Data processing choices can affect findings in differential methylation analyses: an investigation using data from the LIMIT RCT

Jennie Louise Corresp., [1, 2] , **Andrea R Deussen** [3] , **Jodie M Dodd** [3, 4]

[1] Discipline of Obstetrics & Gynaecology and The Robinson Research Institute, The University of Adelaide, Adelaide, Australia

[2] Adelaide Health Technology Asseessment, The University of Adelaide, Adelaide, Australia

[3] Discipline of Obstetrics & Gynaecology and The Robinson Research Institute, The University of Adelaide, Adelaide, South Australia, Australia

[4] Department of Perinatal Medicine, Women's and Babies Division, The Women's and Children's Hospital, Adelaide, South Australia, Australia

Corresponding Author: Jennie Louise
Email address: jennie.louise@adelaide.edu.au

Objective: A wide array of methods exist for processing and analysing DNA methylation data. We aimed to perform a systematic comparison of the behaviour of these methods, using cord blood DNAm from the LIMIT RCT, in relation to detecting hypothesised effects of interest (intervention and pre-pregnancy maternal BMI) as well as effects known to be spurious, and known to be present. Methods: DNAm data, from 645 cord blood samples analysed using Illumina 450K BeadChip arrays, were normalised using three different methods (with probe filtering undertaken pre- or post- normalisation). Batch effects were handled with a supervised algorithm, an unsupervised algorithm, or adjustment in the analysis model. Analysis was undertaken with and without adjustment for estimated cell type proportions. The effects estimated included intervention and BMI (effects of interest in the original study), infant sex and randomly assigned groups. Data processing and analysis methods were compared in relation to number and identity of differentially methylated probes, rankings of probes by p value and log-fold-change, and distributions of p values and log-fold-change estimates. Results: There were differences corresponding to each of the processing and analysis choices. Importantly, some combinations of data processing choices resulted in a substantial number of spurious ‘significant’ findings. We recommend greater emphasis on replication and greater use of sensitivity analyses.

Data Processing Choices Can Affect Findings in Differential Methylation Analyses: An Investigation Using Data from the LIMIT RCT

Authors

Jennie Louise[1,2]
Andrea R Deussen[1]
Jodie M Dodd[1,3]

Affiliations

1. Discipline of Obstetrics & Gynaecology and The Robinson Research Institute, The University of Adelaide, Adelaide, South Australia, AUSTRALIA
2. Adelaide Health Technology Asseessment, The University of Adelaide, Adelaide, South Australia, AUSTRALIA
3. Department of Perinatal Medicine, Women's and Babies Division, The Women's and Children's Hospital, Adelaide, South Australia, AUSTRALIA.

Corresponding Author

Dr Jennie Louise
The University of Adelaide
Women's and Children's Hospital
72 King William Rd
North Adelaide, South Australia, AUSTRALIA 5006
Email: jennie.louise@adelaide.edu.au

Abstract

Objective: A wide array of methods exist for processing and analysing DNA methylation data. We aimed to perform a systematic comparison of the behaviour of these methods, using cord blood DNAm from the LIMIT RCT, in relation to detecting hypothesised effects of interest (intervention and pre-pregnancy maternal BMI) as well as effects known to be spurious, and known to be present.

Methods: DNAm data, from 645 cord blood samples analysed using Illumina 450K BeadChip arrays, were normalised using three different methods (with probe filtering undertaken pre- or post-normalisation). Batch effects were handled with a supervised algorithm, an unsupervised algorithm, or adjustment in the analysis model. Analysis was undertaken with and without adjustment for estimated cell type proportions. The effects estimated included intervention and BMI (effects of interest in the original study), infant sex and randomly assigned groups. Data processing and analysis methods were compared in relation to number and identity of differentially methylated probes, rankings of probes by p value and log-fold-change, and distributions of p values and log-fold-change estimates.

Results: There were differences corresponding to each of the processing and analysis choices. Importantly, some combinations of data processing choices resulted in a substantial number of spurious ‘significant’ findings. We recommend greater emphasis on replication and greater use of sensitivity analyses.

Clinical Trials Registration: ACTRN12607000161426

Word Count: 6298

Acknowledgements

We would like to acknowledge the participants in the LIMIT Randomised Controlled Trial, whose samples and data are used in this study, as well as the LIMIT research team who collected data and samples for the study.

We would also like to acknowledge Shobana Navaneethabalakrishnan, the Academic Editor at PeerJ and two anonymous reviewers, whose helpful suggestions have substantially improved the manuscript.

Introduction and Background

With the advent of high-throughput assays, epigenome-wide DNA methylation studies have become more popular, and researchers are now investigating the effects on DNA methylation (DNAm) of a wide range of environmental exposures and physiological conditions, with particular interest in the contribution of epigenetic mechanisms such as DNAm to the early life origins of health and disease. The ability to perform EWAS is particularly useful in relation to conditions where associated differences in DNAm are likely to be fairly modest (Marabita et al., 2013). However, DNAm data – as with high-dimensional 'omics' data generally – requires substantial pre-processing prior to analysis, including probe and sample filtering, normalisation to remove variation due to technological factors, and correction for other factors which may confound effects of interest, such as batch effects or differences in cell type proportions between samples. Numerous methods exist to perform these processing steps, and many articles have been published which provide useful guidance for the use of analysis pipelines(Marabita et al., 2013; Yousefi et al., 2013; Lehne et al., 2015; Morris & Beck, 2015; Maksimovic, Phipson & Oshlack, 2017), or comparing some alternatives for individual steps in the overall processing pipeline, including probe filtering (Heiss & Just, 2019), normalisation (Wang et al., 2012, 2015; Fortin et al., 2014; Wu et al., 2014; Hicks & Irizarry, 2015), or correction for / avoidance of batch effects. These have led to some general conclusions regarding the need to account for batch effects, the importance of correcting for estimated cell type proportion, and perhaps the greater suitability of within-array normalisation methods compared to between-array methods when global methlation differences are not expected (Maksimovic, Phipson & Oshlack, 2017), but there is no clear overall consensus on the best processing or analysis approach (Price & Robinson, 2018; Zindler et al., 2020), or of the overall advantages and disadvantages of different combinations of processing choices.

We recently investigated the effect of an antenatal diet and lifestyle intervention, and of maternal early pregnancy BMI, on neonatal cord blood DNA methylation in infants of mothers who were overweight or obese in early pregnancy(Louise et al., 2022). We were unable to replicate findings from previous studies which reported a range of loci to be significantly differentially methylated in relation to maternal BMI or diet and lifestyle in pregnancy (Gemma et al., 2009; Sharp et al., 2015, 2017; Thakali et al., 2017; Hjort et al., 2018)and indeed did not find any significant differences In methylation corresponding to either BMI or intervention effects. We were aware of literature suggesting that use of supervised batch-correction algorithms may produce spurious findings (Nygaard, Rødland & Hovig, 2016a; Price & Robinson, 2018; Zindler et al., 2020), and that the number of statistically significant findings may differ according to normalisation method (Wu et al., 2014), adjustment for estimated cell type proportion (Sharp et al., 2017) or stringency of type I error control (Wu et al., 2014), which led to the hypothesis that the discrepancy in findings may be due in part to differences in data processing and analysis methods. Following the common practice in clinical trials of conducting sensitivity analyses to assess robustness of results to various assumptions and decisions (Thabane et al., 2013), we performed several re-analyses with different normalisation methods, methods for batch effect handling, type I error control, and presence vs absence of adjustment for estimated cell type proportions. This confirmed that our findings indeed differed under different data-processing and analysis choices.

While previous studies comparing different methods have also produced different findings, these have tended to consider only one element of the processing and analysis pipeline (such as normalisation, or batch correction) in isolation. In addition, they have tended to concentrate on the tendency for some methods to produce results which are likely to be spurious (false positives), while often being unable to definitively confirm this due to the lack of known truth regarding the presence and magnitude of differential methylation effects.

We therefore set out to investigate the impact of different data-processing choices in a more systematic way, looking at the effect of combinations of data processing choices on findings specifically regarding statistically significant differentially methylated probes (DMPs), and of the behaviour of these combinations in cases where effects are known to be either absent or present, as well as their behaviour in relation to our effects of interest (maternal BMI and lifestyle intervention). We were able to create a scenario in which effects were known to be absent by randomly assigning samples to groupings. We could not similarly ensure a scenario where effects were known to be present (as the truth regarding the existence, location and magnitude of any effects in our samples is not known); however we investigated effects of infant sex as a rough proxy, since infant sex is known to have substantial effects on DNAm which can be detected by the 450K array (Yousefi et al., 2015).

Data and Methods

The LIMIT Randomised Controlled Trial

The LIMIT study was a randomised, controlled trial of an antenatal diet and lifestyle intervention for women with early pregnancy BMI ≥25.0 kg/m^2. The study, and its primary and main secondary outcomes, have been extensively reported elsewhere (Dodd et al., 2014b). Inclusion criteria were early pregnancy BMI ≥25.0 kg/m^2 and pregnancy between 10^{+0} and 20^{+0} weeks' gestation, with exclusion criteria of multiple gestation or previously existing diabetes. The study randomised 2212 women in total to one of two groups: a comprehensive diet and lifestyle intervention (Lifestyle Advice; n=1108) or antenatal care delivered according to local guidelines (Standard Care; n=1104) which did not include information on diet or physical activity. The study was reviewed by the ethics committee of each participating institution including the Women's and Children's Health Network Human Research Ethics Committee (1839 & 2051); the Central and Northern Adelaide Health Network Human Research Ethics Committee (2008033) and the Southern Adelaide Local Health Network Human Research Ethics Committee (128/08). Informed written consent was obtained for all participants to participate in the LIMIT study, and additional written consent was obtained to collect samples of umbilical cord blood at delivery for the purposes of gene expression research related to weight and to the diet and lifestyle intervention.

