Bioinformatics, 2022, 1–8
https://doi.org/10.1093/bioinformatics/btac436
Advance Access Publication Date: 30 June 2022
Original Paper

Gene expression

InterpolatedXY: a two-step strategy to normalize DNA methylation microarray data avoiding sex bias

Yucheng Wang [1], Tyler J. Gorrie-Stone[2], Olivia A. Grant[3], Alexandria D. Andrayas[4], Xiaojun Zhai [1,*], Klaus D. McDonald-Maier[1] and Leonard C. Schalkwyk[3,*]

[1]School of Computer Science and Electronic Engineering, University of Essex, Colchester CO4 3SQ, UK, [2]Diamond Light Source Ltd., Oxfordshire OX11 0DE, UK, [3]School of Life Sciences, University of Essex, Colchester CO4 3SQ, UK and [4]Institute for Social and Economic Research, University of Essex, Colchester CO4 3SQ, UK

*To whom correspondence should be addressed.

Associate Editor: Christina Kendziorski

Received on September 29, 2021; revised on May 31, 2022; editorial decision on June 26, 2022; accepted on June 28, 2022

Abstract

Motivation: Data normalization is an essential step to reduce technical variation within and between arrays. Due to the different karyotypes and the effects of X chromosome inactivation, females and males exhibit distinct methylation patterns on sex chromosomes; thus, it poses a significant challenge to normalize sex chromosome data without introducing bias. Currently, existing methods do not provide unbiased solutions to normalize sex chromosome data, usually, they just process autosomal and sex chromosomes indiscriminately.

Results: Here, we demonstrate that ignoring this sex difference will lead to introducing artificial sex bias, especially for thousands of autosomal CpGs. We present a novel two-step strategy (interpolatedXY) to address this issue, which is applicable to all quantile-based normalization methods. By this new strategy, the autosomal CpGs are first normalized independently by conventional methods, such as funnorm or dasen; then the corrected methylation values of sex chromosome-linked CpGs are estimated as the weighted average of their nearest neighbors on autosomes. The proposed two-step strategy can also be applied to other non-quantile-based normalization methods, as well as other array-based data types. Moreover, we propose a useful concept: the sex explained fraction of variance, to quantitatively measure the normalization effect.

Availability and implementation: The proposed methods are available by calling the function '*adjustedDasen*' or '*adjustedFunnorm*' in the latest wateRmelon package (https://github.com/schalkwyk/wateRmelon), with methods compatible with all the major workflows, including minfi.

Contact: xzhai@essex.ac.uk or lschal@essex.ac.uk

Supplementary information: Supplementary data are available at *Bioinformatics online*.

1 Introduction

DNA methylation microarrays, such as Infinium HumanMethylation450 BeadChip (Bibikova *et al.*, 2011) and Infinium MethylationEPIC BeadChip (Moran *et al.*, 2016), provide cost-effective and high-throughput measurements of the methylation status over half a million CpG sites across the genome will continue to be the first choice by most DNA methylation-related large cohort studies in the near future. Although whole-genome bisulfite sequencing is recognized as the gold standard to measure the methylation patterns across the human genome, the high costs and technical complexity still pose significant challenges that prevent application to large-scale samples (Villicaña and Bell, 2021). Data normalization is an important prerequisite step to reduce unwanted technical variation. Currently, several normalization methods are available for DNA methylation microarray samples. Among them, peak-based correction (Dedeurwaerder *et al.*, 2011), Beta MIxture Quantile normalization (Teschendorff *et al.*, 2013) and noob (Triche *et al.*, 2013) are all within-array normalization methods however they do not reduce between-array variation. In contrast, dasen (Pidsley *et al.*, 2013) and funnorm (Fortin *et al.*, 2014) are the two most widely used between-array normalization methods, which were reported to be able to effectively reducing the variation between samples. Dasen in the wateRmelon package utilizes quantile normalization to normalize methylated and unmethylated intensities separately, and also addresses Types I and II probes separately. Prior to the normalization steps, there are linear regression procedures in dasen to reduce the density distribution difference between Types I and II probes (Pidsley *et al.*, 2013). The functional normalization employed by funnorm is also an extension to quantile normalization that removes variation explained by a set of selected covariates. In funnorm, the covariates are set as the first two principal components of the control probes, and linear regression is used to determine the proportion of variation explained by the covariates (Fortin *et al.*, 2014).

Females have two copies of the X chromosome, while males have one X chromosome and one Y chromosome. To compensate for the different dosages of the X chromosome genes, one X chromosome in female cells is randomly subjected to inactivation in each cell lineage, with most parts of the inactive X being highly methylated (Cotton *et al.*, 2015; Lyon, 1961; Sharp *et al.*, 2011). As a result of this, the mean methylation values of the X chromosomes between sexes are very different (Grant *et al.*, 2022; McCarthy *et al.*, 2014; Wang *et al.*, 2021). The distinct methylation patterns of sex chromosomes between females and males raise a great challenge to unbiasedly normalize sex chromosome data. The existing between-array normalization methods do not provide good solutions for normalizing sex chromosome data. For example, dasen ignores this issue and normalizes autosomes and sex chromosomes together, while funnorm is designed to normalize male samples and female samples separately for X chromosomes and Y chromosomes. Some DNA methylation-related studies simply remove those probes mapped to the X and Y chromosomes prior to the normalization step and do not include them in the downstream analysis. All these strategies come with their own drawbacks, either through losing some potentially interesting and biologically relevant signals from sex chromosomes or by introducing systematic technical differences between sexes.

Here, we first demonstrate that the existing normalization methods used to handle probes mapped on the X and Y chromosomes lead to introducing artificial sex bias into the normalized data. Then, we present a novel two-step strategy, which is designed to unbiasedly normalize both autosome data and sex chromosome data, is applicable to all quantile-based normalization methods.

