Complete pipeline for Infinium® Human Methylation 450K BeadChip data processing using subset quantile normalization for accurate DNA methylation estimation

Background: Huge progress has been made in the development of array- or sequencing-based technologies for DNA methylation analysis. The Illumina Infinium® Human Methylation 450K BeadChip (Illumina Inc., CA, USA) allows the simultaneous quantitative monitoring of more than 480,000 CpG positions, enabling large-scale epigenotyping studies. However, the assay combines two different assay chemistries, which may cause a bias in the analysis if all signals are merged as a unique source of methylation measurement. **Materials & methods:** We confirm in three 450K data sets that Infinium I signals are more stable and cover a wider dynamic range of methylation values than Infinium II signals. We evaluated the methylation profile of Infinium I and II probes obtained with different normalization protocols and compared these results with the methylation values of a subset of CpGs analyzed by pyrosequencing. **Results:** We developed a subset quantile normalization approach for the processing of 450K BeadChips. The Infinium I signals were used as 'anchors' to normalize Infinium II signals at the level of probe coverage categories. Our normalization approach outperformed alternative normalization or correction approaches in terms of bias correction and methylation signal estimation. We further implemented a complete preprocessing protocol that solves most of the issues currently raised by 450K array users. **Conclusion:** We developed a complete preprocessing pipeline for 450K BeadChip data using an original subset quantile normalization approach that performs both sample normalization and efficient Infinium I/II shift correction. The scripts, being freely available from the authors, will allow researchers to concentrate on the biological analysis of data, such as the identification of DNA methylation signatures.

Nizar Touleimat[1] & Jörg Tost*[1,2]

[1]Laboratory for Epigenetics, Centre National de Génotypage, CEA – Institute de Génomique, Bâtiment G2, 2 rue Gaston Crémieux, CP 5721, 91000 Evry, France
[2]Laboratory for Functional Genomics, Fondation Jean Dausset – Centre d'Etude du polymorphisme Humain (CEPH), 75010 Paris, France
*Author for correspondence:
Tel.: +33 01 60878423
Fax: +33 01 60878485
tost@cng.fr

The field of epigenetics and, in particular DNA methylation, has attracted much interest in recent years, as it has become clear that epigenetics plays a key role in normal development, as well as in disease, especially cancer [1]. DNA methylation occurring at CpG dinucleotides is probably the best-studied epigenetic modification due to the extensive mapping of DNA methylation in different cancers that has been performed over the last two decades. The interest in DNA methylation has been raised through multiple studies demonstrating the potential of DNA methylation-based biomarkers, which contain valuable information for early diagnosis of cancer, prognosis, tumor classification and may assist in the prediction of response to therapy [2,3]. A multitude of methods have been devised over the last 20 years for DNA methylation analysis, but only recently, improved technologies for genome-wide DNA methylation analysis, using microarray or second-generation sequencing-based approaches [4,5], now allow epigenome-wide association studies to be carried out, in analogy to the genome-wide association studies performed successfully in many complex diseases over the last decade [6]. These technologies analyze a large number of CpGs with high quantitative resolution in parallel, and have extended the potential of large-scale DNA methylation mapping to autoimmune diseases, such as lupus erythematosus [7] and complex disease including Type 1 [8] and 2 diabetes [9], for which DNA methylation differences are much more subtle than the changes observed in tumorigenesis.

Epigenotyping technologies such as the Infinium® Human Methylation 27K BeadChip (Illumina Inc., CA, USA) generate a methylation-state-specific 'pseudo-SNP' through bisulfite conversion; therefore, translating differences in the DNA methylation patterns into sequence differences that can be analyzed using quantitative genotyping methods [10]. These arrays have the advantage that the probes can be designed for CpG-poor regions, such as CpG island shores and shelves, which are more prone to alter their DNA-methylation status in response to environmental exposure or

Future Medicine part of fsg

during carcinogenesis [11]. Furthermore, they do not depend on restriction enzyme recognition sites and require only a relatively low input of DNA. These arrays have the potential for large-scale high-throughput studies as they use highly standardized protocols that can be implemented with a large degree of automation into existing genotyping pipelines. The analysis of the results is relatively straightforward and does not require correction for CpG density or cost- and time-intensive bioinformatic calculations when compared with sequencing-based DNA methylation analysis. These epigenotyping assays have found to be accurate and display high correlation to reduced representation bisulfite sequencing [12], as well as locus-specific quantitative assays such as MethyLight [13]. However, these arrays analyze only a small number of the 28 million CpG sites of the human genome and, as in all bisulfite-based analysis techniques, they are not able to differentiate between cytosine methylation and hydroxymethylation. Further, the low sequence complexity of bisulfite-treated DNA could induce cross-hybridization events. In addition, there are currently no commercial arrays for the analysis of nonhuman samples available. The newest release, the Illumina Infinium Human Methylation 450K BeadChip (Illumina Inc.), dramatically expands the genome coverage, analyzing more than 480,000 CpG sites, covering 99% of all RefSeq genes with an average of 17 probes per gene. The array is not only focused on CpG islands, but distributed over various functional elements, including, but not limited to, CpG islands shores and shelves, 3′- and 5′-UTRs, gene bodies, DNAse hypersensitive sites, miRNA promoters and other ncRNAs. It provides, for the first time, a good compromise between coverage, throughput, cost, resolution and accuracy permitting genome-wide epigenome analysis by epigenotyping.

The Infinium 450K BeadChip consists of two different chemical assays carrying information about the methylation state in the type of bead for the Infinium I (InfI) probes where the base extension is the same for methylated or unmethylated alleles (i.e., the fluorescent signal does not carry any information on the methylation status). The Infinium II (InfII) probes are attached to a single type of bead and the methylation information is obtained through dual channel single-nucleotide primer extension with labeled dideoxynucleotides on the methylation variable position of a CpG.

The 450K BeadChip, released in spring 2011, has been evaluated in three recent publications [14–16]. The very first publication showed a very high correlation between data obtained on the 27K BeadChip and the corresponding probes on the 450K, of which 90% are present in the novel design, as well as data previously obtained with locus-specific DNA methylation analysis demonstrating the technical accuracy and the potential of the technology for epigenetic studies [14]. However, they did not specifically investigate differences between InfI and InfII probes or provide any suggestions on how to analyze and process the data beyond the software provided by the manufacturer. Illumina confirmed in a second publication, the high reproducibility between technical replicates ($R^2 > 0.992$) and good correlation to 27K BeadChip data ($R^2 > 0.95$), as well as whole-genome bisulfite sequencing data ($R^2 > 0.95–0.96$) [15]. A shift in the density curves between InfI and InfII probes is shown in the paper, but not commented on. In a very recent publication, Dedeurwaerder and colleagues confirmed again the very high quality of the data obtained with the 450K BeadChip, but also showed that the two chemistries (InfI and InfII) have a different dynamic behavior in the assay and that some processing of the data is required to make it comparable [16]. They observed that InfI signals are more stable and have an extended dynamic range of methylation values compared with InfII signals. They devised a relatively simple-to-implement, peak-based approach, which rescales the density distribution of the InfII probes on the basis of InfI density distribution modes. This approach efficiently corrects the InfI/InfII shift and improved the quantitative estimation of the methylation status [16]. The peak-based correction approach was also implemented in a very recently published R package named Illumina Methylation Analyzer (IMA), proposing an analysis pipeline for Illumina's 450K Infinium methylation data, including data preprocessing, differential statistical analysis, annotation of sites/regions of interest and adjustment for confounding factors [17]. However, this correction approach is based on a strong assumption about the bimodal shape of the methylation density profiles and when implementing the algorithm we found its efficiency to be sensitive to variations in the shape of the methylation density curves.

From our point of view, the normalization and shift correction is therefore not yet optimally solved. We also identified additional points that may interfere with the quality of the methylation data produced with the Illumina Infinium Human Methylation 450K BeadChip.

The crucial steps are: data quality control to estimate the quality of a data set after data extraction; probe filtering to eliminate signal variation unrelated to DNA methylation differences or unrelated to the biological context of the study; signal correction for the adjustment of the color balance and background level correction; as well as the InfI/InfII shift correction and between sample normalization.

As, to our knowledge, neither IMA, lumi nor any other tool provides a complete solution for all of these four preprocessing steps, we developed a new pipeline and evaluated the quality of the resulting data by comparing different variants of our pipeline. As reference DNA methylation technology, we compared the results of the different normalization approaches to the results obtained by pyrosequencing, which may be considered as the current 'gold standard' for DNA methylation analysis at single-nucleotide resolution.

Materials & methods
■ Data set descriptions
We used three independent data sets to evaluate our preprocessing pipeline. The analysis for the first data set is found within this article, while figures for the two additional data sets are found in the SUPPLEMENTARY MATERIAL published with this report (see www.futuremedicine.com/doi/suppl/10.2217/epi.12.21).

The first data set contains 42 paired samples divided into 21 tumoral and 21 nontumoral samples taken from the same solid organ. The second and third data sets contain 16 and 19 unpaired samples divided in four nontumoral versus 12 tumoral and six nontumoral versus 13 tumoral, respectively, all samples were taken from the same solid organ.

