frontiers in Genetics

METHODS
published: 07 January 2022
doi: 10.3389/fgene.2021.810985

# GMQN: A Reference-Based Method for Correcting Batch Effects and Probe Bias in HumanMethylation BeadChip

*Zhuang Xiong*[1,2,3†], *Mengwei Li*[1,2,3†], *Yingke Ma*[1,2], *Rujiao Li*[1,2] *and Yiming Bao*[1,2,3*]

*[1]National Genomics Data Center, Beijing Institute of Genomics, Chinese Academy of Sciences/China National Center for Bioinformation, Beijing, China, [2]CAS Key Laboratory of Genome Sciences and Information, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing, China, [3]University of Chinese Academy of Sciences, Beijing, China*

The Illumina HumanMethylation BeadChip is one of the most cost-effective methods to quantify DNA methylation levels at single-base resolution across the human genome, which makes it a routine platform for epigenome-wide association studies. It has accumulated tens of thousands of DNA methylation array samples in public databases, providing great support for data integration and further analysis. However, the majority of public DNA methylation data are deposited as processed data without background probes which are widely used in data normalization. Here, we present Gaussian mixture quantile normalization (GMQN), a reference based method for correcting batch effects as well as probe bias in the HumanMethylation BeadChip. Availability and implementation: https://github.com/MengweiLi-project/gmqn.

**Keywords: DNA methylation, epigenome-wide association studies, batch effect, probe bias, HumanMethylation BeadChip**

***Edited by:***
*Tao Huang,*
*Shanghai Institute of Nutrition and Health (CAS), China*

***Reviewed by:***
*Hongde Liu,*
*Southeast University, China*
*Min Wu,*
*Institute for Infocomm Research (A∗STAR), Singapore*

****Correspondence:***
*Yiming Bao*
*baoym@big.ac.cn*

*†These authors have contributed equally to this work and share first authorship*

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

***Received:*** *08 November 2021*
***Accepted:*** *13 December 2021*
***Published:*** *07 January 2022*

***Citation:***
*Xiong Z, Li M, Ma Y, Li R and Bao Y (2022) GMQN: A Reference-Based Method for Correcting Batch Effects and Probe Bias in HumanMethylation BeadChip. Front. Genet. 12:810985. doi: 10.3389/fgene.2021.810985*

## 1 INTRODUCTION

As a well-known epigenetic marker, DNA methylation plays a crucial role in numerous physiological processes as well as complex traits, such as development, phenotype and cancer (Smith and Meissner, 2013; Xu et al., 2013; Joehanes et al., 2016). With the advancement of epigenetic sequencing technologies and a radical decline in sequencing costs, especially the DNA methylation array, massive samples can be used to the explore epigenetic basis of complex traits, which has also resulted in the accumulation of a large amount of DNA methylation array data in public databases (Barrett et al., 2012; Li et al., 2018; Xiong et al., 2020). According to the statistics of DNA methylation array data in the GEO database, Illumina HumanMethylation450 BeadChip (450 k) has become the most widely used means of large-scale methylation profiling of human samples in recent years. The newly emerging Illumina HumanMethylationEPIC BeadChip (EPIC/850 k) uses the same technology as 450 k but covers nearly double the number of CpG sites and will become the main effective strategy of epigenome-wide association studies (EWAS) in the future (**Figure 1A**). Integrating both large samples from public resources and private data will become a common and main research strategy for future research on potential regulatory mechanisms of complex traits, particularly for EWAS (Yuan et al., 2019). As sample processing and sequencing processes varied amongst laboratories, there are some unavoidable differences which have nothing to do with biological factors but are between-array bias defined as batch effects (Leek et al., 2010; Forest et al., 2018), which will reduce the signal-to-noise ratio and adversely affect downstream analysis.