2 Materials and methods

2.1 Datasets

Two main datasets (Table 1) were used in this study. The first dataset includes 1195 individuals from the Understanding Society: UK Household Longitudinal Survey (UKHLS). Details about this UKHLS dataset are described by Gorrie-Stone *et al.* (2019). In brief, DNA methylation levels in whole blood within 489 male and 686 female healthy individuals were measured by EPIC array. The UKHLS dataset is available under request from the European Genome-phenome Archive under accession EGAS00001002836 (https://www.ebi.ac.uk/ega/home). Since funnorm was developed and tested on 450k array samples, in this study we produce subsets from GSE142512 (Johnson *et al.*, 2020) to evaluate funnorm. GSE142512 includes 87 individuals with Type 1 diabetes (T1D) and 87 individuals without T1D. The peripheral blood samples were collected from the subjects between 1 and 5 time points, with DNA methylation levels measured by either 450k or EPIC array, further details were documented by Johnson *et al.* (2020). We randomly selected 16 450k samples (12 males and 4 females) from GSE142512 as the dataset one which is used to evaluate the performance of funnorm on small size dataset, and randomly selected 48 450k samples (23 males and 25 females) as dataset two to test funnorm's performance on relatively larger size dataset. For reproducibility, the sample IDs in the two subset datasets are listed in Supplementary Table S1. GSE142512 is publicly available from Gene Expression Omnibus (https://www.ncbi.nlm.nih.gov/geo/).

2.2 DNA methylation data process

The DNA methylation raw data (IDAT files) were read into R by either using *iadd2* function in bigmelon or *read.metharray.exp* function in minfi. The methylation level of any given CpG locus is measured by its beta value which is defined as: $\beta = (M)/(M + U + 100)$, where M is methylated intensity and U is unmethylated intensity for a given CpG loci. Basic quality control steps were performed to identify outliers, as recommended by Gorrie-Stone *et al.* (2019). Further, the reported sexes of samples were checked against the predicted sexes from DNA methylation data by using the *estimateSex* function in watermelon package (Pidsley *et al.*, 2013), which predicts sex by comparing the methylation levels on sex chromosomes (Wang *et al.*, 2021). The original dasen normalization is performed by calling the *dasen* function with default settings in the watermelon package, the original funnorm normalization is performed by calling the *preprocessFunnorm* with default settings in the minfi package (Fortin *et al.*, 2016), which is actually applies *noob* method (Triche *et al.*, 2013) as a first step for background correction and then perform the functional normalization.

All analyses were performed using R 3.6.0 under Linux environment.

2.3 A two-step strategy to unbiasedly normalize DNA methylation samples

The framework of the interpolatedXY strategy is illustrated in Figure 1. The explicit procedures of the proposed new strategy to unbiasedly normalize both autosomal CpGs and sex chromosome-linked CpGs are as follows:

1. Step one: normalize the autosomal CpGs by one of the conventional normalization methods, such as funnorm or dasen. It should be noted, the probes mapped to sex chromosomes should not be included in this step to avoid potential influence.
2. Step two: infer the corrected values of sex chromosome-linked CpGs by looking for their nearest neighbors on autosomes, this is achieved by linear interpolation, here is the very efficient implementation:
   a. Sort the corrected values of autosomal CpGs and build a function F which reflects correspondence of the rank of a CpG to its corrected value: $Corrected\_value_i = F(rank_i)$.
   b. Sort and get the ranks of autosomal CpGs based on their raw values.
   c. Estimate the ranks of sex chromosome-linked CpGs by linear interpolation on the rank distribution from the procedure b.
   d. Put the inferred ranks of sex chromosome-linked CpGs into the function F to get their final corrected values.

The above steps are ideally performed on raw signal intensities (M and U) and on each probe type (IGrn, IRed and II in funnorm, I and II in dasen) individually. After that, the normalized intensities can be converted into beta values as: $\beta = (M)/(M + U + 100)$. We name this strategy as interpolatedXY. When dasen is used to normalize autosomal CpGs in the first step, we call this new normalization method as 'interpolatedXY adjusted dasen'. Similarly, 'interpolatedXY adjusted funnorm' refers to another new normalization method in which the functional normalization is applied in the first step.

Table 1. Characteristics of the datasets used in this study

Name	Array type	Samples (female/male)	Age range (years)	Source
Dataset one	450k	16 (4/12)	0.8–13.6	GSE142512
Dataset two	450k	48 (25/23)	0.8–14.1	GSE142512
UKHLS	EPIC	1195 (686/489)	28–98	UKHLS

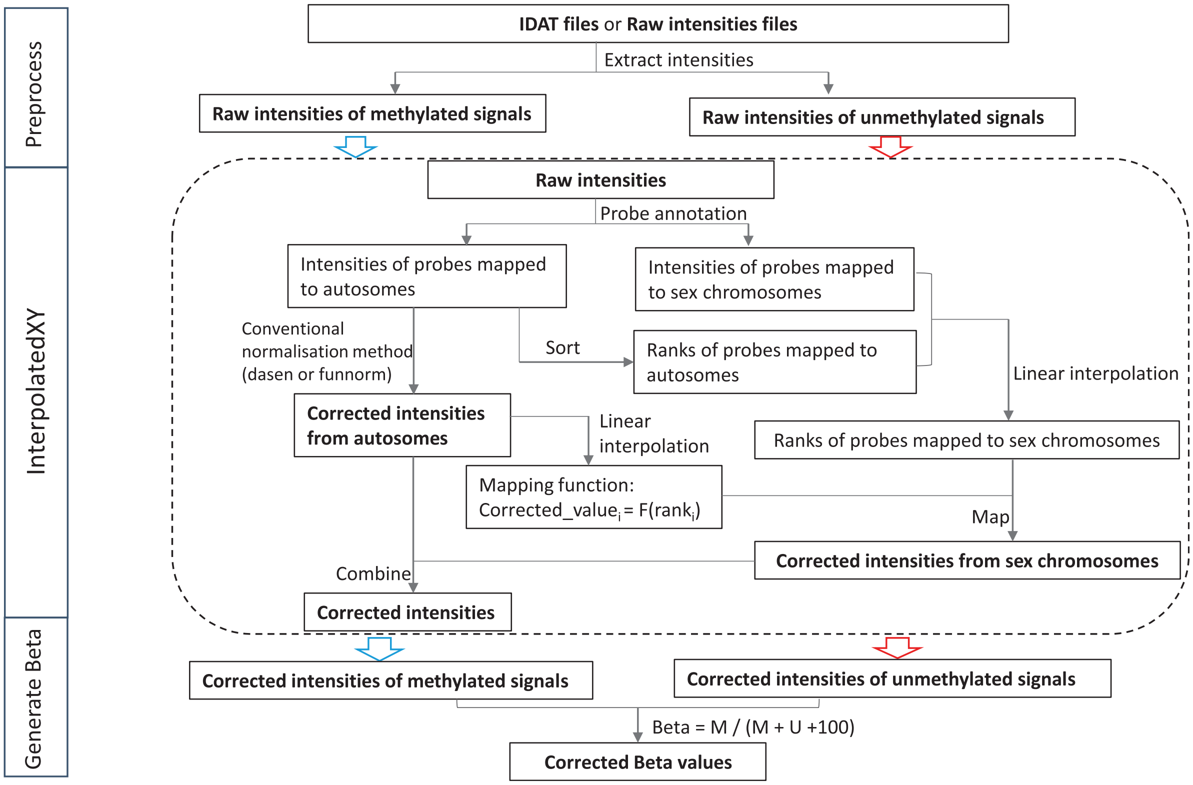
Preprocess
InterpolatedXY
Generate Beta
IDAT files or Raw intensities files
Extract intensities
Raw intensities of methylated signals
Raw intensities of unmethylated signals
Raw intensities
Probe annotation
Intensities of probes mapped to autosomes
Intensities of probes mapped to sex chromosomes
Conventional normalisation method (dasen or funnorm)
Sort
Ranks of probes mapped to autosomes
Linear interpolation
Corrected intensities from autosomes
Linear interpolation
Ranks of probes mapped to sex chromosomes
Mapping function: Corrected_value$_i$ = F(rank$_i$)
Map
Corrected intensities from sex chromosomes
Combine
Corrected intensities
Corrected intensities of methylated signals
Corrected intensities of unmethylated signals
Beta = M / (M + U +100)
Corrected Beta values

