A systematic study of normalization methods for Infinium 450K methylation data using whole-genome bisulfite sequencing data

Ting Wang1,1, Weihua Guan2,1, Jerome Lin3, Nadia Boutaoui1, Glorisa Canino4, Jianhua Luo5, Juan Carlos Celedón1, and Wei Chen1,3,*

1Division of Pulmonary Medicine; Allergy and Immunology; Department of Pediatrics; Children’s Hospital of Pittsburgh of UPMC; University of Pittsburgh; Pittsburgh, PA USA; 2Department of Biostatistics; University of Minnesota; Minneapolis, MN USA; 3Department of Human Genetics; University of Pittsburgh Graduate School of Public Health; Pittsburgh, PA USA; 4Behavioral Sciences Research Institute; University of Puerto Rico; San Juan, Puerto Rico; 5Department of Pathology; University of Pittsburgh; Pittsburgh PA USA

DNA methylation plays an important role in disease etiology. The Illumina Infinium HumanMethylation450 (450K) BeadChip is a widely used platform in large-scale epidemiologic studies. This platform can efficiently and simultaneously measure methylation levels at ~480,000 CpG sites in the human genome in multiple study samples. Due to the intrinsic chip design of 2 types of chemistry probes, data normalization or preprocessing is a critical step to consider before data analysis. To date, numerous methods and pipelines have been developed for this purpose, and some studies have been conducted to evaluate different methods. However, validation studies have often been limited to a small number of CpG sites to reduce the variability in technical replicates. In this study, we measured methylation on a set of samples using both whole-genome bisulfite sequencing (WGBS) and 450K chips. We used WGBS data as a gold standard of true methylation states in cells to compare the performances of 8 normalization methods for 450K data on a genome-wide scale. Analyses on our dataset indicate that the most effective methods are peak-based correction (PBC) and quantile normalization plus β-mixture quantile normalization (QN.BMIQ). To our knowledge, this is the first study to systematically compare existing normalization methods for Illumina 450K data using novel WGBS data. Our results provide a benchmark reference for the analysis of DNA methylation chip data, particularly in white blood cells.

Introduction

DNA methylation, where a methyl group is added to a cytosine or adenine DNA nucleotide, may result in the regulation of gene expression.1 DNA methylation influences numerous biological processes, such as development and cell differentiation, and also plays a crucial role in the progress of complex diseases, including cancer and asthma.2,3 Therefore, exploring the relation between this epigenetic variation and various phenotypes is increasingly favored in scientific research.

Nowadays, Illumina has developed the Infinium HumanMethylation450 (450K) BeadChip (450K) for profiling large-scale DNA methylation at single base resolution by genotyping the C/T polymorphism after sodium bisulfite treatment of DNA.4 Unlike the previous version (27K), this new array incorporates 2 different chemical assays: (1) Infinium I (nD135,501), which uses a single color with 2 different probes to generate methylated (M) and unmethylated (U) measurements for a specific CpG locus; and (2) Infinium II (nD350,076), which uses a single probe with 2 different colors to obtain M and U signal intensities for a CpG locus. Although the 450K chip covers a large number of CpG methylation loci and, thus, can be used in high-throughput methylocmic studies, such a special probe design causes potential problems for data analysis. A previous report showed that Type I and Type II probes usually have different distributions of methylation values, and that Type II probes are relatively less accurate and sensitive for detecting extreme methylation values.5