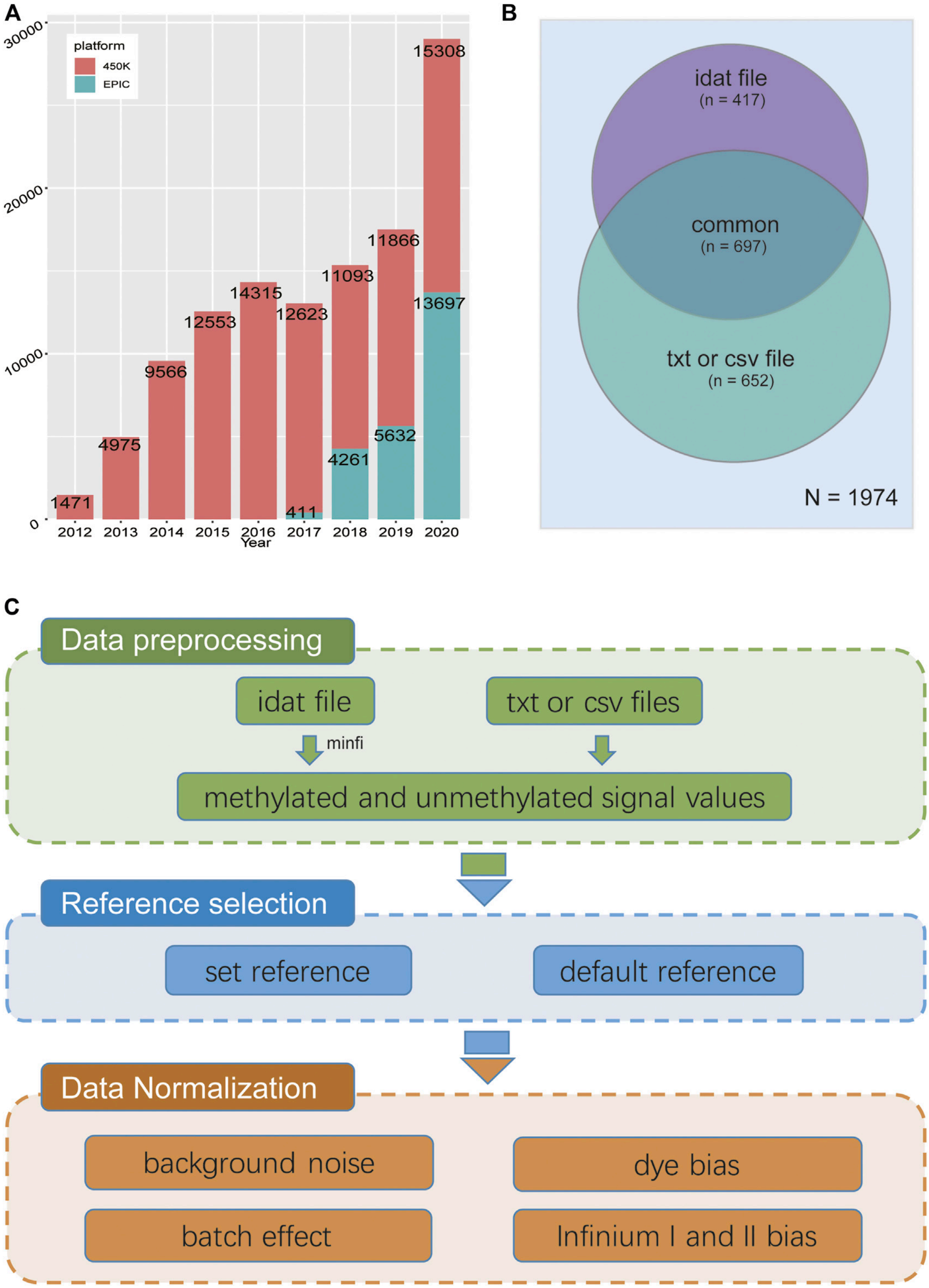

**FIGURE 1 | (A)** Statistics of 450 k and EPIC data by year and project number in NCBI GEO database. **(B)** Distribution of data types of DNA methylation chip projects submitted to the GEO database as of December 2020. There were a total of 1,114 items containing the original “idat” files, and 1,349 items containing TXT or CSV files, indicating that most of the items missed original file. **(C)** The workflow of GMQN.

**TABLE 1 |** Overview of benchmark test dataset.

| Project id | Number of samples | Benchmark test | Annotation | Platform |
|---|---|---|---|---|
| GSE52731 | 56 | batch effects detection | — | 450 k |
| GSE139687 | 27 | batch effects detection | — | EPIC |
| GSE42861 | 689 | case-control study | Rheumatoid Arthritis | 450 k |
| GSE128235 | 537 | case-control study | Depression | 450 k |
| GSE125105 | 210 | regression analysis | Age | 450 k |
| GSE42861 | 335 | regression analysis | Age | 450 k |
| GSE87571 | 732 | regression analysis | Age | 450 k |
| GSE87571 | 732 | comparison of the methylation levels of adjacent sites | — | 450 k |
| GSE42861 | 689 | case-control study (reference evaluation) | Rheumatoid Arthritis | 450 k |
| GSE125105 | 210 | regression analysis (reference evaluation) | Age | 450 k |
| GSE42861 | 335 | regression analysis (reference evaluation) | Age | 450 k |
| GSE87571 | 732 | regression analysis (reference evaluation) | Age | 450 k |

A number of DNA methylation array normalization methods have been proposed, each with its own set of advantages and disadvantages in different study scenarios(Niu et al., 2016; Xu et al., 2017; Wang et al., 2020). Many methods, on the other hand, are not well suited to the analysis of a large amount of public data. The majority of methods rely on data from control probes or OOB (out of band) probes, as a result, cannot be used for public data unless the original data are available. However, only approximately half of the 450 k and EPIC projects in GEO, the largest publicly accessible DNA methylation array database, provide original data (**Figure 1B**). As the well-known normalization method on β-values of DNA methylation, SWAN and BMIQ do not use the information from these two types of probes. Instead, they only deal with within-array bias. (Infinium I and II bias) (Maksimovic et al., 2012; Teschendorff et al., 2013).

Without control probes or OOB, we still have to deal with four types of deviations: Infinium I and II bias, red and green channel signal deviations, background noise, and batch effects. Therefore, we propose a reference-based method for correcting batch effects as well as probe bias in the HumanMethylation BeadChip, which is called Gaussian Mixture Quantile Normalization (GMQN). The method includes four steps: (I) A two-state Gaussian mixture model was fitted to the median values of each Infinium I probe signal intensity from a large single study (GSE105018). For rescaling Infinium I probes, the mean and variance of two components were used as a reference. (II) Fitting of a two-state Gaussian mixture model to the input Infinium I probe signal intensity. (III) Transform the probability of Infinium I probes from each component of input data to quantiles using the inverse of the cumulative Gaussian distribution with the mean and variance estimated from the corresponding reference component. (IV) After reversing the batch effect, GMQN can also normalize Infinium II probes on the basis of Infinium I probes in combination with BMIQ and SWAN, the two well-known normalization methods on β-values of DNA methylation (Maksimovic et al., 2012; Teschendorff et al., 2013) (**Figure 1C**).

# 2 MATERIALS AND METHODS

## 2.1 DNA Methylation Data

Data for method development and testing are taken from the GEO and TCGA databases, which contain 450 k and EPIC records (**Table 1**). Respectively, the sample information is annotated using a combination of automatic grabbing and manual analysis. The R package “minfi” (http://www.bioconductor.org/packages/release/bioc/html/minfi.html) is primarily used to interpret and preprocess the original signal (Fortin et al., 2016). Considering that some public data only have original methylated and unmethylated signal value files, we use the “preprocessRaw” method to extract the original signal values without any processing. To ensure fairness, the methylated and unmethylated signal values of all probes except the control probe are collected and used as the input value in all subsequent tests and comparisons. The methylation level is represented by β, $\beta = M/(M+U)$, where $M$ and $U$ represent the intensity of methylation and non-methylation signal values, respectively.

## 2.2 Reference Data

In GMQN, there are two ways to set the reference signal value distribution. To begin, users can use the function “set reference” in the “GMQN” package to match their own data to fit their own reference distribution. The second option is to use the default reference, which is a two-state Gaussian mixture model fitted to the median values of each Infinium I probe signal intensity from a large single study (GSE105018), including 1,658 whole blood samples obtained from E-Risk cohort participants when they were 18 years old (Hannon et al., 2018). The mean and variance of two components are used as reference for rescaling Infinium I probes.

