An Effective Processing Pipeline for Harmonizing DNA Methylation Data from Illumina's 450K and EPIC Platforms for Epidemiological Studies

Lauren A Vanderlinden (lauren.vanderlinden@cuanschutz.edu)
Colorado School of Public Health https://orcid.org/0000-0002-4019-8395
Randi K Johnson
Colorado School of Public Health
Patrick M Carry
Colorado School of Public Health
Fran Dong
Colorado School of Public Health
Dawn L. DeMeo
Harvard Medical School
Ivana V Yang
University of Colorado Denver
Jill M Norris
Colorado School of Public Health
Katerina Kechris
Colorado School of Public Health

Research note

Keywords: DNA methylation, Illumina 450K, Illumina EPIC, platform harmonization

Posted Date: October 7th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-85951/v1

Abstract

Objective: Illumina BeadChip arrays are commonly used to generate DNA methylation data for large epidemiological studies. Updates in technology over time create challenges for data harmonization within and between studies, many of which obtained data from the older 450K and newer EPIC platforms. The pre-processing pipeline for DNA methylation is not trivial, and influences the downstream analyses. Incorporating different platforms adds a new level of technical variability that has not yet been taken into account by recommended pipelines. Our study evaluated the performance of various tools on different versions of platform data harmonization at each step of pre-processing pipeline, including quality control (QC), normalization, batch effect adjustment, and genomic inflation. We illustrate our novel approach using 450K and EPIC data from the Diabetes Autoimmunity Study in the Young (DAISY) prospective cohort.

Results: We found normalization and probe filtering had the biggest effect on data harmonization. Employing a meta-analysis was an effective and easily executable method for accounting for platform variability. Correcting for genomic inflation also helped with harmonization. We present guidelines for studies seeking to harmonize data from the 450K and EPIC platforms, which includes the use of technical replicates for evaluating numerous pre-processing steps, and employing a meta-analysis.

Introduction

Numerous epidemiological studies have examined DNA methylation due to its important role in physiological processes, and the development and progression of human diseases (1). Microarrays have been widely used for DNA methylation profiling, and are affordable for studies with large sample sizes. In particular, Illumina's methylation array is a common choice in many data repositories such as TCGA with ~12,000 samples, ENCODE with ~250 datasets and GEO with ~7,000 datasets (April 2020).

The DNA methylation array technologies have evolved so more individual methylation CpG sites can be evaluated on a single array. Illumina's BeadChip methylation microarrays are extremely popular, the most recent being the HumanMethylationEPIC BeadChip ("EPIC") released in 2016 that measures ~850,000 CpG sites, which is an increase from the previous array (HumanMethylation450K BeadChip, "450K"). In many studies, both the EPIC and 450K have been used, whether due to technology updates in the middle of large projects, or multiple batches for large studies over time (2–6). Some investigators are interested in maximizing sample size for research questions by analyzing data generated on the current EPIC array and the older 450K array. Numerous publications focus on certain aspects of a pre-processing pipeline such as normalization (7) or probe filtering (8) but there are currently no papers that consider the entire pipeline. Establishing best practices is relevant for other large epidemiological studies that needed to change platforms mid study, in addition to the re-analyses of public data on TCGA, ENCODE and GEO.

In this work we evaluate the performance of common tools on harmonization of 450K and EPIC data at various pre-processing and analytical steps using the Diabetes Autoimmunity Study in the Young

(DAISY), which prospectively follows genetically high-risk children for the development of type 1 diabetes (T1D) (4). We explored normalization, probe level QC and filtering, batch effect adjustment, and genomic inflation by testing methods that were easy to implement from well-established and documented R packages. Finally, we provide evaluation guidelines for studies facing similar harmonization challenges.

Methods

Figure 1 shows a summary of the pre-processing pipeline.

Data

Peripheral whole blood was collected prospectively from individuals enrolled in DAISY. Cases of T1D were frequency matched to controls, and DNA methylation generated on up to five prospective samples per subject (4). The EPIC platform replaced the 450K platform during data acquisition. There were 42/42 and 45/45 matched T1D cases/controls (corresponding to 184 and 211 unique arrays) for the 450K and EPIC platforms, respectively.

