beta$ -values as an alternative to no normalization [11]. With the *lumi* package, a *smooth* quantile normalization can be applied to the intensities or the intensities can be rescaled with a *shift and scaling normalization*. Other methods take into account the design of the Infinium HumanMethylation450 array and process separately the signals from type I and type II probes. For example, in the *wateRmelon* package, the ‘nasen’ method consists in four quantile normalizations between samples, as the data are separated according to probe type (Infinium I or Infinium II) and color channel [16]. The categorical SQN method of Touleimat and Tost takes also the array design into account [14]. Interestingly with Roessler’s data set, which is more or less homogeneous in terms of global methylation level of the samples, all these normalization methods bring no or very little benefit (except for the Touleimat and Tost method, but the benefit is attributable to the within-array normalization component of the method), and with the HCT116 data set, which displays very strong differences in terms of global methylation





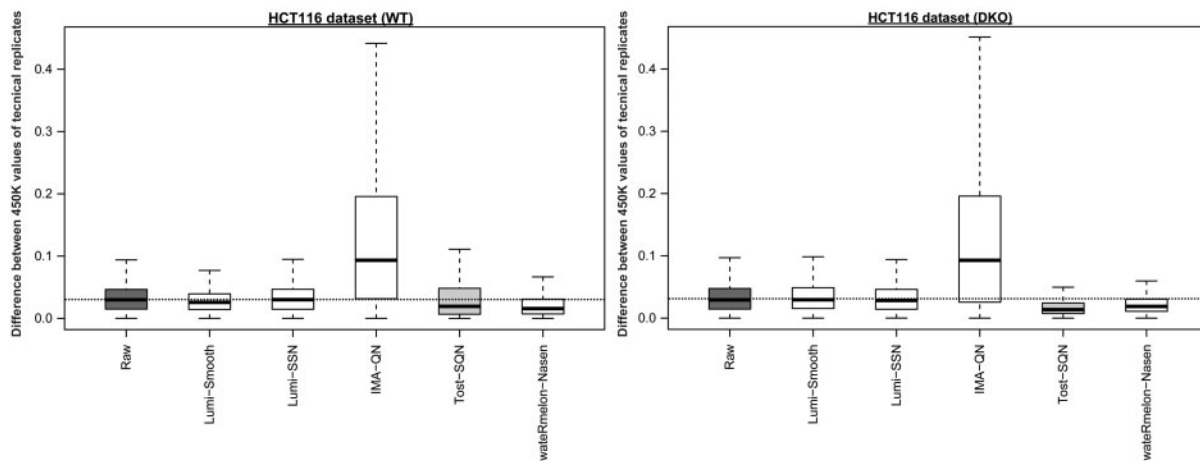
**Figure 4:** Comparison of the different between-array normalization methods using BPS data as referential data. Boxplots show the distribution of the absolute difference between DNA methylation measurements obtained from Infinium HumanMethylation450 and BPS, when Infinium data are subjected (white) or not (dark gray) to between-array normalization, for HCT116 and Roessler’s data sets. Raw: Infinium raw data; Lumi-Smooth: Smooth quantile normalization on intensities from the *lumi* package; Lumi-SSN: Shift and Scaling Normalization on the intensities from the *lumi* package; IMA-QN: Quantile normalization on  $\beta$ -values from the *IMA* package; Tost-SQN: categorical SQN from Touleimat and Tost pipeline (this boxplot is highlighted in light gray to indicate that the normalization method comprises a within-array normalization component in addition to the between-array component); wateRmelon-Nasen: Nasen method from the *wateRmelon* package.

level between samples, they strongly decrease the data quality (Figure 4). The explanation is that all these methods—except the *shift and scaling normalization* of the *lumi* package (which appears as the least bad method)—are quantile-derived methods assuming the same global distribution between samples. This hypothesis is more or less verified for Roessler’s data set but is not verified at all for the HCT116 data set (the HCT116 DKO samples displaying a low global methylation level as compared to HCT116 WT cells). Thus, in our opinion, there is to date no between-array normalization method suited to 450K data that can bring enough benefit to counterbalance the strong impairment of data quality they can cause on some data sets.