## 2.3 GMQN

To eliminate any source of variation that is not related to biology but rather to technical limitations, such as dye bias or batch effects, we must first identify the manifestations of these variations in the data (Dedeurwaerder et al., 2014). To that end, we investigate the signal value distribution characteristics of two types of probes. We found that the signal values of the red and green channels of Infinium I probes can be decomposed into the superposition of two Gaussian distributions, and that the fitting parameters of these Gaussian distributions may efficiently distinguish batches (details in result). Using this feature, we draw on the idea of BMIQ, respectively fit the Gaussian mixture distribution to the signal values of the red and green channels of Infinium I probes, and then adjust the shape of the Gaussian distribution corresponding to different samples to the same shape to the reference to minimize batch effects and other deviations. To achieve this process, GMQN standardizes the data in three steps.

The first step is the establishment of the reference distribution. In order to address the issue of the rapid growth of public data, GMQN adopts a data normalization method based on reference distribution, which is also widely used in the normalization of data in the EWAS Data Hub (https://ngdc.cncb.ac.cn/ewas/datahub/index) (Xiong et al., 2020; Xiong et al., 2021). Usually, we need to average the signal intensity of each probe on the reference data set between samples, and fit the Gaussian mixture distribution to the probe signal intensity on the red and green channels of Infinium I probes respectively. The Expectation-Maximization algorithm is used to estimate the parameters, and the red channel fitting result is expressed as: $\{(\mu_1^{rR}, \sigma_1^{rR}), (\mu_2^{rR}, \sigma_2^{rR})\}$ , the green channel fitting result is expressed as: $\{(\mu_1^{rG}, \sigma_1^{rG}), (\mu_2^{rG}, \sigma_2^{rG})\}$ , where $r$ is the reference, 1 and 2 respectively represent the two states of the mixed model with a smaller and larger mean, and $R$ and $G$ represent the red and green channels, respectively.

The second step is the normalization between arrays. Between-array normalization is carried out separately for the red and green channels of Infinium I probes. Taking the green channel as an example, we first fit the Gaussian mixture distribution to the signal intensity of the green channel of the input Infinium I probe to obtain the fitting parameters $\{(\mu_1^G, \sigma_1^G), (\mu_2^G, \sigma_2^G)\}$. For the state with the smaller mean value, state 1, we perform the following conversion:

$$\rho = \mathrm{F}\left(\mathrm{S}_1 \middle| \mu_1^{\mathrm{G}}, \sigma_1^{\mathrm{G}}\right)$$
$$\mathrm{q} = \mathrm{F}^{-1}\left(\rho \middle| \mu_1^{\mathrm{rG}}, \sigma_1^{\mathrm{rG}}\right)$$

where $S_1$ is the signal belonging to state 1 in the green channel signal, $\rho$ is the cumulative distribution probability of the signal value in the Gaussian distribution, and $q$ is the signal value corresponding to the cumulative probability in the reference distribution. Through this step, we map the input signal to the reference signal and eliminate biases such as dye bias and batch effects. The signal in state 2 and the signal in the red channel are processed using similar steps.

The third step is within-array normalization, which mainly includes Infinium I/II-type bias correction. In the second step, we obtained the normalized Infinium I probes signal. Based on Infinium I probes signal, we used BMIQ or SWAN to standardize the Infinium II probes signal. BMIQ and SWAN were fine-tuned to improve the speed and effectiveness, respectively.

## 2.4 Benchmark Test

Since other methods cannot be used in the absence of original data, in the benchmark test, we compare GMQN, SWAN, BMIQ, and GMQN combined with SWAN and BMIQ (GMQN.SWAN and GMQN.BMIQ). In order to test whether GMQN can improve the effect of SWAN and BMIQ, we designed the following four benchmark tests.

### 2.4.1 Batch Effects Detection

In order to make the method more universal, we searched the GEO database for two sets of technical replicates, including 450 k and EPIC. The first set (EPIC, GSE139687) has nine samples that are replicated three times each, while the second set (450 k, GSE52731) has 56 repetitions of one sample. For the first data set, we measured the variance at the probe level between every three technical replicates and then averaged the variance among the nine samples. For the second, we directly calculated the variance of the sample at the probe level.

### 2.4.2 Case-Control Study

Case-control studies are the most common form of research in EWAS. Researchers classify samples into case and control groups and look for differences in methylation sites between the two groups in this form of study. We used the data of two diseases in public sources to evaluate the performance of GMQN in the case-control studies. To simulate two separate batches, we divide the samples in the data set at a ratio of 2:1 into training and test sets for each disease. In the training set, we aim to keep the samples in the same batch of chips, and the batch effect and other errors are kept to a minimum. Differential methylation analysis was performed in both the training and test sets, with the results of the training set acting as the gold standard for detecting consistency between the training and test sets and drawing the receiver operating characteristic (ROC) curve.

### 2.4.3 Regression Analysis

The term “regression analysis” refers to the process of associating DNA methylation levels with continuous variables such as age, BMI, and so on in order to identify DNA methylation sites that are associated with these variables. Age is a trait that has been reported more frequently in EWAS, and there is a substantial amount of data on it. As a result, we use age as the research object in this study and collect 1,277 sample data sets containing age information from three independent projects. Data from these projects ensure that the sample’s batch effect is high, allowing each standardized method’s effect to be better measured. A large number of studies have reported that there is a linear relationship between DNA methylation and age (Horvath, 2013; Chung et al., 2021), and the Pearson correlation coefficient is particularly suitable for quantifying the linear relationship. Therefore, we calculated the Pearson correlation coefficient between DNA methylation and age as quantitative indicators.

### 2.4.4 Comparison of the Methylation Levels of Adjacent CpG Sites

Studies have reported that DNA methyltransferase has a limited range of action, resulting in nearly identical methylation levels at adjacent CpG sites in the genome (Zhang et al., 2015; Guo et al., 2017). In this part, we selected 141,653 pairs of probes with a genome distance of less than 10 bp on the chip. We determined the average difference in DNA methylation levels of these probes for each sample and chose 141,653 pairs of probes randomly as controls.

