METHODS
published: 27 October 2020
doi: 10.3389/fgene.2020.538492

MGMIN: A Normalization Method for Correcting Probe Design Bias in Illumina Infinium HumanMethylation450 BeadChips

*Zhenxing Wang, Yongzhuang Liu and Yadong Wang**

School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China

The Illumina Infinium HumanMethylation450 Beadchips have been widely utilized in epigenome-wide association studies (EWAS). However, the existing two types of probes (type I and type II), with the distribution of measurements of probes and dynamic range different, may bias downstream analyses. Here, we propose a method, MGMIN (*M*-values Gaussian-MIxture Normalization), to correct the probe designs based on *M*-values of DNA methylation. Our strategy includes fitting Gaussian mixture distributions to type I and type II probes separately, a transformation of *M*-values into quantiles and finally a dilation transformation based on *M*-values of DNA methylation to maintain the continuity of the data. Our method is validated on several public datasets on reducing probe design bias, reducing the technical variation and improving the ability to find biologically differential methylation signals. The results show that MGMIN achieves competitive performances compared to BMIQ which is a well-known normalization method on β-values of DNA methylation.

Keywords: DNA methylation, design bias, normalization, M-value, Gaussian mixture model, Illumina Infinium 450K

OPEN ACCESS

Edited by:
Zhongyu Wei,
Fudan University, China

Reviewed by:
Teng Zhixia,
Northeast Forestry University, China
Zhen Tian,
Zhengzhou University, China

****Correspondence:***
Yadong Wang
ydwang@hit.edu.cn

Specialty section:
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

Received: *27 February 2020*
Accepted: *21 September 2020*
Published: *27 October 2020*

Citation:
Wang Z, Liu Y and Wang Y (2020) MGMIN: A Normalization Method for Correcting Probe Design Bias in Illumina Infinium HumanMethylation450 BeadChips. Front. Genet. 11:538492. doi: 10.3389/fgene.2020.538492

1. INTRODUCTION

DNA methylation, as a well-known epigenetic marker, plays an essential role in biological processes and complex genetic diseases like cancer and diabetes (Irizarry et al., 2009; Paul et al., 2016). The Illumina Infinium HumanMethylation450 (450K) BeadChip (Bibikova et al., 2011) provides measurements of the level of methylation at over 480K CpG sites and has been widely used in epigenome-wide association studies (EWAS) and large-scale projects, such as The Cancer Genome Atlas (TCGA). The probes in the Infinium 450K BeadChip come in two different designs, type I (n = 135,501) and type II (n = 350,076), in order to increase the genomic coverage of the assay. However, the methylation values (β-values or M-values) derived from the two types of designs exhibit different distributions. Particularly, the type I probes possess a larger range of measurement than the type II probes (Dedeurwaerder et al., 2011). The differences between the two types of probe designs may impact the downstream analyses.

Several approaches have been published to correct the probe design bias. A peak-based correction (PBC) method normalizes type II probes to render them comparable with type I probes (Dedeurwaerder et al., 2011). In fact, PBC gets poor performance when the density distribution of methylation values does not show well-defined peaks. SQN (Touleimat and Tost, 2012) and SWAN (Maksimovic et al., 2012) select subset of probes with similar biological category to adjust the probe design bias. Beta MIxture Quantile dilation (BMIQ) is a model-based normalization approach to correct β-values of type II probes according to the beta distribution of β-values of type I probes, which appears to outperform PBC, SQN, and SWAN (Teschendorff et al., 2012).

In this work, we propose a method to correct the probe design bias based on the Gaussian Mixture Model (GMM) of the M-values of DNA methylation, which is called M-value Gaussian-MIxture Normalization (MGMIN). The method includes three steps: (i) fit Gaussian-mixture distributions to type I and type II probes separately, (ii) utilize a transformation of M-values into quantiles, (iii) perform a dilation transformation based on M-values to maintain the continuity of the data. We evaluate MGMIN using several independent datasets in terms of reducing the replicate technical variance and correcting the type II bias. By comparison with BMIQ, the results show that MGMIN improves the overall performance of normalization.

TABLE 1 | Comparison of MGMIN and BMIQ on detecting the differentially methylated probes (DMPs) associated with HPV status was performed by counting the number of DMPs (Dataset 2), the number of validated differentially methylated probes (nTPs) (Dataset 3: GSE38266 and Dataset 4: GSE95036) and corresponding estimates for the positive predictive value (PPV = nTP/nDMPs).

Metric	Raw	BMIQ	MGMIN
nDMP	51 (51[a])	239 (252[a])	220
nTP (GSE38266)	16 (13[a])	55 (51[a])	37
PPV (GSE38266)	0.31 (0.25[a])	0.23 (0.20[a])	0.17
nTP (GSE95036)	3	13	27
PPV (GSE95036)	0.06	0.05	0.12

[a] *Values reported in* Teschendorff et al. (2012).