Normalization and Probe-level QC

First, three well established normalization and probe-level QC methods were evaluated (Fig. 1): subset-quantile within array normalization (SWAN, (9)), normal-exponential using out-of-band probes (Noob, (10)) and single-sample Noob (ssNoob,(11)). The SWAN and ssNoob normalizations were performed within the minfi package (12), while the Noob normalization was performed in the SeSAMe package (7). We examined two different detection above background methods: minfi's default (Aryee et al., 2014) and SeSAMe's pooBAH (7) and coupled it to the normalization that was in the same R package. Filtering on probe-level QC was performed after each normalization prior to evaluation of platform effects.

For evaluating normalization and probe-level QC procedures, we looked at the first 10 principal components (PCs) to determine if there was a large platform effect across components, as well as the three technical replicate metrics mentioned below.

A total of 12 technical replicates were selected to be balanced with respect to sex, age, and islet autoimmunity (IA) status. To examine this data we used three metrics: 1) a difference in methylation Beta value at the individual CpG site (Eqn S1), 2) correlation of all CpG sites across a single technical replicate pair (Eqn S2) and 3) correlation of the technical replicate pairs across a single CpG site (Eqn S3, see Supplementary Methods).

Batch Effect Adjustment

Second, we applied two different lower order (within platform) batch effect methods: ComBat (13) and RUVm (14) (Fig. 1) in the sva package (v3.30.0) and missMethyl package (v1.16.0) respectively.

Additional Probe Filtering

Third, we explored supplemental probe filtering by removing low methylation range probes (< 5% Beta) as suggested by (8) and compared how well the within 450K and across platform technical replicates correlated.

Statistical Analysis

We performed a linear mixed model using T1D status to predict methylation (M-values) while adjusting for age and sex. We used an autoregressive 1 covariance structure to model the repeated subject measures. The model was run using the R/nlme package (v3.1-137) (15).

Genomic Inflation

We corrected for genomic inflation using the R/BACON package (v1.10.1) (16). In brief, this method estimates an empirical null distribution using a Bayesian method to account for the bias and inflation of test-statistics specific to EWAS datasets.

Meta-Analysis

Stouffer's meta-analysis method (17) was used to combine the statistical results from the two different platforms. This method generates a single meta-analysis p-value for each CpG site, and accounts for consistent direction of effects.

Results

For each of the processing Steps 1–4, we compared different options with a variety of evaluation diagnostics and leveraged the technical replicates (Fig. 1).

Normalization Evaluation

First, we explored normalization of both data sets together using ssNoob (coupled with the minfi probe QC), as recommended by Fortin et al (11). After associating the first 10 PCs with platform, we found that the first and second PC had an extremely high association with platform and sex respectively (Fig. 2). Sex differences are expected to be a large contributor to methylation profiles as methylation is known to have a large role in X-chromosome inactivation in females. Applying subsequent batch adjustment did not

reduce the strong platform effect (Supplemental Table S1), regardless of batch adjustment method applied. Therefore, we applied normalization procedures by each platform separately.

To explore the effect of SWAN or SeSAMe on harmonization of the platforms, we examined technical replicates across platforms (see Methods). The correlation across probes for each pair of technical replicates (Eqn S2) was extremely high (> 0.98) for both methods. This is not surprising given the large amount of data points used to calculate each correlation, and similar to the high correlation between random pairs of samples (> 0.97). Individual probe correlations deemed much more informative (Eqn S3). We generated the densities of probe-level correlations across the technical replicate pairs as well as across random sample pairs (Fig. 3). The distribution of the random sample correlations for both the SeSAMe and SWAN are centered around 0 and look more like a Gaussian distribution compared to the distributions for the technical replicate correlations, which look like a mixture of two or more distributions in addition to being centered around a higher correlation coefficient.

We also examined the absolute differences in methylation on the probe level (Eqn S1) of the Beta value (% methylation) for each technical replicate sample (Supplemental Figure S1). In all technical replicate pairs, SeSAMe has a tighter distribution closer to 0 and based on these differences is shown to harmonize the data better, while SWAN has higher absolute differences. Given the results on the technical replicates, we moved forward with the SeSAMe normalization (see supplement for probe QC filtering numbers).

Batch Effect Adjustment

Even normalizing within each platform type, we still have technical batch effects to consider, since a variety of factors can add unwanted technical variation (18). Therefore, we examined within platform batch effect being defined as plate and row location combination. We performed ComBat and RUVm to adjust for within platform batch effects on the SeSAMe dataset. Based on the PC analysis, ComBat performed slightly better, as the top PCs were less associated with batches defined as plate by rows (Supplemental Figure S2, (18)). Our results of ComBat outperforming other methods is consistent with Jiao and colleagues (19).