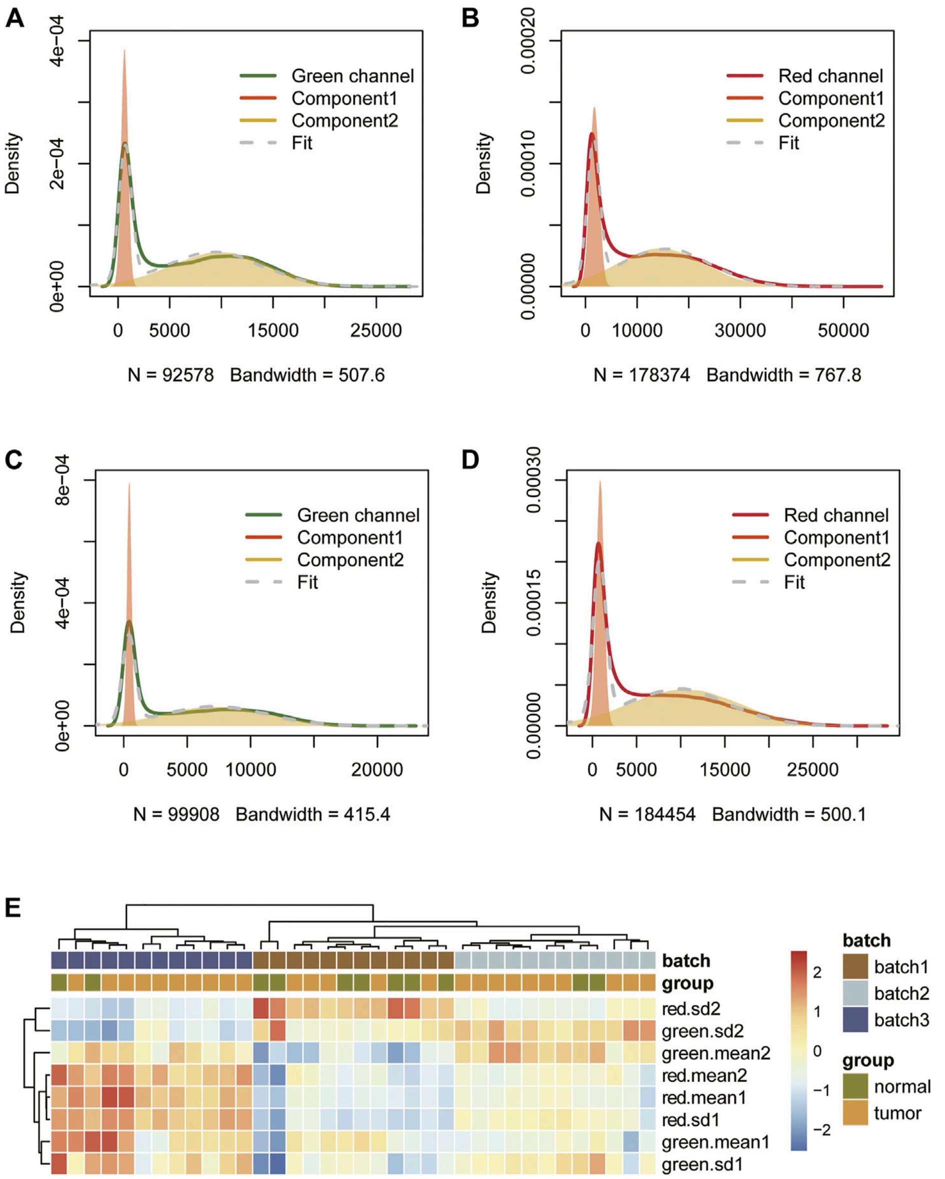

**FIGURE 2 |** The signal intensity distribution characteristics of Infinium I probes (450 k data **(A,B)**, EPIC data **(C,D)**) and clustering results of different batches of samples based on fitting parameters of the Gaussian distributions **(E)**.

# 3 RESULTS

## 3.1 The Signal Intensity Distribution Characteristics of Infinium I Probes and the Principle of GMQN

The signal from the control probe can, ideally, be used to quantify the batch effect between samples. However, most public data lack original data, so we tried to find other manifestations of batch effects. We found that the signal intensity of the red and green channels of Infinium I probes can be approximately decomposed into the superposition of two Gaussian distributions, both in 450 k and EPIC arrays (**Figure 2**). We speculate that this may be related to the bimodal distribution of human DNA methylation levels. When the methylation value is extremely high (>0.8) or extremely low (< 0.2), one of the two Infinium I probes that detects the site's methylation level emits almost no light, and the fluorescence signal intensity of these probes constitutes the first peak of the Gaussian distribution, that is, the peak with the smaller mean. The fluorescence signal intensity of other probes constitutes the second Gaussian distribution. Since the methylation levels of the sites corresponding to these probes are dispersed, the Gaussian distribution variance is larger. We cluster the Gaussian distribution parameters fitted by different

