

For reprint orders, please contact: [reprints@futuremedicine.com](mailto:reprints@futuremedicine.com)

# Evaluation of the Infinium Methylation 450K technology

**Aims:** Studies of DNA methylomes hold enormous promise for biomedicine but are hampered by the technological challenges of analyzing many samples cost-effectively. Recently, a major extension of the previous Infinium HumanMethylation27 BeadChip® (Illumina, Inc. CA, USA), called Infinium HumanMethylation450 (Infinium Methylation 450K; Illumina, Inc. CA, USA) was developed. This upgraded technology is a hybrid of two different chemical assays, the Infinium I and Infinium II assays, allowing (for 12 samples in parallel) assessment of the methylation status of more than 480,000 cytosines distributed over the whole genome. In this article, we evaluate Infinium Methylation 450K on cell lines and tissue samples, highlighting some of its advantages but also some of its limitations. In particular, we compare the methylation values of the Infinium I and Infinium II assays. **Materials & methods:** We used Infinium Methylation 450K to profile: first, the well-characterized HCT116 wild-type and double-knockout cell lines and then, 16 breast tissue samples (including eight normal and eight primary tumor samples). Absolute methylation values ( $\beta$ -values) were extracted with the GenomeStudio™ software and then subjected to detailed analysis. **Results:** While this technology appeared highly robust as previously shown, we noticed a divergence between the  $\beta$ -values retrieved from the type I and type II Infinium assays. Specifically, the  $\beta$ -values obtained from Infinium II probes were less accurate and reproducible than those obtained from Infinium I probes. This suggests that data from the type I and type II assays should be considered separately in any downstream bioinformatic analysis. To be able to deal with the Infinium I and Infinium II data together, we developed and tested a new correction technique, which we called 'peak-based correction'. The idea was to rescale the Infinium II data on the basis of the Infinium I data. While this technique should be viewed as an approximation method, it significantly improves the quality of Infinium II data. **Conclusion:** Infinium 450K is a powerful technique in terms of reagent costs, time of labor, sample throughput and coverage. It holds great promise for the better understanding of the epigenetic component in health and disease. Yet, due to the nature of its design comprising two different chemical assays, analysis of the whole set of data is not as easy as initially anticipated. Correction strategies, such as the peak-based approach proposed here, are a step towards adequate output data analysis.

**KEYWORDS:** bisulfite-based method • DNA methylation • DNA methylome • epigenetics • epigenomics • Infinium I • Infinium II • Infinium Methylation 450K • peak-based correction

DNA methylation of cytosine residues is essential to the normal development and maintenance of gene-expression patterns [1]. In humans, it occurs mainly in the context of CpG dinucleotides, although the presence of 5-methylcytosine in a non-CpG context has been described in embryonic stem cells [2,3]. Throughout the genome 70% of CpG sites are methylated [3], but short regions of high CpG-density – called CpG islands (CGIs) – found in approximately 60% of gene promoters are usually unmethylated [4,5]. In several diseases, DNA methylation landscapes display numerous alterations [6]. For instance in cancers, a global hypomethylation of the genome, paradoxically associated with silencing of tumor suppressor genes through promoter hypermethylation, is a widely reported event [6–8].

Such alterations of DNA methylation in both cancer and other diseases have raised wide interest in developing large-scale DNA methylation profiling methods [7,9]. Bisulfite genomic sequencing remains the gold standard, as it allows mapping at single-base pair resolution [10–12]. Combined with next-generation sequencing, it constitutes the best approach – in terms of accuracy, coverage and resolution – to decipher the complete DNA methylome [10–12]. Such an approach can allow for a comprehensive assessment of a small number of samples [3]. Biomarker research, however, requires effective high-throughput processing of hundreds or thousands of samples, for example, from clinical cohorts. The best compromise thus far in terms of reagent costs, time of labor, sample

Sarah Dedeurwaerder<sup>1†</sup>,  
Matthieu Defrance<sup>1†</sup>,  
Emilie Calonne<sup>1</sup>,  
Hélène Denis<sup>1</sup>,  
Christos Sotiriou<sup>2</sup>  
& François Fuks<sup>\*1</sup>

<sup>1</sup>Laboratory of Cancer Epigenetics, Université Libre de Bruxelles, Faculty of Medicine, Route de Lennik 808, 1070 Brussels, Belgium

<sup>2</sup>Breast Cancer Translational Research Laboratory JC Heuson, Université Libre de Bruxelles, Jules Bordet Institute, Boulevard de Waterloo 127, 1000 Brussels, Belgium

\*Author for correspondence:

Tel.: +32 2555 6235

Fax: +32 2555 6257

[ffuks@ulb.ac.be](mailto:ffuks@ulb.ac.be)

<sup>†</sup>These authors contributed equally.

future  
medicine part of fsg

throughput and coverage may be the recently developed Infinium HumanMethylation450 BeadChip® (Infinium Methylation 450K; Illumina, Inc. CA, USA). This new-generation array constitutes a major extension of the previous Infinium HumanMethylation27 BeadChip (Infinium Methylation 27K; Illumina, Inc. CA, USA) [13]. While it does not yield an integral map of the DNA methylome, Infinium Methylation 450K makes it possible to assess the methylation status of more than 480,000 cytosines distributed over the entire genome. Two recent reports have already shown the accuracy and reproducibility of this technology, notably by comparing the sample methylation profiles obtained using the Infinium 450K with ones produced by two other technologies previously developed by Illumina, Infinium 27K and GoldenGate [14], and by whole-genome bisulfite sequencing [15]. Interestingly, while Bibikova *et al.* report a difference in performance between the two chemical assays used by this technology, Infinium I and Infinium II, they did not measure the consequences of this difference in performance, notably in the detection of differentially methylated cytosines [15]. In the present study, we have further evaluated the Infinium Methylation 450K technology, focusing notably on these two different chemical assays used on this unique array. We proposed a way to overcome the difference in performance of these two assays.

## Materials & methods

### ■ Samples & DNA extraction

HCT116 wild-type (WT) and HCT116 double-knockout (DKO) cells were cultured in McCoy's 5A medium, supplemented with 10% fetal calf serum. Genomic DNA was extracted with the QIAamp DNA Mini Kit (Qiagen, Hilden, Germany) including the recommended proteinase K and RNase A digestions. The eight archival fresh frozen breast tumor samples (BC) used in this study were from patients diagnosed at the Jules Bordet Institute between 1995 and 2003. In addition to the eight tumor samples, eight normal breast tissue samples (N) were selected as well. Genomic DNA from these frozen clinical samples was extracted from 10 µm sections with the Qiagen DNeasy Blood and Tissue Kit according to the supplier's instructions (Qiagen). The procedure included a proteinase K digestion at 55°C overnight. DNA was quantitated with the NanoDrop® ND-1000 UV-Vis Spectrophotometer (NanoDrop Technologies, Wilmington, DE, USA).

### ■ Bisulfite conversion & DNA methylation profiling with Infinium Methylation 450K

Genomic DNA (800 ng) was treated with sodium bisulfite using the Zymo EZ DNA Methylation Kit™ (Zymo Research, Orange, CA, USA) according to the manufacturer's procedure, with the alternative incubation conditions recommended when using the Illumina Infinium Methylation Assay. The methylation assay was performed on 4 µl bisulfite-converted genomic DNA at 50 ng/µl according to the Infinium HD Methylation Assay protocol. Data quality was checked with the GenomeStudio™ Methylation Module software (2010.3) and all samples passed this quality control. Uncorrected β-values were extracted with the same software. Raw data were submitted to the Gene Expression Omnibus (GSE29290) database [101].