Fig. 1. Overview of the interpolatedXY framework. Raw intensities are extracted from IDAT files or intensity text files, then, the raw intensities of methylated and unmethylated signals are processed separately by the interpolatedXY procedure. Above all, chromosome annotation is performed on all probes to separate the raw input intensities into autosome-linked signals and sex chromosome-linked signals. These autosomes linked intensity signals are then normalized by a conventional normalization method, such as dasen or funnorm. These sex chromosomes linked intensity signals are corrected as approximations of their nearest neighbors on autosomes, this is achieved by: (i) obtaining their approximate rankings by linear interpolation on the raw intensity distribution of autosomes mapped probes; (ii) constructing a mapping function which deduces the corrected intensity value from its intensity rank by linear interpolation on the corrected intensities of autosome mapped probes. Finally, the corrected beta values are deduced from the corrected intensities signals

2.4 Performance evaluation for the interpolation approach

The proposed new approach infers the corrected values of sex chromosome-linked CpGs by linear interpolation on autosomal CpGs. To investigate whether the inferred data are accurate and reliable, we need a gold standard to evaluate the estimation accuracy. Females and males have very different methylation patterns on sex chromosomes, that is the main reason that we avoid normalizing female samples and male samples together, with autosomes and sex chromosomes treated indiscriminately. However, when the targeted dataset includes only unisexual samples (only females or only males), then the sex chromosomes should be normalized together with other autosomes.

Inspired by this, we designed single-sex groups: one that includes only female samples and the second that consists of only male samples. First, the two groups are both normalized by conventional methods (e.g. dasen and funnorm) with the sex chromosomes being treated as general autosomes, thus the corrected values of those sex chromosome-linked CpGs could serve as the golden references (i.e. expected values). Second, by our proposed interpolation approach, we infer the corrected values of sex chromosome-linked CpGs by interpolating on the normalized values of the autosomal CpGs. Last, the interpolated values are compared with their corresponding reference values. Root mean squared error (RMSE), which is sensitive to outliers, is used here to measure the deviations from the inferred values to their expected values:

$$RMSE = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(\beta_i - \hat{\beta}_i)^2} \tag{1}$$

where β_i is the methylation beta value of the ith CpG, $\hat{\beta}_i$ represents the expected methylation beta value of the ith CpG, m represents the total number of CpGs studied.

2.5 Evaluation of the technical sex biases

The original dasen performs quantile normalization with autosomal CpGs and sex chromosome CpGs processed together even when the dataset to be normalized is composed of both females and males. To investigate whether such an approach would introduce artificial sex biases, we compared the normalization results of the UKHLS dataset generated by the original dasen and the interpolatedXY adjusted dasen.

The human methylome is not constant but responsive to many internal and external factors, such as genetic backgrounds and environmental factors (van Dongen *et al.*, 2016). As a result, the overall variance of the measured methylation values across all the CpG sites in the studied population can be described as:

$$V_{total} = \frac{1}{n}\frac{1}{m}\sum_{i=1}^{n}\sum_{j=1}^{m}(\beta_{ij} - \overline{\beta_j})^2 \tag{2}$$

Where V_{total} represents the total variance of the studied samples, n is the total number of all samples, m is the total number of studied CpGs, β_{ij} represents the methylation beta value of the jth CpG in the ith sample, $\overline{\beta_j}$ represents the mean methylation beta value of the jth CpG across all samples. Theoretically, we can then split the overall variance into the following two parts:

$$V_{total} = V_{biological} + V_{technical} = \frac{1}{n}\sum_{i=1}^{n}(V_i) \tag{3}$$

The first part $V_{biological}$ represents variance caused by meaningful biological reasons, such as cell types, age, gender, health status and other reasonable factors. The second part $V_{technical}$ represents variance resulting from technical issues, such as batch effect, random fluctuation and other unknown issues.

$$V_{biological} = V_{cell\_type} + V_{age} + V_{sex} + V_{others} \tag{4}$$

$$V_{technical} = V_{batch} + V_{random} + V_{unknown} \tag{5}$$

Sex is one of the major biological factors which influences the methylation status of many autosomal CpGs, as a result, hundreds of autosomal CpGs have been reported showing significant different methylation levels between sexes (Grant *et al.*, 2022; McCarthy *et al.*, 2014; Yousefi *et al.*, 2015). The fraction of variances which are explained by sex can be deduced as follows:

$$\begin{aligned} F_{sex} &= \frac{V_{sex}}{V_{total}} \\ &= 1 - \frac{n_{females} V_{total\_in\_females} + n_{males} V_{total\_in\_males}}{(n_{females} + n_{males}) V_{total}} \end{aligned} \tag{6}$$

Ideally, a good normalization method should be able to not only greatly reduce the variances that are resulted from technical issues ($V_{technical}$), but also need to keep variances which have meaningful biological reasons ($V_{biological}$). This means, after the normalization process, the overall variance should be reduced significantly while the sex explained fraction of variance should be increased. In this article, to study the potential sex bias introduced by the mix normalization method dasen, we compared the mean variance and the fraction of sex explained variances of the methylation values of CpGs after no normalization (raw beta values), dasen normalization and interpolatedXY adjusted dasen normalization within the three chromosome groups (i.e. autosomes, X chromosomes and Y chromosomes).