■ Illumina Methylation 450K data production
A sample of 1 µg of DNA was bisulfite treated using the EpiTect® 96 Bisulfite Kit (Qiagen GmbH, Hilden, Germany), and 200 ng of bisulfite treated DNA was analyzed using the Infinium Human Methylation 450K BeadChips. The samples were processed according to the manufacturer's protocol at the genotyping facility of the Centre National de Génotypage (Evry, France) without any modification to the protocol.

■ Data extraction
We used the GenomeStudio® software (version 2011.1; Illumina Inc.) for the extraction of DNA methylation signals from scanned arrays (methylation module version 1.9.0, Illumina

Inc.). Methylation data were extracted as raw signals with no background subtraction or data normalization. The obtained 'β' values – that is, the methylation scores for each CpG range from 0 (unmethylated, U) to 1 (fully methylated, M) on a continuous scale, and are calculated from the intensity of the M and U alleles as the ratio of fluorescent signals:

$$\beta = \frac{Max(M,0)}{Max(M,0) + Max(U,0) + 100}$$

■ Data processing
All preprocessing, correction and normalization steps were either implemented by us in R (version 2.14.0) with Bioconductor packages (version 2.9) or adapted from the methylumi (version 2.0.1, [101]) and lumi (version 2.6.0, [18]) R packages.

All plots were produced using R functions. Density plots for β-values and standard deviation (SD) of β-values were computed with the density function (kernel density estimation with a Gaussian smoothing function and a bandwidth of 0.05 for β-values and 0.001 for SDs of β-values) to obtain the density plots.

The pipeline is currently a collection of R scripts launched by a 'main' script that sets all local paths and variables. As prerequisites, the latest versions of the methylumi and lumi R packages have to be installed in the R environment. To run the pipeline, we recommend a 64 bit operating system with at least 8 GB of RAM. On such a configuration the runtime to process the whole pipeline for a data set of 42 samples is about 15 min.

■ Differential methylation analysis
Prior to the identification of differentially methylated probes, we applied a nonspecific filtering step to the preprocessing data. This step consists of removing 80% of the least varying probes based on their interquartile range, and with no regard to sample groups. The combination of a preliminary nonspecific filtering with a statistical test has, in the context of gene-expression analysis, been shown to increase discovery rate, while maintaining control over type I errors, thus resulting in a significant increase in detection power [19]. We assumed that a similar benefit can be obtained in the context of DNA methylation analysis. The threshold was determined empirically, demonstrating that a reduction of 80% of the features does not have a significant effect on further univariate statistical tests, but offers a significant gain in time for the identification of differentially methylated genes.

It should be noted that this step is optional and the user can of course use the entire data set for the identification.

We consider a probe as differentially methylated if the absolute value of the difference between robust β-values medians in samples of each phenotypes is higher than 0.2:

$$median(\{\beta_1^{NT}, \beta_2^{NT}, \ldots, \beta_n^{NT}\}) - median(\{\beta_1^{T}, \beta_2^{T}, \ldots, \beta_n^{T}\}) | \geq 0.2$$

where β_i^{NT} and β_i^{T} corresponds to β-values in paired nontumoral and tumoral samples, both associated to detection p-values lower than 0.01. This 0.2 threshold, representing approximately a difference in DNA methylation levels of 20%, corresponds to the recommended difference between samples analyzed with the Illumina methylation Infinium technology that can be detected with 99% confidence [10,15]. From the list of differentially methylated probes, the corresponding gene identities are obtained and to facilitate further analysis a list of differentially methylated genes is presented. However, it is of course possible to work on the level of differentially methylated probes.

Selection of probes for validation with pyrosequencing

To evaluate and validate the results of the different variants of our preprocessing pipeline, we selected a subset of probes for comparison of their methylation status to measurements obtained by pyrosequencing [20]. To be selected, a probe had to match the following criteria after each preprocessing variant: stable methylation values between samples of the same phenotype (β SD < 0.1); differentially methylated (differential methylation > 20%) between samples of different phenotypes; and most importantly large difference between median β-values obtained with each variant of our preprocessing pipeline. The top 13 probes matching these three criteria were selected for validation.

Pyrosequencing

Quantitative DNA methylation analysis was performed by pyrosequencing of bisulfite-treated DNA [20]. CpGs for validation were amplified using 30 ng of bisulfite treated human genomic DNA and 5–7.5 pmol of forward and reverse primer, one of them being biotinylated. Sequences for oligonucleotides for PCR amplification and pyrosequencing are available upon request. Reaction conditions were 1 × HotStar® Taq buffer (Qiagen) supplemented with 1.6 mM MgCl$_2$, 100 µM dNTPs and 2.0 U HotStar

Taq polymerase (Qiagen) in a 25 µl volume. The PCR program consisted of a denaturing step of 15 min at 95°C, followed by 50 cycles of 30 s at 95°C, 30 s at the respective annealing temperature and 20 s at 72°C, with a final extension of 5 min at 72°C. A total of 10 µl of PCR product were rendered single-stranded as previously described [20] and 4 pmol of the respective sequencing primer were used for analysis. Quantitative DNA methylation analysis was carried out on a PSQ 96MD system with the PyroGold SQA Reagent Kit (Qiagen) and results were analyzed using the Q-CpG™ software (V.1.0.9, Biotage, Uppsala, Sweden).

Results

InfI/InfII variability & signal distribution

A shift between the distribution of the β-values of the probes deriving from the two different assay chemistries (InfI and InfII) has previously been reported [16]. To confirm this observation in our data, we analyzed the raw β-values in three data sets composed of two categories of samples: tumoral and nontumoral samples. Figure 1A, Supplementary Figures 1A & 2A show the density plots of the β-values for InfI and InfII probes for each group of samples. All samples exhibit two peaks in their density profile; one peak corresponding to low or unmethylated probes with a β-value close to 0 and the second one corresponding to highly or fully methylated probes with a β-value close to 1, confirming the bimodal shape of the distribution of β-values. Peaks associated with InfII probes were found to have a decreased quantitative dynamic range corresponding to less extreme values. Using a boxplot representation to investigate in more detail the distribution of InfI and InfII probe signals (Figure 1C, Supplementary Figures 1C & 2C), the assay chemistries clearly differed with regard to the medians and interquartile range of the β-values. InfII probes displayed a median β-value close to 0.7, while the median for the InfI probes was closer to 0.1. The third quartiles of InfI and InfII probes were found to be similar, whereas the first quartiles of InfI probes corresponded to a smaller value (approximately β = 0.05) than InfII probes (approximately β = 0.2). InfI probes can therefore report for a wider range of β-values, reflecting all possible methylation states, while InfII probes may not be able to report for completely methylated or unmethylated CpGs with the same sensitivity as InfI probes.

Furthermore, the medians of InfII probes for a given sample display more variation between
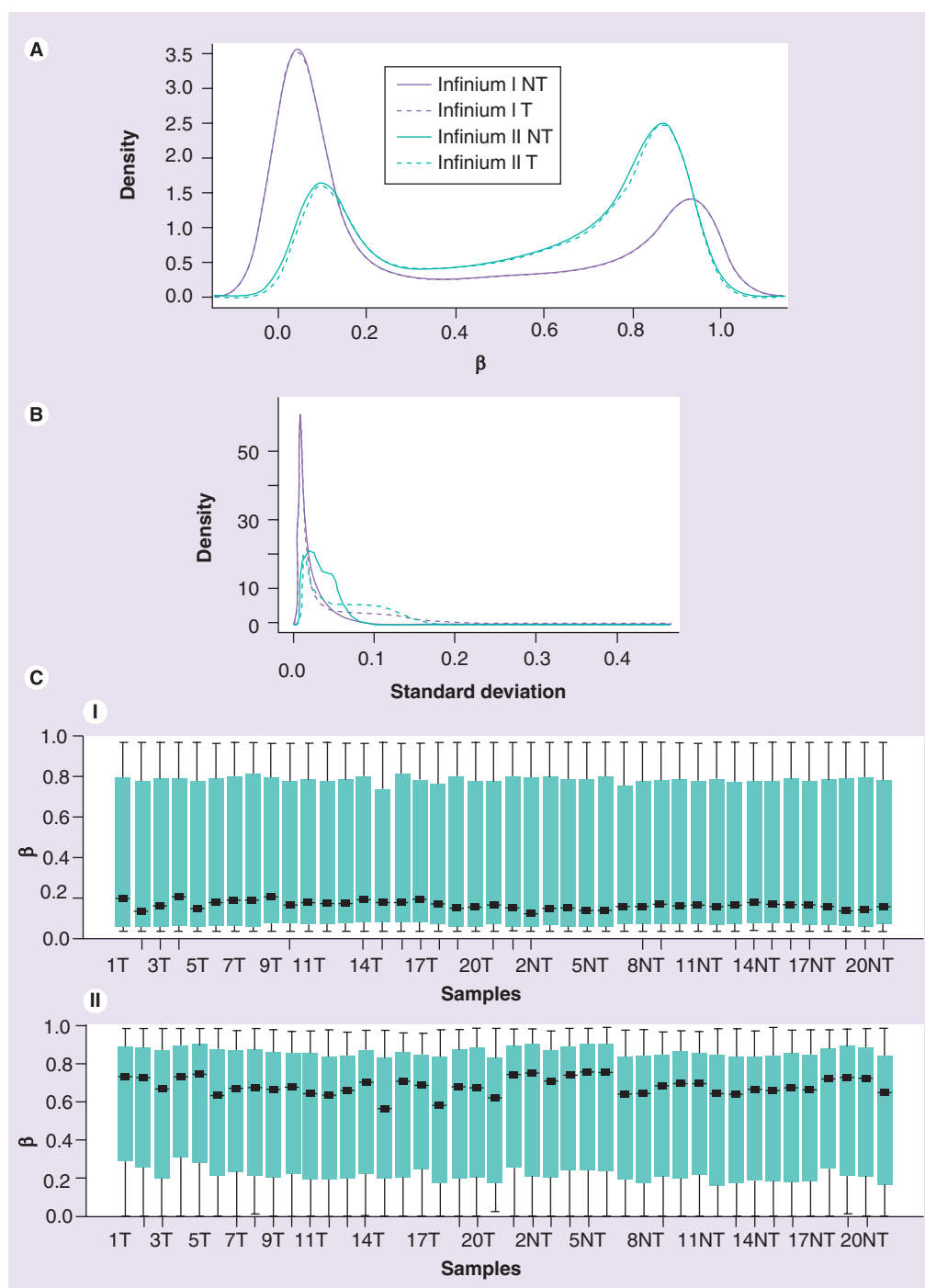
Figure 1. Infinium I/Infinium II (Illumina Inc., CA, USA) signal variability and bias characterization for data set 1. (A) Density plots of tumoral (dashed) and nontumoral (continuous) samples median profiles for Infinium I (purple) and Infinium II (green) probes. Bandwidth = 0.05. **(B)** Density plots of standard deviations of β-values computed for each sample group. Bandwidth = 0.001. **(C)** Boxplots of the distribution of β-values for each sample for **(CI)** Infinium I and **(CII)** Infinium II probes.
NT: Density plot of the median β-value profile for nontumoral samples; T: Density plot of the median β-value profile for tumoral samples.