### ■ Bisulfite pyrosequencing

Fifteen CpGs were selected for technical validation of Infinium Methylation 450K by the bisulfite pyrosequencing technique on HCT116 WT and DKO samples. The methylation levels of six of them were assessed on the Infinium 450K array with the Infinium I assay, and the methylation levels of the other nine were assessed with the Infinium II assay. CpGs were selected so as to cover all CpG categories, in terms of both their relation to CGIs and their genomic location. References of the selected CpGs are indicated in TABLE 1 and SUPPLEMENTARY TABLES 1 & 2 ([www.future-medicine.com/doi/suppl/10.2217/epi.11.105](http://www.future-medicine.com/doi/suppl/10.2217/epi.11.105)). As in the Infinium 450K assay, 800 ng of genomic DNA were bisulfite-converted with the EZ DNA Methylation™ kit (Zymo Research). 2 µl of the converted DNA (corresponding to approximately 20–30 ng) were used as template in each subsequent PCR. Primers for PCR amplification and sequencing were deduced with the PyroMark® Assay Design 2.0 software (Qiagen). PCRs were performed with the HotStarTaq DNA polymerase PCR kit (Qiagen) under the following conditions: 95°C 15 min; 52 cycles of 95°C 1 min; 50°C 1 min; 72°C 1 min; 72°C 10 min. The success of amplification was assessed by agarose gel electrophoresis and pyrosequencing of the PCR products was performed with the Pyromark™ Q24 system (Qiagen). Only blue values (perfect calls) were considered for subsequent analyses. All primer sequences are listed in SUPPLEMENTARY TABLE 1.

For bisulfite pyrosequencing on breast tissue samples, four CpGs assessed on the Infinium 450K by the Infinium II assay were

**Table 1. Absolute methylation values of 15 CpGs given by bisulfite pyrosequencing and Infinium 450K before and after peak-based correction in HCT116 WT (r3) and HCT116 DKO (r3) samples.**

Illumina ID	Sample	Infinium assay	BPS (methylation,%)	Infinium 450K raw data (methylation,%)	Infinium 450K peak-based correction (methylation,%)
cg15487600	HCT116 WT	I	4	9	9
cg15513743	HCT116 WT	I	91	93	93
cg15556380	HCT116 WT	I	96	89	89
cg03144619	HCT116 WT	I	100	95	95
cg15522425	HCT116 WT	I	95	89	89
cg14161399	HCT116 WT	I	89	96	96
cg21395967	HCT116 WT	II	9	15	9
cg06851827	HCT116 WT	II	4	10	5
cg16024801	HCT116 WT	II	89	77	83
cg26158528	HCT116 WT	II	95	85	91
cg26307814	HCT116 WT	II	99	89	94
cg13134535	HCT116 WT	II	3	7	3
cg02874371	HCT116 WT	II	7	17	11
cg06395167	HCT116 WT	II	38	41	39
cg14783814	HCT116 WT	II	85	89	94
cg15487600	HCT116 DKO	I	4	10	10
cg15513743	HCT116 DKO	I	70	61	61
cg15556380	HCT116 DKO	I	48	50	50
cg03144619	HCT116 DKO	I	64	69	69
cg15522425	HCT116 DKO	I	15	19	19
cg14161399	HCT116 DKO	I	50	55	55
cg21395967	HCT116 DKO	II	6	16	9
cg06851827	HCT116 DKO	II	4	10	4
cg16024801	HCT116 DKO	II	73	73	73
cg26158528	HCT116 DKO	II	10	27	19
cg26307814	HCT116 DKO	II	20	32	25
cg13134535	HCT116 DKO	II	4	7	2
cg02874371	HCT116 DKO	II	3	9	3
cg06395167	HCT116 DKO	II	11	26	19
cg14783814	HCT116 DKO	II	46	69	69

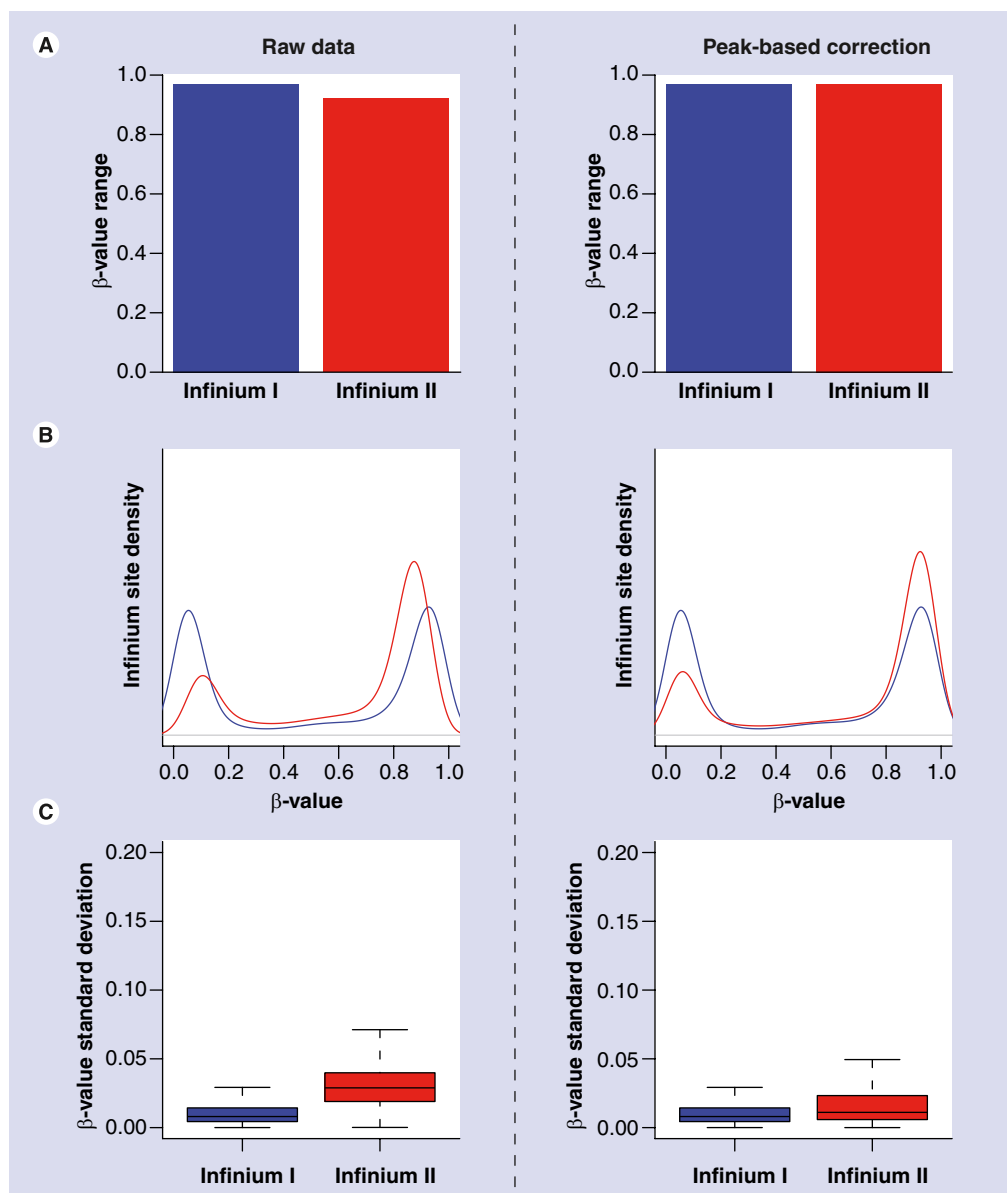
BPS: Bisulfite pyrosequencing; DKO: Double knockout; WT: Wild-type.

selected. Because we only had a limited quantity of material, a preamplification PCR (with EF and ER primers) was performed on bisulfite-converted gDNA before the PCR amplification with biotinylated primers (F and RBio primers) for the pyrosequencing (with S primers). This strategy is described in details in [16]. Sequences of primers used for pyrosequencing on breast tissues are listed in SUPPLEMENTARY TABLE 3.