The primary outcome of the LIMIT study was birth of a Large for Gestational Age (LGA) infant. There were no significant differences observed between the groups in relation to this outcome; however, a significantly lower incidence of high birthweight (>4kg) was observed in the Lifestyle Advice group, with a Relative Risk of 0.82 (95% CI: 0.68, 0.99, p=0.04). Additionally, measures of diet quality and physical activity were improved in women in the Lifestyle Advice group compared to those in the Standard Care group (Dodd et al., 2014a).

As previously outlined in the companion paper (Louise et al., 2022), Cord Blood DNA for a range of secondary studies was collected at the time of birth from consenting participants, and was frozen as whole blood preserved with EDTA. Funding was available to perform DNA methylation analysis for a total of 649 samples, which were randomly selected from the total number of available samples, balanced between the Lifestyle Advice and Standard Care groups (Supplementary Table S1). After DNA extraction, genome-wide DNA methylation was performed using the Illumina Infinium HumanMethylation 450K Bead-Chip array. Results were supplied as raw probe intensities (IDAT files).

For the additional analyses investigating known spurious effects, artificial (fake) groups were created by assigning samples based on random draws from binomial distributions. The first grouping ('Tortoiseshell' vs'Tabby') was generated using a binomial distribution with 50% probability of assignment to each group. The second grouping ('Long'- vs 'Short-Haired') was created to mimic stratified randomisation as well as unequal proportions in each group: within each level of the first fake group, samples were assigned to Long-Haired with 40% probability and Short-Haired with 60% probability.

All data processing and analyses were undertaken using R version 4.0 (R Core Team, 2018).

Probe and Sample Filtering

The *minfi* package (Aryee et al., 2014) was used to read in the raw *idats* (without normalisation), and to calculate both probe-wise and sample-wise 'detection p values'. Samples were identified as 'faulty' if they had a detection p-value ≥0.05. 13 such samples were excluded; however these were due to a known chip failure, and had subsequently been rerun. A further four samples were excluded because the correct corresponding study identifier could not be ascertained, leaving 645 samples for analysis.

Probes were filtered using multiple criteria. Firstly, probes were excluded if they had a detection p-value ≥0.001 in more than 25% of the 645 samples, indicating that their signal could not be accurately detected for a large proportion of samples (Dedeurwaerder et al., 2014; Maksimovic, Phipson & Oshlack, 2017). Secondly, probes were excluded if they were on a list of those previously identified as cross-reactive (Chen et al., 2013); i.e. there was a high probability they may hybridize to locations on the genome different to those for which the probe was designed (Dedeurwaerder et al., 2014; Naeem et al., 2014). Thirdly, probes with an identified SNP within 3 nucleotides of the CpG site and minor allele frequency >1%, and probes on the X and Y chromosomes were excluded. This was done in order to avoid spurious methylation 'differences' due either to SNPs within the CpG targets, or due to X and Y chromosomes (Dedeurwaerder et al., 2014; Naeem et al., 2014). Filtering of cross-reactive probes, probes with a nearby SNP and probes on the X and Y chromosomes was performed using the *DMRCate* package (Peters et al., 2015). This left 426,572 probes available for analysis.

Probe filtering was performed either after normalisation (post-filtered) or prior to normalisation (pre-filtered). The one exception for pre-filtering was when normalising using the BMIQ method, where probes on the X and Y chromosome were retained as this was required in order for the function to run.

Normalisation

Normalisation involves making changes to the raw data in order to remove artifactual variation. In the case of Illumina 450K BeadChip arrays, this requires correcting for the presence of two different probe types. Infinium I probes use the same colour signal for methylated and unmethylated CpG and are often used for regions of high CpG density, while Infinium II probes use different colours to differentiate between methylated and unmethylated states (Pidsley et al., 2013; Wang et al., 2015). Normalisation is performed on β values, or the ratio of methylated to total intensity, defined as $\frac{M}{M+U+offset}$. Here, M is the methylated intensity and U is the unmethylated intensity; the offset is a constant added to regularize the βvalue where both methylated and unmethylated intensities are low. The distribution of β values is bimodal, with peaks corresponding to methylated and unmethylated states, but the distribution of Infinium II probes differs from that of Infinium I, being more compressed towards 0.5 (Pidsley et al., 2013) and hence having a smaller 'dynamic range' (Teschendorff et al., 2013; Dedeurwaerder et al., 2014).

Numerous methods exist for normalising Illumina BeadChip array data, but there is little consensus or guidance on which should be employed in a given context. The main advice is that 'between-array' methods, which normalise across samples, are preferable when global differences between samples are expected, while 'within-array' methods, which normalise probes within each sample, are better suited to effects in which the majority of genes will not be differentially expressed. (Maksimovic, Phipson & Oshlack, 2017) The latter is the context in which many EWAS studies, including the present one, are conducted; as noted above, only modest differences in a small proportion of genes are expected for most early-life exposures. The methods chosen for the present investigation have all been used in the context of studies such as this: Categorical-Subset Quantile Normalisation (SQN) (Wu et al., 2014; Wang et al., 2015), Beta-Mixture Quantile Normalisation (BMIQ) (Teschendorff et al., 2013), and Subset-Quantile Within-Array Normalisation (SWAN) (Maksimovic, Gordon & Oshlack, 2012). While numerous other methods exist, a comparison of all available normalisation methods was beyond the scope of this paper. Further details on the methods are given in the Supplementary Information.

Both Subset Quantile Normalisation and Subset-Within-Array-Normalisation were performed using functions in the *minfi* package (*preprocessQuantile* and *preprocessSWAN* respectively), on raw intensity data. Beta-Mixture Quantile normalisation was performed using the *champ.norm* function in the *ChAMP* package after converting intensities to β values.

Batch Effects

Batch effects arise when samples are processed in separate groups, creating unwanted variation due, for example, to different reagents, different plates or different scanner settings. (Morris & Beck, 2015; Nygaard, Rødland & Hovig, 2016a; Price & Robinson, 2018)
There are 12 Illumina 450K arrays (samples) per chip (this is reduced to 8 arrays per chip for the more recent 850K array); thus most studies involving large numbers of samples must be run on multiple chips. This introduces extra variability to the data, and may also confound the actual effects of interest, if samples from different groups are not evenly distributed between the batches. These effects must be accounted for in order to obtain valid estimates of the effects of interest.

Unlike probe filtering and normalisation, batch effects can be handled at the analysis stage, by adjusting for batch in the analysis model. However, it is also common to address batch effects at the data-processing stage, using a batch-correction algorithm, with the resulting data considered to be free of batch effects (Nygaard, Rødland & Hovig, 2016a). The ComBat algorithm has been widely used and considered the most effective method (Zindler et al., 2020) of removing batch effects in DNAm data; it has been incorporated into various analysis pipelines. Until recently, ComBat could be implemented only as a supervised method, in which the biological factors of interest had to be specified along with the batch variable (Price & Robinson, 2018) (Fortin, Triche & Hansen, 2016); it can now also be implemented as an unsupervised method, in which only the batch variable is specified.

For each of the normalised datasets (i.e. SQN, BMIQ and SWAN normalised datasets, each with probes filtered either before or after normalisation), we handled batch effects in three ways: firstly, by adjusting for batch in the analysis model (BatchAdjust); secondly, implementing the supervised ComBat algorithm (sCB); and thirdly, implementing the unsupervised ComBat algorithm (uCB). For the supervised ComBat algorithm, it was necessary to run the process twice: once with the effects of interest specified as maternal early pregnancy BMI, antenatal intervention group, and their interaction; and again with the effects of interest specified as Fake Group 1, Fake Group 2, their interaction, and infant sex.

Cell Type Proportions

Cord blood, like whole blood, contains a mixture of different cell types, which have different DNA methylation profiles.(Jaffe & Irizarry, 2014; Teschendorff & Zheng, 2017) If samples differ in the proportion of these different cell types, this may confound effects of interest, either hiding true differences in DNAm, or giving rise to spurious differences. Most studies of the effect of BMI, lifestyle interventions, or similar factors on cord blood DNA methylation have not attempted (or have not documented an attempt) to correct for potential differences in cell type composition, perhaps because reference profiles for cord blood were not available until more recently (Bakulski et al., 2016), and the mix of cell types and DNAm profiles may differ in cord blood compared to whole blood, making it inappropriate to apply reference profiles from whole blood to cord blood data.(Cardenas et al., 2016)

We estimated the proportion of B cells, CD4+T, CD8+T, granulocytes, monocytes, natural killer, and nucleated RBCs in the raw data using the *estimateCellCounts()* function in the *minfi* package, with the Cord Blood reference panel. The estimated proportions were then added to the metadata for use as adjustment variables in the analyses. We then undertook analyses either adjusted or not adjusted for estimated cell type proportion.

Figure 1 depicts the combinations of data-processing and analysis choices that were undertaken. In brief, there were six normalised datasets (three different normalisation methods, with probe filtering performed before normalisation or after normalisation). These datasets were either used immediately for analysis, or processed using the ComBat algorithm (in both supervised and unsupervised form) prior to analysis. Non-ComBat-processed data were analysed with three different models: an unadjusted model (containing only the effects of interest), a model adjusted for batch, and a model adjusted for batch and estimated cell type proportions. ComBat-processed data were analysed with two different models: one containing no other adjustment variables (but assumed to be ‘pre-adjusted’ for batch), and one adjusted for cell type proportion.