Several methods and pipelines have been developed to preprocess the 450K array data, in order to adjust for probe type or color bias, subtract background signals, and eliminate systematic errors. For example, quantile normalization (QN), a well-established technique in analyzing gene expression data,6 is often used...
to analyze the Illumina methylation array data. QN replaces the signal intensity of a probe with the mean intensity of the probes that have the same rank from all studied arrays, and thus makes the distribution of probe intensities from each array the same. In general, raw data should be preprocessed by color balance adjustment and background correction before QN. Yousefi et al. proposed a reference factor-based color channel normalization method called all sample mean normalization (ASMN), which used the mean of red/green control probe intensities of all samples to generate the reference normalization factors for the specific color channel and then normalized each sample accordingly. Dedeurwaerder et al. proposed a peak-based correction (PBC) method to estimate the M-value peaks for Infinium I and II independently, and rescale the Infinium II values to match the (PBC) method to estimate the M-value peaks for Infinium I and II independently, and rescale the Infinium II values to match the *PBC* method to estimate the M-value peaks for Infinium I and II independently, and rescale the Infinium II values to match the *PBC* method to estimate the M-value peaks for Infinium I and II independently, and rescale the Infinium II values to match the *PBC* method to estimate the M-value peaks for Infinium I and II independently, and rescale the Infinium II values to match the *PBC* method to estimate the M-value peaks for Infinium I and II independently, and rescale the Infinium II values to match the *PBC* method to estimate the M-value peaks for Infinium I and II independently, and rescale the Infinium II values to match the *PBC* method to estimate the M-value peaks for Infinium I and II independently, and rescale the Infinium II values to match the *PBC* method to estimate the M-value peaks for Infinium I and II independently, and rescale the Infinium II values to match the *PBC* method to estimate the M-value peaks for Infinium I and II independently, and rescale the Infinium II values to match the *PBC* method to estimate the M-value peaks for Infinium I and II independently, and rescale the Infinium II values to match the *PBC* method to estimate the M-value peaks for Infinium I and II independently, and rescale the Infinium II values to match the *PBC* method to estimate the M-value peaks for Infinium I and II independently, and rescale the Infinium II values to match the *PBC* method to estimate the M-value peaks for Infinium I and II independently, and rescale the Infinium II values to match the *PBC* method to estimate the M-value peaks for Infinium I and II independently, and rescale the Infinium II values to match the *PBC* method to estimate the M-value peaks for Infinium I and II independently, and rescale the Infinium II values to match the *PBC* method to estimate the M-value peaks for Infinium I and II independently, and rescale the Infinium II values to match the *PBC* method to estimate the M-value peaks for Infinium I and II independently, and rescale the Infinium II values to match the *PBC* method to estimate the M-value peaks for Infinium I and II independently, and rescale the Infinium II values to match the *PBC* method to estimate the M-value peaks for Infinium I and II independently, and rescale the Infinium II values to match the *PBC* method to estimate the M-value peaks for Infinium I and II independently, and rescale the Infinium II values to match the *PBC* method to estimate the M-value peaks for Infinium I and II independently, and rescale the Infinium II values to match the *PBC* method to estimate the M-value peaks for Infinium I and II independently, and rescale the Infinium II values to match the *PBC* method to estimate the M-value peaks for Infinium I and II independently, and rescale the Infinium II values to match the *PBC* method to estimate the M-value peaks for Infinium I and II independent...
Comparison of CpG methylation differences associated with phenotypes

The four samples, measured by both WGBS and 450K platforms in our data set, originated from 2 monozygotic twin pairs. In every pair of twins, one had asthma while the other one did not. We next compared the performances of these normalization methods in downstream differential methylation analysis. For each twin pair, we computed methylation differences by subtracting the methylation values of healthy samples from the values of asthmatic twin samples, and calculated PCCs of methylation differences between different normalization data and between each twin in the pair. The clustering diagram in Figure 2A shows that data from the same twin pair clustered together, suggesting that biological variation between samples are more significant than that introduced by different normalization methods. Furthermore, we calculated the methylation differences using the WGBS data, and compared the PCCs between WGBS data and normalized 450K data for each twin pair. The results show that the correlations of QN.BMIQ data rank top 3 in both sample pairs (Fig. 2B and C).

Next, we ranked the differentially methylated CpGs according to their absolute methylation differences, and counted the number of top CpGs that were overlapped in both WGBS and 450K data. We used 10 cutoffs from 2000 to 20000 to extract the top CpGs for comparisons. The results showed that the PBC normalized array data had the largest
number of differentially methylated CpGs overlapping with those from the sequencing data in both twin pairs (Fig. 2D and E).