Probe Filtering

Applying the Logue beta range filter (8), removing probes with < 5% Beta methylation range, resulted in removing 15.8% (59,397) and 33.9% (225,342) of probes in the 450K and EPIC platforms respectively. The mean beta values for the probes which were removed fell only on the extremes for both platforms (Supplemental Figure S3), while the probes which passed this criteria had mean beta values throughout the potential 0-100% methylation values. Additional considerations are summarized in the Supplement.

Genomic Inflation in Statistical Analysis

After statistical analysis, it's important to consider the genomic inflation factor lambda (i.e., general inflation of test statistics due to population structure), which is the ratio of the median of the observed distribution of the test statistics to the expected median, and should be close to 1. In the 450K platform, the SWAN normalized dataset resulted in an extreme genomic inflation were the EPIC was deflated (Supplmental Figure S6). However, the genomic inflation was very comparable between the platforms for the SeSAMe normalized datasets. To account for any additional genomic bias and inflation we applied BACON (16), which was developed to control for genomic inflation specifically for epigenomic data. After this adjustment, the genomic inflation factor for the SeSAMe 450K and EPIC datasets were 1.03 and 1.08 respectively.

To perform the meta-analysis, we only kept probes present in both the SeSAMe datasets (199,243 probes). Final results is reported by R. K. Johnson and colleagues (4). This final pipeline as it gives comparable candidate CpG sites to other DNA methylation papers in T1D (20–23).

Discussion

Pre-processing any 'omics dataset including Illumina's BeadChip array can have substantial effects on downstream analyses. The introduction of an updated array adds the additional hurdle of harmonizing more than one platform to leverage all available data. If possible, we recommend including technical replicates in the study design to aid in assessing the quality of pre-processing steps as it was key for our evaluation process of the various methods. Supplemental Figure S7 summarizes our recommended best practices based on the tested approaches. We realize new methods are constantly evolving in this field, and this flow chart aims to help guide analysts in what decisions need to be made throughout this process.

There are some special considerations regarding filtering of probes, which is performed at two stages. The first stage is based on low quality probes identified after normalization, and the second stage is before statistical modeling based on removing probes that are not varying. In the first probe filtering after normalization, the pooBAH method (part of the SeSAMe pipeline) removed a high number of probes compared to that of the SWAN, specifically those on sex chromosomes. However, the resulting genomic inflation values were much more consistent among the platforms and closer to one, which is desired. This suggests that pooBAH is correctly identifying germline and somatic deletions that would be causing this inflated signal. However, the use of pooBAH filtering on sex chromosomes should be considered with caution.

In addition to a meta-analysis, we explored an alternative approach, where data was pooled into one statistical model and including a fixed covariate of platform (categorical variable) and subsequently adjusting for genomic inflation after this single model. For us, results of the single model were similar to

the meta-analysis (results not shown) and provided little evidence to support one method over the other. We recommend that both approaches be considered depending on the type of statistical analysis.

In summary, our evaluation methods relied on technical replicates on both platforms, which we highly recommend. The evaluation metrics on the technical replaces were used to compare methods at different steps, and can be used to evaluate other options as new methods are developed. We hope our guidelines aid others in their endeavors for performing analyses consisting of both 450K and EPIC platforms.

Limitations

Others have explored individual steps in this pipeline (8, 11), therefore we did not examine individual steps in depth using multiple datasets or simulations. The goal of this work was to evaluate the entirety of steps involved in a methylation processing pipeline based on data from both Illumina's 450K and EPIC platforms. We do not claim that our recommended pipeline will be best in all scenarios, but illustrate what factors need to be considered for selecting a pipeline with other data sets, and new methods as they are published.

Abbreviations

QC – quality control

DAISY - Diabetes Autoimmunity Study in the Young

T1D - type 1 diabetes

SWAN - subset-quantile within array normalization

Noob - normal-exponential using out-of-band probes

ssNoob - single-sample Noob

IA - islet autoimmunity

PC - principal component

Declarations

Ethics approval and consent to participate

Data collected from the DAISY study is in accordance with the Declaration of Helsinki and the Colorado Multiple Institutional Review Board approved all study protocols. Informed consent (via written informed consent form) was obtained from the parents of each study subject.