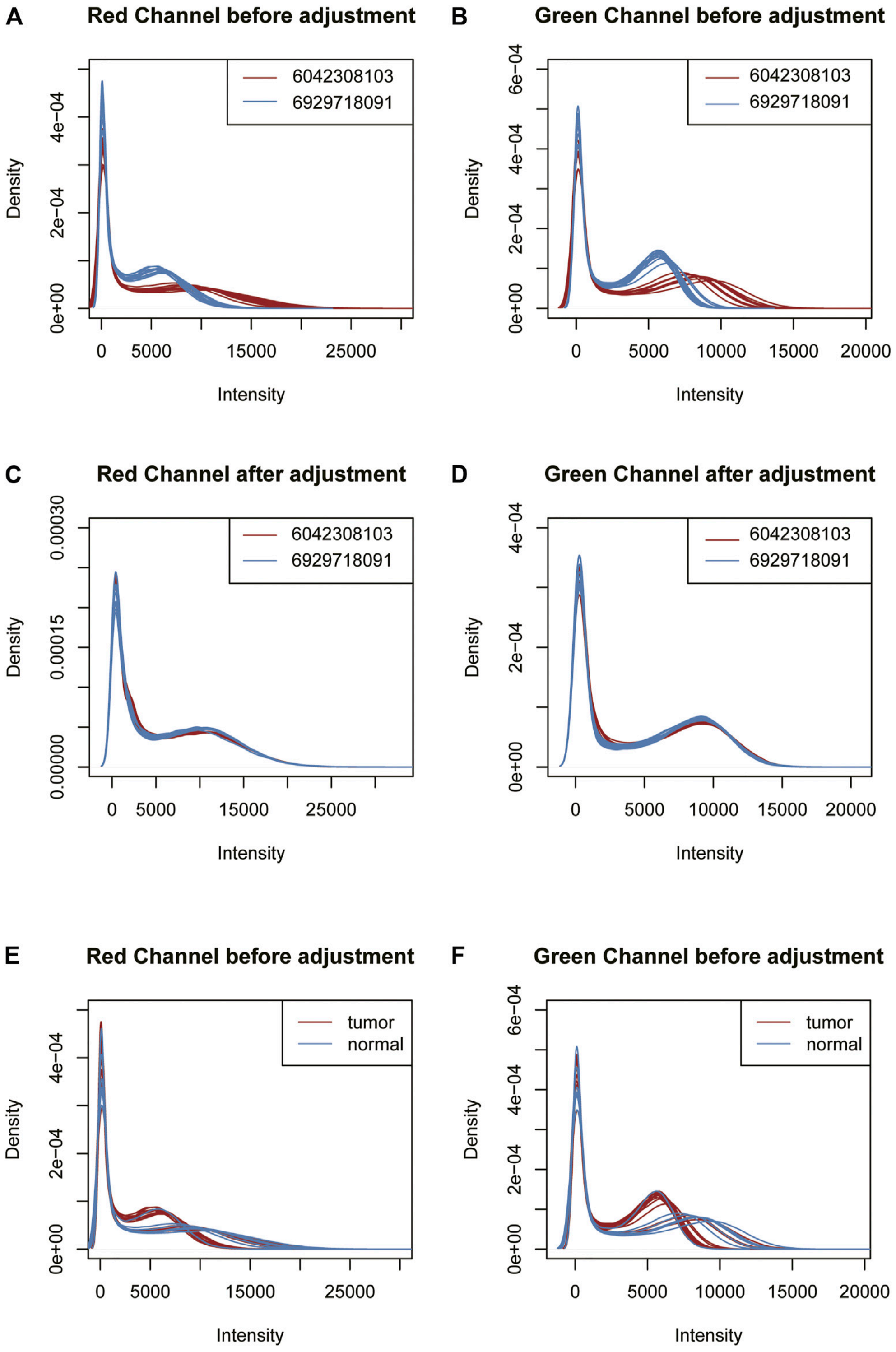

**FIGURE 3 |** The 450 k data signal strength distribution of the red and green channels of Infinium I probes before and after GMQN normalization. The signal intensities of the red **(A)** and green **(B)** channels of the two batches were clearly divided into two batches before being corrected. And the differences Were not due to biological differences (tumor and normal) **(E,F)**. After the GMQN correction, the batch effect problem is significantly reduced **(C,D)**.

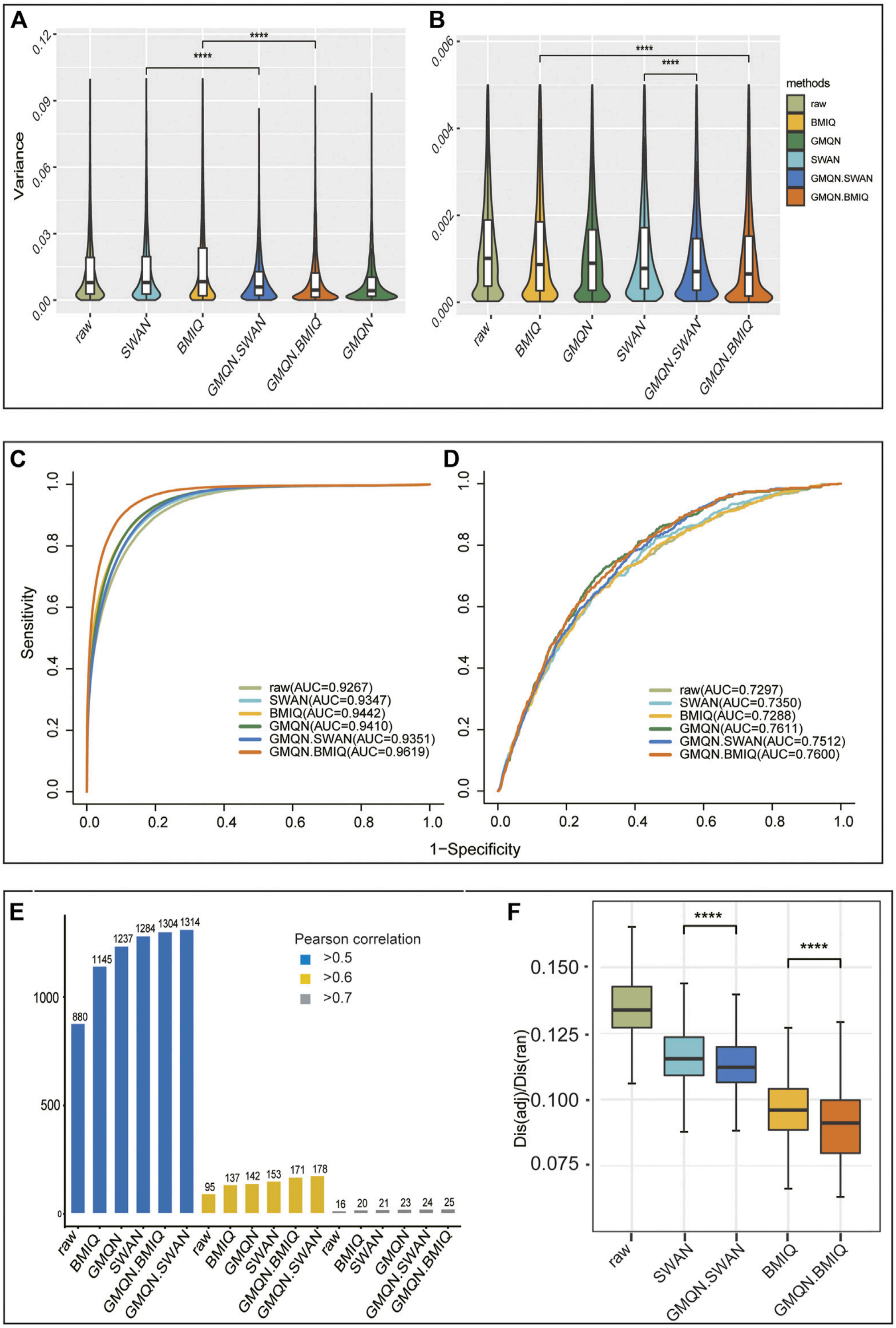

**FIGURE 4 |** The result of Benchmark Test. **(A)** and **(B)**: batch effects detection. **(C)** and **(D)**: case-control study. **(E)**: regression analysis. **(F)**: comparison of the methylation levels of adjacent CpG sites (*****p* < 10-4, *****p* < 10-4)

samples to see if these Gaussian peaks are related to batches. The results show that the fitting parameters of the four Gaussian distributions (two for each of the red and green channels) can be used to distinguish the batches, and that even if the sample difference is large, the parameter difference will be small within the batches (**Figure 2**).