Comparisons based on methylation regions

DNA methylation is usually studied in the context of DNA regions (such as CpG islands, shores, and shelves) rather than single CpG sites, largely because a group of adjacent CpGs that have concordant differential methylation patterns may indicate a robust and significant methylation event associated with phenotype. On the other hand, the SQN method performs normalization based on different probe categories related to CpGs. Therefore, we aggregated CpGs methylation values into regional levels and compared them in the context of regions. According to 450K platform annotations, the measured CpG sites, which were contained in both 450K and WGBS data, belonged to 25,733 CpG islands, 11,536 N_shelves, 20,059 N_shores, 11,003 S_shelves, and 17,893 S_shores. We calculated the average \( b \)-value of the CpGs located in a specific CG region to be the regional methylation level and calculated their average sequencing coverage to be the regional coverage. We then compared those different normalization approaches with previous single-site based analyses. The correlations between WGBS data and 450K data normalized by PBC, BMIQ, and QN.BMIQ were larger, and their mean absolute differences were smaller, than after normalization by other methods. SQN still performed the worst, especially in the context of CpG islands and shores (Fig. 3).

Assessment of reduction of variability in technical replicates

Using the 2 pairs of technical replicates in our data set, we assessed the reduction in variability between replicates following normalization with each method. Figure 4A shows that the PCCs of duplicates from PBC and QN.BMIQ-normalized 450K data are larger than for other methods, while those from ASMN and SQN-normalized data are actually smaller than those of raw data. Consistent results are shown in the MADs of duplicates: QN and QN.BMIQ data showed smaller MADs in replicate samples, while ASMN data showed even larger differences between technical replicates than original raw data (Fig. 4B). These results are similar to those reported by Marabita et al and Dedeuwaerder et al. studies, where they also preferred the QN.BMIQ and PBC methods.

Reduction of probe type bias

An important objective of normalizing the Illuma 450K data is to reduce the bias caused by the use of the 2 types of probes. Therefore, we drew density plots of methylation \( b \)-values for each sample using raw and normalized 450K data (Fig. 5), where the a red line represents the densities of Type I probes for a sample and a blue line represents the densities of Type II probes for a sample. It is shown that QN, ASMN, Dasen, and SWAN do not fully remove probe type bias; in contrast,
QN, BMIQ, BMIQ, PBC, and SQN eliminate the bias between the 2 probe types. These results are similar to those reported by other researchers. However, for SQN, although the bias is reduced, the correlation/difference between the WGBS and its normalized 450K data, and that between technical replicates, is severely compromised, which was already shown in the above analyses. Hence, PBC and QN. BMIQ could be the most efficient normalization methods.

**Discussion**

The Illumina Infinium 450K arrays are now widely used in profiling large-scale intracellular DNA methylation levels in human epigenomic studies. Although many analytic methods and pipelines for data from this platform have been implemented, a consensus on preprocessing methods has not been
reached. In this work, we used whole-genome bisulfite sequencing (WGBS) methylation data as gold standard to compare and evaluate the performance of 8 normalization methods for the 450K data on a genome-wide scale. In particular, microarray data from the 450K platform were obtained from 64 white blood cell samples, among which 4 samples were also measured by WGBS. WGBS data and differently normalized 450K data were compared to evaluate the performance of different normalization approaches. Our results suggest that PBC and QN.BMIQ normalization methods may outperform others when analyzing the Illumina Infinium 450K data.

Similar conclusions have been drawn by other studies. For example, Marabita et al. found the QN plus BMIQ method most effective among their 6 compared methods; Dedeurwaerder et al. recommended the PBC method according to their comparative analyses; Wu et al. found that raw data performed better than several normalization approaches. Our observations are consistent with those reports, although from a novel perspective. However, Maksimovic et al. and Touleimat et al. reported that PBC did not work well when the methylation level distribution did not exhibit well-defined peaks. Teschendorff et al. also showed that PBC could lead to discontinuities in Type II density distribution. So the generalizability of our conclusions may need to be evaluated on data from other tissue types.

The significance of our work lies in using the WGBS data as a gold standard to evaluate the performance of normalizing 450K array data on a genome-wide scale. Previous reports were mainly focused on just a few CpG sites for sequencing-based validation, such as the small bisulfite pyrosequencing (BPS) data sets employed by Dedeurwaerder et al. and Marabita et al. WGBS can capture cytosine methylation states on a whole-genome scale with high accuracy, especially when the sequencing depth is high. Therefore, we generated a whole-genome DNA methylation data set with the WGBS technology, and used it as a gold standard to evaluate the corresponding 450K data normalized by different methods. To our knowledge, this is the first study to systematically compare normalization methods for the Illumina 450K methylation data using WGBS data.