We think these observations are very informative for the 450K users. Generally, to evaluate the effectiveness of a normalization method, researchers look at the agreement between technical replicates. Although this is an important point, it is also crucial to verify that the normalization does not shift the measurements from their true biological values, by double-checking the results obtained using another technology. This is what we did here and why our conclusion could partially contradict the one of others, like Sun and coworkers [21]. Indeed, Sun and coworkers showed that the variation between technical replicates decreases after performing one

of the between-array normalization they evaluated. When looking at our HCT116 data set, we also found that the between-array normalization methods slightly decrease the variation between technical replicates (except the quantile normalization implemented in the *IMA* package, Figure 5). Nevertheless, as described earlier in the text, we also showed that the majority of the normalization methods we have tested shifted the measurements from their true biological values, using BPS data as referential data. It therefore led us to conclude that these normalization methods are not suitable for 450K data.

It is noteworthy that we fully agree with Sun and coworkers on the fact that between-array normalization methods (even if one would exist for 450K data) can partially but not completely remove another type of non-biological variations that we called ‘batch’ and ‘slide’ effects. Batch effects correspond to non-biological variations existing between batches of samples that, for instance, have not been processed the same day, on the same scanner, or by the same experimenter. The position of the array on the slide and the slide itself inside a same batch of samples can also generate non-biological variations. This is what we referred as slide effects. Such batch and slide effects can generate artifacts on measurements at the global level that could be partially removed thanks to



**Figure 5:** Comparison of the different between-array normalization methods using the variation between technical replicates as criterion. Boxplots show the distribution of the median of the absolute differences between DNA methylation measurements obtained with Infinium HumanMethylation450 from three replicates of HCT116 WT cells (left panel) or three replicates of HCT116 DKO cells (right panel), when data are subjected (white) or not subjected (dark gray) to between-array normalization. Raw: Infinium raw data; Lumi-Smooth: Smooth quantile normalization on intensities from the *lumi* package; Lumi-SSN: Shift and Scaling Normalization on the intensities from the *lumi* package; IMA-QN: Quantile normalization on  $\beta$ -values from the *IMA* package; Tost-SQN: categorical SQN from Touleimat and Tost pipeline (this boxplot is highlighted in light gray to indicate that the normalization method comprises a within-array normalization component in addition to the between-array component); waterMelon-Nasen: Nasen method from the *waterMelon* package. For clarity reasons, the boxplots are drawn using whiskers that extend to the most extreme data point, which is no more than 1.5 times the interquartile range from the box.

a good between-array normalization method. Nevertheless, batch effects can generate artifacts that only affect a subset of probes. These artifacts cannot be eliminated by globally normalizing the data. Therefore, an additional normalization (or correction) step should be applied to reduce as much as possible batch and slide effects. Although not evaluated in this study, some methods have been developed to this aim, such as ‘ComBat’ [22] that proved effective on Infinium 27K and 450K data [21, 23]. It is also important to keep in mind that the best way to avoid problems linked to batch and slide effects is to have a good design of the experiment, meaning a good distribution of the samples (cases and controls, for example) on the slides and processing of all the samples on the same day by the same experimenter using the same scanner. Of note, some useful tools, such as the bioconductor package *OSAT* (Optimal Sample Assignment Tool), have been developed to facilitate the allocation of samples to different batches [24].

In conclusion, although we are aware of the importance of between-array normalization for accurate sample comparisons, we do not recommend applying any between-array normalization method to Infinium HumanMethylation450 data for the

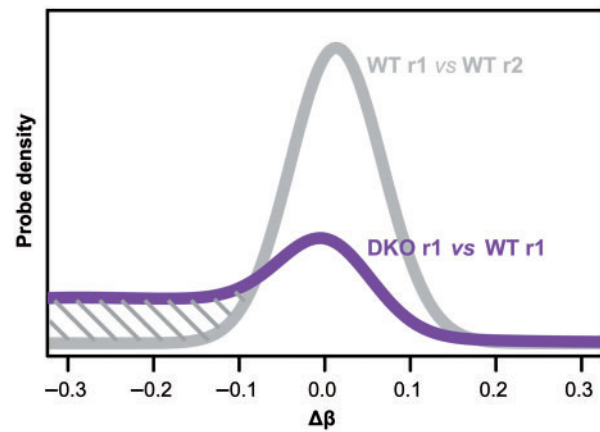
time being because technical variations are weaker for Infinium arrays than for gene expression arrays and, mainly because, from our point of view, there is to date no between-array normalization method suitable for 450K data. We would welcome, of course, the development of a suitable method bringing a real benefit. Methods, such as ‘ComBat’, developed for batch effect removal can be applied, even if possible confounding due to batch and slide effects can be at least partially avoided thanks to a good study design.

