Improving Infinium MethylationEPIC data processing: focus on enhancers and long noncoding RNA genes

Jana Jeschke (jana.jeschke@ulb.be)
Université Libre de Bruxelles (ULB)
Martin Bizet
Université Libre de Bruxelles (ULB)
Matthieu Defrance
Interuniversity Institute of Bioinformatics in Brussels (IB2), ULB
Emilie Calonne
Université Libre de Bruxelles (ULB)
Gianluca Bontempi
Interuniversity Institute of Bioinformatics in Brussels (IB2), ULB
Christos Sotiriou
Institut Jules Bordet
François Fuks
Université Libre de Bruxelles (ULB)

Research Article

Keywords: DNA methylation, 5mC, Infinium, MethylationEPIC, EPIC, 850k, annotation, normalisation, enhancers, long noncoding RNA

Posted Date: March 14th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1439389/v1

Abstract

Background

Illumina Infinium DNA Methylation (5mC) profiling arrays are a popular technology for low-cost, high-throughput, genome-scale measurement of 5mC distribution, especially in cancer and other complex diseases. After the success of its HumanMethylation450 array (450k), Illumina released the MethylationEPIC array (850k) featuring, in addition to the regulatory regions primarily covered by 450k (promoters, gene bodies), increased coverage of enhancers. Despite the widespread use of 850k, analysis of the corresponding data remains suboptimal: it still relies mostly on Illumina's default annotation, which underestimates enhancers and long noncoding RNAs.

Results

We have thus developed an approach, based on the ENCODE and LNCipedia databases, which greatly improves upon Illumina's default annotation of enhancers and long noncoding transcripts. We then compared the re-annotated 850k with both 450k and reduced-representation bisulfite sequencing (RRBS), another high-throughput 5mC profiling technology. We found 850k to cover at least three times as many enhancers and long noncoding RNAs as either 450k or RRBS. We further investigated the reproducibility of the three technologies, applying various normalisation methods to the 850k data. Most of these methods reduced variability to a level below that of RRBS data. We then used 850k with our new annotation and normalisation to profile 5mC changes in breast cancer biopsies. 850k highlighted aberrant enhancer methylation as the predominant feature, in agreement with previous reports.

Conclusion

Our study provides an updated processing approach for 850k data, based on refined probe annotation and normalisation and allowing improved analysis of methylation at enhancers and long noncoding RNA genes. Overall, our findings will help to further advance understanding of the DNA methylome in health and disease.

Background

5-Methylcytosine (5mC) is an abundant epigenetic mark resulting from enzymatic addition of a methyl group (CH_3) to the fifth carbon of a cytosine, typically in the context of a cytosine-guanine dinucleotide (CpG) in human DNA(1). It has a central function in normal human development, as it controls gene transcription and other physiological processes such as chromosome stability, imprinting, and X-chromosome inactivation(1, 2). Several techniques have been developed to measure the genome-wide distribution of 5mC and to study its role in development and disease(3).

Whole-genome bisulfite conversion followed by sequencing (WGBS) offers single-base resolution and provides the most complete 5mC maps. Yet its high cost and the large amount of input material required limit its current use, especially in clinical studies of cancer and other complex diseases, which require profiling large patient cohorts from low amounts of input material(4). For such studies, high-throughput technologies such as reduced-representation bisulfite sequencing (RRBS) and Infinium HumanMethylation microarrays are preferred(5–7). RRBS uses a methylation-insensitive restriction enzyme to limit sequencing to regions of moderate to high CpG density(8). Like WGBS, RRBS allows single-base resolution, but sample input and costs are lower thanks to reduced sequencing.

Illumina's Infinium HumanMethylation arrays assess methylation levels in large portions of the genome at single-base resolution by targeting selected cytosines. The Infinium HumanMethylation27 bead chip (27k) was released in 2008 and covers 27,578 CpGs primarily located within gene promoters(9). The updated Infinium HumanMethylation450 bead chip (450k) covers more than 450,000 CpGs located within promoters, gene bodies, and intergenic regions(10, 11). Although this array shows probe-specific biases that need to be corrected(12), it has proven to be efficient and affordable for genome-scale differential methylation analysis. It was used, for example, by "The Cancer Genome Atlas" (TCGA)(5) and by numerous other large-scale studies on cancer, diabetes, and ageing(13–18) .

High-throughput profiling of patient biopsies has highlighted perturbation of the DNA methylome as a hallmark of cancer. Numerous studies have shown 5mC alterations to contribute to the development and progression of breast and other cancers(19, 20), demonstrating its potential as a cancer biomarker and cancer therapy target(21, 22). With technological progress in recent years, it has emerged that changes in the DNA methylome, under both physiological and pathological conditions, stretch beyond the regions (promoters and gene bodies) primarily covered by the 27k and 450k arrays(23–25). In breast and other cancers, 5mC aberrations have been shown to occur most frequently at enhancers(26–29), and enhancer methylation appears to anti-correlate better than promoter methylation with target gene expression(3, 30, 31). In addition, genomic regions transcribed to noncoding RNAs have been identified as frequent methylation targets. Among these, long noncoding RNAs (lncRNAs, defined as noncoding transcripts exceeding 200 nucleotides) have been shown to regulate key biological processes(32), and both their levels and 5mC patterns have been found altered in cancers(33, 34).

With growing interest in enhancer methylation, Illumina released in 2016 the Infinium MethylationEPIC bead chip (850k), which targets more than 850,000 CpGs and features improved coverage of non-promoter regions, particularly enhancers. Although this array is widely used to study various diseases, analysis of its data remains suboptimal, as it relies mostly on the default probe annotation provided by Illumina(35–37). According to this annotation, which identifies enhancers on the basis of the FANTOM5 database(38), 4.6% of the probes target enhancer-associated cytosines and 2.5% of the targeted cytosines are associated with noncoding RNAs(39). As FANTOM5 enhancers were identified with the restrictive "Cap analysis gene expression" (CAGE) method, the number of enhancers covered by the 850k array is grossly underestimated. Similarly, lncRNA genes are poorly mapped in Illumina's default annotation.

To overcome these limitations, we have developed a novel approach based on the Encyclopedia of DNA element (ENCODE) and LNCipedia databases to re-annotate 850k data. We have demonstrated the advantage of this approach over other technologies in several ways: (1) by comparing its coverage of regions

corresponding to enhancers and lncRNAs with those of two other high-throughput technologies for profiling large patient cohorts (450k and RRBS); (2) by using various normalisation methods developed for 450k to investigate the reproducibility of the data obtained; (3) by using 850k with our new annotation to perform differential methylation analysis on breast cancer biopsies. Our study thus provides an updated analysis pipeline for 850k array data, based on refined probe annotation and normalisation and allowing improved analysis of methylation at enhancers and lncRNA genes.

Results

Re-annotation of the 850k array

Illumina's default annotation of the 850k array provides few clues for identifying enhancer-associated CpGs, even though enhancer coverage is one of its main features(39, 40). To overcome this limitation, we defined enhancers and other regulatory regions on the basis of the hidden Markov model provided by ENCODE (ENCODE-HMM) data. ENCODE-HMM provides enhancer locations based on chromatin immunoprecipitation followed by sequencing (ChIP-seq) of enhancer-specific histone marks such as H3K4me1 and H3K27ac. It features a larger number of enhancers than the restricted Functional annotation of mammalian genome version 5 (FANTOM5) database used by Illumina. With the ENCODE-based annotation, we were able to reduce the number of CpGs associated with 'Intergenic' regions by more than 50% (128,367 reduced to 60,236) and to improve the association of cytosines with 'Dual' regions 1.5-fold (88,822 to 137,952) and with 'Enhancer' regions more than 3.5-fold (68,509 to 259,096) (Fig. 1A). Overall, our new annotation resulted in fewer non-annotated probes and increased the number of CpGs associated with enhancers. Importantly, using the EnhancerAtlas database we were able to associate more than 120,000 enhancer-associated ENCODE CpGs with target transcripts. This means that more than 60% of the EnhancerAtlas target transcripts are associated with at least one CpG present on the 850k array. This essential information is missing from Illumina's default annotation, but is now available through our ENCODE-based annotation as alternative platform annotation on GEO (token will be provided once GEO approved data).

Illumina's annotation was found to associate more CpGs with promoter regions than the ENCODE-based annotation (Fig. 1A). We found CpGs close to a transcription start site (TSS) to be consistently assigned to the 'Promoter' category in both the ENCODE-based and Illumina annotations, but regions outside the − 600 to + 1,200 range showed high discrepancy, increasing with the distance from the TSS and reaching more than 75% for regions located more than 3 kb from the TSS (Fig. 1B). Overall, more than 25% of the CpGs annotated to the 'Promoter' category by Illumina were not confirmed by the ENCODE-based annotation, suggesting that Illumina's annotation associates some CpGs with promoters incorrectly.

Next, to improve the annotation of 850k probes to noncoding transcripts, we used more than 80,000 lncRNAs from the LNCipedia database. In addition, we retrieved coding and small noncoding transcripts from the Ensembl database, so as to produce a global set of more than 250,000 transcripts(41, 42). In addition to the 120,000 CpGs we had already associated with transcription enhancement, we linked more than 300,000 CpGs to transcription initiation by assessing the overlap between a promoter region containing a targeted cytosine and the TSS corresponding to a transcript. Similarly, we identified more than 90,000 850k-targeted CpGs that might regulate enhancer RNAs (eRNA), a type of lncRNA subject to enhancer-controlled transcription initiation. Finally, more than 630,000 CpGs were associated with gene bodies, being located between the TSS and the transcription termination site (TTS) of a transcript-associated gene. As shown in Fig. 1C, this approach allowed us to substantially improve coverage, particularly of lncRNA genes.