#### ■ Peak-based correction method

When studied independently, the  $\beta$ -value distributions for both the type I and type II Infinium assays showed two modes corresponding to the

unmethylated or methylated status of the interrogated cytosines. Plotting the  $\beta$ -value densities (kernel density estimation with a Gaussian smoothing function and a bandwidth = 0.05) for both assay types highlighted two clear peaks (an unmethylated peak and a methylated peak), whose positions seemed to be shifted between the type I and type II Infinium assays (FIGURE 1B, left part). Using the peak summits,  $\beta$ -values could be corrected to match the same values for the methylated and unmethylated status of the interrogated cytosines and to cover the same range (FIGURE 1A & 1B). The correction was performed as follows:



**Figure 1. Peak-based correction method for improving Infinium II data, less accurate and reproducible than Infinium I data. (A)** Bar plots indicating the range of  $\beta$ -values generated for HCT116 wild-type (WT) sample (r3) with the Infinium I and Infinium II assays. **(B)** Density plots of the  $\beta$ -values for the two Infinium assay types considered (blue: Infinium I; red: Infinium II) for HCT116 WT sample (r3). **(C)** Box plots of probe-wise variance between the three replicates of HCT116 WT (r1, r2 and r3) for the Infinium I and Infinium II probes (outliers not drawn). On the left part of the figure,  $\beta$ -values have undergone no correction (raw data); on the right part, they have been subjected to the peak-based correction.

- First, the raw  $\beta$ -values were converted to M-values;
- Using the kernel density estimation the peaks were determined independently for both Infinium I and II (SUPPLEMENTARY FIGURE 1B);
- Corrected M-values were computed by rescaling the raw M-values using the peak summits as references;
- Finally, corrected M-values were rescaled to match Infinium I initial range and then

converted back to  $\beta$ -values (SUPPLEMENTARY FIGURE 1C).

The raw  $\beta$ -values were converted to M-values (see [17] for details) using the relation:  $M\text{-value} = \log_2 (\beta\text{-value}/(1 - \beta\text{-value}))$ . The methylated peaks and unmethylated peak summits were determined for both Infinium I and II using a kernel density estimation with a Gaussian smoothing function and a bandwidth = 0.5. Unmethylated peak summits were computed as  $S_U = \text{argmax} (\text{density M-value})$  for

negative M-values (SUPPLEMENTARY FIGURE 1B) for both Infinium I and II. Similarly, methylated peak summits were computed as  $S_M = \text{argmax}(\text{density M-value})$  for positive M-values. The corrected M-values were then obtained by rescaling independently negative and positive M-values using the distance between the peak summits and zero. For negative M-values the corrected M-values were computed as follows: corrected M-value =  $M\text{-value}/\sigma_U$  where  $\sigma_U$  is the distance between the peak summit and zero ( $\sigma_U = 0 - S_U$ ). Corrected positive M-values were computed using the formula: corrected M-value =  $M\text{-value}/\sigma_M$  with  $\sigma_M = S_M - 0$ . To convert back the corrected M-values to  $\beta$ -values, the M-values were first rescaled to match Infinium I range. Negative M-values were rescaled by the Infinium I  $\sigma_U$  (rescaled M-value = corrected M-value  $\cdot \sigma_U$ ) and positive M-values by the Infinium I  $\sigma_M$  (rescaled M-value = corrected M-value  $\cdot \sigma_M$ ). Finally, rescaled M-values were converted to  $\beta$ -values by means of the relation  $\beta\text{-value} = 2^{M\text{-value}}/(2^{M\text{-value}} + 1)$  (SUPPLEMENTARY FIGURE 1C).

All calculations were performed with the freely available R software [102].

### ■ Bioinformatic analyses

Using Infinium annotation data, Infinium sites (cytosines) were classified according to their relation to CGIs and to the closest annotated gene. Regarding their relation to CGIs, the sites were classified in three categories: sites located inside a CGI, sites located in the vicinity of a CGI (<2000 bp: CGI shores), and sites unrelated to any CGI (>2000 bp: distant). As regards their relation to annotated genes, Infinium sites were categorized as inside the promoter (promoter) if they were inside the 1500-bp region upstream from the transcriptional start site, inside the 5'-UTR region, inside the gene body if their location matched the corresponding gene annotation, and inside the 3'-UTR region. Sites related to multiply annotated transcripts falling into several location categories depending on the considered transcript were classified as multiple locations. Finally, sites unrelated to any annotated gene were classified as intergenic.

Concerning differentially methylated sites, for the HCT116 samples, they were selected on the basis of the absolute methylation change between the means of the three replicates of WT and DKO samples (r1, r2 and r3): a site was considered as differentially methylated if the mean  $\beta$ -value for WT samples was >0.8 and the mean  $\beta$ -value for DKO samples was <0.2. For breast tissue samples, sites were

considered differentially methylated when the relative difference in methylation between the means of normal and cancer samples was greater than 0.2 ( $|\text{abs}[\text{mean}\{\text{BC}\} - \text{mean}\{\text{N}\}]| > 0.2$ ). Furthermore, we retained only the sites showing a statistically significant difference according to the Mann–Whitney test, with a false discovery rate set at 0.05 (computed using the Benjamini Hochberg procedure).

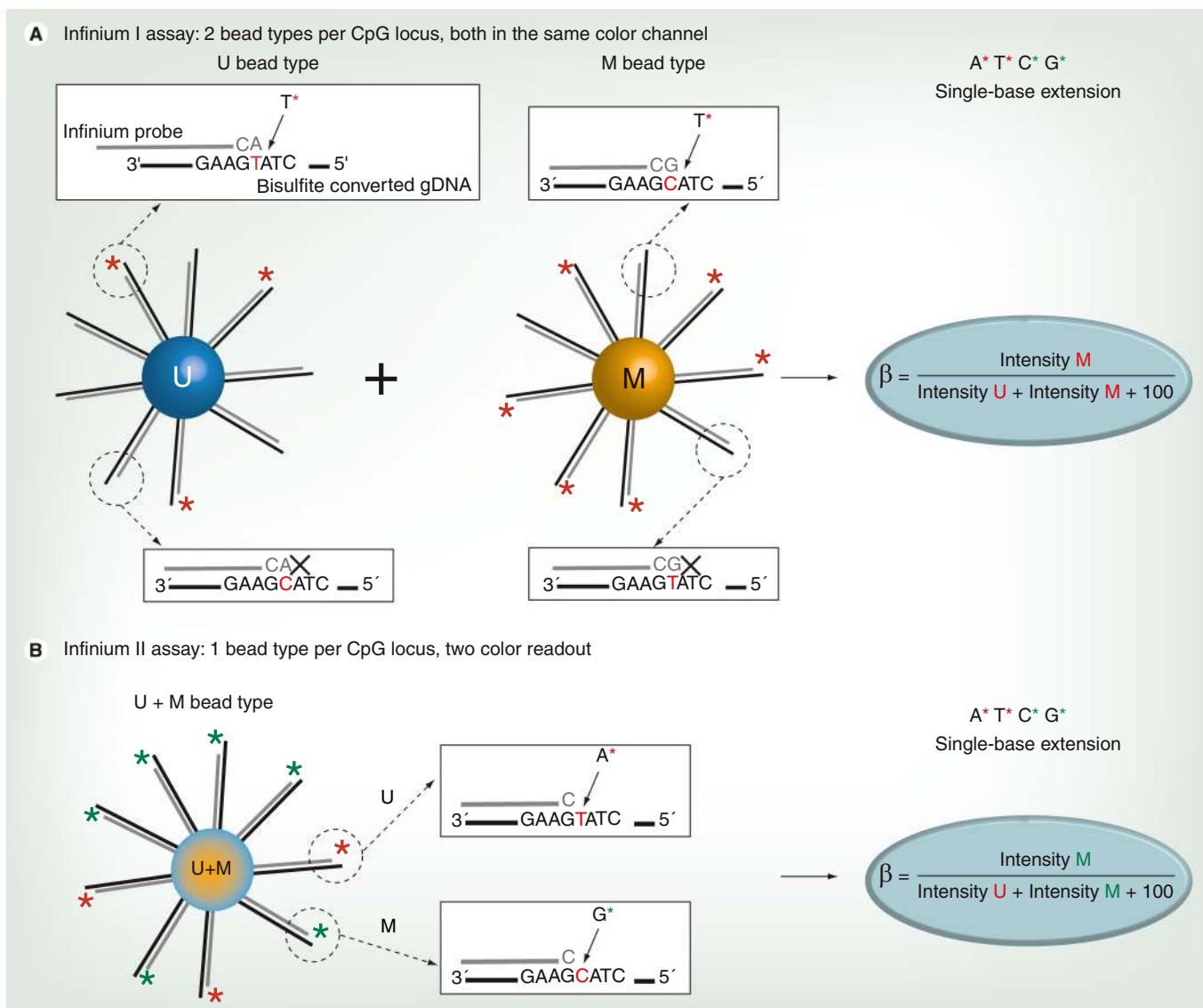
All calculations were performed with the freely available R software [102].