2.6 Artifactual sex differences

If the conventional mixed normalization approaches do introduce systematic artificial sex biases into the autosomal CpGs, then some autosomal CpGs could be falsely sex-associated. Epigenome-wide association studies (EWAS) are commonly used to systematically assess the association between DNA methylation levels at genetic loci across the genome and a phenotype of interest. In this study, we apply EWAS to identify sex-associated CpG sites and then compare the EWAS results resulted from different pre-process approaches.

To perform EWASs for sex, the *champ.dmp* function in champ package (Tian *et al.*, 2017), which utilizes linear regression and *F*-test to identify differentially methylated positions is applied in this study to identify sex-associated CpGs. After Bonferroni multiple comparison correction, those CpG sites with P-value < 0.05 were selected as significantly sex-associate. For simplicity and better comparison, we do not include age, cell type proportions and other covariates within the EWASs.

2.7 Comparison of the original funnorm and the interpolatedXY adjusted funnorm

Funnorm is reported to be suitable for normalizing methylation data with substantial global differences. The main difference between the original funnorm and the proposed interpolatedXY adjusted funnorm is how to normalize the methylation values of sex chromosome-linked CpGs. The original funnorm is designed to normalize X chromosomes separately and differently with Y chromosomes, as well as processes female samples and male samples separately. In contrast, the interpolatedXY adjusted funnorm does not require prior sex annotations and process both genders equally, which generates the corrected values of sex chromosome-linked CpGs by interpolation on the normalized values of autosomal CpGs.

To compare the normalization effects on sex chromosome data between the original funnorm and the adjusted funnorm, we studied both the density distributions and the variances of the methylation values of CpG sites after no normalization (raw beta values), funnorm normalization and adjusted funnorm normalization within three chromosome groups (i.e. autosomes, X chromosomes and Y chromosomes) in two 450k datasets. The first dataset (dataset one) includes 12 male samples and 4 female samples, while the second dataset (dataset two) contains 23 male samples and 25 female samples.

3 Results

3.1 Estimation using the interpolation approach

We first investigated the performance of the interpolation approach employed by the interpolatedXY adjusted funnorm method. The deviations from the inferred values by the interpolation approach to their corresponding reference values are measured by RMSE. As it can be seen from Figure 2, the resulting RMSEs are all very small, especially for those in both X chromosomes and male Y chromosomes: the mean RMSE of X chromosome-linked CpGs is 1.15e−05 (SD = 8.7e−06) in females and is 1.11e−05 (SD = 4.8e−06) in male samples, while the mean RMSE of estimations for male Y chromosomes is 6.61e−06 (SD = 3.2e−06). Though the RMSEs of Y chromosome-linked CpGs in females are slightly higher (mean = 8.98e−04, SD = 6.0e−04), they are still very subtle. With the knowledge that females do not carry Y chromosomes, and those observed signal intensities result from background noises and non-specific hybridization, there is no need to look much into the methylation values of female Y chromosomes. In the same way, we could observe similar performances of the interpolation approach employed by the interpolatedXY adjusted dasen method (Supplementary Fig. S1).

In summary, the above results demonstrate the proposed interpolation approach provides accurate and robust estimations for the corrected values of sex chromosome-linked CpGs.

3.2 Artificial sex biases are introduced into autosomal CpGs by the conventional mixed normalization method

The first round of the UKHLS dataset (Gorrie-Stone *et al.*, 2019) includes 1175 whole blood samples whose DNA methylation levels were measured using the EPIC array. After quality control, 685 female samples and 486 male samples were kept for this analysis. To study the normalization effects, the variance of beta values with three different pre-processing methods (no-normalization, dasen and interpolatedXY adjusted dasen) are compared within three different chromosome groups (i.e. autosomes, X chromosomes and Y chromosomes) separately. As shown in Figure 3, both dasen and adjusted dasen significantly (Wilcoxon signed-rank test, P-value $<$ 2.2e−16) reduce the variance in all three chromosome groups. For instance, the mean variance of autosomes in both sexes decreased from around 0.0025 in non-normalized beta values to about 0.0018 after either dasen or adjusted dasen normalization. The beta values

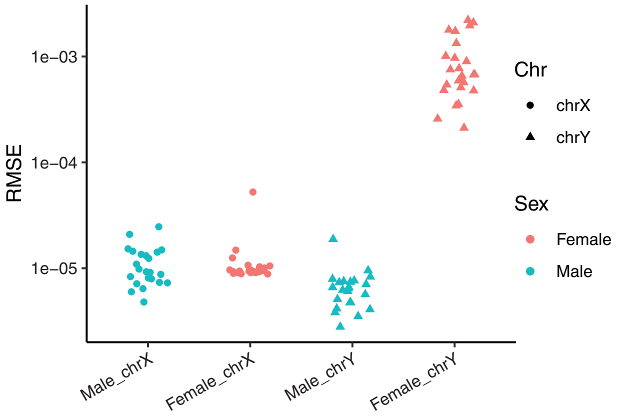

Fig. 2. Difference between interpolated values and expected values within the adjusted funnorm. RMSEs are grouped in four categories: male X chromosomes, female X chromosomes, male Y chromosomes and female Y chromosomes. Female samples are in red colour and male samples are in blue colour. Dots represent X chromosomes, while triangles represent Y chromosomes (A color version of this figure appears in the online version of this article.)

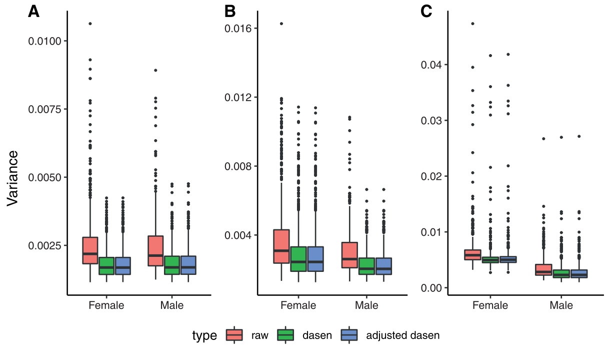

Fig. 3. Variance comparisons in the UKHLS dataset. Boxplots comparing the variance of methylation beta values with three different pre-processing methods (i.e. no normalization, dasen normalization and adjusted dasen normalization) in autosomes (**A**), X chromosomes (**B**) and Y chromosomes (**C**). Females and males are dealt with separately

density plots also demonstrate that both dasen and adjusted dasen greatly reduce the distribution variation (e.g. Supplementary Fig. S2). However, the difference in normalization effects between dasen and adjusted dasen is not significant from the variance level.