Coverage of the re-annotated 850k array in comparison to 450k and RRBS

We then compared the coverages of three high-throughput technologies suited for profiling large patient cohorts: the re-annotated 850k bead chip, its precursor 450k, and RRBS. For this we used in-house data sets for HCT116 cells profiled with triplicate 450k and 850k arrays and a publicly available RRBS data set for Ewing sarcoma. Unreliable CpG measurements were filtered out, including cross-reactive probes (450k and 850k), low coverage reads (RRBS), CpGs located in the sex chromosomes, and CpGs not assessed in all three replicates. After filtering, the proportion of 'reliable' probes remained high for both 850k (92.8%) and 450k (89.2%). In contrast, only 25.6 to 26.3% of the CpGs assessed by RRBS remained after filtering, either because read coverage was insufficient (< 10) or because the methylation level was not available for all three replicates (Fig. 2A). Overall, we assessed, with the 850k array, almost 2 times as many as with the 450k array (803,509 *versus* 433,252 CpGs) and not quite two-thirds as many as with RRBS (803,509 *versus* 1,243,458 CpGs).

Next, we investigated the locations of the CpGs targeted by the three technologies. We first used the UCSC database (which Illumina also used to annotate their 850k platform) to map CpGs to 'CpG Islands' (CGI), 'Shores' (2 kb surrounding a CGI), or 'Open sea'. We found coverage of CpGs at CGIs to be higher with RRBS than with 450k or 850k (Fig. 2B). Shore CpGs were similarly covered by all three technologies. Open-sea CpGs were covered mostly by RRBS, but coverage was much better with 850k (more than 500,000 covered) than with 450k. We then mapped the CpGs to 'Enhancer', 'Promoter', 'Dual', 'Gene body', and 'Intergenic' regions, using the same database, ENCODE-HMM, as for re-annotation of 850k. The 850k array showed better coverage of all ENCODE-HMM-derived categories than the 450k array, especially 'Enhancer' and 'Gene body' (Fig. 2C). About 200,000 850k-targeted CpGs were assigned a 'Gene body' location (25% of the array). The 'Enhancer' category was the dominant one covered by the 850k array, since more than 250,000 of its probes were exclusively located in enhancers. If one adds to this the 'Dual' CpGs (nearly 140,000), almost half of the 850k array probes (49%) were found to target "enhancer CpGs". RRBS predominantly targeted CpGs exclusively assigned a 'Promoter' (24%), 'Gene body' (33%), or 'Intergenic' (15%) location, as compared to about 180,000 CpGs (15%) assigned an 'Enhancer' location and 160,000 (13%) assigned a 'Dual' location.

Using the ENCODE lists of regulatory regions for each cell line, we assessed in the 850k, RRBS, and 450k data the percentage of regulatory regions covered by at least one CpG. The 850k array outperformed both of the other technologies. It covered between 35.8 and 51.5% of the listed enhancers, according to the cell line assessed (Fig. 2D), as compared to 18.3 to 29.1% for 450k and only 13.1–18.6% for RRBS. Interestingly, although RRBS featured more promoter-associated CpGs than the Infinium arrays (Fig. 2B), it covered a lesser percentage of promoters (48.5 to 71.9%) than either 850k (between 59.5 and 81.7%) or 450k (50.1 to 77.0%) (Fig. 2D). This is due to the fact that the promoters assessed by RRBS tended to have a higher CpG density: RRBS covered 25.7 to 47.1% of the promoters with at least 10 CpGs, as compared to 12.8 to 25.7% for 450k and 17.0 to 33.2% for 850k (Additional File 1: Fig. S1).

Finally, we compared the different technologies' coverages of genomic regions corresponding to coding transcripts and lncRNAs. Surprisingly, RRBS covered only slightly more transcripts than 450k (63.4% vs 60.4% of the LNCipedia transcripts and 78.5% vs 78.2% of the Ensembl transcripts). In contrast, we were able to map 850k probes to 75.6% of the LNCipedia transcripts and 86.1% of the Ensembl transcripts (Fig. 2E). Together, our results show that the 850k technology covers more promoters, enhancers, and transcribed sequences than 450k or RRBS.

Normalisation of the 850k array

Like the 450k array, 850k assesses methylated and unmethylated cytosine signals with either two beads emitting light in the same colour channel (Infinium I assay) or one bead emitting light in two different colour channels (Infinium II assay). As the use of two different assay types has been shown to introduce a bias into 450k array data, methods for correcting this bias have been developed(11, 12). We profiled with 850k arrays three replicates of the well-characterised HCT116 human colon cancer wild type (WT) cell line and of its double knock-out (DKO) derivative, in which DNA methylation is strongly reduced because of deletion of the *DNMT1* and *DNMT3B* DNA methyltransferase genes(43). We plotted the beta-value densities separately for the Infinium I and Infinium II assays and observed, as previously shown for 450k data(11), a shift of the unmethylated and methylated signals between the Infinium I and Infinium II assays (Additional File 1: Fig. S2A). When we compared the between-replicate variabilities for each probe, we observed a greater variance for Infinium II probes than for Infinium I probes (median standard deviation = 0.034 for Infinium II vs 0.013 for Infinium I) (Additional File 1: Fig. S2B). Together, these results demonstrate that the use of two different assay types introduces the same bias into 850k data as previously observed with 450k data.

To correct the observed bias of 850k, we applied the normalisation methods developed for the 450k array. Adequate normalisation strategies should both reduce between-replicate and between-technology variance, getting closer to the true value(12). To evaluate the impact of normalisation methods on intra- and inter-technology variance, we computed the absolute pairwise difference for each probe between each 850k replicate and i) the two other 850k replicates (Fig. 3A, *red*), ii) the three replicates of 450k (*green*), and iii) the two replicates of RRBS (*blue*). As shown in the white box of Fig. 3A, we observed in the absence of any correction (raw data) a lesser difference between replicates of 850k and 450k (median = 0.026) than between replicates of 850k and RRBS (median = 0.088). We then tested various within-array normalisation methods (the 'normal exponential convolution model using out-of-band intensities' (NOOB) for correction of the background and the 'peak-based correction' (PBC), the 'subset quantile for within-array normalisation' (SWAN), the 'beta-mixture quantile normalisation' (BMIQ) and the 'regression on correlated probes method' (RCP) for normalisation between type I and type II probes), several methods developed to simultaneously correct within- and between-array biases (the 'quantile normalisation on the intensity signal followed by BMIQ' (QN + BMIQ), the preprocessing pipeline developed by Touleimat and Tost (Tost), the 'NOOB followed by functional normalisation pipeline' (NOOB + Fun) and the 'background correction and quantile normalisation method treating types I and II separately' (Dasen)), and a method developed to correct between-array bias only (local regression based normalisation (LOESS)). As shown in Fig. 3A (within-array normalisation methods in the yellow box and within/between-array normalisation methods in the purple box), all the methods except NOOB + Fun reduced the variability between 850k replicates (intra-technology variance), but only PBC, NOOB, RCP, and BMIQ reduced differences between 850k and the two other technologies (inter-technology variance) (see Fig. 3B for a specific probe).

We next compared the between-replicate variabilities of RRBS, 450k, and 850k (Fig. 3C). We observed a higher median standard deviation between replicates of 850k than of 450k. This is likely due to differences in assay-type composition (Infinium I vs Infinium II) between the 450k and 850k arrays (Additional File 1: Fig. S2B, C). The median standard deviation was slightly higher for 850k raw data than for RRBS data, although some cytosines assessed by RRBS showed higher variability than in 850k. An in-depth analysis demonstrated that the standard deviation between RRBS replicates depends strongly on cytosine coverage. Hence, great sequencing depth is needed for reliable RRBS data (Additional File 1: Fig. S2D). Importantly, when the raw 850k data were corrected with NOOB or PBC, the between-replicate standard deviation decreased to a level comparable to that of corrected 450k data. Together, these findings clearly highlight the need to normalise 850k data prior to downstream bioinformatic analysis.

Statistical tests used to identify differentially methylated CpGs often require constant variability independent of the methylation level. Divergence from this property is called variance heterogeneity. To evaluate the variance heterogeneity of 850k data, we computed the distance between the mean standard deviation, representing the expected profile in a homogeneous variance context, and a local regression model of the standard deviation along the methylation profile, reflecting the observed situation (Additional File 1: Fig. S2E, F). This analysis revealed higher variance heterogeneity for 850k than for 450k data (Fig. 3D). Notably, the normalisation methods developed for 450k did not efficiently correct the variance heterogeneity of the 850k data (Fig. 3D, *pink* rectangles). Some methods (NOOB and SWAN) increased variance at unmethylated sites (Additional File 1: Fig. S2E), others (Dasen or Tost) at methylated sites (Additional File 1: Fig. S2F). PBC was the only method found to reduce variance heterogeneity along the entire profile, but variance heterogeneity remained higher for the corrected 850k than for the corrected 450k data (Fig. 3D *green*). We therefore recommend that 850k users employ a statistical approach less sensitive to variance heterogeneity (*e.g.* a non-parametric test or Welch's version of the t-test) and that they be cautious of a higher risk of false positive results. We did not investigate RRBS data, as their non-continuous nature allows the use of statistical tests that do not require variance homogeneity.

Differential methylation analysis with 850k

Finally, we used the 850k array to perform differential methylation analyses. In addition to HCT116 DKO and HCT116 WT cells, we profiled four fresh-frozen biopsies from normal human breast tissue and ten from human breast tumours. The raw 850k data were normalised by PBC, because this was the only method shown, in our previous tests, to improve both data reproducibility and variance heterogeneity. We then compared the methylation profiles of HCT116 DKO vs HCT116 WT cells and primary breast cancer tissues vs normal breast tissues.

The 850k profiles of HCT116 DKO and HCT116 WT cells revealed 336,934 differentially methylated cytosines covering 41.9% of the array, 97.5% of which were hypomethylated in HCT116 DKO cells (Additional File 1: Fig. S3A). The differentially methylated probes covered all regulatory-region categories. They included both CpGs common to 850k and 450k and CpGs specific to 850k (Additional File 1: Fig. S3B), and it would seem that the CpGs covered only by 850k can reflect 5mC changes as efficiently as the CpGs already targeted by the 450k array. Striking differences in annotation of the differentially methylated probes were observed according to whether we used our ENCODE-based or the default Illumina approach. ENCODE-based annotation identified aberrant DNA methylation mostly in enhancers, whereas Illumina's annotation highlighted promoters as most affected.