samples than the medians of InfI probes. For each category of Infinium probes, we therefore plotted the density of β-value SDs between samples of the same phenotype (FIGURE 1B, SUPPLEMENTARY FIGURES 1B & 2B). The density plots of β-value SDs for InfI probes show a unique peak, close to 0, reflecting the low variability of β-values between samples of the same phenotype. By comparison,

the SDs of the β-values for InfII probes show a smaller peak for low SDs, and have a more spread distribution for higher SDs values, supporting the observation that methylation signals provided by InfI probes are more stable and reproducible across different samples than the signals provided by InfII probes. This difference in signal stability is again independent of the phenotype of the sample, but rather related to the assay chemistry.

We thus confirm the previously observed shift between InfI and InfII probe signals, which may lead to a bias in the analysis of the methylation status requiring a rescaling and normalization step for InfII signals. Therefore, we investigated different approaches to perform a robust and efficient normalization or rescaling that corrects the shift between the two types of probes, and that provides the most robust InfII signals.

■ A subset quantile normalization approach to correct InfI/InfII shift

The identified shift between InfI and InfII β-values may induce a bias in the analysis if the methylation signals corresponding to the two types of assays are analyzed together. This would correspond to comparing probes from two different technologies with different signal distributions using a single statistical approach. InfI and InfII probes can be regarded as two different subarrays, and their β-values distributions should be normalized using standard approaches for intersample normalization, such as quantile normalization. However, three constraints prohibit such a straightforward approach: the number of InfI (28%) and InfII (72%) probes differ and prevent from computing a common set of reference quantiles; the population to 'correct' (InfII) is the larger one and may therefore bias the distribution of the other population (InfI); and there is a large imbalance in the proportions of InfI and InfII probes covering the different CpG and gene-sequence regions. As the variation of the methylation status may be specific for probes covering different subcategories of CpG or gene-sequence regions, a global standardization of methylation values distributions may lead to a dramatic loss of information. In the context of array-based gene expression analysis, a subset quantile normalization (SQN) procedure has been proposed to solve the two first issues, normalizing the gene-expression signal with the help of control probes [102]. In this approach, the reference quantiles of a target set of features are estimated from the smaller set of features used as 'anchors' that are considered to be more

reliable and stable. This approach has previously been adopted by Aryee *et al.* [21] in the context of DNA methylation for the normalization of McrBC-based unmethylated DNA enrichment arrays (CHARM microarray [22]). In the current work, we apply the concept of 'anchor probes' to Illumina Infinium Human Methylation 450K BeadChip data for the correction of the shift between InfI and InfII signals. It should be pointed out, that the implementation in this article is different from the previously described approaches. In brief, our SQN approach modifies the values of the target distribution based on rank equivalence, while the previously published approach modified the target distribution based on the value equivalence. In these approaches, the raw intensity value of any probe on the array equals the qth quantile of the anchor probes on the same array, its normalized intensity is thus defined as the qth quantile from the reference distribution resulting in the fact that, for example, the 95th percentile of nonanchor probes will not have the same value as the 95th percentile of anchor probes. In our approach, we correct the data so that nonanchor and anchor probes of the same percentiles will have the same value.

Owing to the improved stability and dynamic range we decided to use InfI signals as the anchors to estimate a reference distribution of quantiles and to use this reference to estimate a target distribution of quantiles for InfII probes as a means to provide an accurate normalization of InfI/InfII probes and correct for the shift (FIGURE 2).

Illumina provides two types of annotations for the probes: 'relation to CpG' and 'relation to gene sequence', respectively describing which kind of CpG environment or gene region a probe covers. The InfI and InfII probes are represented in different proportions in these sequence categories. These different annotation categories may behave differently in their DNA methylation states, and the imbalance of their proportions of InfI and InfII probes will not allow a direct extrapolation of the distribution of InfI signals to InfII signals. We decided, therefore, to specify as many sets of anchors as there exists probe categories, and to estimate a target set of reference quantiles per probe category for InfII. To test the most informative and efficient way of building probe categories, we implemented two versions of our SQN approach. The first is based on the 'relation to CpG' annotation (S shore, S shelf, N shore, N shelf and distant) and the other based on the 'relation to gene sequence' annotation (body, 5′-UTR, 3′-UTR, 1st exon,

intergenic, multiple annotations, TSS200 and TSS1500). In order to obtain robust 'anchors', the computation of the reference quantiles was based only on InfI probes associated with significant detection p-values only (detection p-values < 0.01 in our study).

As SQN performs at the level of sample batch, this correction approach will also provide a robust between-sample normalization, enabling further comparisons between samples.

## ■ Description of our preprocessing pipeline

In order to obtain the most accurate and robust DNA methylation data, we implemented a complete preprocessing pipeline for the Illumina 450K methylation array. This pipeline is composed of four main steps (FIGURE 3) starting from raw extracted data and resulting in 'ready to analyze' methylation information. We performed a comprehensive investigation on all types of possible bias or issues associated with the 450K data and implemented a solution to account for all of them. Steps 1–3 are based on existing methods, but have never been associated as a unique and complete preprocessing pipeline. However, the last step of the pipeline, the SQN approach, based on a functional annotation system is an original approach, which has, to our knowledge, not been used in the context of methylation data analysis.

Each step of the pipeline is described in detail in the following section and applied to the three data sets.
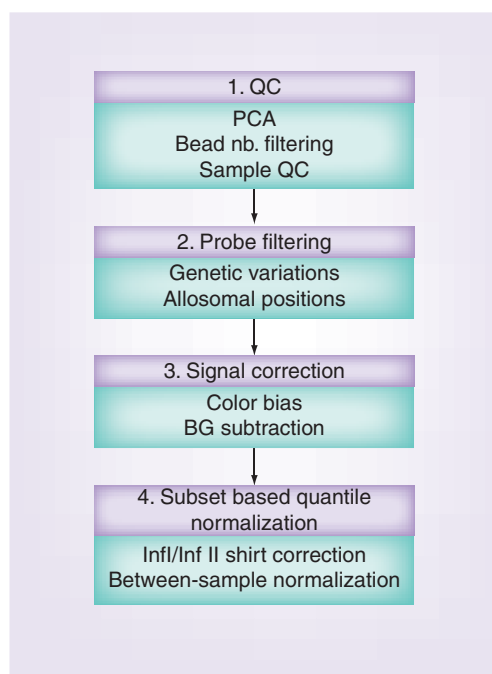
Quality control

This step estimates the quality of a data set and selects reliable probes and samples. A principal component analysis is performed to detect potential batch effects when samples are processed on more than one array. For data set 1, which was processed in parallel on four BeadChips, a small batch effect was detected, which did not abolish the differentiation between nontumoral and tumoral samples as first principal component axis (SUPPLEMENTARY FIGURE 3A). The preprocessing pipeline described below did completely correct for this batch effect, as can clearly be seen by the clustering of the nontumoral samples (SUPPLEMENTARY FIGURE 3B). As we have so far not detected any major batch effect using the completely automated genotyping facility of the Centre National de Génotypage, our pipeline does currently not propose any additional correction approach if such a bias is detected.