Statistical Analysis

Differential methylation was investigated probe-wise using linear models with empirical Bayes variance correction as implemented in the *limma* package (Ritchie et al., 2015; Smyth). For effects of BMI and intervention, models specified BMI (as continuous and mean-centred), intervention (Lifestyle Advice vs Standard Care), and their interaction. Contrasts were specified to estimate the effect of the intervention and of maternal BMI. Because of the presence of the intervention-by-BMI interaction term, this required specification of the BMI values at which the intervention effect was to be estimated, and the intervention groups in which the effect of BMI was to be estimated. For estimating intervention effects, we chose the mean BMI of the cohort (i.e. value of 0 for the mean-centred variable, corresponding to an actual BMI of approximately 33 kg/m^2), and at 5 kg/m^2 above the mean (a value of 5, corresponding to an actual BMI of approximately 38 kg/m^2. For the effect of BMI, we estimated the effect of an increase of 5 kg/m^2 in BMI in each of the intervention groups (Standard Care, Lifestyle Advice) respectively. For effects of fake groups and infant sex, the models specified sex (Female vs Male), Fake Group 1 (Tortoiseshell vs Tabby), Fake Group 2 (Long-Haired vs Short-Haired), and their interaction. Contrasts were specified for infant sex, and for the effect of each fake group separately within levels of the other fake group (i.e. effect of Tortoiseshell in Long-Haired and in Short-Haired; and effect of Short-Haired in Tortoiseshell and Tabby). The model matrix and contrast matrices are shown in the Supplementary Table S4.

For each contrast in each model, the number and identity (where applicable) of any differentially methylated probes (DMPs) were obtained. For detection of DMPs, *limma*'s default method of multiple-comparisons correction (Benjamini-Hochberg) and default alpha of 0.05 was used; this method controls the False Discovery Rate, or the proportion of statistically significant results not corresponding to true effects. Where DMPs were obtained, a comparison was made using the Holm method (retaining alpha of 0.05), which controls the Family-Wise-Error Rate (the probability that at least one statistically significant result does not correspond to a true effect). The Holm method can be considered more stringent than Benjamini-Hochberg, but is less stringent than Bonferroni correction, which is known to be too conservative even outside the context of high-dimensional data and is therefore not generally appropriate for EWAS studies (and has not been used in other studies investigating cord blood DNAm in relation to maternal BMI or diet and lifestyle). The full set of p values and estimated log-fold-changes (for all 426572 probes) corresponding to each contrast were also obtained, in order to compare probe rankings and overall distributions. To make the comparison more tractable, probe rankings were investigated using only those probes ranked in the top 10 (i.e. the probes with the smallest p value, or largest estimated logFC, in a given model).

The findings from different data-processing choices were then compared along five dimensions:

1. Number and identity of differentially methylated probes (DMPs); for infant sex, the direction of differential methylation ('down', corresponding to negative t-statistics or lower methylation in females, versus 'up', corresponding to higher methylation in females) was also examined. For BMI and intervention effects, the number and identity of statistically significant DMPs allows us to see differences in detection of effects, and whether the different analysis pipelines produce consistent results regarding the identity of any DMPs, though the truth is not known. For the fake groupings, the number of statistically significant DMPs is an indication of the tendency to produce spurious findings. For infant sex, while we do not know the actual number and identity of truly differentially methylated sites, differences in the number and identity of DMPs demonstrate that there must be either false positives or false negatives.
2. Consistency of rankings by p value for 'top 10' probes. This gives an indication of whether different methods will give the same results for the probes with the largest differences.
3. Consistency of rankings by logFC for 'top 10' probes, as well as the consistency of the logFC estimates. This gives an indication of whether the estimates of effect are similar between methods.
4. Overall distribution of p values. Under the null hypothesis of no effect, p values should have a uniform distribution between 0 and 1; if effects are present, there will be more p values at the lower end of the distribution, the extent of which will depend on how many DMPs there are and the strength of the effects.
5. Overall distribution of logFC estimates. Under the null hypothesis of no effect, logFC estimates should be roughly normally distributed around 0. If effects are present, there will be more estimates far away from 0 (with the direction depending on whether the effect is one of

hypomethylation or hypermethylation, and the distance depending on the strength of the effect).

Results

All dimensions of data processing choices had some impact on downstream analysis results, in terms of the number (and identity) of differentially methylated probes, rankings of probes (by p value and logFC), estimates of logFC, and overall distribution of p values and logFC, corresponding to both real and spurious effects of interest. In some cases a consistent impact of a particular choice was observed, while in others there was no consistent pattern, or this pattern varied according to the other choices with which it was combined.

Tables 1-3 give information about the number of significantly differentially methylated probes in each of the models fitted for the combinations of filtering, normalisation, batch correction and cell adjustment approaches, for infant sex, maternal BMI (in the Standard Care group) and 'Tortoiseshell' (in the 'Tabby' group) respectively. Supplementary Figures S1-S4 show the degree of overlap in the actual probes found to be significantly differently methylated between models for infant sex, intervention (at the mean BMI of the cohort), BMI (in the Standard Care group), and the effect of 'Short-Haired' in the 'Tabby' group. Figures 2 and Figures 3-5 show the differences in ranking of probes (those which were in the top 10 in any model) by p value and log-Fold-Change for the same set of effects, and Supplementary Tables S4-S6 gives Spearman Rank Correlation matrices for these rankings. The overall distribution of p values, and of log-Fold-Change estimates, for the same set of effects is shown in Figures 6 and 7.

Below we discuss the effect of each dimension (probe filtering, normalisation, batch effects, cell type correction) on results.

Effect of probe filtering pre-normalisation vs post-normalisation

Filtering probes prior to normalisation, compared to filtering after normalisation, led to modest differences in number of DMPs, rankings of probes by logFC and p value, and overall distributions of p values and logFC estimates. Filtering pre-normalisation produced different numbers of DMPs for infant sex, but the nature of the effect differed by normalisation method: in SWAN data there was a consistent pattern of fewer significant probes both negative and positive, while in BMIQ data there were fewer negative but more positive probes, and in SQN data there were more negative but fewer positive probes. In relation to effects of BMI, intervention, and fake groups, differences were harder to discern due to the lack of any DMPs for many models; however, when DMPs were present for an effect, there was a tendency for there to be a greater number of them in the pre-filtered data.

Probe rankings, by logFC and p value, tended to be relatively consistent between pre-filtered and post-filtered data, with some cases of larger discrepancies in rankings for individual probes. The discrepancies were more common, and larger, for fake group, BMI and intervention effects than for infant sex. Similarly, there were no dramatic differences in distributions of p values or logFC estimates for infant sex; there were differences in distribution between pre- and post-filtered data for fake group, intervention and BMI effects, but there was no consistent pattern to these differences.

The question of whether probe filtering should be carried out before or after normalisation is one which has received surprisingly little attention in the literature, but our results suggest that it can make a difference to findings in some contexts. In particular there may be a higher risk of spurious findings in pre-filtered data, but there may also be a risk of failing to detect true differences – either any differences, or specifically hypomethylated or hypermethylated loci, depending upon the normalisation method employed.

Effect of Normalisation Method

Normalisation method had a substantial influence on number and identity of DMPs, rankings of probes and p values, and distributions of p values and logFC estimates. For infant sex, SQN data consistently had the highest number of significant negative probes and the lowest number of significant positive probes, while SWAN data always had the lowest number of significant positive probes. For BMI and intervention effects, only BMIQ data produced DMPs where no ComBat processing was used; in data processed using supervised ComBat, all three normalisation methods resulted in some DMPs, but the number and identity of these probes differed. In fake group data, SQN data produced a large number of significant probes in non-ComBat-processed and supervised-ComBat data, while BMIQ and SWAN data produced a small number of probes in supervised-ComBat data only; again, the number of significant probes differed between the normalisation methods (see Tables 1-3 and Supplementary Figures S1-S4).

There was a fair degree of consistency in rankings of probes by p value for infant sex, but some large discrepancies in rankings for BMI, intervention, and fake group effects. The rankings were less consistent for highest-ranked probes by logFC, with some quite large differences in both rankings and effect estimates (including different directions of effect) for infant sex, BMI, intervention and fake groups. BMIQ estimates tended to be more extreme (further from 0) than the other two methods.

Distributions of p values and logFC estimates also differed between normalisation methods. For p values the differences were not consistent across models and effects, but for logFC there was a clear difference between BMIQ and the other two methods, with the range of estimates in BMIQ data being much more widely dispersed; SQN and SWAN data had more similar distributions, but SQN was moderately narrower than SWAN across all effects and models.

Overall, there was little difference between SQN and SWAN methods when adjusting for batch in the model. There is some evidence that SQN would result in fewer significant DMPs than SWAN for known effects (particularly when using supervised ComBat), but (many) more spuriously significant DMPs than either SWAN or BMIQ where effects are absent. The behaviour of BMIQ was more variable depending on other dimensions of the pipeline, but had a wider dispersion of logFC estimates than the other methods, particularly when adjusting for batch in the model. This tended to result in more DMPs in some scenarios, but in general will lead effect estimates derived from BMIQ data to be more extreme (and probably overestimates of the true effect).

Effect of Batch Correction Method

There were clear differences in all dimensions between batch correction methods. For all effects (infant sex, BMI, intervention and fake groups), supervised ComBat processing produced a larger number of DMPs compared to either unsupervised ComBat processing or adjustment for batch in the analysis model. The difference between unsupervised ComBat and batch-adjustment was less consistent for infant sex effects, but for BMI, intervention and fake group effects, there were no DMPs in unsupervised ComBat models, whereas there were a few for batch-adjusted models.

Rankings of top probes by p value were relatively consistent between batch-adjustment methods for infant sex, but there were some large discrepancies particularly for BMI and intervention effects, and especially in BMIQ data. The same phenomenon was observed for logFC rankings, which also showed a tendency for logFC estimates in unsupervised-ComBat data to be smaller in absolute magnitude (closer to 0).