## PERFORMING THE DIFFERENTIAL METHYLATION ANALYSIS

After correct preprocessing of the data (i.e. filtering out problematic probes and normalizing the data), differential methylation analysis can be performed. Generally, the first approach consists in a single-probe analysis. Statistical tests (such as the *t*-test or Mann-Whitney test) are used, and when the *P*-values obtained are below a given threshold (e.g.  $<0.05$ ), the sites are considered as differentially methylated and referred as differentially methylated positions (DMPs). In this way, several researchers have identified numerous DMPs although the

absolute difference in methylation of the CpG sites between two groups of samples was small (i.e. below 5% of methylation difference). We wish to warn 450K users that technical replicates can frequently display methylation differences up to 10%, as illustrated in Figure 6 using two HCT116 WT replicates of our HCT116 data set. Therefore, very slight observed differences in methylation are more likely due to random technical variations than to true biological differences (Figure 6). Some very slight differences in methylation may be true differences, notably when reflecting a difference in cell-type composition of the tissues analyzed [25, 26], but the technical variability of Infinium HumanMethylation450 makes it unsuitable for confident detection of such differences. Even if the studied data set is large, the technical variability should not be neglected, as the size of the data set will reduce the impact of the technical variability but will not completely eliminate it. Thus, to ensure the selection of CpGs whose methylation difference is not artifactual, we think it is necessary to use, in addition to a statistical criterion, an absolute methylation difference threshold ( $\Delta\beta$ ) that should be determined for each experiment independently, as the technical variability can vary from one experiment to another.

The  $\beta$ -value is the default value retrieved by the Genome Studio software and is simply defined as the ratio of the methylated signal over the total signal (methylated + unmethylated). Yet another type of value, the  $M$ -value, is often used to express the degree of methylation obtained with Infinium. It is defined as the log ratio of the methylated signal over the unmethylated signal. Owing to its construction, the  $\beta$ -value is bounded between 0 and 1 (or 0 and 100%) allowing easy biological interpretation. Although using  $\beta$ -values provides a simple option, its main drawback resides in its bad statistical properties: the  $\beta$ -value has been shown to be highly heteroscedastic [27], implying that the variance across samples at the extremities of the methylation range (close to 0 and 1) is highly reduced. The  $M$ -value has better statistical properties with a lower heteroscedasticity, meaning that the variance across the methylation range is approximately constant. Therefore, we would recommend using the  $M$ -value rather than the  $\beta$ -value to perform statistical tests sensitive to heteroscedasticity, such as  $t$ -test. Other tests, such as the Mann–Whitney test (that is a rank test), are not affected by the monotonic transformation between  $\beta$ - and  $M$ -values and can therefore be applied



**Figure 6:** Small differences of methylation can be observed by chance due to technical variations. Density plot of the  $\Delta\beta$  (difference of methylation) between two technical replicates of HCT116 WT cells (in gray) and between one HCT116 WT sample and one HCT116 DKO sample (in purple). The dashed region ( $< -0.09$ ) indicates the area where random differences are lower than biological differences. A colour version of this figure is available at BIB online: <http://bib.oxfordjournals.org>.

equivalently on the  $\beta$ - or the  $M$ -values. When using an absolute methylation difference criterion, the  $\beta$ -value seems more suitable, as it allows easier biological interpretation.

In addition to the single CpG analysis, a second approach can be used to perform differential methylation analysis to bring further confidence in the results. It consists in looking at regional methylation measurements rather than at single site measurements, and therefore in identifying differentially methylated regions rather than DMPs [28, 29]. The principle of this method resides in the fact that probes being close together (e.g. inside the promoter of the same gene or in a window of a given size) should have the same behavior, i.e. hypomethylated (or hypermethylated) in the cases as compared with control samples. One limitation of this type of approach for analyzing 450K data is that about one-quarter of the array probes is isolated (i.e. located at  $>1$  kb away from any other probe), rendering analysis more difficult to apply than with sequencing data. Nevertheless, this approach can be used for the three other quarters of the array probes, particularly in promoter regions and CGIs that are generally well covered by the Infinium probes.