Following the recommendation from the manufacturer, both methylated and unmethylated
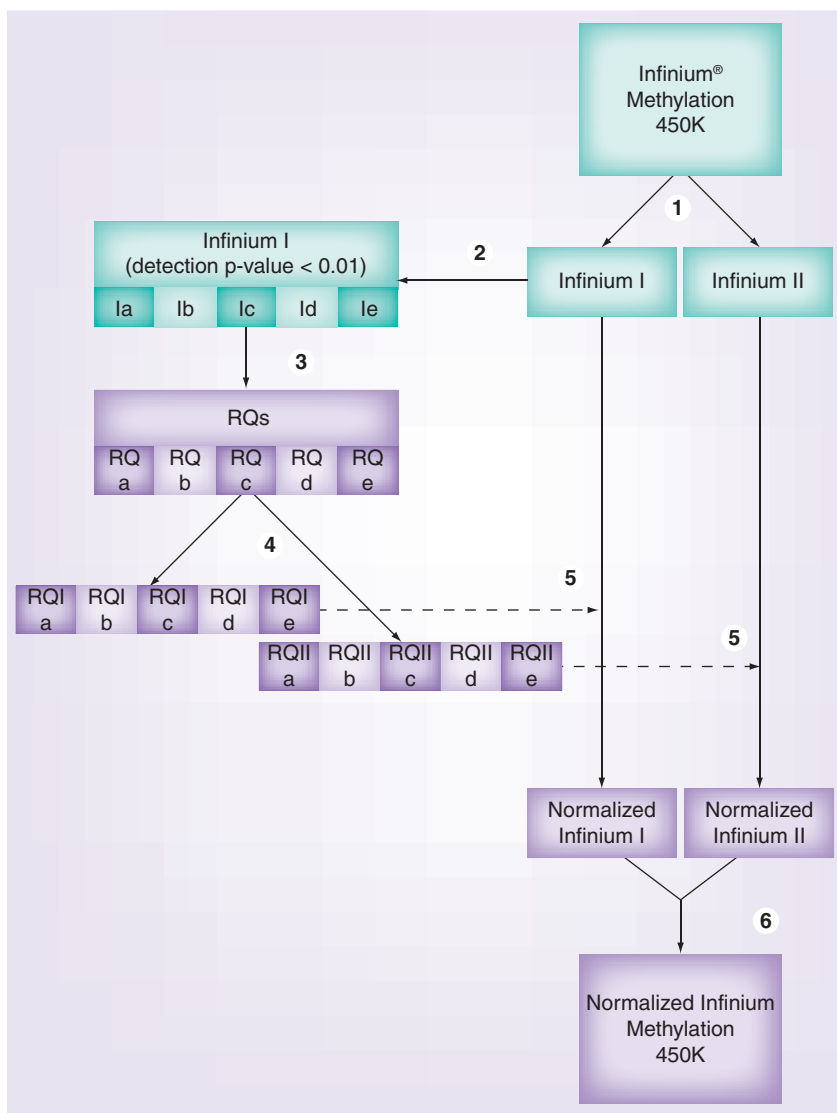
Figure 2. Schematic illustration of our subset quantile normalization approach. Step 1: β-values associated to Infinium I (InfI) and Infinium II (InfII) probes are separated. Step 2: β-values from InfI probes associated with detection p-values higher than a given threshold (e.g., p = 0.01) are removed. Step 3: robust reference quantiles, 'anchors', are computed from the filtered β-values.
Step 4: target reference quantiles are estimated from the anchor reference quantiles to match InfI and InfII probe numbers. Step 5: InfI and InfII associated β-values are normalized with their respective reference quantiles. Step 6: the two normalized data subsets are associated to rebuild a normalized Illumina® (CA, USA) Methylation 450K data set.
RQ: Reference quantile.

signals associated to a probe have to be summarized from at least three functional beads on the array. Rather than eliminating probes for which the unmethylated and methylated signals do not match these criteria, a β-value for these probes is computed but associated with a detection p-value equal to 1. Then the percentage of probes associated with a detection p-value lower than a given threshold is measured for each sample. Samples with less than 90% of high-confidence probes should be removed from further analysis. For

Figure 3. Global scheme of the preprocessing pipeline for Illumina's (CA, USA) Infinium® Human Methylation 450K BeadChip.
BG: Background; Inf: Infinium; nb.: Number; PCA: Principal component analysis; QC: Quality control.

our study, in data sets 1 and 2 all samples were composed of at least 99% of accurate probe signals. For data set 3, 17 samples were composed of at least 99% of accurate probe signals and two samples had only 42 and 61% of accurate probe signals, and were therefore eliminated from the analysis.

Probe filtering

This step does not influence the quality of the signal but may be important depending of the biological context on the study. The first filtering step consists of removing probes, for which the observed variation of the methylation level might be due to genetic variations. Thus, we obtained from the HapMap project [103] the list of all frequent SNPs in the Caucasian population and we removed all probes that potentially contain or extend on one of these SNPs from the 450K methylation data. The minor allele frequency can be chosen depending on cohort size and the desired level of stringency. For our data sets all probes potentially confounded by SNPs with a minor allele frequency of at least 5% were removed. In some studies, there is no information about the sex of the patients and/or patients have not been randomized according to sex. Therefore, variation of DNA methylation levels

for probes interrogating allosomal positions may be due to sex bias in the analyzed sample sets. The user, therefore, has the possibility to remove all probes associated with an allosomal position. However, this step is optional, as for example, in experiments analyzing the very same sample, such as a cell line under different conditions, this step would unnecessarily reduce the number of usable data points.

Signal correction

The Infinium Human Methylation 450 BeadChip is a two-color assay. Owing to the difference in labeling efficiency and scanning properties of the two color channels, the intensities measured in the two color channels might be imbalanced. Furthermore, a color balance adjustment may be important if the color effect is inconsistent across samples. The R package lumi [18] provides a function for color-bias adjustment based on a smooth quantile normalization method, developed first for the 27K version of Infinium Human Methylation BeadChip, and recently adapted to the 450K version of this array.

The lumi package also provides a function for correction of the background level. This function provides two different methods for the estimation of the background. The first one is based on the negative-control probes present on the BeadChip, while the second one is based on the direct estimation of the background from the density modes of the intensities measured by probes interrogating methylation positions. We tested both approaches, and found that the second approach produced, in some instances, aberrant DNA methylation profiles (data not shown). This is probably due to the existence of more than two density modes for some samples on the 450K array. Furthermore, this approach makes a strong assumption about the shape of methylation signal density profile. We therefore implemented the first background estimation approach based on the negative control probes provided by Illumina.

These two correction functions of the lumi package were included into our pipeline and applied to our three data sets.

Robust SQN

We implemented in our pipeline the SQN-based correction approach previously described in this article. This approach performs also a between sample normalization allowing further comparisons of samples.

The entire pipeline is freely available from the authors upon request.

■ Comparison of InfI/InfII shift correction variants

In order to assess the performance of our SQN method, we decided to compare it with the previously published 'peak-based correction' approach [16] and some variants our own approach. We also evaluated the effect of a classic quantile normalization on the InfI/InfII shift using the lumi [18] package function.

We compared the following seven approaches:

- Filtered and corrected β-values followed by no normalization of the β-values at all, this approach uses only steps 1–3 of our pipeline;

- Filtered and corrected β-values further processed with standard sample quantile normalization, this process can be considered as the preprocessing pipeline implemented in the lumi package;

- Filtered and corrected β-values using the peak-based correction approach for InfI/InfII shift correction proposed by Dedeurwaerder et al. [16]. Based on the observation that values from InfII probes were less accurate and reproducible than those obtained from InfI probes, this approach uses the peak summits of the density curves corresponding to the unmethylated status and fully methylated status identified from InfI β-value densities to rescale InfII β-value. With this approach, InfII probe signals match the same values for methylated and unmethylated status and cover the same dynamic range compared with the signals from InfI probes;

- Filtered and corrected β-values using the above described peak-based approach for InfI/InfII shift correction and followed by quantile normalization;

- Filtered and corrected β-values using a global SQN approach, in this variant of our correction strategy, a unique set of reference quantiles is computed from InfI probes and used as anchors for the separate normalization of InfI and correction and normalization of InfII probes;

- Filtered and corrected β-values using separate SQN for each category of CpG-related probes provided by Illumina, in this approach, a different set of robust reference quantiles is computed for each one of the six probe categories, defined by their relative position to the CpG context. Each set of anchors is then used to normalize its respective category of InfI and InfII probes;

- Filtered and corrected β-values using separate SQN for gene sequence-related probes categories provided by Illumina, similar to the above described approach, but probe categories are defined by the relative position of a probe to the nearest gene sequence.