Comparison of the 850k profiles of primary breast cancer and normal breast tissues highlighted 5,617 differentially methylated CpGs, the majority of which were hypermethylated in tumours (Fig. 4A). Of the differentially methylated CpGs, 33.9% (1,905 probes) were associated with 850k-specific non-promoter sites and would thus have been missed with the 450k array (Fig. 4B). Again we observed striking differences in annotation according to the approach used. With Illumina's default annotation, alterations appeared mainly in promoter regions (Fig. 4C). This is likely due to over-estimation of promoters by this annotation. The ENCODE-based annotation revealed differential methylation mainly in enhancers (particularly of 850k-specific CpGs). Lastly, we found our annotation to associate more transcripts (6 times as many lncRNA transcripts [5,416 *versus* 900] and 1.7 times as many other transcripts [17,969 *versus* 10,656]) with differentially methylated CpGs than Illumina's default annotation (Fig. 4D). Together, these findings suggest that our new 850k processing approach, based on refined probe annotation and normalisation, identifies aberrant enhancer methylation as a dominant feature in breast cancer by allowing for a more global view into the DNA methylome of breast cancer (Fig. 5).

Discussion

Illumina's 850k array has become popular for efficient, affordable genome-scale differential methylation analysis on large patient cohorts(44–46). Shortly after its release, Moran et al. (2016) validated its performance, demonstrating high consensus between 850k and 450k data for technical replicates and for fresh-frozen and formalin-fixed paraffin-embedded samples(39). A major drawback of this study is that it was based on Illumina's default probe annotation, which underestimates the number of enhancers covered by the 850k array because it relies on enhancers identified by the restrictive CAGE method (FANTOM5 database). Surprisingly, up to now most studies still rely on Illumina's default probe annotation. To overcome this limitation, we have developed an alternative probe annotation for the 850k array, based on the enhancers defined by ENCODE, identified through a less restrictive approach: ChIP of enhancer-specific histone marks such as H3K4me1 and H3K27ac(47, 48). Using ENCODE has enabled us to increase more than three-fold the number of enhancer-associated 850k array probes and thus to improve the power of this array to study dysregulation of enhancer methylation. Unlike Zhou et al. (2017), who also proposed a basic association between 850k probes and ENCODE chromatin states(40), our 850k re-annotation provides both cell-line-specific categories and a summary of categories for each probe, greatly improving the interpretability of data. More importantly, we have used EnhancerAtlas to associate target transcripts with a large proportion of the enhancers covered by 850k, making it possible to associate methylation changes at enhancers with genes and biological processes.

Our new annotation has also improved the association of probes with promoters. Investigators are still striving to provide a uniform, meaningful, genomic-distance-based definition of a promoter, and previous studies have applied different definitions. Illumina's default annotation provides four categories that can be associated with promoters: 'TSS 1500', 'TSS 200', '5' UTR', and '1st Exon'. Sandoval et al. (2011) merged these four categories into a single promoter region(49). Dedeurwaerder et al. (2011) used an alternative, more restrictive definition, including only the 'TSS 1500' and 'TSS 200' categories and completely ignoring the intragenic part of the promoter(11). Here we have used the ENCODE HMM-based definition of promoters, because a promoter defined experimentally on the basis of histone marks is expected to be closer to the 'true' biological promoter(50). A comparison of our ENCODE-based promoters with genomic-distance-based promoters suggests that the definitions based on Illumina's default categories, as proposed by both Sandoval et al. and Dedeurwaerder et al., are too permissive, particularly for CpGs located further away from a TSS. We reveal a range from − 600 to + 1200 surrounding the TSS, within which discrepancies between the distance-based and ENCODE-based promoter definitions are low. We suggest that this range should be used as promoter definition when only genomic distance is available.

As DNA methylation changes corresponding to noncoding RNAs, particularly lncRNAs, can affect key cancer pathways(33, 34, 51), we have aimed to improve annotation of 850k probes to noncoding transcripts. Using the LNCipedia database, we have tripled the number of lncRNAs linked to 850k probes. Notably, among the transcripts corresponding to 850k probes, we have identified more than 90,000 eRNAs, a type of lncRNA whose synthesis is controlled by enhancers and which is thought to be important in enhancer-promoter looping(52). In addition, using an updated transcriptome version from Ensembl (v93), we have increased by 10% the number of coding-transcript-associated 850k probes. This has enabled us to multiply by 1.6 the number of coding transcripts identified as being differentially methylated between normal and breast cancer tissue.

With assessment of more than 800,000 CpGs, the 850k array covers about 3% of all CpGs in the human genome (more than 28 million). Coverage by RRBS ranges from 1.8 to 71.4%, depending on the sequencing depth(53). The publicly available 'RRBS Ewing' data set used in this study includes methylation values for 4.8 million CpGs, i.e. six times as many CpGs as the 850k data. Yet measuring methylation with RRBS has several limitations. First, as methylation value precision depends on sequencing depth, it is recommended to filter out low-coverage CpGs. Here we have followed ENCODE recommendations, keeping only CpGs covered by at least 10 reads(54). This reduces the number of targeted CpGs by 50 to 60%. Secondly, we show that RRBS suffers from high sequence-coverage-dependent between-replicate variability and from inconsistent coverage of CpGs between experiments. In the 'RRBS Ewing' data, out of 1.6 to 2.29 million sites having survived read-coverage and sex chromosome filtering, only 1.2 million were covered in triplicate. In the 850k data, in contrast, only 24 CpGs were not available in triplicate. Altogether, the 'RRBS Ewing' data covered 1.55 times as many CpGs as the 850k array.

Our analysis further reveals differences in the types of regulatory genomic regions covered by the three technologies. As compared to 450k, the 850k array shows improved coverage of all ENCODE-HMM-derived categories, especially the 'Enhancer' and 'Gene body' categories. Enhancer-located CpGs are in fact the dominant feature of 850k (46% of all probes). RRBS, on the other hand, predominantly covers CpGs associated with CGIs, promoters, and gene bodies. Overall, 850k targets twice as many enhancers as 450k and three times as many as RRBS (cf. Figure 2D). Surprisingly, 850k also targets more promoters than either 450k or RRBS. Yet when interpreting 850k profiles, one must keep in mind that many regulatory regions are covered by only one CpG. RRBS targets fewer promoters and enhancers than 850k (or even 450k), but with several CpGs per region. The two technologies thus provide slightly different pictures of the methylome: RRBS assesses fewer regions but with several CpGs per region, while 850k provides a broader view at the cost of fewer CpGs per region. It is important to consider this difference when assessing methylation at promoters and in distal regulatory elements (e.g. enhancers), as methylation can vary across a region(55). Pidsley et al. have shown that a single 850k probe is not always informative for distal regulatory elements with variable methylation(56). Yet this same study demonstrated that more than 80% of the distal regulatory elements targeted by a single 850k probe accurately represented DNA

methylation across the entire region. Thus, teams aiming to investigate regulatory regions with 850k are advised to use an independent technology to interrogate or validate methylation patterns across critical regions of interest.

As previously shown for the 450k array(12), the 850k array generates biases, as it also relies on two different probe types (Infinium I and II). The 850k-specific probes are primarily Infinium II probes, which are less accurate, less reproducible, and considerably less sensitive for detection of extreme methylation values than Infinium I probes. Using normalisation methods developed for the 450k array, we have succeeded in minimising 850k biases. Among the within-array normalisation strategies examined, methods correcting for differences in methylation distribution modes between Infinium I and II probes (PBC, BMIQ and RCP(11, 57, 58)) and NOOB, a method using 'out-of-bounds' intensities as negative controls to correct for background(59), improved 850k data reproducibility both within and across technologies, as demonstrated previously on 450k data(12, 60). In our data set, methods separating data into subgroups to normalise them independently (SWAN, Tost and Dasen(61–63)) improved 850k data reproducibility within but not across technologies, likely because they depend strongly on the validity of the defined subgroups. This shows a potential systematic bias of these normalisation methods and highlights the necessity of always comparing to another technology when evaluating a normalisation method(12). It is worth stressing that some normalisation methods work on the assumption that most of the array does not change. The hypotheses underlying such methods can be invalid when major changes in methylation are investigated. For instance, HCT116 DKO cells are, by design, strongly depleted of methylated CpGs. Also, highly abnormal methylation patterns can arise through pharmacological DNMT1 inhibition(21) or alteration of the methylation machinery, as observed in cancer(64). It is important to be aware of the risk of bias. On the basis of our comparison of different normalisation methods, we advise 850k users to apply within-array normalisation and preferably PBC, as this method was the only one shown to reduce variance heterogeneity. As variation between arrays is increased in larger cohorts(65), between-array normalisation may be considered, but one should always compare the before- and after-normalisation methylation density profiles to avoid data distortion.

We have further used the 850k array to characterise breast cancer methylomes. The majority of differentially methylated CpGs were detected with 850k-specific probes and would have been missed with the 450k array. On the basis of our new annotation, we have found the largest proportion of 850k-specific probes showing changes in methylation to be associated with non-promoter regions, specifically enhancers. This finding is in agreement with sequencing-based observations on breast and other cancers (66–68). It contrasts with the results obtained with Illumina's default annotation, which highlighted promoters as the most affected category (cf. Figure 4C). Use of our new annotation also led to a substantial increase in the number of noncoding transcripts found to undergo changes in DNA methylation in breast cancer. Indeed, 850k used with the new annotation identifies aberrant enhancer methylation as a dominant feature in breast cancer, but at lower cost, with less input material, and with a higher throughput than WGBS. The 850k array, if its data are analysed with our improved pipeline based on re-annotation and normalisation, is thus the technology best suited for investigating changes in DNA methylation in regulatory regions (enhancers, promoters, lncRNA genes) and coding regions in diseases requiring profiling of large patient cohorts.

Conclusions

We here propose an alternative annotation for Illumina's 850k array, associating its probes with promoters and enhancers identified by ENCODE. We have also used two complementary databases (Ensembl and LNCipedia) to improve association of 850k probes with both coding and noncoding transcripts. Our annotation results in fewer non-annotated probes than Illumina's default annotation and in a higher number of enhancer-associated probes. Even though the total number of CpGs analysed with the 850k array is lower than the number analysed with RRBS, its CpGs are distributed over more regulatory regions, especially enhancers and promoters. This makes the 850k array the technology of choice for studying methylation changes in diseases requiring profiling of large patient cohorts. Furthermore, normalisation improves both the between-replicate and between-technology reproducibility of 850k data. Lastly, with our new annotation, we identify aberrant enhancer methylation as a dominant feature in breast cancer, as reported in the literature. In conclusion, our findings suggest that the 850k array, together with our new annotation, allows improved high-throughput, low-cost analysis of DNA methylation at promoters, enhancers, lncRNA genes, and coding regions.