Table 2 describes the sex explained fraction of variance between three methods in three chromosome categories. We can see that the sex explained variance in sex chromosomes by the three methods all exceeds 70%, while it accounts to only around 0.5% in autosomes. That is in line with our expectation, as sex is a dominant factor causing difference in methylation levels of sex chromosomes, while the majority of autosomal CpGs are not influenced by sex. Interestingly, the sex explained fraction of variance of raw beta values in autosomes is 0.34%, it rises to 0.45% after normalizing by the adjusted dasen, indicating the adjusted dasen method retained the meaningful biological difference when reducing technical variances (Fig. 3A). However, the sex explained variance is much higher (0.57%) by normalizing with the original dasen, can we conclude that the original dasen is better than the adjusted dasen to retain meaningful biological difference? On the contrary, these results indicate the original dasen has introduced artificial sex bias into to the normalized data. Combining the facts that only autosomal CpGs were included to compute the variance, and the difference in normalizing the autosomal CpGs between the two methods is that the correction of autosomal CpGs is affected by the enrolling of sex chromosome data within the original dasen procedures, but not influenced within the adjusted dasen method. We can conclude that the observed higher fraction (sex explained fraction of variance in autosomes) with the original dasen normalization is partly driven by the involvement of sex chromosome data, and this higher figure (i.e. than the adjusted dasen) indicates that technical sex biases have been introduced into to autosomal CpGs by the original dasen.

3.3 Confirmation of the introduced sex biases

We performed EWASs of sex based on autosomal beta values of UKHLS samples with three different pre-processing: no normalization, dasen normalization and interpolatedXY adjusted dasen normalization. The identified number of sex significant (Bonferroni *P*-value < 0.05) differentially methylated positions (saDMPs) are shown in Figure 4.

As illustrated in the Venn diagram (Fig. 4A), there are 10 778 CpG sites been identified as saDMPs in the raw data, with 96.7% of them (10 427) also been captured after adjusted dasen normalization. In addition, compared to raw data, the adjusted dasen approach enables the identification of another 4201 saDMPs. Once again, these results demonstrate that while the adjusted dasen greatly reduces the variation of beta values (Fig. 3A), it preserves the meaningful biological differences.

We found a total of 32 929 saDMPs after the original dasen normalization, which is more than three times the number with no normalization or 2.25 times the number with adjusted dasen normalization. Even so, 1600 CpGs which are identified by both no normalization and adjusted dasen normalization, are missed by the original dasen method. When comparing the dasen and adjusted dasen (Fig. 4B), there are 12 021 saDMPs shared between the two methods. Interestingly, among the 20 908 dasen-specific saDMPs, 96.0% of them (20 070) have higher methylation values in males than that in females. On the contrary, 2318 out of the 2607 adjusted dasen-specific saDMPs (88.9%) show higher methylation values in females than males. Again, with the fact that the interpolatedXY adjusted dasen only differs from the original dasen by not enrolling sex chromosome data when normalizing the autosomal data, the above results suggest the original dasen did introduce artificial sex biases into autosomal CpGs by making the methylation values of many CpGs slightly higher in male samples and lower in female samples. This explains why nearly all the dasen-specific saDMPs have higher methylation values in male samples, and there are more than two thousand CpG sites which have higher methylation values in female samples that were identified as significant saDMPs by the adjusted dasen approach but missed by the original dasen.

3.4 InterpolatedXY adjusted funnorm provides better normalization results for sex chromosome-linked CpGs than the original funnorm

Since the original funnorm has two different designs to deal with different size datasets, we compared the normalization effects between the original funnorm and the interpolatedXY adjusted funnorm in two datasets. The adjusted funnorm does not differ from the original funnorm in normalizing the autosomal CpGs, so the corrected values of autosome data from the two methods are the same, we can thus observe identical results for autosomal CpGs by the two methods (Figs 4C and 5B, Table 3, Supplementary Fig. S3B and C and Table S2).

For the X chromosome-linked CpGs, when applied to small datasets, whose number of female samples or male samples is <10, such as dataset one, funnorm is designed to normalize female X chromosomes and male X chromosomes together by the functional normalization. Compared to the non-normalized raw beta values, the density distributions of the corrected data generated by funnorm turn out to be much discordant in both female samples and male samples (Fig. 5E). On the contrary, after the adjusted funnorm normalization, the density distributions become more consistent in both sexes (Fig. 5F). We can also observe the same trends from the bar plots in Figure 6B, the original funnorm greatly increases the variance in both sex groups, while the adjusted funnorm keeps the variance low. Furthermore, the sex explained fraction of variance was reduced to 82.8% by the original funnorm, which is 92.7% in raw data and 93.0% after the adjusted funnorm normalization (Table 2). Taken together, the above results indicate that the original funnorm is actually adding technical variation into the methylation data of X chromosomes for those small sample size datasets.

When applied to larger datasets, such as in the case of dataset two, funnorm performs separate functional normalizations on female X chromosomes and male X chromosomes, with the underlying consideration that females and males have very different methylation patterns on X chromosomes. When comparing the normalization effects between the original funnorm and the adjusted funnorm based on dataset two, we did not observe any significant differences in the methylation profiles of X chromosomes (Supplementary Figs S3, S4 and Table S2).

For the Y chromosome-linked CpGs, the original funnorm does not use the functional normalization as it does on other chromosomes, such as autosomes. Instead, only quantile normalization is employed by the original funnorm to normalize the Y chromosome data, and with female samples and male samples processed separately. This may explain why the sex explained variance within the original funnorm is much higher (i.e. 97.7%) than that in the raw data (i.e. 88.5%) and adjusted funnorm (i.e. 89.1%) (Supplementary Table S2). We can also observe similar trend from Table 3. These results suggest the separate normalization strategy employed by the

Table 2. The fraction of variance explained by sex in the UKHLS dataset with no normalization (raw), dasen normalization, interpolatedXY adjusted dasen normalization and interpolatedXY adjusted funnorm normalization

Fraction of variance explained by sex (%)	Raw	Dasen	Adjusted dasen	Adjusted funnorm
Autosomes	0.34	0.57	0.45	0.46
X chromosome	73.18	77.24	77.57	76.93
Y chromosome	85.34	87.64	87.50	88.82