■ Influence of normalization/correction variants on the correction of the InfI/InfII shift

FIGURE 4 shows the effect of the five correction and/or normalization approaches on the density profiles of the median β-values for InfI and InfII probes. We consider the results of steps 1–3 of our pipeline, with no normalization, as the reference situation (FIGURE 4A, SUPPLEMENTARY FIGURES 4A & 5A). As expected, the classical quantile normalization of β-values for InfI and InfII probes does not affect the shift between the two types of Infinium probes. Global SQN (FIGURE 4D, SUPPLEMENTARY FIGURES 4D & 5D) and categorical SQN using the gene sequence-related probe categories (FIGURES 4F, SUPPLEMENTARY FIGURES 4F & 5F) seem to correct, too strongly, the shifted DNA methylation profile as InfII profiles overlay the InfI profiles leading to an overestimation of the unmethylated status and an underestimation of the fully methylated status. Categorical SQN using the CpG-related probe categories efficiently corrects the shift, with no influence on InfII signal distribution in all data sets (FIGURE 4E, SUPPLEMENTARY FIGURES 4E & 5E). The peak-based approach performs an efficient correction for data set 1 (FIGURE 4C), but corrects only the peak corresponding to low or unmethylated probes in data sets 2 and 3 (SUPPLEMENTARY FIGURES 4C & 5C). To investigate possible reasons for this lack of correction for the peak-based approach, the β-value densities for individual samples were monitored before and after the peak-based correction for the three data sets noticing that some samples of each data set displayed a 'damaged' profile (SUPPLEMENTARY FIGURES 6B, 7B & 8B). The detailed analysis of the β-value densities of these samples revealed that InfI- and InfII-associated density profiles were slightly different from the expected bimodal shape, with an additional smaller peak for highly methylated probes, providing evidence that the peak-based correction approach may be sensitive to variations in the shape of the β-value density curve. The peak-based correction is performed on sample level, whereas the SQN approach is performed at sample batch level, as in classical quantile normalization. The sensitivity of the peak-based correction approach to variations in the shape of the density curve may be partially corrected by performing a sample quantile normalization after InfI/InfII
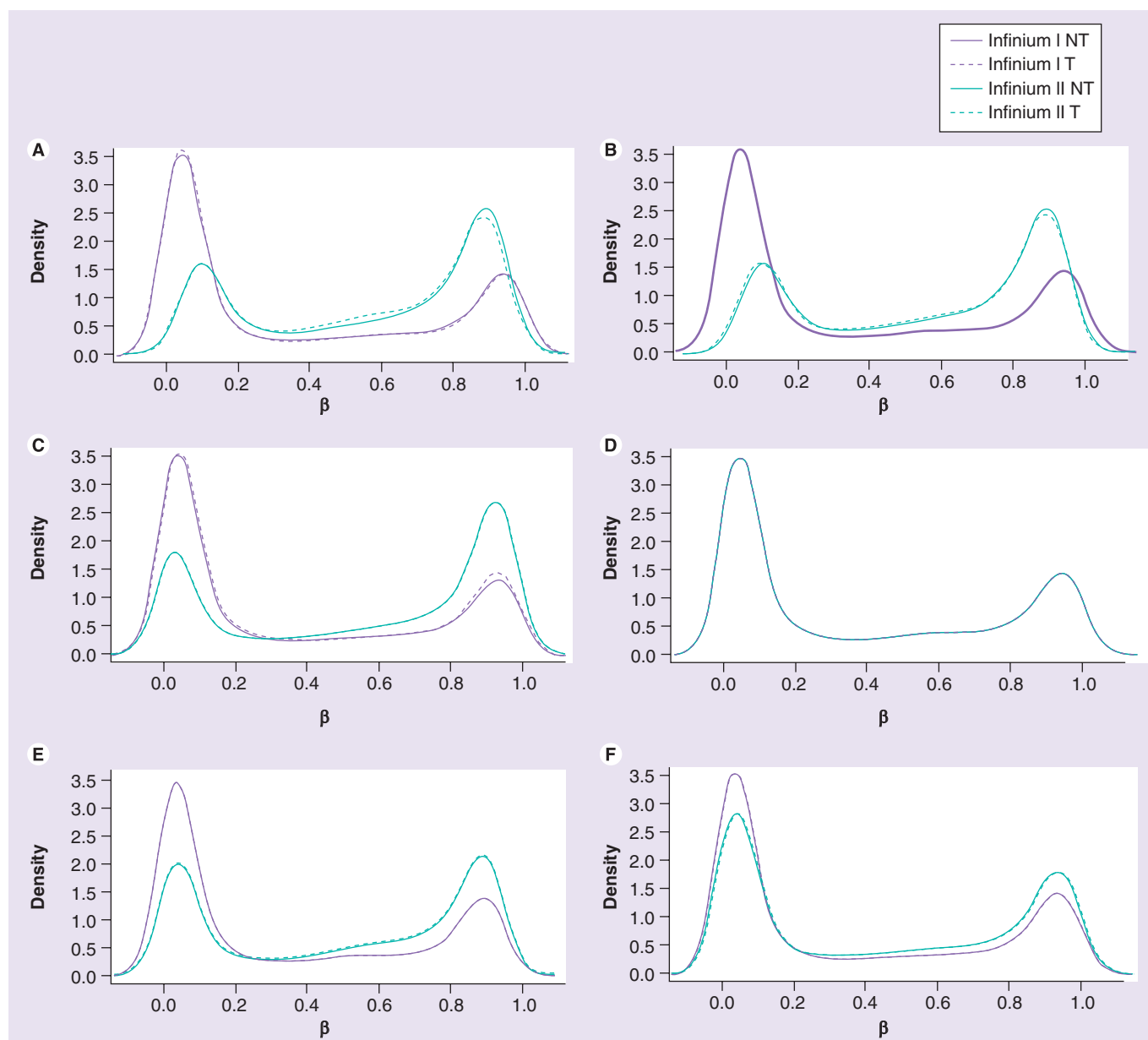
Figure 4. Density plots of β-values after six variants of the fourth step of our preprocessing pipeline. All variants include steps 1–3 of our preprocessing pipeline. **(A)** No normalization. **(B)** Lumi-based classical quantile normalization. **(C)** Peak-based correction followed by quantile normalization. **(D)** Subset quantile normalization with a unique set of reference quantiles computed from Infinium I signals. **(E)** Subset quantile normalization with a reference quantiles set computed from Infinium I signals for each kind of probe category according to the 'relation to CpG' annotations provided by Illumina (CA, USA). **(F)** Subset quantile normalization with a reference quantiles set computed from Infinium I signals for each kind of probe category according to the 'relation to gene sequence' annotations provided by Illumina. Bandwidth = 0.05.
NT: Density plot of the median β-value profile for nontumoral samples; T: Density plot of the median β-value profile for tumoral samples.

shift correction in order to standardize all sample distributions (**SUPPLEMENTARY FIGURES 6C, 7C & 8C**). However, the number of samples whose methylation profile has been 'damaged' after InfI/InfII shift correction may limit this rescue.

■ Methylation density profiles per probe category

The averaged density of the methylation profile was monitored for InfI and InfII probes associated with each probe category for six variants of our pipeline. **FIGURES 5, SUPPLEMENTARY FIGURES 9 & 10** show plots for probes in the vicinity of CpG islands and N shores for the three data sets. As expected, a shift between InfI and InfII signals for all probe categories was observed in the non-normalized signals (panel A), indicating that the signal bias between InfI and InfII technologies is not related to the disequilibrium of the proportions, in which each type of probe