Methods

Samples and DNA extraction

Wild-type and double-knockout HCT116 cells (respectively called HCT116 WT and HCT116 DKO) were cultured in triplicate in McCoy's 5A medium supplemented with 10% foetal calf serum at 37°C under 5% CO2. Genomic DNA was extracted with the QIAamp DNA Mini Kit (Qiagen, Hilden, Germany) and with the recommended proteinase K and RNase A digestions. Ten fresh-frozen breast tumour samples and four normal breast tissue samples, previously profiled with the 450k array(69), were obtained from patients diagnosed with breast cancer at the Jules Bordet Institute between 1995 and 2003, following approval by the Medical Ethics Committee of Institute Jules Bordet, Brussels, Belgium. All patients gave written informed consent before their participation in the study. Genomic DNA from frozen 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 proteinase K digestion at 55°C overnight. DNA was quantified with the NanoDrop® ND-1000 UV–Vis Spectrophotometer (NanoDrop Technologies, Wilmington, DE, USA).

Bisulfite conversion and DNA methylation profiling with the 450k and 850k arrays

Genomic DNA (800 ng) treatment with sodium bisulfite was done with the Zymo EZ DNA Methylation KitTM (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. Bead chip arrays were scanned on Illumina Scan and raw .idat files were generated.

Array data analysis

All analyses were done with R (v3.4.4) except CpG annotation, which was done with python (v2.7.15).

1. Data loading

Depending on the normalisation method, two different loading methods were used. The read.metharray function of the minfi bioconductor R package (v1.24.0) was used to load data into a minfi-specific RGChannelSet object when normalisation methods requiring this object were used. For the other normalisation methods, raw .idat files of 450k data were loaded with the methylumi bioconductor R package (v2.24.1), while 850k data were loaded with the illuminaio package (v0.20.0). The detection p-values were obtained with Genome Studio® software (v1.6) and probes with detection p-values ≥ 0.05 were filtered out. Data quality was checked visually with control probes and all samples passed this quality control.

2. Data normalisation

The following five methods were used to normalise the data:

i) Peak-based correction (PBC)(11), the preprocessing pipeline developed by Touleimat and Tost (Tost)(61), and the background correction and quantile normalisation method(62) treating types I and II separately (Dasen) were run with the wateRmelon package and array-specific annotation files provided by Illumina ('MethylationEPIC_v-1-0_B4.csv': http://emea.support.illumina.com/downloads/infinium-methylationepic-v1-0-product-files.html and 'HumanMethylation450_15017482_v.1.2.csv': http://emea.support.illumina.com/array/array_kits/infinium_humanmethylation450_beadchip_kit/downloads.ht

ii) The normal exponential convolution model using out-of-band intensities (NOOB)(59), the subset quantile for within-array normalisation (SWAN)(63), and the NOOB followed by functional normalisation pipeline (NOOB+Fun)(65) were run with the appropriate functions of the minfi package (*i.e.* preprocessSWAN, preprocessNOOB and preprocessFunnorm, respectively).

iii) The regression on correlated probes method (RCP)(57) was run with the method of the ENmix bioconductor R package (v1.14) after applying the preprocessRaw method of the minfi package.

iv) The pipeline proposed by Marabita et al.(60), consisting in quantile normalisation on the intensity signal followed by beta-mixture quantile normalisation (QN+BMIQ), was run with the lumiMethyN function of the lumi package and the BMIQ function (v 1.1) (code available at https://code.google.com/archive/p/bmiq/).

v) LOESS between-array normalisation from Heiss et al.(70) was adapted from the R code provided with the associated paper.

3. Probe filtering

After normalisation, ambiguous probes from both sex chromosomes were discarded. Cross-reactive probes identified by Price et al.(71) were filtered out of the 450k data and the annotation of McCartney et al.(72) was used to filter out cross-reactive 850k probes. Breast cancer data were also filtered against probes targeting SNPs identified by Price et al. (450k) or McCartney et al. (850k). Raw data (.idat) and preprocessed data were submitted to the Gene Expression Omnibus public database (GEO) (www.ncbi.nlm.nih.gov/geo/) (token will be provided once GEO approved data).

Reduced-representation bisulfite sequencing (RRBS) data

Two already-processed public RRBS datasets were used with different aims:

i) For CpG coverage comparisons, processed RRBS data for three Ewing sarcoma samples were downloaded from GEO (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE89026)(73). CpGs located in both sex chromosomes or covered by less than 10 reads were filtered out. Only CpGs whose methylation status was available for all three samples were analysed.

ii) For the comparison of methylation levels, RRBS data on HCT116 cell samples from two independent experiments were downloaded from ENCODE (https://www.encodeproject.org/experiments/ENCSR000DFS/). To simplify the comparison with bead-array technology, strand specificity was not taken into account. When methylation values were available for both strands of the same CpG site, the values were averaged if the two values were similar (delta < 10%) and discarded otherwise. Only cytosines common to the 450k and 850k arrays and covered by at least 10 reads were kept for subsequent analysis. In order to maximise the number of methylation-level comparisons between 850k and RRBS, cytosines whose methylation levels were available for only one sample were kept in this dataset.

CpG annotation

CpGs were annotated with human genome build hg19. Ensembl positions were lifted from hg38 to hg19 with the 'LiftOver' tool from UCSC (https://genome.ucsc.edu/cgi-bin/hgLiftOver).

1. CpG islands

CpG island (CGI) positions were retrieved from the UCSC database (https://genome.ucsc.edu/cgi-bin/hgTables). CGI shores were generated by adding 2 kb at both ends of each CGI. 850k, 450k, and RRBS data were overlapped with CGIs and Shores, using a custom python script. CpGs not located in a CGI or a Shore region were annotated as 'Open-sea'.

2. Regulatory regions

Regulatory genomic regions were retrieved with the 'UCSC ENCODE Experiment Matrix tool' (https://genome.ucsc.edu/ENCODE/dataMatrix/encodeDataMatrixHuman.html) for all available cell lines (HMEC, HSMM, K562, NHEK, NHLF, HEPG2, HUVEC, HESC, and GM12878). For each cell line, the hidden Markov model provided by ENCODE classifies the whole genome into 15 chromatin states. In order to simplify this annotation, we grouped the '1_Active_Promoter', '2_Weak_Promoter', and '3_Poised_Promoter' states into a 'Promoter' super-region while '4_Strong_Enhancer', '5_Strong_Enhancer', '6_Weak_Enhancer' and '7_Weak_Enhancer' states were assigned to the 'Enhancer' super-region. The other states were ignored. States from the same super-region succeeding each other in the genome were fused. RRBS, 850k and 450k data were then overlapped with the 'Promoter' and 'Enhancer' super-regions in each cell line. As a CpG can be associated with a 'Promoter' super-region in one cell line and an 'Enhancer' in another, we assigned such CpGs to a third category called 'Dual' to reflect their dual role.

3. Transcripts

Coding and noncoding transcript positions were retrieved from the LNCipedia high-confidence set version 5.2 and from Ensembl v93. Duplicates between the two databases were identified with the 'lncipedia_5_2_hc_hg19.gff' file from LNCipedia and only the LNCipedia version of each duplicated lncRNA was kept. Several CpG-to-transcript association types were investigated:

i) Association with the transcription start site (TSS): each transcript whose TSS was located in a 'Promoter' super-region was assigned to that promoter and associated with the CpGs of that promoter. If a TSS for an lncRNA transcript fell into an 'Enhancer' super-region, the transcript was associated with that region and defined as an eRNA (enhancer RNA). The list of 'Promoter'/'Enhancer'-transcript associations is available in Table S1 (Additional File 2).

ii) Enhancer targets: To identify targets of the 'Enhancer' super-regions we used the EnhancerAtlas database (data downloaded in June 2016). This database associates its own set of enhancer positions with Ensembl transcripts for a set of 73 cell lines. For each 'Enhancer' super-region position from ENCODE overlapping a position from EnhancerAtlas in the same cell line, we associated the target provided by EnhancerAtlas with this 'Enhancer'. Remaining 'Enhancers' were characterised as having an unknown target.

iii) Gene body association: Each CpG assessed by RRBS, 850k or 450k was also intersected with LNCipedia and Ensembl transcript positions. The transcription start site (TSS) and transcription termination site (TTS) corresponding to each transcript were located on the genome. CpGs falling between a transcript-associated TSS and the corresponding TTS were described as 'Gene body associated'. If such a CpG was not already classified as 'Promoter', 'Enhancer', or 'Dual', it was classified as 'Gene body', while the remaining CpGs were assigned to the 'Intergenic' category. This annotation is available as alternative platform annotation on GEO (token will be provided once GEO approved data). The list of data source URLs is available in Table S2 (Additional File 3).

4. Promoter and non-promoter regions

Analyses using only the 'Promoter' and 'Non-promoter' categories were carried out with the following grouping: 'Promoter' and 'Dual' cytosines were classified as 'Promoter' cytosines while 'Enhancer', 'Gene body', and 'Intergenic' cytosines were defined as 'Non-promoter' cytosines.

5. Illumina default annotation

To compare the Illumina default annotation with our annotation, we collapsed the different levels of information provided by the Illumina 850k annotation file into the five categories used in our custom annotation. Associations with known transcripts were extracted from the 'UCSC_RefGene_Group' and 'GencodeCompV12_Group' columns. CpGs associated with 'TSS 1500', 'TSS 200', '1st Exon' and '5' UTR' locations were categorised as 'Promoter' CpGs, while the remaining CpGs, associated with 'Body' and '3' UTR', were categorised as 'Gene body' CpGs. Enhancer-associated CpGs were retrieved from the 'Phantom4_Enhancers', 'Phantom5_Enhancers', and '450k_Enhancer' columns. These CpGs were categorised as 'Enhancer' or 'Dual' CpGs, depending on their association with the 'Promoter' category. The remaining CpGs were categorised as 'Intergenic'.