On top of the aforementioned issues, we still wish to add a last key one that is the necessity to verify the reliability of the forecasts. It is important to use at least a negative control, for instance, by mixing up

the samples and by performing additional differential methylation analyses using the mixed-up groups of samples. This allows to assess whether the differences observed between the groups of interest are potentially true or if such differences can be obtained by any random sampling. Furthermore, the empirical false-discovery rate for a specific cutoff (e.g.  $P$ -value chosen) can be estimated more precisely by performing permutation tests, as it has already been proposed in gene expression microarray analysis [30].

## CONCLUSION

The development of Infinium HumanMethylation450 arrays is allowing researchers to perform high-throughput DNA methylation profiling. Increased number of data has already been published and many more are to come. However, Infinium HumanMethylation450 analysis and interpretation appear not as easy as initially thought and this, given the various reasons that we have reviewed and discussed here.

First, it becomes evident that probe annotation has to be improved, as numerous probes seem to generate values that can be confounded by several parameters and need therefore to be filtered out. Some probes were notably identified as cross-reactive, i.e. they co-hybridize at different genomic locations. Others contain known SNPs and therefore evaluate more likely the genotype than the methylation level of the targeted CpG site. Also to be considered is the observation that probes displaying a high average intensity appear less reliable than those displaying lower average intensity.

Second, an adequate sample normalization has to be performed to ensure complete and correct preprocessing of the data. Concerning within-array normalization, numerous methods have been proposed, and it is not that easy to decipher which one is the best one. From our point of view, applying an Infinium type I/II bias correction is essential, as this bias seems to be the most critical one. We would recommend using PBC, BMIQ or the background correction methods developed by Triche and collaborators. Concerning between-array normalization, however, none of the methods available to date seem suitable to 450K data. Methods for batch effect removal, such as 'ComBat', can be used, even if the best way to avoid strong batch effects still resides in a proper experimental design.

Third, concerning differential methylation analyses, different approaches can be used. The single-probe approach is mainly based on statistics. Nevertheless, it is important to keep in mind that Infinium HumanMethylation450 is not suitable for detection of small differences of methylation because of its technical variability in measurements, and therefore, using an absolute methylation difference threshold is strongly recommended. Although this single-probe approach is the most commonly used, regional differential methylation analyses should not be neglected, as they can bring confidence in the results. Also, performing permutations can help to demonstrate the specificity of the results.

In conclusion, Infinium HumanMethylation450 is a nice tool to perform large-scale DNA methylation profiling, and it can be anticipated that its use will likely explode in the near future. Nevertheless, analyzing 450K data is more complex than initially thought, and data processing and interpretation need to be given particular consideration and care. We have summarized here different issues, which we feel as essential to take into consideration for accurate processing of 450K data. Further improvements in 450K data analyses, including benchmarking data sets and standardized preprocessing protocol, would be an important step towards the proper use of this innovative technology.

### Key Points

- Infinium HumanMethylation450 is a popular technology to study the DNA methylome in health and disease.
- Certain types of probes, such as cross-reactive probes and probes containing common SNPs, can generate artifactual data and need therefore to be filtered out.
- The main critical bias that needs to be corrected by within-array normalization is the Infinium type I/II bias.
- No between-array normalization method suitable for 450K arrays is available to date.
- The technical variability of the Infinium measurements should not be neglected and the use of an absolute methylation difference threshold ( $\Delta\beta$ ), in addition to statistical criteria, is strongly recommended.

### Acknowledgements

We wish to thank Ulrich Lehmann and colleagues for giving us access to their data.

### FUNDING

This work was supported by the Brussels Region 'BruBreast' project, the Belgian 'Télévie', the

‘Université Libre de Bruxelles’ (ULB), the ‘Fonds de la Recherche Scientifique’ (FNRS) and the ‘Interuniversity Attraction Poles’ (IAP P7/03).