We acknowledge several limitations of our study. First, the sequencing depth in our experiment is modest, ~13× coverage per CpG site on average. Although we applied a higher cutoff to obtain more accurate measures from sequencing, this reduced the number of CpG sites to be compared. Also, our WGBS validation data set was small, and DNA methylation was only measured in white blood cells. The conclusions may need further confirmation on other tissue data. We also did not evaluate the impact of sequencing errors, which could bias the methylation levels measured by WGBS. However, considering the high cost of WGBS at the current time, the Illumina 450K methylation platform will remain a popular technology for large-scale DNA methylation profiling. Hence, our results will be valuable for the analysis of the 450K array data.

**Materials and Methods**

**Illumina Infinium HumanMethylation450 data**

We quantified DNA methylation levels in white blood cells from 64 children using the Illumina Infinium 450K platform. High quality DNA (750 ng) was bisulfite converted using EZ DNA methylation kit (Zymo Research, Irvine, CA). Bisulfite converted DNA (200 ng) was used for analysis of whole-genome methylation using the HumanMethylation 450K BeadChip.
In brief, bisulfite converted DNA was whole-genome amplified for 23 hours followed by end-point fragmentation. Fragmented DNA was precipitated, denatured, and hybridized to the BeadChips for 18 hours at 48°C. The BeadChips were washed and the hybridized primers were extended and labeled prior to scanning the BeadChips using the Illumina iScan system.

We read methylation data from the raw IDAT files using an R package named methylumi and calculated the $b$-value for each CpG as $b = M/(M + U + \alpha)$, where $M$ and $U$ represent methylated and unmethylated signal intensities at the specific site and $\alpha$ is an arbitrary offset (usually 100) intended to stabilize $b$-values where fluorescent intensities are low. Sometimes, people use another index called M-value to quantify the DNA methylation level, which is calculated as $\log_2 [(M + \alpha)/(U + \alpha)]$. In essence, the M-value is equivalent to a logit transformation of the $b$-value. Note that some normalization methods are performed on $b$-values and some are based on M-values.

Whole-genome bisulfite sequencing data

Four samples were used in the methyl-seq experiment. First, 750 ng of DNA were bisulfite bisulfite-converted using the EZ DNA methylation-lightning kit. The concentration of bisulfite-converted DNA was measured using quantifluor single stranded DNA system (Promega, Grand Island, NY). The epigenome library was amplified (10 cycles) and then purified using AMPure XP system. The library was then quantified using high sensitivity DNA Chip (Agilent Technologies, Santa Clara, CA). The 200-cycle paired-end sequencing experiment was performed by a university facility core, which used Illumina HiSeq2000 and followed a standard protocol from the manufacturer’s protocol. In brief, cDNA was synthesized using random hexamer with tagging sequence. After digesting excess random primer, terminal tagging oligo (TTO) was added to the cDNA. The tagged cDNA was purified using AMPure XP system (Life Technologies, Grand Island, NY). The epigenome library was amplified (10 cycles) and then purified using AMPure XP system. The library was then quantified using high sensitivity DNA Chip (Agilent Technologies, Santa Clara, CA). The 200-cycle paired-end sequencing experiment was performed by a university facility core, which used Illumina HiSeq2000 and followed a standard protocol from the manufacturer’s manual.

We utilized a tool named Bismark v0.12.5 to map the WGBS reads to human genome hg19 and perform methylation calls.20 We then counted the number of methylated (observed as C) and unmethylated (observed as T) cytosines at each CpG and calculated its $b$-value as $#C/(#C + #T)$. The generated $b$ profile was regarded as the benchmark or gold standard of DNA methylation status of the 4 samples.

Normalization of the Infinium 450K data

We pooled the 450K data from 64 samples together as a data set and respectively preprocessed it with 8 representative methods (See Table 1) to produce the normalized DNA methylation profiles.