Variance heterogeneity evaluation

Variance heterogeneity was assessed with a metric we called 'variance heterogeneity measurement' (*h*), which can be described as the distance between the mean standard deviation (representing the expected profile in a context of homogeneous variance) and a local regression model of the standard deviation along the methylation profile (reflecting the observed situation). It was computed as follows:

- The M value was computed for each probe with the formula described elsewhere(11).

- The standard deviation (s_p) and the mean (m_p) of the M-value of each probe across the HCT116 WT triplicates were computed so as to generate the vectors *s* and *m*.

- The observed profile was modelled with a local regression (loess) model fitting *s* as a function of *m* and a loess-smoothed value of the standard deviation was produced for each probe (s_p^*).

- In a context of homogeneous variance, *s* should be independent of *m,* so the expected profile should be modelled as a flat line at s_0, the mean of *s.*

- A measurement of the variance heterogeneity (h_p) was computed for each probe as the absolute difference between s_p^* (value from the observed model) and s_0 (expected value in a homogeneous context).

- Finally, *h* was defined as the mean of all h_p values.

Differential methylation analysis

Differential methylation was assessed with a t-test applied to the M-values. In parallel, a delta beta value was computed as the absolute difference between the median beta value within each category. CpGs showing an adjusted p-value (Benjamini-Hochberg correction) < 0.05 together with an absolute delta beta > 0.2 were reported as differentially methylated.

Abbreviations

27k: HumanMethylation27 array

450k: HumanMethylation450 array

5mC: DNA methylation

850k: MethylationEPIC array

BMIQ: 'beta-mixture quantile normalisation'

CAGE: Cap analysis of gene expression

CGI: CpG Island

ChIP: chromatin immunoprecipitation

CpG: cytosine-guanine dinucleotide

Dasen: 'background correction and quantile normalisation method treating types I and II separately'

DKO: DNMT1, DNMT3B double knock-out

DNMT1: DNA methyltransferase 1

DNMT3B: DNA methyltransferase 3B

ENCODE: Encyclopedia of DNA elements

ENCODE-HMM: hidden Markov-model provided by ENCODE

eRNA: enhancer RNA

FANTOM5: Functional annotation of mammalian genome version 5

GEO: Gene expression omnibus

lncRNA: long noncoding RNA

loess: local regression

LOESS: local regression based normalisation

NOOB: 'normal exponential convolution model using out-of-band intensities'

NOOB+Fun: 'NOOB followed by functional normalisation pipeline' (NOOB+Fun)

PBC: 'peak-based correction'

QN+BMIQ: ‘quantile normalisation on the intensity signal followed by BMIQ’

RCP: ‘regression on correlated probes method’

RRBS: reduced-representation bisulfite sequencing

SWAN: ‘subset quantile for within-array normalisation’

TCGA: The Cancer Genome Atlas

Tost: preprocessing pipeline developed by Touleimat and Tost

TSS: transcription start site

TTS: transcription termination site

WGBS: whole-genome bisulfite sequencing

WT: wild type

Declarations

Ethics approval and consent to participate

Breast specimens were obtained from the tissue bank at the Jules Bordet Institute, following approval by the Medical Ethics Committee of Jules Bordet Institute, Brussels, Belgium. All patients gave written informed consent before their participation in the study.

Consent for publication

Not applicable

Availability of data and materials

Microarray data that support the findings of this study have been deposited in GEO with the accession codes GSE (token will be provided once GEO approved data). All other data supporting the findings of this study are available from the corresponding author on reasonable request.

Competing interests

F.F. is a cofounder of Epics Therapeutics, Belgium. The other authors declare no competing interests.

Funding

This study was supported by the Belgian ‘Fonds de la Recherche Scientifique’ (FNRS), Télévie, the ‘Action de Recherche Concertée’ (ARC; AUWB-2018-2023 ULB-No 7), Wallon Region grants (U-CAN-REST, INTREPID), an FNRS Welbio grant, the ULB Foundation and the Belgian Foundation Against Cancer (FCC 2016-086 FAF-F/2016/872).

Author’s contributions

MB, MD and JJ conceived and designed the study. EC performed the experiments. MB analyzed the data. CS provided patient samples. MB, MD and JJ interpreted data. MB and JJ wrote the manuscript. MD, GB, CS and FF critically reviewed the manuscript. All authors read and approved the final manuscript.

Acknowledgements

We thank Françoise Rothe and Delphine Vincent (Breast Cancer Translational Research Laboratory, Jules Bordet Institute, Brussels, Belgium) for their assistance in tissue collection. We thank the tissue donors, whose help and participation made this work possible. We thank Alexander Koch for his critical review and helpful input on the manuscript.

References

1. Allis CD, Jenuwein T. The molecular hallmarks of epigenetic control. Nat Rev Genet. 2016 Aug;17(8):487–500.
2. Law JA, Jacobsen SE. Establishing, maintaining and modifying DNA methylation patterns in plants and animals. Nat Rev Genet. 2010 Mar;11(3):204–20.

3. Jeschke J, Collignon E, Fuks F. DNA methylome profiling beyond promoters – taking an epigenetic snapshot of the breast tumor microenvironment. The FEBS Journal. 2015;282(9):1801–14.
4. Lister R, Pelizzola M, Dowen RH, Hawkins RD, Hon G, Tonti-Filippini J, et al. Human DNA methylomes at base resolution show widespread epigenomic differences. Nature. 2009 Nov;462(7271):315–22.
5. The Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. Nature. 2012 Oct;490(7418):61–70.
6. Boyle P, Clement K, Gu H, Smith ZD, Ziller M, Fostel JL, et al. Gel-free multiplexed reduced representation bisulfite sequencing for large-scale DNA methylation profiling. Genome Biology. 2012 Oct 3;13(10):R92.
7. Klughammer J, Kiesel B, Roetzer T, Fortelny N, Nemc A, Nenning K-H, et al.The DNA methylation landscape of glioblastoma disease progression shows extensive heterogeneity in time and space. Nat Med. 2018 Oct;24(10):1611–24.
8. Bock C, Tomazou EM, Brinkman AB, Müller F, Simmer F, Gu H, et al. Quantitative comparison of genome-wide DNA methylation mapping technologies. Nat Biotechnol. 2010 Oct;28(10):1106–14.
9. Bibikova M, Le J, Barnes B, Saedinia-Melnyk S, Zhou L, Shen R, et al. Genome-wide DNA methylation profiling using Infinium ® assay. Epigenomics. 2009 Oct;1(1):177–200.
10. Bibikova M, Barnes B, Tsan C, Ho V, Klotzle B, Le JM, et al. High density DNA methylation array with single CpG site resolution. Genomics. 2011 Oct 1;98(4):288–95.
11. Dedeurwaerder S, Defrance M, Calonne E, Denis H, Sotiriou C, Fuks F. Evaluation of the Infinium Methylation 450K technology. Epigenomics. 2011 Dec 1;3(6):771–84.
12. Dedeurwaerder S, Defrance M, Bizet M, Calonne E, Bontempi G, Fuks F. A comprehensive overview of Infinium HumanMethylation450 data processing. Briefings in Bioinformatics. 2014 Nov 1;15(6):929–41.
13. Nordlund J, Bäcklin CL, Wahlberg P, Busche S, Berglund EC, Eloranta M-L, et al. Genome-wide signatures of differential DNA methylation in pediatric acute lymphoblastic leukemia. Genome Biology. 2013 Sep 24;14(9):r105.
14. Horvath S. DNA methylation age of human tissues and cell types. Genome Biology. 2013 Dec 10;14(10):3156.
15. Dayeh T, Volkov P, Salö S, Hall E, Nilsson E, Olsson AH, et al. Genome-Wide DNA Methylation Analysis of Human Pancreatic Islets from Type 2 Diabetic and Non-Diabetic Donors Identifies Candidate Genes That Influence Insulin Secretion. PLoS Genet. 2014 Mar 6;10(3):e1004160.
16. Jeschke J, Bizet M, Desmedt C, Calonne E, Dedeurwaerder S, Garaud S, et al. DNA methylation–based immune response signature improves patient diagnosis in multiple cancers. J Clin Invest. 2017 Aug 1;127(8):3090–102.
17. Ochoa-Rosales C, Portilla-Fernandez E, Nano J, Wilson R, Lehne B, Mishra PP, et al.Epigenetic Link Between Statin Therapy and Type 2 Diabetes. Diabetes Care. 2020 Feb 7;43(4):875–84.
18. Lee H-S, Park T. The influences of DNA methylation and epigenetic clocks, on metabolic disease, in middle-aged Koreans. Clinical Epigenetics. 2020 Oct 15;12(1):148.
19. Baylin SB, Jones PA. Epigenetic Determinants of Cancer. Cold Spring Harb Perspect Biol. 2016 Jan 9;8(9):a019505.
20. Pasculli B, Barbano R, Parrella P. Epigenetics of breast cancer: Biology and clinical implication in the era of precision medicine. Semin Cancer Biol. 2018 Aug;51:22–35.
21. Jones PA, Issa J-PJ, Baylin S. Targeting the cancer epigenome for therapy. Nat Rev Genet. 2016 Oct;17(10):630–41.
22. Berdasco M, Esteller M. Clinical epigenetics: seizing opportunities for translation. Nat Rev Genet. 2019 Feb;20(2):109–27.
23. Klutstein M, Nejman D, Greenfield R, Cedar H. DNA Methylation in Cancer and Aging. Cancer Res. 2016 Jun 15;76(12):3446–50.
24. Smith ZD, Meissner A. DNA methylation: roles in mammalian development. Nat Rev Genet. 2013 Mar;14(3):204–20.
25. Köhler F, Rodríguez-Paredes M. DNA Methylation in Epidermal Differentiation, Aging, and Cancer. J Invest Dermatol. 2020 Jan 1;140(1):38–47.
26. Ehrlich M. DNA hypermethylation in disease: mechanisms and clinical relevance. Epigenetics. 2019 Dec 2;14(12):1141–63.
27. Glass JL, Hassane D, Wouters BJ, Kunimoto H, Avellino R, Garrett-Bakelman FE, et al. Epigenetic Identity in AML Depends on Disruption of Nonpromoter Regulatory Elements and Is Affected by Antagonistic Effects of Mutations in Epigenetic Modifiers. Cancer Discov. 2017 Aug;7(8):868–83.
28. Koldobskiy MA, Abante J, Jenkinson G, Pujadas E, Tetens A, Zhao F, et al. A Dysregulated DNA Methylation Landscape Linked to Gene Expression in MLL-Rearranged AML. Epigenetics. 2020 Aug 2;15(8):841–58.
29. Heyn H, Vidal E, Ferreira HJ, Vizoso M, Sayols S, Gomez A, et al. Epigenomic analysis detects aberrant super-enhancer DNA methylation in human cancer. Genome Biology. 2016 Jan 26;17(1):11.
30. Aran D, Sabato S, Hellman A. DNA methylation of distal regulatory sites characterizes dysregulation of cancer genes. Genome Biology. 2013 Mar 12;14(3):R21.
31. Yao L, Shen H, Laird PW, Farnham PJ, Berman BP. Inferring regulatory element landscapes and transcription factor networks from cancer methylomes. Genome Biology. 2015 May 21;16(1):105.
32. Wilusz JE, Sunwoo H, Spector DL. Long noncoding RNAs: functional surprises from the RNA world. Genes Dev. 2009 Jan 7;23(13):1494–504.
33. Gutschner T, Diederichs S. The hallmarks of cancer: A long non-coding RNA point of view. RNA Biology. 2012 Jun;9(6):703–19.
34. Van Grembergen O, Bizet M, de Bony EJ, Calonne E, Putmans P, Brohée S, et al.Portraying breast cancers with long noncoding RNAs. Sci Adv. 2016 Sep 2;2(9):e1600220.
35. Alsaleh H, Haddrill PR. Identifying blood-specific age-related DNA methylation markers on the Illumina MethylationEPIC® BeadChip. Forensic Science International. 2019 Oct;303:109944.