## Results

### ■ Design of the Infinium Methylation 450K array

The general principle of the Infinium Methylation techniques (both Infinium 27K and Infinium 450K) is to evaluate cytosine methylation through quantitative 'genotyping' of the C/T polymorphism generated by the bisulfite conversion, with a throughput of 12 samples per slide [13]. Whereas Infinium 27K uses only one type of assay (Infinium I) [13], Infinium 450K is a hybrid of two different assays (Infinium I and II) (FIGURE 2) allowing coverage of many more cytosines than the previous version, Infinium 27K. Infinium I exploits two different probes (corresponding to the methylated and unmethylated alleles) located on two different bead types, and the methylated and unmethylated signals are generated in the same color channel (FIGURE 2A). Infinium II uses only one bead type with a unique type of probe allowing detection of both alleles. The methylated and unmethylated signals are generated respectively in the green and the red channels (FIGURE 2B). In both cases, the percentage of methylation of a given cytosine is reported as a  $\beta$ -value corresponding to the ratio of the methylated signal over the sum of the methylated and unmethylated signals.

Infinium 450K covers 96% of the CGIs. Multiple CGI shores and CpG sites located far from islands are also highly represented (FIGURE 3A). In addition, 99% of the RefSeq genes are covered overall, and the cytosine sites interrogated are located across gene regions including not only promoter regions but also 5'-UTRs, gene bodies, and 3'-UTRs (FIGURE 3B). Intergenic regions are included as well. This seemed particularly important in the light of recent data suggesting that one should look beyond CGI promoter methylation [18]. It is noteworthy that, except for those located within CGIs, the methylation status of cytosines is more frequently assessed by the Infinium II than by the Infinium I assay



**Figure 2. Overview of the Infinium I and Infinium II assays.** (A) Infinium I and (B) Infinium II present on the Infinium Methylation 450K array.  
M: Methylated; U: Unmethylated.

(FIGURE 3). It is therefore important to evaluate how efficiently the Infinium II assay, newly introduced on Infinium methylation arrays, detects precise methylation levels.

■ Divergence of results obtained with the Infinium I & Infinium II assays

To evaluate Infinium Methylation 450K, we first profiled the well-characterized HCT116 human colon cancer cell line and a derivative thereof (DKO) where DNA methylation is strongly reduced because of deletion of both the DNMT1 and DNMT3B DNA methyltransferases [19]. As previously shown [14], this technology proved highly reproducible (average Pearson correlation between three replicates (r1, r2 and r3): mean  $R^2 = 0.992$  and  $0.988$

for HCT116 WT and DKO, respectively)  
(SUPPLEMENTARY FIGURE 2).

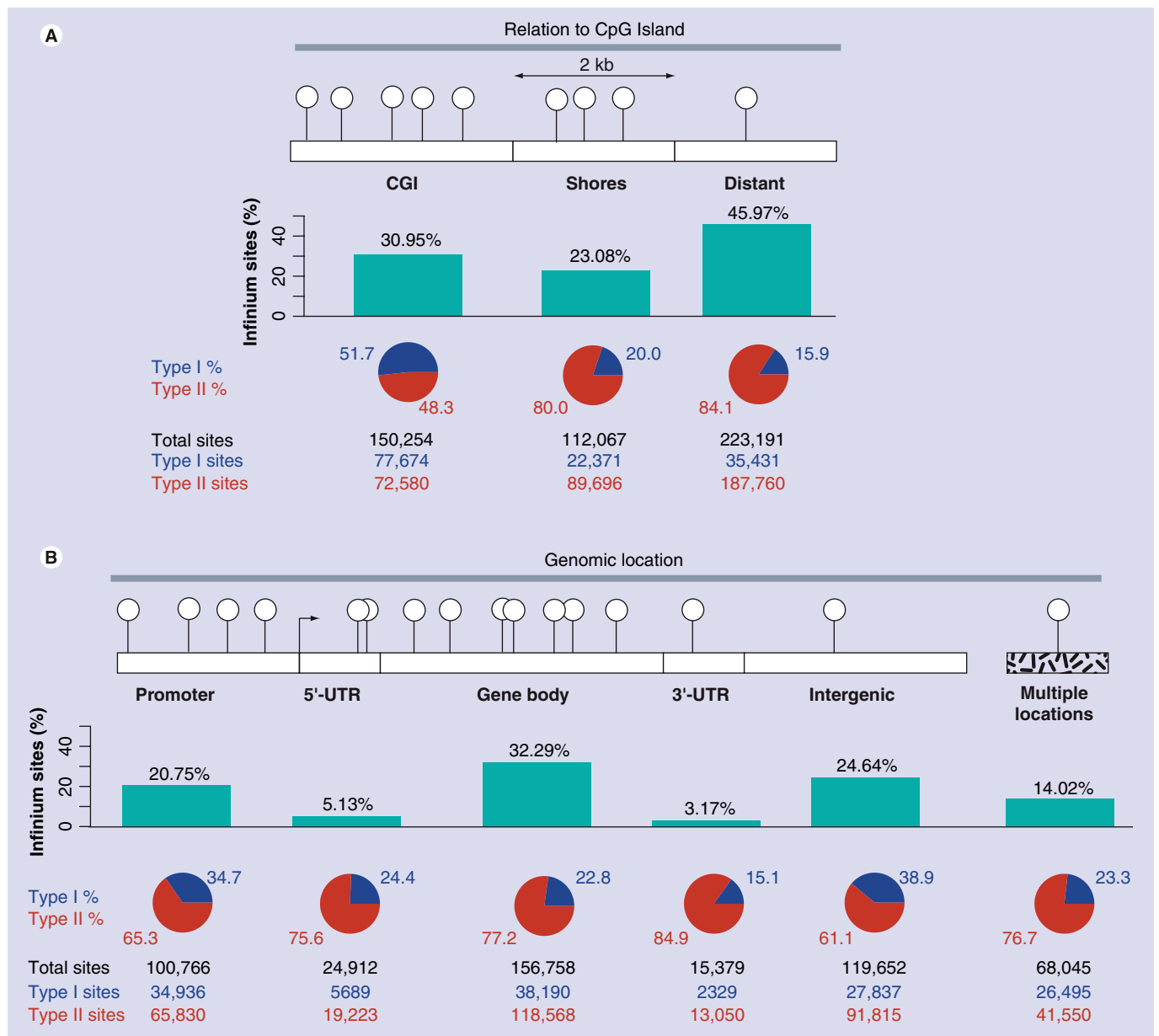
We next considered the data from Infinium I and Infinium II separately in order to compare these assays in terms of performance. Looking at the range of  $\beta$ -values, we remarked that the  $\beta$ -values obtained from the Infinium II probes displayed a range smaller than those obtained from the Infinium I probes (0.971 and 0.922 for Infinium I and II, respectively) (FIGURE 1A, left part). On  $\beta$ -value density plots, both assays displayed a bimodal distribution with two peaks corresponding to unmethylated and fully methylated CpGs, but we noticed a shift of the peaks for Infinium II with respect to those obtained with Infinium I (FIGURE 1B, left part). These differences cannot be explained

by different genomic contexts of the cytosines interrogated by the Infinium I and II assays (SUPPLEMENTARY FIGURE 3). Our data suggest that the Infinium II assay is less sensitive for the detection of extreme methylation values (i.e., 0 and 1) than the Infinium I assay. This is probably due to the dual-channel readout used in the former (FIGURE 2B). In addition, looking at the average probe-wise variance, we observed a greater variance between replicates for the Infinium II probes than for the Infinium I probes (standard deviation = 0.029 for Infinium II vs 0.008 for Infinium I) (FIGURE 1C, left part). Together, these results showed that the Infinium II assay,

although correct, detects absolute methylation levels less efficiently than the Infinium I assay. The divergences between values retrieved from the two assays raise an issue for downstream bioinformatic analysis, as the values are not directly comparable. The Infinium Methylation 450K array should thus be viewed as two different arrays generating results to be treated separately.