The distribution of p values showed clear and consistent differences between batch-correction methods, with the distribution in supervised ComBat data shifted substantially towards 0 relative to both unsupervised ComBat and batch-adjusted models, for all effects. For logFC estimates, supervised ComBat and batch-adjusted data were generally fairly similar, but unsupervised ComBat data generally resulted in a narrower range. This means that supervised ComBat will tend to produce more statistically significant probes, regardless of the presence or absence of an effect. Conversely, effect estimates from unsupervised ComBat may be underestimated; at least, they will tend to be smaller in magnitude than those derived from data where batch is handled differently.

Of particular note is the combination of SQN normalisation and either adjustment for batch in the model, or use of supervised ComBat. These combinations produced an extremely large number of significant DMPs for fake group effects; this was more extreme in the case of supervised ComBat (producing over 6000 DMPs) than when adjusting for batch (somewhat over 2000 DMPs). Additional adjustment for cell type proportion ameliorated this effect, as discussed below, but in the case of supervised ComBat data, did not eliminate spurious findings. This suggests that batch adjustment may be particularly ill-advised in the context of SQN normalisation; since SQN involves between-array as well as within-array normalisation, additional adjustment for batch may be over-correcting.

Effect of Adjustment for Estimated Cell Type Proportion

Adjustment for cell type proportion affected results, but the impact was not consistent across the different types of effects studied. Adjustment for batch resulted in a substantially larger number of DMPs (both negative and positive) for infant sex, but reduced the number of DMPs for fake groups (for models where there were DMPs for fake groups effects). For BMI and intervention, the effect of cell type adjustment was mostly but not entirely to produce more DMPs.

The effect of cell type adjustment on top probe rankings was fairly modest, although some quite large discrepancies were observed for p value rankings, logFC rankings, and logFC estimates. The effect on distribution of p values depended on the effect: for infant sex, adjustment for cell type proportion consistently (for all normalisation and batch-correction methods) shifted the distribution downwards towards 0 (i.e. more statistically significant probes), whereas the differences were less consistent and smaller in BMI, intervention and fake group effects. There were no large or consistent differences in distribution of logFC estimates between cell-type-adjusted and non-adjusted models.

Overall, adjustment for cell type proportion tended to improve model behaviour regarding spurious results: the number of significantly differentially methylated probes decreased with adjustment for cell type proportion, though they were not always eliminated. The number of differentially methylated probes for infant sex was increased, which may reflect either improvement (greater ability to detect true effects due to removal of noise due to cell type differences) or harm (greater number of spurious effects) depending on whether the extra probes are in fact differentially methylated between males and females; without knowing the true number and identity of DMPs, we cannot be certain. Similarly, adjustment for cell type proportion increased the number of DMPs for BMI effects in BMIQ and SWAN data, in one scenario (BMIQ with Supervised ComBat) by a substantial amount (from 99 to 2017 DMPs) and these are most likely to be false positives.

Discussion

Different choices in probe filtering, normalisation, batch handling, and adjustment for cell types resulted in different findings regarding the presence and identity of differentially methylated probes, rankings of probes by p value and log-fold-change, and different overall distributions of p values and log-fold-change estimates. While some of these differences were relatively modest, our results nevertheless show that particular combinations of data processing and analysis choices may result in spurious false positive findings, and/or potentially the failure to detect true effects. Additionally, while the magnitude of effect estimates is often not considered in differential methylation studies, some pipelines may result in an overestimate or underestimate of the true effect. Importantly, the results tended to depend on *combinations* of choices rather than individual elements of the analysis pipeline.

The results of our analyses are consistent with other investigations which have been undertaken into different data-processing and analysis choices. As noted above, the potential for 'false positives' to result from supervised batch-correction methods specifying effects of interest has been previously identified by a number of authors.(Nygaard, Rødland & Hovig, 2016a; Price & Robinson, 2018; Zindler et al., 2020). Our finding that the distribution of p values in the supervised ComBat algorithm tends to shift the p value distribution downward is consistent with the finding of Nygaard et al that, in contexts where the effects of interest are not evenly spread between batches, the distribution of F-statistics will be biased upwards (Nygaard, Rødland & Hovig, 2016a). While implementation as an unsupervised method may be preferable, our findings suggest that this may create a different problem, with the estimates of log-fold-change corresponding to effects of interest biased towards zero.

Wu et al's study (Wu et al., 2014) comparing a variety of normalisation approaches noted a tendency for more statistically significant differences to arise in SQN data, which they hypothesise may be due to reduced overall variance. In our investigation, the main context in which SQN data produced a large number of spurious differentially methylated probes was when supervised ComBat, or adjustment for batch in the model, was used; we additionally found that adjustment for cell type proportion reduced the number of spurious findings (while not necessarily eliminating them). Thus, SQN is not universally more prone to producing spurious findings than other normalisation methods.

Our findings do not suggest that there is one particular combination of methods which can be guaranteed to 'work' in all contexts, and some of the recommendations which have been made by others may need to be modified somewhat. For example, Nygaard et al conclude that adjustment for batch in the model is preferable to the use of batch-correction algorithms,(Nygaard, Rødland & Hovig, 2016b), but our results suggest that this is inadvisable for data that have been normalised using SQN; in our data, this combination resulted in a large number of spurious findings. In general, while our results support others' findings that supervised bach-correction algorithms should not be used, there does not appear to be much difference between unsupervised batch-correction and adjustment for batch in the model. The only caveat here is that some of our results (particularly regarding effects of infant sex) suggest that unsupervised ComBat may underestimate the magnitude of effects, as the distribution of logFC estimates was substantially narrower than other methods. The use of a more stringent method of Type I error control may also help to reduce the number of spurious findings: the use of FDR correction methods such as Benjamini-Hochberg, while very common (Maksimovic, Phipson & Oshlack, 2017), may not be sufficient to deal with higher rates of spurious results (Nygaard, Rødland & Hovig, 2016a). In our data, the use of the Holm method (which controls the Family-Wise Error Rate) reduced, but did not eliminate, spurious findings associated with fake group effects. Investigation of DNA regions, rather than probe-wise analysis, may also help to differentiate true methylation differences from spurious ones (Wang et al., 2015): the statistically significant DMPs for fake group effects (as well as for BMI and intervention effects) tended to be isolated rather than being grouped in the same region, and in our companion paper, we found no significant differences in methylation for groups of probes on candidate genes (Louise et al., 2022).

One limitation of our study is our inability to compare model behaviour in relation to known effects. It was relatively simple to create fake groups to study behaviour of models for effects which were known *not* to exist, but as we do not know the truth about which effects actually exist in our data, we could not compare behaviour of models in their ability to detect these known effects. Simulated data could potentially be used for this purpose; however, the effects in the simulated data would have to be biologically plausible. This was beyond the scope of our study; however, it is a good subject for future research. We used infant sex as the nearest proxy to a known effect, as we knew at least that some effects existed. However, we cannot say whether, and to what extent, the differences observed in relation to infant sex reflect spurious findings versus the failure to detect true effects.

Overall, as many other authors have noted, researchers working with DNAm data should better understand the methods built into standard pipelines (Price & Robinson, 2018; Zindler et al., 2020), and should better document the specific data-processing methods used (Nygaard, Rødland & Hovig, 2016a; Zindler et al., 2020). It is also important, in our view, to pay more attention to the context in which a particular epigenome-wide analysis is performed. For example, a less stringent method of Type I error control may often be chosen because the study is exploratory (hypothesis-generating) rather than confirmatory, and it is considered more important not to miss potential findings than to rule out spurious ones. In this case, the results from such studies should be interpreted accordingly: as suggestive findings which cannot be confidently accepted until they are validated in new data. The validation of existing findings should be treated as a high priority in epigenetics research (Price & Robinson, 2018).

Additionally, the degree of confidence that can be placed in any new discoveries could be enhanced by performing sensitivity analyses – re-performing analyses using different normalisation methods, batch correction methods, or models - which we believe should become standard in this area.

References

Aryee MJ, Jaffe AE, Corrada-Bravo H, Ladd-Acosta C, Feinberg AP, Hansen KD, Irizarry RA. 2014. Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics* 30:1363–1369. DOI: 10.1093/bioinformatics/btu049.

Bakulski KM, Feinberg JI, Andrews SV, Yang J, Brown S, L McKenney S, Witter F, Walston J, Feinberg AP, Fallin MD. 2016. DNA methylation of cord blood cell types: Applications for mixed cell birth studies. *Epigenetics* 11:354–362. DOI: 10.1080/15592294.2016.1161875.

Cardenas A, Allard C, Doyon M, Houseman EA, Bakulski KM, Perron P, Bouchard L, Hivert M-F. 2016. Validation of a DNA methylation reference panel for the estimation of nucleated cells types in cord blood. *Epigenetics* 11:773–779. DOI: 10.1080/15592294.2016.1233091.

Chen Y, Lemire M, Choufani S, Butcher DT, Grafodatskaya D, Zanke BW, Gallinger S, Hudson TJ, Weksberg R. 2013. Discovery of cross-reactive probes and polymorphic CpGs in the Illumina Infinium HumanMethylation450 microarray. *Epigenetics* 8:203–209. DOI: 10.4161/epi.23470.

Dedeurwaerder S, Defrance M, Bizet M, Calonne E, Bontempi G, Fuks F. 2014. A comprehensive overview of Infinium HumanMethylation450 data processing. *Briefings in bioinformatics* 15:929–941. DOI: 10.1093/bib/bbt054.

Dodd JM, Cramp C, Sui Z, Yelland LN, Deussen AR, Grivell RM, Moran LJ, Crowther CA, Turnbull D, McPhee AJ, Wittert G, Owens JA, Robinson JS, For the LIMIT Randomised Trial Group. 2014a. The effects of antenatal dietary and lifestyle advice for women who are overweight or obese on maternal diet and physical activity: the LIMIT randomised trial. *BMC Medicine* 12:161. DOI: 10.1186/s12916-014-0161-y.