Using this feature, we propose a GMQN standardization method. The basic principle of this method is to fit a Gaussian mixture model for Infinium I probes of different batches, and then adjust the Gaussian distribution shapes fitted by different batches to the same to eliminate the batch effect on Infinium I probes. Finally, the Infinium I probes are taken as the standard, and BMIQ or SWAN are used to standardize the Infinium II probes. The signal strength distribution of the red and green channels of Infinium I probes was then measured in two batches of samples in a TCGA tumor project before and after GMQN normalization. We found that the distribution of the two batches differed greatly in both 450 k and EPIC data, and the differences were not due to biological differences (tumor and normal). The distributions of the two batches tend to be consistent after GMQN standardization (**Figure 3**, **Supplementary Figure S3**).

## 3.2 GMQN Reduces Technical Variability

Technical repetition is the most direct way to measure the batch effect. As a result, we chose two different sets of technical replicates. The first set (EPIC, GSE139687) has nine samples that are repeated three times each, while the second set (450 k, GSE52731) has 56 repetitions of one sample (Aryee et al., 2014; Li et al., 2020). The variances of the probes of the two sets of samples were determined separately. While each method decreased the variance of the probe methylation level relative to the original data in the two sets of technical replicates, the variance of the probe methylation level after GMQN + BMIQ and GMQN + SWAN treatment was the lowest and second lowest, respectively (**Figures 4A,B**). In particular, without combining SWAN and BMIQ, GMQN performed best in the first data set (**Figure 4A**). This demonstrates that GMQN, especially when used in combination with BMIQ and SWAN, is capable of effectively reducing batch effects.

## 3.3 GMQN Leads to Better Detection of Differential Methylation

In order to test the effects of GMQN in the case-control studies, we selected normal and disease samples for rheumatoid arthritis and depression (Liu et al., 2013; Zannas et al., 2019). The differential methylation estimation results indicate that there are approximately 50,000 and 1,000 differential methylation positions in the normal and disease samples of these two diseases, respectively (see **Supplementary Table S1**). The ROC curve shows that compared with the original data, the consistency of the training set and the test set results is greatly improved in rheumatoid arthritis, GMQN + BMIQ has the best effect, while SWAN and the original data have poor results, but whether it is BMIQ or SWAN, the effect can be achieved after combination with GMQN, GMQN, GMQN + BMIQ, and GMQN + SWAN all outperform other methods in the depression group (**Figures 4C,D**). In case-control studies, these results suggest that GMQN can enhance SWAN and BMIQ effects.

## 3.4 GMQN Improves the Effectiveness of Regression Analysis

Regression analysis is a crucial form of analysis in EWAS. For continuous traits such as age and BMI, the relevant DNA methylation sites can be found through regression analysis. Compared with case-control studies, the results of regression analysis are often more influenced by data processing methods.

We used data processed by different methods to identify age-related DNA methylation sites to examine the effect of GMQN in regression analysis. Our data in this analysis come from three separate projects, where the batch effect is high and the sample age period is large (from 14 to 94 years old) (Johansson et al., 2013; Liu et al., 2013; Aryee et al., 2014). Using Pearson correlation coefficients of 0.5, 0.6, and 0.7 as thresholds, we measured the number of age-related DNA methylation sites identified by each method (see **Supplementary Table S2**). The findings show that the GMQN + SWAN treatment group can find more age-related methylation sites than other methods under various thresholds, and GMQN can boost the effects of BMIQ and SWAN under a strict threshold, and improve the effect of regression analysis (**Figure 4E**). To ensure that the sites found by GMQN are true positive sites, we further analyzed these sites. Surprisingly, we examined the five sites (cg15448975, cg16419235, cg07416237, cg04875128, cg14692377) with Pearson correlation coefficients less than 0.7 after BMIQ analysis and greater than 0.7 after GMQN + BMIQ analysis in the EWAS Atlas (https://ngdc.cncb.ac.cn/ewas/atlas), a curated knowledgebase of epigenome-wide association studies (Li et al., 2019; Xiong et al., 2021), and discovered that all of them were age-related, indicating that the majority of the newly discovered age-related sites in GMQN are true positives.

## 3.5 GMQN Reduces Differences in Methylation Levels Between Adjacent CpG Sites

The difference in methylation levels between adjacent CpG sites is approximately 13% of that between random sites. Meanwhile, the difference in methylation levels between adjacent CpG sites in the original data group was greater than that in other groups, confirming that this benchmark test is reasonable. The GMQN + BMIQ processed group had the smallest difference in methylation levels between adjacent CpG sites, while the GMQN + SWAN treatment was not as efficient as BMIQ but still better than SWAN (**Figure 4F**).

## 3.5 Selection and Evaluation of Reference Data

To help users better choose reference data, we evaluated the default reference (provided by GMQN) and the user's own data fitting reference by two benchmark test, case-control study and regression analyses (**Supplementary Figure S4**). The evaluation results show that in the case-control study, there is almost no difference between the two methods of establishing references (**Supplementary Figure S4A**, **Supplementary Figure S4B**). In regression analysis, more relevant methylation sites were obtained using the default reference (**Supplementary Figure S4C**, **Supplementary Figure S4D**).

## 4 DISCUSSION

The accumulation of public DNA methylation array data has provided favorable conditions for the advancement of EWAS, allowing data analysts to investigate the association between various traits by massive public data mining without relying on experiments. As a result, we proposed GMQN, a standardized method suitable for massive public DNA methylation array data. In comparison to other DNA methylation array normalization approaches, GMQN has the following advantages: First and foremost, GMQN is a reference-based Gaussian mixture quantile normalization method. It can be used to calibrate a newly added sample to the same level as the previous batch of samples without wasting a lot of computational resources, which will solve the N+1 issue in big data integration. The EWAS data portal of EWAS Open Platform (https://ngdc.cncb.ac.cn/ewas) currently integrates and stores 115,852 methylation chip data using the GMQN (Xiong et al., 2020; Xiong et al., 2021). Second, GMQN will address the issue of batch effect processing and standardization in public data due to missing original data, making it easier for researchers to combine self-produced and public data to investigate epigenetic mechanisms of various phenotypes. Finally, since most DNA methylation chip processing software packages are written in R, GMQN is written in R as well to increase compatibility with other software. Users can easily achieve GMQN standardization using the R package "GMQN". Users can combine SWAN and BMIQ to perform parallel analysis on multiple CPUs using the two functions "gmqn_swan_parallel" and "gmqn_bmiq_parallel".