2. MATERIALS AND METHODS

2.1. Measure DNA Methylation With *M*-value

The β-value of DNA methylation for each probe is defined by the ratio of the methylated intensity (M) and the overall intensity (sum of methylated intensity and unmethylated intensity: M + U):

$$\beta - value = \frac{M}{M + U + \alpha}$$

where α is a constant offset (by default, α = 100) to regularize the β-value when the overall intensity is low. The β-value falls between 0 and 1 which follows a Beta distribution naturally. A β-value of 0 indicates the CpG site of the measured sample is fully unmethylated and a value of 1 indicates that the CpG site is completely methylated.

The M-value is calculated by the log2 ratio of the methylated intensity (M) vs. the unmethylated intensity (U):

$$M - value = \log_2(\frac{M + \alpha}{U + \alpha})$$

where α here is also an offset (by default, α = 1) to counteract the big changes caused by small intensity estimation errors. An M-value close to zero indicates that the measured CpG site is about hemimethylated. A positive M-value suggests that more copies of the measured CpG site are methylated than unmethylated and a negative M-value means more copies of the CpG site are unmethylated. The M-value has been widely used in two-color expression microarray analysis (Du et al., 2010).

Due to more than 95% CpG sites have intensities more than 1,000 in Illumina methylation data, the α in β-value and M-value has an insignificant effect on observed results. So the relationship between β-value and M-value is shown as (with α ignored):

$$\beta = \frac{2^M}{2^M + 1}; M = \log_2(\frac{\beta}{1 - \beta})$$

According to the conclusions in Du et al. (2010), the M-value is more statistically valid in an analysis by modeling the distribution of M-values because of it's *homoscedastic*. So we choose to adjust the M-values of type II probes into the distribution property of type I probes to correct the probe design bias.

2.2. MGMIN: *M*-value Gaussian-MIxture Normalization

Gaussian Mixture Model (GMM) has been widely applied as a clustering method in analyzing gene-expression microarray data (Yeung et al., 2001; Pan et al., 2002) and used to detect differential gene expression (McLachlan et al., 2006). In this paper, we apply GMM to distinguish different methylation states of CpG sites for further correction. The M-values of a single 450K microarray can be viewed as a finite Gaussian mixture model of several methylation states (hypomethylated-U, hemimethylated-H, hypermethylated-F). The probability density function of the M-value for a single CpG site (M_i) is defined as:

$$p(M_i; \theta) = \sum_{k=1}^{K} \pi_k N(M_i | \mu_k, \sigma_k^2) \tag{1}$$

where $p(M_i, \theta)$ represents the model density for M_i with unknown parameter vector θ, K is the number of different methylation states (components), $N(M_i|\mu_k, \sigma_k^2)$ is the probability density function of the kth Gaussian component, and π_k is the mixing proportions which satisfy the constraint that $\sum_{k=1}^{K} \pi_k = 1$ and $0 \leq \pi_k \leq 1$. The parameter vector θ consists of the mixing proportions π_k, the mean value μ_k and the standard deviation σ_k, which can be estimated by the EM algorithm.

Next, we describe MGMIN in detail. First, M-values of type I and type II probes are modeled by GMM separately. Let μ_T^S and σ_T^S denote the mean value and standard deviation where $S \in (U, H, F)$ and $T \in (I, II)$. K_I and K_{II} are the numbers of components for type I and type II probes, which are both set as 3 by default.

Second, each probe is assigned to hypomethylated (U_T), hemimethylated (H_T), or hypermethylated (F_T) states by using the maximum probability criterion. Let U_T^L (U_T^R) denote the U_T probes with M-values smaller (larger) than μ_T^U, and let F_T^L (F_T^R) represent the F_T probes with M-values smaller (larger) than μ_T^F where $T \in (I, II)$. Then, we calculate the probabilities of