Consent for publication

Not applicable

Availability of data and materials

The datasets generated during and/or analysed during the current study are accessible through GEO Series accession number GSE142512 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE142512). All data generated for this study are included in Johnson R.K. and colleagues (4). This manuscript was based on the preprocessing for the analyses used in the Johnson R.K. publication (4).

Competing interests

The authors declare that they have no competing interests.

Funding

We would like to thank the generous contributions made by all participants and families enrolled in the DAISY study as well as the dedicated clinical and research staff at the Barbara Davis Center. This work was funded by NIH R01-DK104351, R01-DK32493.

Authors' contributions

All authors have made a substantial contributions to the conception of the work, interpretation of the data, critical revisions of the manuscript and approved the final draft. L.A.V and R.K.J. were in involved in the analyses of the data. L.A.V., R.K.J, and K.K. have substraintivey drafted and revised the manuscript.

Acknowledgements

We would like to thank everyone involved in the IVYOmics analysis group.

References

1. Jin Z, Liu Y. DNA methylation in human diseases. Genes Dis. 2018 Mar;5(1):1–8.
2. Abdulrahim Jawan W., Kwee Lydia Coulter, Grass Elizabeth, Siegler Ilene C., Williams Redford, Karra Ravi, et al. Epigenome-Wide Association Study for All-Cause Mortality in a Cardiovascular Cohort Identifies Differential Methylation in Castor Zinc Finger 1 (CASZ1). J Am Heart Assoc. 2019 Nov 5;8(21):e013228.

3. Fernandez-Jimenez N, Allard C, Bouchard L, Perron P, Bustamante M, Bilbao JR, et al. Comparison of Illumina 450K and EPIC arrays in placental DNA methylation. Epigenetics. 2019 Dec 2;14(12):1177–82.
4. Johnson RK, Vanderlinden LA, Dong F, Carry PM, Seifert J, Waugh K, et al. Longitudinal DNA methylation differences precede type 1 diabetes. Sci Rep [Internet]. 2020 Feb 28 [cited 2020 Mar 11];10. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7048736/
5. McEwen LM, Jones MJ, Lin DTS, Edgar RD, Husquin LT, MacIsaac JL, et al. Systematic evaluation of DNA methylation age estimation with common preprocessing methods and the Infinium MethylationEPIC BeadChip array. Clin Epigenetics. 2018 Oct 16;10(1):123.
6. Solomon O, MacIsaac J, Quach H, Tindula G, Kobor MS, Huen K, et al. Comparison of DNA methylation measured by Illumina 450K and EPIC BeadChips in blood of newborns and 14-year-old children. Epigenetics. 2018 Aug 15;13(6):655–64.
7. Zhou W, Triche TJ, Laird PW, Shen H. SeSAMe: reducing artifactual detection of DNA methylation by Infinium BeadChips in genomic deletions. Nucleic Acids Res. 2018 16;46(20):e123.
8. Logue MW, Smith AK, Wolf EJ, Maniates H, Stone A, Schichman SA, et al. The correlation of methylation levels measured using Illumina 450K and EPIC BeadChips in blood samples. Epigenomics. 2017;9(11):1363–71.
9. Maksimovic J, Gordon L, Oshlack A. SWAN: Subset-quantile Within Array Normalization for Illumina Infinium HumanMethylation450 BeadChips. Genome Biol. 2012;13(6):R44.
10. Triche TJ, Weisenberger DJ, Van Den Berg D, Laird PW, Siegmund KD. Low-level processing of Illumina Infinium DNA Methylation BeadArrays. Nucleic Acids Res. 2013 Apr 1;41(7):e90–e90.
11. Fortin J-P, Triche TJ, Hansen KD. Preprocessing, normalization and integration of the Illumina HumanMethylationEPIC array with minfi. Bioinforma Oxf Engl. 2017 15;33(4):558–60.
12. Aryee MJ, Jaffe AE, Corrada-Bravo H, Ladd-Acosta C, Feinberg AP, Hansen KD, et al. Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. Bioinforma Oxf Engl. 2014 May 15;30(10):1363–9.
13. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostat Oxf Engl. 2007 Jan;8(1):118–27.
14. Maksimovic J, Gagnon-Bartsch JA, Speed TP, Oshlack A. Removing unwanted variation in a differential methylation analysis of Illumina HumanMethylation450 array data. Nucleic Acids Res. 2015 Sep 18;43(16):e106.
15. Davidian M, Giltinan DM. Nonlinear models for repeated measurement data: An overview and update. J Agric Biol Environ Stat. 2003 Dec 1;8(4):387.
16. van Iterson M, van Zwet EW, BIOS Consortium, Heijmans BT. Controlling bias and inflation in epigenome- and transcriptome-wide association studies using the empirical null distribution. Genome Biol. 2017 27;18(1):19.
17. Stouffer SA, Suchman EA, Devinney LC, Star SA, Williams Jr. RM. The American soldier: Adjustment during army life. (Studies in social psychology in World War II), Vol. 1. Oxford, England: Princeton