**Table 1. Normalization methods for Illumina 450K data**

| Method | Instruction | R Package |
|--------|-------------|-----------|
| QN     | Quantile normalization, 2003\(^6\) | lumi      |
| PBC    | Peak-based correction, 2011\(^5\) | wateRmelon |
| SWAN   | Subset-quantile within array normalization, 2012\(^8\) | miniFi     |
| SQN    | Subset quantile normalization, 2012\(^9\) | wateRmelon |
| BMIQ   | Beta-mixture quantile normalization, 2013\(^10\) | wateRmelon |
| ASMN   | All sample mean normalization, 2013\(^1\) | asmn       |
| QN.BMIQ| QN then BMIQ, 2013\(^13\) | lumi + wateRmelon |

We implemented color balance adjustment, background correction, and quantile normalization (QN) with the R package lumi, which was designed to process the Illumina gene expression and methylation microarray.\(^21\) The fuki, test, BMIQ and dasen functions from the wateRmelon package were respectively utilized for normalizing the 450K data, which are after color and background corrections, based on the PBC SQN, BMIQ, and Dasen methods.\(^10\) ASMN and SWAN were respectively implemented with the asmn\(^7\) and miniFi\(^11\) packages, and QN.BMIQ was achieved by combining the 2 normalization methods with the help of lumi and wateRmelon packages.\(^13\) All computational analyses were implemented with R 3.1.2 and all the packages of normalization were downloaded from Bioconductor 3.0.

From all raw and normalized 450K data, we filtered the SNP-containing probes that were annotated by the Illumina 450K platform and those unreliable probes that had at least one detectable $p$-value larger than 0.05. The remaining methylation profiles were used for our comparative analyses.

Disclosure of Potential Conflicts of Interest

No potential conflicts of interest were disclosed.

Funding

This study was supported by the Competitive Medical Research Fund (WC) from the University of Pittsburgh, and by grants HL07996 and HL117191 (JCC) from the US. National Institutes of Health.

Supplemental Material

Supplemental data for this article can be accessed on the publisher’s website.