By evaluating 450 k and EPIC array data in four separate application scenarios above, we found that GMQN can effectively minimize noise in public data and increase the accuracy of downstream analysis. GMQN will boost the two well-known methylation chip standardization methods, BMIQ and SWAN, even if it does not perform well in some scenarios, especially when the reference methylation distribution and the methylation data distribution to be standardized are vastly different, as in DNA methyltransferase gene knockout samples versus normal samples. Many DNA methylation array data standardization methods have been developed in recent years (Triche et al., 2013; Yousefi et al., 2013; Fortin et al., 2014; Niu et al., 2016; Xu et al., 2016; Xu et al., 2017; Wang et al., 2020), and they have proven to be invaluable in epigenetics research, especially for EWAS (Marabita et al., 2013; Wang et al., 2015). However, we believe that GMQN can improve the normalization effect to some degree, especially when there are no original data.

## DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author.

## AUTHOR CONTRIBUTIONS

YB conceived the study and revised the manuscript. ZX, ML finished the experiments, analyzed the data, designed the tables and figures. ZX wrote the manuscript. YM, RL revised the manuscript. All authors read and revised the final manuscript.

## FUNDING

This work was supported by the Strategic Priority Research Program of the Chinese Academy of Sciences (XDB38030200, XDB38030400); Key Technology Talent Program of the Chinese Academy of Sciences; Genomics Data Center Construction of Chinese Academy of Sciences (WX145XQ07-04); The Professional Association of the Alliance of International Science Organizations (ANSO-PA-2020-07), The Open Biodiversity and Health Big Data Programme of IUBS.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.810985/full#supplementary-material

## REFERENCES

Aryee, M. J., Jaffe, A. E., Corrada-Bravo, H., Ladd-Acosta, C., Feinberg, A. P., Hansen, K. D., et al. (2014). Minfi: a Flexible and Comprehensive Bioconductor Package for the Analysis of Infinium DNA Methylation Microarrays. *Bioinformatics* 30 (10), 1363–1369. doi:10.1093/bioinformatics/btu049

Barrett, T., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., et al. (2012). NCBI GEO: Archive for Functional Genomics Data Sets-Update. *Nucleic Acids Res.* 41 (D1), D991–D995. doi:10.1093/nar/gks1193

Chung, M., Ruan, M., Zhao, N., Koestler, D. C., De Vivo, I., Kelsey, K. T., et al. (2021). DNA Methylation Ageing Clocks and Pancreatic Cancer Risk: Pooled Analysis of Three Prospective Nested Case-Control Studies. *Epigenetics* 16, 1306–1316. doi:10.1080/15592294.2020.1861401

Dedeurwaerder, S., Defrance, M., Bizet, M., Calonne, E., Bontempi, G., and Fuks, F. (2014). A Comprehensive Overview of Infinium HumanMethylation450 Data Processing. *Brief. Bioinform.* 15 (6), 929–941. doi:10.1093/bib/bbt054

Forest, M., O'Donnell, K. J., Voisin, G., Gaudreau, H., Macisaac, J. L., Mcewen, L. M., et al. (2018). Agreement in DNA Methylation Levels from the Illumina 450K Array across Batches, Tissues, and Time. *Epigenetics* 13 (1), 19–32. doi:10.1080/15592294.2017.1411443

Fortin, J.-P., Labbe, A., Lemire, M., Zanke, B. W., Hudson, T. J., Fertig, E. J., et al. (2014). Functional Normalization of 450k Methylation Array Data Improves Replication in Large Cancer Studies. *Genome Biol.* 15 (11), 503. doi:10.1186/s13059-014-0503-2

Fortin, J.-P., Triche, T. J., and Hansen, K. D. (2016). Preprocessing, Normalization and Integration of the Illumina HumanMethylationEPIC Array with Minfi. *Bioinformatics* 33 (4), 558–560. doi:10.1093/bioinformatics/btw691

Guo, S., Diep, D., Plongthongkum, N., Fung, H.-L., Zhang, K., and Zhang, K. (2017). Identification of Methylation Haplotype Blocks Aids in Deconvolution of Heterogeneous Tissue Samples and Tumor Tissue-Of-Origin Mapping from Plasma DNA. *Nat. Genet.* 49 (4), 635–642. doi:10.1038/ng.3805

Hannon, E., Knox, O., Sugden, K., Burrage, J., Wong, C. C. Y., Belsky, D. W., et al. (2018). Characterizing Genetic and Environmental Influences on Variable

DNA Methylation Using Monozygotic and Dizygotic Twins. *Plos Genet.* 14 (8), e1007544. doi:10.1371/journal.pgen.1007544

Horvath, S. (2013). DNA Methylation Age of Human Tissues and Cell Types. *Genome Biol.* 14 (10), R115. doi:10.1186/gb-2013-14-10-r115

Joehanes, R., Just, A. C., Marioni, R. E., Pilling, L. C., Reynolds, L. M., Mandaviya, P. R., et al. (2016). Epigenetic Signatures of Cigarette Smoking. *Circ. Cardiovasc. Genet.* 9 (5), 436–447. doi:10.1161/circgenetics.116.001506

Johansson, Å., Enroth, S., and Gyllensten, U. (2013). Continuous Aging of the Human DNA Methylome throughout the Human Lifespan. *PLoS ONE* 8 (6), e67378. doi:10.1371/journal.pone.0067378

Leek, J. T., Scharpf, R. B., Bravo, H. C., Simcha, D., Langmead, B., Johnson, W. E., et al. (2010). Tackling the Widespread and Critical Impact of Batch Effects in High-Throughput Data. *Nat. Rev. Genet.* 11 (10), 733–739. doi:10.1038/nrg2825