Univ. Press; 1949. xii, 599 p. (The American soldier: Adjustment during army life. (Studies in social psychology in World War II), Vol. 1).

18. Price EM, Robinson WP. Adjusting for Batch Effects in DNA Methylation Microarray Data, a Lesson Learned. Front Genet. 2018;9:83.
19. Jiao C, Zhang C, Dai R, Xia Y, Wang K, Giase G, et al. Positional effects revealed in Illumina methylation array and the impact on analysis. Epigenomics. 2018;10(5):643–59.
20. Belot M-P, Nadéri K, Mille C, Boëlle P-Y, Benachi A, Bougnères P, et al. Role of DNA methylation at the placental RTL1 gene locus in type 1 diabetes. Pediatr Diabetes. 2017;18(3):178–87.
21. Disanto G, Vcelakova J, Pakpoor J, Elangovan RI, Sumnik Z, Ulmannova T, et al. DNA methylation in monozygotic quadruplets affected by type 1 diabetes. Diabetologia. 2013 Sep;56(9):2093–5.
22. Rakyan VK, Beyan H, Down TA, Hawa MI, Maslau S, Aden D, et al. Identification of type 1 diabetes-associated DNA methylation variable positions that precede disease diagnosis. PLoS Genet. 2011 Sep;7(9):e1002300.
23. Stefan M, Zhang W, Concepcion E, Yi Z, Tomer Y. DNA methylation profiles in type 1 diabetes twins point to strong epigenetic effects on etiology. J Autoimmun. 2014 May;50:33–7.