#### ■ Data correction

This observed divergence between the two Infinium assays constitutes a major limitation of the technology. It means that when one wishes to capture the complete DNA methylome of a



**Figure 3. Overview of the coverage and design of the Infinium Methylation 450K array.** Histograms showing the percentage of cytosines covered by this technology according to their relation to (A) CpG islands and (B) to their genomic location. Pie charts indicate the proportion of cytosines in each category whose methylation level were assessed by the Infinium I and Infinium II assays.

single sample, two different types of data are actually generated. To get around this limitation, we looked for a way to deal simultaneously with data coming from the two Infinium assays by correcting the  $\beta$ -values generated from the Infinium II assay. Unfortunately, the methods used classically in gene-expression profiling for interarray normalization (e.g., quantile normalization) cannot be used to correct the above-mentioned data for several reasons:

- The Infinium I and Infinium II assays concern only one sample, not two different ones;
- The two assays do not interrogate the same number of CpGs (FIGURE 3);
- The two assays do not interrogate the same CpGs, resulting in a disproportion of the number of methylated and unmethylated sites evaluated between the two assays (FIGURE 1B).

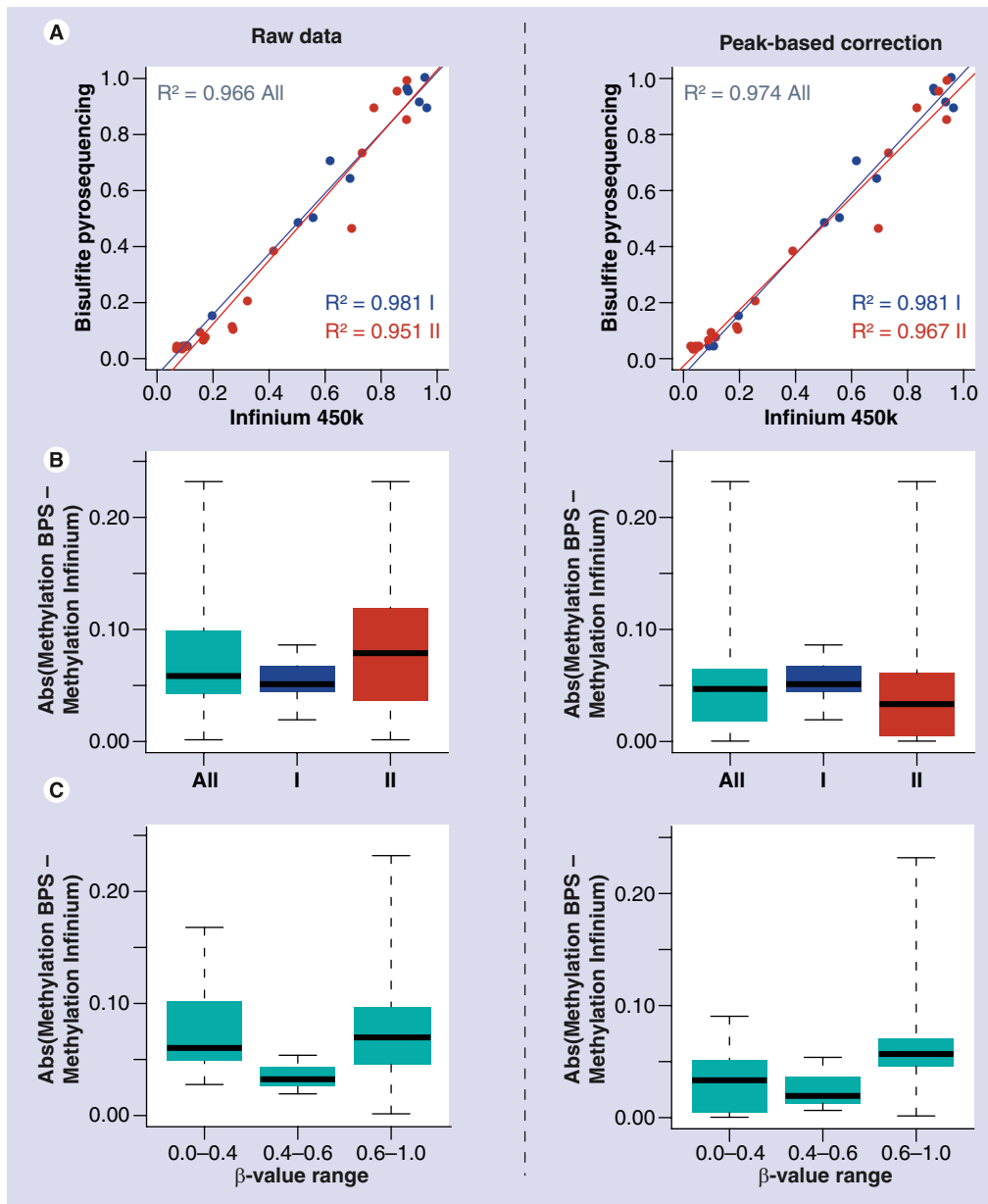
We thus sought another way to correct the data. To this end, we developed and tested a new correction technique, which we called 'peak-based correction'. This method consists in rescaling the Infinium II data to make the unmethylated and methylated peaks of the two assays match (see SUPPLEMENTARY FIGURE 1 and 'Materials & methods'). As shown in the right part of FIGURE 1A for the HCT116 WT sample, this yielded for the Infinium II assay a range of  $\beta$ -values similar to that of the Infinium I assay. As expected, after peak-based correction, the unmethylated and methylated peaks of the two assays matched (FIGURE 1B, right part). In the particular case of DKO cells, displaying no methylated peak, only the unmethylated values were subjected to the correction (see Materials & methods). This resulted in matching of the unmethylated peaks of the two assays without alteration of the methylated values (SUPPLEMENTARY FIGURE 4). One should note that, as only the unmethylated values were stretched, there appeared to be a little break between the methylated and unmethylated values (SUPPLEMENTARY FIGURE 4).

Remarkably, peak-based correction strongly decreased the average probe-wise variance for Infinium II (standard deviation = 0.011), reducing it to the level of variance of Infinium I (FIGURE 1C, right part). Rescaling of the Infinium II  $\beta$ -values to match the Infinium I range tends to move the  $\beta$ -values corresponding to the unmethylated state closer to zero and those corresponding to the methylated state closer to one, reducing the differences between unmethylated or methylated values and thus reducing the variance between replicates. Moreover, although already strong, the

correlation between the Infinium II and bisulfite pyrosequencing data was further enhanced by this correction (Pearson's correlation:  $R^2 = 0.951$  and  $0.967$ , without and with peak-based correction respectively) (FIGURE 4A). FIGURE 4B and TABLE 1 clearly show that applying peak-based correction to the Infinium II data allowed to decrease the differences between methylation values obtained from Infinium and bisulfite pyrosequencing data (see also SUPPLEMENTARY TABLE 2). This suggests that values having undergone the correction are more precise than values which have not. It is of note that, as shown in FIGURE 4C, the benefit of using peak-based correction was specific to extreme methylation values (i.e., close to 0 or 1), being lower for intermediate methylation values (close to 0.5). Thus, while this approach should be viewed as an approximation method, peak-based correction considerably reduces the difference between  $\beta$ -values provided by the Infinium I and Infinium II assays, by correcting the Infinium II  $\beta$ -values. It might not be the perfect method but it makes it possible to consider the two types of data as a single set, thus facilitating downstream bioinformatic analyses.