is represented in the different probe categories. The classical quantile normalization approach (panel B) had little effect on the InfI/InfII bias, but owing to the imbalance between the number of probes from of the two Infinium assays, pulls the InfI profile towards the InfII profile. On the contrary, the three SQN approaches corrected perfectly the InfI/InfII shift and moved the
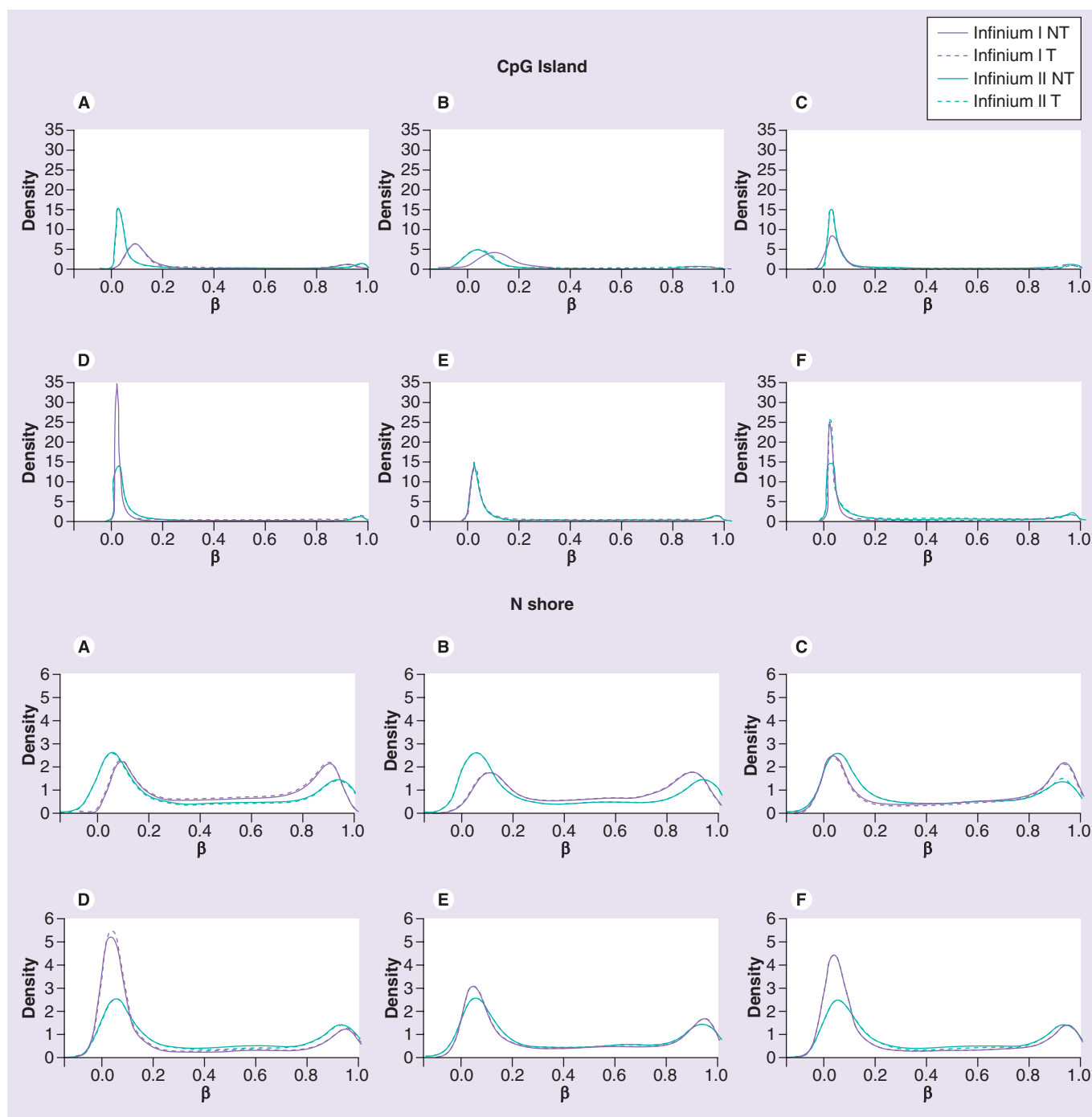
Figure 5. Density distribution of DNA methylation (β-)values for CpG-island and N-Shore annotated probes following six variants of our preprocessing pipeline. All variants include steps 1–3 of our preprocessing pipeline. **(A)** No normalization. **(B)** Lumi-based classical quantile normalization. **(C)** Peak-based correction followed by quantile normalization. **(D)** Subset quantile normalization with a unique set of reference quantiles computed from Infinium I signals. **(E)** Subset quantile normalization with a reference quantiles set computed from Infinium I signals for each probe category according to the 'relation to CpG' annotations provided by Illumina (CA, USA). **(F)** Subset quantile normalization with a reference quantiles set computed from Infinium I signals for each probe category according to the 'relation to gene sequence' annotations provided by Illumina. Bandwidth = 0.05.
NT: Density plot of the median β-value profile for nontumoral samples; T: Density plot of the median β-value profile for tumoral samples.

InfII peaks to the same 'extreme' methylation value of InfI peaks. However, the global SQN approach using all InfI probes as a unique set of anchors or the SQN approach using the relation to gene-sequence annotations to compute different InfI anchor sets tended to overestimate low methylation signals for normalized InfII probes. The SQN approach using the 'relation to CpG' annotations to identify category related 'anchors' perfectly corrected the bias, while retaining InfII signal distribution profiles for the CpG island and N-shore-related probes in all three data sets (FIGURE 5, SUPPLEMENTARY FIGURES 9 & 10). We further confirmed the sensitivity of peak-based correction to the shape of the methylation density profiles, leading only to the efficient correction of the InfI/InfII bias for the first methylation density peak. In addition, we monitored the effect of these six variants of the normalization pipeline in the remaining probe categories defined according to their relation to CpGs (S shore, S shelf, N shore, N shelf and distant) or according to their relative location to the closest gene (body, 5′-UTR, 3′-UTR, 1st exon, intergenic, multiple annotations, TSS200 and TSS1500). All the observations made for the CpG island and N-shore categories

were also valid for all other probe categories, demonstrating the improved performance of CpG-related annotations compared with the gene sequence-related annotation.

■ Comparison of normalized data to pyrosequencing-based DNA methylation values

Pyrosequencing can be considered as one of the gold-standard technologies for DNA methylation analysis owing to its high quantitative precision and its ability to provide data with single nucleotide resolution. It is, therefore, ideally suited and commonly used for validating array- or sequencing-based DNA methylation data. We therefore compared the CpG methylation values derived from the 450K arrays with the quantitative methylation values of the same CpGs provided by the pyrosequencing technology. We selected CpGs as analytical targets that fulfilled three conditions: CpGs were differentially methylated between phenotypes (i.e., tumor or nontumoral staus); stable, that is similarly methylated between samples of the same phenotype; and for which the six different preprocessing approaches we tested in the current study yielded different quantitative values. All of the 13 selected targets

Table 1. Comparison of quantitative DNA methylation values of 13 CpGs obtained by pyrosequencing and the preprocessed Illumina Infinium® Human Methylation 450K BeadChip (Illumina Inc., CA, USA) data.

CpG loci identifier	Pyrosequencing signal							
	A	B	C	D	E	F	G	H
cg00598204	12	10	10	13	11	39	5[†]	7
cg00815198	18	18	18	21	21	19	17	6[†]
cg01189638	11	11	11	14	14	29	10[†]	42
cg02637247	12	13	13	17	17	29	10	5[†]
cg02643580	13	14	14	15	15	23	11[†]	11[†]
cg06580770	6[†]	7	7	12	11	39	14	18
cg08639279	6	3[†]	3[†]	6	4	42	3[†]	17
cg10214581	9	7	7	6	6	43	3[†]	14
cg11149194	8	11	11	17	15	28	11	7[†]
cg16341979	7	4[†]	4[†]	6	6	46	5	50
cg17089019	6	4	5	3[†]	3[†]	50	5	18
cg19867250	6	6	6	9	7	35	5[†]	10
cg25840094	10	10	10	14	13	29	9[†]	42

Values correspond to the median of paired differences (same samples) between pyrosequencing values and the result of a preprocessing variant of the same CpG from the Illumina® 450K data set (Illumina, CA, USA). To facilitate comparison with the pyrosequencing data, the array-based methylation β-values were multiplied by 100 to obtain a value comparable to the methylation percentage. All variants except A include steps 1–3 of our preprocessing pipeline.
A: Raw data; B: No normalization; C: Classical quantile normalization as implemented in the lumi package; D: Peak-based correction [16]; E: Peak-based correction followed by quantile normalization; F: Subset quantile normalization with a unique set of reference quantiles computed from Infinium I (InfI) signals; G: Subset quantile normalization with a reference quantiles set computed from InfI signals for each probe category according to the 'relation to CpG' annotations provided by Illumina; H: Subset quantile normalization, with a set of reference quantiles computed from InfI signals for each probe category according to the 'relation to gene sequence' annotations provided by Illumina.
[†]Indicates the smallest difference between the pyrosequencing signal and the preprocessed Infinium signal.

were analyzed by InfII probes on the BeadChip. Data are presented in SUPPLEMENTARY TABLES 1 & 2.

For all 13 probes, the non-normalized median methylation values were close to the pyrosequencing values and varied in the same range (TABLE 1). As a consequence, differences between the medians of tumoral and nontumoral sample were also comparable between the pyrosequencing and the Infinium data (TABLE 2). These observations reconfirm, at least for this subset of probes, the high quality of the Illumina Infinium Methylation 450K array.

Among all the tested normalization variants, the SQN using the relation to CpG annotations to identify category-related anchors provided the greatest number of closest methylation values (n = 7) to those obtained by pyrosequencing for the very same CpG. This approach, together with the peak-based correction approach, provided also the smallest absolute differences in the methylation values when compared with the pyrosequencing-based methylation values.

■ Identification of differentially methylated probes & genes

For the first data set, we calculated the list of differentially methylated genes for each variant of our pipeline and compared these lists with those obtained from raw (step 1 of the pipeline only) and filtered data (steps 1–3 of the pipeline) (TABLE 3). Raw and filtered data provided lists of differentially methylated genes (3931 and 4665 genes, respectively, with an overlap of 3557 genes) the differences being probably due to the color-bias correction and background subtraction steps. Quantile normalization did not affect the selection of differentially methylated genes. However, the three approaches that correct for the InfI/InfII bias significantly increased the number of differentially methylated genes. The SQN approaches using a unique set of reference quantiles or with the reference quantiles computed according to relation to gene sequence yielded 12,316 and 10,087 differentially methylated genes, respectively, corresponding to approximately three-times that obtained from the raw data. The high number of selected genes might be due to the above-described overestimation of the unmethylated status and the underestimation of fully methylated status for InfII probes. The peak-based correction and the SQN approach with reference quantiles computed according to relation to CpG provide 6469 and 7209 differential genes, with an overlap of 5259 genes. In general, the gene lists obtained with the different approaches

did overlap significantly with each other, indicating that these different approaches do not modify the relative ranking of the most differentially methylated genes, but modify absolute methylation values.