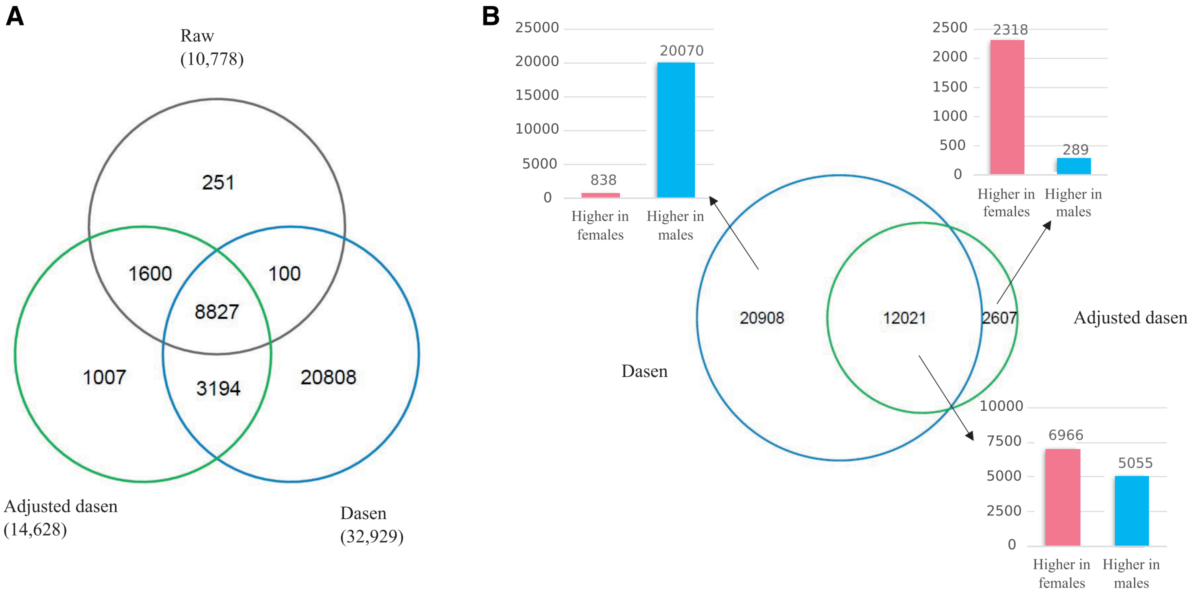

Fig. 4. EWAS results of UKHLS dataset. (**A**) The Venn diagram shows the number of unique and shared saDMPs between three approaches: no normalization (raw), dasen normalization and adjusted dasen normalization. (**B**) The Euler diagram describes the number of unique and shared saDMPs between dasen normalization and adjusted dasen normalization, with the three bar plots showing the number of CpGs which have higher methylation values in females (red) or males (blue) in three categories separately (A color version of this figure appears in the online version of this article.)

original funnorm will increase the difference between the two sex groups, and thus introduce artificial technical bias.

3.5 Comparison between the interpolatedXY adjusted funnorm and interpolatedXY adjusted dasen

We have demonstrated that the fraction of variance explained by sex is very useful to measure the normalization effects for different methods and have also shown that the adjusted the dasen and the adjusted funnorm are both superior than their original versions. Then we compared their normalization effects on a large healthy population: the UKHLS dataset ($n = 1171$). The results are shown in Table 2, the first obvious observation is that both the adjusted dasen and the adjusted funnorm clearly increased the fraction of variance explained by sex in all chromosome groups (i.e. autosomes, X chromosome and Y chromosome) than the raw data, demonstrating that the use of either normalization method is beneficial and worthwhile. As compared to the two adjusted normalization methods, we can see their effects are comparable in the studied dataset (Table 2): the adjusted funnorm marginally outperforms the adjusted dasen in normalizing the autosome data (0.46% versus 0.45%) and Y chromosome data (88.82% versus 87.5%), while the adjusted dasen is slightly better in normalizing the X chromosome data (77.57% versus 76.93%).

4 Discussion

We have described a two-step sex-unbiased data normalization strategy for normalizing DNA methylation microarray samples, which can be applied into almost all quantile-based normalization methods, such as dasen and funnorm. By this strategy, the autosomal CpGs are normalized independently and separately from the sex chromosome CpGs, while the corrected values of sex chromosomes CpGs are estimated as the weighted average of the corrected methylation values of their nearest neighbor atusosomal CpGs.

The two steps are necessary. Since the average methylation levels of CpGs on X chromosome in females are very different from that in males, normalizing them together with the autosomal CpGs, especially by the quantile-based methods, will introduce technical biases for both autosomes and sex chromosomes. By comparing the normalization effects of the original dasen and the interpolatedXY adjusted dasen, we confirmed that the technical sex biases were introduced into the autosomal CpGs by the mix normalization approach (original dasen)—with the sex explained fraction of variance in autosomes rising to 0.57% from 0.44% in the adjusted dasen normalized data. We further propose a rational explanation for this (Fig. 7): within the quantile normalization steps in dasen, there are procedures to sort and return ranks for all the probes, as the mean methylation values of the most X chromosome-linked CpGs in females are higher than nearly half of the autosomal CpGs, whereas the methylation values of the corresponding positions in males are relatively low, thus the quantile normalization algorithm used to make all studied samples fit into a same distribution creating a systematic negative shift for many autosomal CpGs (their methylation values are lower than most X chromosome-linked CpGs) in females and a systematic positive shift for those CpGs in males. As a result of this, when we perform EWAS to look for autosomal sex-associated CpGs, the original dasen approach identified more than two times the number as identified by the adjusted dasen or non-

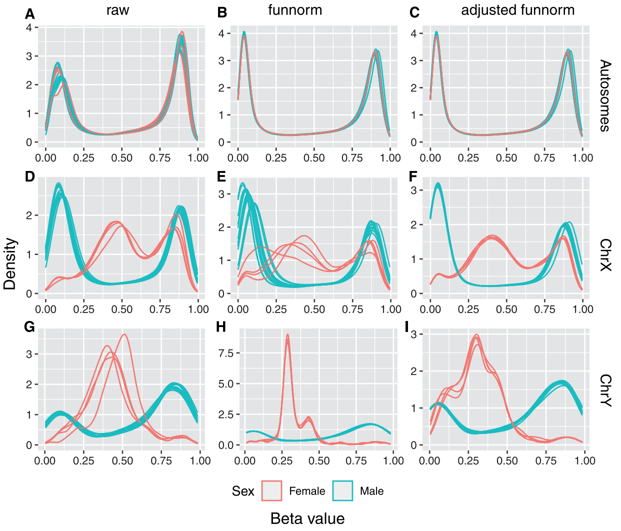

Fig. 5. Comparisons in methylation beta value density distributions for dataset one. The three columns list results from raw data (left column: A, D and G), funnorm normalized data (middle column: B, E and H) and the adjusted funnorm normalized data (right column: C, F and I). The three rows show density distributions of autosomal CpGs (first row), X chromosome-linked CpGs (second row) and Y chromosome-linked CpGs (third row). Red lines represent females and blue lines represent males (A color version of this figure appears in the online version of this article.)