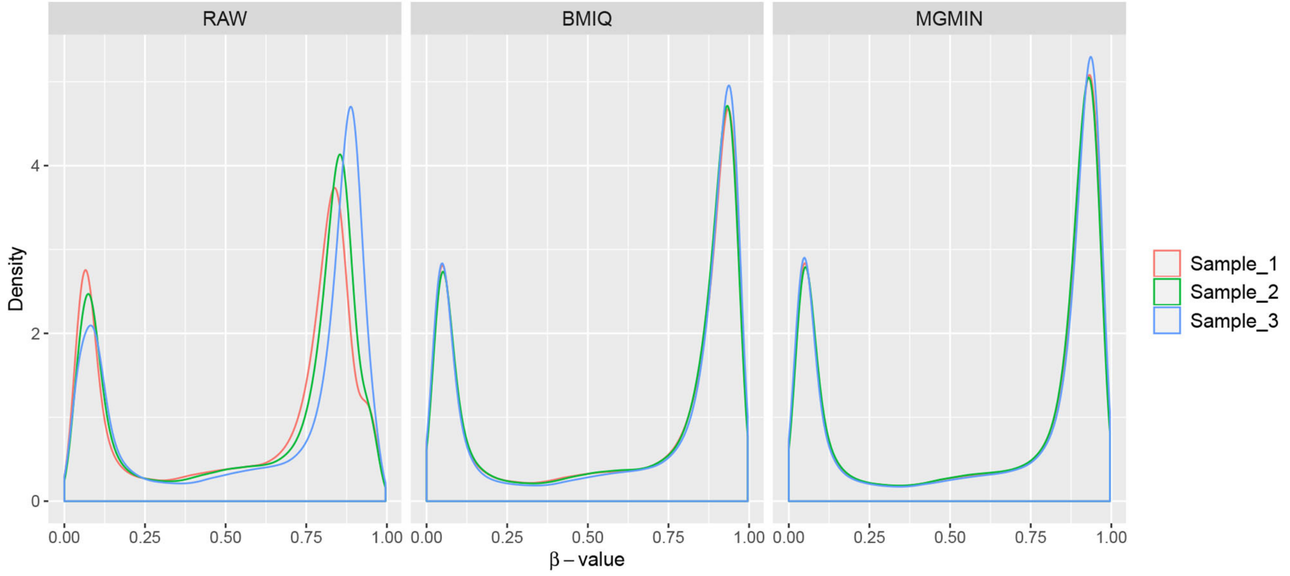

FIGURE 1 | The density curves of β-values for the three replicates in Dataset 1. The left panel is for the case of raw data with no normalization, middle panel for BMIQ and right panel for MGMIN.

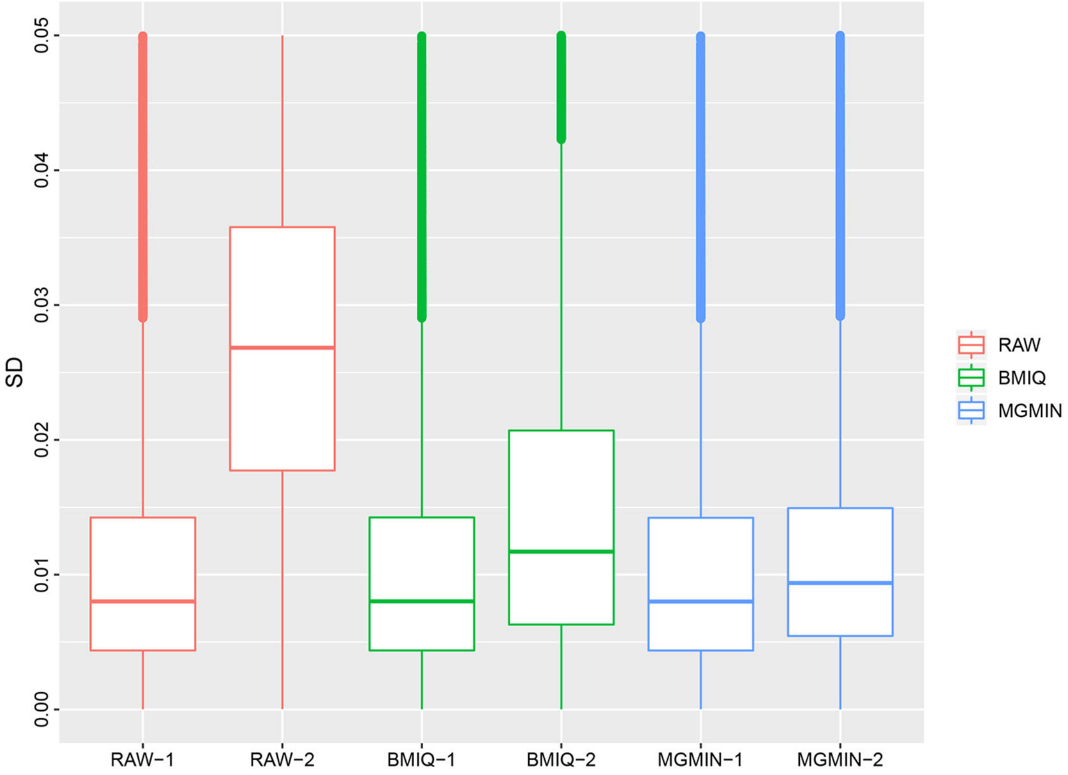

FIGURE 2 | Boxplots of the standard deviations of β-values for the three replicates in Dataset 1, for raw β-values (RAW), normalized β-values by BMIQ (BMIQ), and normalized β-values by MGMIN (MGMIN). RAW-1 represents the type I of raw values and RAW-2 represents the type II of raw values, and so on.