#### ■ Applicability of Infinium 450K & peak-based correction to clinical tissue samples

Having addressed this technical issue linked to the array design, we next examined the applicability of Infinium Methylation 450K to clinical sample profiling. We profiled eight normal breast samples and eight primary breast cancer samples. On density plots, the  $\beta$ -values still showed a bimodal distribution and, as for the cell lines, we observed a shift of the methylated and unmethylated peaks of the Infinium II  $\beta$ -values, with respect to those of the Infinium I  $\beta$ -values (FIGURE 5A, left part). It is noteworthy that, for the tissue samples, both peaks were less pronounced than for the cell lines (see also SUPPLEMENTARY FIGURE 5). In particular, the methylated peaks were broader than those observed for HCT116 cells (compare FIGURE 5A with FIGURE 1B). This likely reflects the mixture of different cell types present in these clinical samples (data not shown). It might render difficult the precise detection of peak summits and therefore might be a problem for applying the peak-based correction. Nevertheless, when we applied the correction to our tissue data, we managed, as for the HCT116 samples, to make the methylated and unmethylated peaks of the two assays coincide (FIGURE 5A, right part). Moreover, upon performing bisulfite pyrosequencing on a few tissue samples, we showed that globally, the

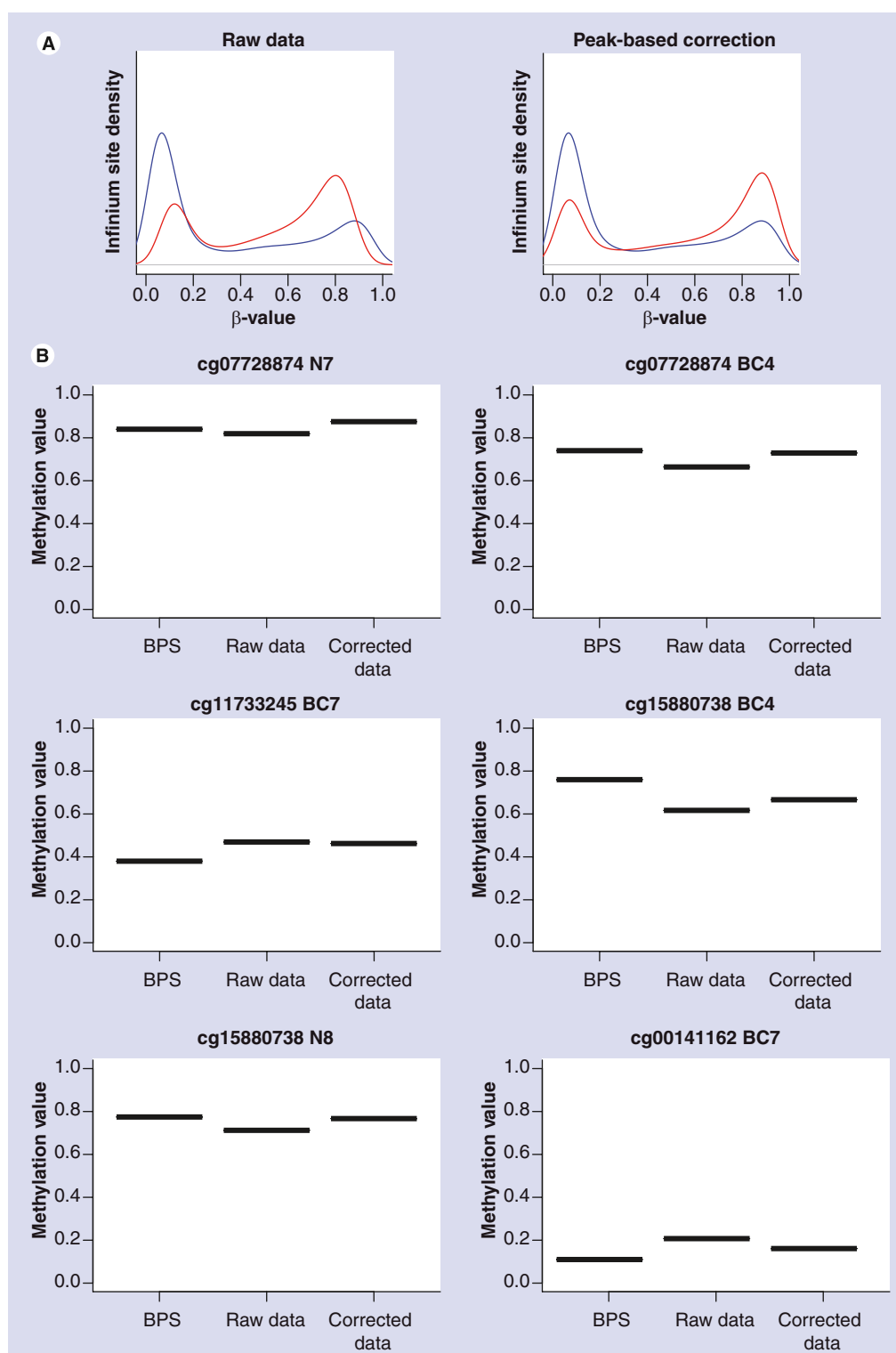


**Figure 4. Enhancement of the correlation between Infinium II and bisulfite pyrosequencing data after peak-based correction.** (A) Correlation for one HCT116 WT sample (r3) between DNA methylation measurements obtained by bisulfite pyrosequencing and with Infinium Methylation 450K when the Infinium data were subjected to no correction (raw data) or peak-based correction. The overall correlation (gray value) and the correlations for the Infinium I and Infinium II assays separately (blue and red, respectively) are given. (B) Box plots indicating the absolute differences between methylation measurements obtained from bisulfite pyrosequencing and Infinium 450K for all Infinium probes and for Infinium I and Infinium II probes separately. (C) Box plots showing the benefit of peak-based correction functions of the original methylation value of Infinium probes for the ranges 0.0–0.4, 0.4–0.6 and 0.6–1.0. On the left part of the figure,  $\beta$ -values have undergone no correction (raw data); on the right part, they have been subjected to the peak-based correction. Abs: Absolute; BPS: Bisulfite pyrosequencing.

Infinium methylation values subjected to peak-based correction were closer to those obtained by bisulfite pyrosequencing than were uncorrected values. (FIGURE 5B). Thus, Infinium 450K is applicable to tissue samples as well, and the peak-based correction allows to improve the data quality.

#### Improving of the detection of differentially methylated cytosines by correcting the Infinium II data

The fact that peak-based correction increases the  $\beta$ -value range and decreases the probe-wise variance for the Infinium II assay suggests that



**Figure 5. Applicability of Infinium 450K and peak-based correction to tissue samples.**

**(A)** Density plots of the  $\beta$ -values for each Infinium assay type considered (blue: Infinium I; red: Infinium II) for one breast tissue sample (N1). On the left part of the figure,  $\beta$ -values have undergone no correction (raw data); on the right part, they have been subjected to the peak-based correction.

**(B)** Absolute methylation levels of four cytosines in a few tissue samples as assessed by bisulfite pyrosequencing and Infinium 450K without and with peak-based correction.