Table 3. The fraction of variance explained by sex in the dataset one ($n = 16$) with no normalization (raw), funnorm normalization and interpolatedXY adjusted funnorm normalization

Fraction of variance explained by sex (%)	Raw	Funnorm	Adjusted funnorm
Autosomes	9.48	10.93	10.93
X chromosome	92.68	82.82	92.99
Y chromosome	91.48	97.09	93.89

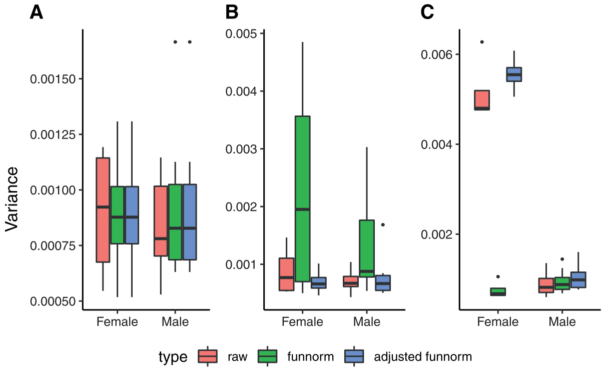

Fig. 6. Variance comparisons in the dataset one. Boxplots comparing the variance of methylation beta values with three different pre-processing methods (i.e. no normalization, dasen normalization and adjusted dasen normalization) in autosomes (**A**), X chromosomes (**B**) and Y chromosomes (**C**). Females and males are dealt with separately

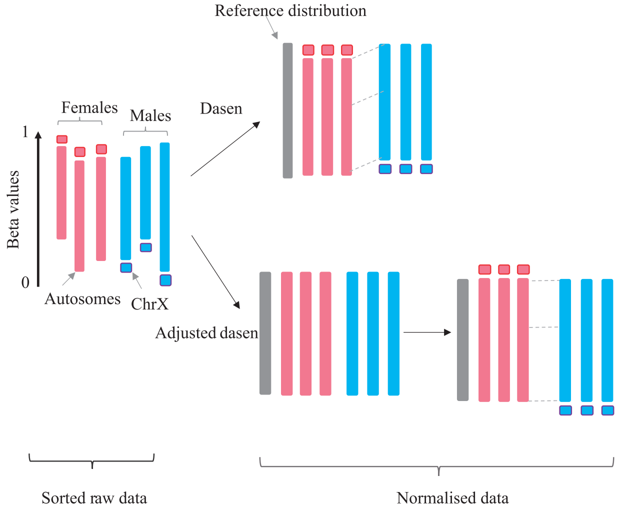

Fig. 7. A simplified schematic diagram illustrates the difference in the normalization process between the original dasen and the interpolatedXY adjusted dasen. The original dasen normalizes autosomes and sex chromosomes together, the mean methylation values of most X chromosome-linked CpGs in females are higher than nearly half of the autosomal CpGs, whereas the values of the corresponding locus in males are relatively very low, thus the quantile normalization algorithm employed by dasen to make all studied samples fit into a same distribution creating a systematic shift for many autosomal CpGs in two sexes. The adjusted dasen manages to avoid such an issue by doing quantile normalization in autosomes separately and independently with sex chromosomes, and infer the corrected values of sex chromosomes by interpolating on autosomes. Red denotes female sample and blue denotes male sample, the long bar represents sorted autosomal CpGs and the short bar represents sorted X chromosome-linked CpGs (A color version of this figure appears in the online version of this article.)

normalized data. Moreover, 96.0% of the dasen-specific saDMPs show higher methylation values in male samples than in female samples, in contrast, the majority of the 2607 CpGs missed by the original dasen but identified by the adjusted dasen have higher methylation values in female samples than male samples.

Estimation of the corrected values for sex chromosomes CpGs by looking at their nearest neighbors on autosomes is made both possible and reliable by the fact that DNA methylation microarrays simultaneously measure over half a million CpG sites across the genome, and only a relatively small portion (i.e. 2.3% in EPIC and 2.4% in 450K) is mapped on the sex chromosomes. Here in this study, we have demonstrated that the linear interpolation approach provides both accurate and robust estimations for the sex chromosome data, with the mean RMSE < 1.2e−5.

Funnorm is favored for normalizing methylation data with substantial global differences, such as cancer samples (Fortin *et al.*, 2014). With the consideration that females and males have distinct methylation patterns for sex chromosomes, funnorm has very explicit rules to normalize X chromosomes and Y chromosomes differently. Within the functional normalization in funnorm, there is a regression step to infer the explainable technical variants based on control probes. The authors may have considered the regression models would be less accurate in the circumstance of only few samples, so funnorm is designed to perform functional normalizations on female X chromosomes and male X chromosomes together when the number of either female samples or male samples is <10. Our results in Section 3.4 have clearly shown that such a mix normalization approach is destructive to the methylation profiles of X chromosomes in both females and males. Though to do functional normalization on females and males separately is a way to avoid such an issue, it may also introduce potential systematic technical bias between the two separate groups.

For the Y chromosome-linked CpGs, the original funnorm does not actually perform the functional normalization as it does on other chromosomes, instead it performs only quantile normalizations on Y chromosomes, and processes female samples and male samples separately. As the proposed interpolatedXY adjusted funnorm could provide near-perfect estimations for corrected values generated by functional normalization, it could be particularly useful for studies that focus on sex chromosomes DNA methylation data, especially when the methylation difference between the studied groups that are known to be very different. Moreover, by the adjusted funnorm method, the corrected values of sex chromosome-linked CpGs are produced by linear interpolating on the distribution of autosomal CpGs, so in theory, they are more comparable with the autosomal CpGs.