U_{II}^{L} probes, i.e.,

$$p = P(M_{U_{II}^{L}} | \mu_{II}^{U}, (\sigma_{II}^{U})^2) \tag{2}$$

where P represents the cumulative distribution function of the Gaussian component. These probabilities are transformed back to quantiles (M-value) by using the parameters μ_{I}^{U} and σ_{I}^{U} of

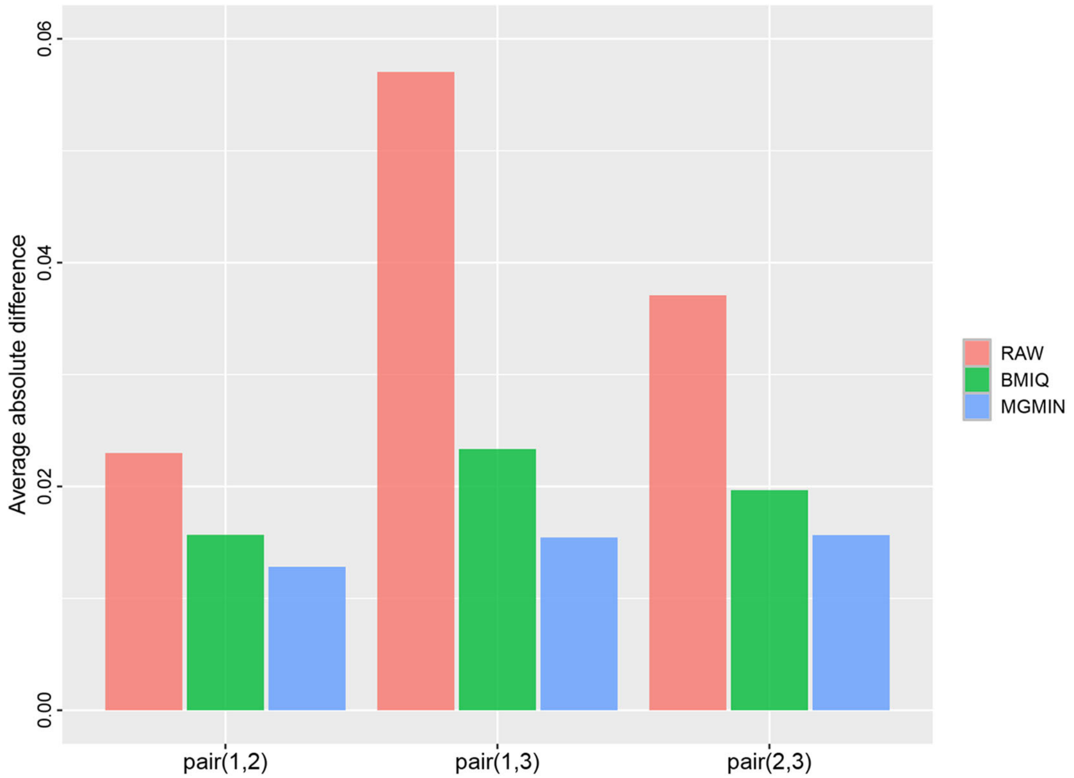

FIGURE 3 | Barplots of the average absolute difference in β-values of type II probes between two samples in each of the three pairs of the three replicates in Dataset 1.

type I probes, i.e.,

$$q = P^{-1}(p|\mu_I^U, (\sigma_I^U)^2) \tag{3}$$

where P^{-1} returns the value of the inverse cumulative density function given the probability p and q is the normalized M-values for U_{II}^L. The similar operation is performed on F_{II}^R probes.

Then, we merge the U_{II}^R, H_{II}, and F_{II}^L probes into one set G on which a conformal (shift + dilation) transformation is performed. Some parameters are identified as $minG = \min\{M_G\}$, $maxG = \max\{M_G\}$ and $\Delta_G^M = maxG - minG$. Similarly, the minimum value of F_{II}^R and the maximum value of U_{II}^L are also identified, i.e., $minF = \min\{F_{II}^R\}$ and $maxU = \max\{U_{II}^L\}$. Two distance values can be calculated as

$$\Delta_{UG} = minG - maxU$$

$$\Delta_{GF} = minF - maxG$$

The new normalized maximum and minimum values of G-probes are expected to satisfy the constraint that

$$maxG' = \min\{F_{II}^{R\prime}\} - \Delta_{GF}$$

$$minG' = \max\{U_{II}^{L\prime}\} + \Delta_{UG}$$

where $F_{II}^{R\prime}$ and $U_{II}^{L\prime}$ are new normalized values for F_{II}^R and U_{II}^L, respectively. So the new normalized range value of set G is $\Delta_G^{M\prime} = maxG' - minG'$. The normalized M-values of set G, $M_{G_{II}}{}'$, is calculated by

$$M_{G_{II}}{}' = minG' + d_f(M_{G_{II}} - minG) \tag{4}$$

where $d_f = \Delta_G^{M\prime}/\Delta_G^M$ is the dilation factor. So, the normalized M-values for type II probes consist of q for U_{II}^L, $M_{G_{II}}{}'$, and q for F_{II}^R.

$$M_{II}{}' = (q_{U_{II}^L}, M_{G_{II}}{}', q_{F_{II}^R})$$

Lastly, the normalized M-values are transformed to β-values.