BPS: Bisulfite presequencing.

it could help to detect differentially methylated cytosines. To address this question, we first

determined the number of differentially methylated cytosines between HCT116 WT

and DKO samples from the raw and corrected Infinium data. A cytosine was considered differentially methylated if it was fully methylated in the WT samples (mean  $\beta$ -value  $>0.8$ ) and unmethylated in the DKO samples (mean  $\beta$ -value  $<0.2$ ) (see Materials & methods). Interestingly, we found more differentially methylated cytosines when dealing with the corrected data than with the raw data (TABLE 2). Furthermore, bisulfite pyrosequencing data applied to two cytosines detected as differentially methylated on the basis of the corrected Infinium data (but not on the basis of the uncorrected Infinium data) confirmed that the two cytosines were really differentially methylated and that they were missed if the correction was not applied (FIGURE 6).

We then performed the same experiment on breast tissue samples by comparing normal tissues and tumors. In this case, a cytosine was considered differentially methylated if the corresponding relative difference in methylation between the mean of the N and BC samples was at least 0.2 and statistically significant (as assessed by the Mann–Whitney test, false discovery rate = 0.05), (see Materials & methods). As previously shown for HCT116 cell lines, applying the peak-based correction made it possible to detect more differentially methylated cytosines (SUPPLEMENTARY TABLE 4). This clearly suggests that applying a correction to the Infinium II data, such as the peak-based correction described here, can help to identify differentially methylated cytosines.

## Discussion

The present report is an evaluation of the recently released Infinium Methylation 450K technology, assessing the quality of the output data generated by the two chemical assays present on this array. The previous version of the array, Infinium 27K, uses only the Infinium I chemical assay. On the new Infinium 450K array, approximately a third of the cytosines are interrogated with Infinium I, but roughly two-thirds are interrogated with another chemical assay, called Infinium II. Here, after examining the data globally and showing their accuracy and reproducibility, we have also directly compared the  $\beta$ -values retrieved from the two Infinium assays and highlighted some significant differences between the two sets of data. Recently, two other groups also studied this new technology and demonstrated its accuracy and reproducibility [14,15]. Our data are in agreement with those two studies. In one of them, the authors noticed a difference between

**Table 2. Number of differentially methylated cytosines between HCT116 wild-type and double-knockout samples when  $\beta$ -values are subjected or not to peak-based correction.**

Infinium probe type	Raw data	Peak-based correction
All	29612	46253
Type I	15901	15901
Type II	13711	30352

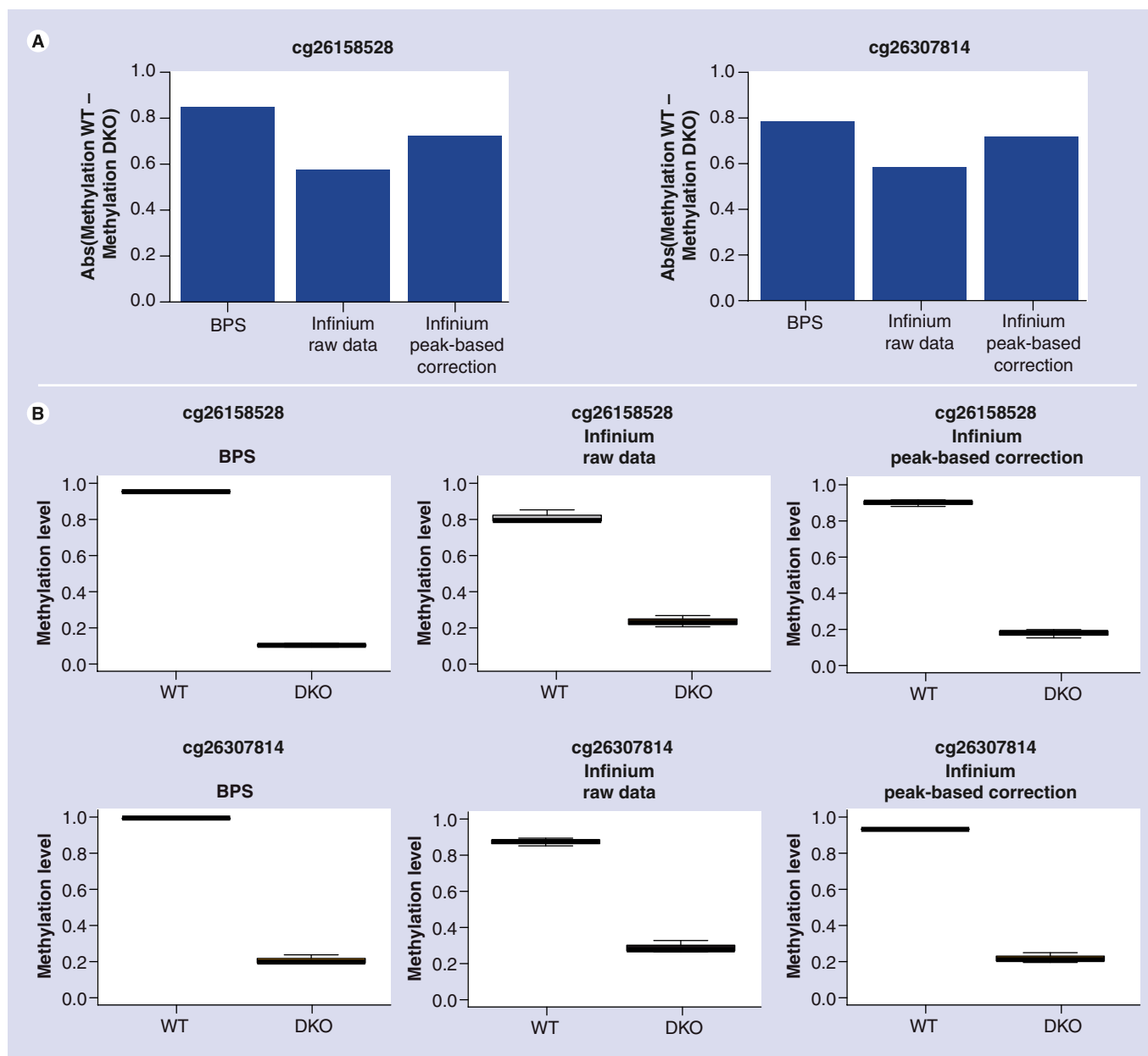
the ranges of  $\beta$ -values retrieved from the two Infinium assays [15]. In our study, we have also noticed this difference in terms of  $\beta$ -values distribution, and in addition, we have highlighted a difference in the average probe-wise variances between replicates. The Infinium II assay is thus less accurate and reproducible, and notably less sensitive for the detection of extreme methylation values (e.g., 0 and 1), than the Infinium I assay. This is really noteworthy, as it means that Infinium I and Infinium II data are not directly comparable.

We thus reveal here an additional bias linked to the use of this technology. Actually, although this was not the focus of the present study, it is important to bear in mind that other potential biases can be observed with this technology:

- As with all bisulfite-based methods, incomplete bisulfite conversion can introduce a bias;
- As with all array-based methods, Infinium Methylation 450K results are probably subject to technical confounding, including a batch effect [20] (although this has yet to be demonstrated);
- SNPs sometimes contained in probes can interfere with methylation level detection [21].

The differences revealed here between Infinium I and II data render downstream bioinformatic analysis more complex. To make the two sets comparable, we have developed a method for correcting the Infinium II  $\beta$ -values. We called it peak-based correction. It consists in rescaling the Infinium II data. Although approximate, this correction method considerably improves the accuracy and reproducibility of Infinium II data, notably for extreme methylation values (close to 0 or 1) and reduces the bias. Furthermore, it facilitates the detection of differentially methylated cytosines that can be missed if the Infinium II data are not corrected.