In this article, we not only present a novel two-step strategy to unbiasedly normalize DNA methylation microarray samples, but also provide a useful concept—the fraction of variance explained by sex, to quantitively measure the normalization effect. Sex is an important biological factor that not only determines the methylation status of sex chromosomes, but also influences many autosomal CpGs. A good candidate normalization method should not only be able to greatly reduce the technical variation between samples, but also should preserve the meaningful variation that has biological reasons (e.g. sex). Even though quantile normalization has been widely employed by several DNA methylation normalization methods, such as SWAN (Maksimovic *et al.*, 2012), dasen (Pidsley *et al.*, 2013) and funnorm (Fortin *et al.*, 2014). There are still concerns about whether the use of between-array normalization methods could bring enough benefits to counterbalance the potential impairment of data quality (Dedeurwaerder *et al.*, 2014). Here, in this study, we demonstrated that the interpolatedXY adjusted dasen and the interpolatedXY adjusted funnorm are two good normalization method candidates, they are able to not only greatly reduce technical variation but also retain the meaningful biological difference, which will be very useful for large cohort EWAS projects.

We believe that the proposed novel two-step strategy may have wider application outside of DNA methylation microarrays and could even be applied in more broader technologies such as RNA-Seq.

5 Conclusion

The proposed two-step strategy of interpolatedXY allows for the normalization of autosomal data and sex chromosome data without bias. The two steps are necessary and reliable, the interpolatedXY approach infers the normalized methylation beta values of sex chromosome-linked CpGs with deviation (RMSE) of around 1.15e−05 to their expected values. With the introducing of the interpolatedXY, the adjusted dasen and the adjusted funnorm both show superior performance than their original versions, i.e. the adjustedDasen avoids the risk of introducing sex bias into the autosomal data when normalizing mixed-sex samples compared to the original dasen; the adjustedFunnorm reduces artificial sex bias in the sex chromosome data as compared to the original funnom. In addition, the sex explained variance analysis reveals the two between-array normalization methods, dasen and funnorm, both enable retaining the meaningful biological difference while reducing technical variation.

Acknowledgements

The authors acknowledge the use of the High-Performance Computing Facility (Ceres) and its associated support services at the University of Essex in the completion of this work.

Funding

The UK Household Longitudinal Study is led by the Institute for Social and Economic Research at the University of Essex and funded by the Economic and Social Research Council [ES/M008592/1]. L.C.S. is funded by the Medical Research Council [MR/R005176/1]. K.D.M. is funded by the Economic and Social Research Council [ES/M010236/1] and the Engineering and Physical Sciences Research Council [EP/P017487/1, EP/R02572X/1 and EP/V000462/1]. X.Z. is funded by the Engineering and Physical Sciences Research Council [EP/V034111/1]. Y.W. is supported by faculty studentship from the University of Essex.

Conflict of Interest: The authors declare that they have no conflict interests.

Data availability

The UKHLS dataset is available under request from the European Genome-phenome Archive under accession EGAS00001002836 (https://www.ebi.ac.uk/ega/home). GSE142512 is availiable in Gene Expression Omnibus (https://www.ncbi.nlm.nih.gov/geo/), and can be accessed with GSE142512.

References

Bibikova,M. *et al.* (2011) High density DNA methylation array with single CpG site resolution. *Genomics*, **98**, 288–295.

Cotton,A.M. *et al.* (2015) Landscape of DNA methylation on the X chromosome reflects CpG density, functional chromatin state and X-chromosome inactivation. *Hum. Mol. Genet.*, **24**, 1528–1539.

Dedeurwaerder,S. *et al.* (2011) Evaluation of the Infinium Methylation 450K technology. *Epigenomics*, **3**, 771–784.

Dedeurwaerder,S. *et al.* (2014) A comprehensive overview of Infinium Human Methylation450 data processing. *Brief. Bioinform.*, **15**, 929–941.

Fortin,J.-P. *et al.* (2014) Functional normalization of 450k methylation array data improves replication in large cancer studies. *Genome Biol.*, **15**, 503.

Fortin,J.-P. *et al.* (2016) Preprocessing, normalization and integration of the Illumina HumanMethylationEPIC array with minfi. *Bioinformatics*, **33**, 558–560.

Gorrie-Stone,T.J. *et al.* (2019) Bigmelon: tools for analysing large DNA methylation datasets. *Bioinformatics*, **35**, 981–986.

Grant,O.A. *et al.* (2022) Characterising sex differences of autosomal DNA methylation in whole blood using the Illumina epic array. *Clin. Epigenet.*, **14**, 1–16.

Johnson,R.K. *et al.* (2020) Longitudinal DNA methylation differences precede type 1 diabetes. *Sci. Rep.*, **10**, 3721.

Lyon,M.F. (1961) Gene action in the X-chromosome of the mouse (*Mus musculus* L.). *Nature*, **190**, 372–373.

Maksimovic,J. *et al.* (2012) SWAN: subset-quantile within array normalization for Illumina Infinium HumanMethylation450 BeadChips. *Genome Biol.*, **13**, R44.

McCarthy,N.S. *et al.* (2014) Meta-analysis of human methylation data for evidence of sex-specific autosomal patterns. *BMC Genomics*, **15**, 981.

Moran,S. *et al.* (2016) Validation of a DNA methylation microarray for 850,000 CpG sites of the human genome enriched in enhancer sequences. *Epigenomics*, **8**, 389–399.

Pidsley,R. *et al.* (2013) A data-driven approach to preprocessing Illumina 450K methylation array data. *BMC Genomics*, **14**, 293.

Sharp,A.J. *et al.* (2011) DNA methylation profiles of human active and inactive X chromosomes. *Genome Res.*, **21**, 1592–1600.

Teschendorff,A.E. *et al.* (2013) A beta-mixture quantile normalization method for correcting probe design bias in Illumina Infinium 450 k DNA methylation data. *Bioinformatics*, **29**, 189–196.

Tian,Y. *et al.* (2017) ChAMP: updated methylation analysis pipeline for Illumina BeadChips. *Bioinformatics*, **33**, 3982–3984.

Triche,T.J. *et al.* (2013) Low-level processing of Illumina Infinium DNA methylation BeadArrays. *Nucleic Acids Res.*, **41**, e90.

van Dongen,J. *et al.*; BIOS Consortium. (2016) Genetic and environmental influences interact with age and sex in shaping the human methylome. *Nat.Commun.*, **7**, 11115.

Villicaña,S. and Bell,J.T. (2021) Genetic impacts on DNA methylation: research findings and future perspectives. *Genome Biol.*, **22**, 1–35.

Wang,Y. *et al.* (2021) DNA methylation-based sex classifier to predict sex and identify sex chromosome aneuploidy. *BMC Genomics*, **22**, 484.

Yousefi,P. *et al.* (2015) Sex differences in DNA methylation assessed by 450K BeadChip in newborns. *BMC Genomics*, **16**, 1–12.