Figures

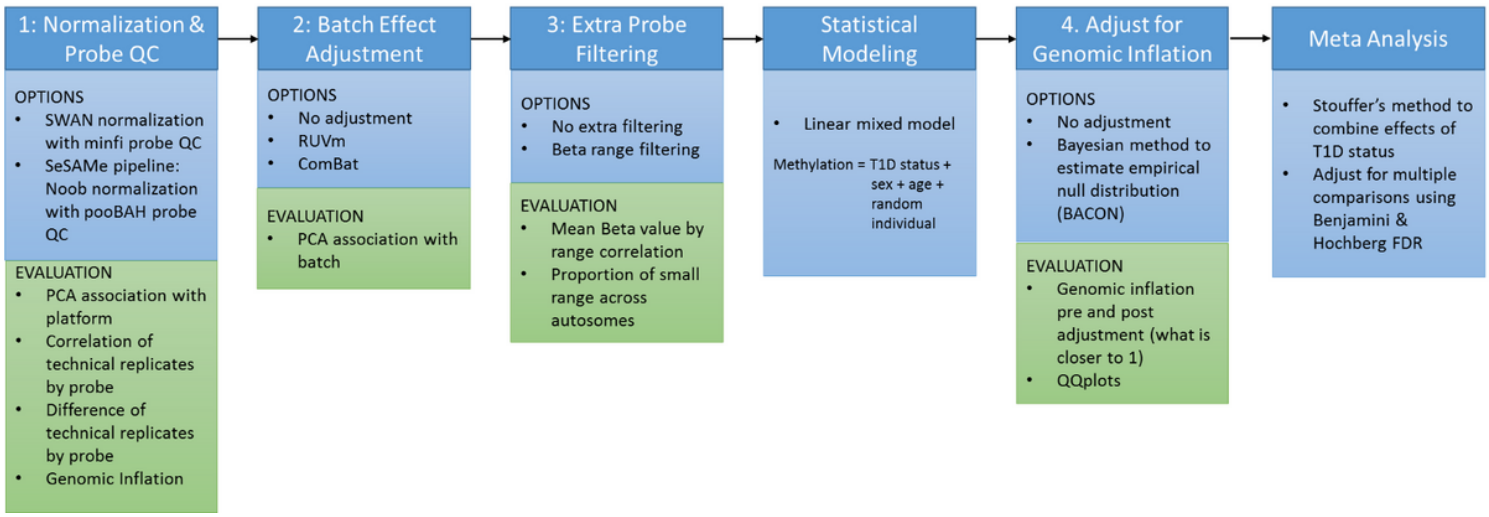

Figure 1

Pipeline Methods Considered. The four main pre-processesing steps are: 1. normalization and probe QC, 2. batch effect adjustment, 3. extra probe filtering and 4. genomic inflation adjustment. The various methods considered for each step is listed along with the evaluation(s) used to assess these methods.

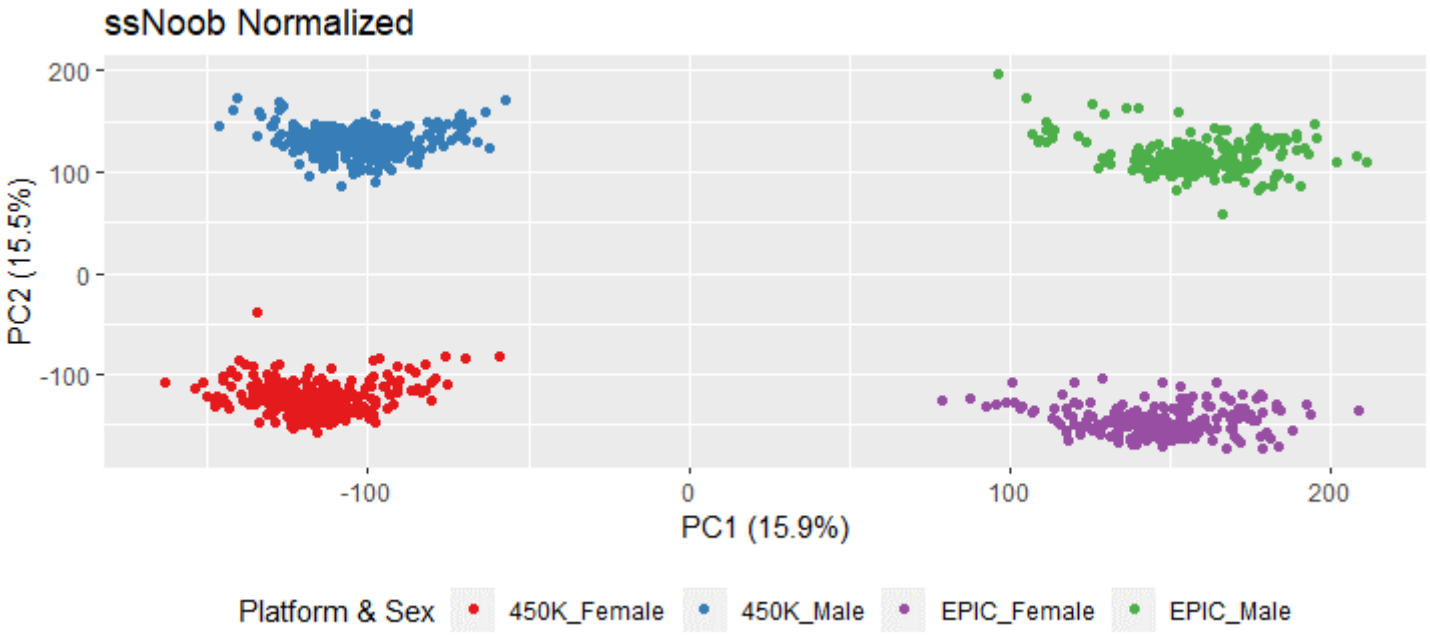

Figure 2

Platform Effect. The 1st and 2nd principal components (PCs) from the ssNoob normalization are plotted with colors symbolizing both platform and sex. Red and blue dots signify the 450K platform while purple and green dots signify the EPIC data. Red and purple dots signifiy females and blue and green dots signify males. Percent variance explained by each PC is noted in parentheses