Li, M., Zou, D., Li, Z., Gao, R., Sang, J., Zhang, Y., et al. (2019). EWAS Atlas: a Curated Knowledgebase of Epigenome-wide Association Studies. *Nucleic Acids Res.* 47 (D1), D983–D988. doi:10.1093/nar/gky1027

Li, R., Liang, F., Li, M., Zou, D., Sun, S., Zhao, Y., et al. (2018). MethBank 3.0: a Database of DNA Methylomes across a Variety of Species. *Nucleic Acids Res.* 46 (D1), D288–D295. doi:10.1093/nar/gkx1139

Li, Y., Hamilton, K. J., Perera, L., Wang, T., Gruzdev, A., Jefferson, T. B., et al. (2020). ESR1 Mutations Associated with Estrogen Insensitivity Syndrome Change Conformation of Ligand-Receptor Complex and Altered Transcriptome Profile. *Endocrinology* 161 (6), bqaa050. doi:10.1210/endocr/bqaa050

Liu, Y., Aryee, M. J., Padyukov, L., Fallin, M. D., Hesselberg, E., Runarsson, A., et al. (2013). Epigenome-wide Association Data Implicate DNA Methylation as an Intermediary of Genetic Risk in Rheumatoid Arthritis. *Nat. Biotechnol.* 31 (2), 142–147. doi:10.1038/nbt.2487

Maksimovic, J., Gordon, L., and Oshlack, A. (2012). SWAN: Subset-Quantile within Array Normalization for Illumina Infinium HumanMethylation450 BeadChips. *Genome Biol.* 13 (6), R44. doi:10.1186/gb-2012-13-6-r44

Marabita, F., Almgren, M., Lindholm, M. E., Ruhrmann, S., Fagerström-Billai, F., Jagodic, M., et al. (2013). An Evaluation of Analysis Pipelines for DNA Methylation Profiling Using the Illumina HumanMethylation450 BeadChip Platform. *Epigenetics* 8 (3), 333–346. doi:10.4161/epi.24008

Niu, L., Xu, Z., and Taylor, J. A. (2016). RCP: a Novel Probe Design Bias Correction Method for Illumina Methylation BeadChip. *Bioinformatics* 32 (17), 2659–2663. doi:10.1093/bioinformatics/btw285

Smith, Z. D., and Meissner, A. (2013). DNA Methylation: Roles in Mammalian Development. *Nat. Rev. Genet.* 14 (3), 204–220. doi:10.1038/nrg3354

Teschendorff, A. E., Marabita, F., Lechner, M., Bartlett, T., Tegner, J., Gomez-Cabrero, D., et al. (2013). A Beta-Mixture Quantile Normalization Method for Correcting Probe Design Bias in Illumina Infinium 450 K DNA Methylation Data. *Bioinformatics* 29 (2), 189–196. doi:10.1093/bioinformatics/bts680

Triche, T. J., Weisenberger, D. J., Van Den Berg, D., Laird, P. W., and Siegmund, K. D. (2013). Low-level Processing of Illumina Infinium DNA Methylation BeadArrays. *Nucleic Acids Res.* 41 (7), e90. doi:10.1093/nar/gkt090

Wang, T., Guan, W., Lin, J., Boutaoui, N., Canino, G., Luo, J., et al. (2015). A Systematic Study of Normalization Methods for Infinium 450K Methylation Data Using Whole-Genome Bisulfite Sequencing Data. *Epigenetics* 10 (7), 662–669. doi:10.1080/15592294.2015.1057384

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

Xiong, Z., Li, M., Yang, F., Ma, Y., Sang, J., Li, R., et al. (2020). EWAS Data Hub: a Resource of DNA Methylation Array Data and Metadata. *Nucleic Acids Res.* 48 (D1), D890–D895. doi:10.1093/nar/gkz840

Xiong, Z., Yang, F., Li, M., Ma, Y., Zhao, W., Guoliang, W., et al. (2021). EWAS Open Platform: Integrated Data, Knowledge and Toolkit for Epigenome-wide Association Study. *Nucleic Acids Res.*, gkab972. doi:10.1093/nar/gkab972

Xu, Z., Bolick, S. C. E., Deroo, L. A., Weinberg, C. R., Sandler, D. P., and Taylor, J. A. (2013). Epigenome-wide Association Study of Breast Cancer Using Prospectively Collected Sister Study Samples. *JNCI: J. Natl. Cancer Inst.* 105 (10), 694–700. doi:10.1093/jnci/djt045

Xu, Z., Langie, S. A. S., De Boever, P., Taylor, J. A., and Niu, L. (2017). RELIC: a Novel Dye-Bias Correction Method for Illumina Methylation BeadChip. *BMC Genomics* 18 (1), 4. doi:10.1186/s12864-016-3426-3

Xu, Z., Niu, L., Li, L., and Taylor, J. A. (2016). ENmix: a Novel Background Correction Method for Illumina HumanMethylation450 BeadChip. *Nucleic Acids Res.* 44 (3), e20. doi:10.1093/nar/gkv907

Yousefi, P., Huen, K., Schall, R. A., Decker, A., Elboudwarej, E., Quach, H., et al. (2013). Considerations for Normalization of DNA Methylation Data by Illumina 450K BeadChip Assay in Population Studies. *Epigenetics* 8 (11), 1141–1152. doi:10.4161/epi.26037

Yuan, V., Price, E. M., Del Gobbo, G., Mostafavi, S., Cox, B., Binder, A. M., et al. (2019). Accurate Ethnicity Prediction from Placental DNA Methylation Data. *Epigenetics & Chromatin* 12 (1), 51. doi:10.1186/s13072-019-0296-3

Zannas, A. S., Jia, M., Hafner, K., Baumert, J., Wiechmann, T., Pape, J. C., et al. (2019). Epigenetic Upregulation of FKBP5 by Aging and Stress Contributes to NF-Kb-Driven Inflammation and Cardiovascular Risk. *Proc. Natl. Acad. Sci. USA* 116 (23), 11370–11379. doi:10.1073/pnas.1816847116

Zhang, W., Spector, T. D., Deloukas, P., Bell, J. T., and Engelhardt, B. E. (2015). Predicting Genome-wide DNA Methylation Using Methylation marks, Genomic Position, and DNA Regulatory Elements. *Genome Biol.* 16 (1), 14. doi:10.1186/s13059-015-0581-9

**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.

**Publisher's Note:** All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.