Dodd JM, Turnbull D, McPhee AJ, Deussen AR, Grivell RM, Yelland LN, Crowther CA, Wittert G, Owens JA, Robinson JS, for the LIMIT Randomised Trial Group. 2014b. Antenatal lifestyle advice for women who are overweight or obese: LIMIT randomised trial. *BMJ* 348:g1285–g1285. DOI: 10.1136/bmj.g1285.

Fortin J-P, Labbe A, Lemire M, Zanke BW, Hudson TJ, Fertig EJ, Greenwood CM, Hansen KD. 2014. Functional normalization of 450k methylation array data improves replication in large cancer studies. *Genome Biology* 15. DOI: 10.1186/s13059-014-0503-2.

Fortin J-P, Triche T, Hansen K. 2016. Preprocessing, normalization and integration of the Illumina HumanMethylationEPIC array. DOI: 10.1101/065490.

Gemma C, Sookoian S, Alvariñas J, García SI, Quintana L, Kanevsky D, González CD, Pirola CJ. 2009. Maternal Pregestational BMI Is Associated With Methylation of the *PPARGC1A* Promoter in Newborns. *Obesity* 17:1032–1039. DOI: 10.1038/oby.2008.605.

Heiss JA, Just AC. 2019. Improved filtering of DNA methylation microarray data by detection p values and its impact on downstream analyses. *Clinical Epigenetics* 11:15. DOI: 10.1186/s13148-019-0615-3.

Hicks SC, Irizarry RA. 2015. quantro: a data-driven approach to guide the choice of an appropriate normalization method. *Genome Biology* 16:117. DOI: 10.1186/s13059-015-0679-0.

Hjort L, Martino D, Grunnet LG, Naeem H, Maksimovic J, Olsson AH, Zhang C, Ling C, Olsen SF, Saffery R, Vaag AA. 2018. Gestational diabetes and maternal obesity are associated with epigenome-wide methylation changes in children. *JCI Insight* 3:e122572. DOI: 10.1172/jci.insight.122572.

Jaffe AE, Irizarry RA. 2014. Accounting for cellular heterogeneity is critical in epigenome-wide association studies. *Genome biology* 15:R31–R31. DOI: 10.1186/gb-2014-15-2-r31.

Lehne B, Drong AW, Loh M, Zhang W, Scott WR, Tan S-T, Afzal U, Scott J, Jarvelin M-R, Elliott P, McCarthy MI, Kooner JS, Chambers JC. 2015. A coherent approach for analysis of the Illumina HumanMethylation450 BeadChip improves data quality and performance in epigenome-wide association studies. *Genome biology* 16:37–37. DOI: 10.1186/s13059-015-0600-x.

Louise J, Deussen AR, Koletzko B, Owens J, Saffery R, Dodd JM. 2022. Effect of an antenatal diet and lifestyle intervention and maternal BMI on cord blood DNA methylation in infants of overweight and obese women: The LIMIT Randomised Controlled Trial. *PLOS ONE* 17:e0269723. DOI: 10.1371/journal.pone.0269723.

Maksimovic J, Gordon L, Oshlack A. 2012. SWAN: Subset-quantile Within Array Normalization for Illumina Infinium HumanMethylation450 BeadChips. *Genome Biology* 13:R44. DOI: 10.1186/gb-2012-13-6-r44.

Maksimovic J, Phipson B, Oshlack A. 2017. A cross-package Bioconductor workflow for analysing methylation array data. *F1000Research* 5:1281. DOI: 10.12688/f1000research.8839.3.

Marabita F, Almgren M, Lindholm ME, Ruhrmann S, Fagerström-Billai F, Jagodic M, Sundberg CJ, Ekström TJ, Teschendorff AE, Tegnér J, Gomez-Cabrero D. 2013. An evaluation of analysis pipelines for DNA methylation profiling using the Illumina HumanMethylation450 BeadChip platform. *Epigenetics* 8:333–346. DOI: 10.4161/epi.24008.

Morris TJ, Beck S. 2015. Analysis pipelines and packages for Infinium HumanMethylation450 BeadChip (450k) data. *Methods* 72:3–8. DOI: 10.1016/j.ymeth.2014.08.011.

Naeem H, Wong NC, Chatterton Z, Hong MKH, Pedersen JS, Corcoran NM, Hovens CM, Macintyre G. 2014. Reducing the risk of false discovery enabling identification of biologically significant genome-wide methylation status using the HumanMethylation450 array. *BMC Genomics* 15:51. DOI: 10.1186/1471-2164-15-51.

Nygaard V, Rødland EA, Hovig E. 2016a. Methods that remove batch effects while retaining group differences may lead to exaggerated confidence in downstream analyses. *Biostatistics* 17:29–39. DOI: 10.1093/biostatistics/kxv027.

Nygaard V, Rødland EA, Hovig E. 2016b. Methods that remove batch effects while retaining group differences may lead to exaggerated confidence in downstream analyses. *Biostatistics* 17:29–39. DOI: 10.1093/biostatistics/kxv027.

Peters TJ, Buckley MJ, Statham AL, Pidsley R, Samaras K, V Lord R, Clark SJ, Molloy PL. 2015. De novo identification of differentially methylated regions in the human genome. *Epigenetics & Chromatin* 8:6. DOI: 10.1186/1756-8935-8-6.

Pidsley R, CC YW, Volta M, Lunnon K, Mill J, Schalkwyk LC. 2013. A data-driven approach to preprocessing Illumina 450K methylation array data. *BMC Genomics* 14:293–293. DOI: 10.1186/1471-2164-14-293.

Price EM, Robinson WP. 2018. Adjusting for Batch Effects in DNA Methylation Microarray Data, a Lesson Learned. *Frontiers in Genetics* 9. DOI: 10.3389/fgene.2018.00083.

R Core Team. 2018. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK. 2015. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic acids research* 43:e47–e47. DOI: 10.1093/nar/gkv007.

Sharp GC, Lawlor DA, Richmond RC, Fraser A, Simpkin A, Suderman M, Shihab HA, Lyttleton O, McArdle W, Ring SM, Gaunt TR, Smith GD, Relton CL. 2015. Maternal pre-pregnancy BMI and gestational weight gain, offspring DNA methylation and later offspring adiposity: Findings from the Avon Longitudinal Study of Parents and Children. *International Journal of Epidemiology* 44:1288–1304. DOI: 10.1093/ije/dyv042.

Sharp GC, Salas LA, Monnereau C, Allard C, Yousefi P, Everson TM, Bohlin J, Xu Z, Huang RC, Reese SE, Xu CJ, Baïz N, Hoyo C, Agha G, Roy R, Holloway JW, Ghantous A, Merid SK, Bakulski KM, Küpers LK, Zhang H, Richmond RC, Page CM, Duijts L, Lie RT, Melton PE, Vonk JM, Nohr EA, Williams-

DeVane CL, Huen K, Rifas-Shiman SL, Ruiz-Arenas C, Gonseth S, Rezwan FI, Herceg Z, Ekström S, Croen L, Falahi F, Perron P, Karagas MR, Quraishi BM, Suderman M, Magnus MC, Jaddoe VWV, Taylor JA, Anderson D, Zhao S, Smit HA, Josey MJ, Bradman A, Baccarelli AA, Bustamante M, Håberg SE, Pershagen G, Hertz-Picciotto I, Newschaffer C, Corpeleijn E, Bouchard L, Lawlor DA, Maguire RL, Barcellos LF, Smith GD, Eskenazi B, Karmaus W, Marsit CJ, Hivert MF, Snieder H, Fallin MD, Melén E, Munthe-Kaas MC, Arshad H, Wiemels JL, Annesi-Maesano I, Vrijheid M, Oken E, Holland N, Murphy SK, Sørensen TIA, Koppelman GH, Newnham JP, Wilcox AJ, Nystad W, London SJ, Felix JF, Relton CL. 2017. Maternal BMI at the start of pregnancy and offspring epigenome-wide DNA methylation: Findings from the pregnancy and childhood epigenetics (PACE) consortium. *Human Molecular Genetics* 26:4067–4085. DOI: 10.1093/hmg/ddx290.

Smyth GK. limma: Linear Models for Microarray Data. *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*:397–420. DOI: 10.1007/0-387-29362-0_23.

Teschendorff AE, Marabita F, Lechner M, Bartlett T, Tegner J, Gomez-Cabrero D, Beck S. 2013. A beta-mixture quantile normalization method for correcting probe design bias in Illumina Infinium 450 k DNA methylation data. *Bioinformatics* 29:189–196. DOI: 10.1093/bioinformatics/bts680.

Teschendorff AE, Zheng SC. 2017. Cell-type deconvolution in epigenome-wide association studies: a review and recommendations. *Epigenomics* 9:757–768. DOI: 10.2217/epi-2016-0153.

Thabane L, Mbuagbaw L, Zhang S, Samaan Z, Marcucci M, Ye C, Thabane M, Giangregorio L, Dennis B, Kosa D, Debono VB, Dillenburg R, Fruci V, Bawor M, Lee J, Wells G, Goldsmith CH. 2013. A tutorial on sensitivity analyses in clinical trials: the what, why, when and how. *BMC Medical Research Methodology* 13:92. DOI: 10.1186/1471-2288-13-92.

Thakali KM, Faske JB, Ishwar A, Alfaro MP, Cleves MA, Badger TM, Andres A, Shankar K. 2017. Maternal obesity and gestational weight gain are modestly associated with umbilical cord DNA methylation. *Placenta* 57:194–203. DOI: 10.1016/j.placenta.2017.07.009.

Wang T, Guan W, Lin J, Boutaoui N, Canino G, Luo J, Celedón JC, Chen W. 2015. A systematic study of normalization methods for Infinium 450K methylation data using whole-genome bisulfite sequencing data. *Epigenetics* 10:662–669. DOI: 10.1080/15592294.2015.1057384.