After demonstrating the robustness of the technique and providing a mean of correcting the bias linked to the use of two different assays on the same array, we have demonstrated the applicability of this technology to profiling



**Figure 6. Improvement of the detection of differentially methylated cytosines by Infinium 450K when applying the peak-based correction. (A)** Bar plots showing, for two cytosines, the differences in methylation between HCT116 WT and HCT116 DKO samples, as assessed by bisulfite pyrosequencing and Infinium 450K without and with peak-based correction. **(B)** Box plots showing the absolute methylation levels of the same two cytosines in HCT116 WT and HCT116 DKO samples, as assessed by bisulfite pyrosequencing and Infinium 450K without and with peak-based correction. Abs: Absolute; BPS: Bisulfite pyrosequencing; DKO: Double knockout; WT: Wild-type.

clinical tissue samples. This is really important as it makes this technology one of the most attractive for large clinical studies. It is worth noting, however, that this technology has a major limitation, that is its inability to differentiate methylation from hydroxymethylation, another chemical modification that can be found on cytosine residues [22]. In the future, a challenge will thus be to develop new genome-wide technologies capable of distinguishing these two types of modified cytosines.

## Conclusion

In conclusion, our study reveals the Infinium Methylation 450K technology as a highly accurate and reproducible technology for genome-scale DNA methylation profiling, even though it presents a bias linked to its particular design. This bias should be corrected with a method – such as the peak-based correction method presented here – before further downstream analysis. Compared with truly genome-wide bisulfite sequencing, the focus of Infinium 450K on

over a half a million cytosines located all over the genome translates into a cost-effective and high-throughput technology. In addition, this technology allows a precise quantification of the methylation level of interrogated cytosines and generates data that are more quickly and easily analyzable than sequencing data. All this and the low input DNA requirements of Infinium make it, in our opinion and that of others [23], the most attractive powerful and cost-effective tool available to date for generating quantitative DNA methylomes in health and disease, notably in the framework of large biomarker discovery studies.

### Financial & competing interests disclosure

S Dedeurwaerder and M Defrance were supported by the Belgian 'FNRS-Télévie' and 'Interuniversity Attraction Poles' (IAP P6/28), respectively. E Calonne and H Denis were supported by the ULB and the Brussels Region 'BruBreast'. C Sotiriou and F Fuks are 'Chercheur Qualifié'

and 'Maître de Recherche' from the FNRS. This work was funded by grants from the FNRS and Télévie, the Brussels Region 'BruBreast' and the 'Interuniversity Attraction Poles' (IAP P6/28), by the EU grant CANCERDIP FP7-200620 and by a European Molecular Biology Organization Young Investigator Programme (EMBO YIP). The authors have no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed.

No writing assistance was utilized in the production of this manuscript.

### 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

### Design of the Infinium Methylation 450K array

- Infinium Methylation 450K is a hybrid of two different assays, Infinium I and II.

### Divergence of results obtained with the Infinium I & Infinium II assays

- Due to its design, Infinium Methylation 450K technology generates a dataset that should be viewed as two distinct datasets.
- Infinium II data are less accurate and reproducible than Infinium I data.

### Data correction

- Peak-based correction is a method consisting in rescaling the Infinium II data on the basis of the Infinium I data.
- This peak-based correction method improves the accuracy and reproducibility of Infinium II data.
- Peak-based correction makes it possible to treat Infinium I and Infinium II data as a single dataset.

### Applicability of Infinium 450K & peak-based correction to clinical tissue samples

- Infinium Methylation 450K and peak-based correction are applicable to clinical tissue samples.

### Improving the detection of differentially methylated cytosines by correcting the Infinium II data

- Applying the peak-based correction to Infinium II data facilitates the detection of differentially methylated cytosines.

### Conclusion

- Infinium Methylation 450K is one of the most attractive powerful and cost-effective tool currently available for generating quantitative DNA methylomes for health and disease, notably in the framework of large biomarker discovery studies.

## References

Papers of special note have been highlighted as:

▪ of interest

▪▪ of considerable interest

- 1 Feinberg AP. Phenotypic plasticity and the epigenetics of human disease. *Nature* 447, 433–440 (2007).
- 2 Laurent L, Wong E, Li G *et al.* Dynamic changes in the human methylome during differentiation. *Genome Res.* 20, 320–331 (2010).
- 3 Lister R, Pelizzola M, Dowen RH *et al.* Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* 462, 315–322 (2009).
- 4 Saxonov S, Berg P, Brutlag DL. A genome-wide analysis of CpG dinucleotides

in the human genome distinguishes two distinct classes of promoters. *Proc. Natl Acad. Sci. USA* 103, 1412–1417 (2006).

- 5 Wang Y, Leung FC. An evaluation of new criteria for CpG islands in the human genome as gene markers. *Bioinformatics* 20, 1170–1177 (2004).
- 6 Portela A, Esteller M. Epigenetic modifications and human disease. *Nat. Biotechnol.* 28, 1057–1068 (2010).
- 7 Jones PA, Baylin SB. The epigenomics of cancer. *Cell* 128, 683–692 (2007).
- 8 Sharma S, Kelly TK, Jones PA. Epigenetics in cancer. *Carcinogenesis* 31, 27–36 (2010).
- 9 Feinberg AP. Epigenomics reveals a functional genome anatomy and a new approach to common disease. *Nat. Biotechnol.* 28, 1049–1052 (2010).
- 10 Bibikova M, Fan JB. Genome-wide DNA methylation profiling. *Wiley Interdiscip. Rev. Syst. Biol. Med.* 2, 210–223 (2010).
- 11 Laird PW. Principles and challenges of genomewide DNA methylation analysis. *Nat. Rev. Genet.* 11, 191–203 (2010).
- 12 Suzuki MM, Bird A. DNA methylation landscapes: provocative insights from epigenomics. *Nat. Rev. Genet.* 9, 465–476 (2008).
- 13 Bibikova M, Le J, Barnes B *et al.* Genome-wide DNA methylation profiling using Infinium assay. *Epigenomics* 1, 177–200 (2009).

- **Technology report on the Infinium I assay used on the Infinium Methylation 27K array.**
- 14 Sandoval J, Heyn HA, Moran S *et al.* Validation of a DNA methylation microarray for 450,000 CpG sites in the human genome. *Epigenetics* 6, 692–702 (2011).
- **The first study describing the Infinium Methylation 450K technology.**
- 15 Bibikova M, Barnes B, Tsan C *et al.* High density DNA methylation array with single CpG site resolution. *Genomics* 98, 288–295 (2011).
- **The second study assessing the accuracy and reproducibility of the Infinium 450K technology.**
- 16 Dedeurwaerder S, Desmedt C, Calonne E *et al.* DNA methylation profiling reveals a predominant immune component in breast cancers. *EMBO Mol. Med.* doi:10.1002/emmm.201100801 (2011) (Epub ahead of print).
- **Shows the applicability and relevance of using Infinium technologies for large biomarker discovery studies.**
- 17 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* 11, 587 (2010).
- 18 Ndlovu MN, Denis H, Fuks F. Exposing the DNA methylome iceberg. *Trends Biochem. Sci.* 36, 381–387 (2011).
- 19 Rhee I, Bachman KE, Park BH *et al.* DNMT1 and DNMT3b cooperate to silence genes in human cancer cells. *Nature* 416, 552–556 (2002).
- 20 Leek JT, Scharpf RB, Bravo HC *et al.* Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat. Rev. Genet.* 11, 733–739 (2010).
- 21 Byun HM, Siegmund KD, Pan F *et al.* Epigenetic profiling of somatic tissues from human autopsy specimens identifies tissue- and individual-specific DNA methylation patterns. *Hum. Mol. Genet.* 18, 4808–4817 (2009).
- 22 Huang Y, Pastor WA, Shen Y, Tahiliani M, Liu DR, Rao A. The behaviour of 5-hydroxymethylcytosine in bisulfite sequencing. *PLoS ONE* 5, e8888 (2010).
- 23 Rakyen VK, Down TA, Balding DJ, Beck S. Epigenome-wide association studies for common human diseases. *Nat. Rev. Genet.* 12, 529–541 (2011).
- **Highlights the pros and cons of each technology available to date that can be used for epigenome-wide association studies.**
- **Websites**
- 101 Gene Expression Omnibus. [www.ncbi.nlm.nih.gov/geo](http://www.ncbi.nlm.nih.gov/geo)
- 102 The R Project for Statistical Computing. [www.r-project.org](http://www.r-project.org)