References

1. Razin A, Riggs AD. DNA methylation and gene function. Science 1980; 210:604-10; PMID:6254144; http://dx.doi.org/10.1126/science.6254144
2. Dedeurwaerder S, Defrance M, Calonne E, Denis H, Sotiriou C, Fuks F. Evaluation of the infinium high density DNA methylation array with single CpG site resolution. Genomics 2011; 98:288-95; PMID:21839163; http://dx.doi.org/10.1016/j.ygeno.2011.07.007
3. Jones PA. Functions of DNA methylation: islands, start sites, gene bodies and beyond. Nat Rev Genet 2012; 13:484-92; PMID:22641018; http://dx.doi.org/10.1038/nrg3230
4. Dedeurwaerder S, Defrance M, Calonne E, Denis H, Sotiriou C, Fuks F. Evaluation of the infinium high density DNA methylation array with single CpG site resolution. Genomics 2011; 98:288-95; PMID:21839163; http://dx.doi.org/10.1016/j.ygeno.2011.07.007
5. Jones PA. Functions of DNA methylation: islands, start sites, gene bodies and beyond. Nat Rev Genet 2012; 13:484-92; PMID:22641018; http://dx.doi.org/10.1038/nrg3230
methyltransferase 450K technology. Epigenomics 2011; 3:771-84; PMID:22126295; http://dx.doi.org/10.2217/epi.11.105
6. Bolstad BM, Irizarry RA, Astrand M, Speed TP. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics 2003; 19:185-93; PMID:12538238; http://dx.doi.org/10.1093/bioinformatics/19.2.185
7. Yousefi P, Huen K, Schall RA, Decker A, Elboudwarej E, Quach H, Barcellou L, Holland N. Considerations for normalization of DNA methylation data by Illumina 450K BeadChip assay in population studies. Epigenetics 2013; 8:1141-52; PMID:23959097; http://dx.doi.org/10.4161/epi.26037
8. Maksimovic J, Gordon L, Oshlack A. SWAN: Subset-quantile within array normalization for Illumina HumanMethylation450 BeadChips. Genome Biol 2012; 13:R44; PMID:22703947; http://dx.doi.org/10.1186/gb-2012-13-6-r44
9. Touleimat N, Tost J. Complete pipeline for Infinium (R) Human Methylation 450K BeadChips data processing using subset quantile normalization for accurate DNA methylation estimation. Epigenomics 2012; 4:325-41; PMID:22690668; http://dx.doi.org/10.2217/epi.12.21
10. 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; 14:293; PMID:23631413; http://dx.doi.org/10.1186/1471-2164-14-293
11. Teschendorff AE, Marabita F, Lechner M, Bartlett T, Tegner J, Gomez-Calderon D, Beck S. A beta-mixture quantile normalization method for correcting probe design bias in Illumina Infinium 450 k DNA methylation data. Bioinformatics 2013; 29:189-96; PMID:23175756; http://dx.doi.org/10.1093/bioinformatics/bts680
12. Dedoussard S, Defrance M, Bizer M, Calone E, Bonnetm P, Fuks F. A comprehensive overview of Illumina HumanMethylation450 data processing. Brief Bioinform 2013; 15(6):929-41; PMID:23990268
13. Marabita F, Aimgren M, Lindholm ME, Ruhrmann S, Fagerstrom-Billai F, Jagodic M, Sundberg CJ, Ekstrom TJ, Teschendorff AE, Tegner J, et al. An evaluation of analysis pipelines for DNA methylation profiling using the Illumina HumanMethylation450 BeadChip platform. Epigenetics 2013; 8:333-46; PMID:23422812; http://dx.doi.org/10.4161/epi.24008
14. Wu MC, Jouhert BR, Kuan PF, Haberg SE, Nystad W, Peddada SD, London SJ. A systematic assessment of normalization approaches for the Illumina 450K methylation platform. Epigenetics 2014; 9:318-29; PMID:24241353; http://dx.doi.org/10.4161/epi.27119
15. Harris RA, Wang T, Cooraf C, Nagarajan RP, Hong C, Downey SL, Johnson BE, Fouse SD, Delaney A, Zhao Y, et al. Comparison of sequencing-based methods to profile DNA methylation and identification of monallelic epigenetic modifications. Nat Biotechnol 2010; 28:1097-105; PMID:20852635; http://dx.doi.org/10.1038/nbt.1682
16. Bell JT, Tsai PC, Yang TP, Pidsley R, Nibber J, Dinanga D, Mangino M, Zhao G, Zhang F, Valdes A, et al. Epigenome-wide scans identify differentially methylated regions for age and age-related phenotypes in a healthy ageing population. PLoS Genet 2012; 8:e1002629; PMID:22532803; http://dx.doi.org/10.1371/journal.pgen.1002629
17. Iriartt RA, Ladd-Acosta C, Wen B, Wu Z, Montano C, Ouyang P, Cui H, Gabo K, Rongione M, Webster M, et al. The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores. Nat Genet 2009; 41:178-86; PMID:19151715; http://dx.doi.org/10.1038/ng.298
18. Shaeta K. DNA methylation profiles of CpG islands for cellular differentiation and development in mammals. Cytogenet Genome Res 2004; 105:325-34; PMID:15237720; http://dx.doi.org/10.1159/000078265
19. Smith ZD, Gu H, Bock C, Gnutk A, Meissner A. High-throughput bisulfite sequencing in mammalian genomes. Methods 2009; 48:226-32; PMID:19442738; http://dx.doi.org/10.1016/j.ymeth.2009.05.003
20. Krueger F, Andrews SR. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. Bioinformatics 2011; 27:1571-2; PMID:21493656; http://dx.doi.org/10.1093/bioinformatics/btr167
21. Du P, Kibbe WA, Lin SM. lumi: a pipeline for processing Illumina microarray. Bioinformatics 2008; 24:1547-8; PMID:18467348; http://dx.doi.org/10.1093/bioinformatics/btn224
22. Ayre R, Jaffe AE, Corrada-Bravo H, Ladd-Acosta C, Feinberg AP, Hansen KD, Iriartt RA. Minfi: a flexible and comprehensive Bioconductor package for the analysis of Illumina DNA methylation microarrays. Bioinformatics 2014; 30:1363-9; PMID:24478359; http://dx.doi.org/10.1093/bioinformatics/btu499