Wang D, Zhang Y, Huang Y, Li P, Wang M, Wu R, Cheng L, Zhang W, Zhang Y, Li B, Wang C, Guo Z. 2012. Comparison of different normalization assumptions for analyses of DNA methylation data from the cancer genome. *Gene* 506:36–42. DOI: 10.1016/j.gene.2012.06.075.

Wu MC, Joubert BR, Kuan P, Håberg SE, Nystad W, Peddada SD, London SJ. 2014. A systematic assessment of normalization approaches for the Infinium 450K methylation platform. *Epigenetics* 9:318–329. DOI: 10.4161/epi.27119.

Yousefi P, Huen K, Aguilar Schall R, Decker A, Elboudwarej E, Quach H, Barcellos L, Holland N. 2013. Considerations for normalization of DNA methylation data by Illumina 450K BeadChip assay in population studies. *Epigenetics* 8:1141–1152. DOI: 10.4161/epi.26037.

Yousefi P, Huen K, Davé V, Barcellos L, Eskenazi B, Holland N. 2015. Sex differences in DNA methylation assessed by 450 K BeadChip in newborns. *BMC Genomics* 16:911. DOI: 10.1186/s12864-015-2034-y.

Zindler T, Frieling H, Neyazi A, Bleich S, Friedel E. 2020. Simulating ComBat: how batch correction can lead to the systematic introduction of false positive results in DNA methylation microarray studies. *BMC bioinformatics* 21:271. DOI: 10.1186/s12859-020-03559-6.

Figure 1

Flowchart of Data Processing and Analysis

Combinations of data-processing and analysis choices, consisting of six normalised datasets (SQN, BMIQ or SWAN, with probe filering before or afterwards), use or non-use of ComBat processing (supervised or unsupervised), and analysis with either an unadjusted model, a model adjusted for batch, and a model adjusted for batch and cell type proportion.

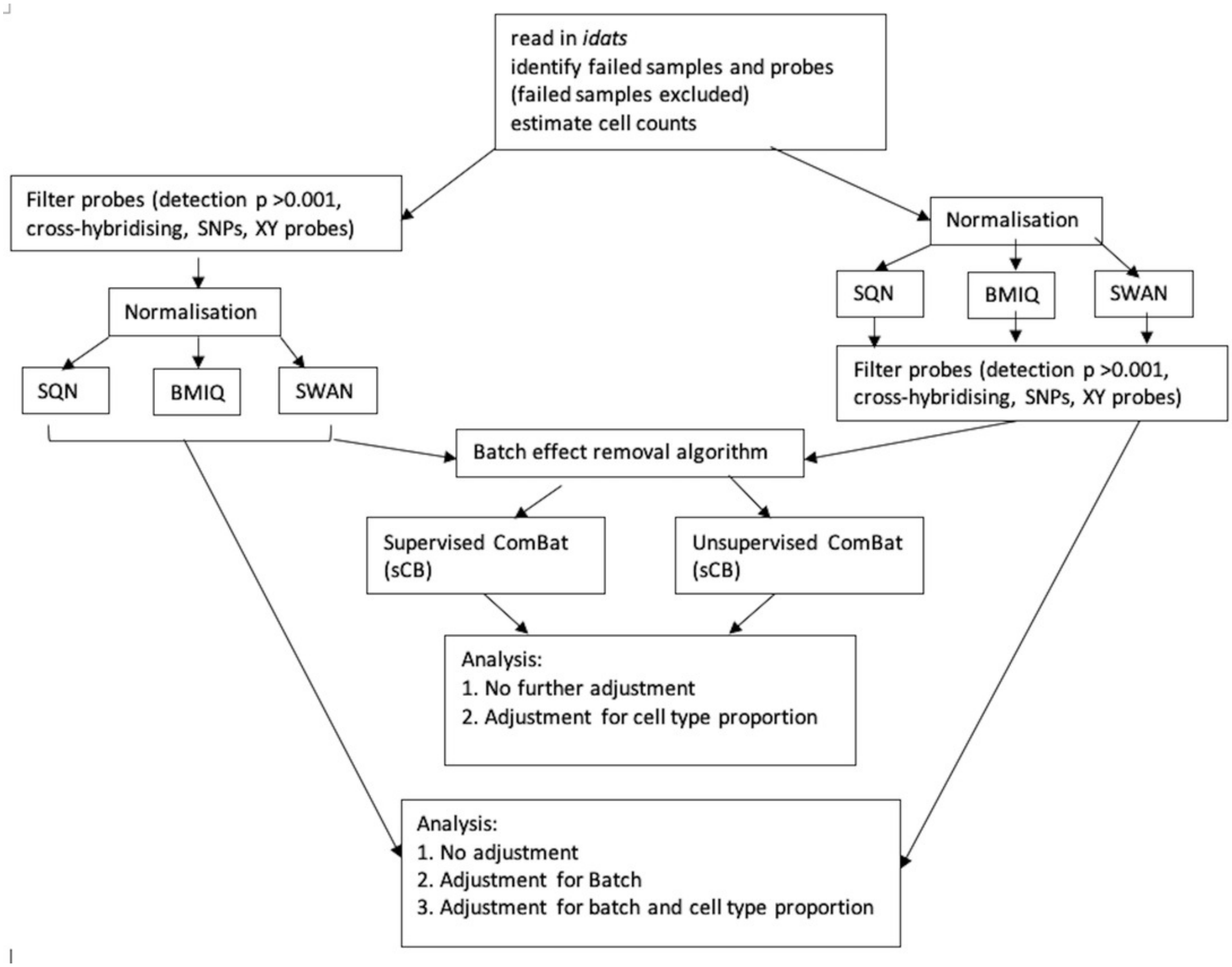
read in *idats*
identify failed samples and probes
(failed samples excluded)
estimate cell counts
Filter probes (detection p >0.001, cross-hybridising, SNPs, XY probes)
Normalisation
SQN
BMIQ
SWAN
Normalisation
SQN
BMIQ
SWAN
Filter probes (detection p >0.001, cross-hybridising, SNPs, XY probes)
Batch effect removal algorithm
Supervised ComBat (sCB)
Unsupervised ComBat (sCB)
Analysis:
1. No further adjustment
2. Adjustment for cell type proportion
Analysis:
1. No adjustment
2. Adjustment for Batch
3. Adjustment for batch and cell type proportion

Figure 2

Probes ranked in top 10 by p-value in Batch+Cell Adjusted Model, for (a) Infant Sex, (b) BMI in Standard Care, (c) Short-Haired in Tabby

For each probe the rank is given by pre- vs post-filtering, normalisation method, and batch-handling method. The model is one adjusting for batch (either explicitly in the model or via batch-correction algorithm) and cell type proportion. Adjust=adjusted for batch in the model; SCB=Supervised ComBat; UCB=Unsupervised ComBat.

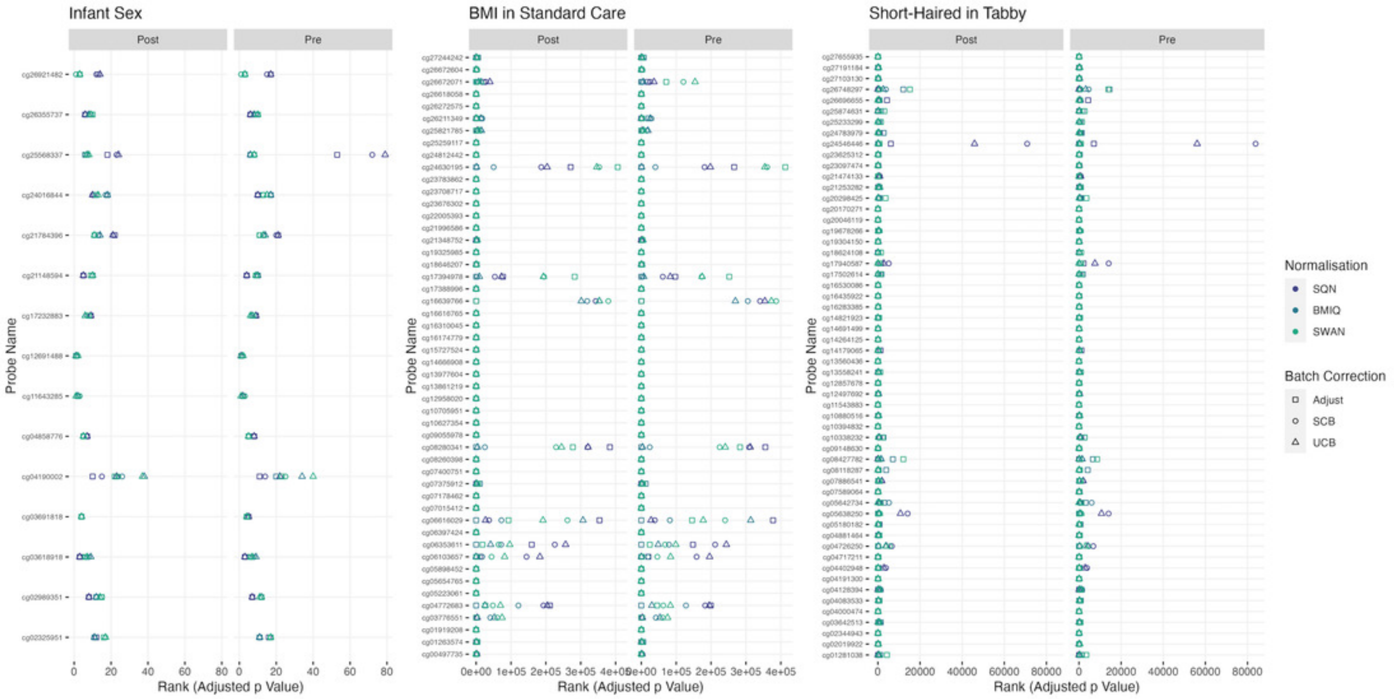

Figure 3

Top 10 Probes by LogFC: Infant Sex

Largest LogFC for Infant Sex (female), by normalisation and batch correction method