There are some important points to notice: (i) the initial values for μ and σ in EM algorithm are set as (−4,0,4) and (1,1,1) and small perturbations to the initial μ and σ will not affect the final model because MGMIN captures the natural property of the M-value of DNA methylation, (ii) K_I will be changed to 4 automatically when $\mu_I^F - \sigma_I^F$ is smaller than $\mu_{II}^F - \sigma_{II}^F$ in order to ensure that μ_I^F can always be larger than μ_{II}^F and avoid the presence of an unexpected peak in transformed M-values of hypermethylated type II probes, (iii) if $K_I = 4$, the F_I will be the set of probes belonging to the component with the largest μ, while the U_I contains the probes belonging to the component with the smallest μ and the other two components are assigned

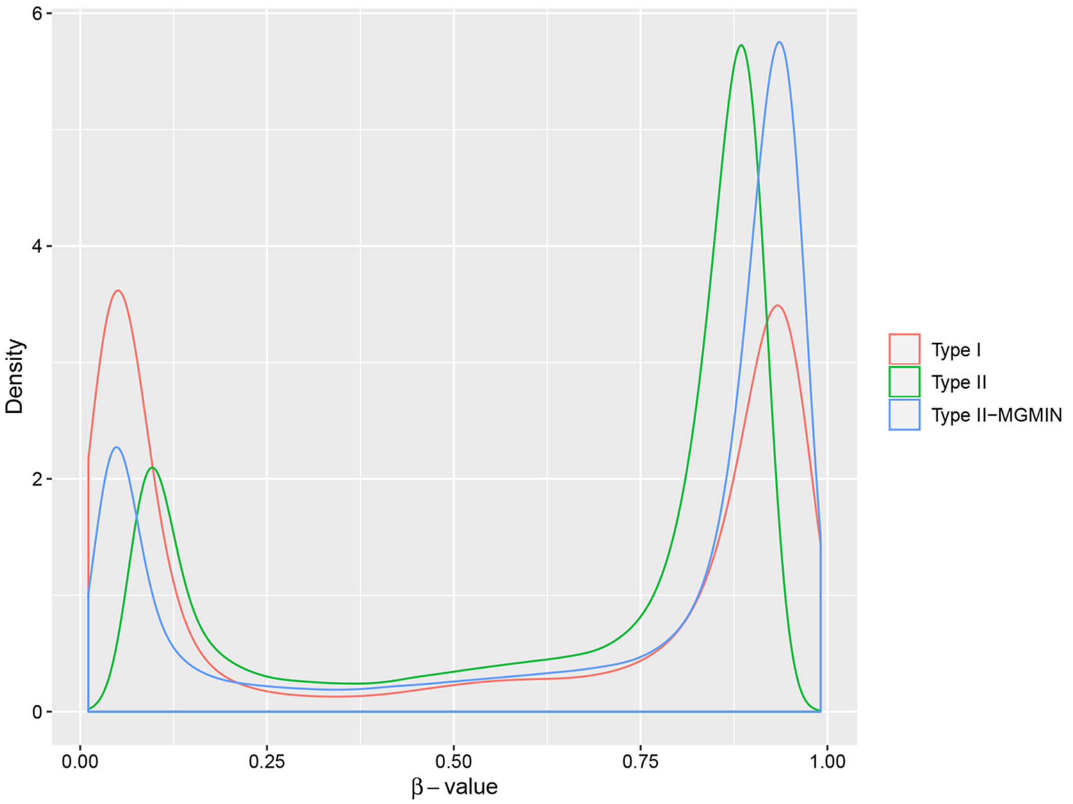

FIGURE 4 | The density curves of β-values for type I probes, type II probes and normalized type II probes (type II-MGMIN) for sample GSM815138 from GEO29290.

to H_I, (iv) no thresholds need to be set by default or estimated by manual to distinguish the three different states of DNA methylation.

2.3. Datasets

We selected several public 450K datasets as following:

Dataset 1: GSE29290 downloaded from GEO considered in Dedeurwaerder et al. (2011). We used the three replicates (GSM15136, GSM15137 and GSM15138) from the HCT116WT cell-line and matched bisulfite pyrosequencing (BPS) date for nine type II probes of sample GSM815138 (r3) (Table 1 in Dedeurwaerder et al., 2011) to evaluate the performance of different methods.

Dataset 2: GSE38268 downloaded from GEO considered in Lechner et al. (2013) which consists of 6 fresh frozen HNC samples. We selected 5 samples as same as (Teschendorff et al., 2012), of which 2 were HPV+ and 3 HPV− (GSM937820 to GSM937824).

Dataset 3: GSE38266 downloaded from GEO considered in Lechner et al. (2013) which contains 21 FFPE HPV+ HNSCC samples and 21 FFPE HPV− HNSCC samples. Note that the entire quality of the dataset GSE38266 is not high.

Dataset 4: GSE95036 downloaded from GEO considered in Degli Esposti et al. (2017) which contains 6 HPV+ HNC samples and 5 HPV− HNC samples.