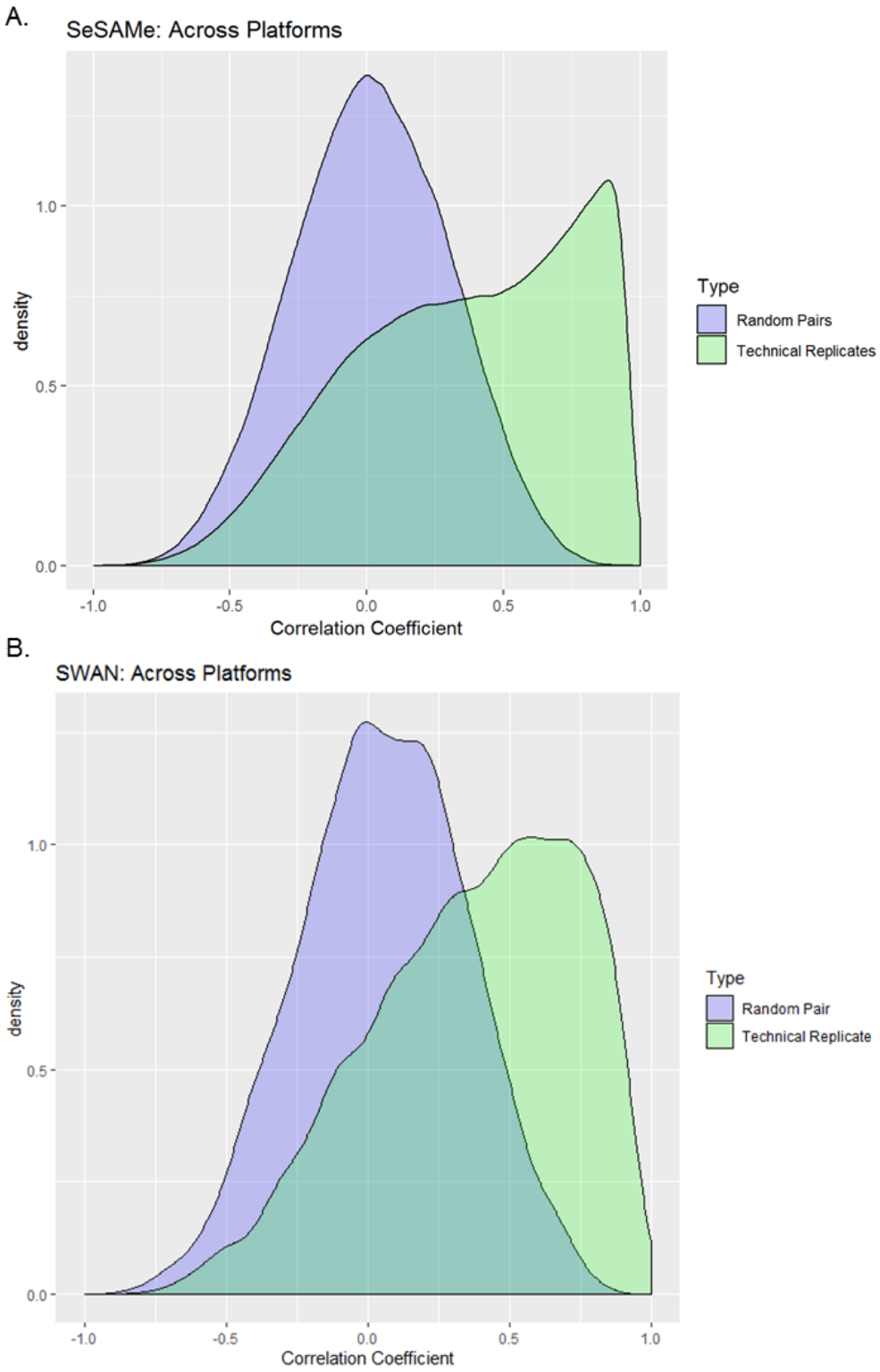

Figure 3

Correlation of Technical Replicates. Density plots of correlations across the platform technical replicates for each probe (n=12, green) as well as a random subset of pairs for comparison (n=12, purple) for the data normalized using A. SeSAMe and B. SWAN. The median correlation coefficient among technical replicates is both 0.41 in the SeSAMe and SWAN methods. The 1st and 3rd quartiles for technical replicates for SeSAMe and SWAN were (0.06, 0.72) and (0.11, 0.67) respectively.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- methylationPipelinePaperSupplementaryv10.docx