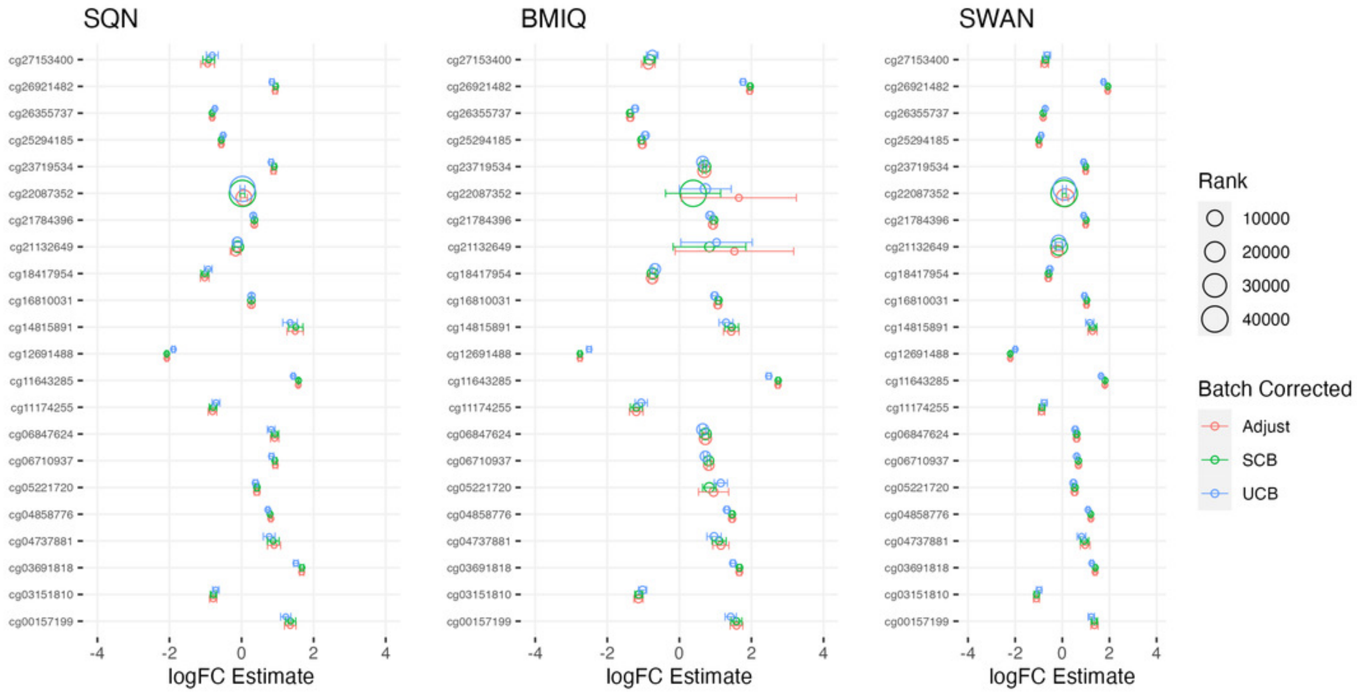

Figure 4

Top 10 Probes by LogFC: BMI in Standard Care

Largest LogFC for effect of BMI in Standard Care group, by normalisation and batch correction method

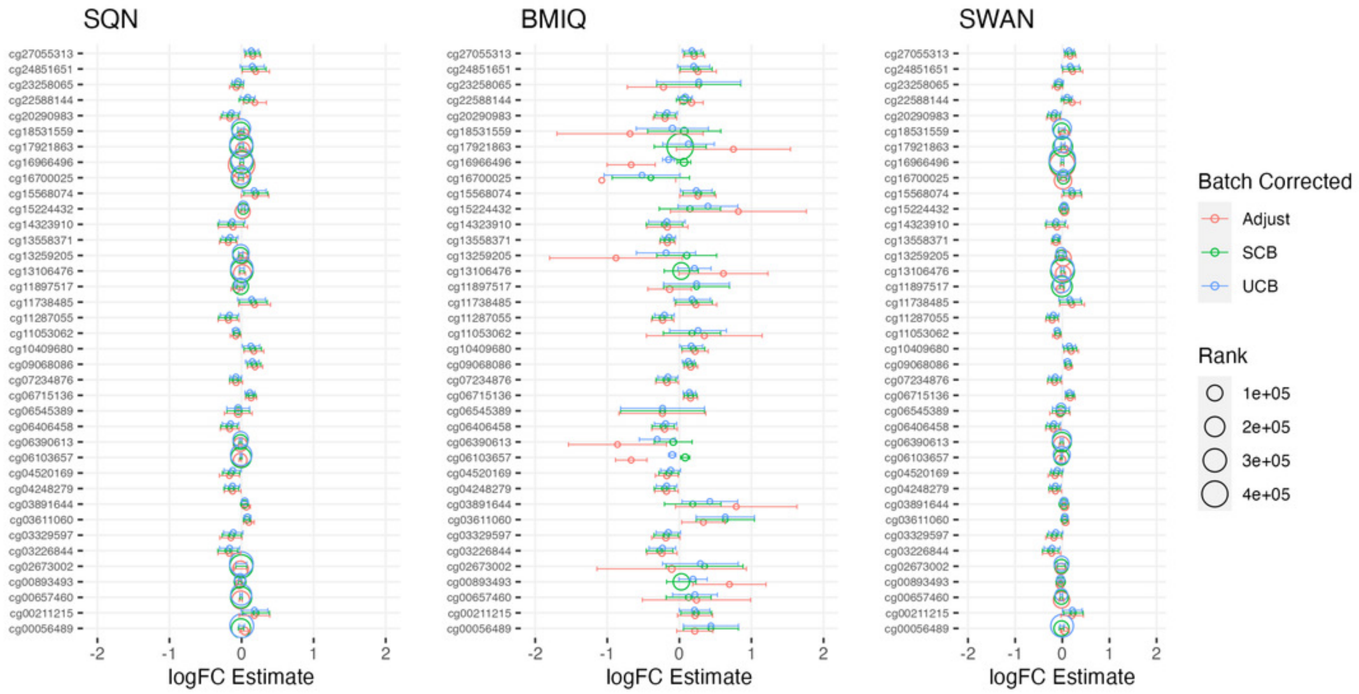

Figure 5

Top 10 Probes by LogFC: 'Short-Haired' in 'Tabby'

Largest LogFC for effect of 'Short-Haired' in 'Tabby' group, by normalisation and batch correction method.

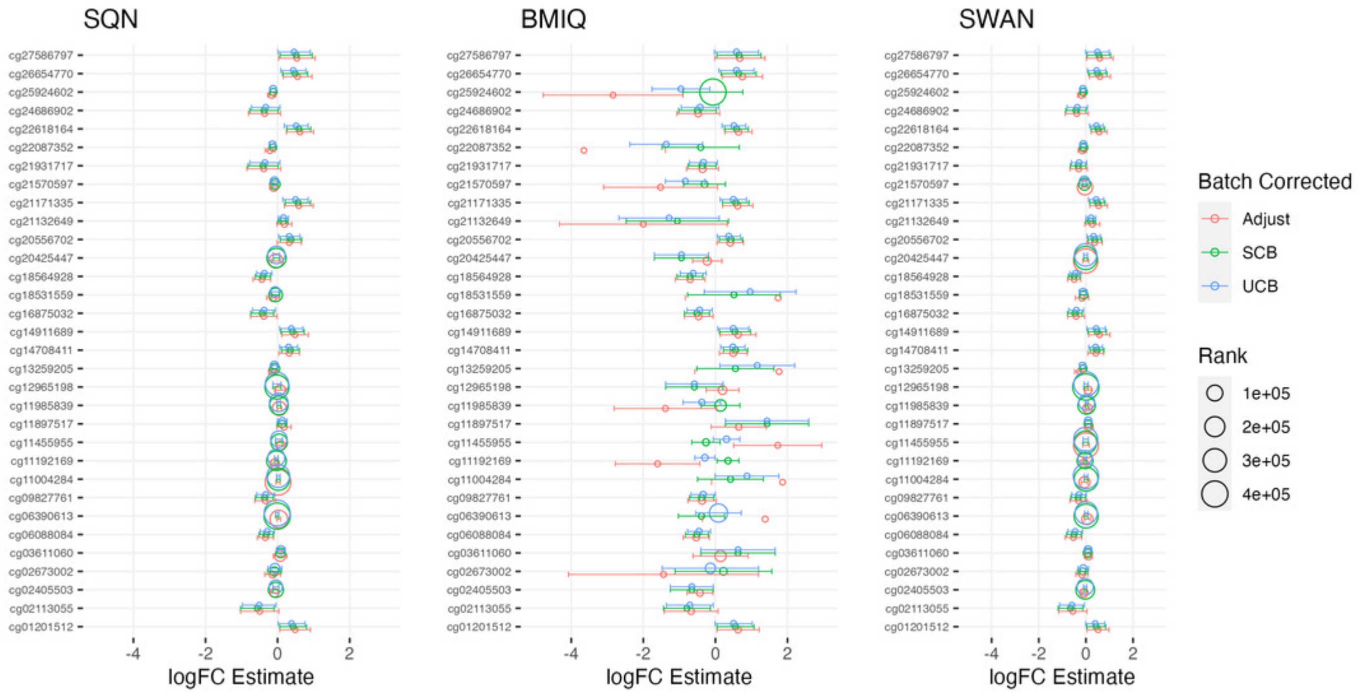

Figure 6

Distribution of Unadjusted P Values by Normalisation and Batch Correction Method, for Batch and Cell Adjusted Models

Only models from data where probe filtering was performed post-normalisation are included. The model is one adjusting for batch (either explicitly in the model or via batch-correction algorithm) and cell type proportion.