3. RESULTS

3.1. MGMIN Needs No Default Initial Values of Parameters

Similar to the mixture model of BMIQ, MGMIN applies Gaussian mixture models for M-values instead of beta-mixture models for β-values. MGMIN also uses quantile information to correct the M-values of the type II probes into a distribution which is comparable with that of type I probes. MGMIN complies the inherent Gaussian mixture distributions for M-values of type I and type II probes to avoid setting any parameters manually, which is different from the default breakpoints in BMIQ. Thus, MGMIN needs less manual intervention than BMIQ. However, MGMIN is slightly inferior to BMIQ on some dataset (**Table 1**) due to the entire low quality of the dataset. Note that the PPV of BMIQ on Dataset 3 is lower than that of no normalization (RAW).

3.2. MGMIN Reduces Technical Variation

MGMIN is applied to Dataset 1 to identify the ability to improve reproducibility. The standard deviation (SD) for each probe across the three replicates was computed using no normalization (RAW), BMIQ, and MGMIN separately. As can be seen in **Figure 1**, both MGMIN and BMIQ almost made the density curves for the three replicates coincide with each other and reduced the technical variation significantly

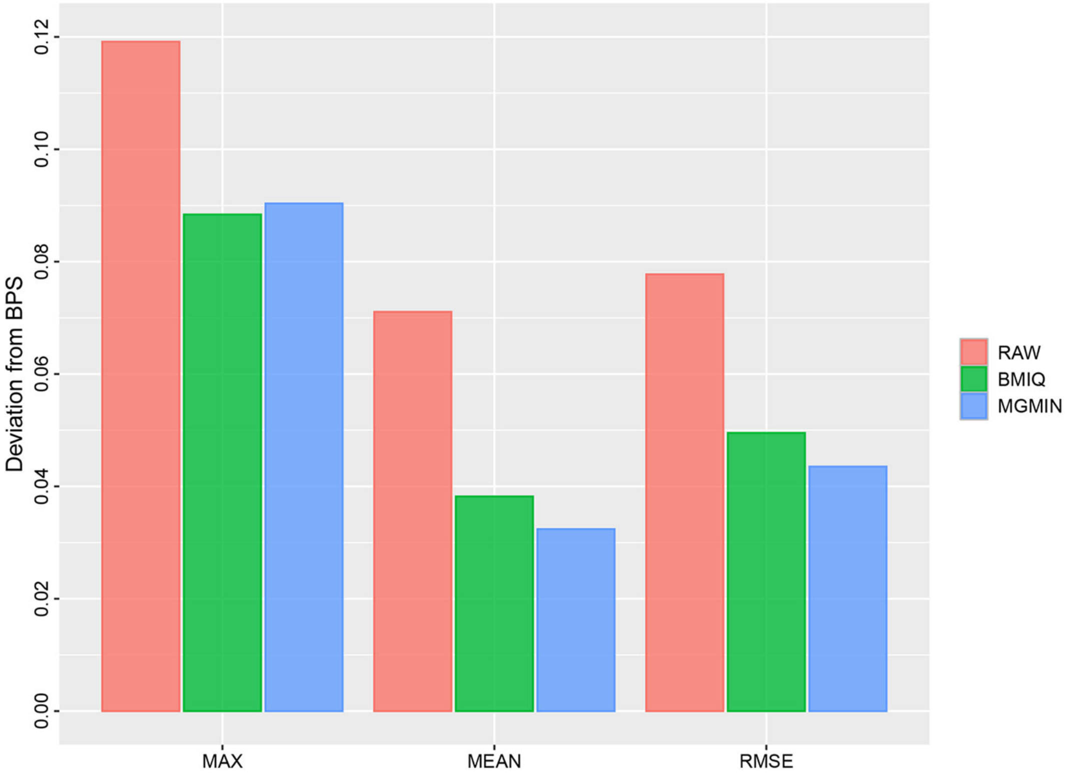

FIGURE 5 | Barplots for the maximum (MAX), mean (MEAN) and root mean square error (RMSE) of the absolute deviation from the matched BPS values of nine type II probes for GSM815138 (r3) in Dataset 1 considered in Dedeurwaerder et al. (2011) using no normalization (RAW), BMIQ, and MGMIN, respectively.

compared to no normalization. Compared to BMIQ, the standard deviation for type II probes adjusted by MGMIN is smaller (**Figure 2**). MGMIN also provided significant reduction of average absolute difference in β-values of type II probes between two samples in each of the three pairs of the three replicates (**Figure 3**).

3.3. MGMIN Reduces Probe Design Bias

MGMIN reduces the probe design bias via correcting the M-values of the type II probes such that the distribution curves for the M-values of the type I and type II probes show similar dynamic ranges and peaks (**Figure 4**). In Dedeurwaerder et al. (2011), the β-values for nine probes of type II by bisulfite pyrosequencing technique for sample GSM815138 (r3) were provided, which can be used as a gold-standard to evaluate the performance of different correction methods. Hence, we compared the normalized results of the nine type II probes in 450K arrays by MGMIN and BMIQ. As shown in **Figure 5**, although MGMIN performed slightly worse than BMIQ at the maximum value of the absolute deviation from BPS data, MGMIN significantly reduced the type II bias than BMIQ and raw data in terms of mean and root mean square error (RMSE) of the absolute deviation from the matched BPS values.