## References

1. Jones PA. Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nat Rev Genet* 2012;**13**:484–92.
2. Bibikova M, Barnes B, Tsan C, *et al.* High density DNA methylation array with single CpG site resolution. *Genomics* 2011;**98**:288–95.
3. Dedeurwaerder S, Defrance M, Calonne E, *et al.* Evaluation of the Infinium Methylation 450K technology. *Epigenomics* 2011;**3**:771–84.
4. Rakyan VK, Down TA, Balding DJ, *et al.* Epigenome-wide association studies for common human diseases. *Nat Rev Genet* 2011;**12**:529–41.
5. Bibikova M, Le J, Barnes B, *et al.* Genome-wide DNA methylation profiling using Infinium(R) assay. *Epigenomics* 2009;**1**:177–200.
6. Price ME, Cotton AM, Lam LL, *et al.* Additional annotation enhances potential for biologically-relevant analysis of the Illumina Infinium HumanMethylation450 BeadChip array. *Epigenetics Chromatin* 2013;**6**:4.
7. Zhang X, Mu W, Zhang W. On the analysis of the illumina 450k array data: probes ambiguously mapped to the human genome. *Front Genet* 2012;**3**:73.
8. Chen YA, Lemire M, Choufani S, *et al.* Discovery of cross-reactive probes and polymorphic CpGs in the Illumina Infinium HumanMethylation450 microarray. *Epigenetics* 2013;**8**:203–9.
9. Roessler J, Ammerpohl O, Gutwein J, *et al.* Quantitative cross-validation and content analysis of the 450k DNA methylation array from Illumina, Inc. *BMC Res Notes* 2012;**5**:210.
10. Siegmund KD. Statistical approaches for the analysis of DNA methylation microarray data. *Hum Genet* 2011;**129**:585–95.
11. Wang D, Yan L, Hu Q, *et al.* IMA: an R package for high-throughput analysis of Illumina’s 450K Infinium methylation data. *Bioinformatics* 2012;**28**:729–30.
12. Du P, Kibbe WA, Lin SM. lumi: a pipeline for processing Illumina microarray. *Bioinformatics* 2008;**24**:15478.
13. Maksimovic J, Gordon L, Oshlack A. SWAN: Subset-quantile within array normalization for illumina infinium HumanMethylation450 BeadChips. *Genome Biol* 2012;**13**:R44.
14. Touleimat N, Tost J. Complete pipeline for Infinium((R)) Human Methylation 450K BeadChip data processing using subset quantile normalization for accurate DNA methylation estimation. *Epigenomics* 2012;**4**:325–41.
15. Teschendorff AE, Marabita F, Lechner M, *et al.* A beta-mixture quantile normalization method for correcting probe design bias in Illumina Infinium 450 k DNA methylation data. *Bioinformatics* 2012;**29**:189–96.
16. Pidsley R, Wong CC, Volta M, *et al.* A data-driven approach to preprocessing Illumina 450K methylation array data. *BMC Genomics* 2013;**14**:293.
17. Triche TJ, Jr, Weisenberger DJ, Van Den Berg D, *et al.* Low-level processing of Illumina Infinium DNA Methylation BeadArrays. *Nucleic Acids Res* 2013;**41**:e90.
18. Wessely F, Emes RD. Identification of DNA methylation biomarkers from Infinium arrays. *Front Genet* 2012;**3**:161.
19. Pan H, Chen L, Dogra S, *et al.* Measuring the methylome in clinical samples: improved processing of the Infinium Human Methylation450 BeadChip Array. *Epigenetics* 2012;**7**:1173–87.
20. Marabita F, Almgren M, Lindholm ME, *et al.* An evaluation of analysis pipelines for DNA methylation profiling using the Illumina HumanMethylation450 BeadChip platform. *Epigenetics* 2013;**8**:333–46.
21. Sun Z, Chai HS, Wu Y, *et al.* Batch effect correction for genome-wide methylation data with Illumina Infinium platform. *BMC Med Genomics* 2011;**4**:84.
22. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* 2007;**8**:118–27.
23. Leek JT, Scharpf RB, Bravo HC, *et al.* Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat Rev Genet* 2010;**11**:733–9.
24. Yan L, Ma C, Wang D, *et al.* OSAT: a tool for sample-to-batch allocations in genomics experiments. *BMC Genomics* 2012;**13**:689.
25. Dedeurwaerder S, Desmedt C, Calonne E, *et al.* DNA methylation profiling reveals a predominant immune component in breast cancers. *EMBO Mol Med* 2011;**3**:726–41.
26. Houseman EA, Accomando WP, Koestler DC, *et al.* DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC Bioinformatics* 2012;**13**:86.
27. Du P, Zhang X, Huang CC, *et al.* Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. *BMC Bioinformatics* 2010;**11**:587.
28. Hansen KD, Timp W, Bravo HC, *et al.* Increased methylation variation in epigenetic domains across cancer types. *Nat Genet* 2011;**43**:768–75.
29. Hansen KD, Langmead B, Irizarry RA. BSmooth: from whole genome bisulfite sequencing reads to differentially methylated regions. *Genome Biol* 2012;**13**:R83.
30. Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci USA* 2001;**98**:5116–21.