Discussion

In this work, we developed and thoroughly evaluated a complete preprocessing pipeline for Illumina Infinium Human Methylation 450K BeadChip data. Our pipeline solves, to the best of our knowledge, most issues related to the processing of DNA methylation data produced with this type of array. Many laboratories perform 'homemade' approaches, or use the dedicated lumi R package [18] to analyze Illumina methylation BeadChip data. However, there is currently no consensus on how to process this type of data. Furthermore, available solutions such as the lumi/methylumi or IMA packages are not complete and do not solve some critical issues of the 450K array, such as the shift between the methylation signals provided by the two types of Infinium assays present on the array.

The IMA [17] package was very recently proposed as a complete analysis pipeline for Illumina's 450K Infinium methylation data. This pipeline offers an integrated and very complete analysis tool focusing on statistical analysis for the identification of differential DNA methylation, annotation of the identified sites/region and adjustment for confounding factors corresponding to the biological analysis of 450K data. However, some preprocessing steps that may be important are missing, such as a quality control related to the number of functional beads associated to a given probe, and the color-bias correction and background subtraction steps. Finally, IMA implements an InfI/InfII shift correction step based on the peak-based correction proposed by Dedeurwaerder et al. [16]. As described in this article, this approach may not be optimal for all samples.

In the present work, we confirmed the existence of a shift between InfI and InfII signals and the lower robustness of InfII signals in comparison to InfI signals. In the current study, we introduced an original normalization approach that corrects the shift between InfI and InfII signals. This approach is based on a SQN principle performed at the level of probe subgroups defined through the CpG region they cover. The SQN outperformed current approaches such as classical quantile normalization or the recently proposed peak-based correction [16] in terms of correction efficiency and estimation of

Table 2. Median DNA methylation differences between tumoral and nontumoral samples obtained by pyrosequencing and after the different variants of the Infinium® Human Methylation 450K BeadChip (Illumina Inc., CA, USA) preprocessing pipeline for 13 CpGs.

CpG loci identifier	Pyrosequencing signal	A	B	C	D	E	F	G	H
cg00598204	40	31	34	35	45	44	19	49	57
cg00815198	38	24	25	25	28	27	41	28	40†
cg01189638	36	29	34	33	41	38	36†	30	17
cg02637247	27	25	29	29	34	32	37	27†	46
cg02643580	43	34	37	37	45	44†	37	41	65
cg06580770	22	31	36	37	47	45	16	48	55
cg08639279	41	34	33	33	44	44	16†	40†	45
cg10214581	39	37	38†	37	51	49	42	49	50
cg11149194	31	31†	34	34	42	39	20	35	53
cg16341979	39	29	36	36	46	40†	21	35	44
cg17089019	39	29	29	29	36†	34	30	30	52
cg19867250	38	26	36	36	45	43	37†	37†	17
cg25840094	32	28	29	30	35	33†	19	31†	57

To facilitate comparison with the pyrosequencing data, the array based methylation β-values were multiplied by 100 to obtain a value comparable to the methylation percentage. All variants except A include steps 1–3 of our preprocessing pipeline.
A: Raw data; B: No normalization; C: Classical quantile normalization as implemented in the lumi package; D: Peak-based correction; E: Peak-based correction followed by quantile normalization; F: Subset quantile normalization with a unique set of reference quantiles computed from Infinium I (InfI) signals; G: Subset quantile normalization with a reference quantiles set computed from InfI signals for each probe category according to the 'relation to CpG' annotations provided by Illumina (CA, USA); H: Subset quantile normalization with a reference quantiles set computed from InfI signals for each probe category according to the 'relation to gene sequence' annotations provided by Illumina.
†Indicates the smallest difference between the pyrosequencing signal and the preprocessed Infinium signal.

DNA methylation signals from InfII probes. The peak-based approach was the only real proposal for InfI/InfII shift correction, which was developed in parallel with our approach. However, this approach requires that two distinct peaks for extreme methylation values for the two types of probes can always be identified, in order to quantify the shift and the same correction will be applied to all values. If more or less than two peaks for the methylation density are present, or if the peaks correspond to partially methylated probes, such an approach will not be applicable and suitable. Accordingly, we found its efficiency sensitive to variation in the shape of the DNA methylation density curve. As this approach relies on the identification of two methylation density modes representing unmethylated and fully methylated probes, it is also highly dependent on the tuning of the bandwidth used as smoothing parameter when estimating the distribution of the methylation density. In our work, we used the default parameter set by Dedeurwaerder *et al.* to 0.5 [16]. We are aware that an optimized tuning of this parameter may have produced better results, but the requirement for tuning this parameter can also be considered as a limitation of the correction approach.

Nonetheless, our pipeline provides the user with the choice of simply performing a classical quantile normalization without InfI/InfII shift correction by selecting the array type (27K or 450K) in the main script before running the preprocessing pipeline. Our pipeline can also be applied to Infinium Human Methylation 27K BeadChip where only the InfI technology is present on the array.

From the results of our study, we can conclude that the SQN method using the relation to CpG annotations successfully achieved two tasks: complete correction of the shift between InfI and InfII probe signals and efficient quantile normalization, without biasing InfII signal toward InfI signal distribution. Furthermore, compared with other methods, this approach provided the closest methylation values to pyrosequencing data, which can be considered as a gold-standard technology demonstrating that our contributions improve the quality of preprocessed data from the Illumina Infinium Human Methylation 450K BeadChip array.

However, some critical points could be raised about the proposed pipeline. We addressed the potential batch effect by a detection approach based on principal component analysis. The small batch effects between BeadChips and/or

sample plates we have so far encountered are easily corrected by the proposed pipeline. However, large batch effects might require a specific correction step and to improve the universal utility of our pipeline further, it might be useful to include such a correction in our pipeline. The work of Gagnon-Bartsch and Speed has already addressed this issue for gene-expression arrays [23]. The principle of their two-step approach consists of using negative-control genes that should not show any variation in a given biological context to estimate what they call 'unwanted' variations. The authors then corrected the whole data set for these unwanted variations. An adaptation of this approach in the context of the lllumina Infinium Human Methylation 450K BeadChip could provide a solution for the batch effect. Another alternative would be to evaluate the empirical Bayes batch adjustment, which has recently been shown to effectively remove nonbiological effects in 27K BeadChip data sets, that is, composed solely of InfI data [24]. If a batch effect is identified but not corrected during the data processing steps, a feature selection approach such as the independent surrogate variable analysis [25], allowing the identification of features correlating with a phenotype of interest in the presence of potential confounding factors such as batch effects, could be used for differential DNA methylation analysis.

In this work, we evaluated the performance of the different variants of our pipeline on the efficiency of the InfI/InfII correction, and on the accuracy of the preprocessed methylation signal (in comparison with pyrosequencing data). However, it would also be very interesting to perform an evaluation at the level of performance of predictive models. This will provide some precise insight on the ability of each pipeline variant to provide predictive signatures. It will also be possible to compare signatures based on InfI probes only with signatures based on InfII or mixing both kinds of probes. As we based our InfI/InfII shift correction approach at the level of probe subgroups defined from the nature of the covered CpG region, it will also be possible to provide a summarized methylation status for genes or functional units based on these functional probe subcategories. It will then be possible to obtain a physical and dynamic insight of the epigenetic variations affecting a given gene.

In our version of SQN, the InfI/InfII shift is corrected by transferring InfI reference quantiles for InfII signal normalization. This procedure is applied whatever the magnitude of the InfI/InfII shift is . In the future, more sophisticated and perhaps more powerful approaches could be developed using a two-step procedure with a first quantification step of the InfI/InfII shift followed by a weighted version of the quantile transfer using a smoothed correction of the distribution of InfII signals.

Our current pipeline consists of different R functions and scripts in the form of text files. One of our next projects will be to release the pipeline as an R package included in the Bioconductor project [26].

Conclusion

The Infinium Human Methylation 450K BeadChip provides a good compromise between coverage, throughput, cost, resolution and accuracy permitting genome-wide epigenome analysis by epigenotyping. However, in addition to the classical signal bias and noise related to microarray data, the existence of two different Infinium assays (InfI and InfII) on the same array induces a signal shift in the methylation profile of a same sample. We therefore developed a complete pipeline that solves, to the best of our knowledge, most issues currently preventing the direct analysis of raw Infinium Human Methylation 450K BeadChip data. This pipeline performs standard quality controls, elimination of unrelated signal variations and filtering of technical noise, as well as a correction of the shift between InfI and II signals using an original

Table 3. Number of differentially methylated genes identified after the different variants of the Infinium® Human Methylation 450K BeadChip (Illumina Inc., CA, USA) preprocessing pipeline and overlaps of the different gene lists.