36. Zaimi I, Pei D, Koestler DC, Marsit CJ, De Vivo I, Tworoger SS, et al. Variation in DNA methylation of human blood over a 1-year period using the Illumina MethylationEPIC array. Epigenetics. 2018 Nov 2;13(10–11):1056–71.
37. McEwen LM, Jones MJ, Lin DTS, Edgar RD, Husquin LT, MacIsaac JL, et al. Systematic evaluation of DNA methylation age estimation with common preprocessing methods and the Infinium MethylationEPIC BeadChip array. Clinical Epigenetics. 2018 Oct 16;10(1):123.
38. The FANTOM Consortium, Andersson R, Gebhard C, Miguel-Escalada I, Hoof I, Bornholdt J, et al. An atlas of active enhancers across human cell types and tissues. Nature. 2014 Mar;507(7493):455–61.
39. Moran S, Arribas C, Esteller M. Validation of a DNA methylation microarray for 850,000 CpG sites of the human genome enriched in enhancer sequences. Epigenomics. 2016 Mar;8(3):389–99.
40. Zhou W, Laird PW, Shen H. Comprehensive characterization, annotation and innovative use of Infinium DNA methylation BeadChip probes. Nucleic Acids Res. 2016 Oct 24;gkw967.
41. Volders P-J, Anckaert J, Verheggen K, Nuytens J, Martens L, Mestdagh P, et al. LNCipedia 5: towards a reference set of human long non-coding RNAs. Nucleic Acids Research. 2019 Jan 8;47(D1):D135–9.
42. Zerbino DR, Achuthan P, Akanni W, Amode MR, Barrell D, Bhai J, et al. Ensembl 2018. Nucleic Acids Research. 2018 Jan 4;46(D1):D754–61.
43. Rhee I, Bachman KE, Park BH, Jair K-W, Yen R-WC, Schuebel KE, et al. DNMT1 and DNMT3b cooperate to silence genes in human cancer cells. Nature. 2002 Apr 4;416(6880):552–6.
44. Talukdar FR, Lima SCS, Khoueiry R, Laskar RS, Cuenin C, Sorroche BP, et al. Genome-Wide DNA Methylation Profiling of Esophageal Squamous Cell Carcinoma from Global High-Incidence Regions Identifies Crucial Genes and Potential Cancer Markers. Cancer Res. 2021 May 15;81(10):2612–24.
45. Salas LA, Zhang Z, Koestler DC, Butler RA, Hansen HM, Molinaro AM, et al. Enhanced cell deconvolution of peripheral blood using DNA methylation for high-resolution immune profiling. Nat Commun. 2022 Feb 9;13(1):761.
46. Chen J-Q, Salas LA, Wiencke JK, Koestler DC, Molinaro AM, Andrew AS, et al. Immune profiles and DNA methylation alterations related with non-muscle-invasive bladder cancer outcomes. Clinical Epigenetics. 2022 Jan 21;14(1):14.
47. Akhtar-Zaidi B, Cowper-Sal·lari R, Corradin O, Saiakhova A, Bartels CF, Balasubramanian D, et al. Epigenomic Enhancer Profiling Defines a Signature of Colon Cancer. Science [Internet]. 2012 May 11 [cited 2022 Jan 15]; Available from: https://www.science.org/doi/abs/10.1126/science.1217277
48. Almamun M, Levinson BT, van Swaay AC, Johnson NT, McKay SD, Arthur GL, et al. Integrated methylome and transcriptome analysis reveals novel regulatory elements in pediatric acute lymphoblastic leukemia. Epigenetics. 2015 Sep 2;10(9):882–90.
49. Sandoval J, Heyn H, Moran S, Serra-Musach J, Pujana MA, Bibikova M, et al. Validation of a DNA methylation microarray for 450,000 CpG sites in the human genome. Epigenetics. 2011 Jun 1;6(6):692–702.
50. Ernst J, Kheradpour P, Mikkelsen TS, Shoresh N, Ward LD, Epstein CB, et al.Mapping and analysis of chromatin state dynamics in nine human cell types. Nature. 2011 May 5;473(7345):43–9.
51. Li Y, Zhang Y, Li S, Lu J, Chen J, Wang Y, et al.Genome-wide DNA methylome analysis reveals epigenetically dysregulated non-coding RNAs in human breast cancer. Sci Rep. 2015 Mar 5;5(1):8790.
52. Li W, Lam MT, Notani D. Enhancer RNAs. Cell Cycle. 2014 Oct 15;13(20):3151–2.
53. Laird PW. Principles and challenges of genomewide DNA methylation analysis. Nat Rev Genet. 2010 Mar;11(3):191–203.
54. De Meyer T, Bady P, Trooskens G, Kurscheid S, Bloch J, Kros JM, et al.Genome-wide DNA methylation detection by MethylCap-seq and Infinium HumanMethylation450 BeadChips: an independent large-scale comparison. Sci Rep. 2015 Oct 20;5(1):15375.
55. Koch A, Joosten SC, Feng Z, de Ruijter TC, Draht MX, Melotte V, et al.Analysis of DNA methylation in cancer: location revisited. Nat Rev Clin Oncol. 2018 Jul;15(7):459–66.
56. Pidsley R, Zotenko E, Peters TJ, Lawrence MG, Risbridger GP, Molloy P, et al. Critical evaluation of the Illumina MethylationEPIC BeadChip microarray for whole-genome DNA methylation profiling. Genome Biol. 2016 Dec;17(1):208.
57. Niu L, Xu Z, Taylor JA. RCP: a novel probe design bias correction method for Illumina Methylation BeadChip. Bioinformatics. 2016 Sep 1;32(17):2659–63.
58. Teschendorff AE, Marabita F, Lechner M, Bartlett T, Tegner J, Gomez-Cabrero D, et al. A beta-mixture quantile normalization method for correcting probe design bias in Illumina Infinium 450 k DNA methylation data. Bioinformatics. 2013 Jan 15;29(2):189–96.
59. Triche TJ, Weisenberger DJ, Van Den Berg D, Laird PW, Siegmund KD.Low-level processing of Illumina Infinium DNA Methylation BeadArrays. Nucleic Acids Res. 2013 Apr;41(7):e90.
60. Marabita F, Almgren M, Lindholm ME, Ruhrmann S, Fagerström-Billai F, Jagodic M, et al. An evaluation of analysis pipelines for DNA methylation profiling using the Illumina HumanMethylation450 BeadChip platform. Epigenetics. 2013 Mar;8(3):333–46.
61. Touleimat N, Tost J. Complete pipeline for Infinium ® Human Methylation 450K BeadChip data processing using subset quantile normalization for accurate DNA methylation estimation. Epigenomics. 2012 Jun;4(3):325–41.
62. Pidsley R, Y Wong CC, Volta M, Lunnon K, Mill J, Schalkwyk LC. A data-driven approach to preprocessing Illumina 450K methylation array data. BMC Genomics. 2013 Dec;14(1):293.
63. Maksimovic J, Gordon L, Oshlack A. SWAN: Subset-quantile Within Array Normalization for Illumina Infinium HumanMethylation450 BeadChips. Genome Biol. 2012;13(6):R44.
64. Pfister SX, Ashworth A. Marked for death: targeting epigenetic changes in cancer. Nat Rev Drug Discov. 2017 Apr;16(4):241–63.
65. Fortin J-P, Labbe A, Lemire M, Zanke BW, Hudson TJ, Fertig EJ, et al. Functional normalization of 450k methylation array data improves replication in large cancer studies. 2014;17.