3.4. MGMIN Robustly Finds Informative Differential Methylation Probes Associated With HPV Status

The goal of a bias correction approach is to reduce the technical variation and identify the biological informative signals at the same time. We used a strategy similar to Teschendorff et al. (2012) to compare the result between MGMIN and BMIQ in identifying the differential methylation probes (DMPs) associated with HPV status. First, Dataset 2 consisting of two HPV+ and three HPV− fresh frozen HNC samples were used as the training set to obtain the DMPs associated with HPV status by the *limma* method (Smyth, 2005) and an FDR threshold 0.35 which was as same as (Teschendorff et al., 2012). Both Dataset 3 and Dataset 4 described in the methods section were used as test set. We reanalyzed Dataset 2 and got similar numbers of DMPs to those reported in Teschendorff et al. (2012) with no normalization method (Raw) or BMIQ method (shown in **Table 1**). The results in **Table 1** shows that the positive predictive value (PPV) of MGMIN is slightly less than BMIQ in terms of GSE38266 (Dataset 3) whereas MGMIN outperforms BMIQ in GSE95036 (Dataset 4). The reason for MGMIN slightly inferior to BMIQ in Dataset 3 may be the entire low quality of the dataset (see **Figure 6**) which is that the ratio of samples passing filters is <0.9 ($r = 0.88$) under the least restrictive condition. Let τ_p represent the p-value threshold for bad probes and τ_r

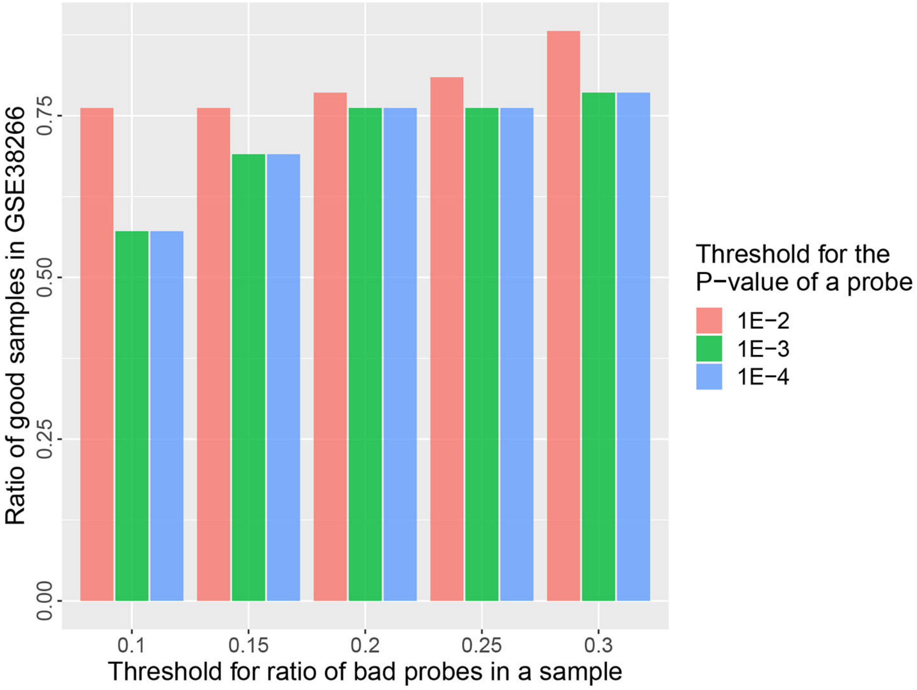

FIGURE 6 | Barplots of the ratio of good samples in GSE38266 under different quality control options (τ_p&τ_r).

represent the threshold for the ratio of bad probes in a sample. The maximum value of τ_r is set to 0.3 here in our opinion because a sample with more than 30% bad probes is vulnerable. We can get the same test dataset from GSE38266 with the one described in Teschendorff et al. (2012) which consists of 18 HPV+ and 14 HPV− samples under the following conditions: (i) $\tau_p = 1e - 4$ or $1e - 3$ and $\tau_r = 0.2$ or 0.25, (ii) $\tau_p = 1e - 2$ and $\tau_r = 0.1$ or 0.15. Overall, MGMIN identified more true positive features than BMIQ.

4. DISCUSSIONS

In this paper, we have proposed a method called MGMIN for correcting the probe design bias of type II probes in Illumina Infinium 450K BeadChips, which can reduce the technical variation and improve the ability to find biologically differential methylation signals. We have shown that MGMIN outperforms BMIQ on multiple evaluation datasets in correcting the type II design bias and improving the data quality.

Similar to BMIQ, MGMIN uses quantile information to correct the *M*-values of type II probes while leaving the *M*-values of type I probes unchanged. The three-state beta-mixture distribution model in BMIQ sets two default breakpoints (0.2, 0.75) to divide the β-values into three classes: hypomethylated, hemimethylated, and hypermethylated, which works well for most cases. However, the result curves of BMIQ show obviously inconsistent in some samples with high heterogeneity. We set 3 or 4 classes for probes depending on the result of $\mu_T^F - \sigma_T^F$ to ensure that the fitted hypermethylated component of type II probes can be located in the left of the hypermethylated component of type I probes, which can partly eliminate the effects of the heterogeneity of samples.