Preprocessing variant	A	B	C	D	E	F
A (3931 genes)						
B (4665 genes)	3557					
C (4663 genes)	3572	4435				
D (6469 genes)	3809	4451	4488			
E (12316 genes)	3158	3755	3756	4809		
F (7209 genes)	3482	4068	4096	5259	5714	
G (10087 genes)	3097	3691	3687	4661	8327	5482

Number of differentially methylated genes are shown in brackets.
All variants except A include steps 1–3 of our preprocessing pipeline.
A: Raw data; B: No normalization; C: Classical quantile normalization as implemented in the lumi package; D: Peak-based correction followed by quantile normalization; E: Subset quantile normalization with a unique set of reference quantiles computed from Infinium (Inf)I signals; F: Subset quantile normalization with a reference quantiles set computed from InfI signals for each probe category according to the 'relation to CpG' annotations provided by Illumina (CA, USA); G: Subset quantile normalization with a reference quantiles set computed from InfI signals for each probe category according to the 'relation to gene sequence' annotations provided by Illumina.

SQN approach. With the tools for preprocessing now available, the next challenges will be the analysis of these data in order to identify predictive methylation signatures, and to provide a deeper insight into global methylome dynamics.

Future perspective

Epigenotyping arrays bear promise of expansion as the current generation of genotyping arrays do now cover up to five million SNPs. It can be expected that the methylation arrays will also be scaled to several million CpGs on a single BeadChip in the near future. Furthermore, the increasing amount of DNA methylation information available through projects like iHEC and TCGA/ICGC will allow the design of optimized arrays selecting the most informative CpG positions for a given question, which will now also, for the first time, allow large-scale DNA methylation-based epidemiological studies requiring dedicated and optimized preprocessing procedures, such as the one described in this article.

Ethical conduct of research

The authors state that they have obtained appropriate institutional review board approval or have followed the principles outlined in the Declaration of Helsinki for all human or animal experimental investigations. In addition, for investigations involving human subjects, informed consent has been obtained from the participants involved.

Executive summary

Infinium® Human Methylation 450K BeadChip (Illumina Inc., CA, USA) data preprocessing

- There is currently no complete pipeline for quality control, elimination of unrelated signal variations, noise filtering and between-sample normalization.
- We confirm the existence of an additional source of bias in 450k data owing to the shift between Infinium I (InfI) and Infinium II (InfII) assays.

Proposition of a complete preprocessing pipeline

- Our pipeline addresses most current preprocessing issues related to Infinium® Human Methylation 450K BeadChip data.
- We propose a new Infinium signal shift correction based on a subset quantile normalization (SQN) approach performed at the level of 'relation to CpG' annotations.
- This approach uses InfI signal as 'anchors' to improve InfII methylation signal stability and accuracy, and to correct for the InfI/InfII shift.

Evaluation of our preprocessing pipeline & comparison with existing approaches for the correction of the Infinium signal shift

- Infinium Human Methylation 450K BeadChip-processed results are comparable with Pyrosequencing data for a selection of CpG sites.
- SQN outperforms the peak-based correction approach in terms of correction of the shift and quantitative estimation of the methylation status.
- SQN performs, at the same time, an efficient correction for the shift of the Infinium signals and a robust normalization between samples.

Conclusion

- We confirm the reproducibility and accuracy of Infinium Human Methylation 450K BeadChip data.
- Our preprocessing pipeline produces 'ready-to-analyze', accurate methylation data.

References

Papers of special note have been highlighted as:
- of interest
- of considerable interest

1   Tost J. DNA methylation: an introduction to the biology and the disease-associated changes of a promising biomarker. *Mol. Biotechnol.* 44(1), 71–81 (2010).

2   Tost J. Analysis of DNA methylation patterns for the early diagnosis, classification and therapy of human cancers. In: *DNA Methylation Research Trends*. Kobayashi TB (Ed.). Nova Science Publishers, NY, USA 87–133 (2007).

3   Deng D, Liu Z, Du Y. Epigenetic alterations as cancer diagnostic, prognostic, and predictive biomarkers. *Adv. Genet.* 71, 125–176 (2010).

4   Boerno ST, Grimm C, Lehrach H, Schweiger MR. Next-generation sequencing technologies for DNA methylation analyses in cancer genomics. *Epigenomics* 2(2), 199–207 (2010).

5   Laird PW. Principles and challenges of genome-wide DNA methylation analysis. *Nat. Rev. Genet.* 11(3), 191–203 (2010).

■■ Comprehensive review of current technologies and their advantages and pitfalls.

6   Rakyan VK, Down TA, Balding DJ, Beck S. Epigenome-wide association studies for common human diseases. *Nat. Rev. Genet.* 12(8), 529–541 (2011).

■■ Describes the basic and statistical considerations for the design and performance of epigenome-wide association studies.

7   Jeffries MA, Dozmorov M, Tang Y, Merrill JT, Wren JD, Sawalha AH. Genome-wide DNA methylation patterns in CD4+ T cells from patients with systemic lupus erythematosus. *Epigenetics* 6(5), 593–601 (2011).

8   Rakyan VK, Beyan H, Down TA *et al.* Identification of type 1 diabetes-associated DNA methylation variable positions that precede disease diagnosis. *PLoS Genet.* 7(9), e1002300 (2011).

■ Shows the power of the Infinium® (Illumina Inc., CA, USA) technology to detect subtle changes in DNA methylation in complex disease.

9   Toperoff G, Aran D, Kark JD *et al.* Genome-wide survey reveals predisposing diabetes type 2-related DNA methylation variations in human peripheral blood. *Hum. Mol. Genet.* 21(2), 371–383 (2012).

10 Bibikova M, Le J, Barnes B *et al.* Genome-wide DNA methylation profiling using Infinium® assay. *Epigenomics* 1(1), 177–200 (2009).

11 Irizarry RA, Ladd-Acosta C, Wen B *et al.* The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores. *Nat. Genet.* 41(2), 178–186 (2009).

12 Bock C, Tomazou EM, Brinkman AB *et al.* Quantitative comparison of genome-wide DNA methylation mapping technologies. *Nat. Biotechnol.* 28(10), 1106–1114 (2010).

13 Campan M, Moffitt M, Houshdaran S *et al.* Genome-scale screen for DNA methylation-based detection markers for ovarian cancer. *PLoS ONE* 6(12), e28141 (2011).

14 Sandoval J, Heyn H, Moran S *et al.* Validation of a DNA methylation microarray for 450,000 CpG sites in the human genome. *Epigenetics* 6(6), 692–702 (2011).

■ First evaluation of the Infinium® Human Methylation 450K BeadChip (Illumina Inc., CA, USA).

15 Bibikova M, Barnes B, Tsan C *et al.* High density DNA methylation array with single CpG site resolution. *Genomics* 98(4), 288–295 (2011).

16 Dedeurwaerder S, Defrance M, Calonne E, Denis H, Sotiriou C, Fuks F. Evaluation of the infinium methylation 450K technology. *Epigenomics* 3(6), 771–784 (2011).

■■ First described the problems associated with the shift between Infinium I and Infinium II probes and proposed a peak-based correction approach.

17 Wang D, Yan L, Hu Q *et al.* IMA: an R package for high-throughput analysis of Illumina's 450K Infinium methylation data. *Bioinformatics* 28(5), 729–730 (2012).

■■ The first pipeline for 450K methylation data preprocessing using the peak-based correction approach and downstream analysis tools.

18 Du P, Kibbe WA, Lin SM. Lumi: a pipeline for processing Illumina microarray. *Bioinformatics* 24(13), 1547–1548 (2008).

19 Bourgon R, Gentleman R, Huber W. Independent filtering increases detection power for high-throughput experiments. *Proc. Natl Acad. Sci. USA* 107(21), 9546–9551 (2010).

20 Tost J, Gut IG. DNA methylation analysis by pyrosequencing. *Nat. Protoc.* 2(9), 2265–2275 (2007).

21 Aryee MJ, Wu Z, Ladd-Acosta C *et al.* Accurate genome-scale percentage DNA methylation estimates from microarray data. *Biostatistics* 12(2), 197–210 (2011).

22 Irizarry RA, Ladd-Acosta C, Carvalho B *et al.* Comprehensive high-throughput arrays for relative methylation (CHARM). *Genome Res.* 18(5), 780–790 (2008).

23 Gagnon-Bartsch JA, Speed TP. Using control genes to correct for unwanted variation in microarray data. *Biostatistics* doi:10.1093/biostatistics/kxr034 (2011) (Epub ahead of print).

24 Sun Z, Chai HS, Wu Y *et al.* Batch effect correction for genome-wide methylation data with illumina infinium platform. *BMC Med. Genomics* 4(1), 84 (2011).

25 Teschendorff AE, Zhuang J, Widschwendter M. Independent surrogate variable analysis to deconvolve confounding factors in large-scale microarray profiling studies. *Bioinformatics* 27(11), 1496–1505 (2011).

26 Gentleman RC, Carey VJ, Bates DM *et al.* Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.* 5(10), R80 (2004).

■ Websites

101 Davis S, Du P, Bilke S, Triche T. Methylumi: handle illumina methylation data (version 2.0.1) (2010). www.bioconductor.org/packages/release/bioc/html/methylumi.html

102 Wu Z. Subset quantile normalization using negative control features. Johns Hopkins University, Department of Biostatistics Working paper 191, (2009). http://biostats.bepress.com/jhubiostat/paper191

103 International HapMap Project. http://hapmap.ncbi.nlm.nih.gov