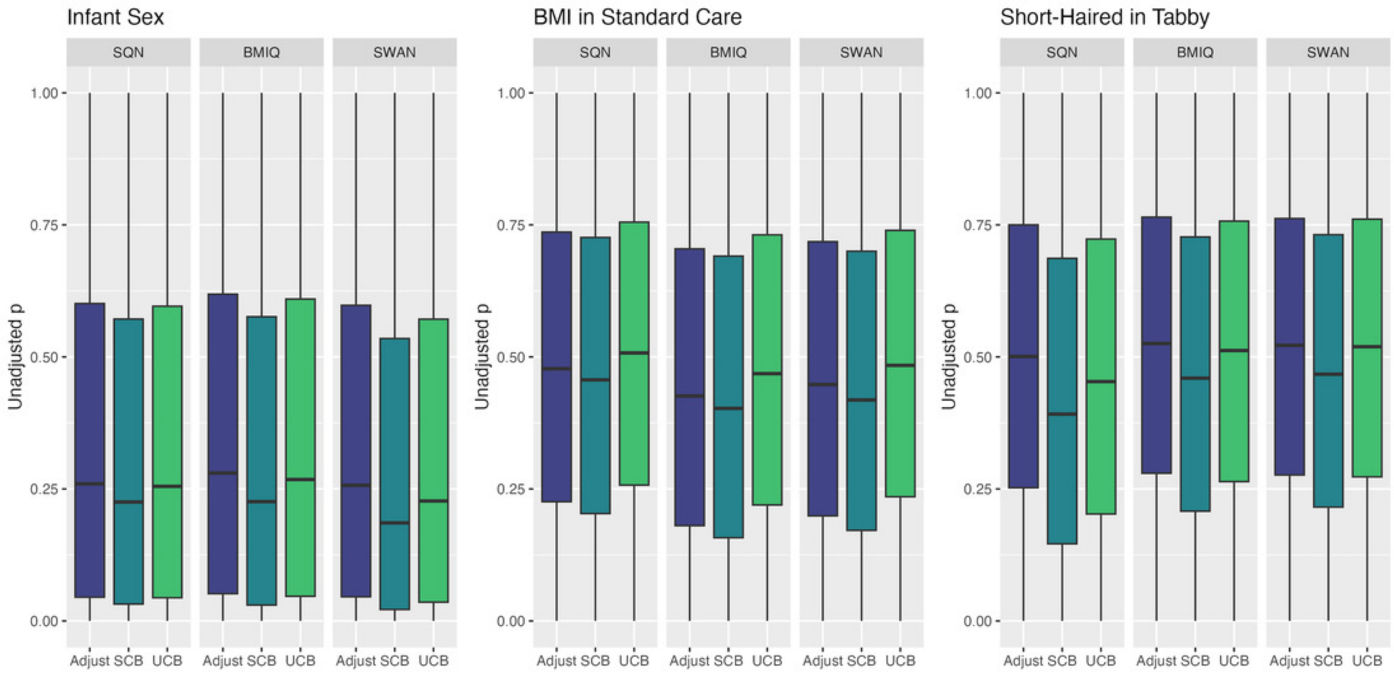

Figure 7

Distribution of Log-Fold-Change Estimates by Normalisation and Batch Correction Method, for Batch and Cell Adjusted Models

Only models from data where probe filtering was performed post-normalisation are included. The model is one adjusting for batch (either explicitly in the model or via batch-correction algorithm) and cell type proportion.

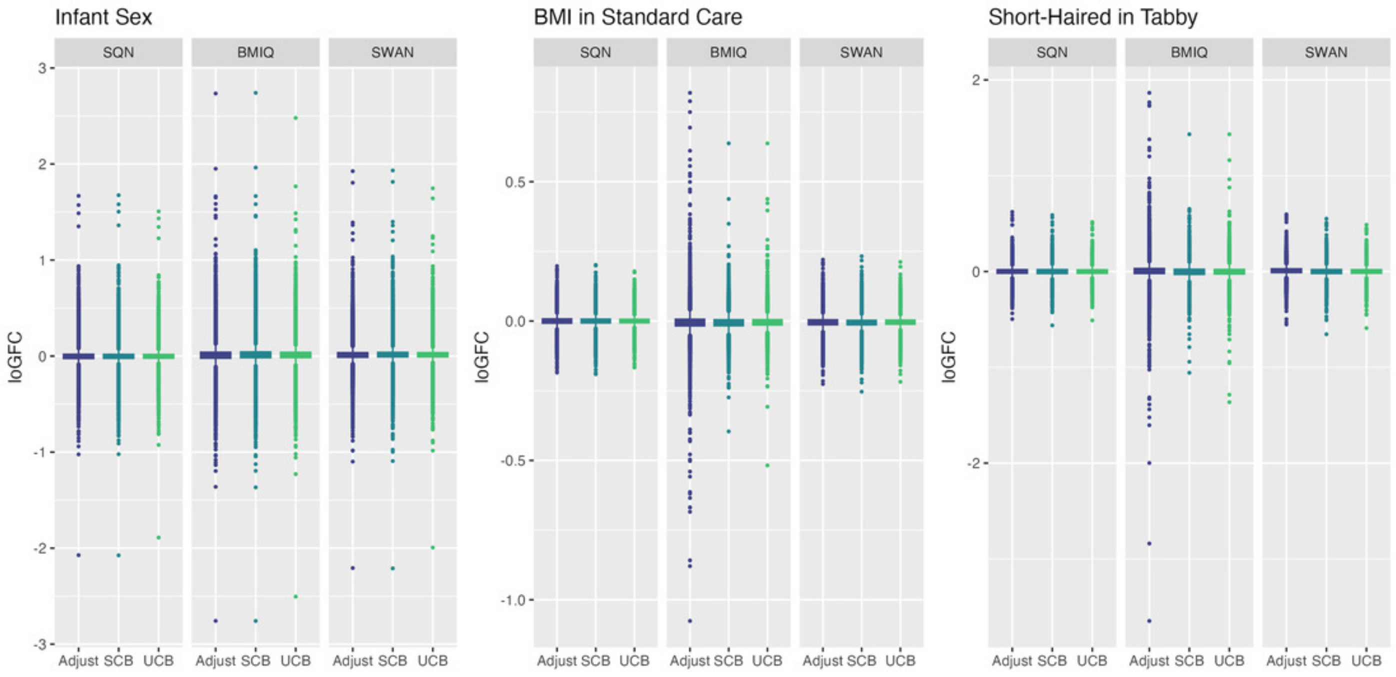

Table 1(on next page)

Number of DMPs for Infant Sex (Female), by Probe Filtering Method, Batch Correction Method, Normalisation Method and Cell Type Method

* No adjustment beyond the correction for batch as implemented in the ComBat algorithm

Table 1. Number of DMPs for Infant Sex (Female), by Probe Filtering Method, Batch Correction Method, Normalisation Method and Cell Type Method

Model		SQN		BMIQ		SWAN	
		Post Filtered	Pre Filtered	Post Filtered	Pre Filtered	Post Filtered	Pre Filtered
No ComBat							
- Unadjusted	Down	15088	14878	7935	7554	7581	6890
	Up	20587	20441	28004	30741	25784	25962
- Adjusted for Batch	Down	13132	13215	6602	6112	7100	6140
	Up	20225	19956	39482	34408	30362	29780
- Adjusted for Batch + Cell	Down	28406	28633	15855	14900	14239	10408
	Up	31973	32204	39719	45255	43155	40709
Supervised ComBat							
- Unadjusted*	Down	20967	21230	11235	10180	10772	9252
	Up	25690	25320	41518	48191	45193	43237
- Adjusted for Cell	Down	35559	36022	18036	16512	16423	12198
	Up	37068	36972	56521	64851	68769	65109
UnSupervised ComBat							
- Unadjusted*	Down	14603	14763	7634	6836	7344	6377
	Up	21336	21041	30961	34882	31892	31170
- Adjusted for Cell	Down	28012	28286	14060	12916	13030	9560
	Up	32478	32447	43370	49520	52037	49102

* No adjustment beyond the correction for batch as implemented in the ComBat algorithm

Table 2(on next page)

DMPs for Intervention and BMI Effects

Table 2 DMPs for Effect of Maternal BMI in the Standard Care group

	SQN		BMIQ		SWAN	
Model	Post Filtered	Pre Filtered	Post Filtered	Pre Filtered	Post Filtered	Pre Filtered
No ComBat						
- Unadjusted	0	0	5	6	0	0
- Adjusted for Batch	0	0	6	0	0	0
- Adjusted for Batch + Cell	0	0	8	0	6	0
Supervised ComBat						
- Unadjusted	0	0	0	10	0	0
- Adjusted for Cell	0	0	99	207	0	0
UnSupervised ComBat						
- Unadjusted	0	0	0	0	0	0
- Adjusted for Cell	0	0	0	0	0	6

Table 3(on next page)

DMPs for Fake Groups

Table 3 DMPs for Fake Groups: ‘Short-Haired’ in ‘Tabby’

	SQN		BMIQ		SWAN	
Model	Post Filtered	Pre Filtered	Post Filtered	Pre Filtered	Post Filtered	Pre Filtered
No ComBat						
- Unadjusted	0	0	0	0	0	0
- Adjusted for Batch	2180	2574	0	0	0	0
- Adjusted for Batch + Cell	0	0	0	0	0	0
Supervised ComBat						
- Unadjusted	6768	7007	3	6	8	8
- Adjusted for Cell	123	133	1	1	0	0
UnSupervised ComBat						
- Unadjusted	0	0	0	0	0	0
- Adjusted for Cell	0	0	0	0	0	0