66. Batra RN, Lifshitz A, Vidakovic AT, Chin S-F, Sati-Batra A, Sammut S-J, et al. DNA methylation landscapes of 1538 breast cancers reveal a replication-linked clock, epigenomic instability and cis-regulation. Nat Commun. 2021 Sep 13;12(1):5406.
67. Huang P, Xu M, Han H, Zhao X, Li MD, Yang Z. Integrative Analysis of Epigenome and Transcriptome Data Reveals Aberrantly Methylated Promoters and Enhancers in Hepatocellular Carcinoma. Frontiers in Oncology [Internet]. 2021 [cited 2022 Feb 23];11. Available from: https://www.frontiersin.org/article/10.3389/fonc.2021.769390
68. Wilson ER, Helton NM, Heath SE, Fulton RS, Payton JE, Welch JS, et al. Focal disruption of DNA methylation dynamics at enhancers in IDH-mutant AML cells. Leukemia. 2021 Dec 6;1–11.
69. Dedeurwaerder S, Desmedt C, Calonne E, Singhal SK, Haibe-Kains B, Defrance M, et al. DNA methylation profiling reveals a predominant immune component in breast cancers. EMBO Mol Med. 2011 Dec;3(12):726–41.
70. Heiss JA, Brenner H. Between-array normalization for 450K data. Front Genet [Internet]. 2015 Mar 10 [cited 2022 Jan 15];6. Available from: http://www.frontiersin.org/Epigenomics_and_Epigenetics/10.3389/fgene.2015.00092/abstract
71. Price EM, Cotton AM, Lam LL, Farré P, Emberly E, Brown CJ, et al. Additional annotation enhances potential for biologically-relevant analysis of the Illumina Infinium HumanMethylation450 BeadChip array. Epigenetics & Chromatin. 2013 Dec;6(1):4.
72. McCartney DL, Walker RM, Morris SW, McIntosh AM, Porteous DJ, Evans KL. Identification of polymorphic and off-target probe binding sites on the Illumina Infinium MethylationEPIC BeadChip. Genom Data. 2016 May 26;9:22–4.
73. Sheffield NC, Pierron G, Klughammer J, Datlinger P, Schönegger A, Schuster M, et al. DNA methylation heterogeneity defines a disease spectrum in Ewing sarcoma. Nat Med. 2017 Mar;23(3):386–95.

Figures

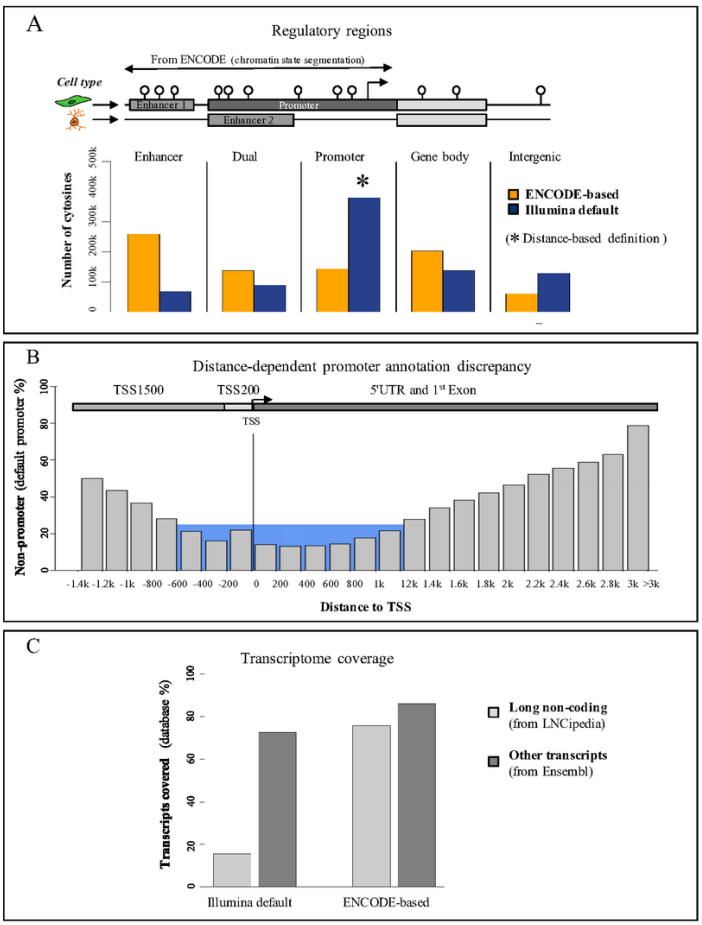

Figure 1 Bizet *et al.*

Figure 1

Reannotation of the 850k bead chip: (A) Barplot of the number of 850k-assessed cytosines associated with 'Enhancer', 'Dual', 'Promoter', 'Gene body', and 'Intergenic' regions using ENCODE-based (*orange*) and Illumina default (*blue*) annotations The ENCODE-based annotation shows better coverage of all regulatory regions ('Enhancer', 'Dual', and 'Gene Body') except 'Promoter' regions (see the * mark), (probably because of an imprecise promoter definition in the Illumina default annotation). *Top*: lollipop scheme of the assessed regions. Cytosines are represented as white lollipops. The scheme shows two example cell

types. Enhancers and promoters, identified through ENCODE chromatin state segmentation, are shown, respectively, in *medium grey* and *dark grey*. The gene body is represented in *light grey*. 'Enhancer' and 'Promoter' regions are defined as regions where only enhancers or promoters, respectively, can be identified through ENCODE chromatin states across all investigated cell lines. 'Dual' regions are associated with a promoter in some cell lines and with an enhancer in others. The TSS is represented as an *arrow*. (B) Barplot of percentages of probes associated with promoters by the Illumina annotation but not the ENCODE-based annotation as a function of distance to TSS. The *grey* bars represent percentages of probes showing a discrepancy between the Illumina and ENCODE-based annotations within a 200-bp window based on distance to the TSS. The *blue* region represents distance to TSS where the discrepancy is lower than 30%.The four promoter-associated regions (according to Sandoval et al(49)) are shown as a scheme at the *top* of the figure: TSS200 (*light grey*), TSS1500 (*medium grey*), 5' UTR and 1st Exon (*dark grey*). The TSS position is specified by a vertical line in the barplot and as an arrow in the scheme. (C) Barplot of the percentage of transcripts associated with at least one 850k-targeted cytosine, according to the reference annotation (Illumina default *left*, ENCODE-based *right*) and transcriptomic database (LNCipedia *light grey*, Ensembl *dark grey*).

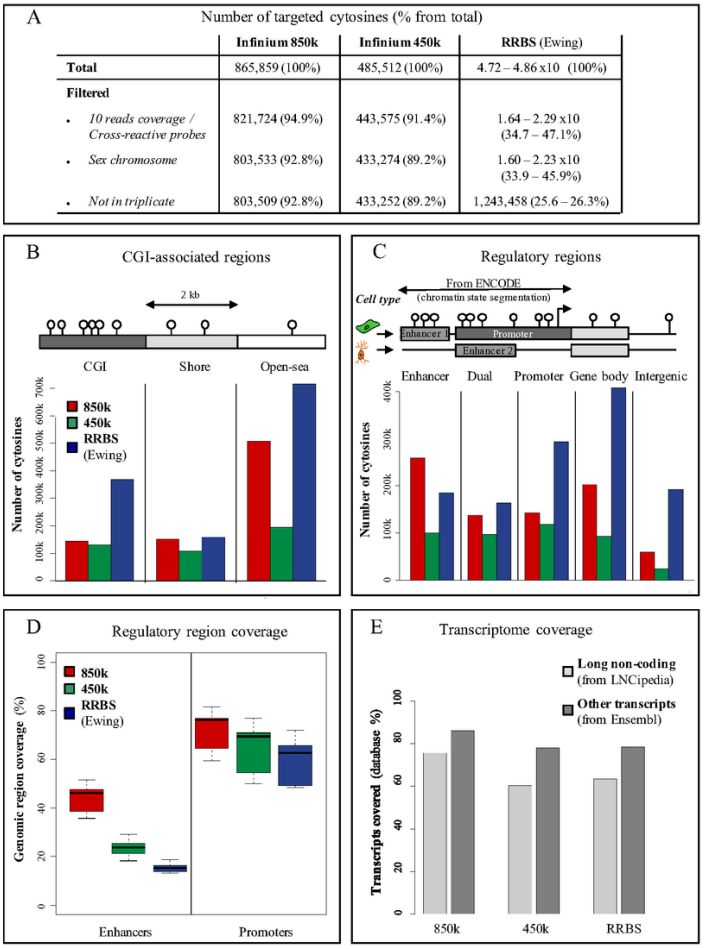

Figure 2 Bizet *et al.*

Figure 2

Coverage of the 850k bead chip as compared to 450k and RRBS: (A) Table showing the number of cytosines targeted by 850k, 450k, and RRBS through the filtering steps. Within parentheses: Percentages of the total number remaining after filtering. (B-C) Barplot of the number of cytosines covered by 850k (*red*), 450k (*green*), and RRBS (*blue*) according to the CGI-associated region (B) or regulatory region (C) where the cytosines are located. *Top*: lollipop scheme of the assessed regions. Cytosines are represented as white lollipops. (B) The scheme highlights that CGIs (*dark grey*) are CpG-dense regions, Shores (*light grey*) are regions located up to 2 kb from a CGI, and the Open sea (*white*) contains the remaining parts of the genome. (C) The scheme shows two example cell types. Enhancers and promoters, identified through ENCODE chromatin state segmentation, are shown, respectively, in medium grey and dark grey. A gene body (from Ensembl or LNCipedia databases) is represented in light grey. 'Enhancer' and 'Promoter' regions are defined as regions where only enhancers or promoters are identified, respectively, through use of ENCODE chromatin states across all investigated cell lines, while 'Dual' regions are associated with a promoter in some cell lines and with an enhancer in others. The TSS is represented as a *black* arrow. (D) Proportion of regulatory regions (enhancers *left*, promoters *right*) covered by at least one cytosine targeted by 850k (*red*), 450k (*green*), or RRBS (*blue*). Each boxplot represents the distribution of the coverage among the nine cell lines provided by ENCODE. (E) Barplot of the percentage of transcripts associated with at least one cytosine targeted by 850k (*left*), 450k (*middle*), or RRBS (*right*), according to the transcriptomic database (LNCipedia *light grey*, Ensembl *dark grey*).