Based on the results of Dataset 3, we think the high quality of dataset is the base of normalization, in other words, there is no meaning to correct the samples with low quality. It should be pointed out that the parameter estimation of MGMIN is slower than that of BMIQ (about 1.5 times), which can be relieved by reducing the number of iterations.

MGMIN can be used in the 450K methylation data preprocessing with other methods to normalize the values of the two type probes and improve the data quality.

DATA AVAILABILITY STATEMENT

The datasets for this study can be found in GEO: GSE29290, GSE38268, GSE38266, and GSE95036.

AUTHOR CONTRIBUTIONS

ZW performed the experiments and wrote the manuscript. All authors read and revised the final manuscript.

FUNDING

This work has been supported by the National Key Research and Development Program of China (Nos: 2017YFC1201201 and 2017YFC0907503).

REFERENCES

Bibikova, M., Barnes, B., Tsan, C., Ho, V., Klotzle, B., Le, J. M., et al. (2011). High density DNA methylation array with single CPG site resolution. *Genomics* 98, 288–295. doi: 10.1016/j.ygeno.2011.07.007

Dedeurwaerder, S., Defrance, M., Calonne, E., Denis, H., Sotiriou, C., and Fuks, F. (2011). Evaluation of the infinium methylation 450K technology. *Epigenomics* 3, 771–784. doi: 10.2217/epi.11.105

Degli Esposti, D., Sklias, A., Lima, S. C., Beghelli-de la Forest Divonne, S., Cahais, V., Fernandez-Jimenez, N., et al. (2017). Unique DNA methylation signature in HPV-positive head and neck squamous cell carcinomas. *Genome Med*. 9:33. doi: 10.1186/s13073-017-0419-z

Du, P., Zhang, X., Huang, C.-C., Jafari, N., Kibbe, W. A., Hou, L., et al. (2010). Comparison of beta-value and *M*-value methods for quantifying methylation levels by microarray analysis. *BMC Bioinformatics* 11:587. doi: 10.1186/1471-2105-11-587

Irizarry, R. A., Ladd-Acosta, C., Wen, B., Wu, Z., Montano, C., Onyango, P., et al. (2009). The human colon cancer methylome shows similar hypo-and hypermethylation at conserved tissue-specific cpg island shores. *Nat. Genet*. 41, 178–186. doi: 10.1038/ng.298

Lechner, M., Fenton, T., West, J., Wilson, G., Feber, A., Henderson, S., et al. (2013). Identification and functional validation of HPV-mediated hypermethylation in head and neck squamous cell carcinoma. *Genome Med*. 5:15. doi: 10.1186/gm419

Maksimovic, J., Gordon, L., and Oshlack, A. (2012). Swan: subset-quantile within array normalization for illumina infinium humanmethylation450 beadchips. *Genome Biol*. 13:R44. doi: 10.1186/gb-2012-13-6-r44

McLachlan, G. J., Bean, R., and Jones, L. B.-T. (2006). A simple implementation of a normal mixture approach to differential gene expression in multiclass microarrays. *Bioinformatics* 22, 1608–1615. doi: 10.1093/bioinformatics/btl148

Pan, W., Lin, J., and Le, C. T. (2002). Model-based cluster analysis of microarray gene-expression data. *Genome Biol*. 3:research0009-1. doi: 10.1186/gb-2002-3-2-research0009

Paul, D. S., Teschendorff, A. E., Dang, M. A., Lowe, R., Hawa, M. I., Ecker, S., et al. (2016). Increased DNA methylation variability in type 1 diabetes across three immune effector cell types. *Nat. Commun*. 7:13555. doi: 10.1038/ncomms13555

Smyth, G. K. (2005). "Limma: linear models for microarray data," in Bioinformatics sand Computational Biology Solutions Using R and Bioconductor, eds R. Gentleman, V. Carey, W. Huber, R. Irizarry, and S. Dudoit (New York, NY: Springer), 397–420. doi: 10.1007/0-387-29362-0_23

Teschendorff, A. E., Marabita, F., Lechner, M., Bartlett, T., Tegner, J., Gomez-Cabrero, D., et al. (2012). A beta-mixture quantile normalization method for correcting probe design bias in illumina infinium 450K DNA methylation data. *Bioinformatics* 29, 189–196. doi: 10.1093/bioinformatics/bts680

Touleimat, N., and Tost, J. (2012). Complete pipeline for infinium® human methylation 450K beadchip data processing using subset quantile normalization for accurate dna methylation estimation. *Epigenomics* 4, 325–341. doi: 10.2217/epi.12.21

Yeung, K. Y., Fraley, C., Murua, A., Raftery, A. E., and Ruzzo, W. L. (2001). Model-based clustering and data transformations for gene expression data. *Bioinformatics* 17, 977–987. doi: 10.1093/bioinformatics/17.10.977

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.