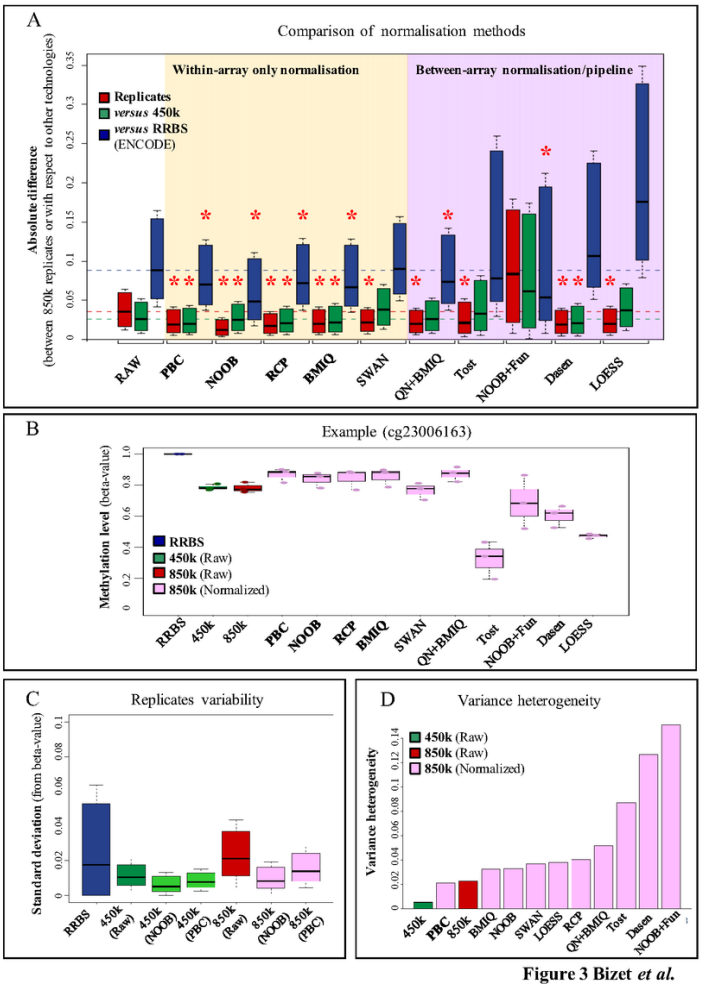

Figure 3 Bizet *et al.*

Figure 3

Evaluation of 850k-bias correction methods: (A) Boxplots showing the distribution of absolute differences between DNA methylation measurements obtained with Infinium 850k from three replicates of HCT116 WT cells (*red*) or between the 850k array and the 450k array (*green*) or between 850k and RBBS (*blue*), when the data are not normalised (*white* background), exclusively within-array normalised (*orange* background), or between-array normalised (possibly in a pipeline also including within-array normalisation) (*purple* background). (B) Example of a probe impacted by the normalisation method. The boxplots show the distribution of cytosine methylation levels assessed on RRBS duplicates (*blue*), 450k triplicates (raw measurement) (*green*), or 850k triplicates (raw measurement in *red*, normalized measurement in *pink*). The RRBS experiment shows a fully methylated cytosine, the raw 450k and 850k data show a beta-value at 1.0 . While some methods (*e.g* PBC, NOOB, and RCP) lead to data more similar to RRBS, others (*e.g* Dasen, LOESS) distort the data toward a hemi-methylated level. (C) Boxplots showing the distribution of the standard deviation obtained upon cytosine methylation level assessment with duplicate RRBS sequencing data (*blue*) or triplicate 850k-array (*red*) or 450k-array (*green*) measurements. The 450k and 850k data were subjected (*dark-coloured*) or not (*light-coloured*) to NOOB or PBC normalisation. (D) Barplot showing levels of variance heterogeneity for HCT116 cell line methylation data: raw 450k data (*dark green*); 850k data subjected (*pink*) or not (*dark red*) to normalisation. RAW: raw Infinium data; PBC: peak-based correction from the wateRmelon package; NOOB: Normal exponential convolution using out-of-bounds; RCP: regression on correlated probes method; SWAN: Subset quantile Within-Array Normalisation from the minfi package; QN+BMIQ: pipeline formed by quantile normalisation on intensities followed by beta-mixture quantile normalisation; Tost: categorical SQN from the Touleimat and Tost pipeline; NOOB+Fun: pipeline composed of NOOB correction followed by functional normalisation; Dasen: Dasen pipeline from the wateRmelon package. LOESS: local-regression between-array normalisation.

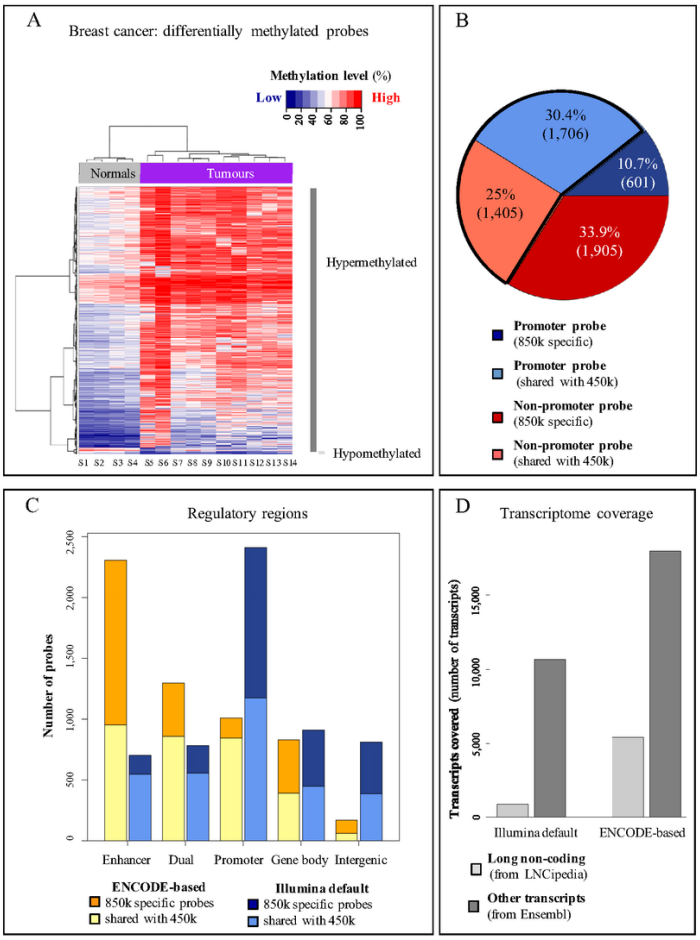

Figure 4 Bizet *et al.*

Figure 4

Differential methylation analysis of breast tumours vs normal samples with 850k: (A) Heatmap of the differentially methylated probes identified on the 850k array upon comparing breast tumour samples (samples S5 to S14, *purple* rectangle) with normal breast tissue samples (samples S1 to S4, *grey* rectangle). The methylation level is represented on a *blue* (unmethylated) to *red* (methylated) scale. Hypermethylated and hypomethylated probes are highlighted respectively a *dark grey* and a *light grey* vertical bar. (B) Pie chart of the proportion of differentially methylated promoter and non-promoter probes. *Blue*: promoter, *red*: non-promoter, *light*: probes common to the 850k and 450k versions, *dark*: probes specific to the 850k array. The part surrounded in black contains probes common to the two versions. (C) Barplot of the number of differentially methylated cytosines that can be assessed exclusively by 850k (*dark*) or that are common to 850k and 450k (*light*). The differentially methylated cytosines are associated with 'Enhancer', 'Dual', 'Promoter', 'Gene body', and 'Intergenic' regions according to the ENCODE-based (*orange*) or Illumina default annotation (*blue*). The ENCODE-based annotation shows a lower proportion of differentially methylated cytosines not associated with any feature (*i.e.* 'Intergenic'). This leads to better interpretability of results. (D) Barplot showing the number of transcripts associated with at least one differentially methylated cytosine, according to the reference annotation (Illumina default *left*, ENCODE-based *right*) and the transcriptomic database (LNCipedia *light grey*, Ensembl *dark grey*). It highlights strong improvement in the number of transcripts identified with the ENCODE-based annotation as compared to Illumina's default annotation.

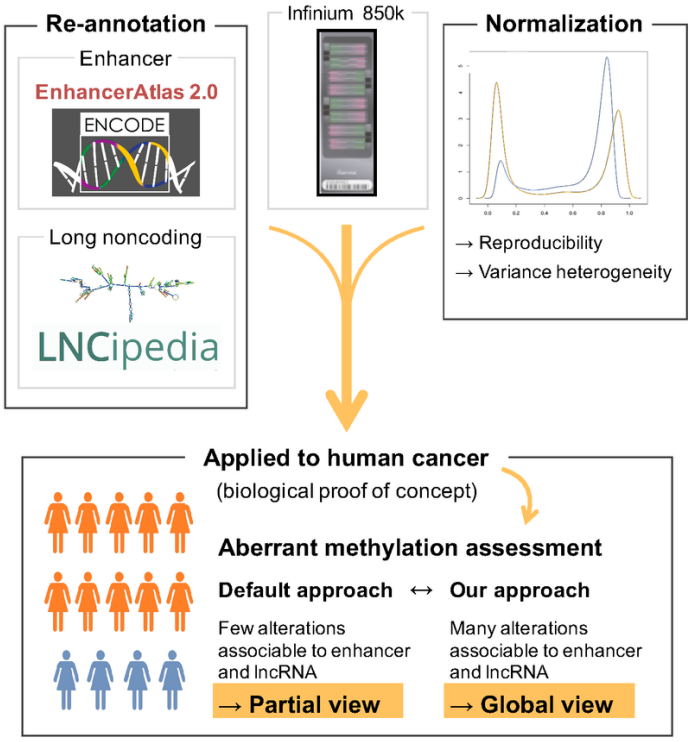

Figure 5 Bizet *et al.*

Figure 5

An updated processing approach for 850k data:

Our new processing approach for 850k data, based on refined probe annotation and normalisation, allows for improved analysis of DNA methylation at enhancers and long noncoding RNA genes. This approach highlights, as previously reported, aberrant enhancer methylation as a dominant feature in breast cancer. Thus, the 850k array, together with our new processing approach, allows for improved high-throughput, low-cost analysis of DNA methylation in clinical samples.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- AdditionalFile1SupplementaryFiguresBizetetal.pdf
- AdditionalFile2TableS1Bizetetal.xlsx
- AdditionalFile3TableS2Bizetetal.xlsx